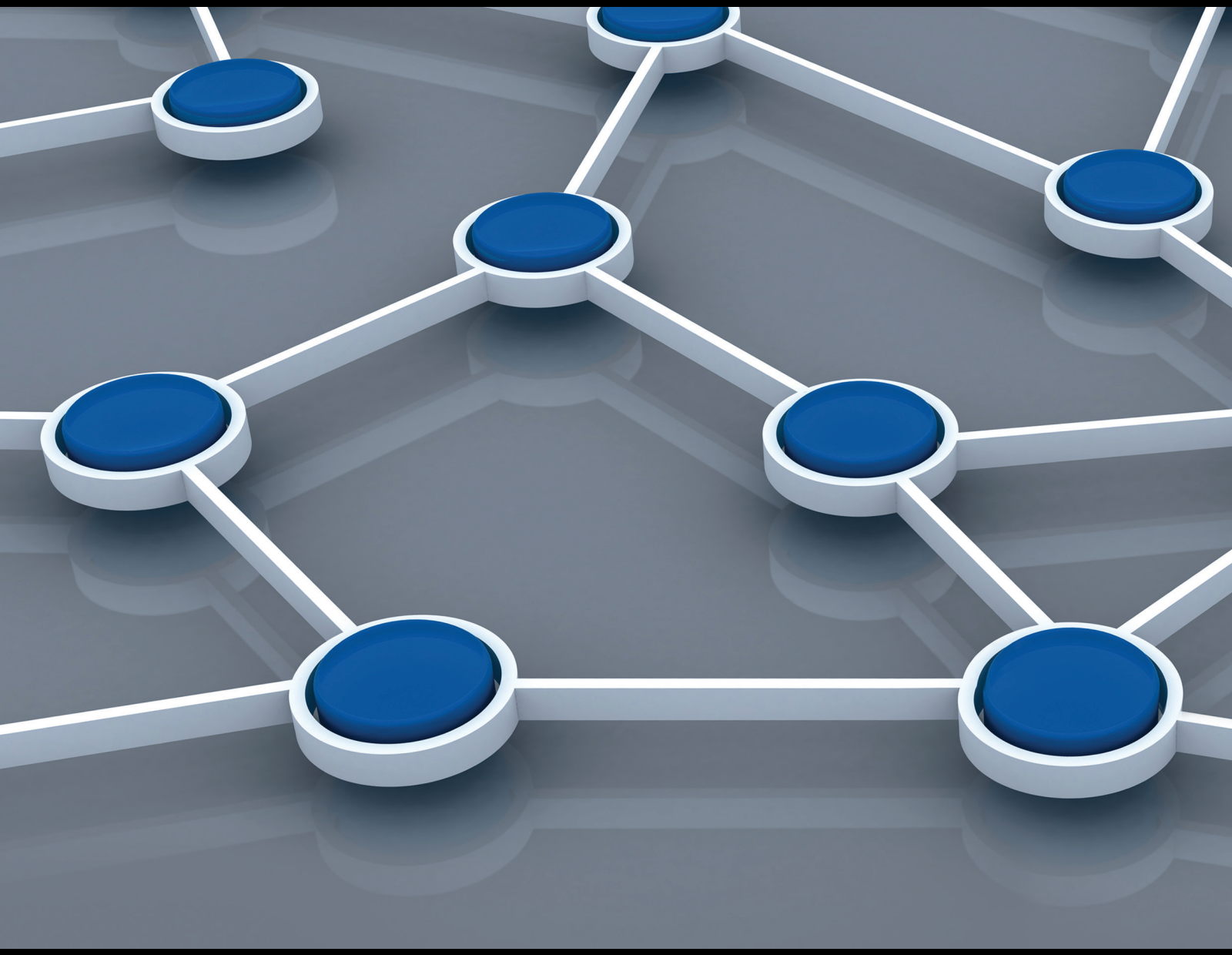
International Journal of Distributed Sensor Networks

Advanced Convergence Technologies and Practices for Wireless Ad Hoc and Sensor Networks

Guest Editors: Jongsung Kim, Ken Choi, and Wook Choi



Advanced Convergence Technologies and Practices for Wireless Ad Hoc and Sensor Networks

International Journal of Distributed Sensor Networks

Advanced Convergence Technologies and Practices for Wireless Ad Hoc and Sensor Networks

Guest Editors: Jongsung Kim, Ken Choi, and Wook Choi

Editorial Board

Contents

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 512985, 1 page
http://dx.doi.org/10.1155/2014/512985

Editorial

Advanced Convergence Technologies and Practices for Wireless Ad Hoc and Sensor Networks

Jongsung Kim,[1] Ken Choi,[2] and Wook Choi[3]

[1] *Kookmin University, 77 Jeongneung-ro, Seongbuk-gu, Seoul 136-702, Republic of Korea*
[2] *Illinois Institute of Technology (IIT), 3300 South Federal Street, Chicago, IL 60616-3793, USA*
[3] *S.LSI S/W Solution Development Team, Samsung Electronics, 2nd Floor DSR B Tower, Hwaseong-si, Gyeonggi-do 445-701, Republic of Korea*

Correspondence should be addressed to Jongsung Kim; jongsungkim02@gmail.com

Received 26 June 2014; Accepted 26 June 2014; Published 13 August 2014

Recently, the massive development and deployment of ubiquitous computing system are expected to provide numerous ubiquitous applications (e.g., smart grid, intelligent building, healthcare, automotive, etc.). A ubiquitous system is necessarily distributed, thus requiring the concepts and technologies of wireless ad hoc and sensor networks to support it. Over the past decade, tremendous technological advances have been made in the fields of wireless ad hoc and sensor networks; however, traditional techniques are not sufficient to accommodate a variety number of ubiquitous applications and services in the right way. In this regard, the aim of this special issue is to foster state-of-the-art research in the development of communications, computing, and security technologies for wireless ad hoc and sensor networks. Particularly, this special issue shows the most recent advancements in the variety type of technological integration of machine-to-machine communication, cloud computing, embedded system, with wireless ad hoc, and sensor network domain.

Acknowledgment

The guest editors are thankful to our reviewers for their effort in reviewing the papers.

Jongsung Kim
Ken Choi
Wook Choi

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 679780, 9 pages
http://dx.doi.org/10.1155/2014/679780

Research Article

Route Prediction Based Vehicular Mobility Management Scheme for VANET

DaeWon Lee,[1] Yoon-Ho Kim,[2] and HwaMin Lee[3]

[1] *Division of General Education, Seokyeong University, Seoul 136-704, Republic of Korea*
[2] *Division of Computer Engineering, Mokwon University, Daejeon 302-318, Republic of Korea*
[3] *Department of Computer Software & Engineering, Soonchunhyang University, Asan, Chungnam 336-745, Republic of Korea*

Correspondence should be addressed to HwaMin Lee; leehm@sch.ac.kr

Received 29 December 2013; Accepted 14 May 2014; Published 16 July 2014

Academic Editor: Jongsung Kim

Since improvement of wireless communication, IP based mobility management protocols have been studied to provide seamless communication and mobility management. The vehicular ad hoc network (VANET) is one of mobility management protocols, especially providing seamless connection with inter/intra/inner vehicle communication. However, each vehicle moves fast that causes short-lived connections with Access Router (AR). Based on vehicles' characteristic, it is hard to provide the availability of IP services in VANET. The most critical issue of the design of scalable routing algorithm is to provide robustness of frequent path disruption caused by vehicles' mobility. In this paper, we pursue the characteristics of vehicles' mobility and analyze them. With the navigation information which is one of vehicles' mobility characteristics, we classify the mobility into intrahighway mobility and global mobility management. Furthermore, we propose mobility management scheme based on route prediction in VANET. Handoffs with intrahighway mobility are managed locally and transparency is provided to CHs, while global mobility is managed with Mobile IPv6. Finally, through the numerical analysis, we show that proposed mobility management protocol reduces handoff latency, signaling costs, and packet loss.

1. Introduction

Nowadays, by improved wireless communication technique, wireless node and sensor are rapidly increased. Variable types of mobile devices which are provided with wireless interface can access Internet anytime and anywhere. IETF proposed IP based mobility management protocols to provide seamless communication and mobility management [1–3]. Depending on each device and environment, several types of mobility are presented (personal mobility, network mobility, sensor mobility, vehicles' mobility, and so on).

The start point of mobility management is the personal mobility. The IETF designed Mobile IPv6 (MIPv6) between wireless IPv6 networks. In MIPv6, mobile nodes are possible to access wireless IPv6 networks without changing their IP address. However, if mobile host (MH) moves frequently, MIPv6 results in high handoff latency and high signaling costs to update the MH's location [1]. Therefore, many mobility management protocols have been proposed to improve handoff performance and reduce signaling overhead. And, the other problem is that if several MHs move to identical route continuously as a group, MIPv6 results in same handoff latency and same signaling costs for each MH. To solve this problem, the IETF NEMO working group proposed network mobility basic support protocol [2] extending MIPv6. Network mobility (NEMO) is designed to support the movement of a mobile network consisting of several mobile nodes where nodes move together as a group, as in a train, car, plane, ship, and so forth. To manage mobility of sensors, ZigBee [3] is proposed that is non-IP layer protocol and TCP/IP is not used. However, sensor networks consisting of too many nodes can be connected to other devices via the wireless communication network. Therefore, an efficient addressing mechanism is needed to communicate with each sensor and wireless node in the network. The IETF IPv6 over low power WPAN (6LoWPAN) working group [4] was organized to define the IPv6 transmission packets over IEEE 802.15.4 [5]. In 6LoWPAN, each node is assigned a global IPv6 address. So, external

IPv6 hosts are able to communicate with sensor nodes in 6LoWPAN [6]. And, the vehicles' mobility management is studied on network environment that is called vehicular ad hoc network (VANET) to provide seamless communication and mobility management within vehicle [7]. In VANET, the mobility can be classified into intervehicle between device of vehicle and other IP network, intravehicle between vehicles, and inner vehicle between vehicle devices.

However, the main problem of intervehicle mobility is short-lived connection to Access Router (AR), that causes additional signaling messages and too much packet loss. So, the most critical issue of the design of scalable routing algorithm is to provide robustness of frequent path disruption caused by vehicles' mobility [8, 9]. Typically, vehicle mobility has several characteristics. We will discuss them at Section 3.

In this paper, we classify the mobility into intrahighway mobility and global mobility management mobility with the navigation information that is one of vehicles' mobility characteristics. Furthermore, we propose mobility management scheme based on route prediction in VANET. Handoffs with intrahighway mobility are managed locally and transparency is provided to CHs, while global mobility is managed with Mobile IPv6. Through the numerical analysis, we show that proposed mobility management protocol reduces handoff latency, signaling costs, and packet loss.

This paper is organized as follows. In Section 2, the related mobile management protocols are introduced. Section 3 explains the environment that we are focusing on. In Section 4 we describe the operation of proposed protocol. Section 5 shows numerical analysis between basic NEMO protocol and proposed protocol. Finally, Section 6 concludes this paper.

2. Related Works

2.1. Mobility Management Schemes. To provide seamless communication for mobile devices, IP based mobility management schemes are proposed [1–5]. Typically, MIP protocol is specified IP routing in mobile environment by the IETF. When an MH changes its point of attachment, it gets new care-of-address (CoA). Then, it announces its binding update (BU) at its home agent (HA) and corresponding hosts (CHs). The HA has its binding cache that BU is mapped between MH's home address (HoA) and MH's CoA. If any packets head to MH, HA intercepts and tunnels them to MH's CoA using IP-in-IP encapsulation. BU is received at CH; then, CH sends packet directly to MH's CoA without triangle routing. However, MIP suffers from several well-known weaknesses such as handoff latency or signaling overhead that have led to macro/micromobility schemes. Thus, [10–15] have been proposed to improve handoff performance and reduce signaling overhead.

2.2. Network Mobility. The NEMO protocol maintains the session continuity for all the groups of MHs [2, 16], even when the mobile network (MN), that consists of MHs, dynamically changes its point of attachment to the Internet. It also manages connectivity for all MHs as it moves. The NEMO protocol has been standardized in RFC 3963 [16] to support network mobility. NEMO is based on IPv6, so all signaling messages such as binding update (BU) and binding acknowledgement (BA) are extended Mobile IPv6 messages. The BU and BA messages have an additional flag R bit to signal the mobile router (MR). In NEMO, explicit and implicit mode are proposed. In the explicit mode, several mobile network prefix options (at least one) should be included in a BU message. In the implicit mode, instead of including mobile network prefix, the HA decides mobile network prefix owned by the MR.

When the MR moves to a new link, the MR sends the BU to its HA with a new CoA, which is the IPv6 address of the MR at its current Internet attachment point. The BU message also includes the mobile network prefix option and an R flag. HA updates the MR's routing table and replies a BA. If the packet is sent to an MH from a CN, the HA intercepts the packet and encapsulates its current CoA in bidirectional tunnel to the MR. Then, the MR decapsulates the packet and forwards to the MH.

2.3. Vehicular Ad Hoc Network. Vehicular communication networks are envisioned for the access to drive through Internet and IP based applications. These services are supported by roadside ARs that connected vehicular ad hoc network (VANET) to external IP networks [7]. However, the VANET suffers from asymmetric links due to variable transmission ranges caused by mobility, obstacles, and dissimilar transmission power, which make it difficult to maintain the bidirectional communications and to provide the random mobility required by most mobile IP devices. Moreover, the mobility of vehicle results in short-lived connections to the ARs, affecting the availability of IP services in VANETs. And, more challenge issues are emerging for seamless communications through multihop VANETs, because of proposing the infrastructure to vehicle to vehicle (I2V2V) communications for infotainment applications, such as IP based services and drive through Internet access [8, 9].

First, due to the dynamic network topology of VANET, vehicles transfer their active connection through different IP networks. Thus, the on-going IP sessions are affected by the change of IP addresses, which causes the session disconnections. Second, additional complexity may be added due to links variability during V2V communications and the presence of asymmetric links caused by irregular transmission ranges between network infrastructure and VANET devices [8, 9].

3. Characteristics of Highway Environment

Nowadays, we can easily find using wireless devices in the moving vehicle. Generally, the seamless communication should be provided to most devices in vehicles. However, the MH and MR in the vehicle could be hard to guarantee seamless communication, when the vehicle moves. And, they are moved into highway, or their speed is over 100 km/h, the seamless communication cannot be guaranteed. To guarantee seamless communication for mobile device in a vehicle, we

must consider vehicle movement features. The one is that the vehicle has its own mobility. The other one is the vehicle moves on route that is structured geographically.

3.1. Characteristics of Vehicles' Mobility. The mobility of vehicle has several characteristics.

(1) Heterogeneous network is available (i.e., GPS, WLAN, WIBRO, LTE/3g, Bluetooth, etc.).

(2) A vehicle has the random mobility.

(3) In the vehicle, several mobile devices have the group mobility.

(4) A vehicle has the group mobility with other vehicles that move in same way.

(5) A vehicle must move on route that is structured geographically.

(6) The vehicle's movement pattern can be predicted by the navigation.

(7) On highway, the random mobility disappeared.

If some of characteristics are satisfied, the mobility of vehicle is determined by the group mobility. Also, the vehicle's movement path can be predicted. Thus, this paper is based on two assumptions that are given as follows.

Assumption 1. The navigation must be used, and the vehicle's own path is set before departure.

Assumption 1 can be easily solved by users. And, the navigation has more than one wireless interface. For example, default wireless interface is GPS for location information, velocity information, and geographic information, and optional wireless interface is WLAN for system's data transfer. If the user inputs his/her destination information, the predicted route path is extracted. In this paper, the vehicle's predicted route path information is provided by Assumption 1.

Assumption 2. The network structure of highway or expressway is virtualized as a single subnet or multiple subnets.

Generally, the structure of highway is consistent with road and loop (entrance and exit). The vehicle must enter and exit through the highway loop. So the logical structure of highway can be several roads that are divided by loop. Typically, the logical network structure of highway can be constructed with multiple subnets that covers each road and single domain that covers the highway. Actually, the logical network structure of highway can be constructed with multiple subnets that covers all highway. However, constructing the network structure of highway is not the responsibility of researcher. In reality, it is the responsible of the highway management agency or other local organizations. In this paper, the network structure of highway is provided by Assumption 2.

3.2. Characteristics of Highway Mobility. The highway environment is fixed route that is consistent with partial route and it is geographically connected. As we discuss in Section 3.1, the logical network structure of highway can be constructed with multiple subnets as a single domain. Existing mobility management schemes [10–15] have been researched for intradomain network. They are proposed to improve handoff performance and to reduce signaling overhead for MH within a single domain. In vehicular environment, because of vehicle's mobility, they could not be adopted without any modifications.

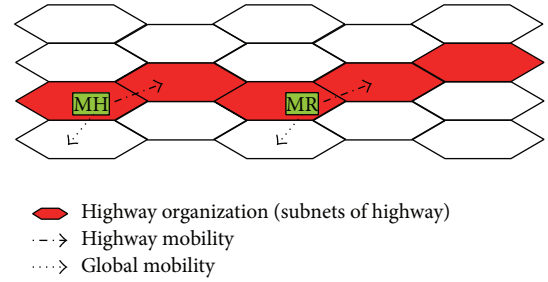


Figure 1: Example of highway organization.

In this paper, we construct highway organization that consists of subnets. We organize geographically distributed subnets into a logical highway organization. Therefore, MH/MR that moves within highway is in physical subnet *i*, but it is logically in highway organization. In Figure 1, there is an example of highway organization that consists of 5 domains.

In this paper, we separate highway mobility from global mobility management by highway organization. The highway mobility does not require binding update signaling to HA. It is managed by highway home AR and minimizes handoff delay on highway environment. Global mobility requires binding update signaling to HA as MIPv6.

3.3. Design of Highway Structure. We define a highway organization that consists of subnets. We organize geographically distributed subnets into a logical highway organization. We propose highway home AR that manages MHs/MRs within the highway organization. Figure 2 shows the example and schematic representation of highway organization.

4. Proposed Highway Mobility Management

4.1. Protocol Overview. In this paper, we separate highway mobility from global mobility management. The highway organization consists of subnets in highway. And the highway organization is connected to the rest of Internet via one or multiple highway home ARs. And the MHs/MRs in the highway organization are managed by the highway home AR.

The proposed highway mobility management scheme consisted of four phases as follows: route_prediction phase, pre_registration phase, registration phase, and packet delivery phase.

4.2. Route_Prediction Phase. Algorithm 1 shows the route prediction algorithm. This algorithm is initiated by

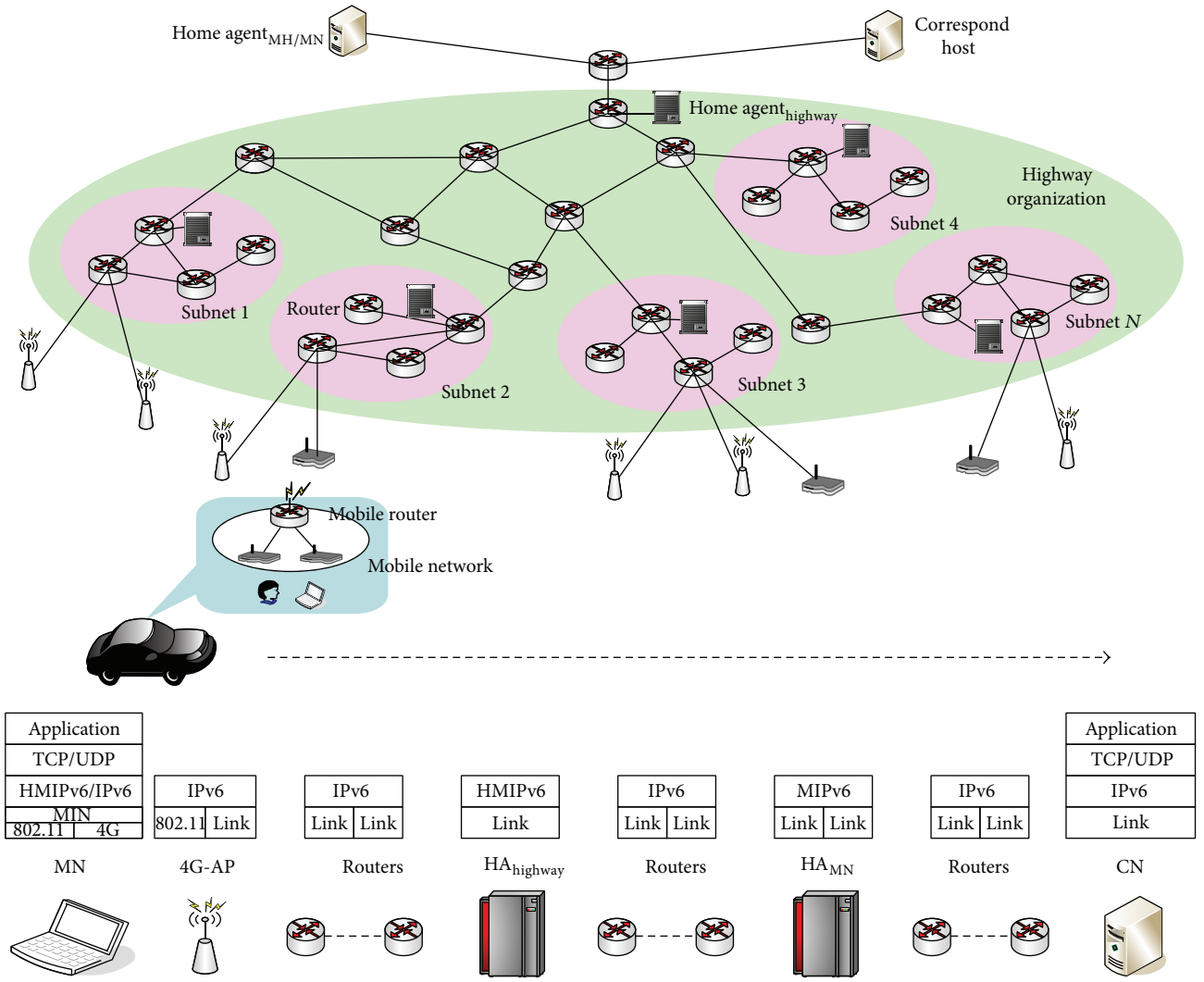


Figure 2: Schematic representation of highway organization.

```
Route_Predictor()
Initialization status;
Get route_path form the navigation;
Begin;
For partial_ path_i
    if bool_operator() is equal to true then
       Send current host id of MH/MR to highway home AR;
    end if
end for
End;
bool_operator()
{
    if (distance > 5 km)                    return true;
    if (current speed > 60 km)              return true;
   return false;
}
```

Algorithm 1: Route prediction algorithm.

```
Highway_Home_AR()
Begin;
if receive current host id of MH then
    Do IP configuration();
   Send prefix of highway home access router address to MH/MR;
   Send reserved host id;
end if
Broadcast reserved host id to all ARs in highway organization
End;
IP configuration()
{
    Do DAD;                                          //Duplicate Address Detection
    if current host id of MH is duplicated then
      discovery available host id;
      Set reserved host id;                          //discovered available host id
    else
      Set reserved host id;                          //current host id of MH
    end if
}
```

Algorithm 2: Algorithm of pre_registration.

MHs/MRs. When the navigation finds out new route path, the route prediction algorithm is started. The $route\_path_i$ consisted of $partial\_path_1$ to $partial\_path_n$. Each $partial\_path_i$ is verified in which there is highway by bool_operator. The bool_operator checks distance and velocity of each $route\_path_i$. And then, if there is one or more $partial\_path_i$ that includes the highway, send pre_registration to highway home AP.

4.3. Pre_Registration Phase. Algorithm 2 shows the pre_registration algorithm. This algorithm is initiated by highway home AR. The pre_registration phase starts with request of pre_registration (). This phase performs the duplicate address detection. If received MH/MR's host id is duplicated, new host id is selected from pool of idle IP. And set it up as a reserved host IP. Then, the reserved host IP is broadcasted to all ARs in highway organization.

4.4. Registration Phase. Figure 3 shows the call-flow of registration. This call-flow consisted of pre_registration phase and after handoff phase. After handoff, the MH/MR does not have to assign its new CoA from new AR. Also, the MH/MR sends only once BU to its home agent when it enters into highway organization.

4.5. Packet Delivery. Figure 4 shows the call-flow of packet delivery. This call-flow consisted of the default packet forwarding phase and the highway packet forwarding phase to which the MH/MR is attached in highway organization.

5. Numerical Analysis

5.1. Signaling Cost and Packet Delivery Cost. We analysis the performance of proposed scheme with respect to the following metrics: location update signaling cost $C_{\text{binding\_update}}$ and packet delivery overhead cost $C_{\text{packet\_delivery}}$. To calculate these metrics, we follow a methodology similar to [17] to calculate the probability that a vehicle moves across ith service areas. We have chosen the MANET centric NEMO scheme [17] for comparison purposes. We define the costs parameters used for the performance analysis as follows:

(i) $C_{\text{binding\_update}}$: the cost of BU/BA,

(ii) $C_{\text{packet\_delivery}}$: the cost of additional IP tunneling header,

(iii) $T_{\text{handoff\_delay}}$: the period of time due to handoff and IP configuration,

(iv) N: the number of network service areas (or AR),

(v) $f_{\text{sa}}(t)$: general distribution of service-area residence time:

$$f_{\text{sa}}(t) = \frac{1}{\mu}, \tag{1}$$

(vi) μ: the rate of service-area crossing:

$$\mu = \frac{vD}{(\pi A)}, \tag{2}$$

(vii) v: the average velocity of MH/MR (vehicle),

(viii) π: the direction of MH/MR (vehicle),

(ix) L: the average length of sessions,

(x) λ_i: the average rate of intersession arrival time,

(xi) ρ_s: the session to mobility ratio:

$$\rho_s = \frac{\lambda_i}{\mu}, \tag{3}$$

(xii) $f_{\text{sa}}(\lambda_i)$: the Laplace transform of service-area residence time distribution,

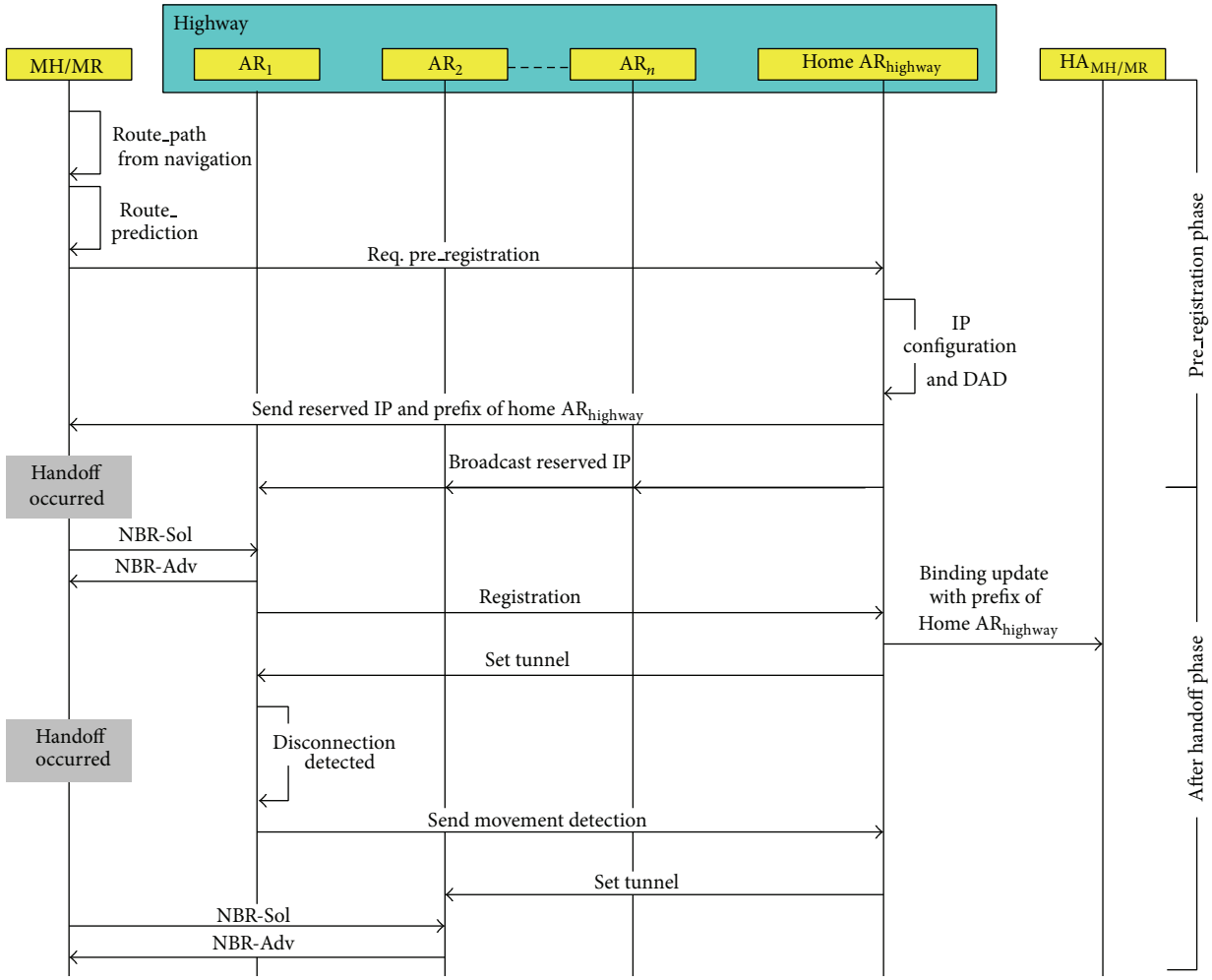


Figure 3: Call-flow of registration.

(xiii) $\alpha(i)$: the probability of crossing i service area:

$$\alpha(i) = \begin{cases} 1 - \dfrac{1}{\rho_s}\left[1 - f^*_{\text{sa}}(\lambda_i)\right], & \text{if } i = 0 \\ \dfrac{1}{\rho_s}\left[1 - f^*_{\text{sa}}(\lambda_i)\right]^2 \left[f^*_{\text{sa}}(\lambda_i)\right]^{i-1}, & \text{if } i > 0, \end{cases} \tag{4}$$

where the session to mobility ratio ρ_s and the Laplace transform of the service-area residence time distribution $f_{\text{sa}}(\lambda_i)$ can be found in [18]. Furthermore, derivation details of (4) can be found in [19].

The location update signaling cost per handoff BU is obtained according to the number of hops the signaling messages have to reach the home agent in default case such as NEMO and the highway home AR in highway organization case that is proposed. It is calculated as follows:

$$\begin{aligned} \text{BU}^{\text{NEMO}} &= \left(n \times \omega + d_{\text{HA}}\right) U^{\text{NEMO}}, \quad \text{if default} \\ \text{BU}^{\text{Proposed}} &= \left(n \times \omega + d_{\text{Highway Home AR}}\right) U^{\text{Highway}}, \quad \text{if handoff within highway,} \end{aligned} \tag{5}$$

where U: the size of BU/BA, $d_{\text{HA}}/d_{\text{Highway\_HA}}$: the number of hops, n: the number of links, and ω: the relative weight of packet transmission on wireless link compared with wired link.

The total location update signaling cost C_{BU} (bytes × hops), incurred by a vehicle moving across several service areas, is calculated as follows:

$$C_{\text{binding\_update}} = \sum_{i=0}^{\infty} i \times \text{BU} \times \alpha(i), \tag{6}$$

where BU is replaced by (5), accordingly. Also, C_{BU} goes to 0 in highway case.

The delivery overhead cost per packet PD accounts for extrainformation and extralinks traversed when delivering a data packet from a server to the vehicle. It is computed as follows:

$$\begin{aligned} \text{PD}^{\text{NEMO}} &= d_{\text{CH}} + H\left(d_{\text{HA}} + n \times \omega\right), \\ \text{PD}^{\text{Proposed}} &= d_{\text{CH}} + H\left(d_{\text{Highway Home AP}} + n \times \omega\right), \end{aligned} \tag{7}$$

where d_{CH} is the distance to the CH, and H is the size of the tunneling IP header.

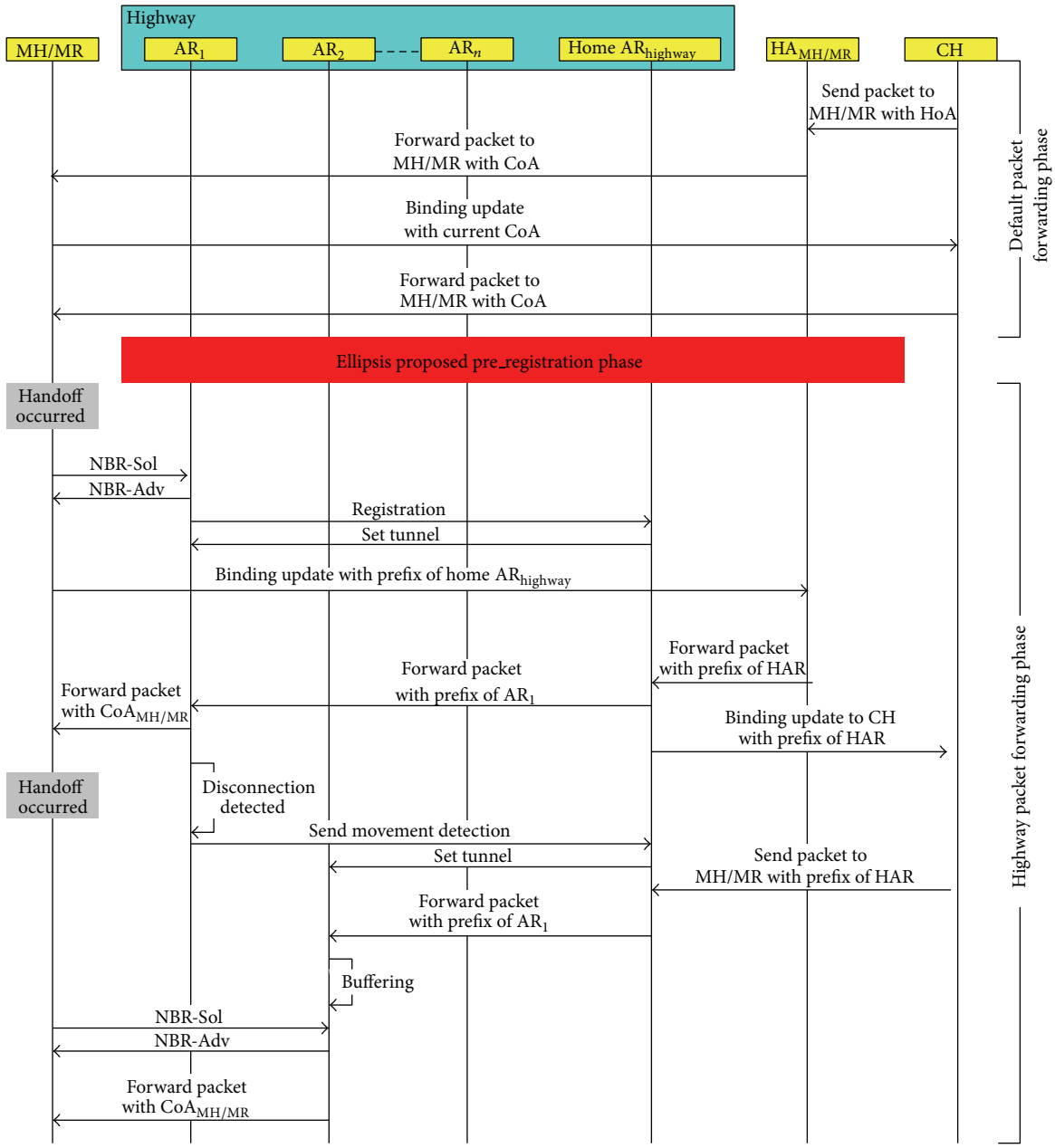


Figure 4: Call-flow of packet delivery.

The total packet delivery cost C_{PD} (bytes × hops) considers the number of active hosts m in the in-vehicle network and the average session length L (packets). L depends on downloading data rate γ, packet size S, and intersession arrival rate λ_i. Thus, C_{PD} is calculated as follows:

$$C_{\text{packet\_delivery}} = m \times \text{PD} \times L, \tag{8}$$

where PD is replaced by (7), accordingly.

Total cost CT is obtained by adding the total location update and total packet delivery cost of each scheme. Therefore,

$$C_{\text{Total}} = C_{\text{binding\_update}} + C_{\text{packet\_delivery}}. \tag{9}$$

Table 1: Numerical analysis parameter.

Parameter	D	A	ω	n
Value	100000	2400 km^2	3	2
$1/\lambda_i$	N	d_{HA}	$d_{\text{Highway Home AR}}$	d_{CH}
10 s–800 s	70	10–15 hops	3–6 hops	15–20 hops
U^{NEMO}	U^{Highway}	P	v	H
124 bytes	124 bytes	124 bytes	30–110 km/h	40 bytes

5.2. Numerical Analysis. In this section, we demonstrate some numerical results. Table 1 shows parameters used in our

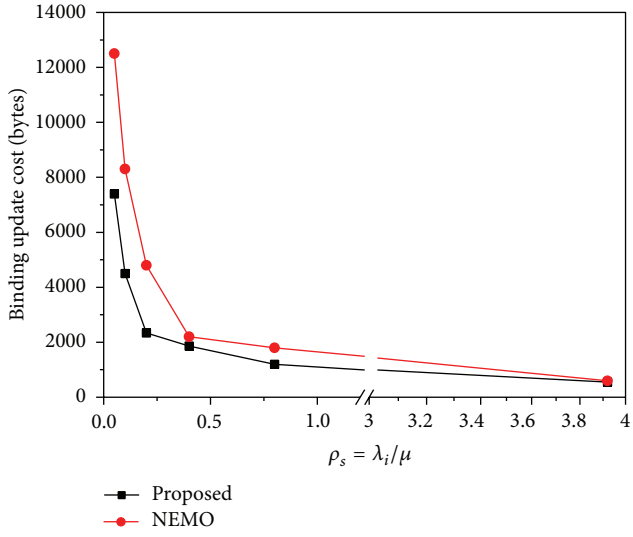


Figure 5: Signaling overhead comparison.

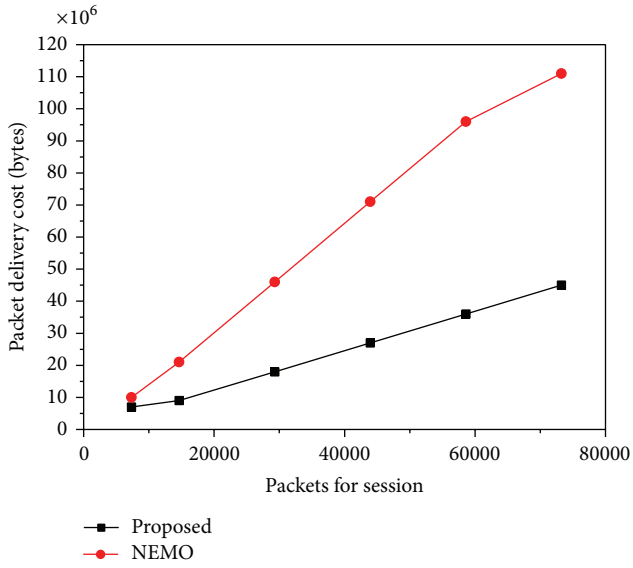


Figure 6: Packet delivery comparison.

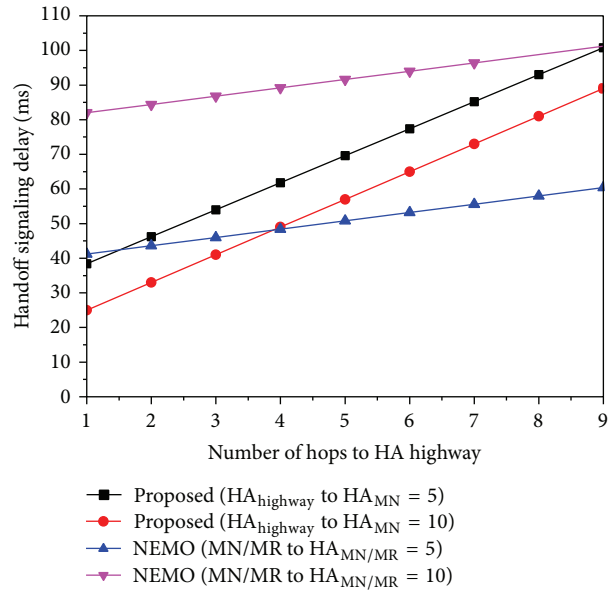


Figure 7: Handoff signaling delay with number of hops.

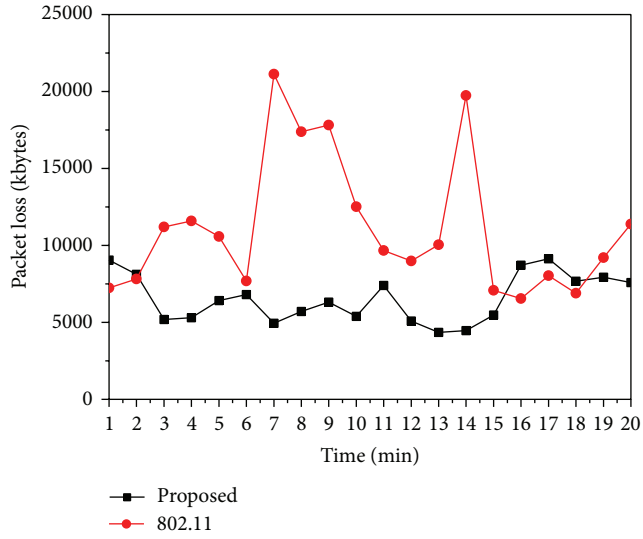


Figure 8: Packet loss in 20 min.

performance analysis [17, 20]. For simplicity, we assume that the distance between mobility agents is fixed and is same.

Figure 5 shows that proposed scheme achieves lower BU cost compared with basic NEMO. When ρ becomes larger, a different result is observed. The result of Figure 5 is compared longer session lengths with mobility, where different session lengths v = 60 km/h, and $(1/\lambda_i)$ = 800 s–10 s. It is also impacted that the reducing packet loss on proposed scheme comes at the cost of a 28.4% increase of BU signaling cost when compared with basic NEMO.

Different downloading data rates (γ = 200 kb/s–4 Mb/s) and session arrival rates ($1/\lambda_i$ = 600 s.) are studied in Figure 6. Figure 6 shows that the packet delivery cost naturally increases for longer data sessions with high velocity. Proposed scheme outperforms basic NEMO on cost of packet delivery. Based on Figure 6, it is observed that the packet overhead is increased with similar rates on the basic NEMO, because packets are affected by the multiple encapsulation, when the highway home AR to subnet AR tunnel is employed.

Figure 7 shows the handoff signaling delay of the proposed protocol. The handoff signaling delay increases linearly with the number of hops. The NEMO does not need to perform the DAD procedure. Also, the size of signaling messages such as BU and BA messages can be reduced by using the pre_registration algorithm.

Figure 8 shows the total packet loss of each method in 20 min. In our simulation scenario the MH moves within highway between 6 min and 15 min. The proposed discovery

process (55,932 kb) outperformed the NEMO process (132,053 kb), showing a 42.3% improvement.

6. Conclusion

In daily life, we can easily see people who are using mobile devices in vehicles. By vehicles movement speed, mobile devices frequently change their point of attachment. So, it is hard to provide seamless connection for mobile devices in vehicles. In this paper, we focus on network mobility management to provide reliable communication within the vehicle that moves in fast moving area such as highway. The most critical issue of the design of scalable routing algorithm is to provide robustness of frequent path disruption caused by vehicles' mobility. For that reason, we pursue the characteristics of vehicles' mobility and analyze them. In this paper, we classified mobility into intrahighway mobility and global mobility management with route prediction by navigation information. Furthermore, we propose efficient mobility management scheme based on route prediction in VANET. Proposed mobility management scheme has several advantages. First, proposed mobility management scheme reduces handoff latency, since handoffs within the highway are managed locally such as handoffs within a single domain. This causes additional advantages increasing handoff speed and minimizes packet loss during transition. Second, proposed mobility management scheme reduces the signaling overhead by BU that each MH/MR initiates and provides transparency to CHs. In this paper, the BU within the highway, the home AP of highway sends the BUs to each MH/MR's HA and their CH. It means proposed mobility management scheme provides transparency to CHs, so the MH/MR does not have to send the BU within the highway. On point of HA/CH, MH/MR stays in the highway. Through the numerical analysis based on the discrete analytic model shows that proposed scheme has superior performance to the basic NEMO scheme within the highway.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by Seokyeong University in 2012.

References

[1] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6," Internet draft (work in progress), draft-ietf-mobileip-ipv6-24.txt, 2003.

[2] V. Devarapalli, R. Wakikawa, and P. Thubert, "Network mobility (NEMO) basic support protocol," Tech. Rep. RFC3963, 2005.

[3] ZigBee Alliance, http://www.zigbee.org/en.

[4] IEEE, IPv6 over low-power WPAN (6LoWPAN), http://www.ietf.org/html.charters/6lowpan-charter.html.

[5] IEEE Computer Society, "Part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (LR-WPANs)," IEEE Std. 802.15.4-2003, 2003.

[6] N. Kushalnagar, G. Montenegro, and C. Schumacher, "IPv6 over low-power wireless personal area networks (6LoWPANs): overview, assumptions, problem statement, and goals," IETF RFC4919, 2007.

[7] F. Li and Y. Wang, "Routing in vehicular ad hoc networks: a survey," *IEEE Vehicular Technology Magazine*, vol. 2, no. 2, pp. 12–22, 2007.

[8] M. E. Mahmoud and X. Shen, "PIS: a practical incentive system for multihop wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 8, pp. 4012–4025, 2010.

[9] N. Lu, T. H. Luan, M. Wang, X. Shen, and F. Bai, "Capacity and delay analysis for social-proximity urban vehicular networks," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '12)*, pp. 1476–1484, Orlando, Fla, USA, March 2012.

[10] H. Soliman, C. Castelluccia, K. E. Malki, and L. Bellier, "Hierarchical MIPv6 mobility management (HMIPv6)," Internet Engineering Task Force draft-ietf-mobileip-hmipv6-04.txt, 2001.

[11] R. Ramjee, K. Varadhan, L. Salgarelli, S. R. Thuel, S.-Y. Wang, and T. la Porta, "HAWAII: a domain-based approach for supporting mobility in wide-area wireless networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 396–410, 2002.

[12] A. T. Campbell, J. Gomez, S. Kim, Z. Turanyi, C.-Y. Wan, and A. Valko, "Cellular IP," Internet Engineering Task Force, draft-ietf-mobileip-cellularip-00.txt, 2000.

[13] C. Williams, "Localized Mobility Management Requirements," Internet Engineering Task Force draft-ietf-mobileip-lmm-requirements-04.txt, 2003.

[14] J. Kempf, "Leveraging Fast Handover Protocols to Support Localized Mobility Management in Mobile IP," Internet Engineering Task Force draft-kempf-mobileip-fastho-lmm-00.txt, June 2003.

[15] N. Xiong, Y. Zhang, L. T. Yang, S. Yeo, L. Shu, and F. Yang, "A fast formation flocking scheme for a group of interactive distributed mobile nodes in autonomous networks," *Mobile Networks and Applications*, vol. 15, no. 4, pp. 477–487, 2010.

[16] I. Ben Jemaa, M. Tsukada, H. Menouar, and T. Ernst, "Validation and evaluation of NEMO in VANET using geographic routing," in *Proceedings of the International Conference on ITS Telecommunications*, November 2010.

[17] R. Chakravorty and I. Pratt, "Performance issues with general packet radio service," *Journal of Communications and Networks*, vol. 4, no. 4, pp. 206–281, 2002.

[18] S. Pack, T. Kwon, Y. Choi, and E. K. Paik, "An adaptive network mobility support protocol in hierarchical mobile IPv6 networks," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 7, pp. 3627–3639, 2009.

[19] J. H. Lee, T. Ernst, and N. Chilamkurti, "Performance analysis of PMIPv6-based NEtwork mobility for intelligent transportation systems," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 1, pp. 74–85, 2012.

[20] J. Xie and I. F. Akyildiz, "A novel distributed dynamic location management scheme for minimizing signaling costs in mobile IP," *IEEE Transactions on Mobile Computing*, vol. 1, no. 3, pp. 163–175, 2002.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 357430, 14 pages
http://dx.doi.org/10.1155/2014/357430

Research Article

PAuthKey: A Pervasive Authentication Protocol and Key Establishment Scheme for Wireless Sensor Networks in Distributed IoT Applications

Pawani Porambage,[1] Corinna Schmitt,[2] Pardeep Kumar,[1] Andrei Gurtov,[3] and Mika Ylianttila[1]

[1] *Centre for Wireless Communications, University of Oulu, P.O. Box 4500, 90014 Oulu, Finland*
[2] *Department of Informatics, University of Zurich, Binzmühlestrasse 14, 8050 Zurich, Switzerland*
[3] *Department of Computer Science and Engineering, Aalto University, 00076 Aalto, Finland*

Correspondence should be addressed to Pawani Porambage; pporamba@ee.oulu.fi

Received 6 November 2013; Revised 2 May 2014; Accepted 7 May 2014; Published 13 July 2014

Academic Editor: Ken Choi

Wireless sensor Networks (WSNs) deployed in distributed Internet of Things (IoT) applications should be integrated into the Internet. According to the distributed architecture, sensor nodes measure data, process, exchange information, and perform collaboratively with other sensor nodes and end-users, which can be internal or external to the network. In order to maintain the trustworthy connectivity and the accessibility of distributed IoT, it is important to establish secure links for end-to-end communication with a strong pervasive authentication mechanism. However, due to the resource constraints and heterogeneous characteristics of the devices, traditional authentication and key management schemes are not effective for such applications. This paper proposes a pervasive lightweight authentication and keying mechanism for WSNs in distributed IoT applications, in which the sensor nodes can establish secured links with peer sensor nodes and end-users. The established authentication scheme PAuthKey is based on implicit certificates and it provides application level end-to-end security. A comprehensive description for the scenario based behavior of the protocol is presented. With the performance evaluation and the security analysis, it is justified that the proposed scheme is viable to deploy in the resource constrained WSNs.

1. Introduction

Wireless sensor network is a key technological building block of Internet of Things, which is considered the future evolution of the Internet. During the past decade, WSN and its security are not only well investigated amongst the industry and academia [1] but also promoted with standardized security solutions [2, 3]. Although the concept and applications of IoT are not novel any longer, IoT security is still in its infancy. However, substantial amount of research work has been done to identify the challenges and possible protection mechanisms for securing IoT, as shown throughout [4–7]. Nevertheless, IoT security protocols are still neither standardized nor commercialized properly due to its novelty and immaturity. Since WSN is an indispensable part of the IoT, it needs to adapt IP technologies to create a seamless and global connectivity with the Internet [6]. The Internet engineering task force (IETF) has contributed significantly to gaining that pervasive connectivity of small objects to IPv6 based Internet. IPv6 over low-power wireless personal area network (6LoWPAN) enables complete integration of WSNs into the Internet [8, 9]. Constrained application protocol (CoAP) and routing protocol for low-power and lossy networks (RPL) are, respectively, proposed for application layer and network layer routing in constrained IoT networks [10, 11]. Physical and MAC layers of low-power networks are defined by IEEE 802.15.4 protocol [2].

In the context of IoT application domains, WSN architectures exist as centralized and distributed approaches [5]. In the centralized approach, a central entity (or a cloud service)

is responsible for acquiring raw data from the sensors, processing received data into required information and format, and providing information for other required entities (e.g., groups of companies and individual customers). In such centralized networks, there is little or no support to access the data sensing network devices directly. In contrast, the distributed networks allow the end-users and other network entities to obtain raw data straightaway from the sensor nodes. Unlike in the centralized approach, in distributed architecture the edge network devices comprise high level intelligence and processing power. Although provisioning of services is located at the edge of the network, different application platforms and end-users can collaborate dynamically with each other. As a result of the decentralized and distributed nature of the network, it is essential to consider the secure management of identity and authentication of connecting devices. In IoT applications, multiple entities (e.g., sensing nodes, service providers, and information processing systems) have to authenticate each other to establish a trusted connection. Not only should the authentication protocols be resistive and robust to malicious attacks, but also they should be lightweight to be deployed in less performing edge devices (i.e., sensors and actuators).

Rather used for generic WSN applications, IoT combined WSN use-cases are widely deployed in smart-home, smart-city, healthcare, and industry monitoring applications [5, 7, 12]. In a hospital environment, there can be different sensors installed in monitoring health conditions of patients (e.g., blood pressure, heart beat, and oxygen concentration). Doctors, who are outside the hospital, might be interested in examining health records of particular patients. Similarly, some medical machinery that maintains the environmental conditions of the ward needs to get the same records. In this scenario, as illustrated in Figure 1, doctors have to access the sensor node as an end-user and the machinery has to collaborate as a sensor node from the same or a distinctive WSN. However, in both cases, the two communication parties need to prove their authenticity to each other before establishing a secure communication link.

In factory automation and power plant monitoring applications, WSNs are deployed inside the factory premises to obtain raw data on machinery vibration, temperature, flow-rate, and light intensity [12]. Their sensed data are used to identify machine abnormalities and to create safety alarms. There can be instances where the users inside and outside the power plant want to acquire raw data directly from the sensor nodes. The end-users and the sensor nodes have to authenticate each other before transferring raw data.

Based on the explained scenarios and the state-of-the-art before, the main contributions of this paper are summarized as follows.

(i) We propose and design a pervasive authentication protocol and a key establishment scheme for the resource constrained WSNs in distributed IoT application, called PAuthKey.

(ii) We implement the PAuthKey protocol and demonstrate its performance measurements on the high resource constrained sensor nodes.

(iii) We conduct a security analysis on PAuthKey, along with performance and security comparisons between it and DTLS scheme, which is currently considered the most appropriate authentication scheme for constrained IoT networks. Moreover, we show the performance comparison results of two phases of PAuthKey with ECDSA and ECDH schemes.

The rest of the paper is organized as follows. Section 2 provides a brief overview about the related work. Section 3 comprehensively describes the system architecture, where the authentication protocol is developed, and the notations used. Section 4 presents the proposed authentication and key management protocol known as PAuthKey. Section 5 gives a detailed explanation about the implementation, performance analysis, security analysis, and scalability of the PAuthKey protocol. Finally, Section 6 concludes the paper.

2. Related Work

In centralized WSN, data from the sensor nodes are transmitted to a single central location, which processes information and combines and provides information acquisition for end-users (i.e., customers) [7]. Due to the high data availability and massive network size, processing of data on a single location might be inefficient, congested, and undertaking a high risk at single entity failure. In the distributed networks, the sensor nodes can retrieve, process, and provide data for other entities and end-users. Figure 1 provides an overview of the distributed IoT approach, which allows the communication among the edge devices, end-users, and IoT server cloud.

Distributed architecture supports the IoT network applications by providing services at local level and collaborating with all the network devices and users to achieve common goals. Due the network heterogeneity and device mobility, there can be many security threats and issues encountered with distributed IoT. In [7] Roman et al. have identified security challenges in distributed IoT. According to their study, network entity identity, authentication, access control, and secure communication channel establishment are major security concerns in distributed IoT. The proposed mechanisms should be robust to node mobility and network scalability due to the dynamic behavior of nodes. Additionally, the network needs to scale up after installation.

Exploitation of a master key for entity authentication for pervasive computing environments would be also a feasible approach to IoT enabled WSNs [13]. According to [14], the authentication mechanisms for WSN applications can be summarized as password based, remote user authentication using one-way hash functions and ticket based authentication. However, most of the work has the sole purpose of enabling end-user authentication in generic WSN architecture and it does not provide the extensibility for the key establishment. In [15, 16], the authors have proposed broadcast authentication schemes for WSNs. Reference [14] presents an effective authentication mechanism for ubiquitous collaboration in heterogeneous computing environment. This is a ticket based user authentication scheme, which is not applicable to the high resource constrained devices due to large

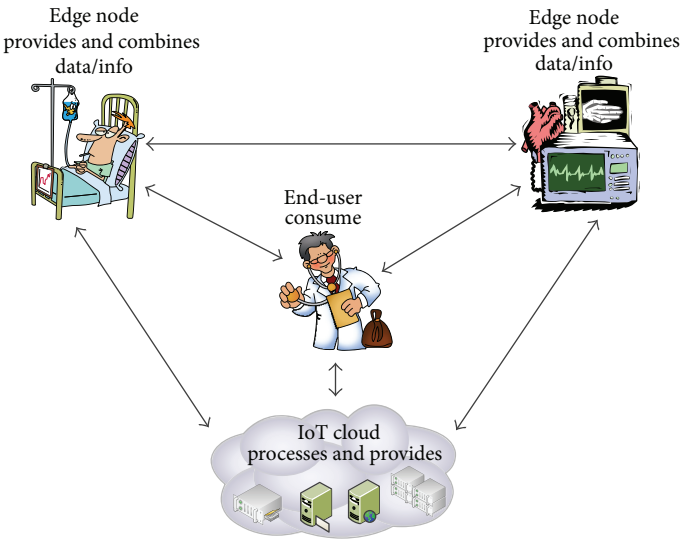


Figure 1: Overview of the distributed IoT approach in a hospital environment.

memory consumption. Nevertheless, these works have less or almost zero contribution to securing IoT combined WSNs. The reason is that they have less addressed network scalability and device mobility issues.

DTLS is an adaptation of TLS protocol and it provides an equal communication security as TLS for datagram protocols [17]. According to [10], the secured version of CoAP (known as CoAPs) is defined with DTLS due to the unreliable communication links in CoAP based IoT networks. In [18], the authors have introduced the first fully implemented two-way authentication scheme for the IoT based on DTLS protocol. However, due to the existence of eight message transfers to complete DTLS handshake, it induces a significant overhead to the network traffic. The main drawback is the utilization of X.509 certificates and RSA public keys with DTLS handshake, which are too heavy for the low performing and high resource restricted sensor nodes.

Due to the high resource demand, public key cryptographic (PKC) algorithms, such as RSA, are not recommended for WSN applications. However, elliptic curve cryptography (ECC) (i.e., a lightweight PKC alternative) based security solutions are not anymore new to WSNs. The utilization of implicit certificates for generating pairwise ephemeral keys is yet an improving realm. There are several implicit certificate generation schemes for WSNs presented in [19, 20]. Elliptic curve Qu-Vanstone (ECQV) is one of such schemes embedded in ZigBee Smart Energy applications [21]. In [22], a fully implemented end-to-end authentication scheme has been introduced to the high constrained embedded devices. TinyECC is a stable ECC implementation for constrained network entities, where in [23] the authors provide implementation details and measurement results for elliptic curve digital signature algorithm (ECDSA) and Diffie-Hellman key establishment (ECDH). Several ECC based security schemes have been proposed for WSNs as published in [15, 19, 24–26].

3. System Model and Notations

In this section, the authors provide details about the system architecture, where the protocol is modeled, and information about the used notations.

3.1. System Model. Figure 2 illustrates the assumed network architecture for the proposed authentication scheme, where end-users can collaborate with different edge devices in order to obtain particular information or service. The edge networks may include heterogeneous devices and the end-users can be humans or virtual entities (e.g., web applications).

According to the distributed IoT architecture, end-users and edge devices (i.e., sensor nodes) should possess the capability of securely accessing an edge device in a WSN. Therefore, based on Figure 2, mutual authentication is considered for four types of communication link establishments, particularly the following.

(1) Two sensor nodes are located in the same cluster (`Link A`).

(2) Two sensor nodes are located in distinctive clusters in the same WSN (`Link B`).

(3) Two sensor nodes are located in distinctive clusters and in distinctive WSNs (`Link C`).

(4) An end-user is linked to a sensor node (`Link D`).

Before starting the actual authentication protocol between two network entities, it is necessary to undergo a registration process by every communication party in order to retrieve cryptographic credentials. Later, the obtained security credentials are to be exploited for mutual authentication. For the given four types of communication link possibilities (1)–(4), every edge device and end-user have to acquire security credentials (e.g., cryptographic suites and implicit certificates) from a trusted third party such as a

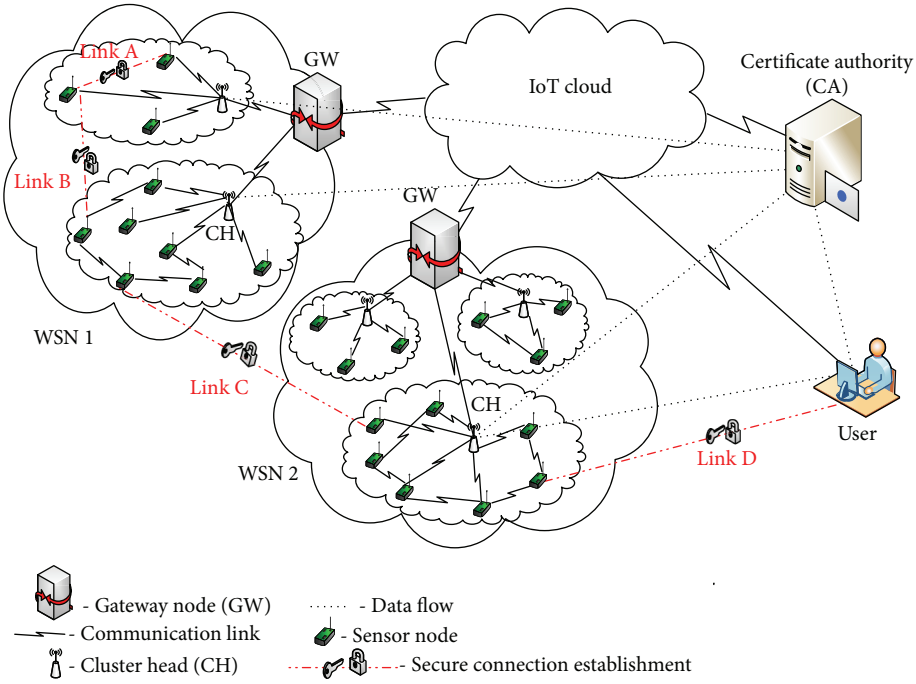


Figure 2: Assumed network architecture.

certificate authority (CA). It is assumed that the CA is a highly resource-rich server and is already known by the edge nodes during the registration phase.

In this architecture, two types of network entities such as resource rich entities (i.e., end-users and cluster heads (CH)) and highly resource constrained network entities (i.e., sensor nodes) are considered. Here, a cluster tree topology of WSNs is assumed, where a CH is the controlling device for the sensor nodes in a particular cluster. Therefore, it is considered that CH is performing as the CA (i.e., to issue implicit certificates) for the same group of sensor nodes in the cluster. However, further details about the authentication between a CH and an end-user (i.e., between two resource-rich entities) are not provided in this paper. The major concerns are the authentication between two constrained nodes or one constrained node and a resource-rich entity. Hence, it is assumed that all the CHs and end-users advocate DTLS for secure end-to-end communication after acquiring X.509 certificates from a common CA. As aforementioned in Section 2, X.509 certificates are only handled by end-users and CHs, due to their complexity and overhead on tiny sensor nodes. As illustrated in Figure 2, resource-rich network entities (i.e., end-users and CHs) first communicate with the common CA along the already established communication links heading through an IoT cloud. If an end-user needs to communicate with a sensor node with a particular cluster, first it needs to establish a secure DTLS connection with the corresponding CH and obtain implicit certificates from the CH. The end-user can obtain the implicit certificate from the CH through that secure link. Then, the end-user can use the obtained implicit certificate to communicate with the sensor node.

Having a valid implicit certificate allows the two entities for mutual authentication irrespective of their local network. Existing nodes can change their locations dynamically after requesting a new certificate. No matter what the size of the network is, adding new nodes can easily extend the data acquisition and service providing networks. It is assumed that CH can recognize the valid identities and communicate with the network entities, which are requesting security credentials [27]. The reason is that the CH has to verify the certificate requestor's identity at the beginning of the handshake and it performs the verification mainly based on the identity of the requestor node. The IPv6 over low-power wireless personal area network (6LoWPAN) identities are considered for the identification. In this paper, an end-to-end authentication is proposed for the application layer, while relying on the security schemes provided from the physical and MAC layers in IEEE802.15.4 standard [2]. Subsequently, the edge devices and end-users can mutually authenticate and establish secure communication channels, due to the distributive nature of the entire architecture.

3.2. Notations. The notations used in this paper are defined in Table 1. Elliptic curve (EC) parameters are denoted by q, a, b, G, n. The variable q is a prime, which indicates finite field F_q. The variables a and b are coefficients of EC $y^2 = x^3 + ax + b$, where $4a^3 + 27b^2 \neq 0$. G is the base point generator with order of n, which is also a prime [28].

4. PAuthKey Work Flow

The PAuthKey protocol mainly consists of two phases: *registration phase* and *authentication phase*. During the registration phase, the sensor nodes in a particular cluster should obtain certificates from the CH and derive their own public-private key pairs. The authentication phase is varying upon

Table 1: Notations used in cryptographic algorithms.

Notations	Description
K	Network-wide symmetric key for initial message authentication
r_U	Secret random integer value generated by U
R_U	EC point for certificate request sent by node U
Cert_U	Implicit certificate of Uth node
e	Integer used to keep hash value of Cert_U
s	Integer used to compute private key of the requestor node
d_U	Node U's private key
Q_U	Node U's public key
N_U	Random cryptographic nonce generated by node U
K_{UV}	Link key between nodes U and V

the type of the communication link between the end-parties (i.e., four communication link possibilities as explained in Section 3). Accordingly, the authentication phase is described for three scenarios with reference to the system model in Figure 2: scenario 1 for link A, scenario 2 for links B and C, and scenario 3 for link D. The upcoming subsections characterize the phases individually.

4.1. Registration Phase (Initial Certificate Acquisition). Initially, the sensor nodes should obtain security credentials from their respective cluster head (CH) as a prerequisite for the actual authentication protocol. All the sensors in a particular cluster consider their certificate authority (CA) as the CH. Upon the certificate request from sensor node U, the CA generates the certificate. The message flow of the certificate acquisition is illustrated in Figure 3. The grey boxes show the change of variables and the white boxes indicate the performed functionality by the entity.

The handshake starts with the `Requestor Hello` message, 6LoWPAN node identity (U), and cipher suites that are supported by the requestor. The cipher suites are common for all the edge devices and would include the available cipher options on the requestor, for example, *CERT_ECC_160_WITH_AES_128_SHA1* requests for certificates in 160-bit EC curves, 128-bit AES for bulk encryption, and SHA1 for hashing. It is assumed that the cipher suites are installed on the sensor nodes during the offline mode before the deployment. CA uses 6LoWPAN node identity to figure out whether it is a valid request from a node that belongs to its cluster. If the requestor identity verification is successful, CA agrees to one cipher suite from the received options and sends `CA Hello` message with its public key Q_{CA} as an unprotected message to approve the initiation of the handshake.

Upon receiving `CA Hello` message, the requestor generates a certificate request EC point R_U. While creating a certificate request, first, the node generates a random number $r_U \in [1, \ldots, n-1]$ and computes $R_U = r_U \times G$. The predefined EC domain parameters are used according to the negotiated cipher suite. Second, the node produces a random cryptographic nonce N_U, calculates message authentication code (MAC) value (i.e., $\text{MAC}[R_U, U, N_U]$), and sends those two along with the `Certificate Request` message. The CA generates the MAC value using the common message authentication key K, which is mentioned in the cipher suite. The random cryptographic nonce is used to ensure the freshness of the message.

After receiving the certificate request, CA verifies the MAC value and nonce N_U in order to identify the integrity and the freshness of it. If the verification is successful, CA computes the certificate Cert_U and private key reconstruction value s for the requestor node U. During this process, CA first generates a random number $r_{\text{CA}} \in [1, \ldots, n-1]$ and computes the certificate $\text{Cert}_U = R_U + r_{\text{CA}} \times G$. Then, CA calculates s using Cert_U, r_{CA}, and its own private key d_{CA}; $e = H(\text{Cert}_U)$ and $s = er_{\text{CA}} + d_{\text{CA}} (\text{mod } n)$. The value e should be computed using the one-way cryptographic hashing function, which is mentioned in the cipher suite (e.g., SHA1). Later, the CA sends the `Certificate` message that includes the certificate Cert_U, s, a random nonce N_{CA}, and the MAC on $[\text{Cert}_U, s, N_{\text{CA}}]$.

While receiving this message, the requestor node U first verifies the MAC and N_{CA}. If they are correct, U calculates $e = H(\text{Cert}_U)$ using the same hash function. Then, the node U can compute its own private key $d_U = er_U + s (\text{mod } n)$ and public key $Q_U = d_U \times G$.

Node U's `Finished` message contains an encrypted message digest of previous handshake messages using the requestor public key Q_U. According to the EC arithmetic operations which are performed for calculating keys [10], CA is also capable of computing U's public key Q_U; $Q_U = e\text{Cert}_U + Q_{\text{CA}}$. The following derivation proves that both equations give exactly the same Q_U as computed by node U:

$$\begin{aligned} Q_U = d_U G &= \left(er_U + s \,(\text{mod } n)\right) G \\ &= \left(er_U + er_{\text{CA}} + d_{CA} \,(\text{mod } n)\right) G \\ &= e\left(r_U + r_{\text{CA}} \,(\text{mod } n)\right) G + d_{\text{CA}} G \\ &= e\left(r_U G + r_{\text{CA}} G\right) + Q_{\text{CA}} \\ &= e\left(R_U + r_{\text{CA}} G\right) + Q_{\text{CA}} \\ &= e\text{Cert}_U + Q_{\text{CA}}. \end{aligned} \tag{1}$$

CA uses public key Q_U for encrypting previous messages and answers with the `Finished` message to complete the handshake of the preauthentication phase.

At the end of the registration phase, the sensor nodes possess the security credentials to start secure communication with the internal and the external network entities (i.e., end-users and sensor nodes).

4.2. Authentication Phase. The authentication phase is described for three scenarios. Scenario 1 (`Link A`) is the authentication between two sensor nodes in the same cluster. Parts of scenario 2 (`Link B and C`) and scenario 3 (`Link D`) also include the principal handshake of scenario 1. Scenario 2 is the authentication between two sensor nodes in distinctive clusters in the same or different WSN. Scenario 3 is the authentication between a sensor node and an end-user. These

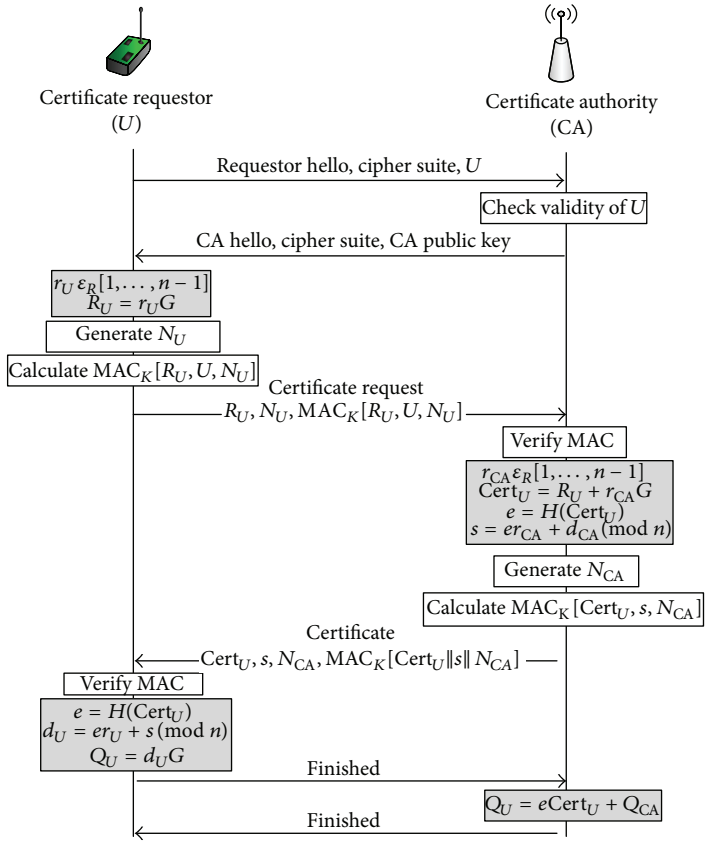


Figure 3: Message flow for the registration phase.

scenarios resemble the communication link possibilities, which are shown in Figure 2.

Scenario 1: Authentication Process between Two Sensor Nodes in the Same Cluster. The first scenario is the exploitation of certificates and public-private keys for node authentication between two sensor nodes in the same cluster. Since the sensor nodes in a particular cluster obtain security credentials from a common CA, they can easily carry out the mutual authentication as depicted in Figure 4. The grey boxes are value ranges and the white boxes are performed operations. The client node U is aware of the 6LoWPAN identity of the server node V, which U needs to acquire the data or the service. As the initial step, the client sends the `Client Hello` message accompanied with its identity U to the server node. The server replies with the `Server Hello` message along with the cipher suites it supports and CH's identity, which is the CA for both nodes. If the client does not have the security credentials from the given cipher suite, it has to retrieve them from the CA. Otherwise, the client can continue the handshake by sending its certificate Cert_U. Similar to the registration phase handshake, random cryptographic nonce N_U and MAC values are used in order to preserve the freshness and integrity of the message.

Upon receiving the client's certificate, the server first verifies the MAC value and then computes the client's public key Q_U using its certificate; $e = H(\text{Cert}_U)$ and $Q_U = e\text{Cert}_U + Q_{\text{CA}}$. This is proven according to (1). Then, the server V generates a random nonce N_V and sends it along with Cert_V and $\text{MAC}_K[\text{Cert}_V, N_V, V]$. In the meantime, the server V computes the pairwise key K_{UV} from its private key d_V and U's public key Q_U, where $K_{UV} = d_V Q_U$. Similar to the server V, upon receiving the message, the client U verifies the MAC and if the verification is successful it computes Q_V and $K_{UV} = d_U Q_V$. Therefore, at the end of two-way message transferring, both parties can derive a common pairwise key for actual secure communication.

Finally, the exchange of the `Finished` messages concludes the handshake. This Finished message is composed of previous handshake messages, which are encrypted by the common key K_{UV}. At the end of six message transfers, the two edge nodes can authenticate each other and establish a common secret key and a secure communication link that can be used for securing further data acquisitions between the client and the server.

Scenario 2: Authentication Process between Two Sensor Nodes in Distinctive Clusters. Here, the node authentication process is demonstrated between two sensor nodes located in distinctive clusters, which might be in the same or different WSNs. In such cases, the nodes cannot use their certificates for mutual authentication since they are generated from two CAs. The messages flow of the authentication protocol is illustrated in Figure 5.

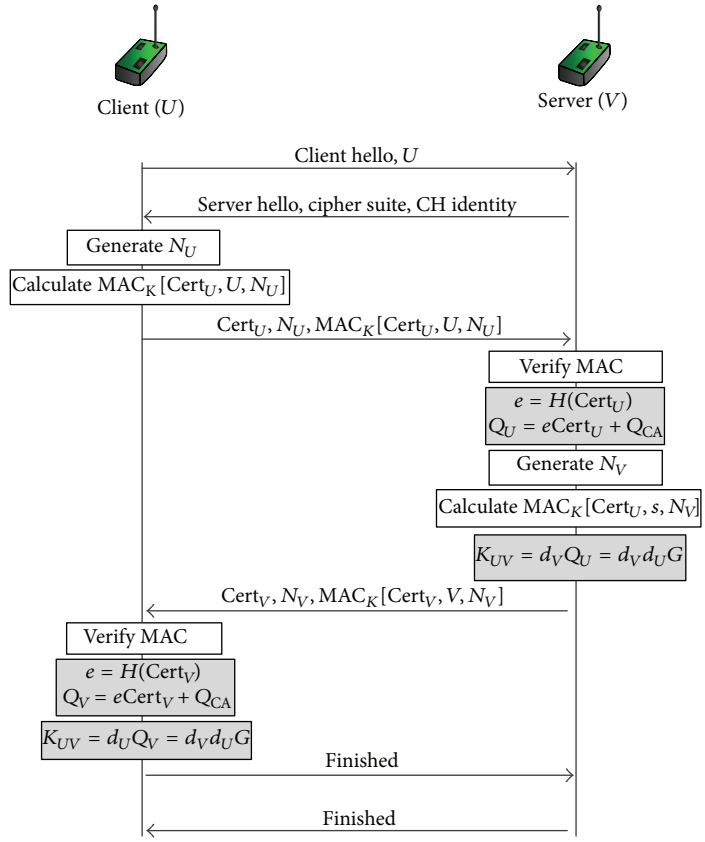


Figure 4: Message flow for scenario 1—authentication process between two sensor nodes in the same cluster.

Similar to scenario 1, the preliminary `Hello` message exchange initiates the handshake. Upon receiving the `Server Hello` message, the client checks the identity of the CH corresponding to the server and verifies the necessity of requesting a new certificate from the given CH. The client U is unable to communicate with CH_V directly and it obtains the security credentials through its own cluster head CH_U. The client forwards the cipher suite and *CH identity* to its cluster head CH_U. CH_U communicates with CH_V on behalf of the node U and requests the security credentials. As explained in the system model, it is assumed that two cluster heads initiate a secure communication link using RSA keys and DTLS handshake [18]. Though CH_V does not communicate with the node U directly, it grants the certificate by being the CA for the node U. The new certificate request and the certificate are exchanged through that established secure channel. Additionally, CH_V sends its public key Q^*_{CA} to CH_U, which is used by U during further computations. Then, CH_U computes the new public-private keys Q^*_U and d^*_U of node U and sends the certificate Cert^*_U and the keys (i.e., Q^*_U, d^*_U, and Q^*_{CA}) to the node U. The security credentials are encrypted by U's primal public key Q_U. After having the required security credentials from the new CA (i.e., cluster head of node V), the client U can continue the authentication protocol as described in scenario 1 handshake.

Scenario 3: Authentication Process between End-User and Sensor Node. Figure 6 demonstrates the flow of the message transactions of the authentication process between an end-user and a sensor node. The difference between scenario 3 and scenario 2 is that here the user directly retrieves security credentials from the given cluster head. However, in scenario 2, the two cluster heads have to communicate first for the acquisition of the implicit certificate for the client node. Similar to the previous case (i.e., scenario 2), it is assumed that the secure link between the user and the CH is established with RSA keys and DTLS handshake [18], and the security credentials are transmitted over that link. Once the end-user obtains the certificate and computes its public-private keys, the rest of the handshake would occur in a similar manner as scenarios 1 and 2.

As explained in the above three scenarios, the end-users and the sensor nodes can establish secure communication links after authenticating each other using implicit certificates. Furthermore, the two-party authentication mechanism enables the nodes to generate a pairwise common secret key. Therefore, this would advocate accessing the data and services in WSNs accommodated in distributed IoT architecture.

5. Analytical Justification of PAuthKey

In this section, a comprehensive analysis of the proposed PAuthKey protocol is presented. The performance analysis is given in terms of memory, energy consumption, and execution time, along with the support for network scalability.

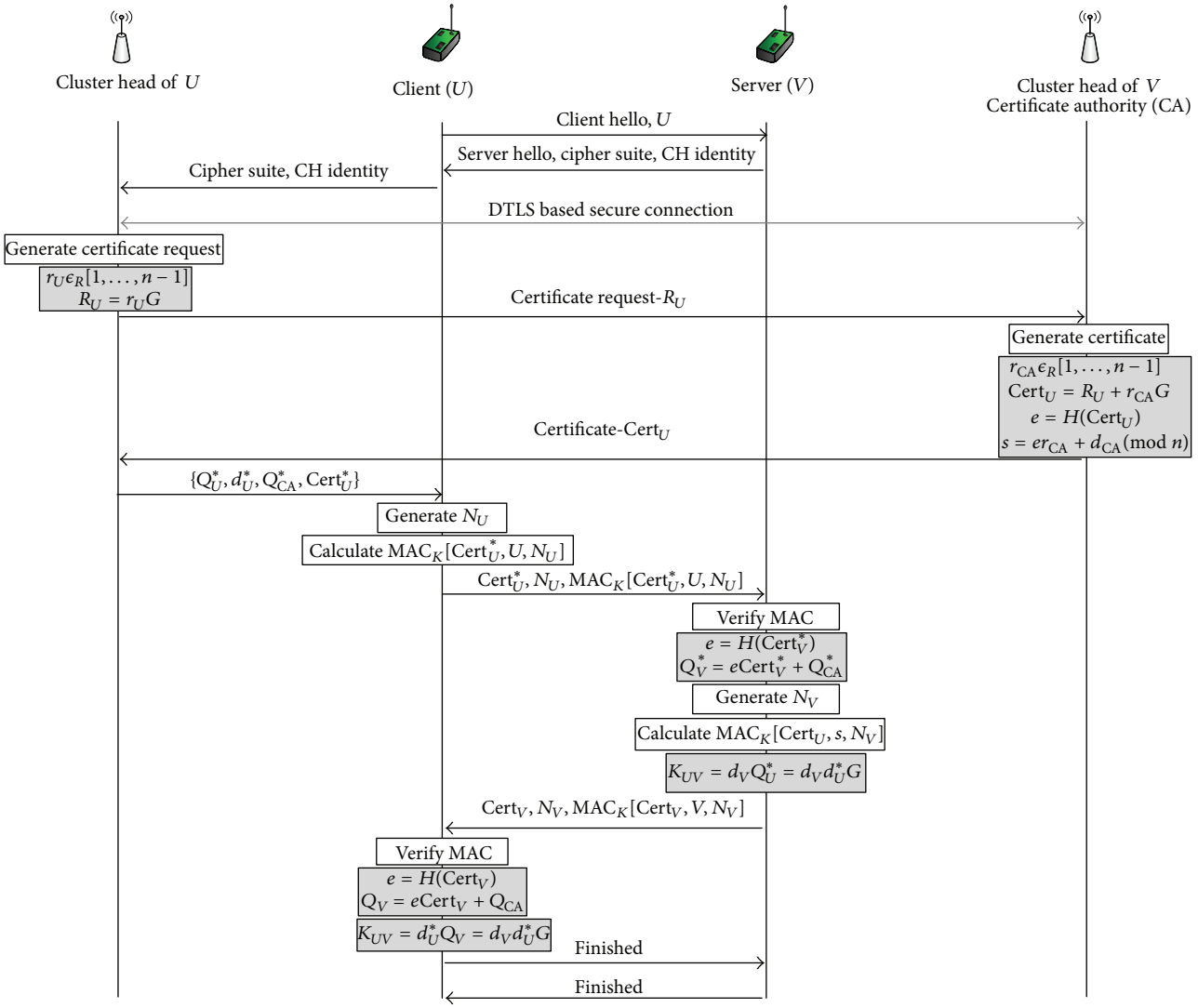


Figure 5: Message flow for scenario 2—authentication process between two sensor nodes in distinctive clusters.

Next, the security of the protocol and the comparisons with the related work are discussed.

5.1. Performance Analysis. Physical and MAC layer security protocols do not provide end-to-end communication security. DTLS is the widely used application level security protocol for authentication in IoT networks. Variants of DTLS handshakes are based on ECC and used with RSA and X.509 certificates [22]. Although the exploitations of RSA and X.509 certificates with DTLS provide interoperability, they are hardly utilized by the high resource constrained devices (e.g., sensors). The major drawbacks are as follows.

(1) RSA has a key size of 2048 bits.

(2) Standard X.509 certificates are in the order of 1 kB in size.

(3) The utilization of RSA and X.509 on constrained sensors consumes resources and induces computation overhead.

The PAuthKey solution is implemented on a simple network with TelosB sensor nodes [29] that have IEEE 802.15.4 compliant CC2420 RF transceivers. The hardware includes 8 MHz, 16-bit MCU with 10 Kbyte RAM and 48 Kbyte ROM. CC2420 RF transceiver has a maximum data rate of 250 kbps and frequency band of 2400 MHz [29]. PAuthKey is developed in NesC on TinyOS 2.1.2 [30]. ECC (i.e., for EC arithmetic operations) and natural number (NN) (i.e., for large natural number operations) interfaces are utilized from TinyECC configurable library [23]. *secp160r1* EC domain parameters are used as defined in [28]. The authors of this paper utilized EC optimization techniques provided in TinyECC such as *Barrett reduction* to speed up modulo operations, *Hybrid Multiplication* and *Squaring* for integer multiplication, *Projective Coordinate Systems* for point addition, and *Sliding Window* for scalar multiplication. SHA-1 is used as the one-way cryptographic function H. ECC operations are extremely costly compared to other cryptographic operations (i.e., SHA-1 and MAC) [23]. Therefore, we have considered the given EC operation optimization techniques.

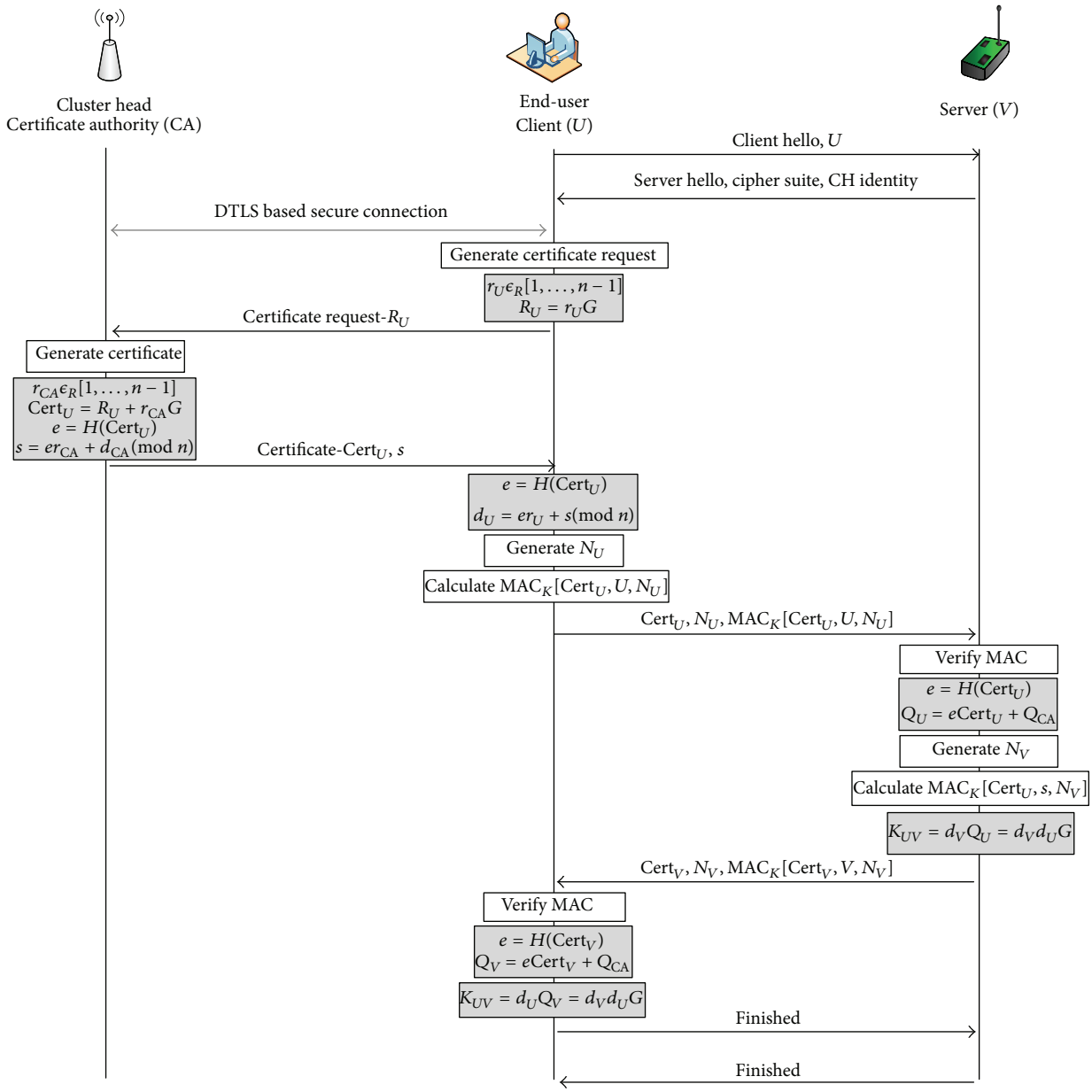


Figure 6: Message flow for scenario 3—authentication process between end-user and sensor node.

The experimental setup comprises three TelosB nodes, one as the CA and the rest as the cluster nodes. For the sake of simplicity and comparison, CA functionalities are also implemented on the sensor node itself. The measurements are taken in terms of execution time, energy, and memory (i.e., RAM and ROM) consumption. The *check_size.pl* script is used to obtain memory consumption values (e.g., for RAM and ROM) required by each operation in registration phase and authentication phase for scenario 1. The execution times are measured directly on the sensor nodes and the energy consumptions are computed using the runtime. The energy consumptions are then calculated as $V \times I \times t$ based on the voltage (V), the current (I), and the execution time (t) on TelosB sensor nodes [29]. Similar to [23], it is also considered that the static values given in [29] are $V = 3$ volts and $I = 1.8$ mA.

Table 2: Memory utilization.

	RAM (bytes)	ROM (bytes)
Registration phase		
Edge node operations	1410	11774
CA operations	2332	16576
Authentication phase		
Operations at one edge node for scenario 1	1530	11650

As given in Table 2, memory utilization values are taken for two phases with respect to the communicating nodes.

Table 3: Execution time and energy consumption.

Operation	Time (ms)	Energy (mJ)
Registration phase		
At the sensor node side		
Initialization	2709	14.6286
Cert Req generation	2764	14.9256
Cert verification	2758	14.8932
`Finished` msg computation	4	0.0216
At the CA side		
Initialization	2709	14.6286
Cert generation	5728	30.9312
`Finished` msg computation	2154	11.6316
Authentication phase (scenario 1)		
At the client or the server side		
Initialization	2672	14.4288
Key computation	5768	31.1472
`Finished` msg computation	4	0.0216

For the registration phase, the total memory consumption is measured for edge node operations (i.e., certificate requestor) and CA operations. Edge node operations include the generations of `Requestor Hello`, `Certificate Request`, and `Finished` messages, certificate verification, and public-private key computation at the sensor node side. Similarly, CA operations include the generation of `CA Hello`, `Certificate`, and `Finished` messages and `Certificate Request` verification.

For the authentication phase, scenario 1 was only considered, since it is almost similar to the major overhead created at the sensor node side for the other scenarios. According to the message flow of the authentication phase (i.e., scenario 1 in Figure 4), the collective operations performed at the client and the server sides are identical. Therefore, the operations at one edge node include the generation of `Hello` and `Finished` messages, MAC computation and verification, public key calculation, and the derivation of the common secret key. As a result, the two phases of the protocol collaboratively consume 2940 bytes of RAM and 23424 bytes of ROM. However, the overall implementation of two phases is still below the 10 kB RAM and 48 kB ROM provided by the high resource constrained TelosB sensor nodes. Although the memory consumption is higher for CA operations, in the real-time deployment it would be tolerable for a resource-rich device.

Since the transmission time depends on the size of the network and the distance between the nodes, only the execution time for the particular operations performed at the edge nodes and CA is measured for the registration and authentication phases for scenario 1. The measured execution time values and the calculated energy consumption values are depicted in Table 3.

During the registration phase, the approximate collective time utilization at certificate requestor's (i.e., the sensor node) side is 8235 ms. This value includes the execution times for initialization (2709 ms), certificate request generation (2764 ms), certificate verification (i.e., private-public key derivation) (2758 ms), and Finished message computation (4 ms). At CA's side, the execution time values are taken for initialization (2709 ms), certificate generation (5728 ms), and `Finished` message computation (2154 ms). Altogether, CA spends 10591 ms for its operations under the registration phase. At CA's side, the `Finished` message computation has a higher execution time due to the derivation of the public key of the sensor node (i.e., EC operation $Q_U = e\text{Cert}_U + Q_{CA}$).

During the authentication phase for scenario 1, each edge node (i.e., the client or the server) takes approximately 8444 ms for initialization, key computation, and `Finished` message computation. For the same operations, we have calculated the energy consumptions at TelosB nodes using $V \times I \times t$. According to the computed values, a TelosB sensor node consumes nearly 44.47 mJ and 45.6 mJ for registration and authentication phases (for scenario 1), respectively. However, these timing, energy, and memory values can be improved by using further optimized basic ECC arithmetic operations. All in all, experimental results show that the proposed authentication mechanism can be easily deployed in low-power less performing devices.

In the proposed two-phase authentication protocol, implicit certificates, which are 160-bit EC points instead of X.509 certificates, were used. Therefore, the size of the certificate is only 44 bytes. Using optimally designed EC curves we can reduce the certificate size and using compression techniques we can further decrease the overall message size. Retransmission clocks can be used at both communicating parties for identifying timeouts and retransmitting when there is a message loss. Furthermore, the authentication protocol supports scaling up the network, since the newly added nodes can authenticate themselves after undergoing the registration phase. As the certificates are not based on the physical locations of the edge devices, they do not have to be alternated according to nodes' mobility.

5.2. Scalability. The proposed authentication protocol supports the scalability of the network (i.e., expanding the network with new node addition) and the location changes of the sensor nodes within the same cluster. When a new node is added to the network, a valid 6LoWPAN node identity, K message authentication key, and cipher suites should be stored while the node is at offline mode. Figure 7 illustrates how PAuthKey protocol supports a new node addition to the network, within a particular cluster. It is illustrated as a three-stage process. In *Stage 1*, at the bootstrapping phase, the newly added node (marked as red rectangular shaped) can send the certificate request and obtain a certificate from the CA for computing its own keys. Hence, the size of the network is not necessary to be predefined during the initial design phase and deployment phase. At a new node request, the CA only needs to verify the validity of the sensor node identities to issue the certificate. In *Stage 2*, the new node receives its certificate. Therefore, *Stages 1* and *2* resemble the registration phase of PAuthKey protocol. Finally in *Stage 3*, the node can undergo the authentication phase and the key establishment using the received certificate.

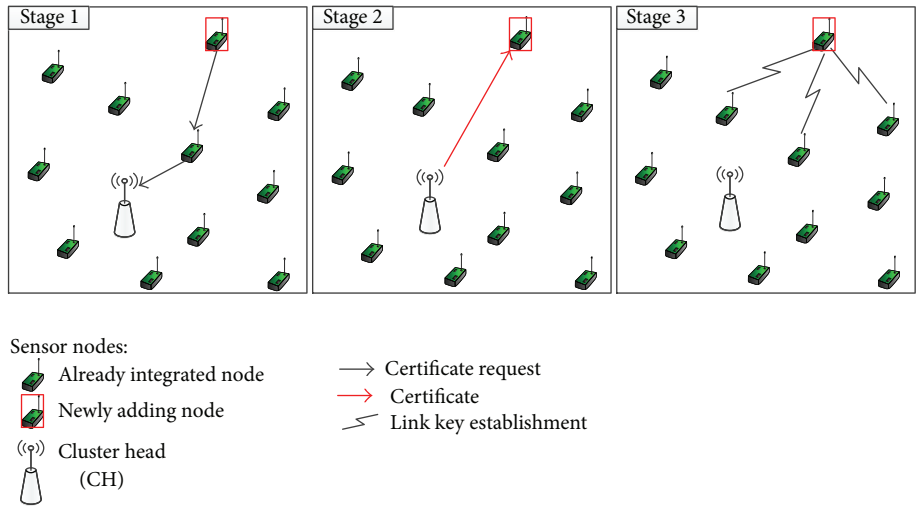


FIGURE 7: Behavior of the protocol when a new sensor node enters the cluster.

Similarly, the sensor nodes do not need prior knowledge of their neighbors. Whenever a new node is added to the network or changes the neighboring set, it can establish the ephemeral pairwise link keys with the corresponding neighbors using the certificate. The certificates always provide an implicit assurance for the sensor nodes that they are legitimate nodes. Even though the sensor nodes are frequently changing their locations (i.e., also the neighboring set), they can authenticate themselves and derive the pairwise keys securely without previous awareness of the new neighboring nodes or end-users. As shown in [31], if the authentication is performed based on pairwise keys between neighbors which are preinstalled, then there should be a large number of stored keys per node, which may not be desirable for large scale networks. However, in PAuthKey protocol, such a large scale key preinstallation is not needed at all since the ephemeral link keys have to be established before starting communication.

5.3. Security Analysis. The proposed implicit certificate based authentication protocol is developed using one of the lightest PKC schemes, ECC. Though it is comparatively more expensive than symmetric key algorithms, it is inherently secured due to the PKC characteristics. However, as shown in Sections 5.1 and 5.2, the proposed scheme is feasible to deploy in real-time WSNs. While using EC scalar-point multiplication, the scheme is provably secured under the random oracle model that the discrete logarithm problem over the subgroup is intractable. The advantage of using ECC is that it provides an equal security as RSA, however, with less overhead (e.g., 160-bit ECC equals RSA with 1024 key size). At the end of the authentication scheme there is a key establishment part which extends the security strength of the standard ECDH key agreement by using mutually authenticated keying materials (i.e., Cert_U and s). Since the authentication is performed using the implicit certificates of edge devices or users and the link keys are derived using two preevaluated values (i.e., d_U and Q_V), the legitimacy and trust between two parties are implicitly assured.

It is very common that denial of service (DoS) attacks can be launched against distributed IoT. Moreover, during the registration phase, the first `Hello` message contains the certificate requestor's identity, which is analyzed by CA. If the unauthorized requestors are trying to access, the CA can identify them at the beginning of identity verification and protect itself from DoS attacks. Similarly, during the authentication phase, the cryptographic credentials are exchanged only after the successful `Hello` message exchange, which provides protection from DoS attacks. In order to overcome illegal message alternations by malicious users and DoS attacks, the subsequent messages contain MAC with the common authentication key K for preserving data integrity. The availability of the proposed protocol is ensured by giving permission to two legitimate nodes, which possess the certificates granted from the CA, to authenticate each other and establish a secure pairwise key for their mutual communication. The freshness of the messages is guaranteed by appending true nonce. In the registration phase, the originator of the certificate (i.e., CA) cannot deny being sent the message (i.e., nonrepudiation property), since the CA uses its key pair to generate the certificate and private key reconstruction value (s). Likewise, during the authentication phase, the sender of the messages cannot deny that the messages are sent by itself since the receiver always uses the certificate of the sender to derive its (i.e., the sender's) public key.

In the security analysis, we are considering three attacks including node compromising attacks, masquerade attacks, and impersonate attacks. In node compromising attacks, an adversary can physically capture a node and obtain its keys. Similarly, in the authentication phase, an attacker can impersonate a legitimate sensor node using its certificate or try to masquerade the key establishment between two legitimate nodes.

5.3.1. Node Compromise Attacks. PAuthKey is resilient to node compromise attacks. If a sensor node U is captured, the adversary can reveal Cert_U, Q_U, d_U, and Q_{CA}. However,

with CA's public key, an adversary cannot create a new valid certificate and private key reconstruction data, since they are derived using d_{CA}, which is only known to CA. Though a node pretends to be a forgery CA and issues certificates with Q_{CA}, eventually the fake certificates and public keys are disclosed during the key establishment in the authentication phase (i.e., calculating $K_{UV} = d_U Q_V$). We assume that the CH can identify compromised nodes using beacon message technique, as explained in [20, 32]. Then, CA will broadcast the compromised node ID to the noncompromised nodes within its cluster and the other CHs and end-users, those who have contacted the CA for acquiring certificates to communicate with that compromised node. Upon receiving the CA's message (i.e., compromised node ID), the other sensor nodes, CHs, and end-users will discard or order to demolish the certificate of the corresponding node and the preestablished pairwise keys. Then, the compromised node cannot appear by itself as a legitimate node in the future because its certificate and user ID are already abandoned by the legitimate users.

5.3.2. Impersonation and Masquerade Attacks. During the key establishment in the authentication phase, nodes are authenticated in order to prevent impersonation attacks and masquerade attacks. Node V computes node U's public key using its Cert_U and the CA's public key Q_{CA}. The pairwise key K_{UV} calculation at both ends will be the same, only if the certificates are issued by the identical valid CA. Therefore, node V has an implicit assurance that the received certificate is genuine (i.e., issued by the CA). Likewise, when two legitimate nodes initiate a pairwise key, an attacker (without a valid certificate) cannot come in between them and masquerade the key establishment. Since the pairwise key is derived on the basis of certificates and private keys of both parties, an attacker is impossible to proceed in it by only using a valid certificate.

5.4. Comparison with Related Work. In this paper, the focus was on authenticating the extreme resource constrained devices, which are deployed in WSNs in distributed IoT applications. Therefore, the proposed authentication protocol was implemented on TelosB sensor nodes and performance measurements were obtained. However, as explained in Section 2, DTLS is considered the prominent authentication protocol for IoT applications. Though we use DTLS in the middle of PAuthKey protocol for certain scenarios (i.e., scenarios 2 and 3), the key foundation of the proposed authentication scheme is explained in scenario 1. In particular, in scenarios 2 and 3, DTLS is also utilized by resource-rich entities such as CHs and end-users for authentication.

As aforementioned in Sections 2 and 4, PAuthKey scheme is inspired by different ECC based security schemes. Among them, ECDSA and ECDH are the most relevant schemes to two phases of PAuthKey protocol. Therefore, the first assessment includes the assessment of PAuthKey scheme with the related work as depicted in Table 4. The memory and timing values of ECDSA digital signature scheme are compared with those of the registrations phase. Similarly, ECDH key establishment performance is contrasted with the key computation of the authentication phase. All the empirical results are measured on TelosB sensor nodes and with the activation ECC optimization techniques as mentioned in Section 5.1.

Table 4: Comparison of PAuthKey, ECDSA, and ECDH schemes.

	RAM (bytes)	ROM (bytes)	Time (ms)
Registration phase			
At the sensor node side	1410	11774	8231
At the CA side	2332	16576	8437
ECDSA scheme [23]	1586	12640	14789
Authentication phase (scenario 1)			
Key computation	1530	11650	5768
ECDH scheme [23]	1866	12102	6146

Table 5: Comparison of PAuthKey and DTLS scheme.

	DTLS scheme [18]	PAuthKey scheme (for scenario 1)
Memory consumption		
ROM	67 kB	22.875 kB
RAM	20 kB	2.871 kB
Time for authentication	4000 ms	8444 ms
Energy for authentication	939 mj	45.59 mj
Key size	2048 bits (RSA)	160 bits (ECC)

According to the given experimental results, the registration phase of PAuthKey scheme at the sensor node side consumes less memory than ECDSA scheme. Although the proposed scheme consumes higher memory values than ECDSA scheme, it would be tolerable for a resource-rich device. However, the execution times of registration phase at both ends (i.e., sensor node and CA) are less than the conventional ECDSA scheme. Similarly, the key computation of the authentication phase utilizes less memory and time than the ECDH scheme. In security aspects, conventional ECDH scheme is vulnerable to impersonation and masquerade attacks, since two communication parties do not have an authentication phase during the key establishment. However, the proposed key establishment is well secured at both types of attacks. Therefore, the given comparison results witness higher performing capability of PAuthKey scheme in the resource constrained sensor nodes than ECDSA and ECDH schemes.

The second assessment presents the comparison results between PAuthKey scheme and conventional DTLS scheme. Thereby, the appropriateness of the proposed protocol for the high resource restricted sensor nodes in WSNs is shown. We use the empirical values, which indicate the performance of DTLS, as given in [18] and the experimental results for PAuthKey. Table 5 shows the comparison results between DTLS scheme and PAuthKey authentication mechanism (i.e., for scenario 1).

The memory utilizations of PAuthKey are much better than the conventional DTLS scheme. This is a convincing remark, which confirms the applicability of PAuthKey scheme for the high resource constrained sensor nodes. Similarly, energy consumption for the authentication in PAuthKey scheme is notably fitting with low-power devices. According to the experimental results, the total time consumption for PAuthKey authentication is nearly double the value of DTLS authentication. However, this can be further reduced by using optimized EC arithmetic operations. Therefore, the authors of this paper believe that the proposed solution PAuthKey extends the existing pool of security solutions concerned with ECC and can optimize the key establishment in WSNs.

6. Conclusion

In this paper, the authors have introduced and analyzed an authentication and key establishment mechanism for WSNs in distributed IoT applications. The proposed PAuthKey protocol comprises two phases: *registration phase* for obtaining cryptographic credentials to the edge devices and end-users and *authentication phase* for authentication and key establishment in mutual communication. The authentication phase is described for three distinctive scenarios, based on the links between two communicating parties. Using PAuthKey protocols, the end-users can authenticate themselves to the sensor nodes directly and acquire sensed data and services. With the experimental results, it is shown that the authentication protocol is feasible to deploy in the low performing resource constrained network devices in WSNs. The protocol supports the distributed IoT applications, since the certificates are lightweight and can be handled by the high resource constrained devices, irrespective of their originality. According to the security analysis, the PAuthKey scheme is secured under certain types of attacks. Finally, a brief comparison between the conventional DTLS scheme and the proposed PAuthKey protocol is presented. This shows the appropriateness of PAuthKey scheme especially on the high resource constrained devices.

In the future, the authors intend to extend the utilization of implicit certificates for access control and multicasting in the massive scale distributed IoT network applications. It is expected to customize the content of the implicit certificates by adding other information, such as the time stamp, location identity, or 6LoWPAN identity, depending upon the application requirements. Furthermore, it is intended to extend the utilization of implicit certificates for group key management in large scale sensor networks.

Disclosure

Part of this work is published at the 10th IEEE International Conference on Embedded Software and Systems, 2013 [33], and the 14th IEEE Wireless Communication and Networking Conference, 2014 [34]. The extensions of this work include the authentication protocol between sensor nodes and the implementation and evaluation of PAuthKey protocol.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work has been supported by Tekes under Massive Scale Machine-to-Machine Service (MAMMotH) project and Academy of Finland project SEMOHealth. Pawani Porambage is also supported by HPY research foundation scholarship granted from Elissa Cooperation, Finland.

References

[1] J. P. Walters, Z. Liang, W. Shi, and V. Chaudhary, "Wireless sensor network security: a survey," in *Distributed, Grid, and Pervasive Computing*, X. Yang, Ed., p. 849, CRC Press, 2007.

[2] "IEEE Standard for Low-Rate Wireless Personal Area Networks (LRWPANs)," IEEE Std 802.15.4. 2011 (Revision of IEEE Std 802.15.4- 2006), 2011.

[3] "ZigBee Specification Version 1.0," ZigBee Alliance, 2008, http://www.zigbee.org/home.aspx.

[4] R. H. Weber, "Internet of things—new security and privacy challenges," *Computer Law and Security Review*, vol. 26, no. 1, pp. 23–30, 2010.

[5] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): a vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[6] M. Brachmann, S. L. Keoh, O. G. Morchon, and S. S. Kumar, "End-to-end transport security in the IP-based internet of things," in *Proceedings of the 21st International Conference on Computer Communications and Networks (ICCCN '12)*, pp. 1–5, IEEE, Munich, Germany, August 2012.

[7] R. Roman, J. Zhou, and J. Lopez, "On the features and challenges of security and privacy in distributed internet of things," *Computer Networks*, vol. 57, no. 10, pp. 2266–2279, 2013.

[8] N. Kushalnagar, G. Montenegro, and C. Schumacher, " IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs): overview, Assumptions, Problem Statement, and Goals," IETF RFC 4919, 2007, http://tools.ietf.org/html/rfc4919.

[9] G. Montenegro, N. Kushalnagar, J. Hui, and D. Culler, "Transmission of IPv6 packets over IEEE 802.15.4 networks," IETF R FC 4944, September 2007, http://tools.ietf.org/html/rfc4944.

[10] Z. Shelby, K. Hartke, and C. Bormann, "Constrained Application Protocol (CoAP)," IETF draft, R FC editor, 2013, http://tools.ietf.org/pdf/draft-ietf-core-coap-18.pdf.

[11] T. Winter, P. Thubert, A. Brandt et al., "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks," IETF RFC 6550, 2012, http://tools.ietf.org/html/rfc6550.

[12] K. Islam, W. Shen, and X. Wang, "Wireless sensor network reliability and security in factory automation: a survey," *IEEE Transactions on Systems, Man and Cybernetics C: Applications and Reviews*, vol. 42, no. 6, pp. 1243–1256, 2012.

[13] F. Zhu, M. W. Mutka, and L. M. Ni, "Private entity authentication for pervasive computing environments," *International Journal of Network Security*, vol. 14, no. 2, pp. 86–100, 2012.

[14] S. Shin, T. Shon, H. Yeh, and K. Kim, "An effective authentication mechanism for ubiquitous collaboration in heterogeneous

computing environment," *Peer-to-Peer Networking and Applications*, 2013.

[15] Y. Liu, J. Li, and M. Guizani, "PKC based broadcast authentication using signature amortization for WSNs," *IEEE Transactions on Wireless Communications*, vol. 11, no. 6, pp. 2106–2115, 2012.

[16] T. Kwon and J. Hong, "Secure and efficient broadcast authentication in wireless sensor networks," *IEEE Transactions on Computers*, vol. 59, no. 8, pp. 1120–1133, 2010.

[17] E. Rescorla and N. Modadugu, "Datagram Transport Layer Security," IETF RFC 4347, April 2006, http://tools.ietf.org/html/http://tools.ietf.org/html/rfc4347.

[18] T. Kothmayr, C. Schmitt, W. Hu, M. Brünig, and G. Carle, "DTLS based security and two-way authentication for the Internet of Things," *Ad Hoc Networks*, vol. 11, no. 8, pp. 2710–2723, 2013.

[19] K. Malasri and L. Wang, "Design and implementation of a secure wireless mote-based medical sensor network," *Sensors*, vol. 9, no. 8, pp. 6273–6297, 2009.

[20] R. Lu, X. Li, X. Liang, X. Shen, and X. Lin, "GRS: the green, reliability, and security of emerging machine to machine communications," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 28–35, 2011.

[21] "SEC 4: Elliptic Curve Qu-Vanstone Implicit Certificate Scheme (ECQV), Version 0.97," August 2013, http://www.secg.org/.

[22] V. Gupta, M. Wurm, Y. Zhu et al., "Sizzle: a standards-based end-to-end security architecture for the embedded Internet," *Pervasive and Mobile Computing*, vol. 1, no. 4, pp. 425–445, 2005.

[23] A. Liu and P. Ning, "TinyECC: a configurable library for elliptic curve cryptography in wireless sensor networks," in *Proceedings of the 7th International Conference on Information Processing in Sensor Networks (IPSN '08)*, pp. 245–256, St. Louis, Mo, USA, April 2008.

[24] X. H. Le, S. Lee, I. Butun et al., "An energy-efficient access control scheme for wireless sensor networks based on elliptic curve cryptography," *Journal of Communications and Networks*, vol. 11, no. 6, pp. 599–606, 2009.

[25] C. T. Li, M. S. Hwang, and Y. P. Chu, "An efficient sensor-to-sensor authenticated path-key establishment scheme for secure communications in wireless sensor networks," *International Journal of Innovative Computing, Information and Control*, vol. 5, no. 8, pp. 2107–2124, 2009.

[26] P. Kotzanikolaou and E. Magkos, "Hybrid key establishment for multiphase self-organized sensor networks," in *Proceedings of the 6th IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM '05)*, pp. 581–587, June 2005.

[27] W. Hu, H. Tan, P. Corke, W. C. Shih, and S. Jha, "Toward trusted wireless sensor Networks," *ACM Transactions on Sensor Networks*, vol. 7, no. 1, article 5, 2010.

[28] D. Hankerson, S. Vanstone, and A. J. Menezes, *Guide to Elliptic Curve Cryptography*, Springer, New York, NY, USA, 2004.

[29] "TelosB Datasheet," Tech. Rep., Crossbow Inc., 2013, http://www.datasheetarchive.com/4–Crossbow∗-datasheet.html.

[30] "TinyOS Documentation," 2013, http://www.tinyos.net.

[31] Y. Zhou, Y. Fang, and Y. Zhang, "Securing wireless sensor networks: a survey," *IEEE Communications Surveys and Tutorials*, vol. 10, no. 3, pp. 6–28, 2008.

[32] S. H. Jokhio, I. A. Jokhio, and A. H. Kemp, "Node capture attack detection and defence in wireless sensor networks," *IET Wireless Sensor Systems*, vol. 2, no. 3, pp. 161–169, 2012.

[33] P. Porambage, P. Kumar, C. Schmitt, A. Gurtov, and M. Ylianttila, "Certificate-based pairwise key establishment protocol for wireless sensor networks," in *Proceedings of 10th IEEE International Conference on Embedded Software and Systems (ICESS '13)*, pp. 667–674, Sydney, Australia, December 2013.

[34] P. Porambage, C. Schmitt, P. Kumar, A. Gurtov, and M. Ylianttila, "Two-phase authentication protocol for wireless sensor networks in distributed IoT applications ," in *Proceedings of the IEEE 14th International Conference on Wireless Communications and Networking (WCNC '14)*, pp. 2770–2775, April 2014.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 527928, 12 pages
http://dx.doi.org/10.1155/2014/527928

Research Article

A New Energy-Efficient Cluster-Based Routing Protocol Using a Representative Path in Wireless Sensor Networks

Hyunjo Lee, Miyoung Jang, and Jae-Woo Chang

Department of Computer Engineering, Chonbuk National University, Jeonju-si, Jeollabuk-do 561-756, Republic of Korea

Correspondence should be addressed to Jae-Woo Chang; jwchang@jbnu.ac.kr

Received 29 December 2013; Revised 28 May 2014; Accepted 3 June 2014

Academic Editor: Thomas Wook Choi

Wireless sensor networks (WSNs) have been broadly studied with advances in ubiquitous computing environment. Because the resource of a sensor node is limited, it is important to use energy-efficient routing protocol in WSNs. The cluster-based routing is an efficient way to reduce energy consumption by decreasing the number of transmitted messages to the sink node. LEACH is the most popular cluster-based routing protocol, which provides an adaptive cluster generation and cluster header rotation. However, its communication range is limited since it assumes a direct communication between sensor nodes and a sink node. To resolve this problem, we propose a new energy-efficient cluster-based routing protocol, which adopts a centralized clustering approach to select cluster headers by generating a representative path. To support reliable data communication, we propose a multihop routing protocol that allows both intra- and intercluster communications. Based on a message success rate and a representative path, the sensor nodes are uniformly distributed in clusters so that the lifetime of network can be prolonged. Through performance analysis, we show that our energy-efficient routing protocol outperforms the existing protocols up to 2 times, in terms of the distribution of cluster members, the energy consumption, and the reliability of a sensor network.

1. Introduction

Wireless sensor networks (WSNs) have been broadly studied in ubiquitous computing environment because of its widespread utilization. The application area of WSNs includes environmental management, health-care services, and military monitoring [1–3]. WSNs are composed of many sensor nodes equipped with processors, memory, and short-range wireless communication. In real applications, the sensor nodes are distributed in the areas of interest, and they sense data from surrounding environments. The sensor nodes cooperate with each other to transmit the sensed data to the central base station, called sink node. A routing protocol is a way of determining a path between a source node and a destination (i.e., sink node) for sensed data transmission. The efficiency of WSNs is highly dependent on routing protocols that directly affect the network lifetime. The main objective of routing protocols is to enhance both reliability and lifetime of WSNs by considering the capability of a sensor node with resource constraints, such as limited power, slow processor, and low communication bandwidth. Hence, the challenging issue of routing protocols is to reduce the communication overhead for data transmission by determining an optimal path.

Clustering is one of the most popular techniques for routing protocols. The cluster-based routing is an efficient way to reduce energy consumption within a cluster by decreasing the number of transmitted messages to the sink node. Hence, there have been many researches on cluster-based routing protocols [4–9]. A popular cluster-based protocol, called LEACH [4], proposes a two-phase operation based on a single-tier network using clusters. LEACH randomly selects a portion of nodes as cluster headers, and the cluster headers gather the neighboring nodes to construct clusters. Each node forwards its sensed data to a cluster header, which collects and delivers data to the sink node. There are several extensions of the LEACH protocol to increase energy efficiency, but the existing protocols have some problems. First, they assume that all sensor nodes can transmit data to the sink node with enough power and network capability. However, this

assumption is not applicable to sensor networks deployed in large regions. In real world applications, the obstacles (i.e., wall) should be considered to provide reliable network communication. Secondly, the random selection of a cluster header causes the skewed distribution of clusters. In general, because the randomly selected cluster headers are not uniformly distributed over the network, some nodes will fail to find a cluster header within their communication range for WSNs.

In this paper, we propose a new energy-efficient cluster-based routing protocol, which adopts a centralized clustering approach to select cluster headers by generating a representative path. Therefore, the burden of network configuration and routing from sensor nodes can be greatly reduced. By using a representative path, the sink node selects cluster headers and generates clusters in a distributed manner. First, we use message success rate for a representative path creation. In initializing phase, a sink node sends flooding messages to gather each sensor node's information. Upon receiving flood message, a sensor node searches for neighboring nodes in its vicinity and sends advertisement messages to the neighbors. Each node receiving the advertisement messages sends back an acknowledge signal. The message success rate of a sensor node is calculated using the number of messages received from the other sensor nodes. The nodes with high message success rate are likely to be included in a representative path. To support reliable data communication, we devise a multihop routing protocol that allows both intra- and intercluster communications. Finally, we propose both cluster split and merge algorithms to maintain the efficiency of network.

The rest of the paper is organized as follows. In Section 2, we present related work. In Section 3, we propose energy-efficient routing protocol in WSNs. Section 4 provides the performance analysis of the proposed approach, in terms of power consumption and communication reliability. Section 5 concludes this paper with final remarks and future directions.

2. Related Work

Many researches have proposed different protocols for energy-efficient routing in WSNs. In general, the routing protocols for WSNs can be divided into flat-based routing [10–13], cluster-based (hierarchical-based) routing [4–9], and location-based routing [14–17], depending on the network structure. In flat-based routing, all nodes are typically assigned equal roles or functionality. In cluster-based routing, however, nodes play different roles in the network. In location-based routing, sensor nodes' positions are exploited to route data in the network. Among the existing protocols, the cluster-based routing protocol is particularly more suitable for continuous data transmission in WSNs. In this section, we present an abbreviated overview of the well-known cluster-based routing protocols for WSNs, along with their limitations.

Heinzelman et al. [4] introduced a hierarchical clustering algorithm for sensor networks, called low energy adaptive clustering hierarchy (LEACH). The idea is to randomly select a few sensor nodes as cluster headers and rotate this role to evenly distribute the energy load among the sensors in the network. The LEACH protocol includes two stages of operation: node clustering and information transmission. First, in the clustering stage (Figures 1(a) and 1(b)), a fraction p of all sensor nodes are chosen to serve as cluster headers. A node that produces the random number being smaller than threshold is selected as cluster header. The other nodes are allocated to the cluster header closest to them. Second, in the information transmission stage (Figures 1(c) and 1(d)), the cluster headers aggregate the data received from their cluster members and send the aggregated data to the base station by single hop communication. LEACH outperforms traditional clustering algorithms by using adaptive clusters and rotating cluster header, which can distribute energy consumption among all the sensor nodes. In addition, LEACH can perform local computation so that the amount of transmitted data can be reduced. However, LEACH assumes direct communication between a node and a base station. This is a high-power operation and shortens the lifetime of the network. Moreover, the random selection of headers does not guarantee optimal cluster construction and may cause rounds of communication when cluster headers are not available.

Heinzelman et al. [4] presented LEACH-C (low energy adaptive clustering hierarchy-centralized) in order to distribute cluster headers evenly over the network and reduce energy dissipation. During the initial stage, each node sends to the sink node information about its current location and energy level. Therefore, sensor nodes whose remaining energy is below the average energy are excluded from becoming a cluster header. For each round, the sink node runs an optimization algorithm to determine cluster headers and to divide the network into clusters. Because LEACH-C requires the position of each node at the beginning of each round, an expensive global positioning system (GPS) is required for sending the position information. In addition, the number of nodes for each cluster is not guaranteed when forming clusters.

Farooq et al. [5] proposed a multihop routing with low energy adaptive clustering hierarchy (MR-LEACH). MR-LEACH partitions the network into different layers of clusters, based on the distance between a sensor node and a sink node. Cluster headers are chosen by the LEACH protocol and transmit the aggregated data to a sink node by using multihop routing. Therefore, it achieves significant improvement on energy consumption, compared with the LEACH protocol. The problem of MR-LEACH is that the selection of a cluster header in a layer solely depends on the energy residue of a sensor node, without considering distances among cluster headers. Figure 2 depicts an example of MR-LEACH where each cluster is generated in (a) and (b) by the multihop communication from a cluster header to a sink node.

Distance-energy cluster structure algorithm (DECSA) [6] was proposed by considering both the distance and residual energy of nodes. Resulting from multihop communication between the base station and cluster heads, DECSA reduces the energy consumption of the cluster head. Javaid et al. [7] proposed Enhanced Developed Distributed Energy-Efficient

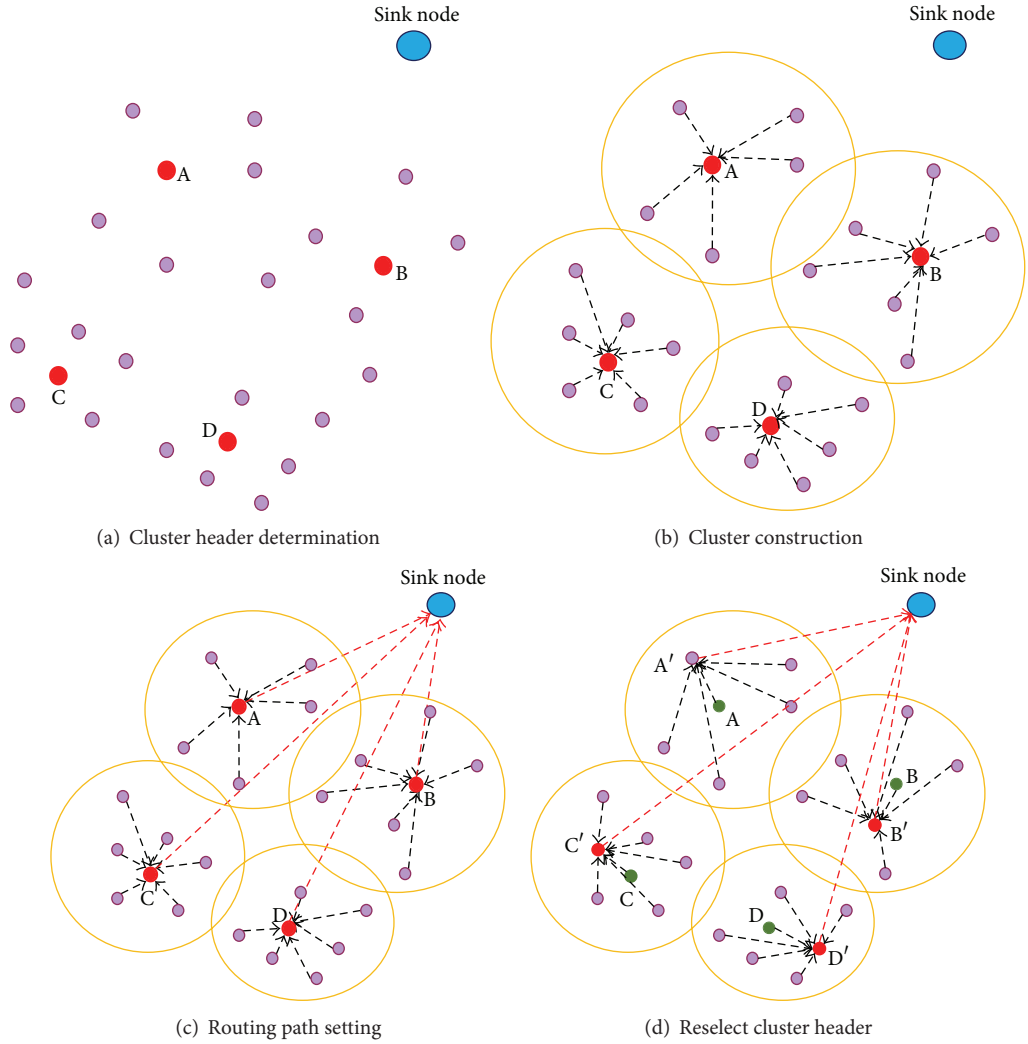


Figure 1: LEACH protocol.

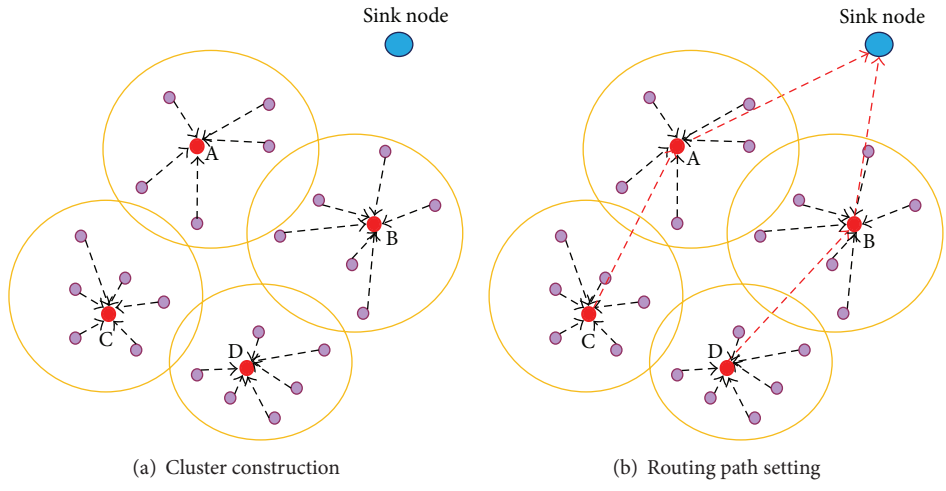


Figure 2: MR-LEACH protocol.

Clustering (EDDEEC), which selects a cluster header based on energy level of the nodes. To distribute the equal amount of energy between sensor nodes, EDDEEC dynamically changes the probability of nodes to be chosen as a cluster header in a balanced way. Jorio el al. [8] proposed K-Way spectral clustering protocol, which uses the principle of spectral clustering. The K-Way clustering approach groups sensor nodes into K disjoint classes based on the Laplacian matrix and its eigenvectors. In this way, it calculates a data transmission path to the base station in an energy-saving manner. Nikolidakis et al. [9] proposed Equalized Cluster Head Election Routing Protocol (ECHERP) that aims at energy conservation through balanced clustering. In order to prolong the network lifetime, ECHERP models a network as a linear system by using the Gaussian elimination algorithm and selects the good combinations of cluster heads for minimizing the energy consumption for data transmission.

3. Energy-Efficient Cluster-Based Routing Protocol

The major problem of LEACH [4] is that it selects cluster headers using the probabilistic approach and so it cannot consider a node's residual energy and its location to the sink node. Moreover, because LEACH assumes a direct connection between a cluster header and a sink node, it limits the size of sensor networks. This kind of approach is not appropriate for an energy-efficient routing protocol in WSNs. To support both energy-efficient routing and a wide range of network connectivity, an algorithm is required to consider a node's connectivity with neighboring nodes and its distance to the sink node.

We propose an energy-efficient cluster-based routing protocol using a representative path. To generate a representative path, we adopt a centralized clustering approach by using the message success rate of a sensor node. By using a representative path, a sink node selects cluster headers and generates clusters in a distributed manner. Therefore, the burden of network configuration and routing from sensor nodes can be greatly reduced. Ideally, nodes with high connectivity should become a cluster header to increase the lifetime of network. Our energy-efficient routing protocol consists of four phases: network information table generation, representative paths construction, cluster generation, and cluster management.

3.1. Network Information Table Generation. In this phase, the network information table of every node is generated by using a message flooding technique [18]. First, a sink node broadcasts flooding messages periodically based on the duration of message flooding and the maximum number of flooding data packets. If a node receives the flooding message, it updates its node information into the sink node, as shown in Table 1. We use a message success rate (MSR) to guarantee the reliability of data communication. For initial data transmission to the sink node, a sensor node selects a parent node as one having the highest MSR among its neighboring nodes. Hence, our protocol can reduce data loss caused by the limited communication range of sensor nodes. Equation (1) shows how the message success rate is calculated where i is a node ID, t is a message flooding duration, *PacketsExpected* means the maximum number of flooding data packets in t, and *PacketsReceived* means the number of messages from the node i in t. Here, *PacketsExpected* is calculated by dividing the message duration t with a predefined time interval between queries. For calculating *PacketsReceived*, the algorithm counts the number of received messages during flooding messages as follows:

$$\text{Average Success Rate}(i) = \frac{PacketsReceived(i \cdot t)}{\max\left(PacketsExpected(i,t), PacketsReceived(i,t)\right)}. \quad (1)$$

Table 1: Node information for a sensor node.

Information	Description
NodeID	ID of a node
HopCount	Minimum number of links from a node to a sink node
NeighborCount	Number of neighbor nodes
Neighbor nodes (ID, MegSuccRate)	Information of neighbor nodes including ID and MSR
ChildnodeCount	Number of child nodes
Child nodes (ID, MegSuccRate)	Information of child nodes including ID and MSR
InitiaParentnodeID	ID of parent node selected initially

Secondly, the initial network is formed to forward each node's information to the sink node. A node selects its parent node as one which has the smallest hop count and the highest MSR among its neighboring nodes. Then, the node sends its network information to the sink node through the parent node. Based on the received information, the sink node generates the whole network information table. Figure 3 shows an example of a sensor network with node information where a dotted edge connecting two nodes shows its message success rates (MSR). Each sensor node selects its parent node using the MSR in the initial stage. For example, since the neighboring nodes of node 15 are nodes 11, 14, 18, and 19 and node 11 has smaller hop count than its own, node 15 selects node 11 as a parent node.

Algorithm 1 shows an algorithm for the network information table generation. The algorithm consists of two parts: a sensor node part (lines 1–6) and a sink node part (lines 7–12). In the sensor node part, each sensor node updates its node information by exchanging flooding messages (lines 1–4). When message flooding is over, all sensor nodes send their node information to the sink node through the parent nodes (line 5). On the other hand, the sink node periodically broadcasts a flooding message for maintaining its network information table up to date (lines 9–12).

3.2. Representative Paths Construction. In this phase, the sink node constructs representative paths (RPs). The representative path means a set of nodes, which have been selected from

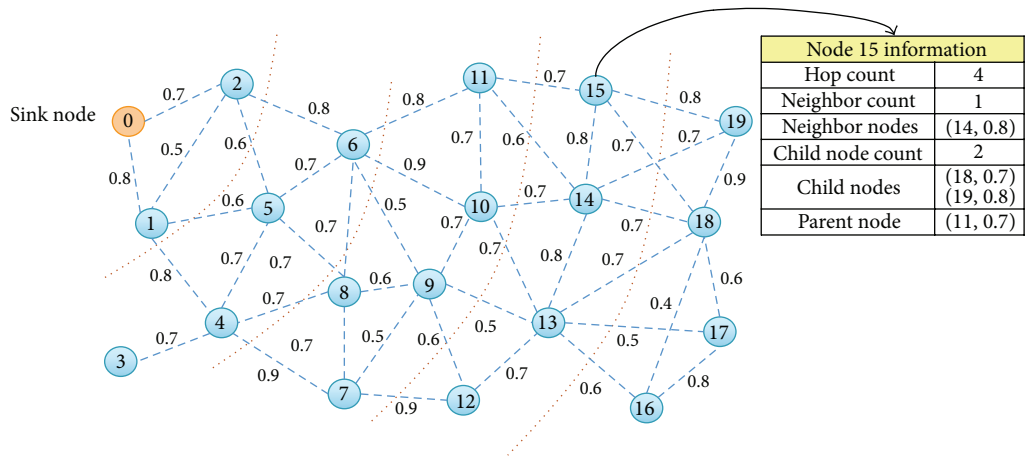


Figure 3: Example of sensor network with node information.

Network information generation algorithm
T_f: Time duration for flooding messages
$numMsg_f$: Maximum number of messages in T_f

```
(1)  Sensor Node(Message msg){
(2)      if (expiredTime < T_f) then
(3)          if (curHopCnt > msg.HopCnt + 1) then curHopCnt = msg.HopCnt + 1
(4)          UpdateNetInfo and flood message
(5)      else    SendNetInfo(parentID, NetInfo(nodeID));
(6)  }
(7)  Sink Node(Message msg){
(8)      while (expiredTime < T_f) FloodingMsg();
(9)      If Receive Message(NetInfo) from Sensor Node then {
(10)         For each entry i of NetInfo
(11)         NetInfoTable[NetInfo.id][parent[i]] = NetInfo.HopCnt[i]
(12) }}
```

Algorithm 1: Network information table generation algorithm.

each hop using their connectivity. The selected node from a hop, called an anchor node, has better connectivity than its neighboring nodes. For example, there are three neighboring nodes A, B, and C in the first hop and their connectivities are A > B > C. Because node A has the highest connectivity among them, it becomes an anchor node. The definitions of anchor node and representative path are as follows.

Definition 1 (anchor node and representative path). There are k number of neighboring nodes in the hth hop ($1 \le h \le m$) = $\{N_{h1}, N_{h2}, \ldots, N_{hk}\}$:

(i) anchor node (AN_1) in the 1st hop = $\{N_{1i} \mid \forall j$, Connectivity $(N_{1i}) \ge$ Connectivity (N_{1j}), where i, j are the integer, $i \ne j$, $i \le k$, and $j \le k\}$;

(ii) AN_h in the hth hop ($h > 1$) = $\{N_{hi} \mid N_{hi}$ and N_{hj} are the child nodes of the anchor node $AN_{h-1}, \forall j$, Connectivity $(N_{hi}) \ge$ Connectivity (N_{hj}), where i, j are the integer, $i \ne j$, $i \le k$, and $j \le k\}$;

(iii) representative path (RP) = $\{AN_1, AN_2, \ldots, AN_m\}$.

To construct a representative path, our protocol measures the connectivity of nodes by using (2). When a node has the higher connectivity, it has the higher reliability of communication since it has the higher message success rate than its neighboring nodes. In (2), *c*node means a current node, node(i) means a neighboring node i of the current node, *MsgSuccRate* means a message success rate (MSR) between the current node and node(i), *NeighborMsgSuccRate* means the MSR between node(i) and node(j), and *ChildnodeCount* means the number of child nodes of node(i). In addition, α, β, γ mean weights for three parts of (2), where the sum of α, β, and γ is 1:

$$\begin{aligned} &\text{Connetivity}_{\text{cnode,node}(i)} \\ &\quad = \alpha \times MsgSuccRate\,(current, node\,(i)) \\ &\qquad + \beta \times \sum \left(NeigborMsgSuccRate_{\text{node}(i),\text{node}(j)}\right) \\ &\qquad + \gamma ChildnodeCount_{\text{node}(i)}. \end{aligned} \tag{2}$$

First, for the nodes in the first hop, the sink node calculates their connectivity. Secondly, a node with the highest

connectivity value is selected as an anchor node for the first representative path. Thirdly, our protocol computes the MSR of the neighboring nodes of the first anchor node. Based on (2), the higher MSR between two nodes can guarantee the better reliability of network communication. If there is a neighboring node whose connectivity value is less than a threshold, the neighboring node is excluded from an anchor candidate. In the same manner, the sink node selects all anchor nodes for a representative path.

Algorithm 2 shows an algorithm for constructing a representative path. First, the sink node selects the node with the highest connectivity from the neighboring anchors. By expanding hop distance, it generates a representative path with the selected nodes (lines 1–10). When the number of representative paths is more than two, the algorithm checks whether or not the selected node is one of the member nodes of other representative paths. If the node is already selected, the algorithm excludes the node and reselects another node from candidates (lines 12-13). The algorithm is terminated when all the representative paths are constructed.

Figure 4 shows an example of representative path construction. In this example, we set the weights of α, β, and γ to 0.4, 0.3, and 0.3, respectively, and the initial threshold of MSR is given to be 0.6. First, the sink node measures the connectivity of 1-hop nodes, that is, nodes 1 and 2, from the sink node. Based on (2), the connectivity between nodes 0 and 1 is calculated as 1.22 and the connectivity between nodes 0 and 2 is calculated as 1.18. Since the MSR between node 1 and node 2 is 0.5, the sink node computes the total number of representative paths as 2. Between two candidate nodes, node 1 is assigned to the anchor of the first representative path (RP_1) because node 1's connectivity value is greater than that of node 2. By expanding RP_1 using the connectivity of child and neighboring nodes, RP_1 is composed of six nodes, that is, 1, 4, 7, 12, 13, and 18. When constructing the second representative path, there is a problem of selecting the same node as anchor at node 10 and node 14. Node 10 selects node 14 as an anchor node, instead of node 13, because node 13 already belongs to RP_1. Similarly, node 14 selects node 19 as an anchor node. As a result, RP_2 is composed of five nodes, that is, 2, 6, 10, 14, and 19.

3.3. Cluster Generation. In the cluster generation phase, our protocol selects cluster headers from the anchor nodes of representative paths and broadcasts a cluster-join message to every sensor node. For minimizing the communication cost, it is important to minimize distances from cluster headers to their cluster members. For this, our protocol considers the possible combinations of anchor nodes of representative paths. Based on the existing work [4], the number of cluster headers is recommended to be set to 5% of the total number of sensor nodes in the network. By using both the number of cluster headers and the number of representative paths, the sink node first calculates the possible combinations for selecting header nodes. For example, we assume that the total number of sensor nodes is 60 and there are two representative paths, RP_1 and RP_2. In this case, since the number of header nodes should be set to 3, we can choose n and 3-n number of anchor nodes from RP_1 and RP_2, respectively, where $0 \leq n \leq 3$. Thus, the possible combinations can be represented as $\{(3,0),(2,1),(1,2),(0,3)\}$, where, in (x, y), x means the number of the selected anchor nodes from RP_1 and y means the number of the selected anchor nodes from RP_2. If RP_1 and RP_2 have 5 anchor nodes and 4 anchor nodes, respectively, the total number of possible combinations is calculated as follows:

$$\begin{aligned}&\left({}_5C_3 * {}_4C_0\right) + \left({}_5C_2 * {}_4C_1\right) + \left({}_5C_1 * {}_4C_2\right) + \left({}_5C_0 * {}_4C_3\right)\\ &\quad = 10 + 40 + 30 + 4 = 84.\end{aligned} \tag{3}$$

Since the computation of all possible combinations for cluster generation is a time-consuming and energy-inefficient one, we propose an approach to filter out some nodes based on hop distances. A candidate node whose hop distance is less than the average can be excluded from the possible combinations. Our protocol measures the communication cost of each combination. The communication cost is calculated based on the hop distances of the shortest path from a sensor node to its nearest header node and a subpath from the cluster header node to the sink node in a representative path. Then our protocol selects a combination with the least cost.

Algorithm 3 shows a cluster generation algorithm. First, the sink node makes possible combinations by using the network information table (lines 1–3). Secondly, the communication cost for each combination is calculated (lines 6–12). Thirdly, the combination with the least cost is selected among them. Finally, the cluster headers send cluster-join messages to all the sensor nodes (line 4).

Figure 5 shows an example of selecting cluster header nodes. Here we assume that the maximum hop distance from a node to its nearest header is only 2. The dotted edge connecting two nodes shows the connectivity of two nodes. Because node 6 and node 13 have the highest connectivity, we consider the combinations that include nodes 6 and 13. In this time, because the threshold of hop distance is 2, we can filter out the combinations that include the neighboring nodes of nodes 6 and 13. Therefore, the combinations (1, 6, 13), (4, 6, 13), (6, 7, 13), and (6, 13, 19) are left as candidates for cluster headers. Finally, among them, nodes 4, 6, and 13 are selected as cluster headers based on the communication cost and connectivity.

3.4. Cluster Management. Cluster head replacement is required to prolong a network lifetime by evenly distributing energy load among sensor nodes. To achieve this, we provide a technique for periodic header replacement and reconfiguration in a cluster. It is obvious that a cluster header consumes a lot more energy than the member nodes of a cluster. Both LEACH and MR-LEACH incur the skewed distribution of clusters because they do not consider the distance between cluster member nodes. Thus, in a dense network, a cluster head is frequently reselected or even the reconstruction of the whole network is required. Our protocol replaces a cluster header periodically by considering the connectivity of nodes in representative paths and their estimated energy residue. Algorithm 4 shows our cluster

```
Make representative path algorithm
maxHop: maximum hop counter
(1) Sink node(NetInfo){
(2)     If nodeID.HopCnt == 1 then Anchors[] = SetAnchor(NetInfo);
(3)     For each node in Anchors[]{
(4)         Set initial RPPath(i) with selected nodes;
(5)     }
(6)     For all RPPath(i){
(7)         If (RPPath(i).CurAnchor → hop ≥ maxhop) then break;
(8)         cand = SetCand(i, RPPath(i).CurAnchor);
(9)         MeasureWeight(cand);
(10)    }
(11)    nextAnc = SelectNextAnchor(i, cand);
(12)    If nextAnc exists in RPpath(i) then nextAnc = ReselectNextAnchor(cand);
(13)    RPPath(i).CurAnchor = nextAnc;
(14) }
```

Algorithm 2: An algorithm for constructing representative paths.

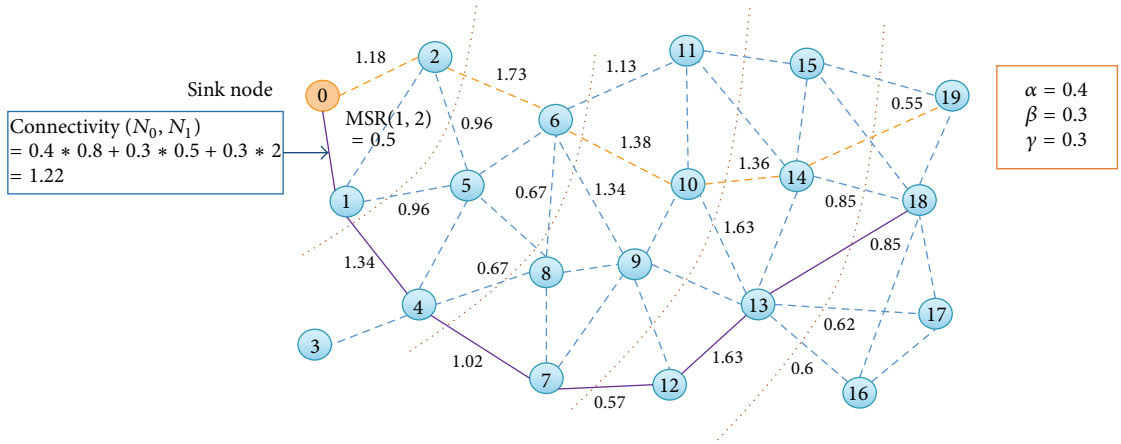


Figure 4: An example of constructing representative paths.

management algorithm. The algorithm periodically checks the estimated energy residue of header nodes (lines 2–5). If there is a header node whose energy level is below the given threshold, the header node is replaced by another anchor node in the representative path. Then, a network modification message is sent to all the sensor nodes (lines 6–9).

4. Performance Analysis

4.1. Comparison. We compare our cluster-based routing scheme with the existing routing protocols, such as LEACH, LEACH-C, MR-LEACH, DESCA, EDDEEC, K-Way, and ECHERP, in terms of mobility of node, cluster header selection, cluster header rotation, supporting multilevel hop, and usage of GPS. Table 2 shows the comparison results. Firstly, because DESCA, K-Way, and ECGERP consider the whole network as a static system, they cannot support the mobility of the node. Secondly, only ECHERP does not support the rotation of the cluster headers since it reconstructs all the clusters when the cluster headers have low energy residue. Thirdly, because MR-LEACH, DESCA, ECHERP, and our protocol can provide a multihop based communication, they can reduce the communication cost by decreasing the distance between the nodes, compared with the other protocols. Finally, even though LEACH-C and K-Way can provide the more efficient cluster optimization algorithm than the other protocols, they cannot be utilized for the typical WSN. This is because they make use of GPS, instead of using the sensors. According to the comparison of routing protocols, it is shown that our protocol is one of the best ones in the typical WSN environment because our protocol not only supports the mobility of the node, but also provides a multihop based communication, without using GPS-equipped sensor nodes.

In our experiment, we compare the performance of our protocol with LEACH and MR-LEACH. We exclude some protocols for the following reasons. First, LEACH-C and K-Way are excluded in our experiment since they utilize GPS-equipped sensor nodes. Secondly, EDDEEC is excluded since it cannot support the multilevel hop based communication. As a competitor, we select LEACH due to its popularity,

```
MakeCluster(NetInfo, *PathList[PathCnt], NumPath){
(1)   For each Path p from NumPath{
(2)     Comm = Combination(PathCnt, TotalHead, PathCnt, ChooseHead)
(3)     If Comm < MinComm then MinComm = Comm
(4)   } FindNNHead(ChooseHead, NetInfo)
(5) }
(6) Combination(PathCnt, TotalHead, q){
(7)   if (r == 0) then add_array(q);
(8)   else{
(9)     buf_route[r − 1] = route[0][n − 1];
(10)    combi(n − 1, r − 1, q);
(11)    combi(n − 1, r, q);
(12) }}
```

Algorithm 3: An algorithm of generating cluster.

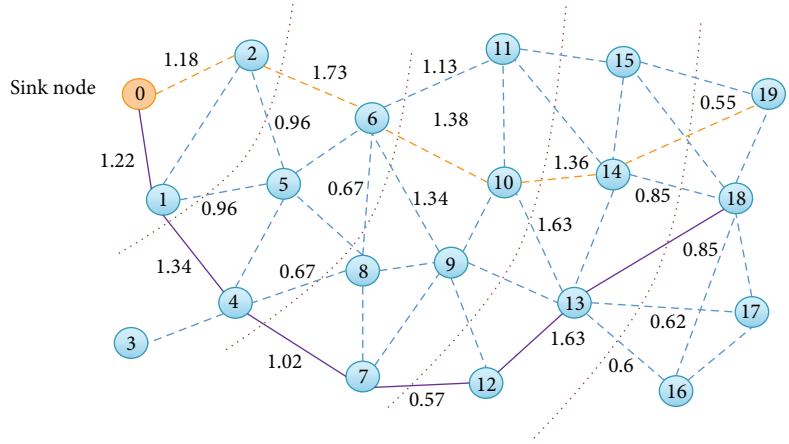


Figure 5: An example of selecting cluster headers (three cluster headers: nodes 4, 6, and 13).

even though it cannot support the communication based on multilevel hop. Thirdly, DESCA and ECHERP are excluded in our experiment because it is impossible to add and remove the nodes in DESCA and ECHERP. It is necessary to support the mobility of nodes in the real world environment since communication problems can occur by obstacles, node failures, and so on.

4.2. Experimental Setup. We implemented our scheme with NesC (network embedded system C) and ran our experiment on TOSSIM simulator [19] in TinyOS using an Intel Xeon 3.0 GHz with 2 GB RAM. We set the maximum communication range of sensor nodes to 25 m, which is the maximum communication range of MicaZ. We consider a scenario where every sensor node generates a message and sends it to the sink node. We use three different types of network distribution, that is, uniform, Gaussian, and skewed distribution, where 100 nodes are distributed within 100 m $*$ 100 m network area. Figure 6 shows three types of sensor node distribution.

We experimentally analyze the optimal weights for choosing a cluster header, discussed in Section 3, for the uniform cluster generation. Figure 7 shows the standard deviation of the number of cluster members for the different weights of α, β, and γ, for instance, 127 meaning $\alpha = 0.1$, $\beta = 0.2$, and $\gamma = 0.7$. Here α is the number of neighboring nodes, β is a message success rate with neighboring nodes, and γ is the number of hops for a sensor node. The total sum of α, β, and γ is 1. As shown in Figure 7, the standard deviation of the number of cluster members is the smallest when α, β, and γ combinations are 3, 2, and 5. So, in this paper, we use the weights of $\alpha = 0.3$, $\beta = 0.2$, and $\gamma = 0.5$ as the optimal weights for choosing a cluster header. In the case when the distribution of nodes is different from that of nodes in our experiment, it is required to set the default values of α, β, and γ. For this, we calculate the averages of the weights being the local optimization values. As a result, we set the default values of α, β, and γ as 0.33, 0.31, and 0.36, respectively.

4.3. Energy Efficiency of Routing Protocols. The primary goal of our scheme is to reduce the communication cost of routing protocol so as to prolong the lifetime of a sensor network. In general, a sensor node has a waiting status, a receiving one, and a transmitting one. Among them, the data transmission status causes the most significant energy loss, since it consumes more than 80% energy of the sensor node.

```
(1)  ManageCluster(NetInfo){
(2)      RoundCheck = CurrRound %/Header Update Interval
(3)      If RoundCheck = 0 then {
(4)          Header = ChangeHeader()
(5)      }
(6)      If Energy of Header < Energy of Threshold then {
(7)          Type of Message = ReConstruction of Routing Table
(8)          MakeNetInfo(Message)
(9)      }
(10) }
```

Algorithm 4: An algorithm of cluster management.

Table 2: Comparison of the protocols.

Protocol	Mobility of nodes	CH selection	CH rotation	Supporting multilevel hop	Usage of GPS
LEACH [4]	Limited	Probability/random	Yes	No	No
LEACH-C [4]	Limited	Probability/random	Yes	No	Yes
MR-LEACH [5]	Limited	Probability/random	Yes	Yes	No
DESCA [6]	No	Distance and energy residue	Yes	Yes	No
EDDEEC [7]	Limited	Random	Yes	No	No
K-Way [8]	No	Random	Yes	No	Yes
ECHERP [9]	No	Random	No	Yes	No
Our protocol	Limited	Connectivity	Yes	Yes	No

Because the amount of transmission energy consumption is influenced by a distance between a sender and a receiver, it is important to have a short communication distance from a sensor node to its parent node in the network hierarchy. In our experiment, we evaluate four measures: the standard deviation of the number of cluster members in a cluster, the average communication range of a sensor node, the lifetime of a sensor node, and the lifetime of the sensor network.

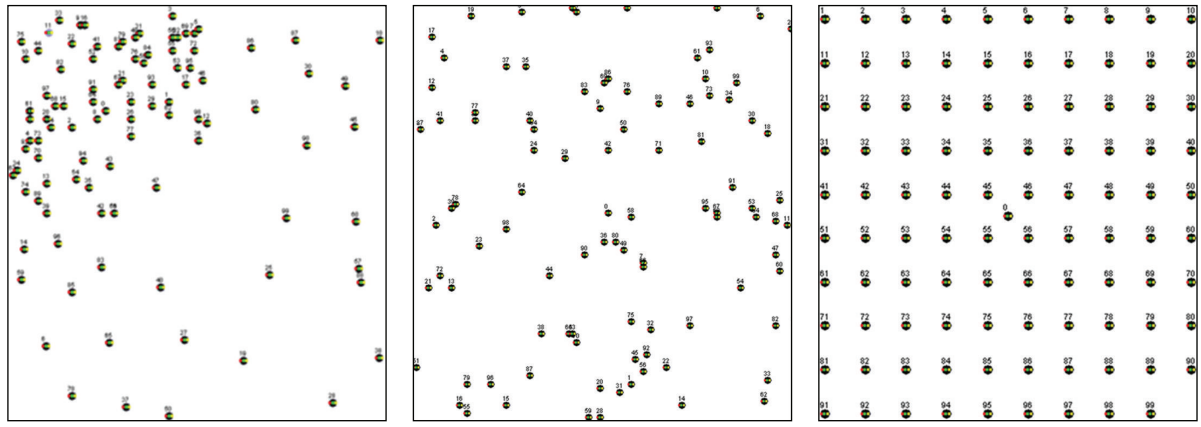
Because imbalanced node distribution incurs a huge amount of energy consumption, we measure the standard deviation of hop distance among the nodes in a cluster. As shown in Figure 8, for all the distribution cases, our scheme has much lower deviation for the number of cluster members than the existing methods. In the uniform network distribution, our scheme achieves 8.07 on deviation whereas LEACH and MR-LEACH have 9.54 and 9.48, respectively. In particular, MR-LEACH generates highly dense clusters in the skewed distribution because the cluster head selection is solely dependent on the energy residue. Our scheme outperforms the existing methods because it performs both cluster splitting and merging algorithms for the optimal cluster construction.

In Figure 9, for random data distribution, the average communication distance of a sensor node is 17.5 m in our scheme, whereas those of LEACH and MR-LEACH are 22.1 m and 21.6 m, respectively. Our scheme has the shortest communication distance because it chooses a cluster header by considering the connectivity of nodes, hop counts, and the number of neighboring nodes, whereas LEACH selects a cluster header using random selection and MR-LEACH chooses a cluster header using the energy residue of a sensor node.

Figure 10 shows the lifetime of a sensor node. In our experiment, we define a round as the period when all the sensed data from sensor nodes are aggregated into a sink node. Because LEACH performs a direct connection between a node and its cluster header, it shows the worst performance among three protocols. For uniform distribution, the average lifetime of sensor nodes for our protocol is 130.3 rounds, whereas those of LEACH and MR-LEACH are 54.1 and 77.8 rounds, respectively. Even though MR-LEACH provides multihop communication between cluster headers and a sink node, our protocol outperforms MR-LEACH because it generates uniformly distributed clusters with the shortest communication distance between sensor node and its header.

For the network lifetime, we measure the number of disconnected sensor nodes in network due to their energy consumption for each round. Figure 11 shows the lifetime of the sensor network with respect to LEACH, MR-LEACH, and our protocol. By following [20], a sensor network is no longer effective and available when the sensor network has less than 50% active sensor nodes in the network. When the number of rounds is 100, there is no active sensor node in LEACH, whereas MR-LEACH and our protocol have 49% and 87% of active sensor nodes, respectively. Because our protocol has the shortest communication distance, it improves the network lifetime up to 150% compared to LEACH and MR-LEACH.

To verify the reliability of routing protocols, we evaluate the response rate of a sensor node under varying communication ranges from 0 m to 150 m. By sending hello messages



(a) Skewed (b) Random (c) Uniform

Figure 6: Sensor node distribution.

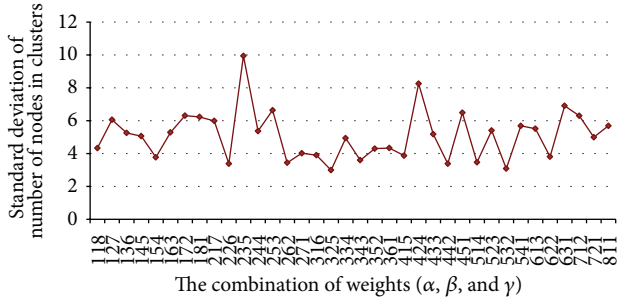


Figure 7: Standard deviation of the number of cluster members for the different weights.

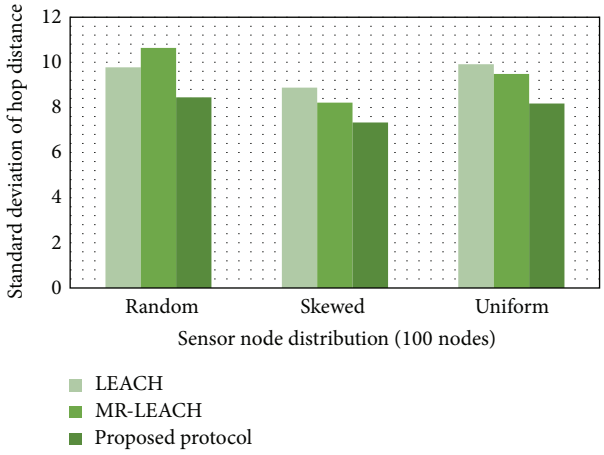


Figure 8: Standard deviation of hop distance among the nodes in the same cluster with varying network distribution.

to all the sensor nodes including cluster headers, the sink node calculates the ratio of the number of the received acknowledge messages to the number of sensor nodes. As shown in Figure 12, when the communication range is 15, the

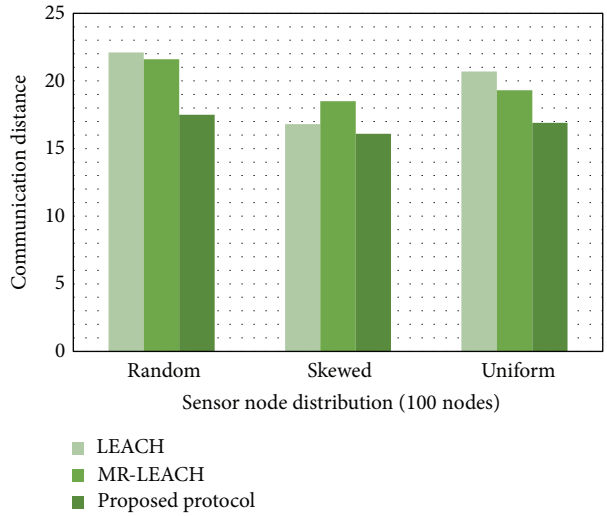


Figure 9: Communication distances with varying network distribution.

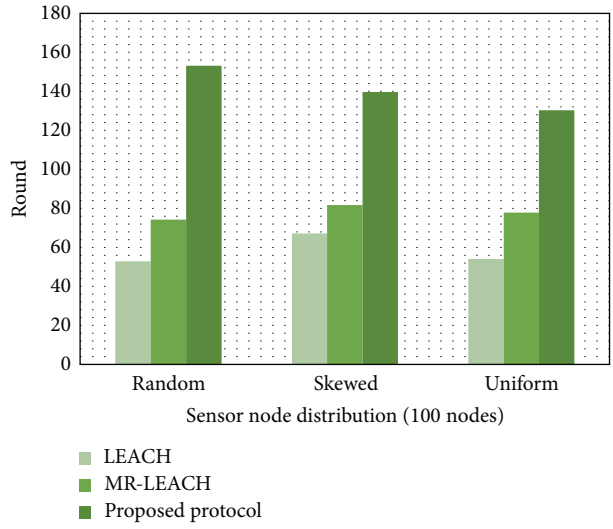


Figure 10: Sensor node lifetime.

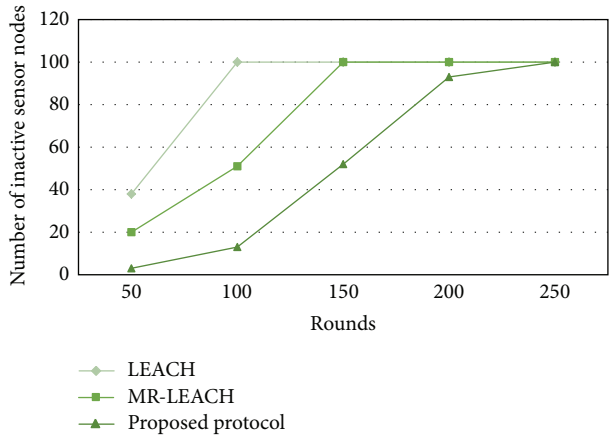


Figure 11: Network lifetime.

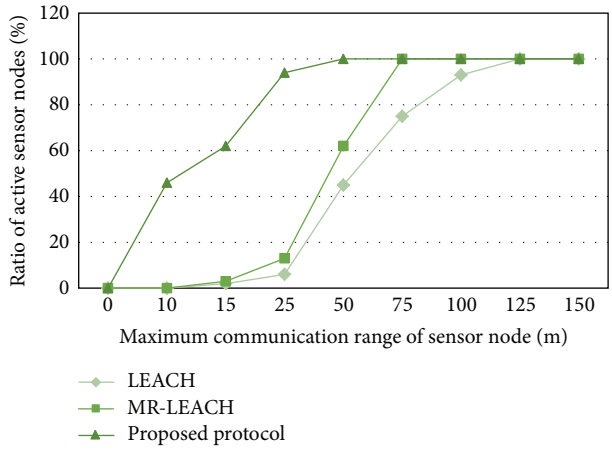


Figure 12: Sensor node response rate with varying communication range.

response ratio of the sensor nodes is 60%, whereas LEACH and MR-LEACH show only 2% and 3%, respectively, in terms of the response rate. Because our scheme provides both inter- and intracluster multihop routing, the sensor nodes with the limited communication range in our scheme can send messages to the sink node, via cluster members and a cluster header. Therefore, our routing protocol can support a whole network communication when the communication range is greater than a quarter of the whole network communication range.

5. Conclusion

In this paper, we proposed an energy-efficient cluster-based routing protocol for WSNs. Our protocol is based on a centralized clustering approach by using a representative path. A representative path is generated to select cluster headers and to form clusters in a distributed manner. To provide reliable network connectivity, we measure the message success rate of a sensor node when generating a representative path. To increase the lifetime of network, we select cluster headers as nodes having high connectivity. Therefore, the burden of network configuration and routing from sensor nodes can be greatly reduced. From our performance analysis, we show that our routing protocol outperforms both LEACH and MR-LEACH, in terms of energy efficiency and network reliability. As a future work, we will study a private data aggregation scheme with the energy-efficient routing protocol.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant no. 2013010099). Also, this work was supported by the Brain Korea 21 PLUS Project, National Research Foundation of Korea.

References

[1] I. F. Akyildiz and M. C. Vuran, "WSN applications," *Wireless Sensor Networks*, pp. 17–35, 2010.

[2] G. Anastasi, M. Conti, M. Di Francesco, and A. Passarella, "Energy conservation in wireless sensor networks: a survey," *Ad Hoc Networks*, vol. 7, no. 3, pp. 537–568, 2009.

[3] Y. Ha, H. Kim, and Y. Byun, "Energy-efficient fire monitoring over cluster-based wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 2012, Article ID 460754, 11 pages, 2012.

[4] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, pp. 660–670, 2002.

[5] M. O. Farooq, A. B. Dogar, and G. A. Shah, "MR-LEACH: multihop routing with low energy adaptive clustering hierarchy," in *Proceedings of the 4th International Conference on Sensor Technologies and Applications (SENSORCOMM '10)*, pp. 262–268, Venice, Italy, July 2010.

[6] Z. Yong and Q. Pei, "A energy efficient clustering routing algorithm based on distance and residual energy for wireless sensor networks," *Procedia Engineering*, vol. 29, pp. 1882–1888, 2012.

[7] N. Javaid, T. N. Qureshi, A. H. Khan, A. Iqbal, E. Akhtar, and M. Ishfaq, "EDDEEC: enhanced developed distributed energy-efficient clustering for heterogeneous wireless sensor networks," *Procedia Computer Science*, vol. 19, pp. 914–919, 2013.

[8] A. Jorio, B. Elbhiri, and D. Aboutajdine, "A new clustering algorithm in WSN based on spectral clustering and residual energy," in *Proceedings of the 7th International Conference on Sensor Technologies and Applications*, pp. 119–125, 2013.

[9] S. A. Nikolidakis, D. Kandris, D. D. Vergados, and C. Douligeris, "Energy efficient routing in wireless sensor networks through balanced clustering," *Algorithms*, vol. 6, no. 1, pp. 29–42, 2013.

[10] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, "Directed diffusion for wireless sensor networking," *IEEE/ACM Transactions on Networking*, vol. 11, no. 1, pp. 2–16, 2003.

[11] D. Braginsky and D. Estrin, "Rumor routing algorithm for sensor networks," in *Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications*, pp. 22–31, Atlanta, Ga, USA, September 2002.

[12] A. Woo and D. E. Culler, "A transmission control scheme for media access in sensor networks," in *Proceedings of the 7th Annual International Conference on Mobile Computing and Networking*, pp. 221–235, ACM, July 2001.

[13] T. Saravanan, G. Saritha, and V. Srinivsan, "A analysis of flat routing protocols in sensor N/W," *Middle-East Journal of Scientific Research*, vol. 20, no. 12, pp. 2566–2570, 2014.

[14] Y. Xu, J. Heidemann, and D. Estrin, "Geography-informed energy conservation for ad hoc routing," in *Proceedings of the 7th Annual International Conference on Mobile Computing and Networking*, pp. 70–84, July 2001.

[15] L. Li and J. Y. Halpern, "Minimum-energy mobile wireless networks revisited," in *Proceedings of the IEEE International Conference on Communications (ICC '01)*, vol. 1, pp. 278–283, June 2000.

[16] L. Lakshmanan and D. C. Tomar, "Location dependent RB multicast routing in Wireless sensor networks using GPS based system," *Indian Streams Research Journal*, vol. 4, no. 1, pp. 1–6, 2014.

[17] S. M. Tornell, E. Hernandez-Orallo, C. T. Calafate, J. C. Cano, and P. Manzoni, "An analytical evaluation of a Map-based Sensor-data Delivery Protocol for VANETs," in *Proceedings of the IEEE 14th International Symposium and Workshops on World of Wireless Mobile and Multimedia Networks (WoWMoM '13)*, pp. 1–6, IEEE, 2013.

[18] Y. Panthachai and P. Keeratiwintakorn, "An energy model for transmission in Telos-based wireless sensor networks," in *Proceedings of the International Joint Conference on Computer Science & Software Engineering (JCSSE '07)*, 2007.

[19] P. Levis and N. Lee, *Tossim: A Simulator for Tinyos Networks*, vol. 24, University of California, Berkeley, 2003.

[20] K. J. Choi, M. J. Yoon, I. B. Sim, and J. Y. Lee, "ECS: energy efficient cluster-head selection algorithm in wireless sensor network," *Journal of Communications and Networks*, vol. 32, no. 6, pp. 1342–349, 2007.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 125269, 6 pages
http://dx.doi.org/10.1155/2014/125269

Research Article

Sensor Relocation Technique Based Lightweight Integrated Protocol for WSN

J. Joy Winston[1] and B. Balan Paramasivan[2]

[1] *PET Engineering College, Vallioor, Tamil Nadu 627 117, India*
[2] *National Engineering College, Kovilpatti, Tamil Nadu 628 503, India*

Correspondence should be addressed to J. Joy Winston; joywinston47@gmail.com

Received 24 November 2013; Accepted 15 May 2014; Published 30 June 2014

Academic Editor: Ken Choi

While deploying sensor nodes for sensitive services, we need to focus on both the coverage and connectivity. The sensor relocation scheme had been proposed earlier, where the sensors themselves move towards the required position to make proper coverage and connection. From our experimental results, we noted that this work achieves good coverage. However, this approach fails to focus on connectivity issue due to node migration and secured communication too, which may compromise our collecting data. To address these serious issues, we proposed an efficient sensor relocation technique based lightweight integrated protocol (LIP). We have implemented and studied thoroughly our proposed work. From our experimental results, our proposed work LIP outperforms the existing sensor relocation scheme in terms of network connectivity, coverage, throughput, packet delivery fairness, energy consumption, energy loss, network resiliency, and secured communication. It is also noted that our proposed model maximizes the lifetime of wireless sensor networks.

1. Introduction

Recent advances in wireless network technologies have made the development of small, inexpensive, low power distributed devices, which are capable of local processing and wireless communication, and those devices are called sensor nodes. Sensors are generally equipped with limited data processing and communication capabilities and are usually deployed in an ad hoc manner in an area of interest to monitor events and gather data about the environment. Among various challenges faced while designing wireless sensor networks, maintaining network connectivity and coverage and maximizing the network lifetime stand out as critical consideration. The connectivity and coverage issues are generally met by deploying a sufficient number of sensor nodes or using specialized nodes with long-range capabilities to maintain a connected graph [1–3]. The network lifetime can be increased through energy conservation methods by using energy efficient protocols and algorithms. Due to various factors, such as the inaccessibility of the terrain and scale of the network, optimal deterministic deployment of the sensor network is often infeasible. A common scenario envisioned for deployment is that of randomly scattering sensor devices over the field of interest [4–6]. Thus, it makes the task of guaranteeing coverage much harder. As an alternative, mobile sensor nodes can be used to heal coverage holes in the network so that the randomness in sensor deployment can be compensated. Mobile platforms are already available in many deployment scenarios, such as soldiers in battlefield surveillance application, animals in habitat monitoring applications, and buses in traffic monitoring applications. In other scenarios, mobile devices can be incorporated into the design of the WSN architecture [1]. Failures in sensor networks [2, 7–9] are common and can be cured by using the redundant nodes in the network, that is, moving mobile redundant nodes to overcome the failure of sensor nodes or activate any sleeping redundant node in the group. Sensor nodes failure may cause connectivity loss and, in some cases, network partitioning.

However, such situation can be corrected by injecting a few mobile nodes in the network which can move to desired locations and repair broken network or using redundant

nodes in the network to heal the network failures. Utilization of redundant mobile nodes plays an important role in prolonging network lifetime. However, reallocating mobile sensor nodes has many challenges and special requirements. First, movement in sensor networks involved communication and can be very expensive in terms of energy. Mobility in WSN would also require network reconfiguration. When a node moves in the network, its relation to the environment and neighboring nodes will change and, thus, cause the network to reconfigure. As a result, mobility will add additional overhead to the network in terms of communication messages and reconfiguration.

Therefore, an energy efficient strategy is required to adopt mobile nodes in the network. Second, the reallocation of redundant mobile sensor nodes should have minimum effect on network sensing topology. Third, reallocation should be localized to achieve quick response time. For example, failure of sensor nodes monitoring a patient should be replaced immediately. To address the above discussed challenges, Asim et al. proposed distributed cellular architecture that partitioned the whole network into a virtual grid of cells. The initial design of the cellular architecture [10] was proposed by Asim et al., where a cell manager is chosen in each cell to perform management tasks. These cells combine to form various groups and each group chooses one of their cell managers to be a group manager. This model was used in [11]. However, from our research work, we have realized that the modified cellular architecture [11] consumes more energy, which minimizes the lifetime of sensor networks. And also it is identified that this approach fails to focus on secured communication, which may compromise our collecting data. To address these serious issues, this research work has proposed an efficient sensor relocation technique based on lightweight integrated protocol (LIP).

2. Related Work

Mobility and its effects on the sensor network operation have been extensively studied and emerged as an important requirement for wireless sensor networks. Wang et al. [12] presented a proxy-based sensor relocation algorithm for the sensor networks composed of both static nodes and mobiles. Mobile from nearby locations moves to fill the coverage hole. This results in the emergence of new holes. Thus, more and more sensors are involved in relocation. This approach relies on flooding for replacement and uses a direct relocation method that can produce inconsistent relocation delay. Wang et al. [13] presented a grid-quorum-based relocation protocol for mobile sensor networks. In this protocol, the network field is geographically partitioned into grids. In each grid, a node as grid head runs the quorum-based location service to fund the redundant sensor nodes in the network. Then the discovered replacement is relocated along a carefully selected path in a cascaded way, that is, in the shifted way. Asim et al. [11] have proposed sensor relocation scheme which consists of two main phases, namely, (i) identifying redundant nodes and (ii) sensor relocation. In this model, the cell manager is responsible for collecting information of its cell members and determines the existence of redundant sensors based on their location. For redundant sensors located on the boundary of the cells, the cell managers coordinate to make decisions. The cell manager can also monitor its cell members and initiate a relocation process in case of new event or sensor failure.

Redundant nodes may be sent to a sleep mode to save or conserve energy. In other words, in some cell areas, there may be more sensor nodes than others and, hence, they need to maintain nodes intensity. That is, some nodes can be sent to a sleep mode to adjust the cell size. Cell size is affected by factors such as the transmission range of the transmitter or the transmission power and the sensing range of the sensor nodes. Varying the cell size in the network affects the lifetime of the network.

In this cellular architecture [6, 9, 11], the cell size is a user defined parameter, which can be adjusted to meet the required cell-head density. Also, to keep the hierarchical structure efficient, load for each cluster head should be equivalent. Thus, the cluster size is a key parameter to achieve balanced load among clusters.

Cell-head density will be defined according to application requirements. Appropriate cell-head density plays an important role in maximizing the performance of the network. However, for most sensor networks applications, it is important to support fast delivery of important and urgent data. Also, maximizing cell-head density may put extra burden on cell manager for certain operations, that is, data aggregation. Therefore, it is extremely important for the performance of sensor network to carefully define cell size and cell-head density.

The average number of static sensors needed to cover a cell is represented by p and is maintained by the cell manager [11]. However, some cells may contain fewer sensors than p due to the randomness in deployment or node failures. If cell I contains static nodes (Ni) $< p$, mobiles nodes need to move into the cell to fill in the vacancies. The cell managers within the same group represent a virtual grid structure towards their group manager. Instead of flooding subscribe/publish messages across the network and polling information from hundreds of thousands of nodes, the cell manager contacts its group manager in the virtual grid structure to track the redundant mobile nodes. This design minimizes the number of communication messages and thus conserves node energy. Our proposed framework is based on finding redundant sensor nodes in a localized fashion. We believe that adopting localization to a certain degree reduces network traffic whenever possible. Additionally, such an approach also has a quick response to events that occurred in the network. Each group manager [11] maintains information about the publisher cells within its group and shares this information with the closest neighboring group managers only. This supports the short distance movement of mobile sensor nodes. If the mobile sensor node travels a long distance to replace a faulty node or fill the coverage, it may run out of power and create a new coverage hole. When a cell has redundant sensor nodes, the cell manager propagates this information to its group manager. When a cell wants more sensors, the cell manager only needs to contact its group

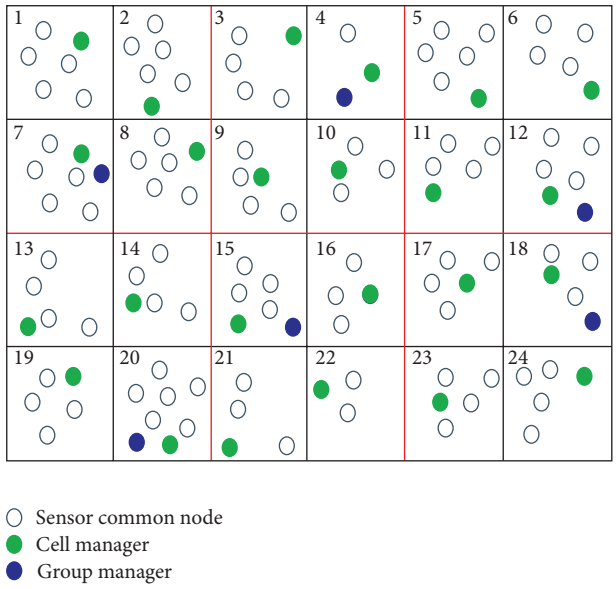


Figure 1: Finding redundant nodes.

manager. Group manager will first look for redundant sensor nodes within a group and if there are no redundant nodes within its group, it then searches which nearest group has redundant sensor nodes.

For example, as shown in Figure 1, suppose Cell 3 and Cell 9 have redundant sensors, while Cell 4 needs information for its group manager. The cell manager of Cell 4 puts forward its demand for more sensors to its group manager. The group manager finds the distance between the subscriber cell and all possible publishing cells. The publishing cell with the shortest distance to the subscribing cell will get the priority. The group manager will notify the selected publisher cell to move its redundant sensor nodes to the subscriber cell. The idea and the finding of the redundant nodes [11] are shown in Figure 1.

From our literature survey, we concluded that the sensor relocation scheme proposed by Asim et al. [11] is the best model, which defined the problem of sensor relocation that can be used to deal with sensor coverage holes or sensor failures [14]. This cellular hierarchical architecture located redundant mobile sensor nodes with minimum message complexity. Information about the redundant sensor nodes is only available at some intermediate nodes. This helps to reduce message complexity through message filtration and avoid message flooding, and it reduces the energy consumption also.

3. Identified Problem

Though this sensor relocation scheme reduces the message complexity through message filtration and avoiding message flooding and it reduces the energy consumption, this model could not address the important sensors networks' challenges such as connectivity and network security. That is, this model could not achieve fair connectivity as the sensor nodes are involving migration from one cell to another cell for achieving coverage. The mobility pattern is also to be considered to provide assured connectivity. This is one of the major issues to be addressed. And for the second one, as the nodes are moving from one cell to another cell, this model might be compromised and vulnerable to system security. To address these issues, this research work proposed an efficient lightweight integrated protocol (LIP) which is integrated with the existing sensor relocation scheme, that is, focusing both connectivity and security as well. This proposed sensor relocation technique based lightweight integrated protocol (LIP) addresses secured coverage, connectivity, communication, and path security.

4. Proposed Sensor Relocation Technique Based Lightweight Integrated Protocol (LIP)

As discussed in the previous section, we have understood that the sensor relocation scheme could not support ensuring connectivity and security.

Thus, this research work has proposed an efficient lightweight integrated protocol (LIP) which is integrated with the sensor relocation technique. This section briefly describes the principle and the procedure of our proposed technique.

4.1. Principle of Sensor Relocation Technique Based Lightweight Integrated Protocol (LIP). A wireless sensor may consist of hundreds to thousands of sensor nodes and is usually deployed randomly, and, hence, this may result in some areas having more sensor nodes than others. Hence, the proposed sensor relocation scheme identifies the redundant nodes.

After locating the redundant sensor nodes, the sensor relocation scheme moves the sensor to the new destination, where the density of sensor nodes is less. The nodes are moving to destination as follows.

(i) Direct movement, where nodes are moving between two direct neighboring cells, which heals the coverage. But connectivity is the issue.

(ii) Cascaded movement, where nodes are moving from one cell to another remote cell through neighbor cell.

While moving nodes, the above described procedure does not focus on location awareness and connectivity awareness as well. Thus, the proposed lightweight integrated protocol is working along with the sensor relocation scheme as follows.

The first phase of the proposed work is registering the IDs of all created nodes with a WSN-security server. During the migration of nodes between cells, group manager and cell manager will permit the nodes to enter into any cell if that node's ID is registered with WSN-security server.

It divides the sensor networks into virtual rings of optimal width $\mu = R_C/2.45$ [11]. Here, R_C is the communication range of the node. This ensures that while moving nodes from one cell to another cell, the connectivity is ensured.

(i) Then to provide guaranteed connectivity coverage, the distance between nodes is maintained

as $\min\{\sqrt{3R_S}, R_C\}$ where R_C and R_S are the communication and sensing ranges of nodes, respectively.

(ii) It also divides the communication range into two different threshold levels, namely, Threshold Th1 and Threshold Th2. The Th1 is at 40% and Th2 is at 80% of R_C. Accordingly, nodes that receive a signal stronger than Th1 (i.e., they are within 40% of R_C) make the first ring, while nodes that receive a signal weaker than Th1 but stronger than Th2 (i.e., they are within 40% to 80% of R_C) make the second ring. A third ring is possibly defined for nodes that receive a signal weaker than Th2 (i.e., they are beyond 80% of R_C), which may be updated on receiving a stronger signal afterwards.

(iii) The sensor relocation scheme through LIP will move the redundant sensors only when the previous step conditions are satisfied, that is, within Virtual Ring 1 or Virtual Ring 2. Otherwise, this scheme does not move the sensor nodes.

Instead of moving sensor nodes, this proposed scheme will change the state of nodes as follows. Sensing only: nodes in this state can sense their environment but cannot transmit or receive data as their transceivers are switched off. A very low energy is used by the nodes in this state.

Sleeping: nodes in this state can neither sense their environment nor transmit and receive data. Sleeping nodes consume extremely low amount of energy.

4.2. Procedure of Sensor Relocation Technique Based Lightweight Integrated Protocol (LIP). After the network boots up, all nodes in the network run the WSN-security server to register their IDs for getting authentication while migrating to another cell in future for communication.

While migrating, group manager/cell manager will get authentication from WSN-security server for permitting node to enter into any cell.

Then, these nodes run the sensor relocation scheme as follows.

Call the Publication Phase

Collect the availability of redundant sensor nodes through publication phase called subscription phase.

Find the sensing hole [15].

Request the redundant nodes.

Move towards hole with the help of group and cell managers if condition cdn satisfied Cdn.

Divide the sensor nodes into virtual rings with width = $\mu = R_C/2.45$. Make move if distance between nodes is $\min\{\sqrt{3R_S}, R_C\}$.

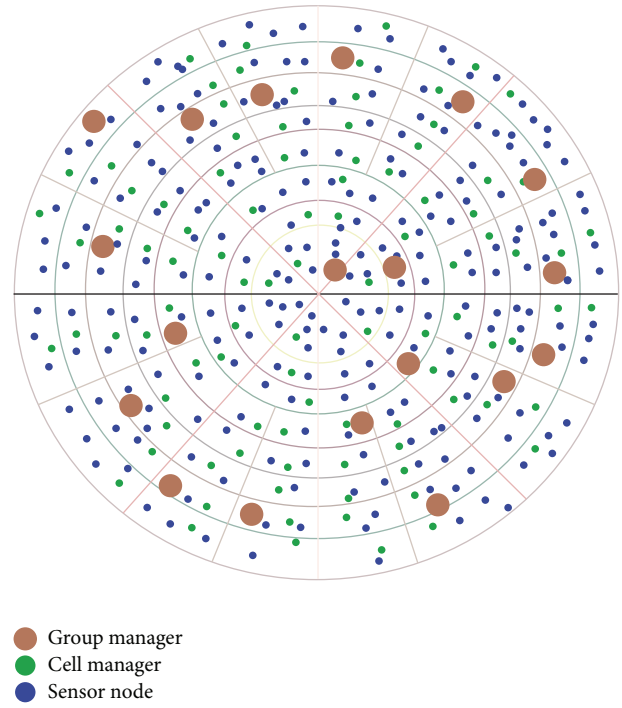


Figure 2: Simulation setup for framework.

Set Node to Fully Active

If communication range level is less than $R_C{}^*.4$, call direct move.

If communication range level is greater than $R_C{}^*.4$ and less than $R_C{}^*.8$, call cascaded move.

Otherwise restrict the move.

Change the node state to sensing or sleeping state.

If redundant nodes broadcast message within the group and its communication range is $\{\sqrt{3R_S}\}$, change state into sensing.

If sensing nodes receive the same broadcast message (sense) more than one time, then change state into sleeping for period T.

Move towards hole with the help of group and cell managers if condition cdn is satisfied.

The sensor relocation technique based lightweight integrated protocol (LIP) is executing nodes as shown in the above procedure.

5. Experimental Setup and Performance Evaluation

As shown in Figure 2, the sensor network is divided into rings and groups/clusters as proposed by the lightweight integrated protocol (LIP). Each cell is managed by cell manager and each group/cluster is managed by group manager. This model is implemented with QualNet 5.0 Simulator and is studied

Table 1

S. number	Simulation	Description
1	Number of nodes	500
2	Topology	Random
3	Deployment area	100 m × 100 m to 400 m × 400 m
4	Mobility	Random
5	Channel	Wireless
6	MAC	IEEE802.15.4/ZigBee
7	Transmission/sensing	5 m–50 m
8	Sensor's initial	2000 mJ
9	Involved protocols	Cellular based protocol (grid-quorum-based and layered diffusion based protocol)

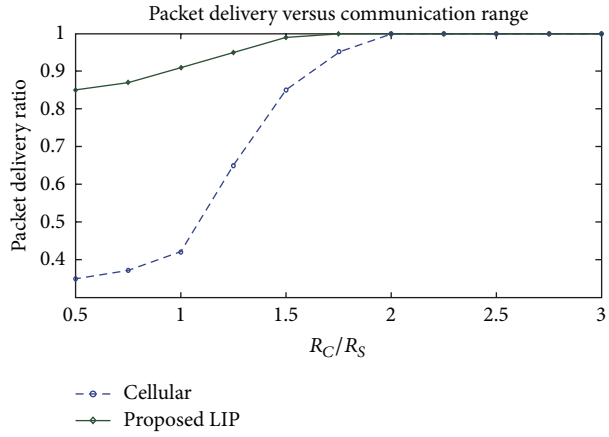


Figure 3: Throughput (packet delivery).

thoroughly. We have considered simulation parameters in Table 1 for our simulation.

From Figure 3, it is observed that the throughput of the proposed work is better and almost fair compared with the existing sensor relocation technique based cellular model. The proposed technique achieves higher fair throughput because this system does not face any connectivity issues while moving from one cell to another. That is, before moving the sensor, our proposed model examines the transmission, coverage, and sensing range between two nodes based on their distances. The proposed work retains the coverage range of the existing network, which is shown in Figure 4.

Our proposed model saves considerable energy by restricting sensor movements between cells. The proposed model permits the redundant node to migrate from one cell to another cell to ensure coverage provided that the system ensures connectivity and hence the spending energy for migration will not be wasted. The energy consumption too is limited as the migrations of nodes are restricted as shown in Figure 6. Thus, the proposed model has maximized the lifetime of sensors and sensor networks, which is shown in Figure 5.

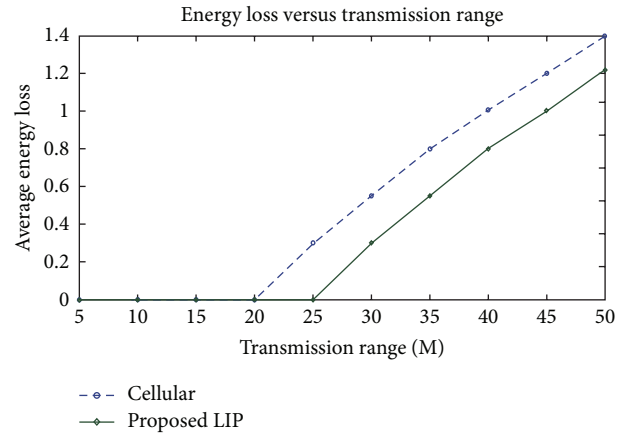


Figure 4: Coverage range.

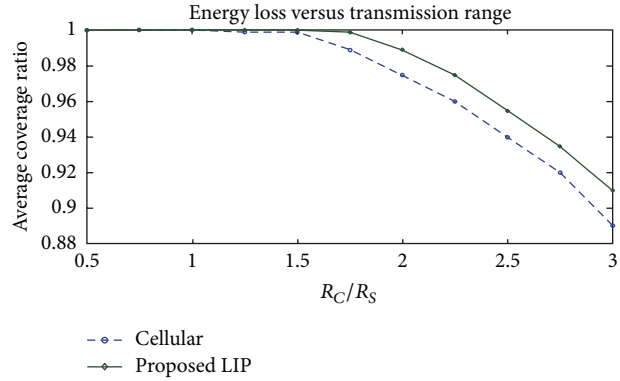


Figure 5: Energy loss.

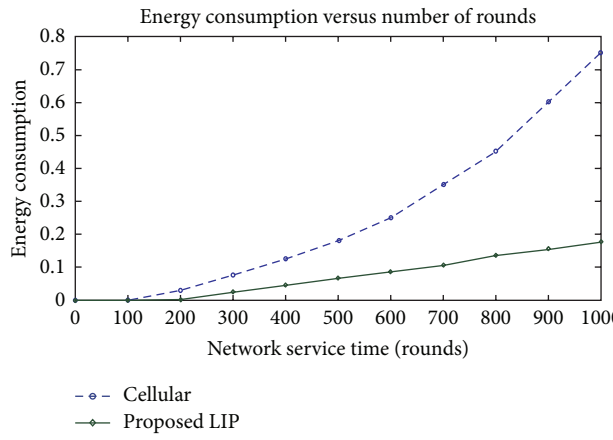


Figure 6: Energy consumption.

6. Conclusion

This research work has studied sensor relocation scheme thoroughly and is implemented through QualNet 5.0 Simulator. The experimental results established that this sensor relocation scheme performs better in terms of coverage. It is found that this scheme is consuming more energy for node

migration, it faces connectivity issue too, and it leads to minimizing of the lifetime of sensor networks. With respect to this issue, our research work has proposed and implemented an efficient sensor relocation technique based lightweight integrated protocol (LIP). From our study, it is observed that our proposed work LIP performs better than the existing sensor relocation scheme in terms of network connectivity, coverage, throughput, packet delivery fairness, energy consumption, energy loss, network resiliency, and secured communication. It is also noted that our proposed model maximizes the lifetime of wireless sensor networks.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] E. Ekici, Y. Gu, and D. Bozdag, "Mobility-based communication in wireless sensor networks," *IEEE Communications Magazine*, vol. 44, no. 7, pp. 56–62, 2006.

[2] Y. Zhou and M. Medidi, "Sleep-based topology control for wakeup scheduling in wireless sensor networks ," in *Proceedings of the 4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON '07)*, pp. 304–313, San Diego, Calif, USA, June 2007.

[3] P. M. Wightman and M. A. Labrador, "A family of simple distributed minimum connected dominating set-based topology construction algorithms," *Journal of Network and Computer Applications*, vol. 34, no. 6, pp. 1997–2010, 2011.

[4] W. Wang, V. Srinivasan, and K. C. Chua, "Trade-offs between mobility and density for coverage in wireless sensor networks," in *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking (MobiCom '07)*, pp. 39–50, Montreal, Canada, September 2007.

[5] S. Jauregui-Ortiz, M. Siller, and F. Ramos, "Node localization in WSN using trigonometric figures," in *Proceedings of the IEEE Topical Conference on Wireless Sensors and Sensor Networks (WiSNet '11)*, pp. 65–68, January 2011.

[6] M. Akhlaq and T. R. Sheltami, "The recursive time synchronization protocol for wireless sensor networks," in *Proceedings of the IEEE Sensors Applications Symposium (SAS '12)*, pp. 62–67, February 2012.

[7] M. Asim, H. Mokhtar, and M. Merabti, "A cellular approach to fault detection and recovery in wireless sensor networks," in *Proceedings of the 3rd International Conference on Sensor Technologies and Applications (SENSORCOMM '09)*, pp. 352–357, Athens, Greece, June 2009.

[8] M. Z. Khan, M. Madjid, B. Askwith, and F. Bouhafs, "A fault-tolerant network management architecture for wireless sensor networks," in *Proceedings of the 11th Annual PGNet Conference*, Liverpool, UK, 2010.

[9] K. Srinivasan and P. Levis, "RSSI is under appreciated," in *Proceedings of the 3rd Workshop on Embedded Networked Sensors (EmNets '06)*, Cambridge, Mass, USA, 2006.

[10] M. Asim, H. Mokhtar, and M. Merabti, "A cellular self-organization architecture for wireless sensor networks," in *Proceedings of the (PG NET '08)*, August 2008.

[11] M. Asim, H. Mokhtar, M. Z. Khan, and M. Merabti, "A sensor relocation scheme for wireless sensor networks," in *Proceedings of the 25th IEEE International Conference on Advanced Information Networking and Applications Workshops (WAINA '11)*, pp. 808–813, Singapore, March 2011.

[12] G. Wang, G. Cao, and T. L. Porta, "Proxy-based sensor deployment for mobile sensor networks," in *Proceedings of the IEEE International Conference on Mobile Ad-Hoc and Sensor Systems*, pp. 493–502, October 2004.

[13] G. Wang, G. Cao, T. La Porta, and W. Zhang, "Sensor relocation in mobile sensor networks," in *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '05)*, vol. 4, pp. 2302–2312, Miami, Fla, USA, March 2005.

[14] M. Akhlaq, T. R. Sheltami, and E. M. Shakshuki, "An integrated protocol for coverage, connectivity and communication (c3) in wireless sensor networks," in *Proceedings of the 27th IEEE International Conference on Advanced Information Networking and Applications (AINA '13)*, pp. 606–612, March 2013.

[15] T.-W. Sung and C.-S. Yang, "A cell-based sensor deployment strategy with improved coverage for mobility-assisted hybrid wireless sensor networks," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 5, no. 3, pp. 189–198, 2010.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 151493, 21 pages
http://dx.doi.org/10.1155/2014/151493

Research Article

Flexible Capturing Application for Enhanced Generation of EPCIS Events

Fengjuan Jia,[1] Seungwoo Jeon,[1] Bonghee Hong,[1] Joonho Kwon,[2] and Yoon-sik Kwak[3]

[1] *Department of Computer Engineering, Pusan National University, Busan 609-735, Republic of Korea*
[2] *Institute of Logistics Information Technology, Pusan National University, Busan 609-735, Republic of Korea*
[3] *Department of Computer Engineering, Korea National University of Transportation, Chungbuk 380-702, Republic of Korea*

Correspondence should be addressed to Joonho Kwon; jhkwon@pusan.ac.kr

Received 30 November 2013; Accepted 19 March 2014; Published 25 June 2014

Academic Editor: Ken Choi

Radio frequency identification (RFID) technology and electronic product code (EPC) technology have been widely used to identify and keep track of physical objects. EPCglobal proposed the EPC network which consists of several components such as application level events (ALE) and EPC information service (EPCIS) to deal with the captured data from different layers. Many studies mostly concentrated on dealing with the RFID tag data in ALE, as well as querying and sharing EPCIS events in EPCIS. However, there is no well-known study on specifying how to generate higher level EPCIS events. The event types and semantic event fields are both uncertain for the capturing application to generate EPCIS events. Therefore, this paper proposes the flexible capturing application (FCA) to solve the problem that the event types and semantic event fields are both uncertain. FCA specifies generation rules about the four EPCIS event types. All the generation rules are matched for the incoming tag data to determine the event types. Event fields are generated with tag data and other sources of data after deciding event types. Thus, FCA can generate EPCIS events arising from supply chain activity. We evaluate our approaches by means of simulation and real experiments. Our experimental results indicate that FCA can be effective in processing EPCIS events data. We conclude with suggestions for future work.

1. Introduction

Radio frequency identification (RFID) [1–3] is a technology that allows an electronic product code (EPC) [4, 5] tag attached to an item to carry an identity for that item. Due to the reductions in cost and size of device components, RFID technology has been widely used to identify and keep track of physical objects in many RFID applications [6–9]. Industrial enterprises and government organizations such as Wal-Mart [10], Tesco, and even the US Department of Defense gain practical benefits by using RFID technology.

EPCglobal [11] proposed EPC network [12–14] to provide real-time data on physical objects by using RFID technology. Figure 1 shows the simple architecture of EPC network, which includes several components to deal with the captured data from different layers. A reader protocol [12, 15] specifies how to collect raw tag data from readers and delivers them as physical events. The application level events (ALE) [12, 16] filters and collects raw tag data read and delivers them as logical events. The capturing application [12, 17] generates high level EPC-related business events and delivers them as EPCIS events. EPCIS events differ from low level RFID tag data by providing dynamic semantic tracking information about items as they move through the supply chains. Thus, business applications know how and why physical events occurred and what state the physical objects are in by using EPCIS events. The EPC Information Services (EPCIS) [12, 17] stores and retrieves EPCIS events generated by capturing cpplication. Thus, EPCIS events can be shared both within and across enterprises.

Example 1. Consider Figure 1 which shows a generation and flow of RFID events through the EPCglobal framework. After producing a RFID tagged item, manufacturer A sent the item to wholesaler B. Suppose that the item is being tagged with RFID tag with EPC number "epc:1200.123.1123." A RFID reader collects the tag information when it passes through according to steps of reader protocol. Then ALE will

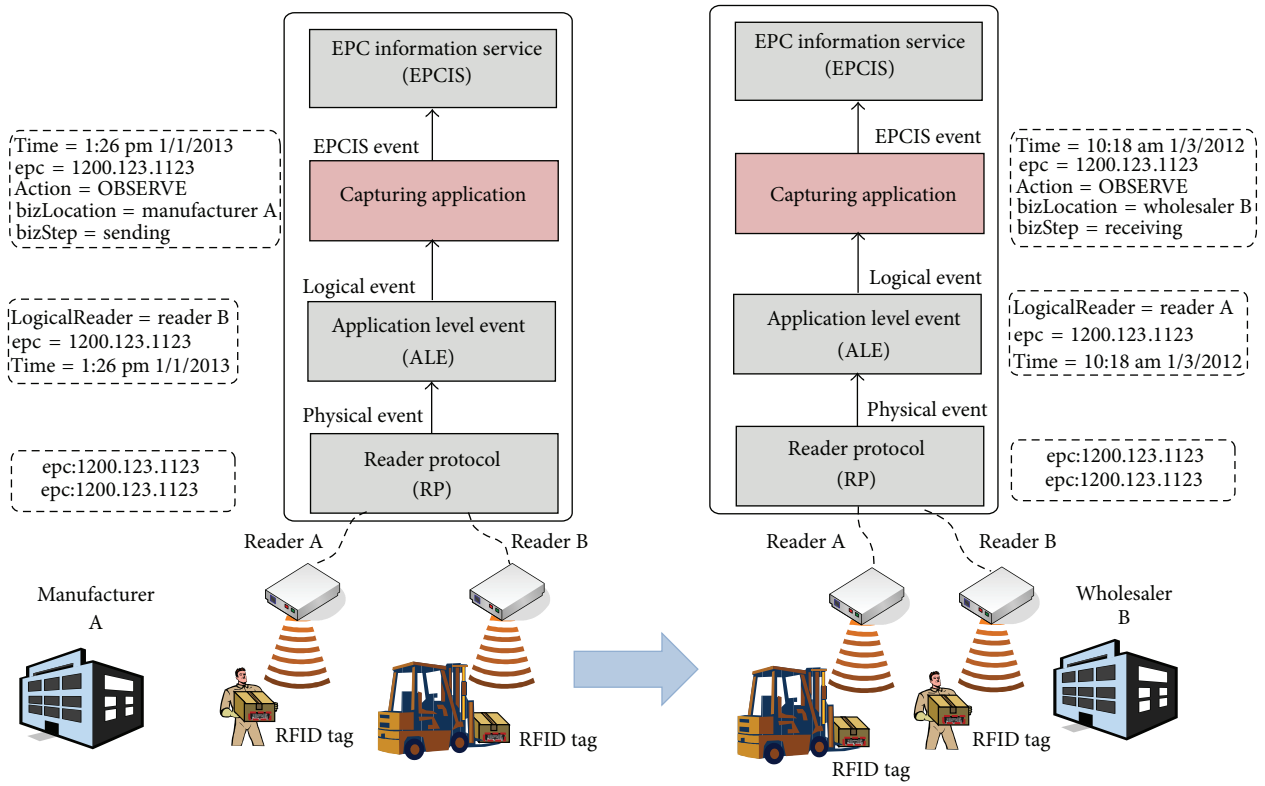


Figure 1: The simple architecture of EPC network [12].

transform a physical event, which is "epc:1200.123.1123," into a logical event by adding a logical reader name and timestamp. ALE also filters and smoothes duplicated physical events. Next, a capturing application generates an EPCIS event. The business logic information is added to the EPCIS event. The values of action, bizLocation, and bizStep are "OBSERVE," "manufacturer A," and "sending," respectively. Lastly this EPCIS data is stored into EPCIS and shared to the partners of manufacturer A.

Suppose wholesaler B is expecting the delivery from manufacturer A. When the item arrives, it will pass through the RFID reader in the vicinity of wholesaler B. A physical event containing "epc:1200.123.1123" is generated and then it is transformed into a logical event with a logical reader name and timestamp. An EPCIS event is generated as the same way with the business information. At this time, the values of action, bizLocatio, and bizStep are "OBSERVE," "wholesaler B," and "receiving," respectively.

Since the goal of RFID system is to collect and share tracking information about physical objects, the capturing application should generate correct EPCIS events for all the incoming RFID tag data. However, the problem of capturing application is that the event types and semantic event fields are both uncertain when a capturing application tries to generate EPCIS events after receiving RFID tag data and time information from ALE. The capturing application does not know how to get the semantic fields which cannot be obtained from the incoming RFID tag data either. Therefore, a proper capturing application is of great importance to generate correct event types with desired event fields.

Many studies mostly concentrated on dealing with the low level RFID tag data in ALE [18, 19], as well as querying and sharing EPCIS events in EPCIS [20, 21]. There are few studies on capturing application to generate EPCIS events. Several papers [22–25] mapped ALE ECReports to an EPCIS event type for each logical reader and configured the event fields. However, these existing methods cannot generate EPCIS events flexibly. Different EPCIS event types for different incoming ECReports cannot be generated at a specific reader. APDL (AspireRFID process description language) [26] generated EPCIS events based on the newly defined XML schema. However, the condition and timing for generating each event type are unknown.

In this paper, we propose the flexible capturing application (FCA) system to generate EPCIS events flexibly, which solves the problem that the event types and semantic event fields are both uncertain. The key contributions of our work are summarized as follows.

(i) *Modeling of EPCIS Events*. We formally define the modeling of transaction data, business context data, and situation context data to represent the RFID semantic data required to generate EPCIS events. Since our model supports user defined EPCIS event, we can handle product return cases of the supply chain.

(ii) *Flexible Capturing Application*. The event generation rules are proposed based on ECReports, RFID semantic data, and other related data to generate EPCIS events. In our proposed FCA, the event types for the incoming ECReports are determined automatically and flexibly by matching the specified event generation rules on four EPCIS event types. We have provided the detailed algorithms for generating different EPCIS event types.

(iii) *Proofs for Correctness*. We also proved the completeness of our FCA approach. The event fields are automatically generated with ECReports and RFID semantic data if an EPCIS event generation rule is satisfied. One or more EPCIS event types are generated for one ECReport. All generated EPCIS events are delivered to EPCIS repository after all the four EPCIS event generation rules are matched.

(iv) *Experimental Evaluation*. We have presented a detailed experimental evaluation to verify the proposed FCA. For this purpose, we defined the flexibility for evaluation. The experiments results show that FCA can generate correct EPCIS events more flexibly than the previous approach.

The remainder of the paper is organized as follows. Section 2 provides the background and motivation. Section 3 proposes modeling of RFID semantic data. Section 4 proposes the flexible capturing application. Section 5 presents the experiments. Section 6 introduces the related work. Section 7 concludes the paper.

2. Background and Motivation

In this section, we first explain the background knowledge of a RFID system and then describe the problem definition and motivations of our work.

2.1. RFID System Architecture. A general RFID system architecture which consists of (RFID) tags, (RFID) readers, application level event (ALE), capturing application, and EPC information service (EPCIS) is depicted in Figure 2. A tag is an identification device attached to the physical object we want to track. A reader is a device that can recognize the presence of RFID tags and read the information stored on them. The reader can inform another system such as reader protocol (RP) about the presence of the tagged physical objects. After RP collects the raw tag data from readers, ALE filters these raw tag data. For receiving RFID tag data from ALE, capturing application sends ECSpec to ALE; then ALE reports the tag data with ECReports. ECSpec is used to specify which RFID tag data from ALE are to be sent to capturing application. ECReports are the output of ALE and the input of capturing application, which includes the RFID tag data described in ECSpec. Then capturing application generates EPCIS events with the ECReports from ALE and delivers the EPCIS events to EPCIS repository. EPCIS stores and retrieves EPCIS events. Business partners can share the EPCIS events by retrieving EPCIS events.

Table 1: Summarized fields of EPCIS events.

Category	Fields
What	EPC
	EPCClass + quantity (QuantityEvent)
	BusinessTransactionList (TransactionEvent)
When	EventTime
	RecordTime
Where	ReadPoint
	BusinessLocation
Why (business context)	BusinessStep
	Disposition
Other	Action

2.2. EPCIS Events. In this subsection, we will explain the details of EPCIS events since they have key roles in RFID system. EPCIS events differ from low level RFID tag data by providing semantic information. For this purpose, four event types are specified as a subclass of EPCIS event in the EPCIS standard as shown in Figure 3. A user can also define a new event type if needed.

The four EPCIS events describe how and why RFID tag data occurred and what state the physical objects are in. ObjectEvent describes events pertaining to one or more EPCs in a supply chain from birth (ADD) through middle life (OBSERVE) to death (DELETE). AggregationEvent explains events that child EPCs have been physically aggregated to a parent EPC, including EPCs from an aggregation (ADD); see EPCs in an aggregation (OBSERVE) or remove EPCs from an aggregation (DELETE). QuantityEvent describes events pertaining to a specified quantity of an object class. TransactionEvent describes the association (ADD) or disassociation (DELETE) of physical objects to one or more business transactions. The verbs in the parenthesis mean action fields of EPCIS events which are explained in the following paragraph.

Event fields of an EPCIS event carry descriptive information for the event in detail. Table 1 summarizes the event fields specified in the EPCIS standard which can be divided into four dimensions named what, when, where, and why, as well as the action field. EPC can be an epcList or parentID/childEPC. `EPCClass` and `quantity` are the class and quantity of object in QuantityEvent, respectively. `BusinessTransactionList` includes the business transactions with a type and a number. `EventTime` is the time when an event took place. `RecordTime` is the time when the event was received through the EPCIS capture interface. `ReadPoint` indicates the location where an event occurred. `BusinessLocation` describes the place where the object is immediately after the event occurs. `BusinessStep` is the business operation occurring at the time of the event. Disposition is the business state of the object immediately after the event occurs. The `action` field has three values: ADD, OBSERVE, and DELETE, describing how the event relates to the lifecycle of the EPCs in an ObjectEvent, how the event relates to the aggregation in an AggregationEvent,

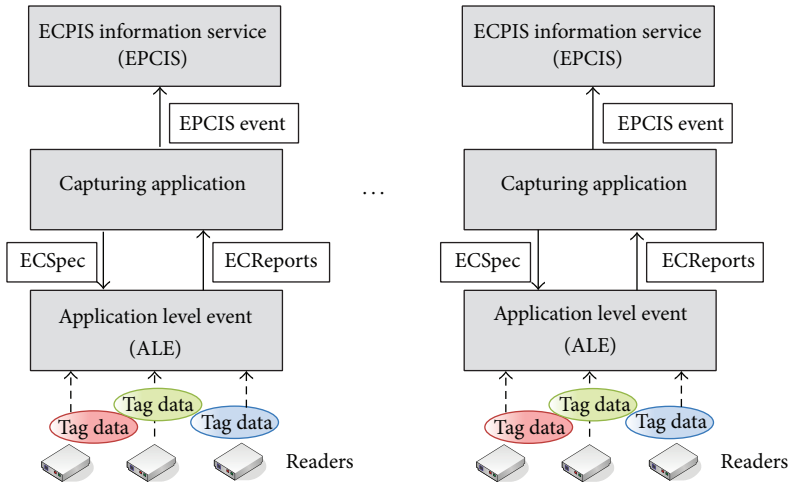


Figure 2: RFID system architectures.

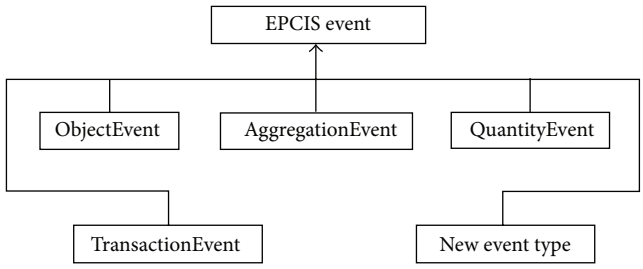


Figure 3: Event types of EPCIS events.

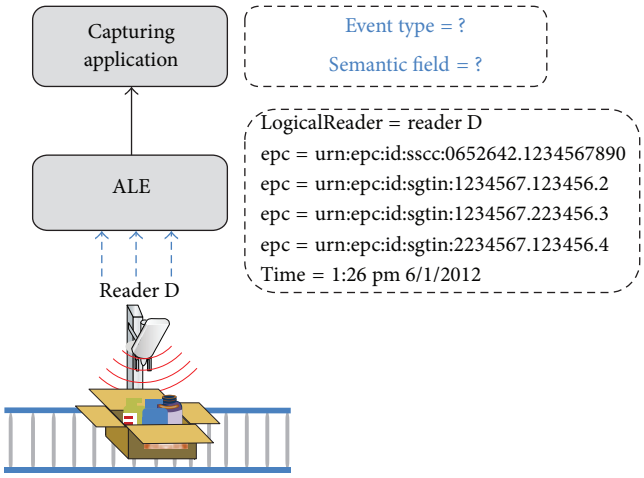


Figure 4: A single reader case.

or how the event relates to the business transaction in a TransactionEvent.

2.3. Motivation. In this subsection, we explain motivation and the problem definition of our work.

For digging out the problem, let us consider two different cases: simple single reader case and general multiple readers case.

Example 2 (a single reader case). Figure 4 shows a simple single reader case of a RFID system. Three heterogeneous items with different object classes and a case are coming through reader D. In some cases, we just use the case to deliver the item. Thus, there is no aggregation between the items and the case. Therefore,`ObjectEvent` should be generated here. In some other cases, we want to pack the items as a set for further usage. Thus, the items are aggregated into the case. In this case,`AggregationEvent` should be generated here. A capturing application should generate correct EPCIS events with incoming ECReports from ALE. However, it is impossible for a capturing application to decide whether to generate ObjectEvent or heterogeneous AggregationEvent without further semantic and business information.

The problems of general multiple readers case are explained in the following example.

Example 3 (multiple readers case). Consider again Figure 1 which shows a manufacturer-wholesaler supply chain including multiple readers. Each of two trading partners has its own EPCIS repository. With all the incoming ECReports from ALE, all capturing applications should generate correct EPCIS events for readers in each trading partner. However, the capturing application cannot decide to generate which event types correctly for each reader since it cannot obtain exact business steps information such as "receiving" and "sending." The values of action fields are unknown to capturing application either.

Our Motivation. As explained in Examples 2 and 3, a capturing application cannot determine which types of EPCIS events will be generated. The reason is that the incoming ECReports from ALE exclude the semantic event fields such as action, bizLocation, bizStep, disposition, and bizTransactionList. These shortcomings have motivated us to propose a flexible capturing application system that generates EPCIS events effectively based on the semantic and business context data.

Thus, we can formally state a flexible capturing application problem as follows.

Given a set of ECReports R which include RFID tag data from ALE and a set of context data C, generate a set of EPCIS events e with desired event fields for each $r \in R$.

3. Modeling of RFID Semantic Data

In this section, we describe a RFID semantic data model employed in our flexible capturing application (FCA). The RFID semantic data model can be categorized into context data and transaction data.

3.1. Context Data. Context data is used to provide some semantic information. We define and explain two context data: situation context data and business context data.

3.1.1. Situation Context Data. We describe the necessity of situation context first by using an example. A RFID system handles various types of products. For example, consider two situations in Figure 5. In Figure 5(a), four items are being packaged into a case under reader D. For this, we can generate an "ADD" AggregationEvent for the packaged case and five "OBSERVE" ObjectEvents for items and case. However, we think that an "ADD" AggregationEvent is enough for describing the jobs that happened and five ObjectEvents are unnecessary. Figure 5(b) describes another situation. Assume that two cases come through reader D in turn. The case under the reader D contains the same types of items and the other case contains three heterogeneous items as a set. Here, we need heterogeneous AggregationEvent rules. In other words, a capturing application must understand the situation difference between these packaged cases and items.

Before introducing a definition of situation context data, we begin by summarizing some of key notational conventions used in our discussion in the remainder of the paper in Notation Section. Additional notation will be introduced when necessary.

Situation context data is used to define extra situation rules for whether generating some event types or not according to the users' requirements. It is represented as a list of event generation rules, where each rule can have time, epcList, readPoint, bizLocation, bizStep, and disposition fields. More formally, situation context data SCD is described as SCD=(oe, ae, qe, te, and ne). Definition 4 shows more details of the situation context data.

Definition 4 (situation context data). The situation context data is represented as a list of event generation rules (*oe*, *ae*, *qe*, *te*, and *ne*), where

(i) oe is an ObjectEvent generation rule in 2 tuples (isOE, fieldList);

(ii) ae is an AggregationEvent generation rule in 3 tuples (isAE, hePattern, and fieldList);

(a) hePattern=(parent, children) is to generate heterogeneous AggregationEvent;

(1) parent represents a parent ID;
(2) children are a set of child EPCs;

(iii) qe is a QuantityEvent generation rule in 3 tuples (bSList, actionList, and fieldList);

(iv) te is a TransactionEvent generation rule in 2 tuples (isTE, fieldList);

(v) ne is a new event type generation rule in 3 tuples (isExist, eventType, and fieldList);

(a) eventType is the type name of the new event type.

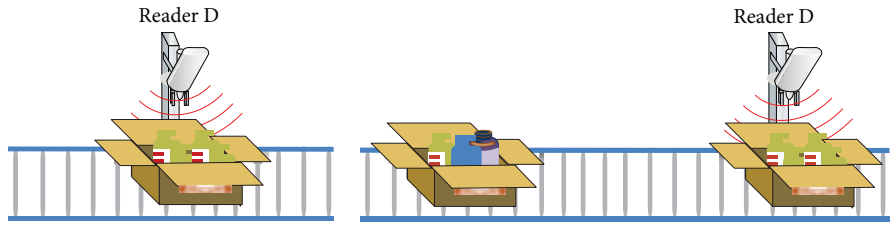
Table 2: Default value of situation context data.

Rule field	Subfield	Default value
oe	isOE	False
	fieldList	Null
ae	isAE	False
	hePattern	Null
	fieldList	Null
qe	bSL	Storing, retail_selling
	actionL	ADD
	fieldList	Null
te	isTE	False
	fieldList	Null
ne	isExist	False
	eventType	Null
	fieldList	Null

Here, the values of parent and children fields should be object class EPC based on the schema specified in the TDS standard [5]. The values in bSL field should be the business step values specified in the CBV standard [27]. The values of an action field should be ADD or OBSERVE.

The situation context data is the extra generation rules. Thus, it is optional for the FCA, which means there can be no situation context data for an FCA. If some rule fields of situation context data are null or incomplete or there is no situation context data for FACA, then default values of each rule field are used. Table 2 specifies the default value of each field. That is, FCA does not generate unnecessary OBSERVE ObjectEvent and there are no extensible fields in ObjectEvent. FCA does not generate unnecessary OBSERVE AggregationEvent and heterogeneous AggregationEvent. There is no extensible field in AggregationEvent either. FCA generates QuantityEvent if "ADD" AggregationEvent is generated for the incoming ECReports or if bizStep is storing or retail_selling. There is no extensible field in QuantityEvent either. FCA does not generate OBSERVE TransactionEvent and there are no extensible fields in TransactionEvent. FCA does not generate new event type.

An example of situation context data is shown in Figure 6.



(a) Four items are being packaged into a case

(b) Two cases come through reader ID in turn

Figure 5: Necessity of situation context data.

Table 3: The functions in business context data.

Function	Descprition
`checkSCD()`	If the field values of situation context data are valid, return true.
`copySCDFileds`(*e*, scd.*e*)	After checking the validity of scd.*e*, then copy fields of scd.*e* into appropriate fields of an event *e*.

Example 5. As Figure 6(a) shows, a user wants to pack the three heterogeneous items as a set. That is, a heterogeneous AggregationEvent for a set should be generated and unnecessary AggregatEvent for each item should not be generated here. Then we can use situation context data shown in Figure 6(b), which specifies the heterogeneous AggregationEvent rules. The parent ID class is in the pattern of urn: epc: id: sscc: 0652642, while the child EPCs classes are in the pattern of urn: epc: id: sgtin: 1234567. 123456, urn: epc: id: sgtin: 1234567. 223456, and urn: epc: id: sgtin: 2234567. 123456.

In order to generate EPCIS events, the situation context data needs to provide a function for checking errors. Table 3 shows the function.

3.1.2. Business Context Data. Business context data is used to obtain the bizLocation, bizStep, and disposition fields for a specific logical reader. It is represented in the format of 4 tuples $bcd = (reader, bizLoc, bizStep, \text{and } disposition)$. Definition 6 shows more details of the business context data. Here, the value of bizStep field should be the business step values specified in the CBV standard [27]. The values of disposition field should be the disposition values specified in the CBV standard [27].

Definition 6 (business context data). Business context data is represented as 4 tuples $bcd = (reader, bizLoc, bizStep, \text{and } disposition)$, where

(i) *reader* is a name of the reader;

(ii) *bizLoc* is a bizLocation;

(iii) *bizStep* is a bizStep;

(iv) *disposition* is a disposition.

Table 4: The functions in business context data.

Function	Description
`obtainBCDFields`(*e*, bcd)	Obtain the bizLocation, bizStep, and disposition fields of EPCISEvent *e* from bizLoc, bizStep, and disposition fields of business context data bcd.
`getBCD`(*r*, bcd)	Get the bizLoc, bizStep, and disposition fields from a business context data set bcd for ECReports *r*.
`checkBCD()`	If the field values of all the business context data are valid, return true. Otherwise, return false.

In order to generate EPCIS events, the business context data needs to provide several functions. Table 4 shows the functions.

Business context data is mandatory information for the FCA. All the elements of business context data for an FCA must be specified in an XML file.

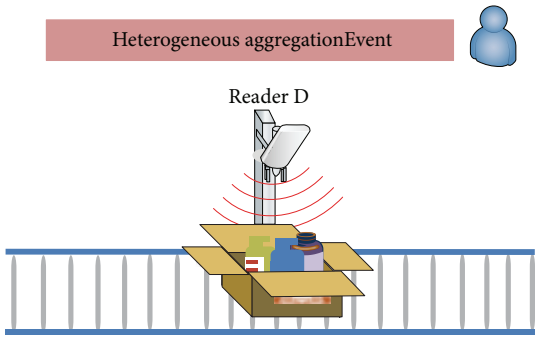
Example 7. Figure 7 shows an example of a business context data. There are four readers in an imaginary manufacturer supply chain in Figure 7(a). The corresponding business context must be specified for the four readers in an XML file. Figure 7(b) depicts only two business context data for reader A and reader D.

3.2. Transaction Data. Transaction data determines TransactionEvent type and provides the action field and bizTransactionList field of TransactionEvent. It is represented in the format of $td = (type, id, epcList, reader, \text{and } sr)$. Definition 8 shows the specification of transaction data. Here, the value of a type field should be obtained from the business transaction type values described in the core business vocabulary (CBV) standard [27]. The value of an epcList field should be based on the schema specified in the EPC tag data (TDS) standard [5].

Definition 8 (transaction data). Transaction data is represented as 4 tuples $td = (type, id, epcList, reader, \text{and } sr)$, where

(i) *type* is a transaction type;

(ii) *id* is a transaction ID number;

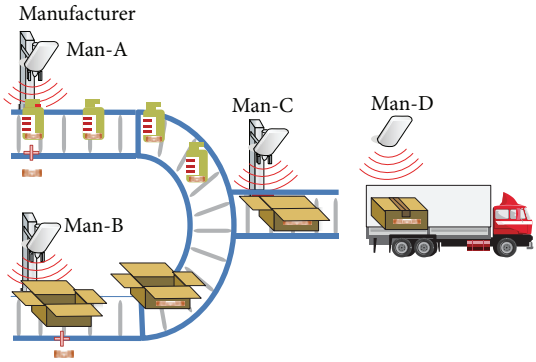


(a) The heterogeneous AggregationEvent case

```
<?xml version="1.0" encoding="UTF-8"?>
<SituationContextData>
  <ae>
    <isAE>false</isAE>
    <hePattern>
      <parent>urn:epc:id:sscc:0652642</parent>
      <children>
        <string>urn:epc:id:sgtin:1234567.123456</string>
        <string>urn:epc:id:sgtin:1234567.223456</string>
        <string>urn:epc:id:sgtin:2234567.123456</string>
      </children>
  </ae>
</SituationContextData>
```

(b) The situation context data for the heterogeneous AggregationEvent

Figure 6: An example of situation context data.



(a) A manufacturer supply chain

```
<?xml version="1.0" encoding="UTF-8"?>
<BusinessContextList>
  <BusinessContext>
    <reader>Man-A</reader>
    <bizLoc>Man-commissionA</bizLoc>
    <bizStep>commissioning</bizStep>
    <disposition>active</disposition>
  </BusinessContext>
  ...
  <BusinessContext>
    <reader>Man-D</reader>
    <bizLoc>Man-ship</bizLoc>
    <bizStep>shipping</bizStep>
    <disposition>in_progress</disposition>
  </BusinessContext>
</BusinessContextList>
```

(b) Business context data

Figure 7: An example of business context data.

(iii) *epcList* is a set of EPCs;

(iv) *reader* is a name of the reader;

(v) *sr* specifies whether the reader is the start (S), end (E), or intermediate reader (I) of the transaction.

Transaction data is optional for the FCA, which means there can be no transaction data for an FCA if there is no business transaction existing. All the transaction data for an FCA can be included in an XML file.

Example 9. Here is an example of transaction data. In a wholesaler supply chain as shown in Figure 8(a), two cases are picked up at reader Who-C when associated with the transaction order from a retailer. Figure 8(b) shows the corresponding transaction data.

In order to generate EPCIS events, transaction data needs to provide several functions. Table 5 shows the functions.

Table 5: The functions in transaction data.

Function	Description
`isRelevant`(*r*, td)	If the transaction data td is related to ECReports *r*, then return true. Otherwise, return false.
`addBizTran`(td, *t*)	Obtain the transaction type and ID of transaction data td and add them to the bizTransactionList of TransactionEvent *t*.
`checkTD()`	If the field values of all the transaction data are valid, return true. Otherwise, return false.

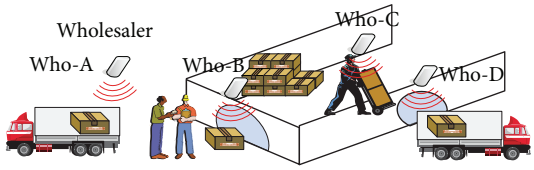
4. Flexible Capturing Application

In this section, we present our flexible capturing application (FCA) for generating correct EPCIS events. We first describe the architecture of FCA (Section 4.1) and then explain how our approach can generate correct EPCIS events and provide the correctness of our approach (Section 4.2). We also provide a running example of FCA (Section 4.3).

4.1. Architecture. Figure 9 shows the proposed architecture of the FCA. FCA consists of three principal components: ALE accessor, event generator, and EPCIS accessor.

ALE accessor deals with ECSpec and ECReports related to ALE. ECSpec manager configures one or more ECSpecs to request ECReports from ALE, and then ECReports handler processes the received ECReports.

After receiving ECReports from ALE, event generator generates EPCIS events with the processed ECReports



(a) A wholesaler supply chain

```
<?xml version="1.0" encoding="UTF-8"?>
<TransactionList>
  <TransactionData>
    <type>po</type>
    <id>1</id>
    <epcList>
      <string>urn:epc:id:sscc:0614141.1234567890</string>
      <string>urn:epc:id:sscc:0614141.1234567891</string>
    </epcList>
    <reader>Who-C</reader>
    <sr>S</sr>
  </TransactionData>
<TransactionList>
```

(b) Transaction data

Figure 8: An example of transaction data.

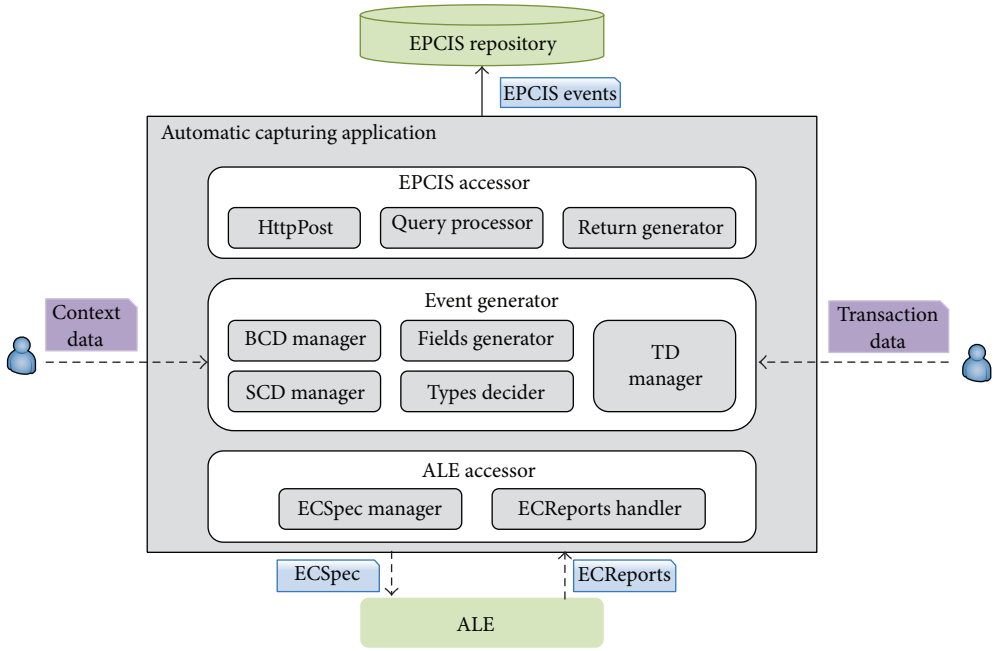


Figure 9: The architecture of FCA.

and other sources of data such as context data, transaction data (TD), and relevant AggregationEvent in EPCIS repository. TD manager deals with the input transaction data. BCD manager and SCD manager deal with input context data, which can be divided into business context data (BCD) and situation context data (SCD). Event generation rules about the four EPCIS event types are specified with ECReports, situation context data, transaction data, and relevant AggregationEvent in EPCIS repository. Type decider matches all event generation rules to decide the event types for the incoming ECReports. If an event generation rule is satisfied, fields generator generates the event fields with ECReports, business context data, transaction data, and relevant AggregationEvent in EPCIS repository. After matching all the four event generation rules, event generator generates all EPCIS events for the current ECReports.

EPCIS accessor provides the communication with EPCIS. Query processor retrieves the relevant AggregationEvent in EPCIS repository queried by event generator to generate AggregationEvent. HttpPost delivers all generated EPCIS events to EPCIS repository. Return generator returns the verified information of EPCIS events format from ECPIS. After delivering the generated EPCIS events for the current ECReports, event generator turns to deal with the next incoming processed ECReports.

4.2. Event Generation. In this subsection, we will explain the ECPIS event generation algorithm of FCA.

4.2.1. Event Generation Algorithm. The event generation algorithm first checks the context data and tries to generate appropriate EPCIS events by invoking all EPCIS event generation algorithms. The detailed steps of event generation algorithm are described in Algorithm 1. With the incoming ECReports (r), all the transaction data, the business context data, and the situation context data, the algorithm generates EPCIS

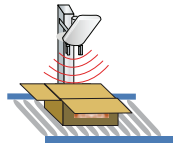
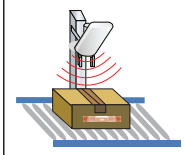
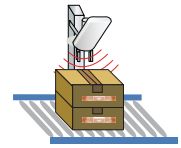
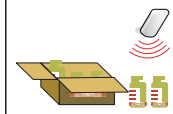
	Case 1	Case 2	Case 3	Case 4	Case 5
Explanation	EO and EC are aggregated for the first time	EO and EC are first read into a new enterprise and have been aggregated	EO and EC have been aggregated and EC has one element	EO and EC have been aggregated and EC has several elements	EO and EC are disaggregated
Figure					

FIGURE 10: The cases for AggregationEvent.

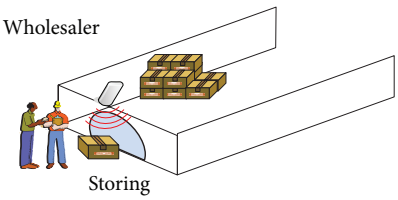
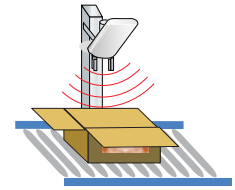
	Case 1	Case 2
Explanation	s.qe = null and bizStep is storing	s.Qe ≠ null and "ADD" AggregationEvent is generated
Figure		

FIGURE 11: The cases for QuantityEvent.

events (e). If the incoming ECReports include epcs in line 2, then event generation starts to match each event algorithm to generate all possible EPCIS events in lines 3–11. The events generated in each event algorithm are added into the final EPCIS events in lines 7–11. The`ObjectEvent generation` algorithm is matched after the other three EPCIS events for the reason that an ObjectEvent is generated if none of other three EPCIS event generation rules are satisfied.

4.2.2. AggregationEvent Generation. An AggregationEvent occurs when one or more items physically aggregated to each other. It establishes a parent-child relationship between shipping containers such as case and the items which are included within it. Thus we need to divide the incoming EPCs into containers (EC) and objects (EO) according to the encoding schema and epcClass.

In a supply chain, there exist five cases for the AggregationEvent as shown in Figure 10.

Case 1. This includes the first aggregation of item(s) and a container.

Since this is the first time of aggregation, there is no previous AggregationEvent for EC and EO. We generate an "ADD" AggregationEvent if the bizStep is not receiving which indicates that EO and EC are aggregated.

Case 2. This includes the first observation at a different business partner.

This is similar to Case1; the difference is due to BizStep. In a supply chain, whenever a wholesaler receives item(s) and a container, it is the first observation at wholesaler. Since it is already aggregated before the different business partner, we generate an "OBSERVE" AggregationEvent.

Case 3. A container and several items are recognized.

Since only one EPC is included in EC and the relevant previous AggregationEvent of EC is also found, we generate an "OBSERVE" AggregationEvent setting the parent ID and child EPCs to EC and EO, respectively.

Case 4. Several containers and items are recognized.

Since EC includes many EPCs of containers, we need to search all relevant AggregationEvents for each element EC_i in EC. Then, we can generate an OBSERVE AggregationEvent after setting the parent ID and child EPCs to EC_i and the child EPCs of previous aggregation event of EC_i, respectively.

Case 5. Items are removed from the container (disaggregation).

At this case, only EO is read and there exists a relevant previous AggregationEvent of EO. Then, we generate DELETE AggregationEvent. The child EPCs and parent ID are set to EO and the parent ID of the relevant (previous) AggregationEvent, respectively.

Algorithm 2 shows the `AggregationEvent generation`. It first divides EPCs into EC and EO and then generates

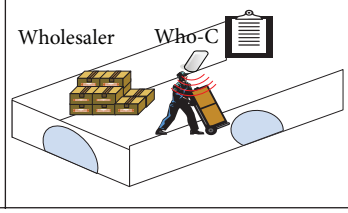
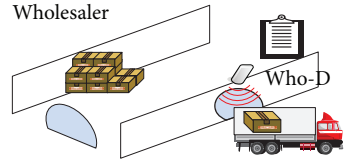
	Case 1	Case 2	Case 3
Explanation	isRelevant(r, tdi) = true and tdi.sr = S	isRelevant(r, tdi) = true and tdi.sr = I	isRelevant(r, tdi) = true and tdi.sr = E
Figure			

Figure 12: The cases for TransactionEvent.

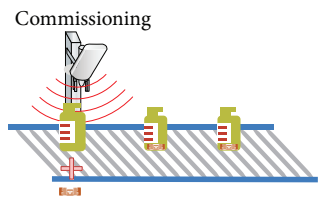
	Case 1	Case 2	Case 3
Explanation	bizStep is commissioning	bizStep is decommissioning	bizStep is shipping
Figure			

Figure 13: The cases for ObjectEvent.

a corresponding AggregationEvent according to five cases. The exact fields for the generated AggregationEvent are provided by EC, EC, relevant (previous) AggregationEvent, bizStep, situation context data, and business context data.

4.2.3. QuantityEvent Generation. A quantity event provides information on the number of items identified by RFID readers. But it does not give the individual identities of recognized items.

There exist two cases in a supply chain as shown in Figure 11. A wholesaler would like to know the quantity of boxes/items to expect upon receiving the shipment. This is an independent usage case of the QuentityEvent. For this case, although no situation context data is given, our FCA generates a QuantityEvent by using the default values of situation context data. On aggregating items into containers, a manufacturer would like to check the number of aggregated objects. FCA first checks an action field of situation context data. If the situation context data includes proper bizStep and action field (value of "ADD"), then FCA generates a QuantityEvent.

Algorithm 3 shows the `QuantityEvent generation` algorithm which deals with two usage cases of the QuantityEvent.

4.2.4. TransactionEvent Generation. A TransactionEvent occurs when an EPC-tagged item becomes associated or disassociated with a business transaction such as a purchase order.

As shown in Figure 12, each business transaction consists of three stages: (1) start, (2) intermediate, and (3) end. However, a TransactionEvent is generated for the start and end stages. A value "ADD" is set to the action value of a start transaction and "DELETE" is set to the action value of an end transaction. FCA obtains bizLocation, bizStep, and disposition fields from business context data and a bizTransactionList field from the relevant transaction data td_i. Details of `TransactionEvent generation` algorithm are described in Algorithm 4.

4.2.5. ObjectEvent Generation. Basically, an ObjectEvent applies to one or more objects identified by an EPC. However, when considering objects identification in a supply chain, there might be three different cases for ObjectEvent as shown in Figure 13. If bizStep is commissioned such as Case1, it means that an object is just identified at this time (birth of the object). Thus we will assign the value "ADD" to the action field. If bizStep is decommissioned such as Case 2, it means the end of life for an object. Thus we will assign the value "DELETE" to action field. Otherwise, it means that an object has not been changed and has been just recognized. The value "OBSERVE" will be assigned to an action field.

Algorithm 5 shows the `ObjectEvent generation` algorithm. At the first step, ObejctEvent type is decided by

```
Input: {r, td, bcd, s}-r is ECReports; td is a set of Transaction Data;
                    bcd is a set of Business Context Data;
                    s is Situation Context Data;
Output: {e}-e is a set of EPCIS events;
procedure EventGeneration(r, td, bcd, s)
(1)  e ← null
(2)  if r includes epc then
(3)     CheckTD();
(4)     CheckBCD();
(5)     CheckSCD();
(6)     bc ← getBCD(r, bcd);
(7)     e.add(AggregationEventGeneration(r, bc, s));
(8)     e.add(QuantityEventGeneration(r, bc, s));
(9)     e.add(TransactionEventGeneration(r, td, bc, s));
(10)    e.add(ObjectEventGeneration(r, bc, s));
(11)    e.add(NewEventTypeGeneration(r, bc, s));
     end if
(12) return e;
```

Algorithm 1: Event generation algorithm.

a situation context data, bizStep, and other event types. If a condition is satisfied, then, at the second step, the action field is obtained from by bizStep and other fields are obtained from ECReports, business context data, and situation context data.

Theorem 10. *Giving the ECReports r, business context data bc, and situation context data s, algorithm ObejctEvent correctly generates ObejctEvent o.*

Proof. In algorithm ObejctEvent generation, if no other event types are generated, if s.oe.isOE=true which means generate unnecessary OBSERVE ObjectEvent, if bc.bizStep=shipping or receiving which means to identify shipped objects, or if bc.bizStep=commissioning or decommissioning which means to commission or decommission EPCs to objects, we generate ObejectEvent. If bizStep is commissioned such as Case 1, then action value is ADD. If bizStep is decommissioned such as Case 2, then action value is DELETE. Otherwise, action value is OBSERVE such as Case 3. The epcList field can be obtained from ECReports. We generate the bizLocation, bizStep, and disposition fields from business context data. We generate extensible fields if s.oe.fieldList is not null. □

4.2.6. New Event Type Generation. Although the EPCglobal provides core event types such ObjectEvent, AggregationEvent, QunatityEvent, and TransactionEvent, a user defined event type is needed to meet the requirements of an industry or application area. For example, a manufacturer wants to quickly find returned items events; then it can define a new event type for returned items.

Example 11. A manufacturer ships the cases to a wholesaler. However, if the wholesaler finds out that the received case is a wrong product, then it sends the case back to the manufacturer. When the manufacturer receives the returned case, it generates the new event type named ReturnCaseEvent as shown in Figure 14(b) depending on situation data in Figure 14(a). The epcList field means the returned case. The bizLocation, bizStep, and disposition fields can be obtained from business context data. There is an extensible field with the name "reason" and the value "wrong product," which can be obtained from situation context data.

Algorithm 6 shows how new event type generation algorithm works. First, it defined the exact fields for a new event type from situation data. Then, it obtains the values of fields from ECReport, business context data, and situation context data.

Theorem 12. *Given an ECReport r, business context data, bc and situation context data s, algorithm new event type correctly generates new event type n.*

Proof. In algorithm new event type generation, if *s.ne.isExist* is true, we generate the new event type. The name of the new event type is obtained from SCD.getNEType.getEventType. The epcList field can be obtained from ECReports. We generate the bizLocation, bizStep, and disposition fields from business context data. We generate extensible fields if s.ne.fieldList is not null. □

4.3. A Running Example for Event Generation. After specifying the event generation rules in four EPCIS event generation algorithms, we present how FCA works to generate EPCIS events by using a manufacturer supply chain and business context data in Figure 15. There are no transaction data and situation context data for manufacturers FCA. Thus, transaction data is null here and the default values for situation context data are used.

A reader Man-A reads a number of items carrying EPC tags. FCA starts to generate EPCIS events after

```
   Input: {r, bc, s}-r is ECReports; bc is Business Context Data;
                      s is Situation Context Data;
   Output: {a}-a is (a set of) AggregationEvent;
Procedure AggregationEventGeneration(r, bc, s)
(1)   a ← null;
(2)   divide r into container (EC) and objects (EO);
(3)   if EC ≠ null and EO ≠ null then
(4)      if EC.size = 1 and findAggregationEvent(EC) = null and
         findAggregationEvent(EO) = null then
(5)            if getBizStep(bc) ≠ receiving then a.action ← ADD;          // Case 1
(6)            else a.action ← OBSERV E;                                   // Case 2
(7)            a.parentID ← EC;
(8)            a.childEPCs ← EO;
(9)            obtainBCDFields(a, bc);
(10)           copySCDFields(a, s.ae);
         end if
(11)     if s.ae.isAE = true then
(12)           if EC.size = 1 and findAggregationEvent(EC) ≠ null then     // Case 3
(13)              a.action ← OBSERV E;
(14)              a.parentID ← EC;
(15)              a.childEPCs ← EO;
(16)              obtainBCDFields(a, bc);
(17)              copySCDFields(a, s.ae);
               end if
(18)           if EC.size > 1 then                                         // Case 4
(19)              for each epc EC_i in EC do
(20)                 if findAggregationEvent(EC_i) ≠ null then
(21)                    generate AggregationEvent sub;
(22)                    sub.action ← OBSERV E;
(23)                    sub.parentID ← EC_i;
(24)                    sub.childEPCs ← findAggregationEvent(EC_i).getchildEPCs();
(25)                    obtainBCDFields(a, bc);
(26)                    copySCDFields(sub, s.ae);
(27)                    a.add(sub);
                     end if
                  end for
               end if
         end if
      end if
(28)  if EC = null and EO ≠ null and findAggregation(EO) ≠ null then       // Case 5
(29)     a.action ← DELETE;
(30)     a.parentID ← findAggregation(EO).getParentID();
(31)     a.childEPCs ← EO;
(32)     obtainBCDFields(a, bc);
(33)     copySCDFields(a, s.ae);
      end if
(34)  return a;
```

Algorithm 2: AggregationEvent generation algorithm.

receiving the ECReports from ALE. Use the `AggregationEvent generation` algorithm to match each generation rule. After dividing the incoming epcs, only EO is read. EO is the items. However, there is no relevant AggregationEvent about EO in manufacturers EPCIS repository, so no AggregationEvent generation rule is satisfied. Use the QuantityEvent generation algorithm to match each generation rule. No "ADD" AggregationEvent is generated and the bizStep is not storing or retail_selling, so no QuantityEvent generation rule is satisfied. Use the TransactionEvent generation algorithm to match each generation rule. There is no transaction data here, so no TransactionEvent generation rule is satisfied. Use the ObjectEvent generation algorithm to match each generation rule. The bizStep is commissioned, so generate ADD ObjectEvent. The epcList is the EPCs of items. The bizLocation, bizStep, and disposition fields can be obtained from manufacturers business context data. Thus, only ObjectEvent is generated. FCA delivers the generated ObjectEvent to manufacturers EPCIS repository.

```
    Input: {r, bc, s}-r is ECReports; bc is Business Context Data;
                        s is Situation Context Data;
    Output: {q}-q is QuantityEvent;
procedure QuantityEventGeneration(r, bc, s)
(1)   q ← null;
(2)   if s.qe = null then
(3)     if (bc.bizStep is storing or receiving) then                    // Case 1
(4)       obtain the epcClass and quantity fields of q from r;
(5)       obtainBCDFields(q, bc);
        end if
      else
(6)     if s.qe.bSList.contains(bc.bizStep) or s.qe.actionList.contains(ADD) then   // Case 2
(7)       obtain the epcClass and quantity fields of q from r;
(8)       obtainBCDFields(q, bc);
(9)       copySCDFields(q, s.qe);
        end if
      end if
(10)  return q;
```

Algorithm 3: QuantityEvent generation algorithm.

```
    Input: {r, td, bc, s}-r is ECReports; td is a set of transaction data;
                        bc is business context data;
                        s is situation context data;
    Output: {t}-t is a set of TransactionEvent;
procedure TransactionEventGeneration(r, td, bc, s)
(1)   t ← null;
(2)   for each transaction data td_i in td do
(3)     if isRelevant(r, td_i) = true and s.te.is TE = true then
(4)       if td_i.sr! = I then                                          // not Case 2
(5)         if td_i.sr = S then sub.action ← ADD;                       // Case 1
(6)         if td_i.sr = E then sub.action ← DELETE;                    // Case 3
(7)         sub.epcList ← r.epcList;
(8)         obtainBCDFields(sub, bc);
(9)         addBizTran(td_i, sub);
(10)        copySCDFields(sub, s.te);
(11)        t.add(sub);
          end if
        end if
      end for
(12)  return t;
```

Algorithm 4: TransactionEvent generation algorithm.

Similarly, only ObjectEvent is generated when reader Man-B reads an EPC-equipped case. AggregationEvent and QuantityEvent are generated when reader Man-C reads a number of items carrying EPC tags and an EPC-equipped case. Only ObjectEvent is generated when reader Man-D reads a number of items carrying EPC tags and an EPC-equipped case.

Table 6 summarizes the generated EPCIS event types of the manufacturer supply chain by using FCA.

Table 6: Generated event types for the manufacturer supply chain example.

Reader	Generated event types	Action
Man-A	ObjectEvent	ADD
Man-B	ObjectEvent	ADD
Man-C	AggregationEvent	ADD
	QuantityEvent	X
Man-D	ObjectEvent	OBSERVE

5. Experimental Results

To demonstrate the flexibility of FCA, we implemented the FCA according to the architecture in Figure 9 and defined the equation of flexibility. We selected BizAF [25] as the naive approach to be compared with FCA. We used

```
Input: {r, bc, s}-r is ECReports; bc is Business Context Data; s is Situation
                Context Data;
Output: {o}-o is ObjectEvent;
procedure ObjectEventGeneration(r, bc, s)
(1)  o ← null;
(2)  if no other event type generates or s.oe.isOE = true or bc.bizStep = shipping or receiving then
(3)     if bc.bizStep = commissioning then o.action ← ADD;                    // Case 1
(4)     else if bc.bizStep = decommissioning then o.action ← DELETE;          // Case 2
(5)     else o.action ← OBSERV E;                                             // Case 3
(6)     o.epcList ← r.epcList;
(7)     obtainBCDFields(o, bc);
(8)     copySCDFields(o, s.oe);
        end if
(9)  return o;
```

Algorithm 5: ObjectEvent generation algorithm.

```
Input: {r, bc, s}-r is ECReports; bc is Business Context Data; s is Situation
                Context Data;
Output: {n}-n is New Event Type;
procedure NewEventTypeGeneration(r, bc, s)
(1)  n ← null;
(2)  if s.ne.isExist = true then
(3)     n.name ← s.ne.eventT ype;
(4)     n.epcList ← r.epcList;
(5)     obtainBCDFields(n, bc);
(6)     copySCDFields(n, s.ne);
        end if
(7)  return n;
```

Algorithm 6: New event type generation algorithm.

two different datasets in the performance evaluation of FCA: (1) a synthetically generated dataset and (2) a real dataset.

5.1. Implementation. We implemented FCA in Java language using Eclipse 3.6 and complied the code using JDK 1.6. All experiments were conducted on an Intel Core 2 Duo 3.00 GHz machine with 2 GB Ram running Windows 7 in 32 bits.

Figure 16 shows the GUI of the implemented FCA. ECSpecs, business context data, situation context data, and transaction data should be provided. The ECSpecs file includes the name and path of all readers ECSpec. The user can input the path of required data, as well as loading them in the load tab. What is more is that the user can draw his own data about ECSpecs, business context data, and transaction data in the panel of draw tab. When the user clicks the start button, FCA starts working. After ECSpecs are sent to ALE, ECReports are received from ALE. Then, FCA handles the incoming ECReports and generates EPCIS events for incoming ECReports. At last FCA delivers the generated events to EPCIS repository and the EPCIS returns the verified information of generated events format to FCA.

5.2. Evaluation Metric. We selected two factors to compute flexibility. One factor is the total number of tag records in generated EPCIS events. The other factor is the number of tag data involved in event generation. We want to generate more tag records in generated EPCIS events for all tag data. Therefore, we compute the flexibility f as shown in (1). Here, $\sum_{i=1}^{n} N_i$ is total number of tag records in generated EPCIS events, n is the number of generated events, N_i is the number tags included in a generated event, and N_T is the number of tag data involved in event generation:

$$f = \frac{\sum_{i=1}^{n} N_i}{N_T}. \tag{1}$$

5.3. Experiments with Synthetic Data. In this section, we analyze the performance of FCA.

5.3.1. A Synthetic Dataset. Since the collection for RFID tag data from multiple readers is time-consuming and money-consuming, we generated the virtual tag data using a synthetic RFID data generator developed by us. For this purpose, we made an imaginary manufacturer-wholesaler-retailer supply chains scenario shown in in Figure 17. We assume that each enterprise has its own flexible capturing application and EPCIS. Thus, three virtual tag data sets are used to evaluate the FCA.

```
<?xml version="1.0" encoding="UTF-8"?>
<SituationContextData>
  <ne>
    <isExist>true</isExist>
    <eventType>ReturnCaseEvent</eventType>
    <fieldList>
      <field>
        <name>reason</name>
        <value>wrong product</value>
      </field>
    </fieldList>
  </ne>
</SituationContextData>
```

(a) Situation context data

ReturnCaseEvent

-eventTime

-epcList:

urn:epc:id:sscc:0652642.1234567890

-readPoint: Man-RP:E

-bizLocation: Man-return

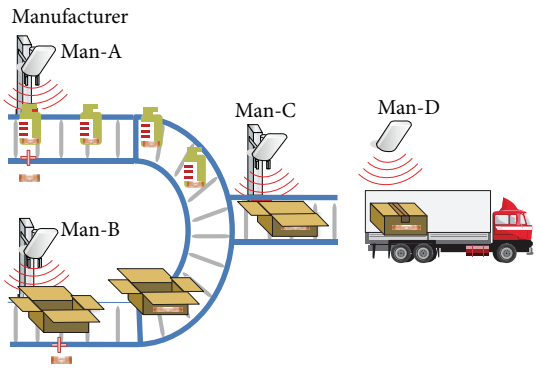
-bizStep: holding

-disposition: non_sellable_recalled

-reason: wrong product

(b) A generated new event type

Figure 14: New event type generation.



(a) A manufacturer supply chain

```
<?xml version="1.0" encoding="UTF-8"?>
<BusinessContextList>
  <BusinessContext>
    <reader>Man-A</reader>
    <bizLoc>Man-commissionA</bizLoc>
    <bizStep>commissioning</bizStep>
    <disposition>active</disposition>
  </BusinessContext>
  ...
  <BusinessContext>
    <reader>Man-D</reader>
    <bizLoc>Man-ship</bizLoc>
    <bizStep>shipping</bizStep>
    <disposition>in_progress</disposition>
  </BusinessContext>
</BusinessContextList>
```

(b) Business context data

Figure 15: A running example of EPCIS events generation.

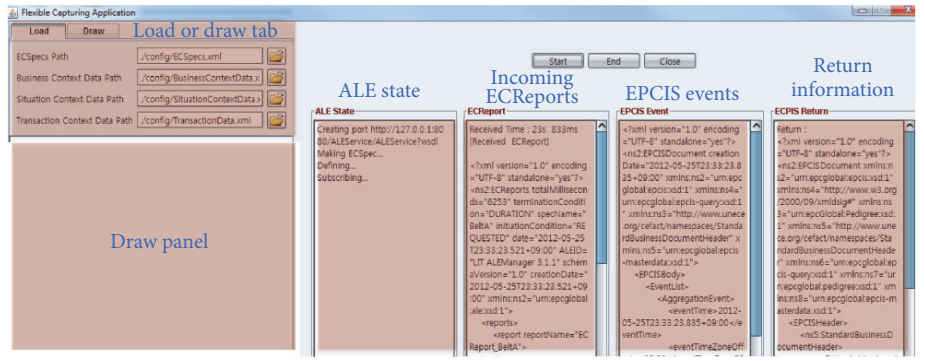


Figure 16: The GUI of implemented FCA.

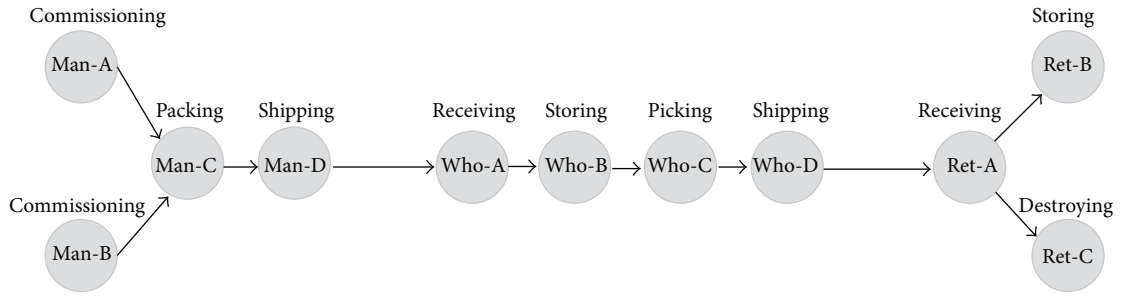


Figure 17: Supply chains scenario for the synthetic data.

The first data set consists of 10000 tag data in the retailer supply chain as shown in Figure 18(a). The tags can arrive at a reader in one ECReport at once or be divided into several ECReports for multiple reads at a reader. There are R ($R < 10000$) item tags in reader Ret-B in the format of urn:epc:id:sgtin:1234567.123456.X; X is from 1 to R. There are 10000-R case tags in reader Ret-C in the format of urn:epc:id:sscc:0641414.12345X; X is from 00001- to 10000-R.

The second data set consists of 20000 tag data in the manufacturer supply chain as shown in Figure 18(b). The tags can arrive at a reader in one ECReport at once or be divided into several ECReports for multiple reads at a reader. There are M ($M < 20000$) item tags in reader Man-A in the format of urn:epc:id:sgtin:1234567.123456.X; X is from 1 to M. There are 20000-M case tags in reader Man-B in the format of urn:epc:id:sscc:0641414.12345X; X is from 00001- to 20000-M.

The third data set consists of 30000 tag data in the wholesaler supply chain as shown in Figure 18(c). The tags can arrive at a reader in one ECReport at once or be divided into several ECReports for multiple reads at a reader. For each reader in the wholesaler supply chain, there are W ($W < 30000$) item tags in the format of urn:epc:id:sgtin:1234567.123456.X; X is from 1 to W. There are 30000-W case tags in the format of urn:epc:id:sscc:0641414.12345X; X is from 00001- to 30000-W.

5.3.2. Performance Analysis. Figure 19(a) shows the performance comparison in terms of the flexibility. The flexibility of FCA is higher than BizAF, which means that our proposed FCA provides more flexibility for generating EPCIS events. For the data set in the retailer supply chain, FCA generates ObjectEvent for identifying the tags, AggregationEvent for aggregation between items and case, and TransactionEvent for transaction at reader Ret-A. FCA generates AggregationEvent for disaggregation and QuantityEvent for storing at reader Ret-B, while BizAF generates only one event type at each reader. Thus, the flexibility of FCA is higher than BizAF for the data set in the retailer supply chain. For the data set in the manufacturer supply chain, FCA generates AggregationEvent and QuantityEvent at reader Man-C and AggregationEvent and TransactionEvent at reader Man-D, while BizAF generates only one event type at each reader. Thus, the flexibility of FCA is higher than BizAF. For the data set in the wholesaler supply chain, FCA generates ObjectEvent and AggregationEvent at reader Who-A, AggregationEvent and QuantityEvent at reader Who-B, TransactionEvent and AggregationEvent at reader Who-C, and ObjectEvent and AggregationEvent at reader Who-D. Thus, the flexibility of FCA is higher than BizAF.

Tag data are included in an ECReport as the input of capturing application to be processed. The relationship between the number of ECReports N_c and the number of tag data involved in event generation N_T is as in shown (2). Here N_t is the number of tags (item and case should be both included) which arrive at a reader at the same time. N_r is the number of readers. For example, 9 item tags and 1 case tag arrive at reader Who-A at the same time; then N_t is 10. There are four readers in the wholesaler supply chain; thus N_r is 4. If N_T is 30000, then value of N_c is 12000:

$$N_c = \frac{N_T}{N_t} * N_r. \quad (2)$$

Figure 19(b) shows the comparison of execution time using the wholesaler supply chain. 9 item tags and 1 case tag are included in one ECReport. We used 25000, 50000, 75000, and 100000 tag data to obtain 10000, 20000, 30000, and 40000 ECReports, respectively. FCA matches event generation rules to decide the event types and processes semantic data to obtain some event fields, while BizAF configures the event type and event fields. Therefore, our FCA takes a little more time than BizAF.

The memory is around a specific value while running FCA and BizAF. Thus, we listed the value of memory usage for comparison in Table 7. There is no so big difference in memory usage. The memory of our FCA is a little smaller than BizAF.

From the above comparison, we can know that FCA can generate EPCIS events more flexibly with a little more time.

5.4. Experiments with Real Data Set. In this subsection, we conducted experiments to verify flexibility of FCA using real RFID data set.

5.4.1. Real Dataset. To obtain the real data from the real RFID devices, we designed a virtual manufacturer's scenario and installed a belt conveyor as shown in Figure 20(a), a roller conveyor as shown in Figure 20(b), and three RFID readers (two Alien 9800 readers and one Intermec reader) according to the scenario.

Details of the virtual manufacturer's scenario in Figure 21 are as follows.

(i) *Conveyor Belt A*. It includes putting 2 separated items and 2 items of product at the start position (belt A).

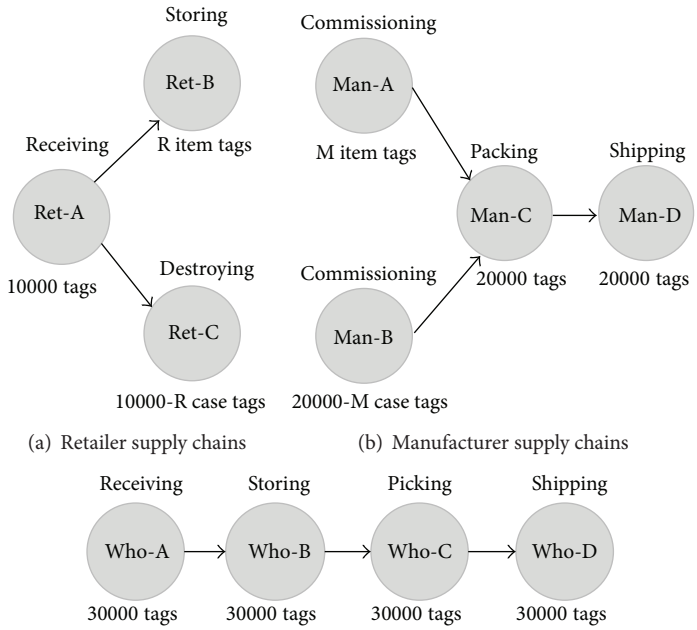
(ii) *Passing Read Point 1 (Reader 1)*. The antenna will read and transmit the EPC tags of the items to ALE (application level event) and pass them to FCA (flexible capturing application) for generating object event.

(iii) *Passing Round-Belt*. The product will be aggregated and packed together in a box (containing 4 items) and then moves on to belt B.

(iv) *Passing Read Point 2 (Reader 2)*. The antenna will read and transmit the EPC tags of the items to ALE and pass them to FCA for generating aggregation event.

(v) *To Roller Conveyor*. After passing through belt B, the box will be moving on to a roller conveyor.

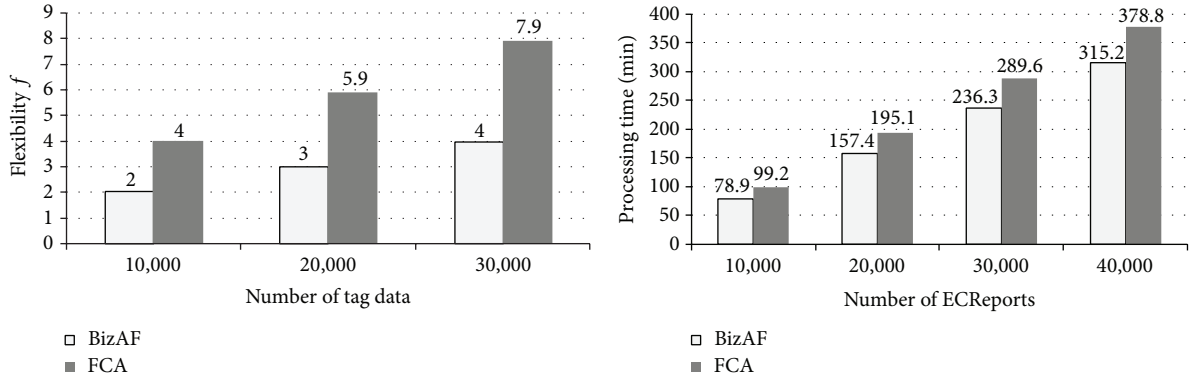
(vi) *Passing Read Point 3 (Reader 3)*. The antenna will read and transmit the EPC tags again and transmit them to ALE and pass them to FCA for generating transaction event.



(a) Retailer supply chains (b) Manufacturer supply chains

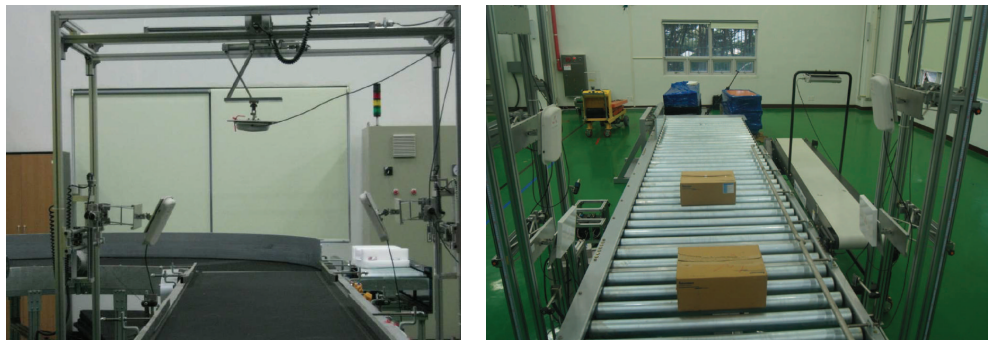
(c) Wholesaler supply chains

Figure 18: Data sets in retailer/manufacturer/wholesaler supply chains.



(a) Flexibility (b) Execution time

Figure 19: Experimental results.



(a) Installed belt conveyor (b) Installed roller conveyor

Figure 20: Installed devices.

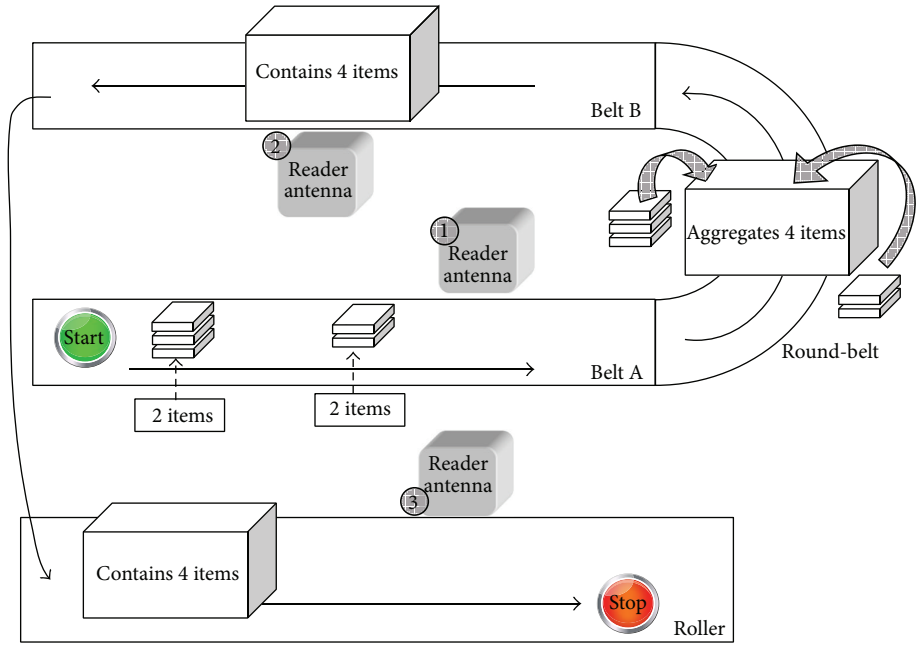


Figure 21: A virtual manufacturer's scenario.

Table 7: The comparison of memory usage.

	Memory usage
BizAF	120 MB
FCA	100 MB

Table 8: The experimental results.

Description	Number of events	Generated event types
Reader 1	3	ObjectEvent, AggregationEvent, QuantityEvent
Reader 2	1	ObjectEvent
Reader 3	2	AggregationEvent, QuantityEvent

5.4.2. Experimental Results. Algorithm 7 shows the generated EPCIS events of reader 2. As we expected, the reader detected the 5 items (1 box in *sscc* format and 4 items in *sgtin* format) and generated an ObjectEvent for 5 items. It generated an AggregateEvent describing information about 4 tag items. It also generated a QuantityEvent containing the number of included items.

The results of the virtual manufacturer's scenario are summarized in Table 8. We can perform automatic event generation according to various actions that happened during the virtual manufacturer's scenario by providing business context data and situation context data.

6. Related Work

The RFID technology is widely used to track and trace objects. Due to the scenario diversity and EPCIS events generation complexity, there are few studies on capturing application [22–26, 28, 29] to generate EPCIS events arising from supply chain activity.

Fosstrak et al. [22, 23] provide a custom capturing application and a generic capturing application to generate EPCIS events for an EPCIS repository. The custom capturing application can only generate ObjectEvent in the specified scenario, which is not appropriate for diverse supply chains to generate four EPCIS event types. In the generic capturing application case, a user must define and implement a set of JBoss rules (Drools) for generating EPCIS events.

WebSphere [24] uses an EPCIS connector to generate EPCIS events from incoming ECReports. When the EPCIS connector receives an ECReport, it creates an EPCIS event with the incoming ECReports and the metadata in a specified table. The metadata contains the event type and event fields to be generated, which is configured by the user.

BizAF [25] is proposed to develop RFID business applications cost-effectively in the EPC network. The capturing service in BizAF can play the role of capturing application to generate EPCIS events with RFID tag data. The user should define the condition and dataset to specify the event types and event fields. Thus, the capturing service collects real-time RFID tag data and generates EPCIS events according to the specific event type and event fields. However, the existing configuration method such as BizAF has the limitation that one event type is mapped to one reader name. That is, only one event type is generated at a specific logical reader. Different EPCIS event types cannot be generated for different incoming ECReports flexibly at a specific reader. Problems occur when different EPCIS event types are required.

The RSN tool [28, 29] can be used to simulation of RFID middleware by extending Petri nets. Users can virtually test RFID environments by setting the physical parameters of

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
...
<EPCISBody>
   <EventList>
      <ObjectEvent>
         <eventTime>2013-10-21T16:59:01.631+09:00</eventTime>
         <eventTimeZoneOffset>+09:00</eventTimeZoneOffset>
         <epcList>
            <epc>urn:epc:id:sgtin:2234567.223456.11</epc>
            <epc>urn:epc:id:sscc:0614141.1234567891</epc>
            <epc>urn:epc:id:sgtin:2234567.223456.3</epc>
            <epc>urn:epc:id:sgtin:2234567.223456.2</epc>
            <epc>urn:epc:id:sgtin:2234567.223456.15</epc>
         </epcList>
         <action>ADD</action>
         <bizStep>commissioning</bizStep>
         <disposition>active</disposition>
         <readPoint>   <id>BeltA</id>   </readPoint>
         <bizLocation>   <id>Belt-commission</id>   </bizLocation>
      </ObjectEvent>
      <AggregationEvent>
         <eventTime>2013-10-21T16:59:01.631+09:00</eventTime>
         <eventTimeZoneOffset>+09:00</eventTimeZoneOffset>
         <parentID>urn:epc:id:sscc:0614141.1234567891</parentID>
         <childEPCs>
            <epc>urn:epc:id:sgtin:2234567.223456.11</epc>
            <epc>urn:epc:id:sgtin:2234567.223456.3</epc>
            <epc>urn:epc:id:sgtin:2234567.223456.2</epc>
            <epc>urn:epc:id:sgtin:2234567.223456.15</epc>
         </childEPCs>
         <action>ADD</action>
         <bizStep>commissioning</bizStep>
         <disposition>active</disposition>
         <readPoint>   <id>BeltA</id>   </readPoint>
         <bizLocation>   <id>Belt-commission</id>   </bizLocation>
      </AggregationEvent>
      <QuantityEvent>
         <eventTime>2013-10-21T16:59:01.631+09:00</eventTime>
         <eventTimeZoneOffset>+09:00</eventTimeZoneOffset>
         <epcClass>urn:epc:id:sgtin:2234567.223456</epcClass>
         <quantity>4</quantity>
         <bizStep>commissioning</bizStep>
         <disposition>active</disposition>
         <readPoint>   <id>BeltA</id>   </readPoint>
         <bizLocation>   <id>Belt-commission</id>   </bizLocation>
      </QuantityEvent>
   </EventList>
  </EPCISBody>
</ns2:EPCISDocument>
```

Algorithm 7: The generated EPCIS events of reader 1.

RFID readers. However, the RSN tool does not provide the functionalities for EPCIS events.

APDL [26] system can generate the EPCIS events based on the newly defined language. The BEG module in APDL plays the role of a capturing application. BEG automates the mapping between reports stemming from ALE and EPCIS events. APDL describes a business process in a coherent way that combines the ECSpec, LRSpec, and master data together. APDL captures the data and semantics of RFID processes. Several ECReports are defined at any ECSpec for generating the four EPCIS events. Thus, BEG generates EPCIS events and stores the generated events at the EPCIS repository. The BEG in APDL only specifies how to generate event fields for each EPCIS event. However, the condition and timing for each event type are unknown. That is, BEG does not know on which condition to generate which event type. The action

field is complex for each event type; BEG does not specify how to get the value of the action field either.

7. Conclusion

There is no well-known study on specifying how to generate high level EPCIS events. Capturing application does not know to generate which EPCIS event types and how to get each semantic field value for an incoming ECReport. That is, the event types and semantic event fields are both uncertain.

In this paper, the FCA is proposed to solve the problem. FCA specifies generation rules about the four EPCIS event types with ECReports, situation context data, and transaction data to decide the event types for the incoming ECReports. If an event type generation rule is satisfied, event fields are generated with the ECReports, business context data, transaction data, and relevant AggregationEvent. After matching all the four EPCIS event type generation rules, FCA sends the generated EPCIS events to the EPCIS repository and turns to deal with the next incoming ECReports.

The incoming RFID tag data may be wrong due to the read problem caused by device. The event cycle defined to communicate between ALE and capturing application may also cause some tag data lost. Thus the generated EPCIS events are wrong due to the wrong incoming RFID tag data. However, after generating EPCIS events and delivering to EPCIS repository, capturing application cannot delete or modify EPCIS events. The only way is to generate subsequent EPCIS events. In future work, we plan to extend our work to generate subsequent EPCIS events for retracting or correcting prior to EPCIS events.

Notation

oe/ae/qe/te/ne:	ObjectEvent/AggregateEvent/ QuantityEvent/TransactionEvent/ NewEvent generation rule
isOE/isAE/isTE:	a variable for generating unnecessary OBSERVE ObjectEvent/Aggregation/ TransactionEvent/or not
isExist:	a variable for generating a new event type
fieldList:	a set of extensible fields and each field has a pair of name and value
actionList:	a set of actions
bSList:	a set of bizSteps.

Disclosure

This paper is an extended version of a master thesis [30].

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by Bioindustry Technology Development Program, Ministry for Food, Agriculture, Forestry, and Fisheries, Republic of Korea.

References

[1] H. Bhatt and B. Glover, *RFID Essentials*, O'Reilly, Sebastopol, Calif, USA, 2006.

[2] S. Ahson and M. Ilyas, *RFID Handbook: Applications, Technology, Security and Privacy*, CRC Press, Boca Raton, Fla, USA, 2008.

[3] S. Kim and G. Garrison, "Understanding users'behaviors regarding supply chain technology: determinants impacting the adoption and implementation of RFID technology in South Korea," *International Journal of Information Management*, vol. 30, no. 5, pp. 388–398, 2010.

[4] S. Sarma, D. Brock, and D. Engels, "Radio frequency identification and the electronic product code," *IEEE Micro*, vol. 21, no. 6, pp. 50–54, 2001.

[5] EPCglocal, "Gs1 epc tag data standard 1.6," 2011, http://www.gs1.org/gsmp/kc/epcglobal/tds/.

[6] A. S. Martínez-Sala, E. Egea-López, F. García-Sánchez, and J. García-Haro, "Tracking of returnable packaging and transport units with active RFID in the grocery supply chain," *Computers in Industry*, vol. 60, no. 3, pp. 161–171, 2009.

[7] J.-L. Chen, M.-C. Chen, C.-W. Chen, and Y.-C. Chang, "Architecture design and performance evaluation of RFID object tracking systems," *Computer Communications*, vol. 30, no. 9, pp. 2070–2086, 2007.

[8] Y. Gao, D. Yang, and W. Ning, "RFID application in tire manufacturing logistics," in *Proceedings of the 2nd IEEE International Conference on Advanced Management Science (ICAMS '10)*, pp. 109–112, July 2010.

[9] D. Yue, X. Wu, and J. Bai, "RFID application framework for pharmaceutical supply chain," in *Proceedings of the IEEE International Conference on Service Operations and Logistics, and Informatics (IEEE/SOLI '08)*, pp. 1125–1130, October 2008.

[10] M. Bustillo, "Wal-mart radio tags to track clothing," 2010, http://online.wsj.com/article/SB10001424052748704383213061-198090.html.

[11] EPCglocal, http://www.gs1.org/epcglobal/.

[12] EPCglocal, "The epcglobal architecture framework epcglobal final version 1.4," 2010, http://www.gs1.org/gsmp/kc/epcglobal/architecture/.

[13] F. Thiesse, C. Floerkemeier, M. Harrison, F. Michahelles, and C. Roduner, "Technology, standards, and real-world deployments of the EPC network," *IEEE Internet Computing*, vol. 13, no. 2, pp. 36–43, 2009.

[14] D. Sundaram, W. Zhou, S. Piramuthu, and S. Pienaar, "Knowledge-based RFID enabled Web Service architecture for supply chain management," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7937–7946, 2010.

[15] EPCglocal, "The reader protocol standard, version 1.1," 2006.

[16] EPCglocal, "The application level events (ale) specification, version 1.1.1 part I: core specification," 2009, http://www.gs1.org/gsmp/kc/epcglobal/ale/.

[17] EPCglocal, "Epc information services (epcis) version 1.0.1 specification," 2007, http://www.gs1.org/gsmp/kc/epcglobal/epcis/.

[18] F. Wang, S. Liu, and P. Liu, "A temporal RFID data model for querying physical objects," *Pervasive and Mobile Computing*, vol. 6, no. 3, pp. 382–397, 2010.

[19] H. Zhang, J. Kwon, and B. Hong, "A graph model based simulation tool for generating rfid streaming data," in *Proceedings of the 13th Asia-Pacific Web Conference*, pp. 290–300, 2011.

[20] T. Nguyen, Y.-K. Lee, B.-S. Jeong, and S. Lee, "Event query processing in epc information services," in *Proceedings of the 3rd IEEE International Conference on Signal Image Technologies and Internet Based Systems (SITIS '07)*, pp. 159–166, December 2007.

[21] D. Guinard, M. Mueller, and J. Pasquier-Rocha, "Giving RFID a REST: building a web-enabled EPCIS," in *Proceedings of the 2nd International Internet of Things Conference (IoT '10)*, pp. 1–8, December 2010.

[22] C. Floerkemeier, C. Roduner, and M. Lampe, "Rfid application development with the accada middleware platform," *Systems Journal*, vol. 1, no. 2, pp. 82–94, 2007.

[23] Fosstrak, http://code.google.com/p/fosstrak/wiki/AleCapturingApp/.

[24] I. WebSphere, "Websphere premises server v6.1 epcis support," 2008.

[25] T. Nam and K. Yeom, "Business-aware framework for supporting RFID-enabled applications in EPC Network," *Journal of Network and Computer Applications*, vol. 34, no. 3, pp. 958–971, 2011.

[26] N. Kefalakis, J. Soldatos, N. Konstantinou, and N. R. Prasad, "Apdl: a reference xml schema for process-centered definition of rfid solutions," *Journal of Systems and Software*, vol. 84, no. 7, pp. 1244–1259, 2011.

[27] EPCglocal, "Core business vocabulary standard," 2010, http://www.gs1.org/gsmp/kc/epcglobal/cbv/.

[28] W. Ryu, J. Kwon, and B. Hong, "A simulation network model to evaluate RFID middlewares," *International Journal of Software Engineering and Knowledge Engineering*, vol. 21, no. 6, pp. 779–801, 2011.

[29] W. Ryu, J. Kwon, and B. Hong, "Generation of RFID test datasets using RSN tool," *Personal and Ubiquitous Computing*, vol. 17, no. 7, pp. 1409–1419, 2013.

[30] F. Jia, *Flexible cpaturing application for generating extensible epcis events [M.S. thesis]*, Pusan National University, Busan, Korea, 2012.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 832510, 7 pages
http://dx.doi.org/10.1155/2014/832510

Research Article

Real Time Traceability and Monitoring System for Agricultural Products Based on Wireless Sensor Network

Daesik Ko,[1] Yunsik Kwak,[2] and Seokil Song[2]

[1] *Department of Electronic Engineering, Mokwon University, Seo-gu, Daejeon 302-729, Republic of Korea*
[2] *Department of Computer Engineering, Korea National University of Transportation, Chungju, Chungbuk 380-702, Republic of Korea*

Correspondence should be addressed to Seokil Song; sisong@ut.ac.kr

Received 29 November 2013; Accepted 2 May 2014; Published 24 June 2014

Academic Editor: Ken Choi

A system to monitor and trace the yields and distribution of agricultural products is an important precision agricultural application. The system can be used to predict the next year's yields or provide the distribution channel of agricultural products from farmers to customers. Existing traceability and monitoring systems, usually, are implemented by using both wireless sensor network (WSN) and radio frequency identification (RFID) techniques. In this paper, we propose a system architecture that only requires WSN techniques and implement it. The implementing system consists of sensor nodes, communication hubs, a communication protocol, and an event detection engine. We define events to trace and monitor agricultural products, and our event detection engine detects those events by using the current location and the status of each sensor node. We describe the overall architecture of the proposed system and implementation details.

1. Introduction

During the past couple of decades, precision agriculture (PA) has emerged with the advances of electronic equipment which has enabled farmers to increase the efficiency of their operations and develop new farming practices. PA is based on numerous technologies and infrastructures such as data gathering and management systems, geographic information systems (GIS), global positioning systems (GPS), microelectronics, wireless sensor networks (WSNs), and radio frequency identification (RFID) technologies [1–6]. Various applications of PA have been proposed and researched extensively. Among many applications, we focus on the agricultural products' trace and monitoring systems.

Monitoring and tracing systems for the yields and distribution of agricultural products are one of the important PA applications. The results of the systems can be used to predict the next year's yields. Also, we can provide the distribution channel of agricultural products from farmers to customers by using those systems. Most of the existing systems are based on both RFID systems and WSNs. Usually, RFIDs are used to trace agricultural products and measure their yields, and WSNs are for monitoring the environment during the distribution and storage [1].

In this paper, we propose a trace and monitoring system for agricultural products' yields and distribution based on WSNs. Without RFID technologies, we efficiently trace agricultural products and also measure the yields of agricultural products. If humidity, temperature, and other sensors are added to our sensor nodes, the applicable area of our system can be extended to quality oriented applications. We propose an overall architecture of our proposed system and describe the implementation details of the proposed system.

This paper is organized as follows. In Section 2, we explain the existing WSNs and RFID technologies and various applications of PA based on WSNs and RFIDs. In Section 3, we explain the physical and logical architecture of the proposed system with some diagrams. In Section 4, we describe the implementation details of our proposed system. Finally, in Section 5, we conclude our paper.

2. Related Work

2.1. Wireless Sensor Networks. A wireless sensor network is a system that consists of radio frequency (RF) transceivers,

sensors, microcontrollers, and power sources. Wireless sensor networks have some properties, such as self-organizing, self-configuring, self-diagnosing, and self-healing properties. These properties solve problems or enable applications that traditional technologies could not address [3–5].

An obvious advantage of wireless transmission is a significant reduction and simplification in deployment. Wireless sensor nodes allow impossible sensor applications, such as monitoring dangerous, hazardous, unwired, or remote areas and locations. This technology provides nearly unlimited installation flexibility for sensor nodes and increased network robustness. Furthermore, wireless technology reduces maintenance complexity and costs. Wireless sensor networks allow faster deployment and installation of various types of sensors because many of these networks provide self-organizing, self-configuring, self-diagnosing, and self-healing capabilities to the sensor nodes [3–5].

Wireless sensor node technology allows MEMS (micro-electro-mechanical systems) sensors to be integrated with a sensor node with extremely low cost, small size, and low power requirement. MEMS pressure sensors, temperature sensors, humidity sensors, and various capacitive sensors for proximity, position, velocity, acceleration, and vibration measurements have been integrated to wireless sensor nodes [7]. These MEMS sensors enable quality oriented applications.

2.2. Radio Frequency Identification (RFID). RFID systems are comprised of three main components such as the tag or transponder, the reader or transceiver that reads and writes data to a transponder, and the computer containing database and information management software. RFID tags can be active, passive, or semipassive. Passive and semipassive RFIDs send their data by reflection or modulation of the electromagnetic field that was emitted by the reader. The typical reading range is between 10 cm and 3 m [1, 8].

The major drawback of RFID is the limited reading range of about 20 cm, compared to other 13.56 MHz passive RFID tags (ISO15693). Sensor tags with UHF interfaces extend the reading range to a few meters and allow for automated readout during the unloading of the transport, but their signals cannot penetrate metals or liquids. Accessing passive tags during transport in a packed container is far beyond technical feasibility.

2.3. PA Applications. According to [3], there are various PA applications that have been studied and addressed. Particularly, we briefly describe RFID-based traceability systems of [3] in the following. Inexpensive, disposable RFID biosensor tags were studied. These tags were used for history checking and contamination and inventory control on food products. The biosensor tags were based on an acoustic wave platform and used antigen-antibody reaction to detect bacteria. Also, some researchers introduced the potential of RFID tags for smart packaging, automatic checkout, smart appliances, smart recycling, and marketing [3].

An on-the-road monitoring system for animals during transportation was developed. The system included sensors installed in the animal compartment to identify the animals and to monitor the air-quality, vibration, and animal behaviors. A GPS provided the location of the vehicle. A data transfer unit regularly sent data to a service center via the GSM network. It was reported that the system greatly improved animal welfare during handling and transportation [3].

A field data acquisition system was developed to collect data for crop management and spatial-variability studies. The system consisted of a data collection vehicle, a manager vehicle, and data acquisition and control systems on farm machines. The system collected data of soil water availability, soil compaction, soil fertility, biomass yield, leaf area index, leaf temperature, leaf chlorophyll content, plant water status, local climate data, insect-disease-weed infestation, grain yield, and so forth. The data collection vehicle retrieved data from farm machines via a WLAN and analyzed, stored, and transmitted the data to the manager vehicle wirelessly [3].

A greenhouse monitoring and control system was developed to collect outdoor and indoor climate data in Portugal. Several solar-powered data acquisition stations (SPWAS) were installed indoor and outdoor to measure and monitor the climate data. RF links were established among multiple SPWASs and a base station, which was used to control the SPWASs and to store the data [3].

3. Architecture of the Proposed System

The architecture of our proposed yields and distribution trace system for agricultural products is in Figure 1. We assume that workflow is as follows. First, picked agricultural products are stored in farms' local cold warehouses for a time. Picked agricultural products are packed in boxes that sensors are attached to. Agricultural products stored in farms are transported to an APC (agricultural products processing center). The APC sorts and repacks agricultural products transported from farms and stores to its cold warehouse. The agricultural products are shipped whenever wholesalers require products.

Each farm has boxes that are equipped with sensor nodes. A sensor node has some functionalities such as temperature sensors, humid sensors, gyro sensors, infrared sensors, CPU, and wireless communication functions. CPU and wireless communication functions are mandatory. Other functions may be optionally applied to a sensor node. Also, a communication hub is placed in each farm's local cold warehouse. The communication hub gathers data from sensor nodes in boxes and processes sensor data. Also, transport vehicles that deliver agricultural product boxes may have one or more communication hubs. Communication hubs send gathered data to a central server with their location data. The server stores sent data from each communication hub to a local database.

Sensor nodes of the proposed system are inactive status when they are not used. Infrared sensor and gyro sensor of a sensor node detect whether boxes are used or not. The sensor node is activated when agricultural products are put into a box and the box is moved. A communication hub sends gathered data from sensor nodes and location data of the communication hub. The location data is used to detect and trace the boxes that are equipped with sensor nodes.

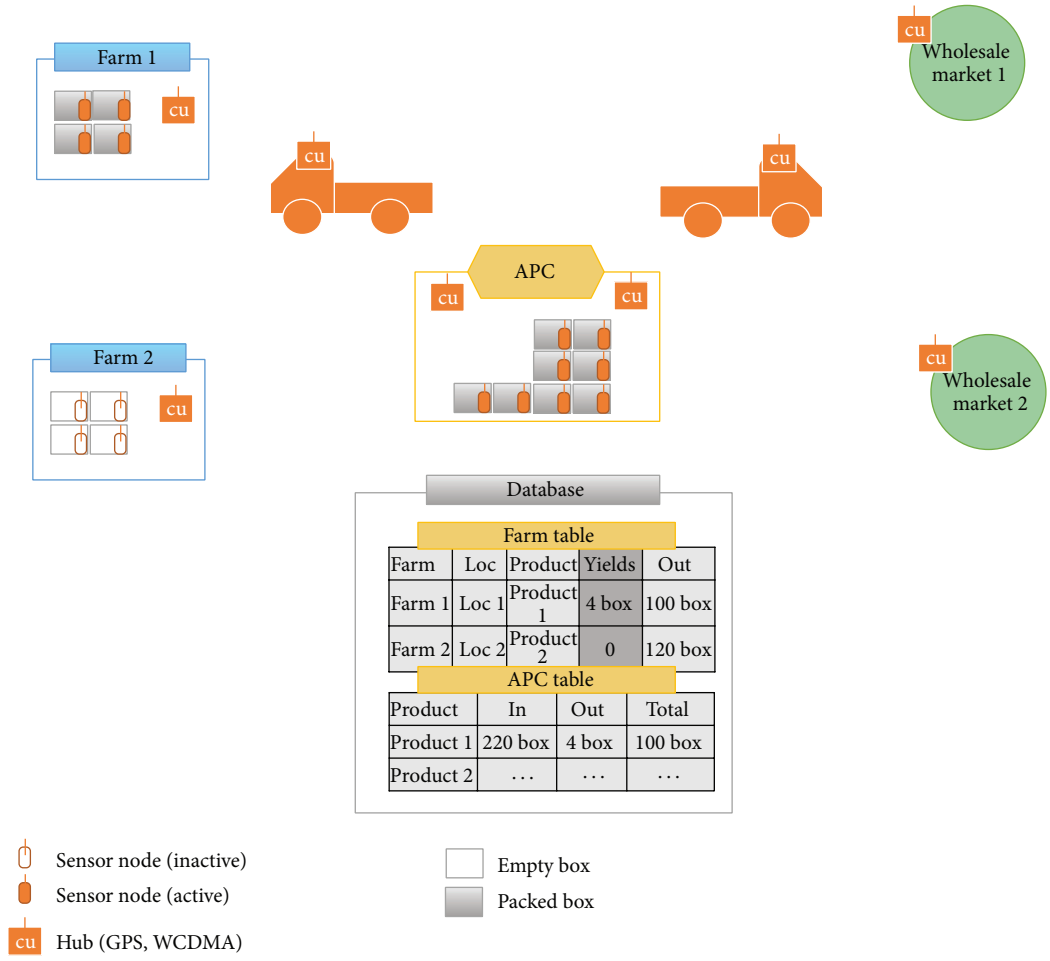


Figure 1: The overall architecture of the proposed system.

(a) Sensor nodes (b) Box with a sensor node (c) Hub

Figure 2: Deployed sensor nodes and a communication hub.

4. Implementation of the Proposed System

In this section, we describe the implementation details of our proposed system. We focus on software components rather than hardware components. Mandatory requirements of hardware components like sensor nodes and communication units are shown in Table 1. Figure 2 shows sensor nodes, a hub, and box equipped with a sensor node. A sensor node is installed in a box as in Figure 2(b) to detect whether the box is filled with agricultural products.

Figure 3 shows the database schema for the proposed system. Temperature_humidity and CO2 tables are used when a sensor node has humidity, temperature, and CO2 sensors. Hub table is used to register hubs. Each hub has cell phone number for its WCDMA module. Hub table stores the phone number and the ID of a hub. Location table stores the

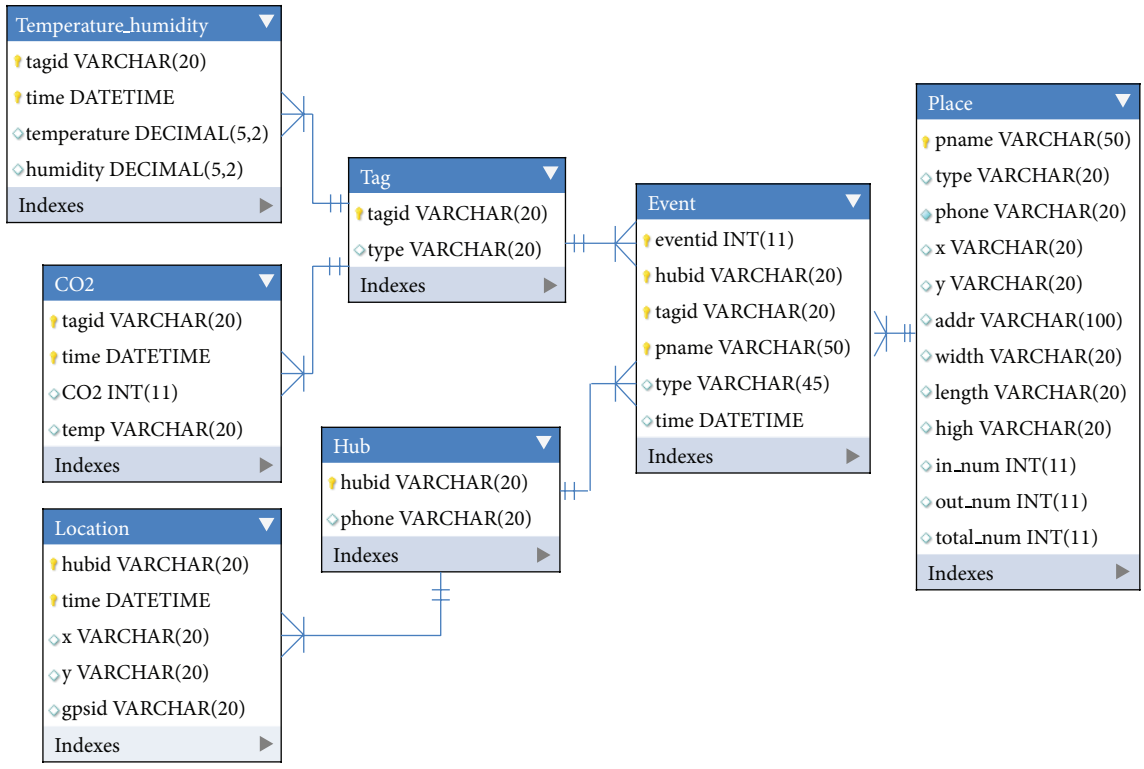


Figure 3: Database schema for the proposed system.

Table 1: Required specifications of hardware components.

Hardware components	Required specifications
Hub	GPS module
	RF module, 424 MHz
	WCDMA module
	Axis accelerometer
	Flash memory module
	RTC
	CPU (main, RF)
	SRAM
Sensor node	RF module, 424 MHz
	Infrared sensor
	Gyro sensor
	CPU

Table 2: Definition of events.

Event ID	Event
0	Sensor activated; a box is packed with agricultural products
1	Agricultural products box has moved into a place
2	Agricultural products box has moved out of a place
3	Agricultural products box is moving

locations of hubs. Each hub periodically sends its location with gathered sensor data. Tag table is used to register sensor nodes.

Place table registers the location of farms, APCs, and wholesalers. Our system traces and monitors the distribution path and yields of agricultural products by using this table and locations of hubs. Event table stores detected events by our event detection engine. We define four events as shown in Table 2. The event detection engine uses those tables and input data stream from hubs to detect these events. Event 1 occurs when a box containing agricultural products has moved into a registered place. Event 2 occurs when the box has moved out of a registered place. Finally, event 3 means that the box is moving from a place to another place. Detected events of a sensor node are stored in event table.

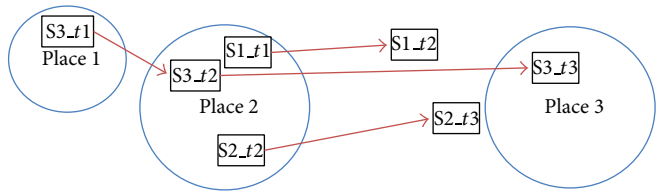
Algorithm 1 shows the algorithm of our event detection engine. This algorithm works whenever a central server receives data from a hub. First, it finds a place where a sensor exists by querying which geo_fence contains the location of the sensor. Geo_fence is given by users, and it is used to compensate the GPS error. Then, the algorithm checks the previous event of the sensor and decides the event of the sensor as in Figure 3. If previous event is null, it means that the sensor is activated at the place, and event 0 is assigned. If previous event is 1, and the current place is null, the event of the sensor node is 3 which means that the sensor node is moving. If the previous event is 3, and the current place is not null, the event of the sensor node is 1. Also, if the previous event is 2, and the current place is null, the sensor node starts moving out of a place.

```
//detect current place
Sql = select*, min(dist)
      from (select sqrt(pow(A.x-loc.latitude, 2) + pow(A.y-loc.longitude,2)) dist, A.*
            from place A) B
      where B.dist < geo_fence;
Place = Excute_query(Sql);

//check the previous event of a sensor and detect event
Sql = select* from event where tagid = 'sensorid' order by time desc limit 1;
Prev_event = Excute_query(sql);

if (Place == null)
{
   if (Prev_event == null)
      insert event 0;
   else if (Prev_event.event == 1)
      insert event 2, Prev_event.pname;
}
else
{
   if (Prev_event.event == 0 && Prev_event == null)
      insert event 1, Place.pname;
   else if (Prev_event.event == 1 && Prev_event.pname != Place.pname)
   {
      insert event 2, Prev_event.pname;
      insert event 3, Prev_event.pname;
      insert event 1, Place.pname;
   }
   else if (Prev_event.event == 2)
   {
      insert event 3, Prev_event.pname;
      insert event 1, Place.pname;
   }
}
```

Algorithm 1: Algorithm for detecting events.



Sensor ID \ Time	$t1$	$t2$	$t3$
S1	1 (p2)	2, 3 (null)	1 (p3)
S2	0 (null)	1 (p2)	2 (null)
S3	1 (p1)	2, 1 (p2)	2, 1 (p3)

Event 0 (none)

Event 1 (in)

Event 2 (out)

Event 3 (end and in)

Figure 4: Example of detecting events.

Figure 4 shows an example of event detection algorithm. There are three places and transport vehicles with CUs. Three boxes with sensor nodes are moved between places by transport vehicles. At time $t1$, sensor 3 (S3) sends data through the hub of place 1. Since the previous event of S3 is null and the current place is not null, we decide that event 1 occurs. At time $t2$, S3 sends its data through the hub of place 2. The previous event of S3 is 1, and the current place is place 2 so we decide the events are 2 and 1. It means that the sensor node was moved out of place 1 and moved in place 2. At $t2$, sensor 1 (S1) sends its data through the hub of a transport vehicle. The previous of S1 is 1, and the current place is null, so the events

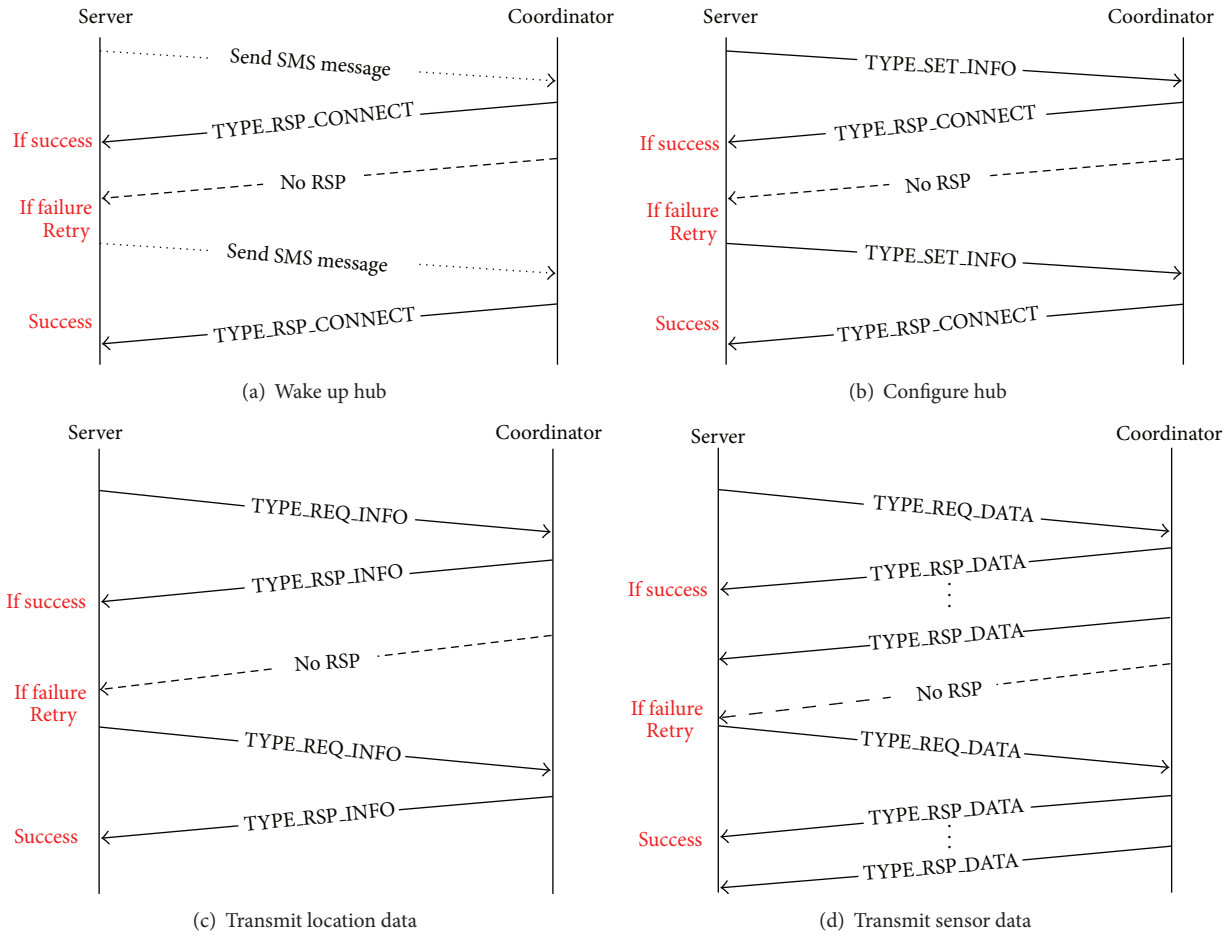


Figure 5: Communication protocol between hub and server.

of S1 are 2 and 3 which means that the sensor node has been moved out and on the road.

Figure 5 shows our communication protocol between a server and hub (coordinator in this figure). A sleeping hub is woken up by the server. The server sends an SMS (short message service) message to the hub to wake up. Then, the hub sends a connection request to the server. After the connection between the server and the hub is established, the server sends configuration information to the hub such as transmission period. The hub starts to send the location data and gathered sensor data to the server periodically.

5. Conclusion

In this paper, we proposed a system to trace and monitor agricultural products' yields and distribution channel with WSN techniques. Also, we implemented the proposed system. The implemented system consists of hardware components such as sensor nodes and hub and software components such as communication protocol between hub and server and event detection engine. We defined events to trace and monitor agricultural products. The event detection engine detects events by using the current location of each sensor node, previous event, and registered places. If the sensors such as temperature, humidity, or others are added to our sensor nodes, the applicable area can be extended.

Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2A10042015) and Technology Development Program for "Agriculture and Forestry" or "Food" or "Fisheries," Ministry for Food, Agriculture, Forestry and Fisheries, Republic of Korea.

References

[1] K. Sugahara, "Traceability system for agricultural products based on RFID and mobile technology," in *Computer and Computing Technologies in Agriculture II*, vol. 3, pp. 2293–2301, Springer, New York, NY, USA, 2009.

[2] J. A. López Riquelme, F. Soto, J. Suardíaz, P. Sánchez, A. Iborra, and J. A. Vera, "Wireless sensor networks for precision horticulture in Southern Spain," *Computers and Electronics in Agriculture*, vol. 68, no. 1, pp. 25–35, 2009.

[3] N. Wang, N. Zhang, and M. Wang, "Wireless sensors in agriculture and food industry—recent development and future perspective," *Computers and Electronics in Agriculture*, vol. 50, no. 1, pp. 1–14, 2006.

[4] R. Beckwith, D. Teibel, and P. Bowen, "Report from the field: results from an agricultural wireless sensor network," in *Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks (LCN '04)*, pp. 471–478, Tampa, Fla, USA, November 2004.

[5] A. Baggio, "Wireless sensor networks in precision agriculture," in *Proceedings of the ACM Workshop on Real-World Wireless Sensor Networks (REALWSN '05)*, Stockholm, Sweden, June 2005.

[6] L. Ruiz-Garcia, L. Lunadei, P. Barreiro, and I. Robla, "A review of wireless sensor technologies and applications in agriculture and food industry: state of the art and current trends," *Sensors*, vol. 9, no. 6, pp. 4728–4750, 2009.

[7] http://www.crossbow.com/.

[8] R. Jedermann, L. Ruiz-Garcia, and W. Lang, "Spatial temperature profiling by semi-passive RFID loggers for perishable food transportation," *Computers and Electronics in Agriculture*, vol. 65, no. 2, pp. 145–154, 2009.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 876914, 10 pages
http://dx.doi.org/10.1155/2014/876914

Research Article

Wireless Monitoring of Household Electrical Power Meter Using Embedded RFID with Wireless Sensor Network Platform

Wasana Boonsong and Widad Ismail

Auto-ID Laboratory, School of Electrical and Electronic Engineering, University of Science Malaysia (USM), Engineering Campus, 14300 Nibong Tebal, Penang, Malaysia

Correspondence should be addressed to Widad Ismail; eewidad@usm.my

Received 12 December 2013; Revised 20 May 2014; Accepted 21 May 2014; Published 18 June 2014

Academic Editor: Thomas Wook Choi

Tracking and monitoring system using radio frequency identification (RFID) have gained a lot of improvements especially for applications that need automation with reduction in human intervention and become more interesting nowadays with the increasing market demand for internet of things (IoT) technologies. The objective of this study is to improve the machine-to-machine (M2M) communication using active RFID with wireless sensor networks (WSNs) with heterogeneous data transfer (regardless of power meter type) for monitoring and identification of household electrical consumption. M2M is a popular technology and has become a part of our daily life. WSN through ZigBee technology is applied to monitor data and to read the load consumption from the electrical household power meter by embedding a tag module into the meter. In-house built-in tag was embedded into an electrical power meter with a power management circuit communicating to a reader at an RF signal of 2.45 GHz based on the star network topology with the sleep mode function. The experimental results indicated that the electrical power meters with embedded tags successfully worked wirelessly with acceptable mean value of electrical consumption difference in Watts more than those without tags (standalone meter), which is in the range of 2.44 W to 2.5 W based on 10-hour measurements.

1. Introduction

New technologies have become very useful and important in our daily life, and they were made easier and more manageable with their low cost, as well as their ability to reduce manpower requirement. Thus, obtaining higher system reliability is an important objective. Therefore, a growing interest in sensor application has created the need for protocols and algorithms in large-scale self-organizing ad hoc network that consists of hundreds or thousands of nodes. Hence, in the past decades, WSNs have been the subject of considerable research because of their potential for civilian and military applications and their applicability to M2M networks. M2M networks do not only consist of sensors. WSNs play an important role in M2M communication as they are the key components. Therefore, sensor networks are sometimes referred to as M2M networks [1, 2].

The RFID technology plays an important role in any kind of applications because of its ability to identify and track information. RFID tags are the most important components in this technology [3]. RFID is an automatic identification (ID) method whereby ID data are stored in electronic devices called RFID tags (or transponders). These data are retrieved by RFID readers (or interrogators) using RFs [4]. RFID mainly consists of a tag and a reader, and these two devices are wirelessly connected. The signal sent out by the tag is read by the reader when both of the devices are within the prescribed distance range. The read signal is then sent for data utilization [5]. RFID comes generally with three types of RFID tags. The active RFID tags contain a battery and can transmit signal autonomously. The passive RFID tags do not have a battery and require an external source to start signal transmission. The battery-assisted passive RFID tags require an external source to power them up but have significant higher forward link capability that provides longer range [3].

RFID technology has been used for ID and monitoring in many applications, such as in the Universiti Kebangsaan Malaysia (UKM) campus bus ID and monitoring, which uses RFID and geographical information system (GIS) [5]. RFID is used to monitor and identify the location of a certain bus,

which is presented as a map aided by the GIS software. RFID has also been used in monitoring enterprise activities stream using RFID tags [6] and in an accurate location tracking based on active RFID for health and safety monitoring [7].

In this paper, the wireless data monitoring of electrical power meter in a household by using the embedded active RFID tag module with WSN platform is employed. Active RFID by ZigBee protocol is used to monitor the value and to identify the electrical power meters. The ZigBee modules embedded into the electrical power meter act like wireless sensors that monitor the electricity consumption value of electrical power meter and transmit the data value with the RF signal to the portable reader. The relayed signal is applied to facilitate some daily life processes, save time, and reduce cost and error in information system that can be committed by humans. This work is the first attempt to combine and embed RFID and WSN technology into a single platform with household power meter and improve the heterogeneous functionality in data transfer regardless of power meter type. The scope of this paper is limited to the proposed hardware communication reliability compared to the standalone system communication.

The aims of this research are as follows: to develop an M2M communication system using RFID system concept for household power meter monitoring, to study the performance of the proposed household electrical power meters with embedded active RFID tags based on WSN platform in comparison to standalone design, and to analyze the power consumption of the embedded power meter with RFID tags under the same load for characterization.

Sequentially, this paper is organized as follows. In Section 2, the research methodology is divided into four parts, namely, (Section 2.1) the principle of electrical power meter; (Section 2.2) the embedded active RFID tag architecture; (Section 2.3) the portable reader architecture; and (Section 2.4) the proposed WSN. Section 3 describes the experimental results and presents the discussion. Finally, the conclusions are summarized in Section 4.

2. Methodology

In this study, the wireless monitoring of electrical power meter using embedded RFID tag with WSN platform is proposed. This research used an active RFID tag embedded into the power meter to monitor the value on real-time function and to read the electrical power meter using the star network topology based on the sleeping-mode function. The end part of the RFID tag sends the data to the reader tag, which can communicate with the reader several meters away. The main principle of the study is presented in the details of the following sections.

2.1. Principle of Electrical Power Meter. In this section, the basic theory of the electrical power meter is presented. Electrical power meter is an energy provided by a source to a load in a unit time. Let us consider an alternating voltage $v(t)$ with amplitude V_o applied to a load and current $i(t)$ circulating through the load with an amplitude I_o and phase difference φ between them. The instantaneous power is given by the following equation [8]:

$$P = \frac{1}{2} \cdot V_o \cdot I_o \cdot \cos\varphi - \frac{1}{2} \cdot V_o \cdot I_o \cdot \cos\left(2 \cdot \omega \cdot t + \varphi\right). \quad (1)$$

Active power is the energy used by the load to perform work in a unit of time. This power is the one actually used by the circuits and the loads, and it is the average of the instantaneous power:

$$P = \frac{1}{T} \int_0^T p\,(t) \cdot dt = V_o \cdot I_o \cdot \cos\varphi. \quad (2)$$

Reactive power is used in the generation of the electric and magnetic fields. It is expressed as

$$Q = V_o \cdot I_o \cdot \sin\varphi. \quad (3)$$

Apparent power, S, is the combined power value that is obtained by allowing for the different values of current and voltage or the quadratic sum of the active and reactive powers as defined above in (2) and (3). Thus, S can be found as follows:

$$S = V_o \cdot I_o = \sqrt{P^2 + Q^2}. \quad (4)$$

A power meter is an instrument that measures the electric power or the supply rate of electrical energy (in Watts) in any given circuit. An instrument that measures the electrical energy in Watt-hour (electricity meter or energy analyzer) is essentially a Wattmeter that accumulates or averages the power consumption. Such instruments can measure and display many parameters such as the following: voltages, current, apparent instantaneous power, actual power, power factor, energy consumption over a period of time, and cost of electricity consumed [9].

2.2. Embedded Active RFID Tag Architecture. Embedded system has become a centrally important element in a wide variety of applications, ranging from hand-held devices to household appliances and RFID tags [10]. Tags need power to perform computations. These tags can obtain power from a battery or from electromagnetic waves emitted by readers that induce an electric current in the tags. The power requirements of a tag depend on several factors, including the operating distance between the tag and the reader, the frequency being used, and the functionality of the tag. In general, the more complex the function supported by the tags is, the larger it is for the power requirement. For example, tags that support cryptography or authentication require more energy than those that are limited to transmitting an identifier [11].

The active RFID tag can function as a sensor. On the other hand, the components that compose a wireless sensor node are similar to those that compose the active RFID tag. A sensor node is equipped with an onboard battery and transmits sensing information to a sensor network router or coordinator, whereas the active RFID tag transmits ID information to the active RFID reader using the same components.

If we consider a tag's ID as one type of sensing information, the concept of the sensor is extended to involve the active tags [12]. Therefore, in this section, the application of an active RFID tag embedded in the power meter to monitor the value in real time is presented. The active RFID tag contains some types of power source [13]. It enables a greater communication range and can be applied to metal objects. It also allows easy addition of sensing modules [14, 15].

In this part, the proposed embedded active RFID tag module architecture is presented. The proposed embedded active RFID tag module is designed to support the heterogeneous electrical household power meter. Smart RFID tag modules are capable of measuring instantaneous voltage and current of the electrical circuit which it is connected to. The RFID tag module draws its power supply from the electrical power meter, which contains a power management circuit faction that transforms the voltage source of active RFID power supply of 3.3 V. The embedded device contains a microcontroller as well as a software programme that determines the behavior of the active RFID tag, and communication is created between devices. In this work, ZigBee is applied to act as a wireless sensor to monitor the value of the power consumption from the electrical household power meter and send information to the reader as shown in Figure 1.

The details of RFID tag module are shown in Figure 2 illustrating the elements of embedded module which consists of microcontroller unit, current sensor, voltage sensor, real-time clock, memory, display unit, and RF transceiver.

In the proposed study, the embedded RFID module is designed to minimize the power consumption. The functions of each part are described as follows.

Current Sensor. It is a sensor to detect the electrical current consumption from the electrical power meter.

Voltage Sensor. It is a sensor to detect the voltage signal to the microcontroller, which uses the basic of voltage divider circuit.

Microprocessor. It is necessary to have some processors in the circuit to communicate with the electrical power meter through the ZigBee interface and retrieve the necessary information.

RF Transceiver. Wireless ZigBee Pro Series 2 module is used for wireless communication between the smart embedded RFID tag module and the reader which are available in the utility company/office.

Display Module. The LCD module shows the data value as the process.

Real-Time Clock. The function of this clock is to generate the integration period for the embedded module.

Memory. It is a unit to record and store data such as current, voltage, time, date, and month.

The principle of data monitoring for embedded RFID tag module is shown in the flowchart in Figure 3.

2.3. Portable Reader Architecture. RFID is a communication system between two nonequivalent nodes: the RFID reader and the RFID tag. The RFID reader is stationary, large, and expensive and has a direct current (DC) supply. On the other hand, the RFID tag has almost the opposite characteristics [19].

All readers have an RF subsystem interface to communicate with tags. Most of them have a second interface to communicate with the enterprise subsystem. The enterprise subsystem interface supports transfer of RFID data from the reader to enterprise subsystem's computer for processing and analysis. In most cases, the enterprise subsystem interface is used for remote management of the readers. The interface may be a wired (e.g., Ethernet) or a wireless (e.g., Wi-Fi or satellite) link. Many systems use Simple Network Management Protocol (SNMP) to monitor the readers and alert administrators of conditions that warrant attention [11].

In this section, basic principle of the RFID tag-reader communication is presented. The architecture of the active RFID reader is shown in Figure 4.

Figure 4 shows the architecture of the proposed active RFID reader, which is similar to the embedded active RFID tag module. The functions of each part are explained as follows.

Microprocessor. It controls every part of the architecture, which is connected to the RF transceiver, to receive the data signal and send it to the RS-232 USB interface to a personal computer.

RF Transceiver. ZigBee Pro Series 2 module is adopted to be as RF transceiver, to send the wake-up signal to the end device tags, and to receive the data signal from the end device tag to the microprocessor.

LCD Module. It displays the value received from the end device tag, which is embedded within the electrical power meter.

Clock. It generates the time of the database system.

Memory. It records and stores data such as current, voltage, time, date, and month.

USB Interface. A USB peripheral is also available on this portable active RFID reader which would be useful in exporting data read from the smart meter to a PC for further processing or monitoring.

2.4. The Proposed WSN. The tag-reader communication protocols are often specified in the RFID standards. The prominent international standards include the ISO/IEC 18000 series for item management and the ISO/IEC 14443 and ISO/IEC 15693 standards for contactless smart cards. The most recent EPC global Class-1 Generation-2 standard is essentially equivalent. It is inexpensive, has power savings features, and uses the low data-rate wireless star technology design which uses the LR-WPAN type. Its lower layers are based on the IEEE 802.15.4 LR-WPAN standard. The

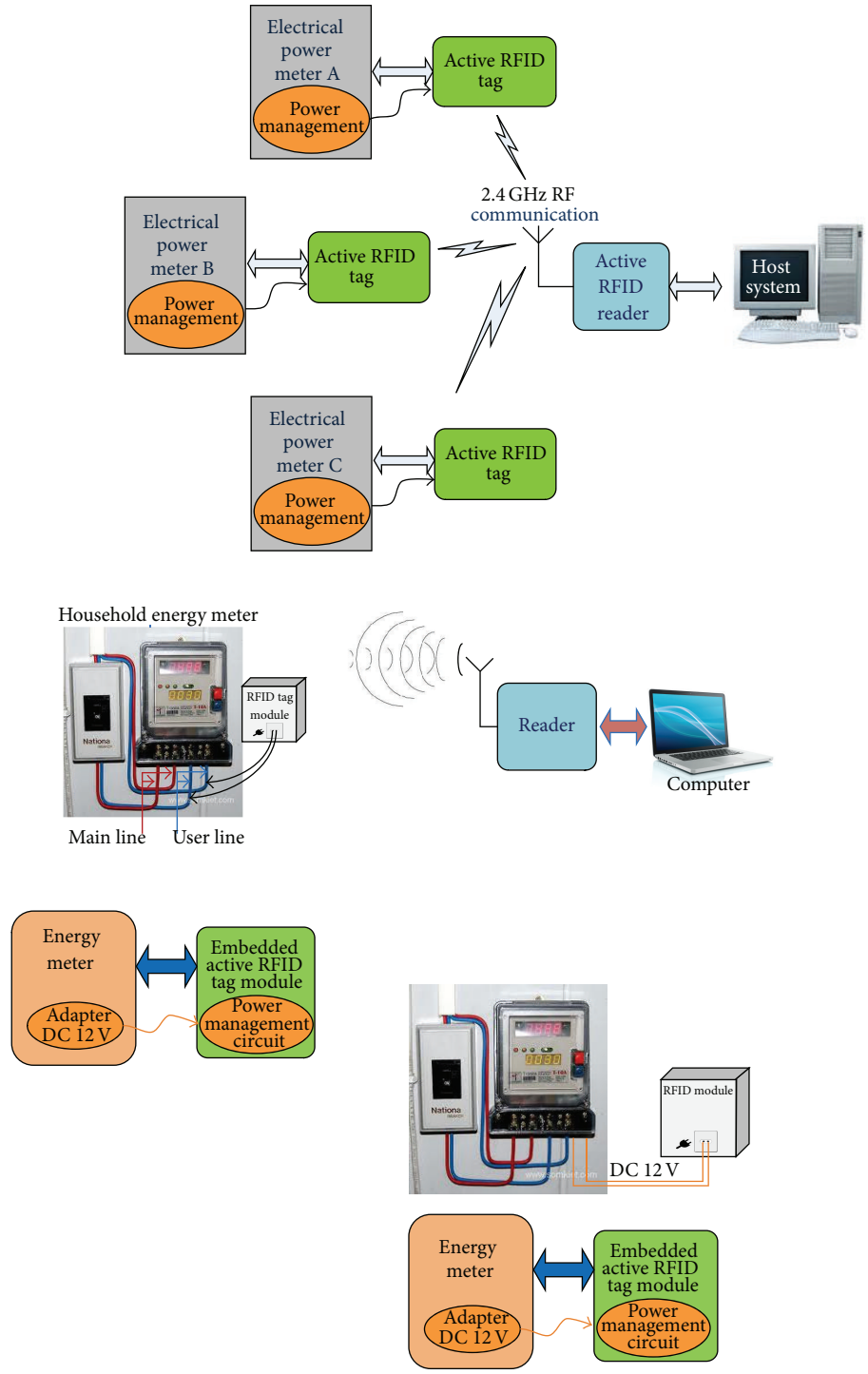


Figure 1: Electrical power meter with embedded RFID tag module.

electrical power meter with embedded active RFID tag communicates with the tag reader at 2.45 GHz to support the wireless network communication by developing a fully automatic and embedded system to monitor the value and ID.

An ideal wireless sensor should have the following properties: being smart and scalable, possibility to be incorporated into a network, having very low power consumption and programmable software, being capable of fast data acquisition, reliable, accurate over a long period, and economical to acquire and install, and requiring no maintenance [19]. Selecting the optimum sensor and wireless communications link requires the knowledge of the application and the problem definition. Battery life, sensor update rates, and size are all major design considerations.

Table 1: Comparison of three electrical power meters without and with embedded RFID tags.

Number of hours	Electrical power meter without tag			Electrical power meter embedded with tag			Power difference, D (W)					
							Power meter					
	Meter A load = 100 W	Meter B load = 200 W	Meter C load = 300 W	Meter A load = 100 W	Meter B load = 200 W	Meter C load = 300 W	A $\lvert D\rvert$	B $\lvert D\rvert$	C $\lvert D\rvert$	A % D	B % D	C % D
1	100.00	200.00	300.00	102.45	202.45	302.45	2.45	2.45	2.45	2.39	1.21	0.81
2	200.00	400.00	600.00	202.46	402.45	602.45	2.46	2.45	2.45	1.22	0.61	0.41
3	300.00	600.00	900.00	302.45	602.44	902.45	2.45	2.44	2.45	0.81	0.41	0.27
4	400.00	800.00	1200.00	402.45	802.45	1202.46	2.45	2.45	2.46	0.61	0.31	0.20
5	500.00	1000.00	1500.00	502.44	1002.45	1502.45	2.44	2.45	2.45	0.49	0.24	0.16
6	600.00	1200.00	1800.00	602.46	1202.46	1802.44	2.46	2.46	2.44	0.41	0.20	0.14
7	700.00	1400.00	2100.00	702.45	1402.43	2102.45	2.45	2.43	2.45	0.35	0.17	0.12
8	800.00	1600.00	2400.00	802.44	1602.45	2402.46	2.44	2.45	2.46	0.30	0.15	0.10
9	900.00	1800.00	2700.00	902.45	1802.45	2702.45	2.45	2.45	2.45	0.27	0.14	0.09
10	1000.00	2000.00	3000.00	1002.45	2002.45	3002.44	2.45	2.45	2.44	0.24	0.12	0.08
			Average				**2.45**	**2.44**	**2.45**	**0.71**	**0.36**	**0.24**

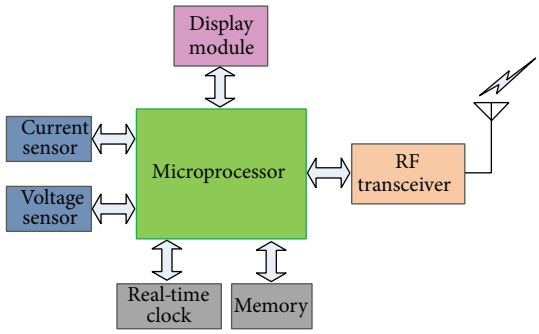


Figure 2: The proposed embedded RFID module.

In this study, the ZigBee protocol based on IEEE802.15.4 standard with star network topology in sleeping-mode function is proposed (single point-to-multipoint) where a single base station can send and/or receive a message from a number of remote nodes as illustrated by Figure 4. The remote nodes can only send or receive a message from the base station but they are not permitted to send messages to one another [20].

The embedded RFID in the proposed module for reading and sending data wirelessly is based on radio frequency 2.45 GHz. Each of the three RFID tag which are embedded inside communication modules is programmed to be an END DEVICE. The portable reader also uses the same protocol that is set to be as COORDINATOR in sleeping-mode function. The end device tags can communicate by sending packet data to the reader automatically under the system installed as shown in the structure in Figure 5.

The star network is adopted in this study because of its low latency communication between the remote node and the base station [20]. The proposed study takes three electrical power meters with embedded active RFID tag module using a two-way communication between the end device tags and the portable reader. The electrical power meters with an embedded RFID tags module can communicate wirelessly to the portable reader whenever the reader sends the wake-up signal to the end device tags. The RFID tag stays in the power-up mode until it builds a DC voltage. When the DC voltage is generated, the RFID reader sends a continuous RF signal to all the RFID tags in the neighborhood, and it starts to send the address of the RFID tags to the chosen one. The RFID reader sends the address of the signal from the RFID reader and enters the reading mode. Subsequently, the RFID tag sends the information to the RFID reader until it receives a STOP command from the reader or until the RF signal vanishes. In this study, the tag-reader communication based on the sleeping-mode function is used because sleep mode minimizes the power consumption.

3. Experimental Results and Discussion

In experiments for performance evaluation, the important parameter to determine is the difference between the electrical power meters in the household with the absence and presence of embedded RFID tags. The experimental results are presented in Table 1.

The results are divided into three parts, which consist of the following: (Section 3.1) electrical power meters without embedded RFID tags; (Section 3.2) electrical power meter with embedded RFID tags; and (Section 3.3) performance analysis. The details are explained as follows.

3.1. Electrical Power Meter without Embedded RFID Tag. In this section, the experimental results of the electrical power meter without tag for three electrical power meters, namely, A, B, and C, are presented. Each electrical power meter is tested under different loads. Electrical power meters A, B, and C had a load of 100 W, 200 W, and 300 W, respectively. The experimental results are obtained for an hour per day for ten days. The experimental results are shown in Table 1. The

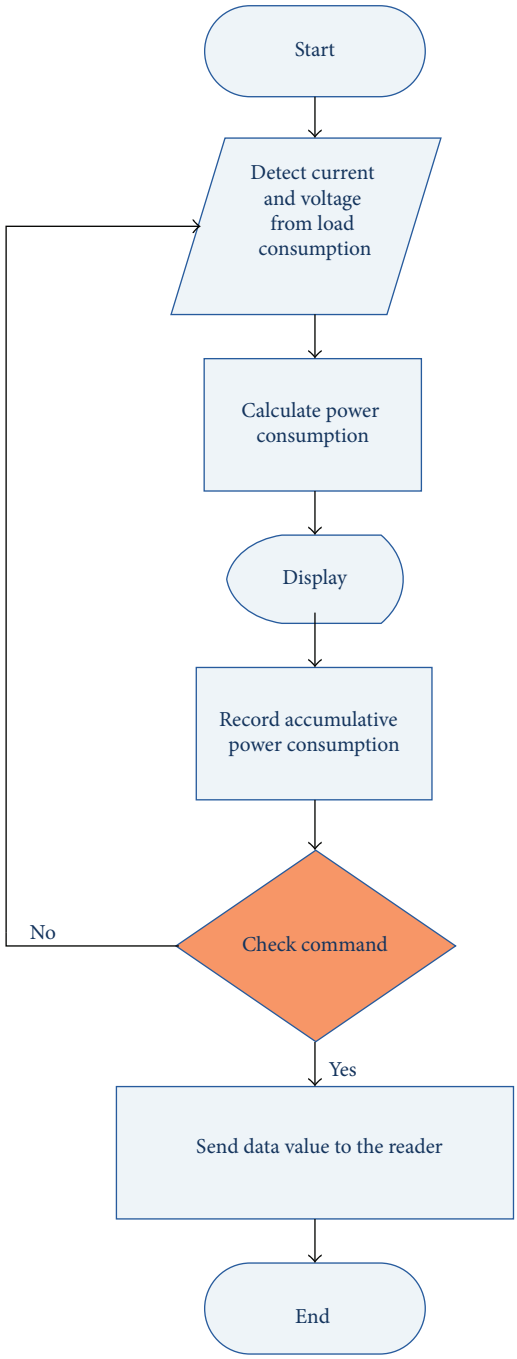


Figure 3: Flowchart of data monitoring.

obtained results indicated that the electrical power meters without RFID tag modules had no error and they can show the correct value according to the real load consumption.

3.2. Electrical Power Meter with Embedded RFID Tag. In this section, the experimental results of electrical power meters A, B, and C with embedded RFID tags under the same loads as those of the electrical power meter without embedded tags are presented. The results present that electrical consumption is higher than the one without RFID tag modules ranging from 2.44 W to 2.46 W.

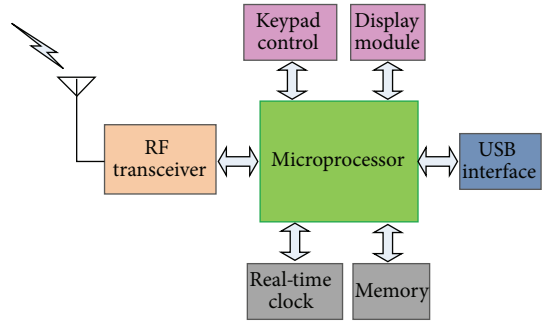


Figure 4: The proposed architecture of the active RFID reader.

3.3. Performance Analysis. This section is to compare the power difference of electrical power meters without and with embedded RFID tag modules. The power consumption differences are shown in Table 1.

Figure 6 shows the comparison graph of electrical power meters A, B, and C without and with embedded RFID tags under the loads of 100, 200, and 300 W, respectively. The graph shows that the power differences in terms of Watt for three of electrical power meters have fluctuated between 2.44 and 2.46 W with time of testing for 10 hours.

The graph in Figure 7 shows the power difference results of electrical power meters with load consumption of 100 W, 200 W, and 300 W, respectively. The graph compares three electrical power meters with difference loads, which shows that when the electrical consumption is increased, the power differences in terms of percentage (%) will gradually decrease to be more stabilized.

The electrical power consumption of monitoring RFID module is approximately 2.43 W, which can be explained by the illustration of individual block in Figure 8.

The total power consumption of the RFID tag module consists of many sections, such as current sensor, voltage sensor, microcontroller, embedded active RFID tag, display module, real-time clock, memory, and power supply, and can be calculated as follows:

$$\begin{aligned}&\text{Total power consumption (watt)}\\&\quad = 0.15\,\text{W} + 0.15\,\text{W} + 0.462\,\text{W} + 0.001\,\text{W}\\&\qquad + 0.15\,\text{W} + 0.0165\,\text{W} + 0.65\,\mu\text{W} + 1.5\,\text{W}\\&\quad = 2.43\,\text{Watts}.\end{aligned} \tag{5}$$

Therefore, the calculated total electrical consumption which is about 2.43 W verifies the measured power difference (ranging from 2.44 to 2.46 W) and it is within the acceptable limits meaning that it is at least +/− 0.07 W from theory. In addition, this also indicates that the addition of the embedded RFID tag into the electrical power meter only increases the electrical usage by 0.24% to 0.71% only for 10 hours, which is very small compared to electrical consumption of other household appliances like common 14 W for bulb and 55 W for fan per hour.

The research is to be useful, which can be applied in real environment. The billing electrical charges companies can take into consideration to deduct the additional billing prices (due to additional power consumption of embedded

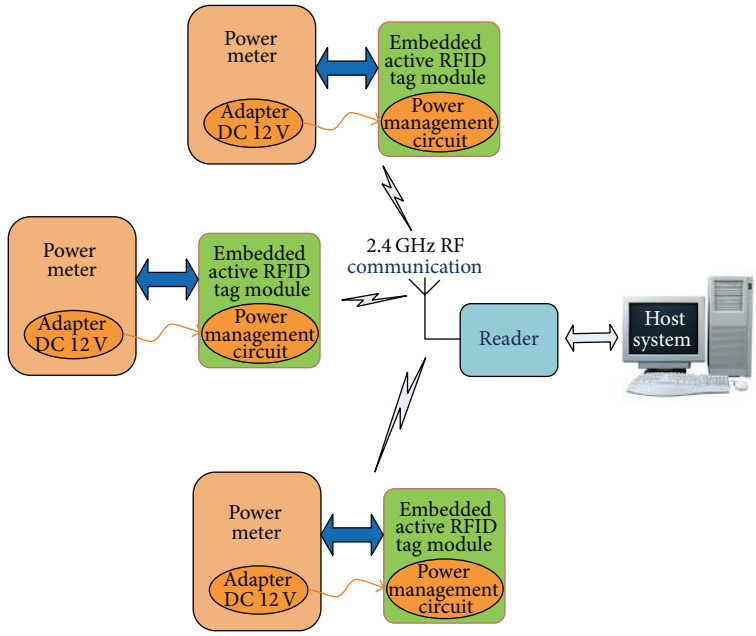


Figure 5: The proposed wireless communication system based on star network topology.

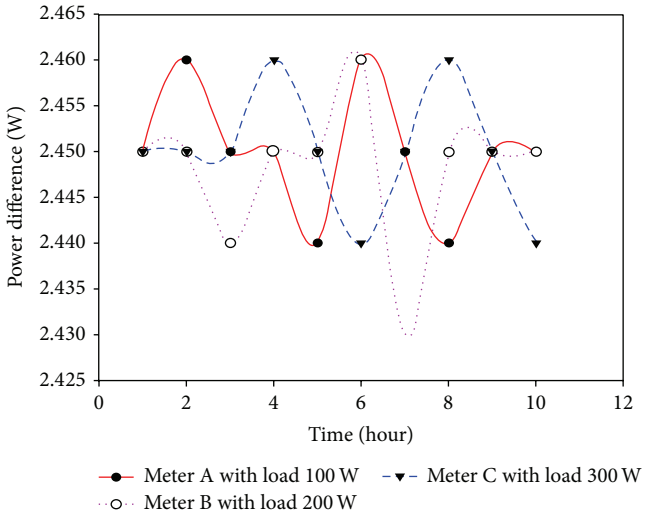


Figure 6: Magnitude of power difference for the electrical power meters A, B, and C between those without and with embedded RFID tags.

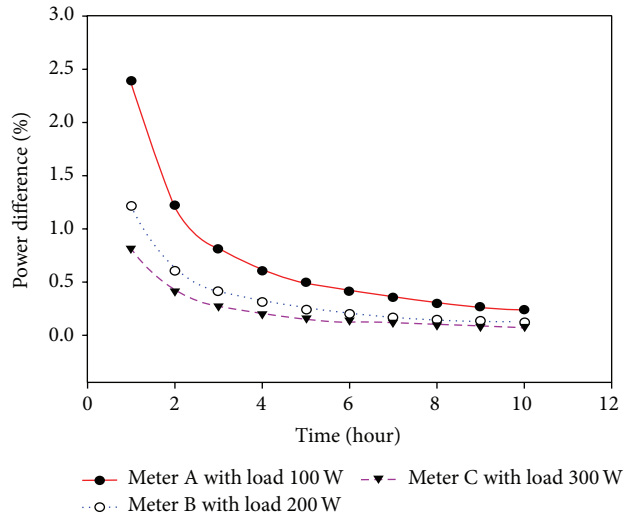


Figure 7: Power difference in terms of percentage (%) between electrical power meters A, B, and C without and with embedded RFID tags.

tag which can be negligible) since the proposed wireless monitoring reduces the operational cost for manual electrical power meter collection at the premises.

From the proved results, the RFID system based on WSN application is best suited where the business processes are dynamic or unconstrained, the movement of the tagged assets varies, and more sophisticated security, sensing, and/or data storage capabilities are required [21]. In this paper, smart embedded module is proposed to read accumulated power consumption for the month and current demand through the user interface. On the other hand, it is also effective to monitor the electrical power consumption in real time with more security of information, which sends the wireless data monitoring to the users. The WSN based on ZigBee is adopted in this system because of lower cost and lower power consumption than the Bluetooth [22] and Wi-Fi in the 802.11 standard [23]. The ZigBee can be defined as a low tier, ad hoc, terrestrial, and wireless standard; in some ways, it is similar to Bluetooth. It is promoted by ZigBee Alliance and incorporated in the IEEE 802.15.4 standard, although ZigBee has some features in addition to those of the 802.15.4 standard. Another protocol is needed because other existing

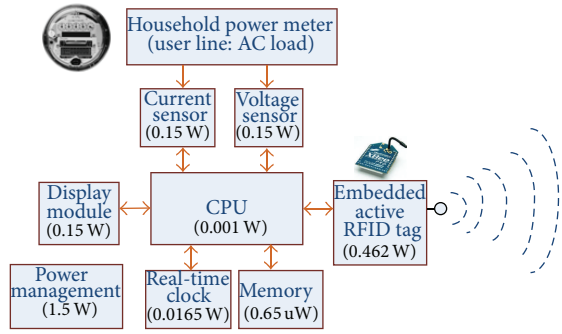


Figure 8: Total power consumption of the proposed embedded RFID tag based on individual blocks.

Table 2: Comparison of the proposed system with previous works.

Developed system	Technology used	Data transfer	Performance
Smart RFID system for Oil Palm Bio-Laboratory The smart RFID-integrated sensors for biolaboratory system are to serve as interfaces among users, which focus on critical parameters such as temperature, humidity, liquid, phase, and gas compositions (Aziz et al., 2008) [16]	Smart RFID system with high frequency 13.56 MHz, no WSN capability	RS 232/RFID	Medium data rate and read ranges up to about 1.5 meters
Hardware development for smart meter based innovations Smart meter is interfaced with developed hardware using ANSI C12.19 standard. The design communicates with the smart meter through a wireless link, which allows the user to read the accumulated consumption data from meter as well as to monitor the instantaneous demand real time (Kulatunga et al., 2012) [17]	ZigBee and high frequency at 900 MHz, not RFID based	900 MHz high-frequency radio communication	The accumulated consumption of the smart meter was read through the communication module interface based on advanced meter infrastructure (AMI) and sent to the utility company
Perpetual and low-cost power meter for monitoring residential and industrial appliances A wireless current sensor node (WCSN) for measuring the current consumption with low-power, 32 microcontroller for data processing, and a wireless transceiver to send data via the IEEE 802.15.4 standard protocol (Porcarelli et al., 2013) [18]	A 2.4 GHz IEEE 802.15.4 compliant radio transceiver, not RFID based	Not specified	Minimum error of WCSN is 1.6%
Proposed work Wireless monitoring of household electrical power meter using embedded RFID with wireless sensor network platform Data monitoring based on RFID communication system consists of embedded RFID module and portable reader which sends data to PC using USB peripheral based on star network topology	RFID system based on IEEE802.15.4 protocol + WSN based on star topology + universal household power meter	WSN via star network topology	(1) Applicable for universal power meter which provides the convenience and flexibility for the users (2) Improvement on system security which prepares the data backup, in case of surge or brownout of electricity with integrated memory unit and real-time clock (3) Improvement on M2M wireless communication in monitoring of household electrical power meter with a maximum of 90 m for indoor environment and 1600 m for outdoor environment

short-range protocols such as the 802.11 and 802.15 use too much power and are too complex (and, thus, too expensive) to be embedded in virtually every kind of device imaginable. The data rate of ZigBee is 250 kbps at 2.4 GHz. ZigBee allows small, low-cost devices to transmit quickly the small amount of data such as temperature readings for thermostats, on/off requests for light switches, or keystrokes for a wireless keyboard. ZigBee devices [24], typically the battery-powered ones, can actually transmit information up to 90 m for indoor environment and 1600 m for outdoor environment because each device within the listening distance passes the message along to any other device within the range, and only the intended device acts upon the message. Meanwhile, Bluetooth is mainly used for short-range communications, for example, from a laptop to a nearby printer or from a cell phone to a wireless headset. Its normal range of operation is 10 m (at 1 mW transmit power), and this range can be increased to 100 m by increasing the transmit power to 100 mW [22]. Moreover, to highlight the advantages and benefits of the proposed RFID system based on WSN platform, a comparison is done with the previous works as shown in Table 2. This shows that the combination of RFID, WSN, and power meter technology heterogeneously into a single platform is feasible. This paper provides guidelines for design requirements employing the proposed mentioned technologies.

4. Conclusions

In this paper, the wireless data monitoring of electrical power meters using embedded RFID module with WSN platform is proposed with consideration given to fulfill the requirement for universality usage of household power meter types. Smart measuring of embedded RFID module provides data to the utility office and can be more frequent. These data may contain a considerable amount of confidential information based on RFID system with WSN platform application, which is useful for monitoring the value and ID of the electrical power meters to facilitate some daily life processes, saving time, and reduce the operating cost because of the reduction in the manpower requirement and error in information system that can be omitted through humans, thus improving the M2M communication and providing higher reliability on the communication system because the current development will focus on local control strategies. This study can be guidelines to the electrical power utility company and consumers for alternatives in electrical consumption billings in the future.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank the USM RU (Research University) Grant Secretariat and Malaysia Ministry of Higher Education (LRGS Fund) for sponsoring the development of the in-house built-in RFID devices.

References

[1] R. Kwok, "Use of (1) sensors and (2) radio frequency ID (RFID) for the national children's study," Tech. Rep., RTI International, 2004.

[2] N. Tekbiyik and E. Uysal-Biyikoglu, "Energy efficient wireless unicast routing alternatives for machine-to-machine networks," *Journal of Network and Computer Applications*, vol. 34, no. 5, pp. 1587–1614, 2011.

[3] M. Zhang and P. Li, "RFID application strategy in agri-food supply chain based on safety and benefit analysis," *Physics Procedia*, vol. 25, pp. 636–642, 2012.

[4] W.-J. Yoon, S.-H. Chung, and S.-J. Lee, "Implementation and performance evaluation of an active RFID system for fast tag collection," *Computer Communications*, vol. 31, no. 17, pp. 4107–4116, 2008.

[5] A. Mahani Mustapha, M. A. Hannan, A. Hussain, and H. Basri, "UKM campus bus identification and monitoring using RFID and GIS," in *Proceedings of the IEEE Student Conference on Research and Development (SCOReD '09)*, pp. 101–104, November 2009.

[6] E. Masciari, "SMART: stream monitoring enterprise activities by RFID tags," *Information Sciences*, vol. 195, pp. 25–44, 2012.

[7] C.-S. Cheng, H. H. Chang, Y.-T. Chen et al., "Accurate location tracking based on active RFID for health and safety monitoring," in *Proceedings of the 3rd International Conference on Bioinformatics and Biomedical Engineering (ICBBE '09)*, pp. 1–4, Beijing, China, June 2009.

[8] D. Ramírez Muñoz, D. Moro Pérez, J. Sánchez Moreno, S. Casans Berga, and E. Castro Montero, "Design and experimental verification of a smart sensor to measure the energy and power consumption in a one-phase AC line," *Measurement*, vol. 42, no. 3, pp. 412–419, 2009.

[9] Federal Energy Regulatory Commission, "Assessment of demand & response advanced metering," Staff Report, 2008.

[10] C. Paar and A. Weimerskirch, "Embedded security in a pervasive world," *Information Security Technical Report*, vol. 12, no. 3, pp. 155–161, 2007.

[11] T. Karygiannis, B. Eydt, G. Barber, L. Bunn, and T. Phillip, *Guidelines for Securing Radio Frequency Identification (RFID) Systems*, National Institute of Standards and Technology, Technology Administration U.S. Department of Commerce, 2007.

[12] H. Yang, L. Yang, and S.-H. Yang, "Hybrid Zigbee RFID sensor network for humanitarian logistics centre management," *Journal of Network and Computer Applications*, vol. 34, no. 3, pp. 938–948, 2011.

[13] E. Masciari, "SMART: stream monitoring enterprise activities by RFID Tags," *Information Sciences*, vol. 195, pp. 25–44, 2012.

[14] K. Finkenzeller, *RFID Handbook*, John Wiley & Sons, 2nd edition, 2003.

[15] K. Michael and L. McCathie, "The pros and cons of RFID in supply chain management," in *Proceedings of the International Conference on Mobile Business (ICMB '05)*, pp. 623–629, 2005.

[16] N. H. A. Aziz, A. J. Alias, A. T. Hashim et al., "Smart RFID system for oil palm bio-laboratory," in *Proceedings of the IEEE International RF and Microwave Conference (RFM '08)*, pp. 247–251, December 2008.

[17] N. A. Kulatunga, S. Navaratne, J. Dole, C. Liyanagedera, and T. Martin, "Hardware development for Smart Meter based innovations," in *Proceedings of the IEEE Innovative Smart Grid Technologies—Asia (ISGT Asia '12)*, pp. 1–5, Tianjin, China, May 2012.

[18] D. Porcarelli, D. Balsamo, D. Brunelli, and G. Paci, "Perpetual and low-cost power meter for monitoring residential and industrial appliances," in *Design, Automation & Test in Europe Conference & Exhibition (DATE '13)*, pp. 1155–11160, Grenoble, France, March 2013.

[19] C. Paar and A. Weimerskirch, "Embedded security in a pervasive world," *Information Security Technical Report*, vol. 12, no. 3, pp. 155–161, 2007.

[20] C. Townsend and S. Arms, *Wireless Sensor Networks*, MicroStrain, 2005.

[21] C. Z. Zulkifli, W. Ismail, and M. G. Rahman, "Implementation of embedded active RFID," *Electronics World*, vol. 117, no. 1908, pp. 28–36, 2011.

[22] H. Lehpamer, *RFID Design Principles*, Artech House, London, UK, 2nd edition, 2012.

[23] C. Smith and J. Meyer, *3G Wireless With WiMAX and Wi-Fi: 802.16 and 802.11:802.11*, Professional Engineering, McGraw-Hill, 2004.

[24] IEEE 802.15.4 RF Modules by Digi International. XBee/XBee-Pro RF Modules -802.15.4 - v1.xEx, 2009.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 181760, 7 pages
http://dx.doi.org/10.1155/2014/181760

Research Article

Dynamic Access Control Model for Security Client Services in Smart Grid

Sang-Soo Yeo,[1] Si-Jung Kim,[2] and Do-Eun Cho[3]

[1] *Division of Convergence Computer & Media, Mokwon University, Daejeon 302-729, Republic of Korea*
[2] *College of General Education, Hannam University, Daejeon 306-791, Republic of Korea*
[3] *Innovation Center for Engineering Education, Mokwon University, Daejeon 302-729, Republic of Korea*

Correspondence should be addressed to Do-Eun Cho; decho@mokwon.ac.kr

Received 3 January 2014; Accepted 15 May 2014; Published 18 June 2014

Academic Editor: Jongsung Kim

In the next-generation intelligent power grid, known as the smart grid, various objects can access systems in several network environments, and, accordingly, access control security becomes critical. Thus, to provide users with secure services in the smart grid, a new access control security model is needed. This paper proposes a dynamic access model for secure user services in the smart grid environment. The proposed model analyzes the user's various access contexts and chooses an appropriate context type among the predefined context types. And then it applies the context-based user security policy to allow the user's access to services dynamically. Therefore, it provides stronger security services by permitting context information-applied security services and flexible access control in various network environments. It is expected that this study will be used to solve important access control issues when establishing the smart grid.

1. Introduction

Recently, with the development of renewable energy, interest in efficient energy management has increased. The next-generation intelligent power grid, smart grid, unlike existing provider-centered one-way energy operation systems, is a two-way operating system in which consumers participate in energy use and operation [1–3]. In addition, smart grid technology, interlocked with home networks, allows the control of information appliances no matter when or where the user is. However, to securely provide these and other services, it is important to secure home network security, protect private information, and restrict access to home devices.

For example, if a user requesting services receives authorization with his or her user information and receives the same services regardless of location, time, or access device, a severe security issue may occur if the user's authority is stolen or the device is lost.

In the infrastructure of existing services, access rights to a resource are granted only after the execution of a user authorization phase. In contrast, for the resources or services executed in various network environments like the smart grid, the user's accessibility should change depending on the ambient context information.

Currently, the smart grid has a variety of security vulnerabilities. In particular, security measures for various network environments and corresponding new services are lacking [4–6].

Therefore, to provide secure user services according to ambient context, it is also necessary to provide dynamic, context-adaptive services. To this end, various sensors and computers should collect and effectively share environment information, find the contexts of the user and macroenvironment, and provide appropriate services for them.

Context refers to the information that characterizes and defines the state of entities in the real world. Context awareness is a technical method of interacting with this context and characterizing a human's current context [7]. Context awareness computing application technology includes methods based on the correlation between the user and services. Its implementation and application technologies can be devised in various forms [8–10]. Recently, security-related

areas that consider context awareness have received attention, and various studies on several security models have been actively carried out [11–13]. It is necessary to study a new access control security model, applying this to the smart grid environment to provide security services according to time, space, and user context.

Access control is a well-known security mechanism to give access permission or denial message to an access request according to the predefined access policies, in which the system monitors and controls who can access the specific data and also what they can do onto that data. Unlike general access control, dynamic access control uses place (where), time (when), and purpose (why) according to context information as the conditions for access permission [14].

This paper proposes a dynamic access control model to provide users with secure services in the smart grid environment. The model proposed in this paper analyzes ambient context information according to context type and, accordingly, dynamically manages service authority for the users. In addition, the security levels are applied differently depending on the users' context information, even to users with service authority. Thus, the proposed model provides context-adaptive security services and flexible access controls in the various network environments of the smart grid. In addition, it inspects ambient conditions in real time, dynamically grants access right differently depending on them, and provides more powerful security services than existing resource security services.

This study is organized as follows. Section 2 examines two related models from among the existing access control security models. Section 3 proposes a model limiting dynamic access rights depending on changes in the ambient context information. Section 4 describes an application of the proposed model to the smart grid environment. Lastly, Section 5 concludes the paper and proposes future research topics.

2. Related Work

This section examines two related areas of research: role-based access control (RBAC) and the context awareness access control (CAAC) model. Additionally, it describes the necessity of a security model providing dynamic security services according to context in the smart grid environment.

2.1. Role-Based Access Control (RBAC). The RBAC model is a technology that does not give access rights to system resources by user or predefined access control rules, but by the group to which the user belongs, that is, the user's role [15–19]. This model classifies rights not to the user unit but to the user's role. In addition, the roles have a hierarchical structure, and through the structure ancestor's access rights can be inherited to its descendants easily, and hence, access rights can be more effectively managed in this hierarchical structure. Figure 1 shows the characteristics of a general role-based control model. Sandhu et al. proposed role-based access control by classifying models into the following four kinds [20].

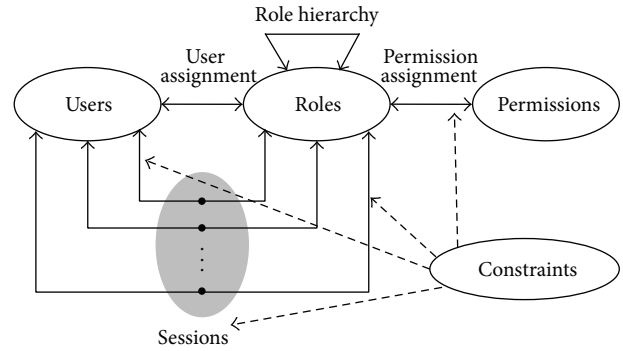


Figure 1: Role-based access control model.

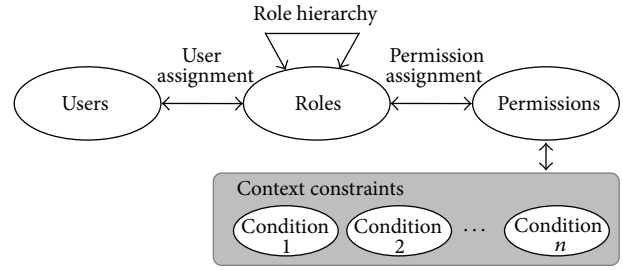


Figure 2: xoRBAC model.

(i) $RBAC_0$: role-based access control basic model.

(ii) $RBAC_1$: basic model with the addition of role hierarchy, an inheritance concept.

(iii) $RBAC_2$: basic model with the addition of context constraint conditions.

(iv) $RBAC_3$: model integrating $RBAC_1$ and $RBAC_2$

The RBAC model classifies access rights by role and grants the responsibilities and rights of the individual user accordingly. Thus, by providing security services through the access control of the user for resources, it maximizes the efficiency of security management. However, the RBAC model cannot perform dynamic access control based on contexts such as time and space.

Neumann and Strembeck proposed the xoRBAC model that limits role-based access control to use context information in access control decisions [21]. A context constraint describes the condition that satisfies a context information attribute to permit a specific calculation by limiting role-based access control. Access control is limited by comparing the real value of the context information attributes with predefined conditions. The context constraints are formed of a tuple of context attribute, function, and condition. The decision regarding rights is made according to the rights of a specific subject or role. Thus, as in Figure 2, the context is a condition that limits the granting of rights. Rights relate to several context constraints, and, when all context constraints have true values, access is permitted.

2.2. Context Aware Access Control (CAAC). The CAAC model is an access control technology that uses context

awareness by dynamically measuring the current context of the user's access demand and evaluating it. In other words, it is a model that decides rights by adding context information to the existing RBAC model [22]. The CAAC model access control methods are given by the following four definitions [23, 24].

(i) Context type (CT): an element of context constraint that defines context information.

(1) Context set (CS): a set of all context types in an application

$$\mathrm{CS} = \{\mathrm{CT}_1, \mathrm{CT}_2 \cdots \mathrm{CT}_n\}, \quad 1 \le i \le n. \tag{1}$$

(2) Context implementation (CI): a function of context types defined by

$$\mathrm{CI} : \mathrm{CT}_1 \times \mathrm{CT}_2 \times \cdots \times \mathrm{CT}_n \longrightarrow \mathrm{CT}, \quad n \ge 0. \tag{2}$$

(ii) Context constraint (CC): the definition of context information using CT in a formulaic form.

(1) CC := $\mathrm{Clause}_1 \cup \mathrm{Clause}_2 \cup \cdots \cup \mathrm{Clause}_n$.
(2) Clause := $\mathrm{Condition}_1 \cap \mathrm{Condition}_2 \cap \cdots \cap \mathrm{Condition}_i$.
(3) Condition := <CT><OP><VALUE>.
(4) CT is an element of CS.
(5) OP is a logical operator in set $\{>, \ge, <, \le, \ne, =\}$.
(6) VALUE is a specific value of CT.

(iii) Authorization policy (AP): a policy providing access rights (R) to resources for the user or role (P) according to context constraint (C).

(1) An authorization policy as a triple, AP = (R, P, C), where
(2) R is the subject in this policy, which could be a user or a role,
(3) P is the permission in this policy, which is defined as a pair $< M, O >$, where M is an operation mode defined in {READ, APPEND, DELETE, UPDATE} and O is a data object or data type,
(4) C is a context constraint in this policy.

(iv) Data access (DA): an attempt to access specific information using the user's role and context information.

(1) DA = (U, P, RC) where
(2) U is a user in the user set that issues this data access,
(3) P is the permission this user wants to acquire,
(4) Runtime context (RC) is a set of values for every context type in the context set.

DA (U, P, RC) is granted if there exists an AP (R, P', C),

(1) $U \in R$ and

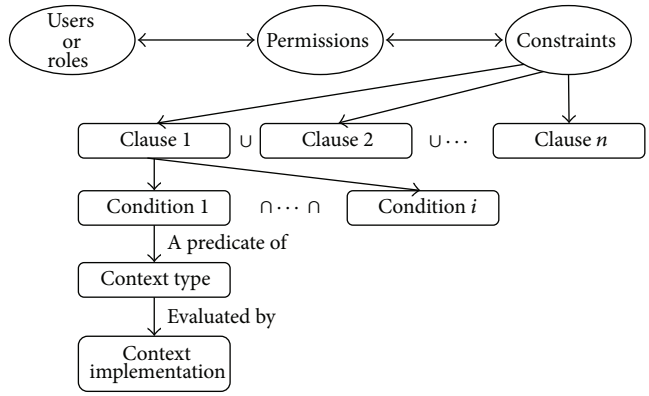


Figure 3: CAAC authorization policy structure.

(2) $P = P'$ and
(3) C is evaluated as true under RC.

Figure 3 shows the CAAC model's policy decision process. This is similar to the RBAC model, but, by adding a context constraint element, it decides whether to grant rights according to a context value.

2.3. Demands for Dynamic Access Controlling. Recently published access control schemes have various characteristics for providing flexibility and security. Generalized temporal role-based access control (GTRBAC) can give access rights under the time constraints and the periodical configuration [25, 26], Privacy role-based access control (PRABAC) can provide stronger privacy policy to the access time [27, 28]. And Geo-RBAC model is considering the user location information before giving the access permission [29]. Nonetheless, the smart grid is not a simple architecture and it has many kinds of context and circumstances, the existing access control models can cover all aspects of the smart grid environment.

In the smart grid environment, each model is distributed and arranged for cooperative performance and various objects may access the systems. This access control management for each object is very closely related to security issues. Thus, for efficient access control of smart grid, it is necessary to systematically analyze security requirements and a policy to solve them is needed. In addition, to apply access control policy more efficiently and consistently, an access control mechanism is necessary.

3. Dynamic Access Control Modeling (DACM)

This section describes a dynamic access control model that can be applied to the smart grid environment for secure user services. Access control in the smart grid environment should consider scalability, accessibility for many users, and distinctiveness of two-way communication through a variety of equipment. The existing RBAC model controls access based on many roles in various contexts, so it has been difficult to prevent dynamic access. Therefore, this proposed model provides a dynamic access control for each context-based CAAC model.

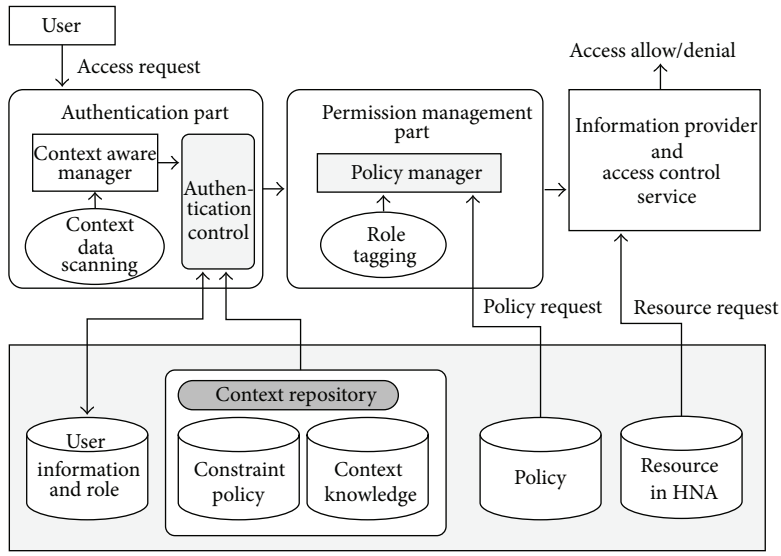


Figure 4: Overview of dynamic access control model.

3.1. Proposed DACM Structure. The proposed DACM model collects context information in the user authentication phase via the context awareness manager. In each domain, it performs mapping and follows the policy of the relevant DB. In the access right decision phase, it maps each domain, classifies the task, tags the roles, and applies the role in context.

Context information type for dynamic access control in the proposed model is defined as follows. Context information is obtained by the user by scanning for environment information at the time of services access.

Source context data types are listed below:

(i) Regular_Role_ID,
(ii) Password,
(iii) Time_Stamp_Value,
(iv) Location Type,
(v) Location Value,
(vi) Access_Device,
(vii) Access_Format,
(viii) Access_Network_type,
(ix) Task_Attribute.

Figure 4 shows the structure of the proposed dynamic access control model. The model performs access control by real time context information as follows.

Step 1. The user attempts access to a certain data entity using an already issued authentication key. For providing dynamic access, Access Ticket is issued with UserID, Session Key for runtime context (RC), and Share Key. Session Key can be created using the user's current runtime context and its mapped information. Share Key is calculated from the user access key value.

$$\text{Access Ticket} = \{\text{UserID} \parallel \text{Session\_Key for RC} \parallel \text{Share Key}\}$$

Step 2. The use of public services does not require an access license. For services for which different access license levels have been assigned, the user asks for an access right to the management server and waits for a response.

Step 3. The system applies the metadata value entered in the basic role to grant a new Role_ID.

$$\text{Role\_ID} = \{\text{Regular\_Role\_ID}, \text{Time\_Stamp\_Value}, \text{Location Type}, \text{Access\_Device or Access\_Format}, \text{Task\_Attribute}\}$$

Step 4. The granted Role_ID forms a tuple in which the metadata are stored, and role tagging is carried out.

Step 5. The tuple relevant to the tag-granted Role_ID satisfies the condition specified in the relevant domain and the user receives the access license. In this case, even for already licensed Role_IDs, the DB domain is decided by a Role_new_ID and regenerated according to the metadata value generated in the access and authorization is checked.

Tuple format is

(i) Role_ID,
(ii) Service Name,
(iii) Data,
(iv) Access Permission Check Value,
(v) Rule Domain

3.2. Policy Management for a Secure Client Service. The constraint conditions for the policy management of the proposed model are as follows. The regular ID and the Domian ID of the user are verified, and then the user password is also verified. If the two values are correctly verified, access is granted.

After the access is made, the data necessary for the user's context awareness is scanned. The input values

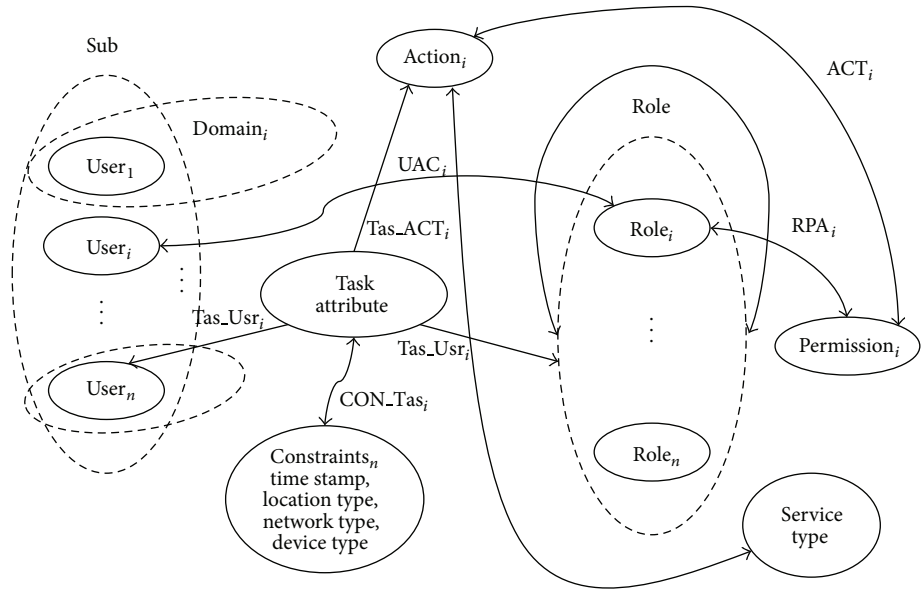


Figure 5: Overview of policy for access control model.

include Time_Stamp_Value, Location Type, Location Value, Access_Device, Access_Format, Access_Network_Type, and Task_Attribute.

First, Time_Stamp_Value checks if the time for the user access is authorized. Regarding the user's Location Type, the user's access network type is checked. Different security levels can be granted by network type. The Access_Device has a limited range of available services according to device, so the access of the services and the results are decided by Access_Format. In addition, the user's basic right regarding the Task_Attribute access services is checked.

Definition of user's context data constraint:

(i) Role_ID = if (DB_Domain_1 ‖ DB_Domain_2 ‖ DB_Domain_n)

(ii) Password = if ((Password$_{\text{input}}$ = Trust_Password) = TRUE)

(iii) Time Constraint = if (Time_Stamp$_{\text{input}}$ > TimeC_{low} && Time_Stamp$_{\text{input}}$ < TimeC_{high})

(iv) Location Type = Switch (case 1 (in HAN), case 2 (in LAN), case 3 (in WAN), etc.)

(v) Network Type = Switch (case 1 (use Zigbee), case 2 (use WCDMA), case 3 (use WiFi), case 4 (use Wibro), etc.)

(vi) Access Device = Switch (case 1 (use Remot Contorller), case 2 (use Cellular Phone), case 3 (use Pc & Mobile), etc.)

(vii) Access Format = if (Type$_{\text{input}}$ = (Type_1 ‖ Type_2 ‖ ⋯ Type_n))

(viii) Task_Atribute = ServiceRequest_Task_Type (Public ‖ Private ‖ Administrate)

Figure 5 shows the policy management process of the proposed model.

(i) Sub, Domain: sets subject a and domains.

(ii) Role$_i$, Permission$_i$: sets of Role, Permissions, Constraints and Actions in the ith domain for each i member of Domain.

(iii) User$_i$: a function that determines the set of users in the ith domain for each i member of Domain.

(iv) RPA$_i$: Roles X Permissions, a many-to-many role-to-permission assignment relation.

(v) UAC$_i$: a function mapping each user in the ith domain to a set of Actions.

(vi) AC T$_i$: Actions X Roles, a many-to-many Action to permission assignment relation.

4. DACM for a Secure Client Service in the Smart Grid

This section shows how to apply the suggested dynamic access control model to the smart grid environment specifically. The general existing access control model is designed properly for a single system, so some parts must be modified to handle the complexity of the smart grid. A proper access control model for the smart grid should efficiently manage many users, devices, and systems and should be able to conduct subtle control. The DACM flow suggested for the smart grid environment is shown in Figure 6.

Role A can be defined as the user with upper network access rights in the smart grid environment. If the user requests services remotely to the home or office, the user's access rights change dynamically with context.

If the user requests services, then context information collection can be conducted at the same time as user certification. The user's context information is collected by the context aware manager, and the context information follows the constraint rule of the context aware policy described in

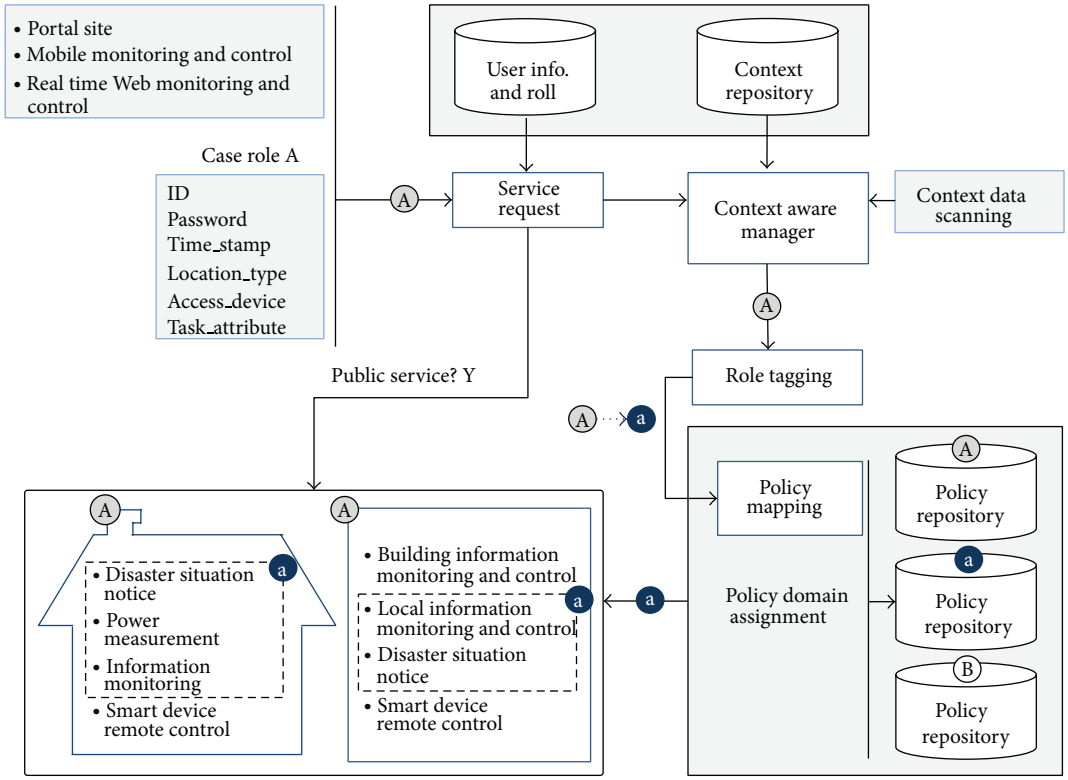


Figure 6: A control flow with the proposed DACM in smart grid.

Section 3. After that, role tagging is conducted by the policy manager, and the rights for Role A are decided. The mapping of policy DB about Role A is conducted by tagging Role A. Through this process, the relevant defined services and information regarding each role can be accessed in the policy DB.

This results in an increase of the security of services for remote access or control services by the user in the smart grid. In addition, the user with a manager role in a home or enterprise network can provide proper services in a dynamic way by setting the various roles for service access or providing information about a variety of contexts.

When the user requests remote services, the entered metadata value and the user's context data are scanned for services access. The user can access to the data entities in various roles such as a general manager or system control manager in a home network or as a service provider in the smart grid environment. General-share services can be provided through a direct policy repository. When the home network user remotely accesses the smart grid environment by a basic Role ID, the rights policy limits the access to dynamic services according to the access context information.

5. Conclusions

This paper proposes a novel dynamic access control model which provides security-enhanced data access services in the smart grid environment. The proposed model identifies each user's role and current context. The user's context can be mapped to a certain predefined context type of the proposed model, and that context type is associated with an access policy which can control the user's access privilege. The context-aware manager can manage this mapping process, collecting information about the user role, context, and requested service, and mapping the proper context type and access policy to the user. And the policy manager controls the role-tagging process for the user and applies the exact roles to the user finally. Consequently, the proposed access control model can control dynamically the user's data access permissions.

The proposed model applies different access security policies depending on context information even for the same user by judging whether to provide authority management and services dynamically according to the user's context information. This provides security enhancements for overall smart grid services and resource access. Unlike the existing power grids, in the smart grid, various access objects such as users, devices, and systems can access systems along with two-way communication, and, accordingly, issues of access control and relevant security become important.

Recently, various security models have been studied with respect to access control using context awareness, but the various services provided in a smart grid and access control in such an environment still have serious vulnerabilities. In consideration of the lack of studies on smart grid access control, it is expected that the model proposed in this study will be used to solve important access control issues when establishing the smart grid in the future.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (no. 2014R1A1A1A05008391).

References

[1] B. Vaidya, S. S. Yeo, D. Y. Choi, and S. Han, "Robust and secure routing scheme for wireless multihop network," *Personal and Ubiquitous Computing*, vol. 13, no. 7, pp. 457–469, 2009.

[2] H. Shen and Y. Cheng, "A semantic-aware context-based access control framework for mobile web Services," *Applied Mechanics and Materials*, vol. 195-196, pp. 498–503, 2012.

[3] S. ben Ayed and F. Teraoka, "Collaborative access control for multi-domain cloud computing," *IEICE Transactions on Information and Systems*, vol. 95, no. 10, pp. 2401–2414, 2012.

[4] S.-S. Yeo, D.-J. Kang, and J. H. Park, "Intelligent decision-making system with green pervasive computing for renewable energy business in electricity markets on smart grid," *Eurasip Journal on Wireless Communications and Networking*, vol. 2009, Article ID 247483, 12 pages, 2009.

[5] NIST, Smart Grid Cyber Security Strategy and Requirements, CSCTG (Cyber Security Coordination Task Group), 2009.

[6] Cisco White Paper, Security for the Smart Grid, 2009.

[7] D. E. Cho, B. S. Koh, and S. S. Yeo, "Secure D-CAS system for digital contents downloading services," *Journal of Supercomputing*, vol. 64, no. 2, pp. 477–491, 2013.

[8] M. Younas and I. Awan, "Mobility management scheme for context-aware transactions in pervasive and mobile cyberspace," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 3, pp. 1108–1115, 2013.

[9] R. Tan, J. Gu, Z. Zhong, and P. Chen, "Metadata management of context resources in context-aware middleware system," in *Web Information Systems and Mining*, vol. 7529 of *Lecture Notes in Computer Science*, pp. 350–357, Springer, Berlin, Germany, 2012.

[10] L. Zhou, N. Xiong, L. Shu, A. Vasilakos, and S. S. Yeo, "Context-aware middleware for multimedia services in heterogeneous networks," *IEEE Intelligent Systems*, vol. 25, no. 2, pp. 40–47, 2010.

[11] E. Tong, W. Niu, H. Tang, G. Li, and Z. Zhao, "Reasoning-based context-aware workflow management in wireless sensor network," in *Service-Oriented Computing - ICSOC 2011 Workshops*, vol. 7221 of *Lecture Notes in Computer Science*, pp. 270–282, Springer, Berlin, Germany, 2012.

[12] D. J. Xue and W. X. Zhang, "Design and implement of dynamic context-aware monitoring system based on OWL," *Advanced Materials Research*, vol. 532, pp. 1022–6680, 2012.

[13] M. Netter, S. Hassan, and G. Pernul, "An autonomous social web privacy infrastructure with context-aware access control," in *Trust, Privacy and Security in Digital Business*, vol. 7449 of *Lecture Notes in Computer Science*, pp. 65–78, Springer, Berlin, Germany, 2012.

[14] M. Nakamura, S. Matsuo, and S. Matsumoto, "Supporting end-user development of context-aware services in home network system," *Studies in Computational Intelligence*, vol. 443, pp. 159–170, 2013.

[15] D. F. Ferraiolo, J. A. Cugini, and D. R. Kuhn, "Role-Based Access Control (RABC): features and motivations," in *Proceedings of the 11th Annual Computer Security Applications Conferences*, pp. 241–248, 1995.

[16] M. J. Moyer and M. Ahamad, "Generalized role-based access control," in *Proceedings of the 21st IEEE International Conference on Distributed Computing Systems (ICDCS '01)*, pp. 391–398, April 2001.

[17] J. Barkley, K. Beznosov, and J. Uppal, "Supporting relationships in access control using role based access control," in *Proceedings of the 4th ACM Workshop on Role Based Access Control*, pp. 55–65, 1999.

[18] R. Sandhu, D. Ferraiolo, and R. Kuhn, "NIST model for role-based access control: towards a unified standard," in *Proceedings of the 5th ACM Workshop on Role-Based Access Control*, pp. 47–63, Berlin, Germany, July 2000.

[19] D. F. Ferraiolo, J. F. Barkley, and D. R. Kuhn, "A role based access control model and reference implementation within a corporate intranet," *ACM Transactions on Information and System Security*, vol. 2, no. 1, pp. 34–64, 1999.

[20] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, "Computer role-based access control models," *Computer*, vol. 29, no. 2, pp. 38–47, 1996.

[21] G. Neumann and M. Strembeck, "An approach to engineer and enforce context constraints in an RBAC environment," in *Proceedings of 8th ACM Symposium on Access Control Models and Technologies (SACMAT '03)*, pp. 65–79, Como, Italy, June 2003.

[22] G. Ahang and M. Parashar, "Dynamic context-aware access control for grid application," in *The 4th International Workshop on Grid computing*, pp. 101–108, 2003.

[23] D. M. Kim and J. O. Kim, "Design of emergency demand response program using analytic hierarchy process," *IEEE Transactions on Smart Grid*, vol. 3, no. 2, pp. 635–644, 2012.

[24] J. Hu and A. C. Weaver, "A dynamic, context-aware security infrastructure for distributed healthcare applications," in *Proceedings of the 1st Workshop on Pervasive Privacy Security, Privacy, and Trust*, 2004.

[25] J. B. D. Joshi, E. Bertino, and A. Ghafoor, "Hybrid role hierarchy for generalized temporal role based access control model," in *Proceedings of the 26th Annual International Computer Software and Applications Conference*, pp. 951–956, August 2002.

[26] J. B. D. Joshi, E. Bertino, U. Latif, and A. Ghafoor, "A generalized temporal role-based access control model," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 1, pp. 4–23, 2005.

[27] A. F. A. Dafa-Alla, E. H. Kim, K. H. Ryu, and Y. J. Heo, "PRBAC: An extended role based access control for privacy preserving data mining," in *Proceedings of the 4th Annual ACIS International Conference on Computer and Information Science (ICIS '05)*, pp. 68–73, July 2006.

[28] Q. Ni, E. Bertino, J. Lobo, and S. B. Calo, "Privacy-aware role-based access control," *IEEE Security and Privacy*, vol. 7, no. 4, pp. 35–43, 2009.

[29] E. Bertino, B. Catania, M. L. Damiani, and P. Perlasca, "GEO-RBAC: a spatially aware RBAC," in *Proceedings of 10th ACM Symposium on Access Control Models and Technologies (SACMAT '05)*, pp. 29–37, swe, June 2005.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 594652, 8 pages
http://dx.doi.org/10.1155/2014/594652

Research Article

Secure Model against APT in m-Connected SCADA Network

Si-Jung Kim,[1] Do-Eun Cho,[2] and Sang-Soo Yeo[3]

[1] *College of General Education, Hannam University, Daejeon 306-791, Republic of Korea*
[2] *Innovation Center for Engineering Education, Mokwon University, Daejeon 302-729, Republic of Korea*
[3] *Division of Convergence Computer & Media, Mokwon University, Daejeon 302-729, Republic of Korea*

Correspondence should be addressed to Sang-Soo Yeo; sangsooyeo@gmail.com

Received 4 January 2014; Accepted 1 April 2014; Published 17 June 2014

Academic Editor: Jongsung Kim

Supervisory control and data acquisition (SCADA) networks for the remote control and operation of various industrial infrastructures are currently being used as main metropolitan infrastructures, especially smart grid and power plants. Most of the existing SCADA networks have fortified securities because of their powerful access control based on closed and private networks. However, recent SCADA networks are frequently connected to various IT-based systems and also to other conventional networks, in order to achieve the operational convenience of SCADA systems, as well as the execution requirements of various applications. Therefore, SCADA systems have acute needs for secure countermeasures against the ordinary network vulnerabilities and for tangible preparations against ever-changing intrusion attacks such as advanced persistent threat (APT). This paper introduces the concept of m-connected SCADA networks, analyzes various security vulnerabilities on such networks, and finally proposes an integrated secure model having an APT managing module and a rule-based intrusion detection system (IDS) for internal and external network access.

1. Introduction

Currently, most of the major core infrastructures, including power supply chains, are managed and operated through the supervisory control and data acquisition (SCADA) system. Various types of metropolitan infrastructure networks consist of IT-based network systems. Consequently, cyber terrors aimed at these systems, as well as malfunction and information leakage due to virus infections and hacking, and unauthorized remote control have resulted in greatly increased damage. Large-scale plant facilities, such as power plants and dams, are gradually moving toward information systemization for the effective operation of management systems located remotely in major social infrastructure and industry fields. Therefore, concerns regarding the information security of control systems are increasing.

Thus far, national infrastructure control systems have operated based on the closed SCADA system. The SCADA system manages and controls major national infrastructures, including oil, gas equipment, and water and sewage equipment, and it is the technology that operators can use to collect data from remote infrastructure equipment, as well as transfer commands to control such equipment [1, 2].

Most countries operate SCADA systems in closed networks, and it can be said that such networks are secure from cyber attacks because the vendor's own operating system and protocols are used. However, in recent years, connections with open networks have been implemented for work effectiveness and operational convenience. Because most SCADA networks include sensor devices, network communication functions, remote monitoring facilities, data acquisition systems, they can be easily connected to wide area networks and to the public networks.

The traditional connecting method is to use exclusively private networks, but now it has been changed into the dual structures consisting of its own intranet and Internet. Consequently, much more security vulnerability appears due to the interlocking of intranet and Internet connections with various IT systems, and the possibility of critical damage caused by cyber attacks might increase. For example, a new malicious code called Stuxnet has infected some essential mechatronic devices at the Bushehr Power Plant in Iran and

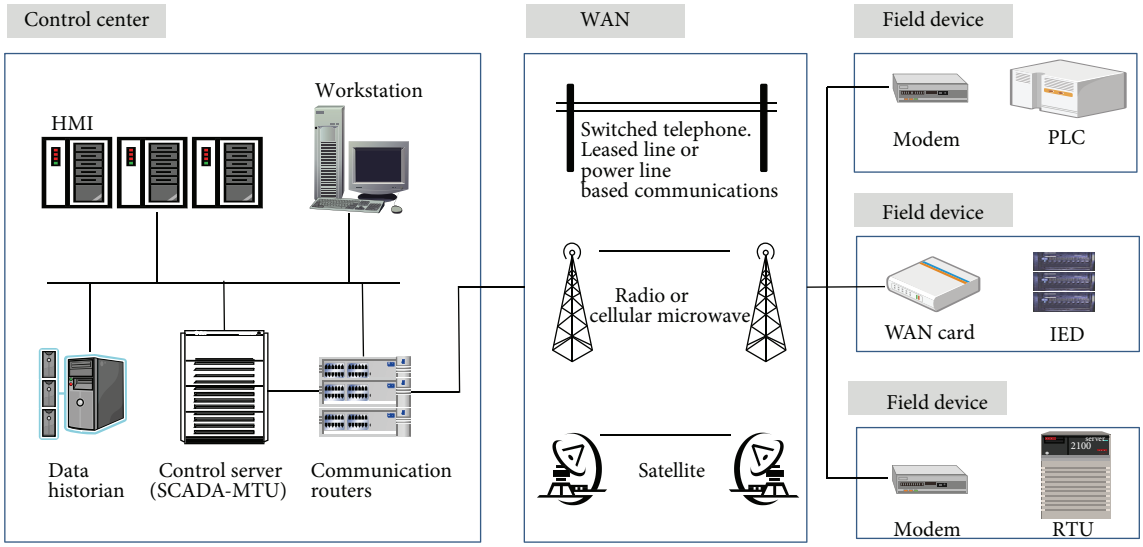


Figure 1: A typical structure of SCADA network [20].

caused them to malfunction; this type of malicious code has attracted the attention of security professionals around the world. The SCADA system which was then allowed to operate in a closed private network disconnected from the Internet for not being a victim to cyber danger now strongly requires various types of external connections along with remote maintenance and usage of mobile storage mediums such as USB flash memory. Such an environmental change means that the securities of the SCADA system, which are rooted in the characteristics of closed networks, can no longer be maintained [3].

Our proposed m-connected SCADA network is defined in this paper in order to analyze the weak points that can take place while closed SCADA jobs are performed and to present a security model against advanced persistent threat (APT) attacks.

In this paper, we will discuss security through a more enhanced SCADA network security model by presenting a security model for APT attacks in the form of anomaly detection of m-connected SCADA networks that operate in open or closed structures. In Section 2, the structure of existing SCADA systems is examined, and the existing research on its security is discussed. In Section 3, the definition of m-connected SCADA networks is discussed, and attacks against these networks are examined. In addition, Section 3 analyzes vulnerabilities of the SCADA system and responses to attacks are examined. In Section 4, a security model against vulnerabilities is proposed. Our conclusion is presented in Section 5. This paper includes our initial research result published in [3] and gives more detailed explanations and elaborations onto it.

2. Related Work

In general, the system structure of SCADA networks can be changed according to their usage and objectives. Figure 1 shows a typical structure of SCADA networks, including the SCADA control servers, human machine interfaces (HMIs), data historians, and field devices, and such structure should be well designed and organized in order to process sensing data and control commands securely [2].

(i) Human machine interface (HMI) transmits the collected information to the system operator and transfers the control command under the secure user authentication.

(ii) SCADA control server collects and analyzes the measured information transferred from the field devices and transfers the control command through the HMI to the field devices.

(iii) Field device transfers various status signals or information of the target network to the SCADA server. In addition, it transfers the command signals that are suitable to the actual target device for control by analyzing the SCADA control command transmitted from the server to the various file site devices, including the remote terminal unit (RTU), programmable logic controllers (PLC), intelligent electronic device (IED), and programmable automation controller (PAC).

2.1. SCADA Network Security Invasions. Recently, APT attacks, using very intelligent, viable, and malicious codes such as Duqu, Flame, and Gauss, which are variants of Stuxnet, occurred on SCADA networks. In addition, the appearance of the malicious code Flamer (W32.Flamer), which targets national infrastructures and can leak information after an attack, has increased concerns regarding information security [1–3].

Similarly, new attack methods, such as the polymorphism attack that occurs, namely, through the server from where other attackers automatically generate mutagenic malicious code that was obtained from the attacker's website, have

appeared and constantly take place, along with high-skilled targeted attacks, including APT attacks.

APT attack methods are performed in stages (internal intrusion, searching, collection, and leakage) through spear phishing, after a target has been elaborately checked through preinvestigation.

Approximately 100,000 personal computers (PCs) have been infected by Stuxnet all over the world; similar damage to PCs in India and the USA has also been reported. Since the advent of Stuxnet, it is impossible to assert that SCADA systems are secure from cyber attacks; closed networks are no exception.

Generally, this type of attack intrudes networks after normal PCs have been infected. When the infected PCs then connect to and interface with closed networks, the network becomes infected. Even before the advent of Stuxnet in 2010, SCADA network security invasions occurred several times [4–6].

Existing closed SCADA systems should be controlled effectively, because they are frequently accessed from outside the system through mobile storage mediums, including universal serial bus (USB) storage devices. Control systems that are operated only within closed networks can be exposed to infection threats from malicious code based on the use of mobile storage mediums, including USBs. In addition, policies are required that block connections with mobile storage mediums through USB ports, including operator PCs (human machine interface, HMI) and central servers that can regulate control equipment.

Stuxnet and other malicious codes have even been transmitted through mobile storage mediums such as USBs and have invaded facilities operated within closed networks.

Strengthening the security measures for outside personnel is also necessary. Notebooks and portable PCs that belong to outside personnel should only be used after being checked for viruses and illegal programs. Fundamentally, such equipment should be checked by clean PCs that are stored by the management organization.

User account management and authentication processes require strengthening. Major systems, including HMI PCs that operate and control equipment, central servers, and network devices should only be accessed by authorized managers. In addition, control access, including the provision of IDs as well as the registration, change, and disposal procedures performed according to authorized managers and users, should be strengthened. In some organizations, manager IDs and passwords are frequently shared among all operators, and systems are automatically logged into for convenience; however, such practices should be discontinued and avoided, without exception.

The latest security patch should be maintained through regular security updates, along with vaccine installation. Vaccines in the control system should be kept current through regular updates; pretests of software security patches should be performed to study their effects on the system, prior to the next offline update.

Other security processes that should occur are the detection of unauthorized modems and wireless LANs that might have been installed through internal or external access; if found, such modems and wireless LANs should be disconnected immediately and constantly monitored, because they could be operated while an external connection is open for remote maintenance [7–9].

2.2. m-Connected SCADA Network. As we explained above, even a closed SCADA network can be momentary online status which is defined as "*m-connected*" status, and such temporary pseudoconnections are made by portable mediums such as external flash memory, floppy disks, and CD-ROMs which are used to perform maintenance tasks including patching, upgrades, and migration [3]. So "m-connected SCADA network" is a closed, isolated, and private SCADA network which has, however, similar levels of vulnerabilities to open online SCADA networks in a long-term observation. Moreover, such m-connected status of a closed SCADA network can be formed by an official update/patch server attached in the SCADA network, and if the update/patch server is infected by malware through portable mediums, this will make a big disaster on the SCADA system.

3. Security of m-Connected SCADA Network

The SCADA network generally performs services through the proper interface, according to each network type. In addition, security vulnerabilities appear after general-purpose hardware and software begin to be used. An m-connected SCADA network that operates within both open platforms and closed platforms shows serious weak points [3].

Currently, SCADA systems based on exclusively closed protocols and their own dedicated interfaces are no longer secure but have the lack of awareness of security and authentication in the design, disposition, and operation of the SCADA network. Consequently, any belief that the SCADA network is secure because of its physical isolation is not true anymore.

In particular, some of the weak points in the managerial aspect of SCADA networks are a weak security connection, passwords shared by several people, impossibility in tracing it when an attack has occurred, and not knowing where the responsibility lies. Technical weak points include an OS, whose security is not strengthened, applications, system operation, and damage from attacks. Therefore, SCADA networks are quite complex in their security measures [10–14].

3.1. Threat of Attack to SCADA Network. Security threats from attacks to SCADA networks are described as follows [14].

(i) Threat to the Use of Platform Technology with Standard Protocol and Vulnerability. Organizations have the same exposure to vulnerability known by the use of famous operating systems from the use of exclusive systems. In addition, standard network protocols, such as transmission control protocol/Internet protocol (TCP/IP), are used for cost reduction and performance increase. The uses of these protocols and technologies do present advantages in the economic and

technological aspect but are extremely vulnerable to attacks from effective hacking tools.

(ii) Increased Access among Networks. Internal and external organizations generally connect the SCADA system to a network system in order to fulfill various objectives, including operation and information management. In such organizations, there is a system manager or technology supporting personnel responsible for monitoring the external system. These same managers or supporting personnel set up remote access channels; in addition, access among different networks is increased to collect information about the system operation. The SCADA system uses wide area networks and the Internet for the operation of remote or local devices; this structure can increase network vulnerability.

(iii) Connection of Various Access Devices. System maintenance is responsible for authorizing wireless communication in cases where remote access is permissible and related services are established. Illegal access or authentication can be tried to test access to system or to test authentication procedures. Dangers to security might not be recognized because of the complexity that exists among different networks when a given network and the SCADA network attempt to gain access, and this could result in weaknesses in the control access to the network.

3.2. Vulnerability of the m-Connected SCADA Network. This section explores various weaknesses in the m-connected SCADA networks.

(i) Structural Weak Points in the SCADA Network. The control network is secured with powerful access controls based on isolation from the commercial network. However, such security control will not be meaningful anymore after the control network is connected to the commercial network.

In addition, it has become easier to acquire information to an attacking path to the SCADA network, because information about the SCADA network structure has been revealed in the Internet. Access to separate devices has been avoidable in cases of emergency and during system maintenance in the generally closed status of closed SCADA networks. This is an exceptional access that can be expected to occur periodically, and this means that abnormal access or the possibility to be exposed to various attack methods has greatly increased.

(ii) Security Vulnerabilities for Physical Connections. Data might be exposed during data transfer through wireless connections when communication is established between remote devices and the control center. Moreover, data might be accessed through HMIs without passing through the proper authentication process during communication with the telephone network.

(iii) Security Vulnerabilities in EMS. When commercial networks and SCADA networks are used through their connections, security threats exist due to attacks to several devices. Authentication systems that use passwords for remote terminal units (RTU) and intelligent electronic devices (IED) are vulnerable to attacks because of password exposure and management carelessness.

3.3. APT Attacks. According to the recent threat analysis of cyber attacks, advanced persistent threats (APTs) attacks are the most common attacks in SCADA systems. APT attacks are also defined as advanced targeted persistent threats (ATPTs), and they are an attack type where attackers with professional technology having elaborated levels or vast resources use an attacking path. APTs mainly target large organization networks, and the damage they cause is more considerable than any other attack types [14, 15].

The goals of APT attacks are to leak information constantly by providing and expanding the internal foundations of information technology infrastructures in the general organization, to obstruct important aspects in the organization, or to later acquire its foundations. Repetitive and continuous attacks are performed for a long time, while constant threats adapt to the defending resistance and persistently maintain the necessary level of interaction to execute their objectives. The differences between intelligent, constant threats and existing attacks are as follows. First, the attacker attempts to assail continuously a specific field or organization. Second, the attacker abuses weak points until new ones are discovered, or large-scale attacks are rearranged by combining small weak points. Third, there is an incubation period that makes attacks difficult to detect. Invasions are relatively easy to detect because general security invasions tend to steal large amounts of data in a short time. Conversely, existing security systems can become incapacitated because intelligent and constant threats use a method to leak the target data over several months or years.

Figure 2 shows a general process of an APT attack. APT attacks are a type of attack that utilizes malicious code to attack large-scale networks and target specific organizations. APT attacks deliberately choose a target, and the attacking group is strategically flexible for the target. Major attacks focus on large-scale organization networks using worms that leak information for lengthy periods or provide a foundation for other invasion attacks using an evasion in the network of the target organization [16–19].

The security requirements against APT attacks are described as follows [19].

Step 1. Requirements for continuous network traffic analysis are

(i) traffic analysis of protocols, including Internet relay chats (IRC) and hypertext transfer protocols (HTTP), and traffic monitoring through secure socket layer (SSL) communication analysis;

(ii) upgrading the platform system operation file through a network vulnerability analysis;

(iii) general usage pattern analysis for the network user, namely, execution of action analysis.

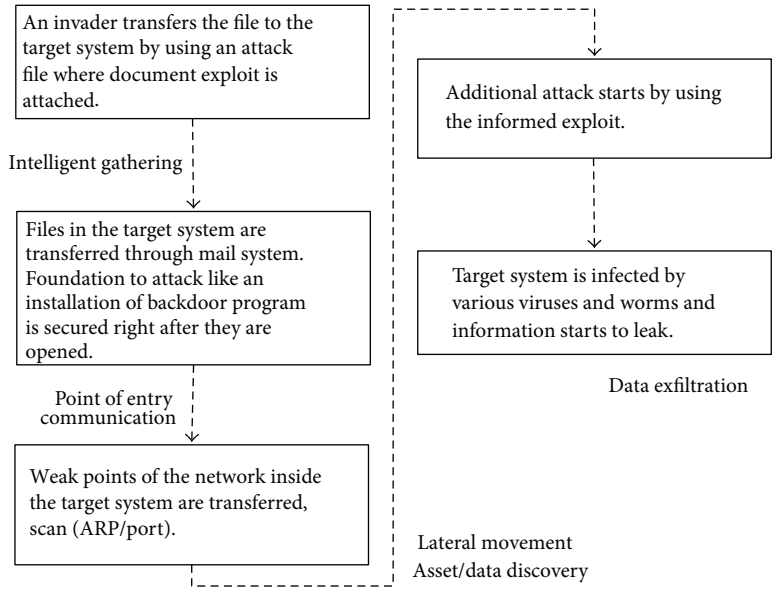


Figure 2: General process of APT attack.

Step 2. Requirements for context analysis based on network operation are

(i) analysis of known vulnerabilities through analysis of protocols based on the platform;

(ii) application program detection and analysis in the cyber space;

(iii) access log file analysis and extraction of the corresponding access data.

Step 3. Content analysis based on access includes

(i) structural weak point analysis of data file;

(ii) analysis of attached files and monitoring of the running code in case of downloads.

Similarly, security measures by phase for APT attacks taking place in the SCADA network are required. Analysis and monitoring of all network traffic should be performed, and separate handling routines should be maintained to detect malicious contents and executable code. Test processes by phase are required to determine whether an attack has occurred. Adequate intrusion detection system (IDS) rules should be applied.

This type of regular monitoring system and detection routine rule should be updated in real time. When an attack is detected, a handling process for emergency control operations should be performed. Abnormal access should be detected and blocked through a more dynamic operation in the existing system.

4. Secure Model for m-Connected SCADA Networks

Currently, SCADA separated from extranets is connected to networks for general work and to IT system networks for the efficient management and operation of information. Consequently, many types of accidents related to security can occur frequently. Most SCADA networks are set up with general and fundamental network technology, and they may be exposed to various attacks caused by vulnerabilities, just as is the case with existing IT systems. Therefore, security measures of SCADA networks require the operation of security management programs, establishment of measures according to risk evaluation of vulnerability analysis, application of secure security modules, and establishment of security policies.

4.1. Intrusion Detection Module for the Proposed Security Model. Most devices do not consider system securities because SCADA, which is a converging network that consists of various application programs and devices, is set up with security solutions that are applicable to the existing networks. In addition, currently operated servers have security vulnerabilities because access authentication is performed with a simple password in cases of remote access.

Moreover, protocols used in the SCADA network, such as distributed network protocols (DNP), intercontrol center communication protocols (ICCP), and Modbus, are becoming the target of attackers, because such protocols are not guaranteed with integrities that are important security elements.

The SCADA network requires access between a network and another network for efficient operation. Currently, the secure measure is an introduction of IDS for the most efficient access. This is to guarantee the security of transferred data by placing IDS in case of data transfer inside and outside the network. It needs to guarantee security by placing the IDS module at the access point of internal and external SCADA networks. Presently, host-based security modules are applied to the IDS internal module; application of the IDS module should be performed after security assets of the SCADA

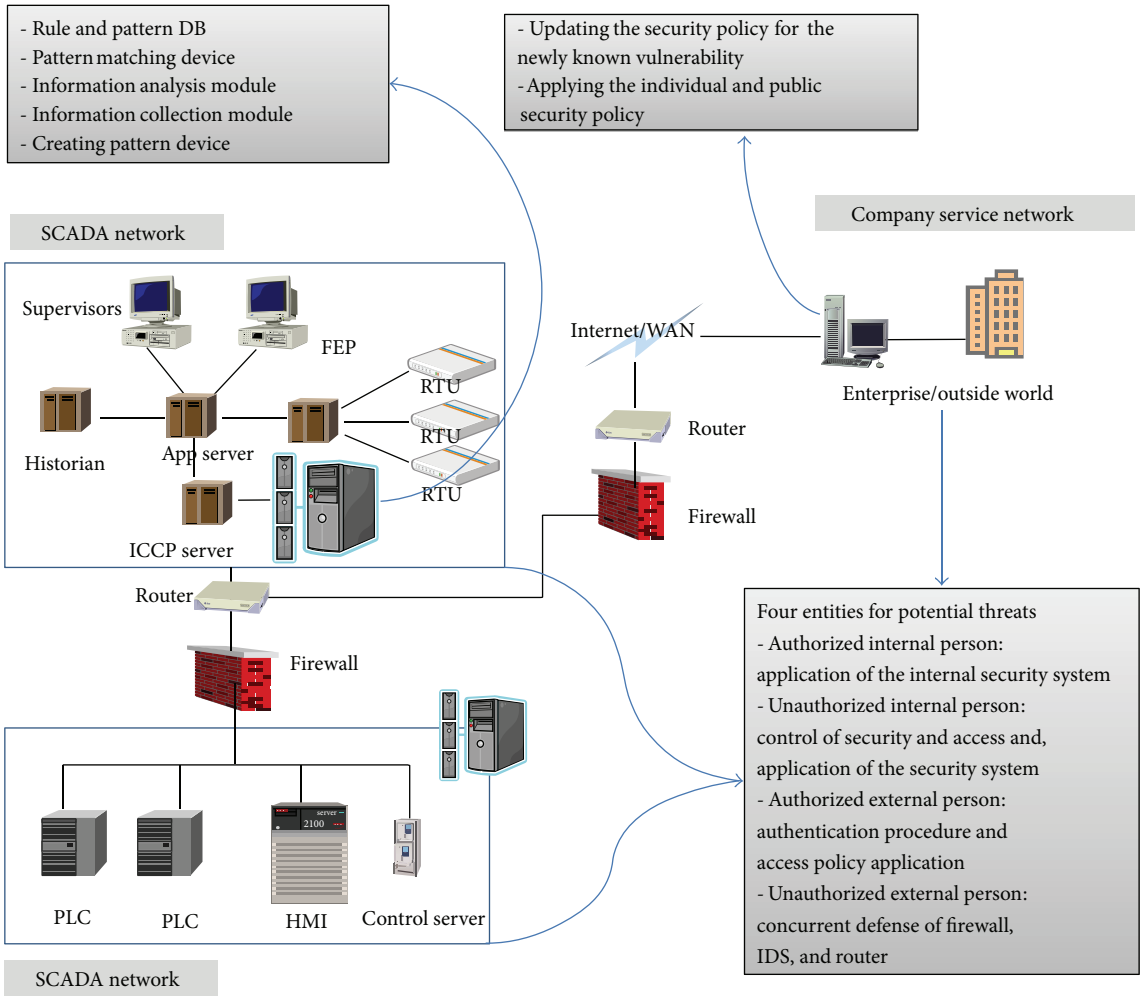


Figure 3: Intrusion detection module in the proposed secure model.

network are validated by applying the network-based module to the module of internal network access points.

Malicious code and web viruses are detected on the network through the checksum of metadata files and the state data monitoring of log data processes in the host IDS module of the SCADA network. To this end, it is necessary to execute the application module of the IDS system based on the host. In addition, effects on the process of the existing SCADA network service should be considered.

As shown in Figure 3, new attack patterns can be generated by information analysis module and collection module, and then these patterns and their matching rules are registered to the rule and pattern database. These new patterns and rules are applied to the pattern matching devices. The rule and pattern DBs located in internal IDS network are updated continuously, and this updating mechanism allows the whole systems to get active defensing characteristics and more powerful detecting capability. In the proposed IDS model, access entities are categorized into four kinds of entity types, and this classification definitely reduces unnecessary security policy enforcements and makes the system really up to date and protective.

4.2. Countermeasures against APT Attacks in the Proposed Security Model. Vulnerability studies on the security of SCADA networks are mostly conducted for various network platforms and communication protocols. Integrated control policy is needed through data surveillance and analysis to test illegal access to the total network system and for detection and measures against intrusion or malicious code. Changing data and access status should always be analyzed through network monitoring, and reporting processes should be performed through such monitoring.

A security defense method where an elaborate pattern matching process for known malicious code is applied is required for measuring APT attacks. Measuring strategies for polymorphic malicious code should be applied through the analysis function of network traffic and processes. At the core of APT attacks are unknown system vulnerabilities and new malicious codes. Therefore, updating previous intrusion pattern data or rules is not significant. In order to measure APT attacks, it is necessary to evaluate risks through processes such as rapid and precise virtual execution, in case of detection of unknown intrusion patterns or malicious code. Additionally, it is required to design countermeasures against

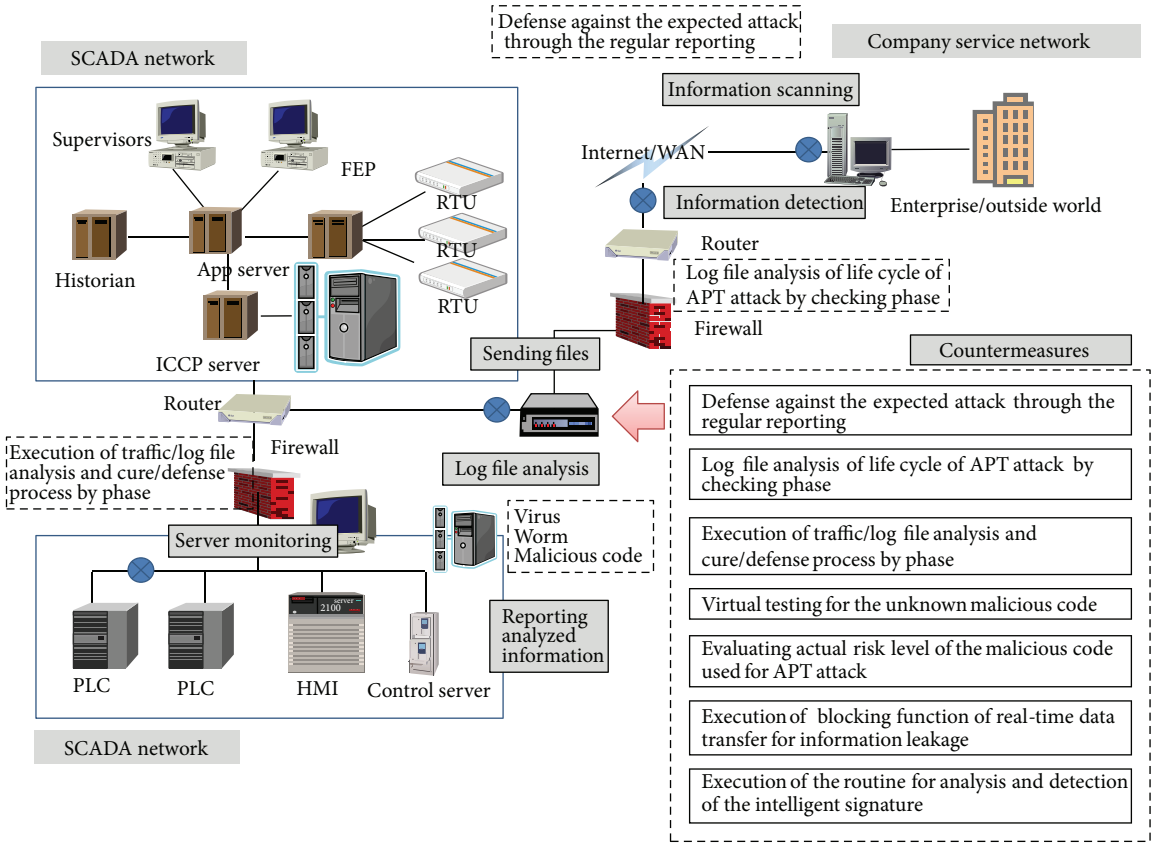


Figure 4: Countermeasures against APT attacks in the proposed security model.

the security vulnerabilities of file transferring and various data attachments for inbound and outbound protocols of the SCADA system.

To this end, inspection processes of security items must be executed.

(i) Integrated monitoring on network data includes

(a) files to be updated: data and registry files residing in the system;

(b) updating data key values of the application program that is necessary for platform operation;

(c) updating log file and file device data.

(ii) Inspection on log data includes

(a) comprehensive analysis of the log file for application programs that run in the system;

(b) determining whether log data has been invaded after the intrusion detection process, which is based on the rule generated from the secure model, has been executed.

Figure 4 shows the proposed security model and its various countermeasures against APT attacks. The seven sorts of countermeasures are considered in the attack detecting process and also can be applied to the attack detection policy. These countermeasures are designed to protect the SCADA system from unknown or unexpected APT attacks, and they can be elaborated for providing virtual test environment for unknown malicious codes, evaluating actual risk levels, and blocking real-time information leakage. Consequently, the overall security model is very strong at managing unexpected threats and its various attack patterns, and it is very useful for controlling already infiltrated code and its risk level change. The proposed model is conducting continuous security checking phases including the log file analysis and reporting, traffic analysis, and information scanning. Analyzed results are used for updating protective mechanism which is monitoring APT life cycle.

5. Conclusions

This paper has analyzed security vulnerabilities of the existing closed network SCADA, which is one of the industrial control systems, and shows that such SCADA network can be an m-connected SCADA network.

SCADA networks that typically operate within a closed network have recently been connected to several intranets, extranets, and other devices in order to achieve operational effectiveness and convenience; therefore, the security of SCADA network cannot be guaranteed anymore by just using its isolation property. Because of several APT attacks, damage has been reported for the control systems of large-scale organizations, including nuclear power plants under

SCADA networks. In establishing security countermeasures of the SCADA network, a concrete security design should be made under consideration about how to apply new policy on the existing services and how big its main and side effects are.

The security model presented in this paper recognized the connecting status through asset analysis of SCADA networks, analyzed the connection type of intra- and extranetworks, applied the well-defined host/network-based intrusion detection module, provided continuous monitoring on data as a countermeasure against APT attacks, and designed the security module for surveillance analyses and effective controls.

The proposed security model counters the existing vulnerabilities through the well-made IDS rules which are refined through asset analysis for integrated security measures to the SCADA network, IT devices, and field devices. In addition, the proposed model analyzes all possible paths of APT attacks constantly, monitors any changes in the systems and networks in real time, reports novel intrusion patterns, and applies new IDS rules to its own rule database. In our on-going research, more detailed design elaborations will be taken into our proposed security model to be used in practical SCADA networks.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2011-0014394).

References

[1] W. Kim, H. K. Kim, K. Lee, and H. Y. Youm, "Risk analysis and monitoring model of urban SCADA network infrastructure," *Journal of the Korea Institute of Information Security and Cryptology*, vol. 21, no. 6, pp. 67–81, 2011.

[2] K. Stouffer, J. Falco, and K. Scarfone, *Guide to Supervisory Control and Data Acquisition (SCADA) and Industrial Control Systems Security*, NIST Special Publication 800-82, 2008.

[3] S.-J. Kim, B.-H. Kim, S.-S. Yeo, and D.-E. Cho, "Network anomaly detection for m-connected SCADA networks," in *Proceedings of International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA '13)*, pp. 351–354, October 2013.

[4] R. J. Robles and M.-K. Choi, "Assessment of the vulnerabilities of SCADA,control systems and critical infrastructure systems," *International Journal of Grid and Distributed Computing*, vol. 2, no. 2, pp. 27–34, 2009.

[5] S. Cheung, B. Dutertre, M. Fong, U. Lindqvist, K. Skinner, and A. Valdes, "Using model-based intrusion detection for SCADA network," in *Proceedings of the SCADA Security Scientific Symposium*, pp. 127–134, January 2007.

[6] J. Verba and M. Milvich, "Idaho national laboratory supervisory control and data acquisition intrusion detection system (SCADA IDS)," in *Proceedings of the IEEE International Conference on Technologies for Homeland Security (HST '08)*, pp. 469–473, May 2008.

[7] P. Oman and M. Phillips, "Intrusion detection and event monitoring in SCADA networks," in *Critical Infrastructure Protection*, E. Goetz and S. Shenoi, Eds., vol. 253 of *IFIP International Federation for Information Processing*, pp. 161–173, Springer, 2007.

[8] K. E. Holbert, A. Mishra, and L. Mill, "Intrusion detection through SCADA systems using fuzzy logic-based state estimation methods," *International Journal of Critical Infrastructures*, vol. 3, no. 1-2, pp. 58–87, 2007.

[9] D. Choi, S. Lee, D. Won, and S. Kim, "Efficient secure group communications for SCADA," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 714–722, 2010.

[10] J. O. Kwon and Y. J. Hong, "A study on the security management plan of industrial control system," *Samsung SDS Journal of IT Services*, vol. 8, no. 2, pp. 112–135, 2011.

[11] A. Carcano, I. N. Fovino, M. Masera, and A. Trombetta, "State-based network intrusion detection systems for SCADA protocols: a proof of concept," in *Critical Information Infrastructures Security*, E. Rome and R. Bloomfield, Eds., vol. 6027 of *Lecture Notes in Computer Science*, pp. 138–150, Springer, 2010.

[12] International Electro Technical Commission, "Data and communication security—profiles including MMS," IEC Standard IEC 62351-4, 2007.

[13] International Electro Technical Commission, "Data and communication security—security for IEC 61850," IEC Standard IEC 62351-6, 2007.

[14] V. Jyothsna, V. V. R. Prasad, and K. M. Prasad, "A review of anomaly based intrusion detection systems," *International Journal of Computer Applications*, vol. 28, no. 7, pp. 26–35, 2011.

[15] D.-J. Kang, J.-J. Lee, Y. Lee, I.-S. Lee, and H.-K. Kim, "Quantitative methodology to assess cyber security risks of SCADA system in electric power industry," *Journal of Korea Institute of Information Security and Cryptology*, vol. 23, no. 3, pp. 445–457, 2013.

[16] NIST Special Publication 800-30 Revision 1, "Guide for Conducting Risk Assessments," 2012, http://csrc.nist.gov/publications/drafts/800-30-rev1/SP800-30-Rev1-ipd.pdf.

[17] IDG, *Blue Code Security Report; APT Attack*, 2012.

[18] IDG, *Understanding of Next-Generation Security Evasion Techniques APT*, 2012.

[19] IDG, *The Start of the Next-Generation Hacking, Malicious Code and Understanding How to Respond*, 2012.

[20] NIST, *Guide to Industrial Control Systems (ICS) Security*, 2008.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 617346, 8 pages
http://dx.doi.org/10.1155/2014/617346

Research Article

Optimization of Processor Clock Frequency for Sensor Network Nodes Based on Energy Use and Timing Constraints

Youngmin Kim, Heeju Joo, and Chan-Gun Lee

Department of Computer Science and Engineering, Chung-Ang University, Dongjak-gu, Seoul 156-756, Republic of Korea

Correspondence should be addressed to Chan-Gun Lee; cglee@cau.ac.kr

Received 16 December 2013; Accepted 16 April 2014; Published 16 June 2014

Academic Editor: Jongsung Kim

The effectiveness of sensor networks depends critically on efficient power management of the sensor nodes. Dynamic voltage frequency scaling (DVFS) and dynamic power management (DPM) have been proposed to enable energy-efficient scheduling for real-time and embedded systems. However, most power-aware scheduling algorithms are designed to deal with only those cases in which the task execution time is determined solely by the clock frequency of the processor. In this study, we propose an extended task execution model that is appropriate for the sensor nodes and an algorithm that determines the optimal clock frequency for a node's processor. We analyze the extended model and verify that our algorithm calculates the clock frequency that optimizes energy savings while satisfying the timing constraints.

1. Introduction

A typical sensor network consists of multiple sensor nodes and wireless networks that connect these nodes. Each sensor in a sensor network runs on a battery with a limited power supply [1]. Hence, it is considered critical for the sensor node to operate in an energy-saving manner. Numerous approaches have been reported recently for saving the sensors' energy [2, 3].

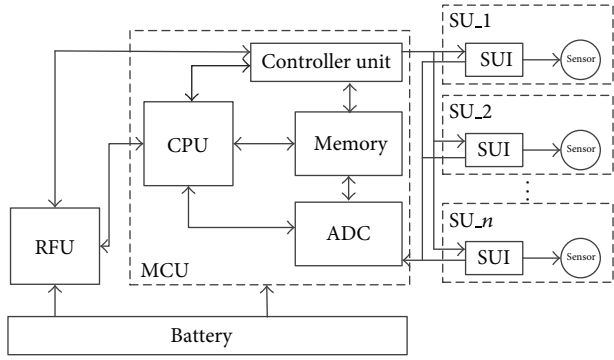
Each sensor node is composed of units for sensing, processing, radio frequency (RF) transmission, and battery power supply. A typical sensor executes real-time applications with timing constraints; the application periodically operates sensing units, processes the collected data, and transmits the processed data to the wireless network; thus, the challenge at the sensor node is to finish the above tasks with minimal energy expenditure while satisfying the timing constraints. There are real-time scheduling algorithms that are designed to address energy issues; among them, DVFS [4] and DPM [5, 6] are the most widely used schemes in the field.

DVFS is a technique in which the clock frequency and the operating voltage of a processor are adjusted during the execution. A processor operates at different clock frequencies depending on the core voltage. Reducing the core voltage and the clock frequency not only reduces energy use, but also reduces execution speed; thus, in any reduction scheme of this type there is a limit on the energy savings that can be realized without violating the timing constraints.

Real-Time DVFS technology has also been developed to exploit DVFS while also addressing the issue of timing constraints. Various other approaches have been reported in the literature to exploit DVFS, including the use of various models to describe the relevant tasks and systems [7–9].

The DPM technique functions differently, switching a device into a sleep mode when there is no task to execute. A device consumes minimal energy while in sleep mode but is also inactive, and waking the device typically entails a certain delay. This issue should be considered when developing a scheduling algorithm that not only uses sleep mode but also needs to guarantee real-time responsiveness. Benini et al. [10] have studied the question of when to switch a device into sleep mode in real-time scheduling.

Recently, Devadas and Aydin [11] and Zhao and Aydin [12] studied the interplay of DVFS and DPM for a real-time embedded system. They analyzed the combined use of DVFS and DPM for a real-time application that uses devices. However, these approaches employ an execution model in which the execution time of a real-time task depends only



MCU: Microcontroller unit
ADC: Analog digital convertor
SU: Sensor unit
RFU: Radiofrequency unit

Figure 1: Architecture of the sensor node.

on the clock frequency of the processor, which means that every device is impacted by the clock frequency. In practice, there are devices that do not depend on the clock frequency. We argue that an extended task execution model should be studied that takes both the processor and the devices into account.

In this study, we propose an extended energy model in which both of clock-dependent and clock-independent task executions are considered and present analysis results on this model. In addition, we provide an algorithm for calculating the clock frequency of the processor that optimizes the system's energy consumption; this algorithm was developed considering the interplay of DVFS and DPM for the extended model. This paper is extended from our preliminary work [13] for the nodes of sensor network.

The rest of this paper is organized as follows. Section 2 describes the system architecture of a sensor node. Section 3 proposes an extended model that considers both the clock-dependent and clock-independent task executions. In Section 4, we analyze the extended model and derive an algorithm to determine the optimal clock frequency; that is, the frequency that minimizes energy consumption while satisfying the timing constraints. Section 5 presents simulation results and analysis. We summarize the paper and suggest future work in Section 6.

2. Sensor Node Architecture

Figure 1 shows the system architecture of a typical sensor node with multiple sensor units.

As shown in Figure 1, a sensor node consists of a microcontroller unit (MCU), a RF Unit (RFU), a battery, and multiples sensor units (SUs). The analog signal detected by each SU is translated into digital data after being processed by the analog digital convertor (ADC) of the MCU. Hence, the resources such as CPU, memory, ADC, and RFU are shared by multiple SUs. The control unit controls the other components within the MCU.

It should be noted that each SU is designed to operate at a fixed frequency that is typically determined by the system designer; hence, the sensor unit is a system clock-independent device. In addition, the RFU operates at yet another frequency and depends neither on the clock frequency of the CPU nor on that of the sensor unit.

3. System Models

In this section, we present models for the task execution, the device's operation, and the system's energy consumption. The task execution model is designed to consider the different clock frequencies of the processor and the device. We normalize the clock frequencies of the processor and the device into the range of [0, 1]. In addition, our discussion will be based on the systems running on frame-based real-time scheduling.

3.1. Task Execution Model. Recent studies on the interplay of DVS and DPM for a real-time embedded system [11, 12] proposed the following model for task execution:

$$C(f) = \frac{C_{\max}}{f}. \tag{1}$$

By using this model, we can calculate a task's execution time by a processor with the clock frequency f. In the equation, $C_{\max}$ is the execution time for the case with the processor running at the clock frequency $f_{\max}$, which is normalized to 1.

However, the above model does not consider the case in which the execution time also depends on the clock frequency of the device. We present an extended model for considering the frequencies of both the processor and the device as shown in (2). In this equation, $C_{\max}$ represents

the execution time for a task when the processors and the devices run at their maximum clock frequencies:

$$C(f_1,\ldots,f_n) = \frac{C_{\max}}{f_1}\emptyset_1 + \cdots + \frac{C_{\max}}{f_n}\emptyset_n + C_{\max}\left(1-(\emptyset_1+\cdots+\emptyset_n)\right). \quad (2)$$

Each $\emptyset_i$ represents the ratio of the dependency of the task on f_i; these also ranges from 0 to 1. The expression $(1-(\emptyset_1+\cdots+\emptyset_n))$ is a ratio representing the device's dependency on its clock frequency.

Equation (3) shows how to determine $\emptyset_i$. $C^i_{\min}$ is calculated by setting only f_i to its minimum, which is referred to as $f^i_{\min}$, and the others to their maximums. Note that $C^i_{\min}$ is the task execution time when the ith device operates at its minimum clock frequency:

$$\emptyset_i = \frac{C_{\max}-C^i_{\min}}{C_{\max}}\frac{f^i_{\min}}{1-f^i_{\min}}. \quad (3)$$

Equations (4) shows that $C_{\max}$, the execution time when the ith device operates at its maximum clock frequency, is derived from (2) by setting f_i to $f^i_{\max}$. Similarly, $C^i_{\min}$ can be derived by setting f_i to $f^i_{\min}$ as shown in (5):

$$C_{\max} = \frac{C_{\max}}{f^1_{\max}}\emptyset_1 + \cdots + \frac{C_{\max}}{f^i_{\max}}\emptyset_i + \cdots + \frac{C_{\max}}{f^n_{\max}}\emptyset_n + C_{\max}\left(1-(\emptyset_1+\cdots+\emptyset_n)\right), \quad (4)$$

$$C^i_{\min} = \frac{C_{\max}}{f^1_{\max}}\emptyset_1 + \cdots + \frac{C_{\max}}{f^i_{\min}}\emptyset_i + \cdots + \frac{C_{\max}}{f^n_{\max}}\emptyset_n + C_{\max}\left(1-(\emptyset_1+\cdots+\emptyset_n)\right). \quad (5)$$

3.2. Device Model. In this paper, we assume that the devices attached to the system support both DVFS and DPM. There are two device modes: active mode and sleep mode. In active mode, the device is ready to process requests, while in sleep mode, the device goes into a low-power mode and cannot process requests. In addition, we assume that the device is in an active mode at the beginning of the task and stays in this the mode until the task execution ends; this is typically referred to as intertask device scheduling. The following list gives various notations for device parameters:

(i) P_a: power consumption in active mode;

(ii) P_s: power consumption in sleep mode;

(iii) E_{sd} and T_{sd}: power consumption and time delay, respectively, that are incurred in changing from active mode to sleep mode;

(iv) E_{wu} and T_{wu}: power consumption and time delay, respectively, that are incurred in changing from sleep mode to active mode.

The break-even time B represents the minimum duration of the sleep mode that will compensate for the added power consumption incurred in changing between the active and sleep mode. Devadas and Aydin [11] calculated the break-even time as shown in (6) and we adopt this calculation in our extended execution model:

$$B = \frac{E_{sd}+E_{wu}-(T_{sd}+T_{wu})P_s}{P_a-P_s}. \quad (6)$$

3.3. Energy Model. The system energy E is partitioned into the static energy E_s and the dynamic energy E_d; specifically, E_s is the static energy consumed in operating the system clock, while E_d is dynamically varying energy that relates directly to the clock frequency of the processor. In this paper, we focus on dynamic energy consumption. We propose an extended energy model that utilizes (1) from [11]. However, as some portion of the task does not depend on any processor frequency, it is difficult to exactly determine the execution time of a task. In this paper, the energy model considers the execution time model represented by (2). Equation (7) represents the energy model, which is an extension of that presented in [11]. $\delta(f)$ represents the slack time in the frame at the frequency f. $\mathbf{D}_a$ and $\mathbf{D}_s$ represent the sets of devices transitioned to active and sleep mode, respectively, during the slack time of the frame. The total active power of devices P_{ind} is independent of the frequencies of CPUs and devices:

$$E_d(f_1,\ldots,f_n) = \left(af^3+P_{\text{ind}}\right)C(f_1,\ldots,f_n) + \sum_{i|D_i\in\mathbf{D}_a} P^i_a\delta(f_1,\ldots,f_n) + \sum_{i|D_i\in\mathbf{D}_s}(E_{sd}+E_{wu}). \quad (7)$$

4. Optimal Frequency Decision in SUs

In order to simplify the discussion, we assume that a sensor node is equipped with a sensing unit that runs at a fixed frequency. The following equation represents an extended energy model for this scenario:

$$E_d(f) = \left(af^3+P_{\text{ind}}\right)C(f) + \sum_{i|D_i\in\mathbf{D}_a}\delta(f)P^i_a + \sum_{i|D_i\in\mathbf{D}_s}(E_{sd}+E_{wu}). \quad (8)$$

The analysis of our energy model is illustrated in the following sections. In addition, we discuss how to determine the optimal frequency that minimizes the sensor node's energy consumption. Note that the analysis result is very similar to [11] because our model is extended from theirs. In the following, we shall refer the left and the right areas based on $f = f_b$ to areas α and β, respectively, where f_b represents the break-even point of the device.

4.1. Minimum Energy Consumption Frequency Decision in Area α. In area α, the device is never switched into the sleep mode during its idle time because the duration of sleep mode would be shorter than the break-even time. Therefore, the

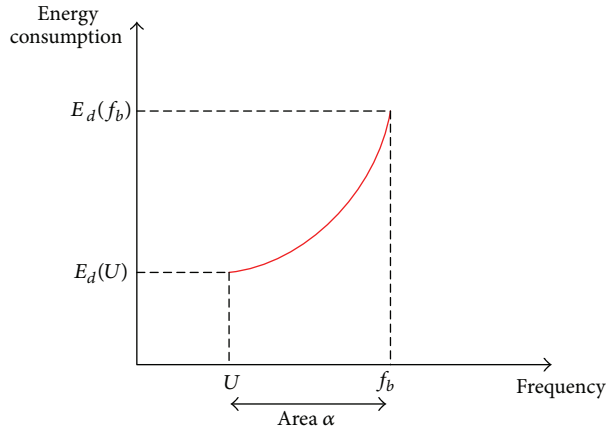


Figure 2: Energy consumption in area α.

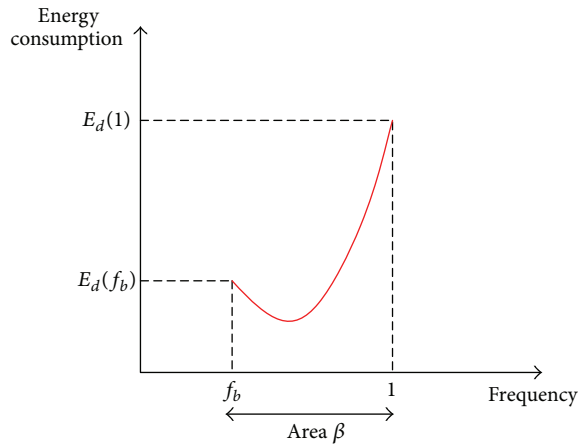


Figure 3: Energy consumption in area β.

```
(1) function FrequencyDecision()
(2)    f_opt1 ← U                          ▷the optimal frequency decision in area α
(3)    if x_1 > 0 then                     ▷the optimal frequency decision in area β
(4)       f_ee ← x_1
(5)    else
(6)       f_ee ← x_2
(7)    end if
(8)    if f_ee < f_b then
(9)       f_opt2 ← f_b
(10)   else if f_ee > 1 then
(11)      f_opt2 ← 1
(12)   else
(13)      f_opt2 ← f_ee
(14)   end if
(15)   E_diff ← E_d(f_opt1) − E_d(f_opt2)
(16)   if E_diff > 0 then                  ▷the optimal frequency decision on an entire system
(17)      f_opt ← f_opt2
(18)   else
(19)      f_opt ← f_opt1
(20)   end if
(21)   return f_opt
(22) end function
```

Algorithm 1: The optimal frequency decision algorithm.

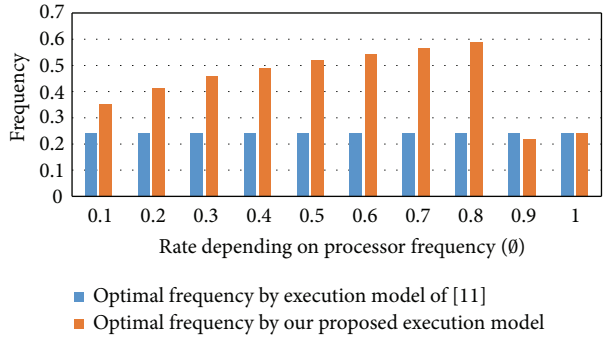


Figure 4: Optimal clock frequencies for various ∅, as determined by a previous approach and the proposed approach.

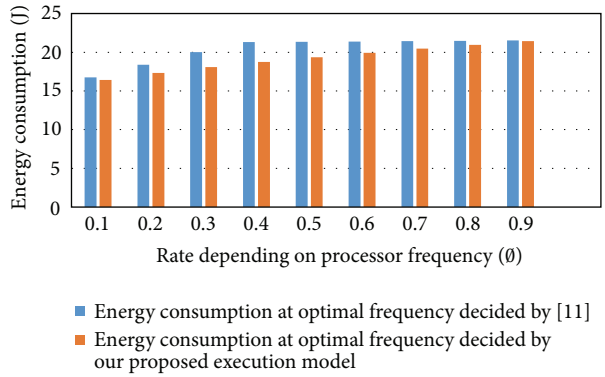


Figure 5: Energy consumption for various ∅.

energy consumption of the device is a constant value $P_a d$, where d is the deadline of the task. On the other hand, the energy consumption of a processor is proportional to the cube of processor frequency f^3; hence, lower frequencies correspond to lower energy consumptions.

Note that the execution of the task is finished exactly at its deadline when $f = U$, the lowest clock frequency that meets the timing constraint. Hence, U is the optimal clock frequency of the processor for this area. The utilization U is derived as shown below:

$$U = \frac{C_{\max}\emptyset}{d - C_{\max}(1-\emptyset)}. \tag{9}$$

Figure 2 shows the energy consumption trend in area α. The energy consumption increases monotonically as a function of frequency. Hence, $E_d(U)$ is minimal energy consumption in area α.

4.2. Minimum Energy Consumption Frequency Decision in Area β. In area β we need to consider the tradeoff between the energy consumption by the processor and the device with respect to the clock frequency. The energy model can be obtained as follows:

$$E_d(f) = \left(af^3 + P_a\right)\left(\frac{C_{\max}}{f}\emptyset + C_{\max}(1-\emptyset)\right) + E_{sd} + E_{wu}. \tag{10}$$

Equation (10) is strictly convex for $f > 0$; hence, there must be a minimum point in $f > 0$. The frequency that minimizes (10) can be determined by setting its derivative Equation (11) to zero. Figure 3 shows the relationship between frequency and the corresponding energy consumption within area β:

$$E_d'(f) = f^{-2}\left(3a(1-\emptyset)C_{\max}f^4 + 2a\emptyset C_{\max}f^3 - P_a\emptyset C_{\max}\right). \tag{11}$$

Solving the quartic formula yields four values. The derivative of our energy model consists of a second-order term that is positive infinite at $f = \infty$ and a negative second-order term that is negative infinite at $f \approx 0$. Thus, our energy model must have a minimum point for $f > 0$. One of the solutions for (11), which corresponds to x_1 or x_2 in (12), represents the minimum point.

Two Solutions of (11)*:*

$$
\begin{aligned}
x_1 &= -\frac{2a\emptyset C_{\max}}{12a(1-\emptyset)C_{\max}} \\
&\quad -\frac{1}{2}\left(\left(\frac{4a\emptyset C_{\max}}{12a(1-\emptyset)C_{\max}}\right)^2+\left(\frac{\sqrt[3]{2}R}{81(a(1-\emptyset))^3\sqrt{\left(108a^2P_a\emptyset^3C_{\max}^3\right)+\sqrt{-4\left(36aP_a(1-\emptyset)C_{\max}^2\right)^3+\left(108a^2P_a\emptyset^3C_{\max}^3\right)^2}}}\right.\right. \\
&\qquad \left.\left.+\frac{\sqrt{\left(108a^2P_a\emptyset^3C_{\max}^3\right)+\sqrt{-4\left(36aP_a(1-\emptyset)C_{\max}^2\right)^3+\left(108a^2P_a\emptyset^3C_{\max}^3\right)^2}}}{9\sqrt[3]{2}a(1-\emptyset)C_{\max}}\right)\right)^{1/2} \\
&\quad +\frac{1}{2}\left(\left(\frac{\sqrt{32}a\emptyset C_{\max}}{12a(1-\emptyset)C_{\max}}\right)^2-\left(\frac{\sqrt[3]{2}R}{81(a(1-\emptyset))^3\sqrt{\left(108a^2P_a\emptyset^3C_{\max}^3\right)+\sqrt{-4\left(36aP_a(1-\emptyset)C_{\max}^2\right)^3+\left(108a^2P_a\emptyset^3C_{\max}^3\right)^2}}}\right.\right. \\
&\qquad \left.\left.+\frac{\sqrt{\left(108a^2P_a\emptyset^3C_{\max}^3\right)+\sqrt{-4\left(36aP_a(1-\emptyset)C_{\max}^2\right)^3+\left(108a^2P_a\emptyset^3C_{\max}^3\right)^2}}}{9\sqrt[3]{2}a(1-\emptyset)C_{\max}}\right)\right) \\
&\quad -\left(2a\emptyset C_{\max}\right)^3\times\left(108\left(a(1-\emptyset)C_{\max}\right)^3\left(\left(\frac{4a\emptyset C_{\max}}{12a(1-\emptyset)C_{\max}}\right)^2+\left(\frac{\sqrt[3]{2}R}{81(a(1-\emptyset))^3\sqrt{\left(108a^2P_a\emptyset^3C_{\max}^3\right)+\sqrt{-4\left(36aP_a(1-\emptyset)C_{\max}^2\right)^3+\left(108a^2P_a\emptyset^3C_{\max}^3\right)^2}}}\right.\right.\right. \\
&\qquad \left.\left.\left.+\frac{\sqrt{\left(108a^2P_a\emptyset^3C_{\max}^3\right)+\sqrt{-4\left(36aP_a(1-\emptyset)C_{\max}^2\right)^3+\left(108a^2P_a\emptyset^3C_{\max}^3\right)^2}}}{9\sqrt[3]{2}a(1-\emptyset)C_{\max}}\right)\right)^{1/2}\right)^{-1}\right)^{1/2}, \\
x_2 &= -\frac{2a\emptyset C_{\max}}{12a(1-\emptyset)C_{\max}}+\frac{1}{2}\left(\left(\frac{4a\emptyset C_{\max}}{12a(1-\emptyset)C_{\max}}\right)^2+\left(\frac{\sqrt[3]{2}R}{81(a(1-\emptyset))^3\sqrt{\left(108a^2P_a\emptyset^3C_{\max}^3\right)+\sqrt{-4\left(36aP_a(1-\emptyset)C_{\max}^2\right)^3+\left(108a^2P_a\emptyset^3C_{\max}^3\right)^2}}}\right.\right. \\
&\qquad \left.\left.+\frac{\sqrt{\left(108a^2P_a\emptyset^3C_{\max}^3\right)+\sqrt{-4\left(36aP_a(1-\emptyset)C_{\max}^2\right)^3+\left(108a^2P_a\emptyset^3C_{\max}^3\right)^2}}}{9\sqrt[3]{2}a(1-\emptyset)C_{\max}}\right)\right)^{1/2} \\
&\quad +\frac{1}{2}\left(\left(\frac{\sqrt{32}a\emptyset C_{\max}}{12a(1-\emptyset)C_{\max}}\right)^2-\left(\frac{\sqrt[3]{2}R}{81(a(1-\emptyset))^3\sqrt{\left(108a^2P_a\emptyset^3C_{\max}^3\right)+\sqrt{-4\left(36aP_a(1-\emptyset)C_{\max}^2\right)^3+\left(108a^2P_a\emptyset^3C_{\max}^3\right)^2}}}\right.\right. \\
&\qquad \left.\left.+\frac{\sqrt{\left(108a^2P_a\emptyset^3C_{\max}^3\right)+\sqrt{-4\left(36aP_a(1-\emptyset)C_{\max}^2\right)^3+\left(108a^2P_a\emptyset^3C_{\max}^3\right)^2}}}{9\sqrt[3]{2}a(1-\emptyset)C_{\max}}\right)\right) \\
&\quad +\left(2a\emptyset C_{\max}\right)^3\times\left(108\left(a(1-\emptyset)C_{\max}\right)^3\left(\left(\frac{4a\emptyset C_{\max}}{12a(1-\emptyset)C_{\max}}\right)^2\right.\right. \\
&\qquad \left.\left.+\left(\frac{\sqrt[3]{2}R}{81(a(1-\emptyset))^3\sqrt{\left(108a^2P_a\emptyset^3C_{\max}^3\right)+\sqrt{-4\left(36aP_a(1-\emptyset)C_{\max}^2\right)^3+\left(108a^2P_a\emptyset^3C_{\max}^3\right)^2}}}\right.\right.\right. \\
&\qquad \left.\left.\left.+\frac{\sqrt{\left(108a^2P_a\emptyset^3C_{\max}^3\right)+\sqrt{-4\left(36aP_a(1-\emptyset)C_{\max}^2\right)^3+\left(108a^2P_a\emptyset^3C_{\max}^3\right)^2}}}{9\sqrt[3]{2}a(1-\emptyset)C_{\max}}\right)\right)^{1/2}\right)^{-1}\right)^{1/2}
\end{aligned}
\tag{12}
$$

We can choose a clock frequency that minimizes the energy consumption in area β as that frequency that minimizes energy consumption for $f > 0$. We first find out whether the frequency that corresponds to the minimum point is in area β. The frequency of the minimum point and the optimal frequency of area β are denoted by f_{ee} and f_{opt2}, respectively, in the following:

(i) if $f_{\text{ee}} < f_b$, $f_{\text{opt2}} = f_b$;

(ii) if $f_b \leq f_{\text{ee}} \leq 1$, $f_{\text{opt2}} = f_{\text{ee}}$;

(iii) if $1 < f_{\text{ee}}$, $f_{\text{opt2}} = 1$.

4.3. Optimal Frequency Decision. Now, we can choose a frequency that optimizes the energy consumption by considering the results from both areas. The difference in the optimal energy consumption of each area is denoted by E_{diff}, and is calculated as shown in (13). If $E_{\text{diff}} > 0$, we set $f_{\text{opt}} = f_{\text{opt2}}$; otherwise, we set $f_{\text{opt}} = f_{\text{opt1}}$:

$$E_{\text{diff}} = E_d\left(f_{\text{opt1}}\right) - E_d\left(f_{\text{opt2}}\right). \tag{13}$$

Algorithm 1 shows an algorithm that derives the optimal clock frequency for a sensor node. After calculating the optimal frequency for each area using (10), we choose the optimal clock frequency between the two.

5. Simulation Results

In this section we evaluate our algorithm in various scenarios. Throughout this section, we assume a real-time task with $d = 42$ and $C_{\max} = 10$. The task uses a device with the following characteristics: $P_a = 0.5$, $E_{sd} = 5$, $E_{wu} = 5$, $T_{sd} = 10$, $T_{wu} = 10$, and $B = 20$. The switching capacitance of the processor is assumed to be $a = 1$. The parameters are set to the values used in [11] for the comparison purpose.

Figure 4 shows that the previous approach fails to handle the case in which the execution time of the task is not determined solely by the clock frequency. In the figure, the clock frequency dependency $\emptyset$ changes from 0.1 to 1, and only when the entire system depends on the clock frequency, that is, $\emptyset = 1$, does the previous approach [11] achieve an the optimal solution. In contrast, our proposed algorithm accurately determines the optimal frequency in all cases. In the worst case, using the result of the previous approach consumes 2.5 J more energy than the optimal solution.

This simulation results in Figure 5 show that our algorithm enables better power management than the previous approach [11]. This is because our algorithm is based on an extended model that more accurately reflects the architecture of the sensor node.

6. Conclusion and Future Work

A sensor node employs various devices, including a microcontroller unit for processing, a radio frequency unit for transmissions, a battery to supply power, and multiple sensor units for sensing. Some of these devices depend directly on the clock frequency and others do not.

In this study, we proposed an extended model that considers the case in which the execution time of the task is not determined only by the clock frequency of the processor.

Based on the proposed model, we presented an algorithm that calculates the optimal clock frequency of the processor to achieve system-wide energy savings. We validated our approach by both analysis and simulation results in various scenarios.

In future work, we shall extend the model to handle multiple devices; we also intend to perform experiments with actual sensor nodes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by the Chung-Ang University Excellent Student Scholarship, the National Research Foundation (NRF-2011-0013924), and the MSIP (Ministry of Science, ICT, and Future Planning) under the ITRC (Information Technology Research Center) support program (NIPA-2014-H0301-14-1023) supervised by the NIPA.

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–105, 2002.

[2] R. Poornachandran, H. Ahmad, and H. Çam, "Energy-efficient task scheduling for wireless sensor nodes with multiple sensing units," in *Proceedings of the 24th IEEE International Performance, Computing, and Communications Conference (IPCCC '05)*, pp. 409–414, April 2005.

[3] B. Hohlt, L. Doherty, and E. Brewer, "Flexible power scheduling for sensor networks," in *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks (IPSN '04)*, pp. 205–214, April 2004.

[4] M. Weiser, B. Welch, A. Demers, and S. Shenker, "Scheduling for reduced cpu energy," in *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation*, 1994.

[5] H. Cheng and S. Goddard, "Online energy-aware I/O device scheduling for hard real-time systems," in *Proceedings of the Design, Automation and Test in Europe (DATE '06)*, March 2006.

[6] V. Devadas and H. Aydin, "Real-time Dynamic Power Management through device forbidden regions," in *Proceedings of the 14th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS '08)*, pp. 34–44, April 2008.

[7] S. Saewong and R. Rajkumar, "Practical voltage scaling for fixed priority rt-systems," in *Proceedings of the IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS '03)*, 2003.

[8] H. Aydin, R. Melhem, D. Mossé, and P. Mejia-Alvarez, "Power-aware scheduling for periodic real-time tasks," *IEEE Transactions on Computers*, vol. 53, no. 5, pp. 584–600, 2004.

[9] M. Marinoni and G. Buttazzo, "Elastic DVS management in processors with discrete voltage/frequency modes," *IEEE*

Transactions on Industrial Informatics, vol. 3, no. 1, pp. 51–62, 2007.

[10] L. Benini, A. Bogliolo, and G. de Micheli, "A survey of design techniques for system-level dynamic power management," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 8, no. 3, pp. 299–316, 2000.

[11] V. Devadas and H. Aydin, "On the interplay of dynamic voltage scaling and dynamic power management in real-time embedded applications," in *Proceedings of the 7th ACM International Conference on Embedded Software (EMSOFT '08)*, pp. 99–108, October 2008.

[12] B. Zhao and H. Aydin, "Minimizing expected energy consumption through optimal integration of DVS and DPM," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD '09)*, pp. 449–456, November 2009.

[13] Y. Kim and C. G. Lee, "Determining frequency for energy saving in consideration of the frequency-dependent code ratio for real-time embedded systems," in *Proceedings of the Korea Conference on Software Engineering (KCSE '13)*, 2013.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 839710, 10 pages
http://dx.doi.org/10.1155/2014/839710

Research Article

Game-Theoretic Camera Selection Using Inference Tree Method for a Wireless Visual Sensor Network

Yeong-Jae Choi,[1] Go-Wun Jeong,[1] Yong-Ho Seo,[2] and Hyun S. Yang[1]

[1] *Department of Computer Science, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea*
[2] *Department of Intelligent Robot Engineering, Mokwon University, 88 Doanbuk-ro, Seo-gu, Daejeon 302-729, Republic of Korea*

Correspondence should be addressed to Yong-Ho Seo; yhseo@mokwon.ac.kr

Received 28 November 2013; Accepted 17 April 2014; Published 15 June 2014

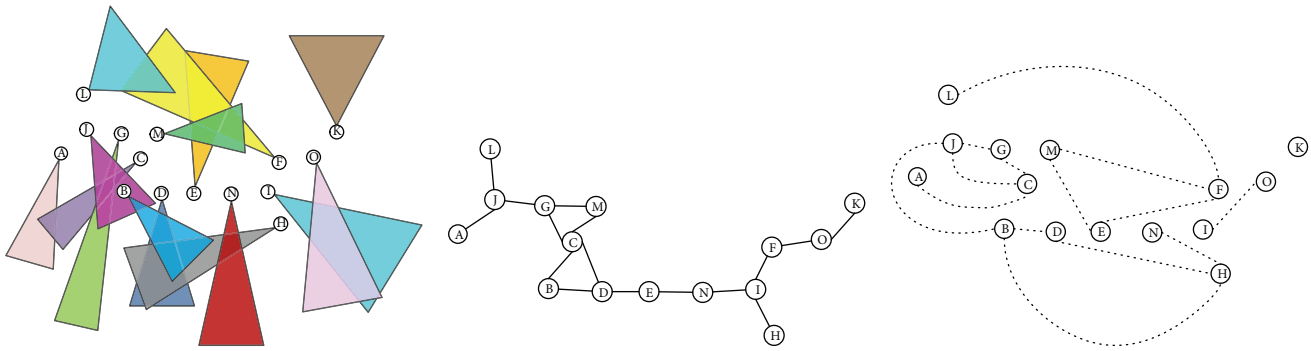
Academic Editor: Jongsung Kim

In a wireless visual sensor network consisting of wireless, battery-powered, and field-of-view (FoV) overlapping and stationary visual sensors, trade-offs exist between extending network lifetime and enhancing its sensing accuracy. Moreover, aggregating individual inferences from each sensor is essential to generate a globally consistent inference, because these individual inferences can be biased by noise or other unexpected conditions. Those challenges can be addressed by reducing the amount of data transmission among the sensors and by activating, in a timely manner, only a desirable camera subset for given targets. In this paper, we initialize an optimal data transmission path among visual sensors using the inference tree method, which is vital for collecting individual inferences and building a global inference. Based on the optimal data transmission path, we model the camera selection problem in a cooperative bargaining game. In this game, based on the serial dictatorial rule, camera sensors cooperatively attempt to raise the overall sensing accuracy by sequentially deciding their own mode between "sleep" and "active" in descending order of their bargaining power. Simulated results demonstrate that our proposed approach outperforms other alternatives, resulting in reduced resource overhead and improved network lifetime and sensing accuracy.

1. Introduction

There has been an increasing necessity to extract relevant information for multiple targets moving around inside wide areas for surveillance purposes. Moreover, these requirements must be fulfilled in a cost-efficient manner. A visual sensor is primarily equipped with an image sensing device, several processing units, communication facilities, and a set of batteries. This composition is very suitable for surveillance, because of the advantageous characteristics such as a wide monitoring area, rich visual information, and human-friendly data. Following the development of these inexpensive, powerful, and easily-deployable visual sensors, wireless visual sensor networks (WVSN) consisting of wireless, battery-powered, and field-of-view (FoV) overlapping and stationary visual sensors have been widely employed for surveillance in public places [1, 2]. Compared with other types of wireless sensors, visual sensors are impacted more by their limited bandwidth, lifespan, computation, and storage capabilities, because they contend with high-dimension data sets containing rich information generated from images [3]. Thus, it will be necessary to initialize an optimal data transmission path to reduce the amount of data transmission among the sensors for a global inference and to efficiently activate only selected cameras, which optimizes their collective coverage of given targets in a timely manner. The latter is referred to as *Camera Selection* (CS).

In this paper, we initialize an optimal data transmission path among visual sensors utilizing the inference tree method, which is a key component in aggregating individual inferences and building a global inference with minimized transmissions [2]. Based on the optimal transmission path, every visual sensor can exchange data with other sensors. Additionally each sensor can autonomously switch its mode between "sleep" (in *sleep* mode, the sensor stops capturing data; it will continue to transmit data) and "active" only with local knowledge, during advanced target analysis beyond basic tracking. This local rationale can be feasible under the practical assumption that FoV overlapping cameras can directly communicate with each other; additionally, the view



(a) Configuration (b) Physical topology (c) Logical topology

Figure 1: A wireless visual sensor network.

of a target is shared only between neighboring cameras. Thus, the camera's local knowledge is sufficient to measure its sensing contribution towards the global sensing accuracy, considering its neighbors' contributions. As discussed in [4, 5], each of the multiple cameras can cooperatively bargain for an optimal collective coverage. Our proposed approach utilizes the serial dictatorial rule, in which preferred cameras are prioritized to select their mode in earlier steps to achieve efficient computation.

The remainder of this paper is organized as follows. Section 2 introduces and discusses several advanced CS solutions; Section 3 describes the inference tree method used to initialize the data transmission path; Section 4 models a CS problem in a cooperative game; Section 5 describes our proposed serial dictatorial rule-based solution to force every camera to select its optimal mode using its local knowledge; Section 6 simulates and analyzes our approach in several network performance metrics and resource overhead; Section 7 compares our solution with its representative alternatives employing our concerned metrics; and Section 8 concludes the paper.

2. Related Work

Because of the lack of CS-related studies at this time, we note that several considerable research efforts have proposed comparable solutions, mainly based on greedy selection (GS) or potential game (PG).

The GS-based approaches select a camera best satisfying the criteria of interest at the time until l cameras are selected for a certain l that has been heuristically determined in advance. For example, a candidate camera may be selected based on the extent to which it improves the current visual hull and thus reduces occlusion [6, 7]; candidate cameras may also be selected based on the degree to which their images are different from images of already-selected cameras, with a goal of producing varied images that reduce redundancies [8]. Although this selection process may provide richer information about targets, it does not adaptively cope with the dynamics in the targets' locations because of the stationary l.

For seamless tracking, the PG studies assign a camera to every target based on the maximum utility between the camera and the associated target. A camera's utility for a target is quantified by how large the camera records the target and its face in stationary camera networks [4] or the degree to which the camera can closely observe the target in active camera networks [5]. Until a set of probabilities describing how effectively a target is tracked by a camera for every camera and every target is converged, a set of these utilities and the probability set are alternatingly updated; additionally, they both influence each other. This one-for-one selection could efficiently produce results for tracking or locating given targets; however, it may be insufficient to produce advanced target information, such as multitarget interaction analysis.

All of the discussed techniques have approached a CS solution in a centralized manner, with global knowledge of every camera and every target. To obtain such global knowledge, high bandwidth consumption will necessarily occur at a central operator in centrally controlled networks or at every camera sensor in distributed controlled networks. To avoid such dissipation, our approach employs only limited knowledge; however, it aims to perform similarly to, or improve, the previously mentioned alternatives.

3. Inference Tree Method for Initializing Transmission Path

A WVSN can be described as displayed in Figure 1. Figure 1(a) displays the configuration of a WVSN, Figure 1(b) illustrates a physical topology based on wireless connection reachability (or geometric proximity), and Figure 1(c) represents a logical topology based on the FoV overlapping constraint. In this paper, however, we assume that FoV overlapping cameras can directly communicate with each other. Thus, the physical topology graph can be ignored. This assumption is feasible in practical applications and enables us to focus on the data transmission and camera selection challenges.

As displayed in Table 1 [9], power consumption costs are higher for a broadcasting network compared with a unicasting network. In order to minimize data transmissions for building a global inference in a WVSN, we need to convert the logical topology from a broadcasting/multicasting network to a unicasting network. To initialize a transmission path as

Table 1: Energy consumption for IEEE 802.11 11 Mbps wireless network card.

Network	Data	Energy cost per bit (uWs/byte)
Broadcast	Send	2.1
	Receive	0.26
Unicast	Send	0.48
	Receive	0.12

a unicasting network, we use the inference tree method [2]; the process is illustrated in Figure 2. When a logical topology is provided as an input, a weight value is calculated for each edge based on the amount of FoV overlapping between two visual sensors. The result of this initial step is a weighted graph. From the weighted graph, a maximum spanning tree is produced, and the center node of the maximum length path is selected as a root node. Utilizing the root node selected in the second step, the weighted graph is converted into a minimum depth tree employing a breadth-first search algorithm. The result of this step is an initial inference tree. Lastly, it is optimized with up and down actions necessary to build a balanced tree. We utilized the final resulting tree as an optimal data transmission path for the WVSN.

The results from our implementation are displayed in Figure 3. Figure 3(a) illustrates that there is a significant reduction in the number of data transmissions for both the leaf and internal nodes. Figure 3(b) illustrates that energy consumption for data transmissions has decreased for all nodes. Figures 3(a) and 3(b) successfully support that the inference tree method is effective for initializing an optimal data transmission path in WVSN.

4. Cooperative Game for CS

Consider c cameras, indexed by i, statically deployed, and t targets, indexed by j, randomly moving inside a geographical area. As previously stated, we also assume that any two neighboring cameras are able to communicate with each other if their FoVs overlap. The locations of cameras are initially calibrated and stationarily fixed. The locations of targets are updated based on any object localization algorithm of [10], utilizing the most recently recorded images at each time instant, provided to their associated cameras. Whenever new locations are provided, the expected target locations at the next time instant are also estimated by the extended Kalman filter as in [11]. At this point, every camera i is aware of the set of its observable targets to move into its FoV, termed T_i.

Illustration of parameter notations in camera i's FoV with targets j and j' can be described as displayed in Figure 4. Safe region sr_i is the set of every 2D point inside the dotted-line square, where a target is seen observed safely enough. Unsafe distance l_j of j not located in sr_i is the distance from the center of sr_i to j's location. Distinction angle $\theta_{jj'}$ of j and j' is the included angle between the i-to-j vector and the i-to-j' vector.

Declaring that m_i represents the mode of camera i between 0 for sleep (in *sleep* mode, the sensor stops capturing data; it will continue to transmit data) and 1 for active, we modeled our defined CS problem in the form of a classic, normal game given as ⟨*player, action, utility*⟩ in [12]. If function *utility* is identical for every player, the players are encouraged to be cooperative to maximize their shared utility [13]. These conditions are analogous to a CS where the objective, from the point of view of a game designer, is to obtain a minimal set of active cameras that can achieve a high sensing accuracy for given targets, equal to the accuracy provided by an entire network [14]. To enable every camera to autonomously select its mode for the objective, we replace the given game form by $\langle \{i, j\}, m_i, U_g \rangle$, where U_g is the global utility equally shared within the entire network and can be quantified as follows:

$$U_g\left(\{m_i\}\right) = \sum_{j=1}^{t} U_j\left(\{m_i\}_j\right). \tag{1}$$

For $\{m_i\}_j$, the mode set of the cameras, which are able to observe target j, (1) quantifies the global sensing accuracy given t targets by summing the extent to which each target j is well observed by its associated cameras in global target utility U_j. This value generally becomes larger as more cameras are active, but not necessarily. A small number of images may omnidirectionally cover a target; conversely, images that are too similar, produced by closely located and similarly oriented cameras, redundantly dissipate resources to transmit and process these images. By setting the upper bound for global target utilities to 1 to restrict the redundancies as in (3), we observe that the global target utility of target j is constructed by individual target utilities $\{\mathrm{tu}_j\}$ obtained by each of the associated cameras as in the following equation:

$$U_j\left(\{m_i\}_j\right) = \sum_{\{m_i\}_j} \mathrm{tu}_j\left(m_i\right), \tag{2}$$

such that

$$U_j\left(\{m_i\}_j\right) \le 1, \tag{3}$$

$$\mathrm{tu}_j\left(m_i\right) = m_i\, \mathrm{safe}_i\left(j\right) \begin{cases} 1 & \text{if } \left|T_i\right| = 1 \\ \min\limits_{j' \in T_i \setminus j} \mathrm{dist}_i\left(\theta_{jj'}\right) & \text{if } \left|T_i\right| \ge 2, \end{cases} \tag{4}$$

$$\mathrm{safe}_i\left(j\right) = \begin{cases} 1 & \text{if } j \in \mathrm{sr}_i \\ \dfrac{1}{l_j} & \text{if } j \notin \mathrm{sr}_i, \end{cases} \tag{5}$$

$$\mathrm{dist}_i\left(\theta_{jj'}\right) = \begin{cases} 1 & \text{if } \theta_{jj'} \ge \dfrac{90}{A_i} \\ \sin\left(A_i \theta_{jj'}\right) & \text{if } \theta_{jj'} < \dfrac{90}{A_i}. \end{cases} \tag{6}$$

The target utility of target j by camera i represents the smallest likelihood that j is sufficiently observed without any occlusion in i's FoV according to i's mode. This likelihood is determined utilizing three values: m_i of whether or not i is

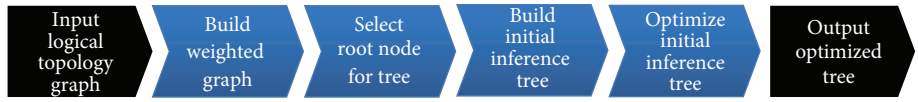
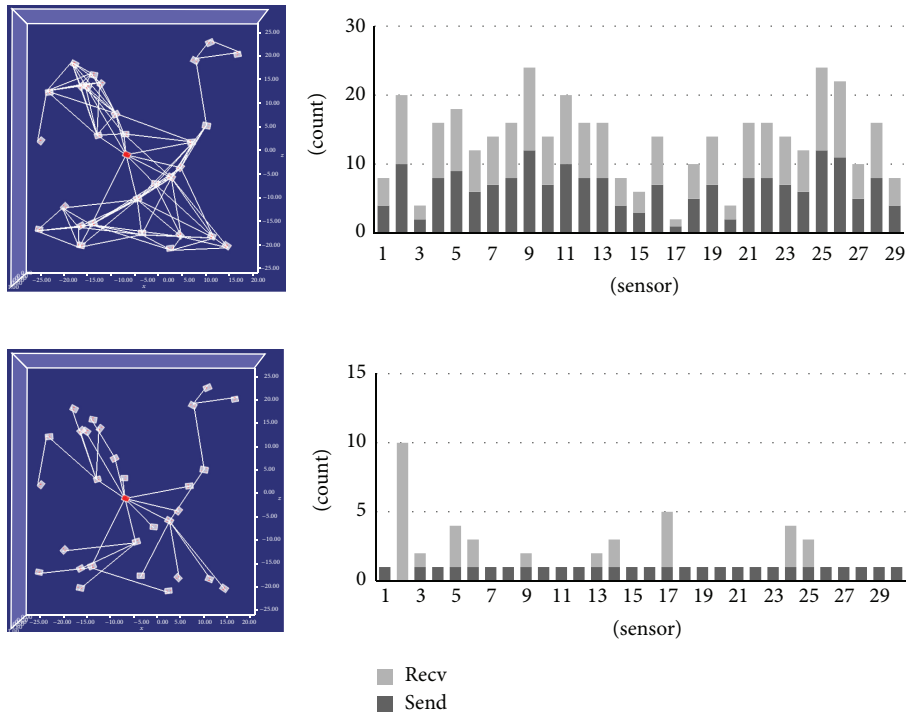
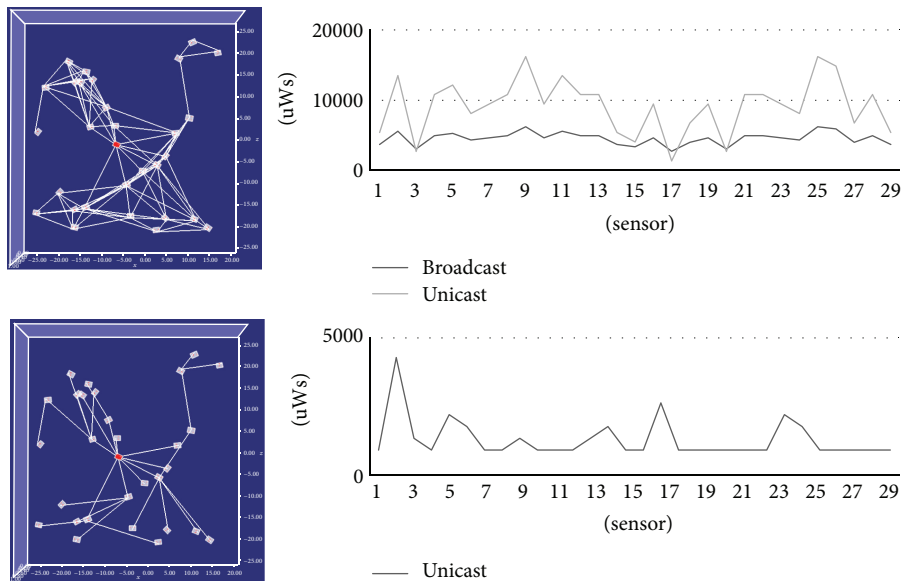


Figure 2: The process of inference tree method.



(a) Message count for 30 visual sensors



(b) Energy consumption using 1 Kbyte packet for 30 visual sensors

Figure 3: The results of the inference tree method.

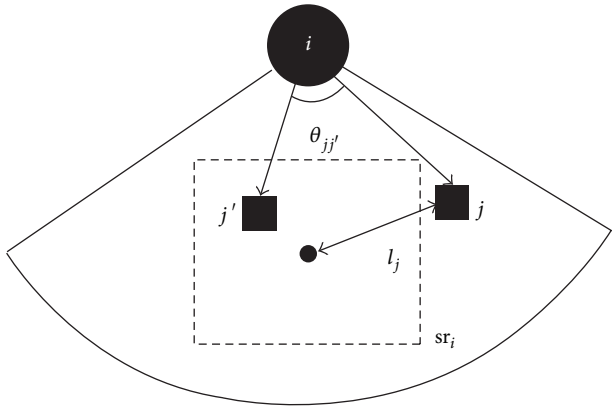


Figure 4: Illustration of parameter notations in camera i's FoV with targets j and j'. Safe region sr_i is the set of every 2D point inside the dotted-line square, where a target is observed safely enough. Unsafe distance l_j of j not located in sr_i is the distance from the center of sr_i to j's location. Distinction angle $\theta_{jj'}$ of j and j' is the included angle between the i-to-j vector and the i-to-j' vector.

active, safe_i of the degree to which j can be safely observed by i, and dist_i of the degree to which j is occluded by other targets in i's FoV, as in (4). As stated in Figure 4, j can be *safely* observed if it is located in i's safe region sr_i; otherwise, the safety value for j inversely decreases as j's unsafe distance l_j becomes greater as in (5). To measure the occluded degree of j in i's FoV, we define the distinction probability between images simultaneously taken by j and j' as in (6), where θ_{jj}, is the distinction angle of j and j' defined in Figure 4 and A_i is the scaling factor of i, because a greater $\theta_{jj'}$ will result in greater reduction of the occlusion between j and j'. Utilizing these definitions, both U_j and tu_j can have a value between 0 and 1, indicating that even a single camera can provide a thorough observation of a target if the camera safely monitors the target without occlusion, as we have assumed.

Forthwith, we discuss the extent to which an individual camera contributes to a certain global utility. Camera utility of camera i evaluates its contribution degree to observe the set of its observable targets T_i according to m_i by summing the target utilities of every j in T_i as in the following equation:

$$\mathrm{cu}_i\left(m_i\right) = \sum_{j \in T_i} \mathrm{tu}_j\left(m_i\right). \tag{7}$$

The equations from (4) to (7) reflect the following characteristics that only camera sensors can possess.

(i) Every camera can be assumed to have, in its FoV, its own safe region where any target is observable in sufficient detail to give desirable information with minimal distortion.

(ii) Owing to the 3D-to-2D projection of imaging, the occlusion among multiple targets in a camera's FoV obstructs the extraction of the targets' information [3, 4].

(iii) The interaction analysis among multiple targets provides more relevant information about the targets beyond their locations [15].

While (i) and (ii) are, respectively, represented by (5) and (6) for (4), more relevant information is provided by cameras with a higher camera utility, which becomes greater as it observes more targets or its associated target utilities are greater, as in (7).

Subsequently, every camera selects its mode to cooperatively maximize their payoff and the global utility U_g, for the least number of active cameras while considering the condition (3). The mode selection process, to be discussed in the following section, is greatly enhanced by taking the camera utilities into account.

5. Serial Dictatorial Rule-Based Bargaining Solution

According to [13], the serial dictatorial rule is a sequence of dictatorial rules conducted by individual players whose exercising order is statically arranged by their bargaining powers. By evaluating camera utilities provided in (7), we consider that a camera has greater bargaining power if it observes a greater number of targets, in a less occluded manner, in the corresponding safe region. Given that all cameras possessing greater bargaining power have already determined their modes and a camera must presently select a mode, it will select the mode maximizing its payoff, U_g, according to the dictatorial rule. The camera is under the assumption that other cameras that have not determined their modes are in sleep mode [14]. This bargaining process is serially performed until every camera determines its mode while communicating with neighboring cameras as follows.

5.1. Order Cameras by Their Camera Utilities. Given estimated locations of T_i, every camera i computes the target utilities of every target in T_i and its own camera utility assuming $m_i = 1$ and subsequently transmits them to its neighboring cameras. Thereafter, i obtains its position in the dictatorial ordering list of it and its neighbors in descending order of their camera utilities, while initializing the mode set for the list, $M_i = \{0\}$.

5.2. Select the Current Mode. Prior to mode selection, every i waits for all the modes of its more bargaining-powerful neighbors to be announced while updating M_i if it is not the first on the list. Otherwise, it instantly assumes its mode by (8) for $M_{i,j}^{m_i}$ which is the target j's associated subset of M_i where only the mode of i is replaced by m_i. Consider

$$m_i = \begin{cases} 1 & \text{if } \sum_{j \in T_i} \left(U_j\left(M_{i,j}^1\right) - U_j\left(M_{i,j}^0\right)\right) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

A camera i will decide to be active only if it can improve the total of the global target utilities for every target of T_i by its contribution. Specifically, *a camera covers its FoV only*

if at least one target in its FoV is not sufficiently observed at the moment, indicating that it will not be responsible for the targets already well covered by other cameras. Subsequent to this mode decision, i announces its selected mode followed by i; accordingly every neighbor, termed i', updates $M_{i'}$ with i's new mode and drops every j such that $U_j(M_{i',j}) = 1$ obtained by the new $M_{i'}$ from $T_{i'}$ for rapid computation.

This local reasoning with limited knowledge soundly and completely extends maximizing U_g for the following theorem.

Theorem 1. *For every camera i with its mode m_i, it always holds that*

$$U_g(M^*) - U_g\left(M^{*,-i}\right) \geq \sum_{j \in T_i} \left(U_g\left(M_{i,j}^{m_i}\right) - U_g\left(M_{i,j}^{m_{-i}}\right)\right), \tag{9}$$

where M^* is the bargained mode set for every camera, m_{-i} is the opposite mode of m_i, $M^{*,-i}$ is the mode set where only m_i is replaced by m_{-i} in M^*, and $M_{i,j}$ is i's assumed mode set of i and i's neighbors associated with target j for the bargaining process.

Proof. The following equation (10) holds by the global utility definition and (11) is derived because the change of camera i's mode affects only the global target utilities of every j in T_i. As M_j^* and $M_j^{*,-i}$ are, respectively, restated as $M_{i,j}^{m_i}$ and $M_{i,j}^{m_{-i}}$, we must demonstrate that (12) always holds for every i with T_i for our claim. Consider that

$$U_g(M^*) - U_g\left(M^{*,-i}\right) = \sum_{j=1}^{t} \left(U_j\left(M_j^*\right) - U_j\left(M_j^{*,-i}\right)\right) \tag{10}$$

$$U_g(M^*) - U_g\left(M^{*,-i}\right) = \sum_{j \in T_i} \left(U_j\left(M_j^*\right) - U_j\left(M_j^{*,-i}\right)\right), \tag{11}$$

$$\sum_{j \in T_i} \left(U_j\left(M_j^*\right) - U_j\left(M_j^{*,-i}\right)\right) \geq \sum_{j \in T_i} \left(U_j\left(M_{i,j}^{m_i}\right) - U_j\left(M_{i,j}^{m_{-i}}\right)\right). \tag{12}$$

To more easily understand the claim, we restate the target argument as the following if-then rule.

By the nature of our bargaining process, $U_g(M^*) \geq U_g(M^{*,-i})$ always holds for every camera i. Given the condition, (12) also always holds for every i.

Subsequently, we verify this claim for both its soundness and completeness.

Soundness. We demonstrate that (12) and $U_g(M^*) \geq U_g(M^{*,-i})$, respectively, hold in the following two cases.

(a) case of $m_i = 1$ and $m_{-i} = 0$.

When i decides its mode as active while assuming that every mode for the less bargaining-powerful neighboring cameras is sleep, it is believed that it can improve the global target utility of any in T_i. Let us say that $\{j'\}$ is the target set each global target utility of which is actually raised by i. The difference resulting from i's mode change is given as follows:

$$\sum_{\{j'\} m_{i'} \in M_{i,j'} \setminus m_i} \mathrm{tu}_{j'}\left(m_{i'}\right). \tag{13}$$

Because any of less bargaining-powerful neighbors could be active to contribute to the improvement of the concerned global target utilities, $M_{j'}^*$ is likely to contain an equal or greater number of active cameras than $M_{i,j'}$. Thus, (12) is valid. In addition, (13) is always greater than or equal to 0 by the definition of the target utility, which eventually leads to the conclusion that $U_g(M^*) \geq U_g(M^{*,-i})$ is valid.

(b) Case of $m_i = 0$ and $m_{-i} = 1$.

When i determines its mode as sleep, it believes that every j in T_i is sufficiently covered by more bargaining-powerful neighbors, which derives $\sum_{j \in T_i}(U_j(M_{i,j}^1) - U_j(M_{i,j}^0)) = 0$ by (8). Similarly, it holds that $\sum_{j \in T_i}(U_j(M_j^*) - U_j(M_j^{*,-i})) = 0$ regardless of the other modes in M_j^*. Thus, (12) is valid in this case and $U_g(M^*) = U_g(M^{*,-i})$ is valid, too.

Completeness. We demonstrate that M^* exists, bargained by our bargaining process satisfying (12).

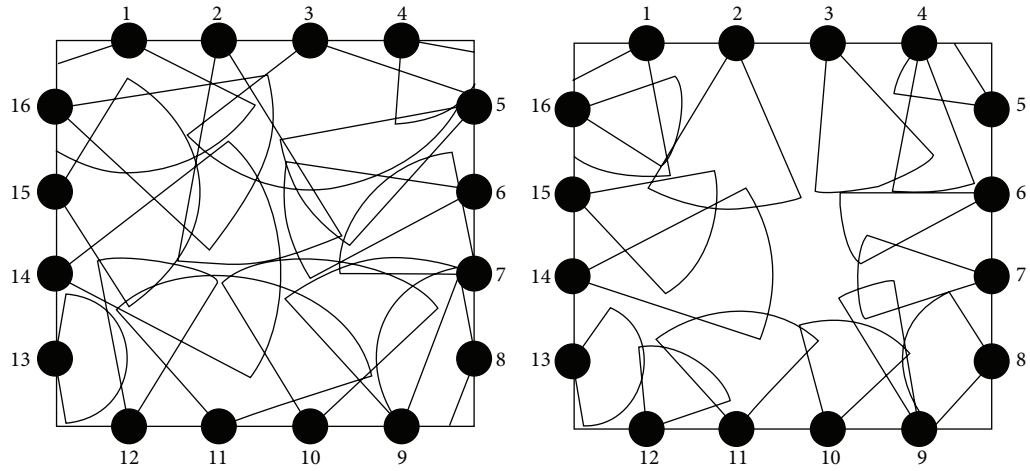
Consider any M^* where every mode is deterministically decided by its associated camera at its order, without possessing actual knowledge of the entire M^*. For every i, regardless of its decided mode, the entire mode set M_j^* for any target j will have more or the same number of active cameras compared to the virtual $M_{i,j}$ because i assumes that its less bargaining-powerful cameras adopt the sleep mode, as we discussed above. Therefore, the difference resulting from its mode change in M_j^* would always be equal to or greater than in $M_{i,j}$. In summary, M^* can be efficiently and deterministically obtained by (8) and any M^* satisfies (12) for every camera. □

Therefore, the selected mode by (8) with limited knowledge for every camera optimizes the global utility by Theorem 1.

6. Simulations and Analyses of Our Approach

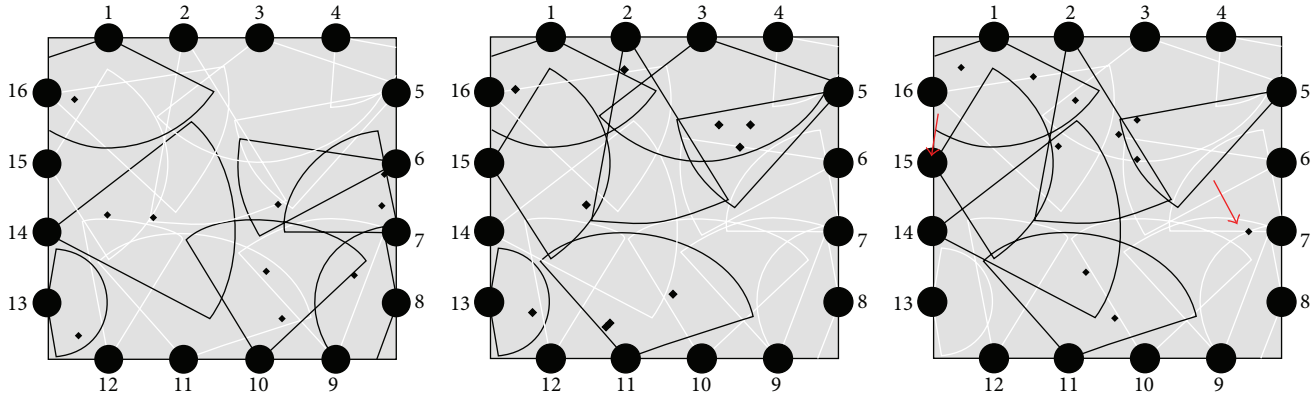
In this section, we evaluate our proposed serial dictatorial rule-based game for CS utilizing several network performance metrics in different simulations and quantitatively analyze the complexity required to achieve each step of our approach's design.

6.1. Simulated Performance Analysis. Our proposed approach has been simulated in the following multicamera and multitarget environment.



(a) Dense deployment of $d = 7$ (b) Sparse deployment of $d = 3$

FIGURE 5: Two different camera deployment environments with $d = 7$ of the left and $d = 3$ of the right.



(a) Distantly located targets (b) Closely located targets (c) Error case

FIGURE 6: Two successful and a single-failed examples of $d = 7$ and 10 targets.

(i) 16 stationary cameras of different FoVs, labeled 1 through 16, are deployed and calibrated in a square-shape area of 230×230 cells as in Figure 5.

(ii) Every camera is provided with its own safe region and FoV center point in advance.

(iii) The scaling factor A_i for every camera i is fixed to 2, because we assume that any two targets having a distinction angle of 45 or more degrees in the camera's FoV can be perfectly and separately monitored.

(iv) Every camera is initially charged with 1000 power bars and only dissipates a bar per time instance if it is active.

(v) The maximum neighboring density for every camera, d, is set to 7 as in Figure 5(a) or 3 as in Figure 5(b).

(vi) In a scenario, 10 or 20 targets freely move inside the area, under a speed of 15 cells per time instance, until the network lifetime has expired.

(vii) The target locations at the previous time instance obtained by any image-based localization of [10] are always announced to their associated cameras. After receiving this location information, every camera computes the expected locations of its observable targets at the current time instance by the extended Kalman filter as in [11]. To add authenticity to the simulation, we implement imperfect localization by adding a small amount of noise to our target localization.

To permit greater focus on the network performance of our interest achieved by our approach, we vary only the simulation situations with different neighboring densities and different target cardinalities, whereas the energy and scaling specifications of the cameras are not changed. In this environment, our approach is evaluated by observing the number of cameras that are active on average, *#Active*, the number of cameras that are redundantly active, *#RActive*, the number of

Table 2: Simulation results in the four cases of two densities and two target cardinalities.

(d, t)	*#Active*	*#RActive*	*#Missing*	*Lifetime*
(7, 10)	6.2	0.85	0.32	1364
(3, 10)	5.7	1.2	0.78	1481
(7, 20)	8.9	0.79	0.72	1163
(3, 20)	9.1	0.93	1.5	1249

Table 3: Analyzed complexity by each operation.

	Operation	Complexity
Computation	(1) Sensing and processing images	α
	(2) Estimating the current target locations	$O(t^2)$
	(3) Computing utilities	$O(t^2 + td)$
	(4) Selecting the mode	$O(t)$
Communication	(5) Transmitting processed data	β
	(6) Exchanging the target locations	$O(td)$
	(7) Exchanging utilities	$O(td)$
	(8) Exchanging selected modes	$O(td)$

targets that are missed, *#Missing*, at a time instance, and the amount of time that a camera network survives, *Lifetime*. To assist in understanding the three metrics, *#Active*, *#RActive*, and *#Missing*, we present two successful and a single-failed examples of $d = 7$ and 10 targets as in Figure 6. Generally, distantly located targets are well covered as in Figure 6(a); however, Cameras 3 and 5 of Figure 6(b) collectively monitor their observable targets closely located, because they could be occluded in the FoV of Camera 5. For both cases, *#Active* is 7. As in Figure 6(c), cameras may miss targets as Cameras 7 and 9, *#Missing* = 1, or be redundantly active as Camera 15, *#RActive* = 1, because of wrongly given previous, or differently estimated current, target locations.

Table 2 illustrates the average simulated values of our four primary performance metrics (in two or more significant figures) over 100 random scenario tests. The following claims can be derived from the results.

Claim 1. On average, six cameras for 10 targets and nine cameras for 20 targets are active. Accordingly, Lifetime for 10 targets is longer than that for 20 targets.

Claim 2. As d is larger, *#RActive* and *#Missing* are smaller. It would be more advantageous for more neighboring cameras to synthesize more accurate target locations by exchanging different information and the full network coverage of $d = 7$ is wider than that of $d = 3$.

Claim 3. A smaller numbers of active cameras in all cases somewhat extend the network lifetime from 1000 to 1481.

Table 4: Simulation results of the five approaches.

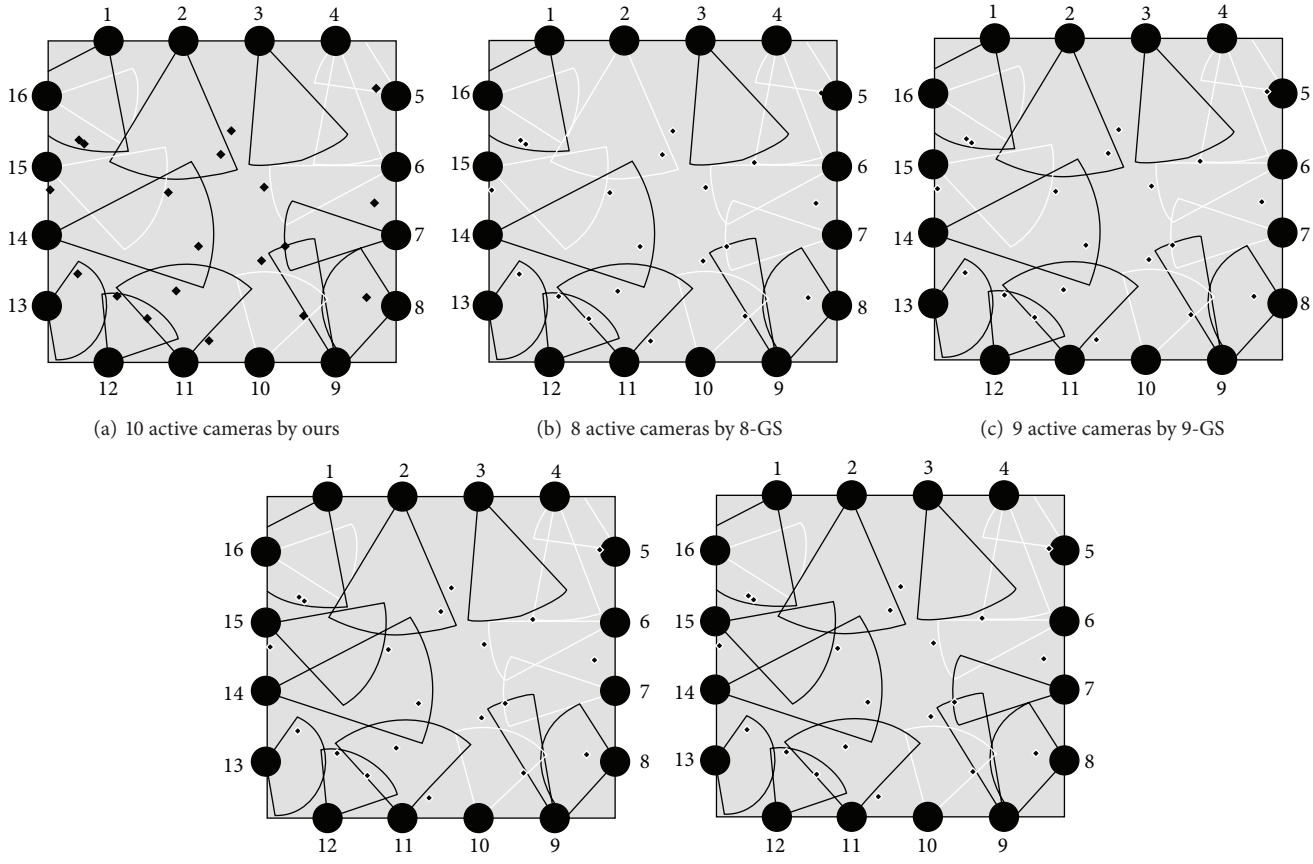
(d, t)			(7, 10)		
App.	Ours	5-GS	6-GS	7-GS	PG
#Active	6.2	5	6	7	6.7
#RActive	0.85	0.21	0.75	1.29	0.84
#Missing	0.32	1.5	0.74	0.40	0.32
Lifetime	1364	1482	1363	1287	1280
(d, t)			(3, 10)		
App.	Ours	5-GS	6-GS	7-GS	PG
#Active	5.7	5	6	7	5.8
#RActive	1.2	2.3	1.7	0.93	1.5
#Missing	0.78	1.6	1.1	0.90	0.80
Lifetime	1481	1912	1732	1489	1474
(d, t)			(7, 20)		
App.	Ours	8-GS	9-GS	10-GS	PG
#Active	8.9	8	9	10	9.9
#RActive	0.73	0.14	0.59	0.91	0.69
#Missing	0.72	1.7	1.1	0.82	0.72
Lifetime	1163	1201	1139	1107	1125
(d, t)			(3, 20)		
App.	Ours	8-GS	9-GS	10-GS	PG
#Active	9.1	8	9	10	9.2
#RActive	0.92	2.1	1.6	0.88	1.2
#Missing	1.5	2.6	2.1	1.8	1.5
Lifetime	1249	1446	1335	1304	1243

Claim 4. Because of the low *#RActives* and *#Missings* in all cases, our approach might be able to work over some localization errors, which necessarily occur in any existing localization techniques.

6.2. Complexity Analysis. Because we consider wireless cameras, we must discuss the resource overhead required by our proposed method. For each time instance in our design, energy consumption occurs according to the operations listed in Table 3.

The complexities for (1) and (5) depend on the models or algorithms camera sensors employ, and we leave them as α and β. We strongly emphasize that the two operations are conducted only by active cameras, and our proposed approach, on average, activates 0.57 (=9.1/16) times fewer cameras, including the worst case. The computational complexity for (2) is referred to as $O(t^2)$ [16]. Each target utility, each camera utility, and each global target utility respectively consume $O(t)$, $O(t^2)$, and $O(td)$ computations, which leads to $O(t^2 + td)$ computation for (3). Given such utilities, a camera determines its mode by searching in the $O(t)$ space for (4). Conversely, the communication complexity for (6) to (8) is equal to $O(td)$ because a camera maximally exchanges t pieces of information with d neighbors.

The energy consumption of camera sensors is dominated by (1) and (5) because of the significant size of image data [3, 17]. This supports our simple assumption about power consumption that only active cameras can monotonously



(a) 10 active cameras by ours (b) 8 active cameras by 8-GS (c) 9 active cameras by 9-GS

(d) 10 active cameras by 10-GS (e) 11 active cameras by PG

FIGURE 7: Screen shots of the five approaches in the simulation of $(d, t) = (3, 20)$.

consume a single power bar per time instance and indicates that our approach can be considered as fairly competitive if targets are not highly crowded in any of the FoVs.

7. Comparison Work

Utilizing the same environments employed in our simulations, we simulated the representative alternatives, PG of [8] and GS of [5], while measuring the four network performance metrics, *#Active*, *#RActive*, *#Missing*, and *Lifetime*. For GS in particular, we have heuristically assumed that the optimal number of active cameras for 10 targets and 20 targets are, respectively, 6 and 9 by *#Actives* of Table 2. Because the numbers are not consistently optimal, we examined three GS tests for the two different target sets as {5-GS, 6-GS, 7-GS} for 10 targets and {8-GS, 9-GS, 10-GS} for 20 targets. Table 4 lists each of the network performance results for the five different approaches, on average, after 100 tests. A smaller number of active cameras typically result in longer network lifetime; correspondingly, the GSes with smaller numbers of active cameras provided the network with longer life. However, their performance observed for *#RActive* and *#Missing* was lower than analogous results from our approach and PG. Compared with our approach, PG produces similar results for every factor. However, PG requires each camera to compute required utilities for every camera and to communicate with every other camera. This consumes greater resources compared with our approach. Therefore, we emphasize that our approach generally provides more advantageous trade-offs between {*#Active*, *Lifetime*} and *#Missing* and between *Lifetime* and resource overheads compared with the alternate approaches.

As representative instances of this comparison process, we provide five simulation screenshots for each approach in one simulation of $(d, t) = (3, 20)$ as in Figure 7. Aside from unobservable targets, our approach covers every target by 10 active cameras, whereas PG activates one redundant camera, Camera 15. As previously stated, because GS cannot adaptively select the number of active cameras, it misses targets as in Figures 7(b) and 7(c) or it additionally activates redundant cameras as in Figure 7(d).

8. Conclusion

In this paper, we addressed trade-offs between extending network lifetime and enhancing its sensing accuracy. To minimize the energy consumption necessary for data transmission while aggregating individual inferences to build

a global inference, we utilized the inference tree method to initialize an optimal data transmission path, and we demonstrated that it is very effective for reducing the number of data transmissions and energy consumption. We modeled a CS in the context of a cooperative bargaining game, where every participating camera serially optimizes the global utility, employing only local knowledge based on the serial dictatorial rule. The simulated results demonstrated that our approach extends network lifetime and performs accurately over limitedly accurate target locations. Moreover, our approach is energy-efficient for uncrowded targets, compared with the alternative representative conventional studies.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by the Basic Science Research Program (2013R1A1A2064233) and by the Converging Research Center Program (2013K000358) through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning, Korea. This work was partly supported by the IT R&D Program of MSIP/KEIT, (Development of personalized and creative learning tutoring system based on participational interactive contents and collaborative learning technology).

References

[1] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A survey on wireless multimedia sensor networks," *Computer Networks*, vol. 51, no. 4, pp. 921–960, 2007.

[2] Y.-I. Cho, *Collaborative data aggregation using inference tree for occupancy reasoning in visual sensor networks [Ph.D. thesis]*, KAIST, Daejeon, South Korea, 2011.

[3] S. Soro and W. Heinzelman, "A survey of visual sensor networks," *Advances in Multimedia*, vol. 2009, Article ID 640386, 21 pages, 2009.

[4] Y. Li and B. Bhanu, "Utility-based dynamic camera assignment and hand-off in a video network," in *Proceedings of the 2nd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC '08)*, pp. 1–9, Stanford, Calif, USA, September 2008.

[5] C. Soto, B. Song, and A. K. Roy-Chowdhury, "Distributed multi-target tracking in a self-configuring camera network," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1486–1493, Miami, Fla, USA, June 2009.

[6] D. B. Yang, J. W. Shin, A. O. Ercan, and L. J. Guibas, "Sensor tasking for occupancy reasoning in a network of cameras," in *Proceedings of the 1st International Workshop on Broadband Advanced Sensor Networks*, 2004.

[7] H. Lee, L. Tessens, M. Morbee, H. Aghajan, and W. Philips, "Sub-optimal camera selection in practical vision networks through shape approximation," in *Advanced Concepts For Intelligent Vision Systems*, vol. 5259 of *Lecture Notes in Computer Science*, pp. 266–277, 2008.

[8] R. Dai and I. F. Akyildiz, "A spatial correlation model for visual information in wireless multimedia sensor networks," *IEEE Transactions on Multimedia*, vol. 11, no. 6, pp. 1148–1159, 2009.

[9] L. M. Feeney and M. Nilsson, "Investigating the energy consumption of a wireless network interface in an ad hoc networking environment," in *Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3, pp. 1548–1557, Anchorage, AK, USA, April 2001.

[10] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Efficient subwindow search: a branch and bound framework for object localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2129–2142, 2009.

[11] J. Lin, W. Xiao, F. L. Lewis, and L. Xie, "Energy-efficient distributed adaptive multisensor scheduling for target tracking in wireless sensor networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 6, pp. 1886–1896, 2009.

[12] Y. Shoham and K. Leyton-Brown, "3. Introduction to noncooperative game theory: games in normal form," in *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundation*, pp. 47–88, Cambridge University Press, New York, NY, USA, 2009.

[13] O. Kibris, "Cooperative game theory approaches to negotiations," in *Handbook of Group Decision and Negotiation*, Springer, New York, NY, USA, 2010.

[14] G. Jeong, Y.-H. Seo, S.-S. Yeo, and H. S. Yang, "Serial dictatorial rule-based games for camera selection," in *Proceedings of the 4th FTRA International Conference on Mobile, Ubiquitous, and Intelligent Computing*, 2013.

[15] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, "Understanding transit scenes: a survey on human behavior-recognition algorithms," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 1, pp. 206–224, 2010.

[16] R. Triebel, "6. Online estimation: the kalman filter," in *Information Processing in Robotics*, ETH, Zürich, Switzerland, 2009.

[17] C. B. Margi, R. Manduchi, and K. Obraczka, "Energy consumption tradeoffs in visual sensor networks," in *Proceedings of the 24th Brazilian Symposium on Computer Networks (SBRC '06)*, Curitiba, Brazil, May 2006.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 379181, 8 pages
http://dx.doi.org/10.1155/2014/379181

Research Article

A Zone-Based Self-Organized Handover Scheme for Heterogeneous Mobile and Ad Hoc Networks

Murad Khan and Kijun Han

School of Computer Science and Engineering, Kyungpook National University, Daegu 702-701, Republic of Korea

Correspondence should be addressed to Kijun Han; kjhan@knu.ac.kr

Received 19 November 2013; Revised 2 May 2014; Accepted 15 May 2014; Published 12 June 2014

Academic Editor: Thomas Wook Choi

Uninterrupted internet services are need of the day and their demand will increase in future by manifold. However, providing uninterrupted services under heterogeneous networks environment is a challenging task. One of the major challenges in this regard is the management of handover system between various networks. This paper proposes a handover management scheme by dividing the total coverage area of the base station (BS) or access point (AP) into three different zones (strong, average, and weak zones), respectively, on the basis of receive signal strength indication (RSSI) to provide fast vertical handover support in heterogeneous wireless networks. Furthermore, a new event is proposed and added to the IEEE 802.21: media independent handover (MIH) standard to integrate the functionality of multithreshold. The proposed scheme is implemented in NS2 and it is shown that our scheme provides a superior performance over the previous methods used for similar purposes.

1. Introduction

Handover management is one of the most important tasks under the wireless networks which have received a great attention from many researches in recent years [1, 2]. Various algorithms have been proposed by researchers with the aim to transfer sessions between networks without losing connection and data.

A handover process starts by a mobile node (MN) when it receives weak RSSI from a BS/AP. After getting weak signals from the current BS/AP, an MN starts searching for available networks. The handover time is mainly dependent on the scanning delay of the available networks. Furthermore, an optimum network can be selected for handover from the available networks on the basis of price, security, transmission rate, and quality of service (QoS).

Due to aforementioned constraints, using different technologies for wireless communication leads to different problems such as selection of best network for handover, incompatibility among different networks, and handover delay. To overcome problem of moving across different networks, an efficient handover management scheme is needed. When an MN leaving from one BS to another, it first executes a discovery mechanism for searching nearby BSs and then MN makes a connection to it. This delay can be minimized by adjusting different factors like RSSI, data rate, available bandwidth, and signal to interference and noise ratio (SINR) from a BS [3, 4].

In 2008, International Telecommunication Union-Radio Sector (ITU-R) defines new specifications for 4G standard called International Mobile Telecommunication Advanced (IMT-Advanced). IMT-Advanced supports 100 Mbits/s for high mobility connections and 1 Gbits/s for low mobility connections [5].

As the data rate increases in this new standard, new technology starts participating in the long run of new access technology modifications. Moving around different access technologies becomes a problem because of compatibility issues between these networks.

The IEEE 802.21 MIH standard was published in 2008 for seamless handover between networks of the same type and also networks of different types [6]. The MIH standard consists of different services like media independent information service (MIIS), media independent event service (MIES), and

media independent command service (MICS). MIIS server is used for supporting various information services that can provide available networks within a geographical area.

After publishing of IEEE 802.21 standard, a lot of research works have been carried out in modifying it for different improvements [7–10]. The standard is experimentally tested on test beds in [11]. A recommendation is made by MIH standard for an MN is that it must be installed with all of the interfaces necessary for accessing these networks. The integration of link layers and upper layers is made through different triggers used in MIH standard [12]. When an MN detects a new network, it triggers a particular in event. The current BS of the MN performs a specific action on sensing the trigger of MN.

Recently, much research work is done on purifying the MIH standard [13–15]. The MIH standard has still many problems that can be addressed like: (1) high handover time is needed if the MIIS server is located many hops away, (2) time required for handover process is very short if the number of handovers is frequent in a handover region, and (3) failure of a hop requires alternate routes to MIIS server which can increase handover time.

In the MIH standard, an MN initiates handover when it receives RSSI below the predefined threshold. The time necessary for handover is constant even if the MIIS server is located many hops away. When an MIIS server is located many hops away, a longer time is required for an MN to get the information of the available networks. If an alternate route available to MIIS server consists of many hops in case of route failure, the time required for handover will be increased and it can lead to the breaking of connection during handover in the worst case.

In order to provide an efficient solution to above problems, this paper proposes an efficient handover scheme in which a BS collects information of available networks for an MN in advance. An MN does not wait for more time in a handover region for collection of information of available network. Now, whenever an MN needs the information of available networks, it will be available one hop away from it.

The rest of the paper is organized as follows. Literature review is presented in Section 2. The proposed scheme is explained in Section 3. Simulation results are presented in Section 4, and finally conclusion is given in Section 5.

2. Related Work

Recently, researchers have shown much interest in minimizing handover delay in wireless networks for fast handover [16]. In traditional approaches, a border between different networks is ignored, while focus is remain only on the handover strategies. Different techniques have been presented in past for better and fast handover [17, 18].

A border between two BSs or APs is a region where the probability of handover is high. As the networks are growing rapidly, load on a single MIIS server is increased. This problem is identified in [19] with a solution of dividing the network in different mobility zones. Each mobility zone is connected with a zone MIIS server which is further connected to a local MIIS server. And further, this local MIIS server is connected to a global MIIS server. This technique reduces access load on the MIIS server by dividing MIIS server functionalities into a sub-MIIS servers. However, frequent handovers in overlapping region from multiple MNs can lead to overflow of MIIS server cache and breaking of connection from an AP.

A technique to balance the number of handovers in a high probable region is used in [20]. The proposed scheme assumes a border zone between two networks and then connects one or more mobility anchor points (MAP) to border zone. Further, as long as an ongoing session continues an MN is connected with one particular MAP inside a border zone. Connection with a single MAP during border zone effectively reduces the handover in a particular zone. A vertical handover technique based on data rate is presented in [2]. This technique efficiently adjusted traffic load of different networks for smooth handover from one network to another.

A scheme based on finding best point of attachment (PoA) for handover has been proposed in [21]. The decision of best PoA is taken on the basis of RSSI and SINR of available networks. A BS obtains the information of RSSI and SINR and then passes it to the resource manager. The resource manager decides the best available network by generating a report called Report_Best_PoA. The report is sent to the MN for handover.

Mobile IP version 6 (MIPv6) cannot support two connections at the same time. A scheme based on adding a new entity called "added entity" (AE) has been proposed in [22]. This new entity can enable MIPv6 to support two connections at the same time. When an MN is going to handover from one network to another, then one of its connections is attached with current network and another makes a connection with the new network with the help of proposed AE.

The handover across heterogeneous networks is performed by collecting information of link layers and maps this information to a generic one by providing compatibility of an MN from one link layer device or interface to another link device or interface [23]. A media independent handover function (MIHF) is responsible for processing and modifying information obtained from different events and passes it to the upper layers. This exchange of different events is carried out by different media independent handover services. These services are called service APs (SAP) [24].

All of the techniques available in the literature enable handover routine when an MN crosses a particular threshold of RSSI. The area for handover defined in IEEE 80.21 is not enough to handle maximum number of handovers in an area of frequent handovers. When an MN detects a new network with strong RSSI in the available premises, then it sends a link going down event message to the current BS. The BS makes another event message and sends it to the Media independent server (MIIS) for collecting information of available networks. If the MIIS server is away by many hops from the BS, then it takes a long amount of time for a response from MIIS server. There are other cases when the probability of handover in a particular area is high. In such situations, handling handover is difficult for an access network (AN).

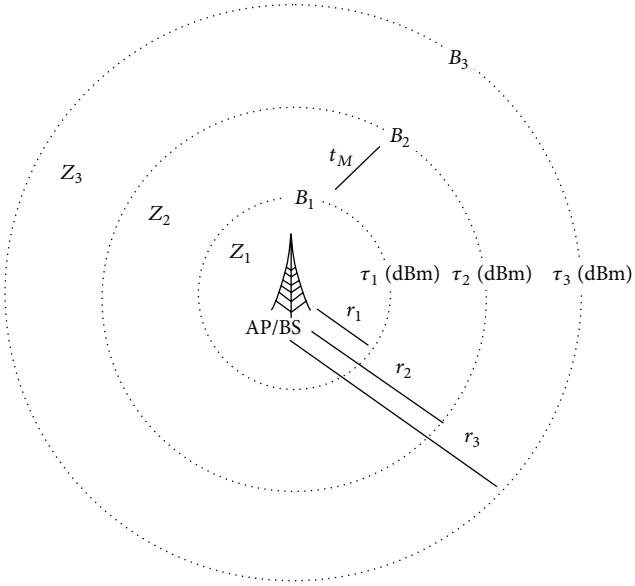


Figure 1: Division of zones with three different thresholds of RSSI.

3. Proposed Multilevel Threshold Scheme

3.1. Basic Concept of Proposed Scheme. The overall goal of the MIH standard is to provide handover management in heterogeneous networks. The MIH standard uses different events messages to control the process of handover management. These events messages are used by an MN for communication with a BS/AP and also by AP/BS for communication with an MN, new AP/BS, and MIIS server. We propose a new event message (MIH_LINK_INFO) and add this event message to the existing MIH standard.

The purpose of adding this new event message is to integrate a new functionality of multithreshold into the MIH standard. The total coverage area by each BS/AP is divided into three zones, that is, strong (Z_1), average (Z_2), and weak (Z_3) zones, respectively, on the basis of RSSI. In addition, we define three different thresholds, denoted by τ_1, τ_2, and τ_3, on the RSSI at the boundaries B_1, B_2, and B_3 of each zone. Figure 1 depicts division of region into three different zones based on different thresholds of RSSI.

When the MN approaches the boundary B_1 while moving in the coverage area of a BS/AP, it generates the new MIH_LINK_INFO event message and sends this event message to the current BS/AP. When BS/AP gets this event message, it sends an information packet to the MIIS server for getting information of available BSs/APs. After it gets reply from MIIS server, it sends the information of available BS/AP to MN upon approaching the boundary B_2. MN makes a connection with the new BS/AP and terminates the connection with old BS/AP. When the MN gets the boundary B_3, it indicates successful completion and termination of connection with the old network.

3.2. The Proposed Scheme. This section describes the proposed scheme and its working mechanism, such that how it reduces the handover time required for an MN moving over heterogeneous networks.

Actually, we are only interested in τ_1 and τ_2 in the proposed scheme, because τ_3 is only used for an indication of successful completion of handover process. The MN obtains the information of available APs and BSs much before the handover is initiated . In MIH standard only one MIIS server is responsible for the processing of handover information of every MN of each AN. It sometime leads to the congestion of information on the MIIS server and hence an MN suffer from longer handover delay. The proposed scheme efficiently fetched the handover information from MIIS server when an MN crosses τ_1. The proposed MIH_LINK_INFO event message consists of the following information:

(1) MN MAC address;

(2) MAC address of the associated BS;

(3) unique event identifier.

The MN MAC address is used by BS for identifying MN and saved for further processing when the MN approaches the boundary B_2.

This event message is broadcast over the network, and a BS can use its MAC address for identifying whether the event message is for this BS or not. If a BS gets an event message and it is not for this BS, it simply discards the event message. In the MIH, every event message has a unique identifier; a BS can process an event message on the basis of event message's identifier. BS receives an event message from an MN; it checks the information and forwards this event to the MIIS server on the basis of available information. MIIS server generates a response in the form of event message that sends the information of the new BS to the old BS. Upon receiving the event message by the old BS, the old BS extracts the new BS ID from the event message and generates a handover information table which is composed of three attributes: ID of the MN generating MIH_LINK_INFO event message, expiration time (denoted by T_{expire}), and new BS ID which will be used later by MN for handover.

T_{expire} is the time during which the current BS/AP should hold the new BS/AP information for the MN. After time duration of T_{expire} is elapsed, the information of the new BS/AP is deleted from the current BS/AP.

Consider two cases to justify necessity of T_{expire}: (i) the MN stops its movement while moving from B_1 to B_2 and (ii) MN disconnects from the current BS/AP before getting to B_2.

In both cases, when T_{expire} is expired, the BS/AP deletes the information of the new BS/AP for the MN that has already sent the request to the current BS/AP upon arriving at B_1.

Let t_M mean the elapsed time for the MN to move from B_1 to B_2 very slowly (at a speed of 1 m/s approximately). Then, T_{expire} is given double value of t_M to provide an MN with sufficient time. That is,

$$T_{\text{expire}} = 2t_M. \tag{1}$$

When the MN approaches the boundary B_2 while moving from one network to another, the entry of the MN will be deleted from handover information table. By this strategy,

Table 1: Simulation parameters.

Number of MNs	1 to 50
MN movement	Random Waypoint Mobility Model (10 m/s)
Propagation channel model	Two-Ray Ground
Wired links	1 Gbps
UMTS network	500 m
r_1, r_2, r_3 (UMTS network)	290, 350, 485 (meters)
Wi-Fi coverage	100 m
r_1, r_2, r_3 (WiFi network)	50, 75, 90 (meters)
τ_1, τ_2, τ_3	(−56), (−62), (−64) dBm
t_M (UMTS network)	12 sec
t_M (Wi-Fi network)	5 sec
Traffic type	CBR
Packet size	512 Bytes

buffer of BS is dynamically updated for supporting the maximum number of handovers and MNs.

When the MN gets to the boundary B_2 before the expiration time, it generates a link going down event (MIH_LINK_GOING DOWN) available in the MIH standard [25]. When BS/AP receives MIH_LINK_GOING DOWN event message, it checks the MN ID in the handover information table and sends the entry against it in the NEW BS ID field to the MN. In this way, the MN will not wait until the BS/AP forwards this event to the MIIS server as the BS/AP already has the information of new BS. Therefore, the time required to discover the new network and switching to it can be significantly reduced by bringing information of the new network close to the MN.

To elaborate the handover triggering procedure in our scheme, a scenario is illustrated in Figure 2. When the MN gets to the boundary B_1, it sends MIH_LINK_INFO event to AP/BS. This AP/BS is further connected to the gateways (GW) of the AN. MIIS server is available in core network, and the core network gateway is connected to the Internet. Directed arrows in Figure 2 show the flow of data in a particular direction.

Furthermore, we divide the handover region into three parts: handover starting point, nearby BS info point, and successful handover point. These points are synchronized between MN and BS using flags in the header of the packet. When the MN crosses a point, it switches on the required flag in a packet after that point. This synchronization is used for identification of successful handover. Now, the proposed scheme can efficiently perform fast handover even if the available handover region is small enough. And utilizing the effect of these points, we can design a topology for future networks.

4. Simulations and Results

The proposed scheme is compared with the existing MIH standard. We used a simulation scenario which consists of one universal mobile telecommunications system (UMTS) network (consists of 5 BS) and one Wi-Fi network (consists of 20 APs). Initially, an MN is attached to the UMTS. Then, the MN performs a handover from UMTS to Wi-Fi network. First, an MN performs a handover form UMTS to WiFi using the MIH standard events. Secondly, the MN performs a handover using the proposed scheme. Finally, different number of handovers is performed from Wi-Fi to UMTS network. Simulation scenario is shown in Figure 3, where each scenario uses different number of MNs.

The MIIS server is placed a number of hops away from both networks. Four different scenarios were tested with different number of hops. The MIIS server is 2, 3, 6, and 9 hops away, respectively, in four scenarios. Similarly, MNs moving in different directions were tested. Different numbers of handovers were performed by different number of MNs. Table 1 represents some important parameters used in the simulation. Handover time, delay, throughput, and network load were measured for different scenarios against the number of MNs, the number of handovers and simulation time, respectively.

Figure 4 depicts the comparison of the proposed scheme against the MIH standard in terms of the average handover time taken by different MNs. We can see that the proposed scheme efficiently reduces the handover time taken by an MN. It is also observed, from this experiment as the number of MNs increases, that the handover time taken by MIH standard also increases linearly. For example, the proposed scheme reduces the handover time by 25% or 15% when the numbers of nodes are 35 or 50, respectively. This improvement is mainly because the handover information required by the MN during handover process is available one hop away and hence the MN requires less time to obtain this information from the current BS/AP.

Figure 5 delineates the comparison of the proposed scheme against the MIH standard in terms of the average throughput. We tested 50 MNs having the simulation time of 2 hours in which different numbers of handovers are performed at once. We reduce the packet size from 512 bytes to 256 bytes in order to simulate the proposed scheme for longer period of time. For the sake of clarity, we use boxes at the places where handover is performed. It is shown that the proposed scheme experiences less packet loss as compared with the MIH standard that results in higher throughput. In addition, the proposed scheme also smoothly retains the throughput by redirecting the traffic from the old BS to the new BS in an efficient manner at the time when handover occurs. On the other hand, MIH standard take a significant amount of time to redirect the traffic from the old BS to the new BS.

Figure 6 depicts handover delay of the proposed scheme against the MIH standard as the number of hops is varied. Handover delay in 2, 3, 6, and 9 hops scenarios is shown in Figure 6. It is seen that the handover delay increases as we increase the number of hops between a BS and an MIIS server. A high delay time is observed as the number of hops is increased from 6 to 9 in the MIH standard. But, in the proposed scheme, the handover delay is approximately constant since the information is always available one hop away from an MN.

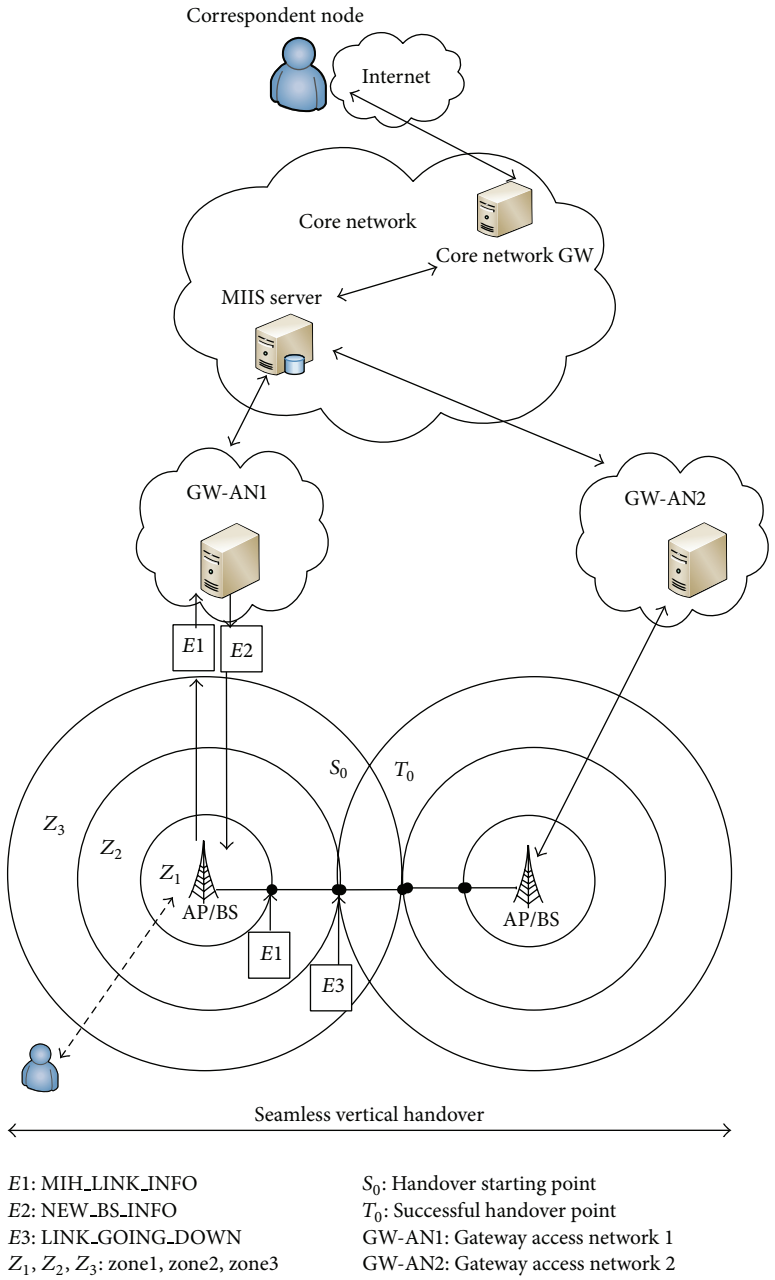


Figure 2: Operation of the proposed scheme.

Figure 7 illustrates the handover delay required by MN when the network load increases. In case of proposed scheme, the increase in handover delay is very low as compared to MIH standard. The proposed scheme obtains the available information of new network when the MN gets to the boundary B_1. The time required for handover when MN gets to boundary B_2 is very less in case of proposed scheme. Therefore, the proposed scheme performs better even if the network load increases due to increase in number of handovers by high number of MNs. Figure 7 also depicts that as the number of MNs in a particular region increases; then the proposed scheme shows better result in case of handover latency against the MIH standard.

5. Conclusions and Future Work

When an MIIS server is many hops away from a BS/AP, it requires long time to discover a new network in the MIH standard. The proposed scheme brings the information of the new network closer to the MN and can significantly reduce the time for discovering new network than the standard MIH.

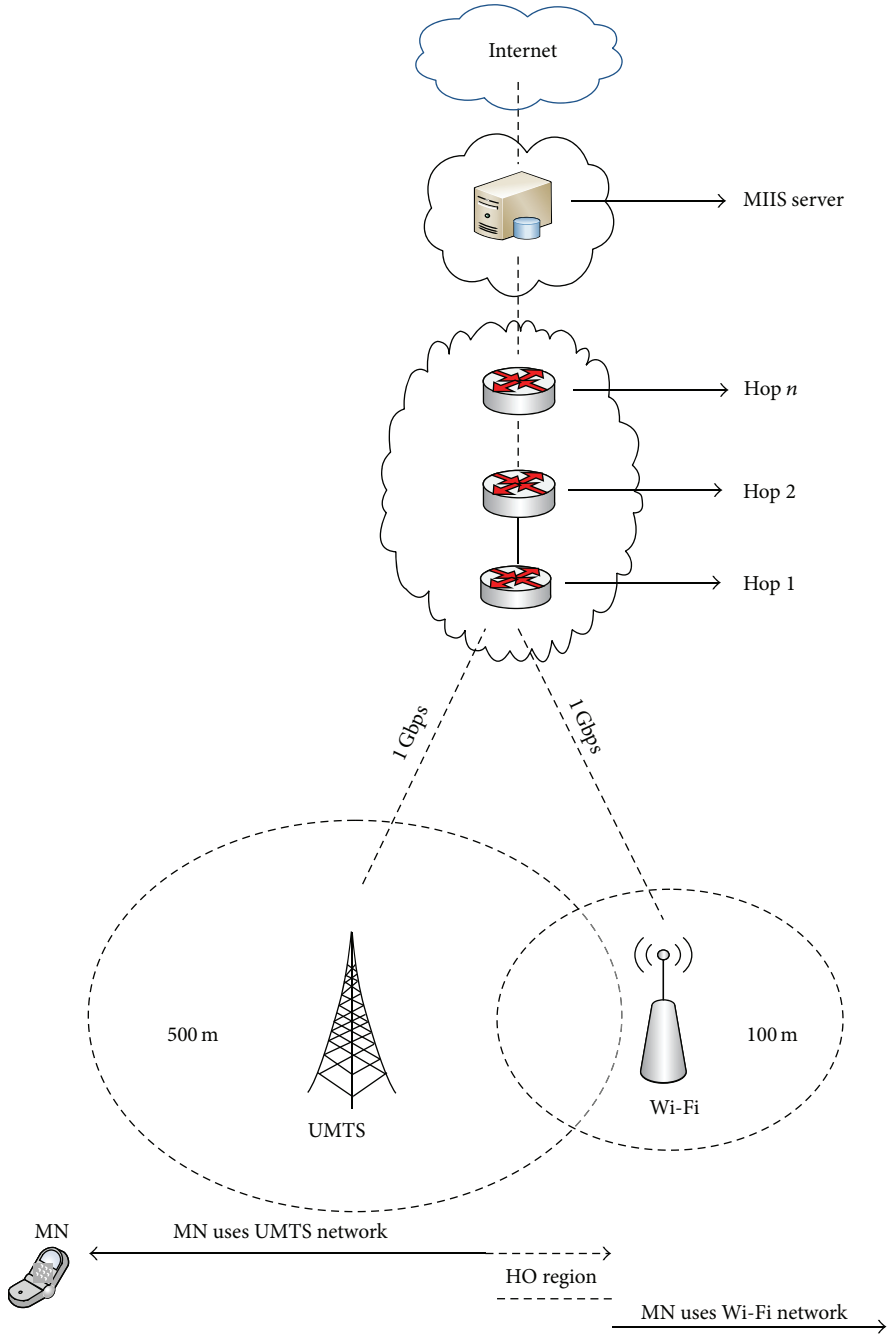


Figure 3: Simulation scenario.

A more sophisticated scheme can be designed while using the previous handover information for future handovers by an MN.

The threshold mechanism used in MIH standard is not enough to perform better in case of high network load and MNs. The triggering mechanism used by MIH standard is also not supportable when the number of MNs increases. Our proposed scheme performs better in case of both high network traffic and MNs.

The experimental results show that the proposed scheme saves 10% to 35% of the handover time as compared with the MIH standard. Furthermore, bringing the information of new BS from MIIS server to the associated BS is possible over only one hop in the proposed scheme. Similarly, the results and simulation show that, as the number of MNs is increasing, the handover delay and time are slightly increasing in case of proposed scheme. The packet loss ratio is significantly reduced which shows the accuracy of the proposed scheme. The network load is increasing as more numbers of MNs are injected in the scenario, but in case of proposed scheme, the network load remains balanced and tolerable. This trade-off between different parameters shows

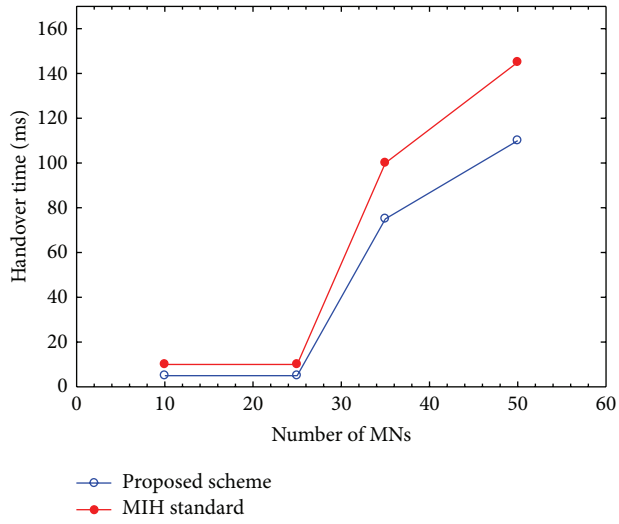


Figure 4: Handover time as the number of MNs is varied.

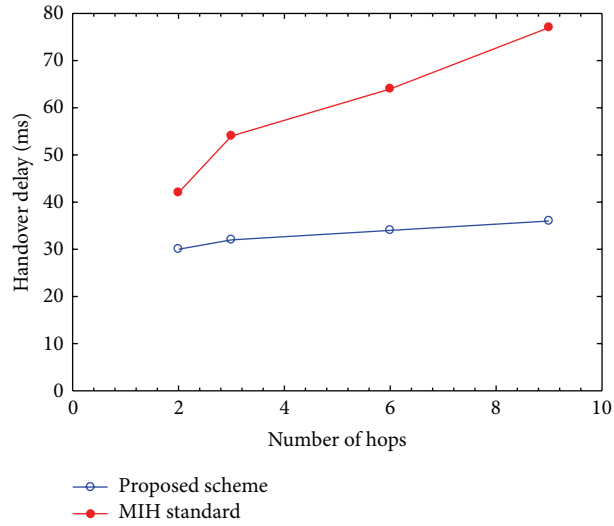


Figure 6: Handover delay with different number of hops.

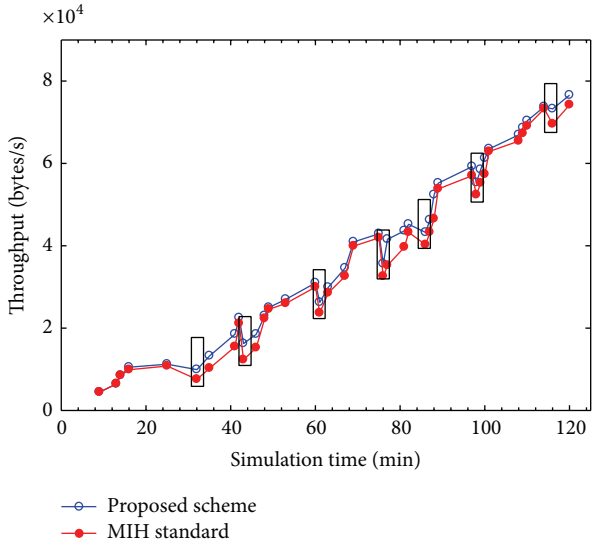


Figure 5: Throughput performance.

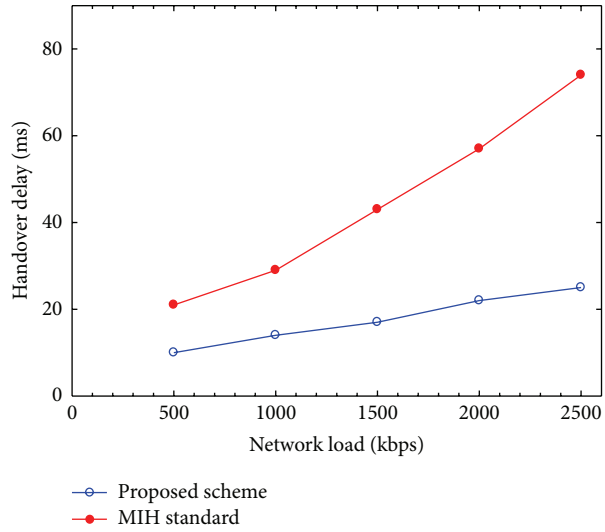


Figure 7: Handover delay as the network load varied.

that the proposed scheme performs better than the MIH standard and it can be easily adopted for the next generation of networks.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the IT R&D Program of MSIP/KEIT (10041145, Self-Organized Software Platform (SoSp) for Welfare Devices). This work was supported by the BK21 Plus Project (SW Human Resource Development Program for Supporting Smart Life) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Republic of Korea (21A20131600005).

References

[1] R. Tamijetchelvy and G. Sivaradje, "An optimized fast vertical handover strategy for heterogeneous wireless access networks based on IEEE 802.21 media independent handover standard," in *Proceedings of the 4th International Conference on Advanced Computing (ICoAC '12)*, pp. 1–7, Chennai, India, December 2012.

[2] S. J. Bae, M. Y. Chung, and J. So, "Handover triggering mechanism based on IEEE 802.21 in heterogeneous networks with LTE and WLAN," in *Proceedings of the International Conference on Information Networking (ICOIN '11)*, pp. 399–403, Kuala Lumpur, Malaysia, January 2011.

[3] G. Ciccarese, M. de Blasi, P. Marra et al., "Vertical handover algorithm for heterogeneous wireless networks," in *Proceedings of the 5th International Joint Conference on INC, IMS and IDC (NCM '09)*, pp. 1948–1954, Seoul, Republic of Korea, August 2009.

[4] A. A. Bathich, M. D. Baba, and M. Ibrahim, "IEEE 802.21 based vertical handover in WiFi and WiMAX networks," in *Proceedings of the IEEE Symposium on Computers and Informatics (ISCI '12)*, pp. 140–144, Penang, Malaysia, March 2012.

[5] ITU, http://www.itu.int/pub/R-QUE-SG07.

[6] IEEE 802.21: Media Independent Handover, January 2009, http://www.ieee802.org/21/.

[7] A. de la Oliva, A. Banchs, I. Soto, T. Melia, and A. Vidal, "An overview of IEEE 802.21: media-independent handover services," *IEEE Wireless Communications*, vol. 15, no. 4, pp. 96–103, 2008.

[8] C. Christakos and P. D. Allen, "A scalability and performance analysis of preauthentication algorithms for wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 7, pp. 3166–3176, 2012.

[9] Q. B. Mussabbir, W. Yao, Z. Niu, and X. Fu, "Optimized FMIPv6 using IEEE 802.21 MIH services in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 6, pp. 3397–3407, 2007.

[10] A. Dutta, D. Famolari, S. Das et al., "Media-independent pre-authentication supporting secure interdomain handover optimization," *IEEE Wireless Communications*, vol. 15, no. 2, pp. 55–64, 2008.

[11] P. Neves, J. Soares, and S. Sargento, "Media independent handovers: LAN, MAN and WAN scenarios," in *Proceedings of the IEEE Globecom Workshops*, vol. 15, pp. 1–6, December 2009.

[12] A. B. Pontes, D. dos Passos Silva, J. Jailton Jr., O. Rodrigues Jr., and K. L. Dias, "Handover management in integrated WLAN and mobile WiMAX networks," *IEEE Wireless Communications*, vol. 15, no. 5, pp. 86–95, 2008.

[13] L. Eastwood, S. Migaldi, Q. Xie, and V. Gupta, "Mobility using IEEE 802.21 in a heterogeneous IEEE 802.1 6/802.11-based, IMT-advanced (4G) network," *IEEE Wireless Communications*, vol. 15, no. 2, pp. 26–34, 2008.

[14] G. Lampropoulos, A. K. Salkintzis, and N. Passas, "Media-independent handover for seamless service provision in heterogeneous networks," *IEEE Communications Magazine*, vol. 46, no. 1, pp. 64–71, 2008.

[15] A. B. Pontes, D. dos Passos Silva, J. Jailton Jr., O. Rodrigues Jr., and K. L. Dias, "Handover management in integrated WLAN and mobile WiMAX networks," *IEEE Wireless Communications*, vol. 15, no. 5, pp. 86–95, 2008.

[16] S. Pack, J. Choi, T. Kwon, and Y. Choi, "Fast handoff support in IEEE 802.11 wireless networks," *IEEE Communications Surveys & Tutorials*, vol. 9, no. 1, pp. 2–12, 2007.

[17] A. Stephane, A. Mihailovic, and A. H. Aghvami, "Mechanisms and hierarchical topology for fast handover in wireless IP networks," *IEEE Communications Magazine*, vol. 38, no. 11, pp. 112–115, 2000.

[18] A. E. Xhafa and O. K. Tonguz, "Dynamic priority queueing of handover calls in wireless networks: an analytical framework," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 5, pp. 904–916, 2004.

[19] F. Buiati, L. J. Garcia Villalba, D. Corujo, J. Soares, S. Sargento, and R. L. Aguiar, "Hierarchical neighbor discovery scheme for handover optimization," *IEEE Communications Letters*, vol. 14, no. 11, pp. 1020–1022, 2010.

[20] K. Samdanis and A. H. Aghvami, "Scalable inter-area handovers for hierarchical wireless networks," *IEEE Wireless Communications*, vol. 16, no. 6, pp. 62–68, 2009.

[21] A. Vulpe, O. Fratu, and R. Craciunescu, "Performance evaluation of heterogeneous interworking using IEEE 802.21," in *Proceedings of the 20th Telecommunications Forum (TELFOR '12)*, pp. 498–501, Belgrade, Serbia, November 2012.

[22] S. Benoubira, M. Frikha, S. Tabbane, and K. Ayadi, "Vertical handover based on IEEE802.21 and mobile IPv6 in UMTS/WLAN networks," in *Proceedings of the 1st International Conference on Communications and Networking (ComNet '09)*, pp. 1–6, Hammamet, Tunisia, November 2009.

[23] G. Lampropoulos, A. K. Salkintzis, and N. Passas, "Media-independent handover for seamless service provision in heterogeneous networks," *IEEE Communications Magazine*, vol. 46, no. 1, pp. 64–71, 2008.

[24] A. de la Oliva, I. Soto, A. Banchs, J. Lessmann, C. Niephaus, and T. Melia, "IEEE 802.21: media independence beyond handover," *Computer Standards & Interfaces*, vol. 33, no. 6, pp. 556–564, 2011.

[25] IEEE 802.21, "IEEE Standard for Local and metropolitan area networks—media independent handover services," Institute of Electrical and Electronics Engineers, New York, NY, USA, May 2008.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 527307, 10 pages
http://dx.doi.org/10.1155/2014/527307

Research Article

Design of a Circularly Polarized *Z*-Slot Antenna with Isotropic Pattern for the UHF RFID Reader of WSN

Jongan Park,[1] Jonghun Chun,[2] Gwangwon Kang,[3] Sungkwan Kang,[1] and Youngeun An[4]

[1] *Department of Information & Communications Engineering, Chosun University, Gwangju 501-759, Republic of Korea*
[2] *Department of Information & Communications Engineering, Jeonnam Provincial College, 152 Juknokwonro Damyang-eup, Damyang-gun, Jeollanam-do 517-802, Republic of Korea*
[3] *Department of Green Transportation, Honam Institute for Regional Program Evaluation (HIRPE), Gwangju 500-706, Republic of Korea*
[4] *The Division of Undeclared Majors, Chosun University, Gwangju 501-759, Republic of Korea*

Correspondence should be addressed to Jonghun Chun; jhchun@dorip.ac.kr

Received 19 December 2013; Accepted 25 February 2014; Published 4 June 2014

Academic Editor: Jongsung Kim

In an antenna for a UHF RFID reader of wireless sensor networks (WSN), receiver sensitivity in sensing multitags from remote distances is an important performance index. This study designed a dual structured *Z*-slot antenna with optimized receiver sensitivity to enhance the sensitivity to a circularly polarized antenna with an isotropic pattern for a UHF RFID. Through analysis of performance in the designed antenna, the following was verified: return loss (S_{11}) was about −62.21 dB at 925.25 MHz, antenna gain was 7.36 dBi, and ΔP_r, isotropic gain deviation, was 1.3 dB. Impedance matching was about 50.069 Ω at 925.25 MHz and VSWR was from 1.001 to 1.028. Through this research it was discovered that this can be applied to the design of all RFID readers of WSN. Based on the above results, it is suggested that a circularly polarized *Z*-slot antenna which can enhance receiver sensitivity over a wide range can be widely applied to UHF RFID readers of WSN.

1. Introduction

RFID is an electronic tag and detection system that confirms information on things and detects surrounding conditions as one of the several tasks. A variety of radio frequencies and techniques are used in RFID systems [1, 2].

There have been a variety of studies in using the HF band (13.56 MHz), the UHF band (860 ~ 960 MHz), and the ISM band (2.4 GHz) as the standard RFID frequency bands for WSN [3]. As HF band RFID communication uses magnetic coupling, the receiver area of the antenna is very narrow and as the ISM band RFID is sensitive to the surrounding environment, the performance and sensitivity of the RFID system varies [1, 2]. However, the UHF band is the most outstanding for recognition rates and distance, and many tags can be recognized quickly using the radiation of electric waves. Also, as the signal is very stable in the surrounding environment and the tags and tag chips for this frequency can be produced at low prices, it is known that this is the most appropriate frequency for the realization of RFID technology and sensor networks [4, 5].

However, for a plane polarized antenna in the UHF band, errors may occur due to reflection and the interference of signals when multipath signals from the RFID tag are detected, and to recognize tags, the location of the tag has to be in line with the polarized plane [6, 7]. That is, for the plane polarized antenna, as electronic waves are radiated in a linearly polarized way, it is hard to recognize many tags simultaneously according to the locations of tags and the directionality of tag antennas and multipaths can be causes of disruptions [8]. To minimize errors due to multipath signals and gain high recognition rates regardless of the directions tag antennas face, a circularly polarized antenna is used. For a circularly polarized antenna, when two right-angled signals with different status and the same amplitude are polarized, electromagnetic waves are radiated circularly [9, 10]. This is more effective for recognizing several tags regardless of the locations of the tags and the directions of the antenna. That is,

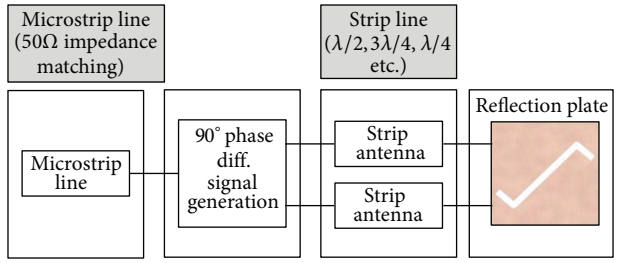


Figure 1: Composition of antennas for the UHF RFID reader of WSN designed.

for a directional pattern antenna, gain is high but recognition ranges of tags are low [11]. For an isotropic pattern, antenna gain is low but the recognition range of tags is wide, so studies to enhance receiver sensitivity with isotropic antenna over a wide range have been carried out [12–15].

This study designed an isotropic antenna with high receiver sensitivity and an isotropic pattern for UHF RFID. To design and manufacture circularly polarized isotropic antennas, the signals of readers were divided into two signals of the same magnitude and an incidence difference of 90°. Electricity was supplied to a strip line with a length of $\lambda/2$. The radiation board was connected to the feeding point of the strip line to enhance receiver sensitivity and the Z-slot was designed in a section of the radiation board to enhance receiver sensitivity so that passive tags could be recognized over a wide range.

This study is composed of the following: Section 1 describes the design of circularly polarized antennas with an isotropic pattern for UHF RFID readers of WSN, Section 2 describes the design, manufacture, and simulation of the antennas for readers, Section 3 measures the parameters of the antennas manufactured and their performance, and Section 4 concludes the antennas designed.

2. Design and Simulation of an Antenna for Readers

We have designed and manufactured a circularly polarized antenna and Z-slot antenna with isotropic pattern.

The processes are as follows.

2.1. Design and Manufacture of the Proposed Antenna. The block diagram of circularly polarized antennas for UHF RFID readers of WSN with an isotropic pattern is presented in Figure 1. For compatibility with different RFID readers a microstrip line was designed as shown in Figure 2. The inductance value (L_1) was adapted and impedance matching was 50 Ω.

As seen in Figure 2, a hybrid coupler with insertion loss of 0.16 dB and a status difference of 90° was used to give a status difference of 90° from the original signal. To divide signals into two right-angled linear signals with the same amplitude, two strip lines with a length of $\lambda/2$ were designed. We called the shape the dual structured antenna. The circularly polarized antenna designed was characterized by left-turn circular polarization (LHCP). For the PCB, a 1 mm thick FR4 epoxy substrate with a relative permittivity of 4.8 was used. To receive the minute power reflected from tags, two strip line feeding points with a length of $\lambda/2$ and a Z-slot radiation plate with cross section of 13.4 cm × 13.4 cm were used for a double antenna.

To achieve an isotropic pattern, the diagonal length of the radiation plate was $\lambda/4$ (80 mm), and length of each line (d) was 30 mm and the slot home span (t) was 6 mm as seen in Figure 3.

2.2. Simulated Results for Radiation Pattern Characteristics. We used the ADS (advanced design system) 2004A of Agilent company for simulating the manufactured antenna. The simulation direction of the designed circular polarization antenna is as shown in Figure 4. The simulation result presented that it had high power density at 0360° direction to x-, y-, and z-axis (isotropic pattern) as shown in Figure 5.

3. Parameter Measurement of the Designed Antenna and Performance Analysis

We measure and analyze the performance of the designed antenna as shown in the following.

3.1. Measurement of Return Loss (S_{11}). If there are unmatched impedance points in the transmission system, power reflection occurs there and part of the input power is reflected. Here, the ratio of the input power to reflected power is the return loss.

Measurement results for the return loss (S_{11}) of the designed UHF RFID reader of WSN are presented in Figure 6 and as shown in Table 1, the return loss decreased at 925.25 MHz to −62.21 dB.

The wideband of the return loss of −10 dB was between 720.25 MHz and 1.12 GHz and the wideband of the axis rate of 3 dB was within 1.2 dB.

3.2. Measurement of the Impedance Matching Parameter. As a result of measuring the matching parameters of the designed antenna, matching was carried out with 50.069 Ω at 925.25 MHz as shown in Figure 7 and VSWR was from 1.001 to 1.028. It was discovered that this is compatible with different RFID readers of WSN. The impedance of the designed antenna is presented in Table 2.

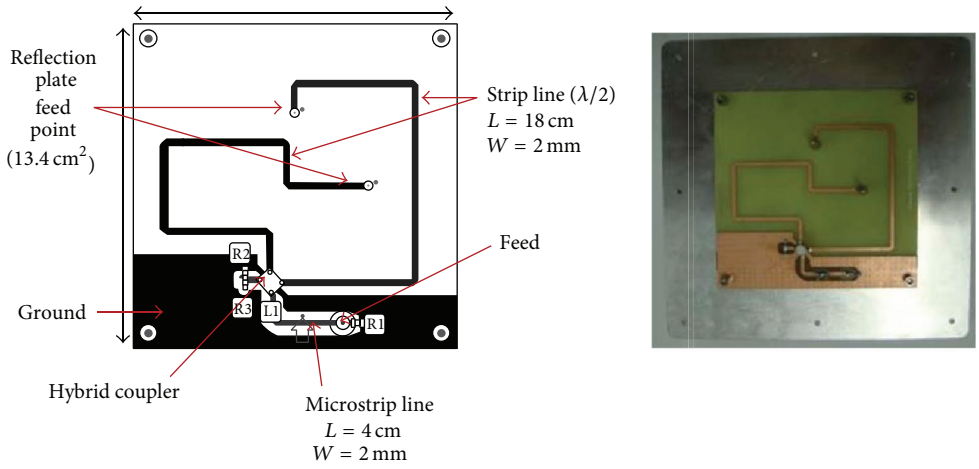


Figure 2: Design and manufacture of the PCB.

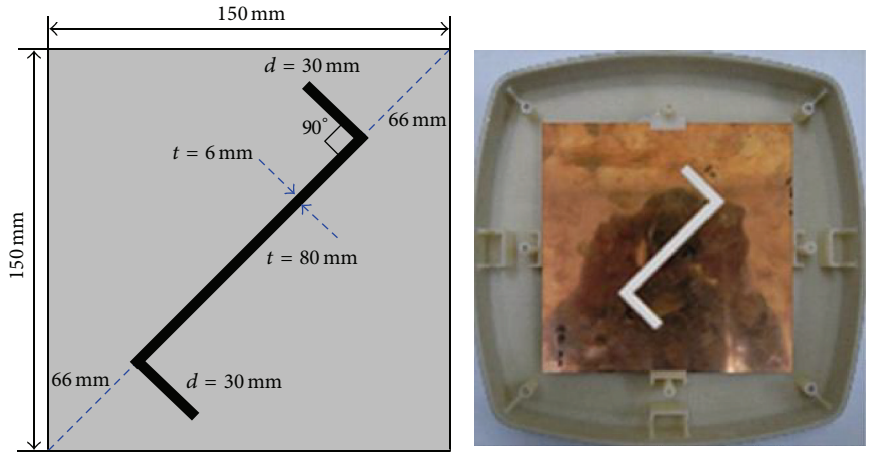


Figure 3: Design and manufacture of a Z-slot reflection plate.

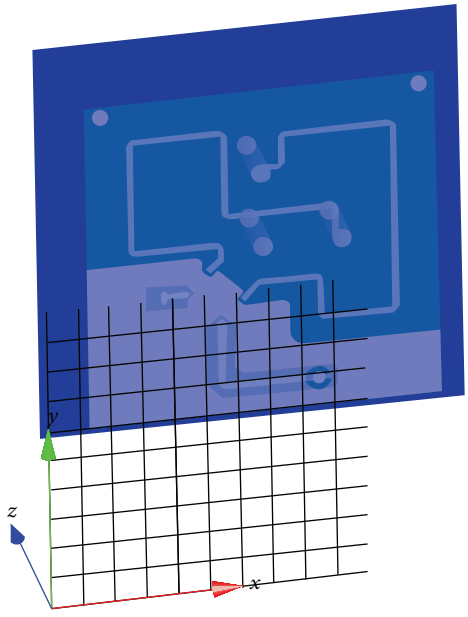


Figure 4: Simulation of the direction of the designed antenna.

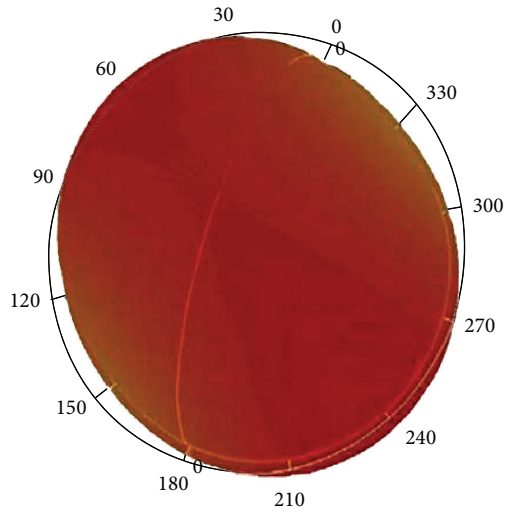


Figure 5: 3D simulation results of the z-axis radiation pattern.

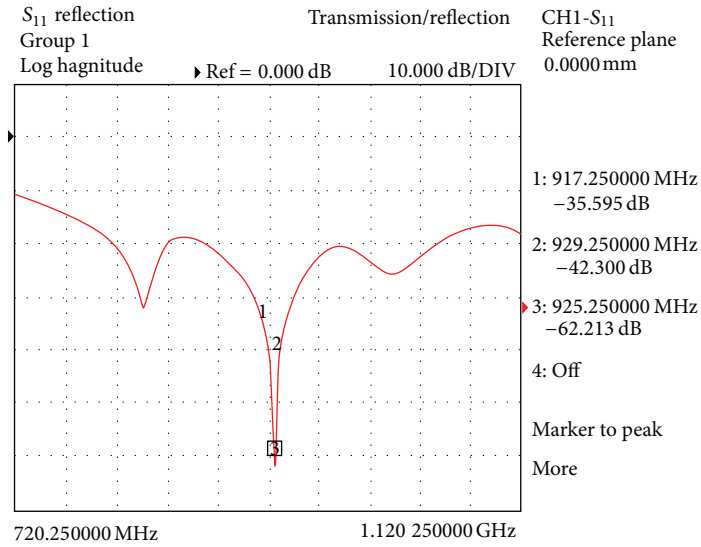


Figure 6: Return loss (S_{11}) of a manufactured UHF RFID antenna @span 400 MHz.

Table 1: Return loss (S_{11}) of the designed antenna.

Frequency	Return loss (S_{11})
917.25 MHz	−35.595 dB
925.25 MHz	−62.213 dB
929.25 MHz	−43.300 dB

Table 2: Measured impedance values.

Frequency	Impedance
917.25 MHz	51.415 Ω
925.25 MHz	50.069 Ω
929.25 MHz	49.486 Ω

3.3. Test Environment and Radiation Pattern Measurements

3.3.1. Test Environment for Antenna Radiation Pattern. The designed antenna radiation pattern was measured by an absorber system. To measure performance, the circularly polarized Z-slot antenna was placed as a target, as seen in Figure 8, and a horn antenna was used as a standard for comparison. The two antennas had a separation distance of 2.36 m. For the standard, a BBHA-9120-D made by SCHWARZBECK was used and the parameters are as presented in Table 3.

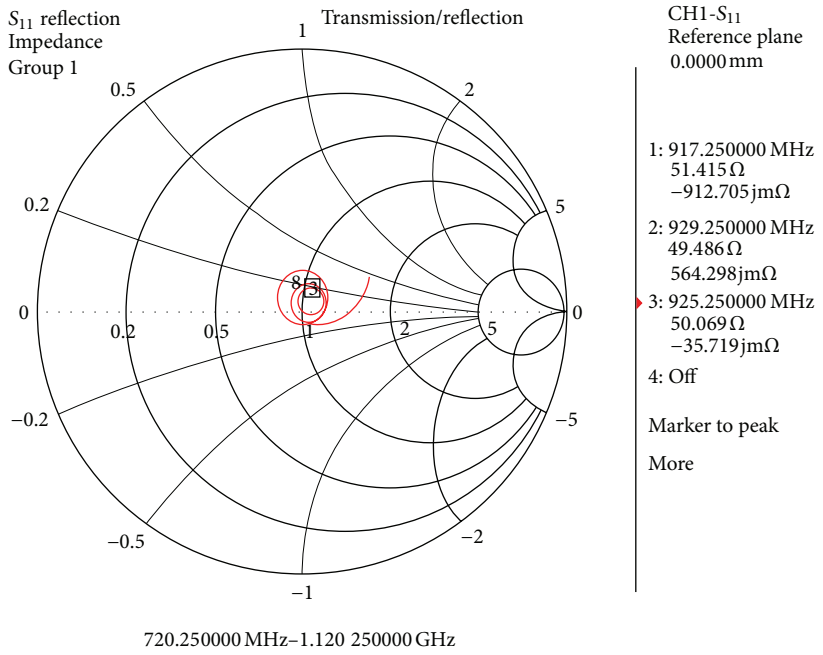


Figure 7: Measurement results of impedance parameters.

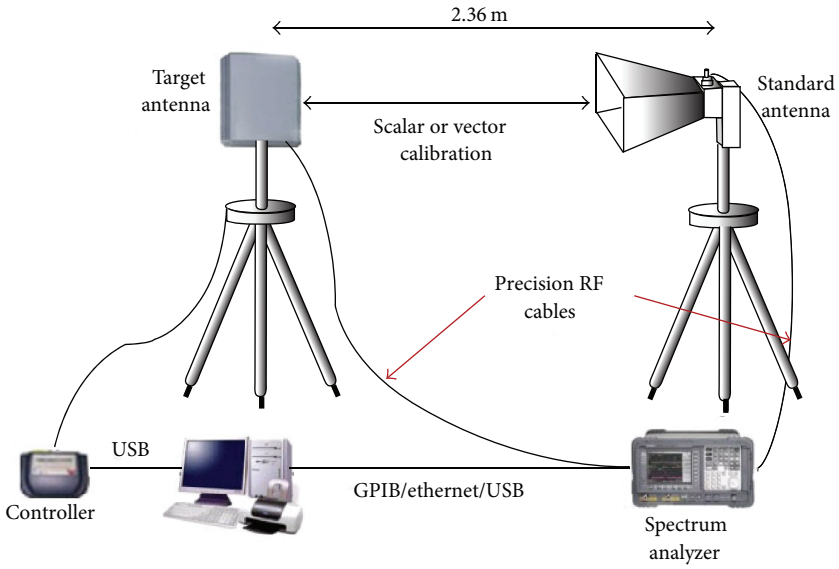


Figure 8: Test environment for the radiation pattern.

The equation defined by H. T. Friis which describes this wave behavior in "free space," called the Friis Transmission Equation, is [16–18] as follows:

$$P_r = \left(\frac{\lambda}{4\pi d}\right)^2 G_t G_r P_t, \tag{1}$$

where P_r represents received power level, P_t represents transmit power level, λ represents transmit wave length, G_t represents gain of the transmit antenna, G_r represents gain of the receive antenna, and d represents separation distance between antennas.

It is convenient to express Friis formula in terms of $S_{21}^2 = P_r/P_t$ and dB:

$$S_{21\,\mathrm{dB}} = P_L^{\mathrm{dB}} + G_t^{\mathrm{dB}} + G_r^{\mathrm{dB}}, \tag{2}$$

where the path loss is defined as

$$P_L^{\mathrm{dB}} = 20 \log\left(\frac{\lambda}{4\pi d}\right). \tag{3}$$

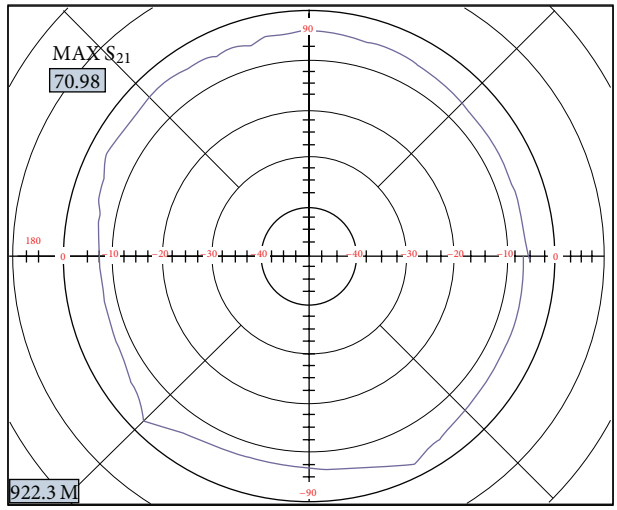


Figure 9: Radiation pattern for the antenna gain 1 (the standard antenna: horizontal and the designed antenna: frontal).

Table 3: Parameters of the standard antenna.

Frequency (f)	Distance (d)	Wavelength (λ)
925 MHz	2.36 m	0.324 m
Att. (dB)	Gain (Isotr.)	Ant.-factor
26.8 dB	5.09 dBi	24.21 dB/m

3.3.2. Measurements of Antenna Radiation Pattern and the Gain Parameter. To measure the radiation pattern and antenna gain according to the directions of the designed antenna, the standard antenna and the designed antenna were rotated horizontally and vertically and the pattern and gain were measured in the four different aspects.

Figure 9 presents the results of the test where the standard antenna was set horizontally and the designed antenna was set in front. The antenna gain was 5.28 dBi and ΔP_r, a deviation of isotropic gain, was 3.2 dB.

Figure 10 presents the results of the test where the standard antenna was set vertically and the designed antenna was set in front. The antenna gain was 5.83 dBi and ΔP_r, the deviation in isotropic gain, was 7.2 dB.

Figure 11 presents the results of the test where the standard antenna was set horizontally and the designed antenna was turned at 90°. The antenna gain was 7.36 dBi and ΔP_r, a deviation of isotropic gain, was 1.3 dB.

Figure 12 presents the results of the test where the standard antenna was set vertically and the designed antenna was turned to 90°. The antenna gain was 7.46 dBi and ΔP_r, a deviation of isotropic gain, was 16.4 dB.

Table 4 presents the results of comparing the antenna gains measured and the deviations of the isotropic gains under the four different test environments. The gain of the antenna manufactured through a Z-slot formed on the radiation plate was about 7.36 dBi. The largest gain is a little low, but the deviation in isotropic gain was 1.3 dB, which indicates that the radiation pattern is superior.

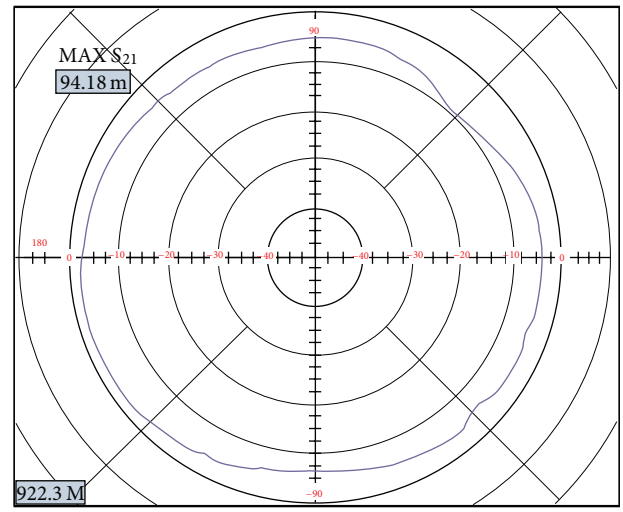


Figure 10: Radiation pattern for the antenna gain 2 (the standard antenna: vertical and the designed antenna: frontal).

Table 4: Measured antenna gains and the deviations of the isotropic gains under the four different test environments.

	Standard antenna direction	Designed antenna direction	Antenna gain	Deviations of isotropic gain
Environment 1	Horizontal	Frontal	5.28 dBi	3.2 dB
Environment 2	Vertical	Frontal	5.83 dBi	7.2 dB
Environment 3	Horizontal	Turned 90°	7.36 dBi	1.3 dB
Environment 4	Vertical	Turned 90°	7.42 dBi	16.4 dB

3.3.3. Radiation Pattern Comparison according to the Z-Slot Plate Parameters. The design parameters of the Z-slot formed on the antenna radiation plate, that is, the diagonal length (L),

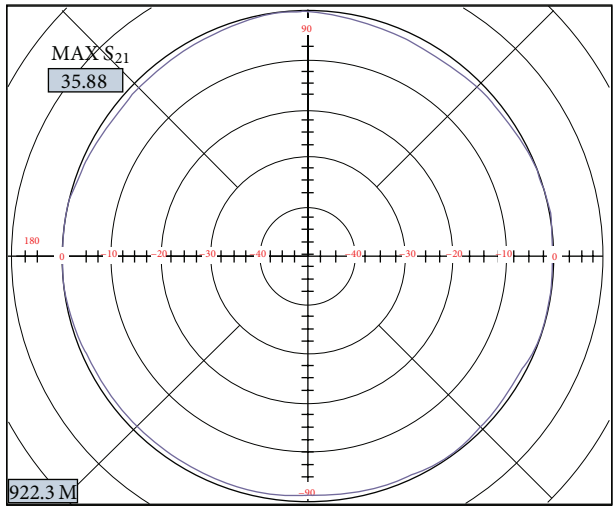


Figure 11: Radiation pattern for the antenna gain 3 (the standard antenna: horizontal and the designed antenna: turned 90°).

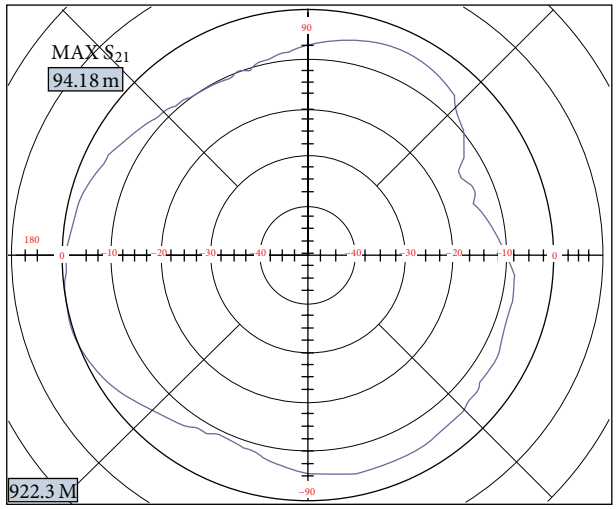


Figure 12: Radiation pattern for the antenna gain 4 (the standard antenna: vertical and the designed antenna: turned 90°).

the length of the side line (d), and the slot span (t), varied as seen in Table 5. Figure 13 is the designed antenna radiation plate.

Figure 14 shows the results of the test where the standard antenna with an unsatisfactory radiation pattern was set vertically and the designed antenna was set in front to measure the radiated power according to the Z-slot parameters. Figure 15 shows the results of the test where the standard antenna was set vertically and the designed antenna was turned to 90°. As a result of measuring the radiated power according to the Z-slot parameters, the larger the slot span was, the bigger the radiation gain was. When the slot span was larger than 6 mm resonance point, ΔP_r was higher. This indicates that the radiation pattern is not satisfactory. As a result

Table 5: Design parameters of the Z-slot radiation plate.

Antenna samples	Diagonal length (L) [mm]	Line length (d) [mm]	Slot span (t) [mm]
1	$\lambda/4$ (80)	$3\lambda/32$ (30)	2
2	$\lambda/4$ (80)	$3\lambda/32$ (30)	4
3	$\lambda/4$ (80)	$3\lambda/32$ (30)	6
4	$\lambda/4$ (80)	$\lambda/8$ (40)	2
5	$\lambda/4$ (80)	$\lambda/8$ (40)	4
6	$\lambda/4$ (80)	$\lambda/8$ (40)	6

of the measurements, when the diagonal length was $\lambda/4$, d, the line length, was $3\lambda/32$ (30 mm), and t, the slot span, was 6 mm, the isotropic pattern was the most satisfactory.

3.3.4. Analysis of the Radiation Pattern according to the Radiation Plate. To analyze radiation patterns according to the presence of the Z-slot on the radiation plate, the radiation pattern was measured with the standard antenna set vertically and the designed antenna turned 90°.

As a result, it was discovered that the maximum antenna gain without a Z-slot was about 9.09 dBi and ΔP_r was about 24 dB. In the case when there is a Z-slot on the plate, the maximum antenna gain was about 7.36 dBi and ΔP_r was about 1.3 dB. Therefore, the maximum antenna gain was about 1.73 dBi and ΔP_r, a deviation of isotropic gain, was about 22 dB. Therefore, it was discovered that the isotropic pattern was superior when the Z-slot was designed on the radiation plate and the receiver sensitivity of passive tags in the wide range of the antenna for the RFID reader of WSN was superior (Figure 16).

3.3.5. Performance Comparison. Table 6 shows the electric characteristics of the designed antenna with the RFID antenna. The known antenna was more than 21 cm^2 in size while the designed antenna is only about 13.4 cm^2. However, the antenna gain was the same as that of the known antenna, 5.28 ~ 7.36 dBi, although its size is smaller. The VSWR of the designed antenna was 1.028, higher than that of the known antenna. The tag recognition distance of the known antenna was about 3 m in a narrow range while that of the designed antenna was about 3 m in a wider range.

4. Conclusion

This study designed a dual structured and circularly polarized Z-slot antenna for UHF RFID readers of WSN with an isotropic pattern.

To recognize many tags at a remote distance at the same time without errors, the designed antenna had an insertion loss of 0.16 dB and used a hybrid coupler with a status difference of 90° so there was a status difference of 90° from the original signal. Also, to separate signals into two right-angled linear signals of the same amplitude, there were two strip lines with a length of $\lambda/2$. The designed dual antenna had a left-turn circularly polarized pattern and to receive

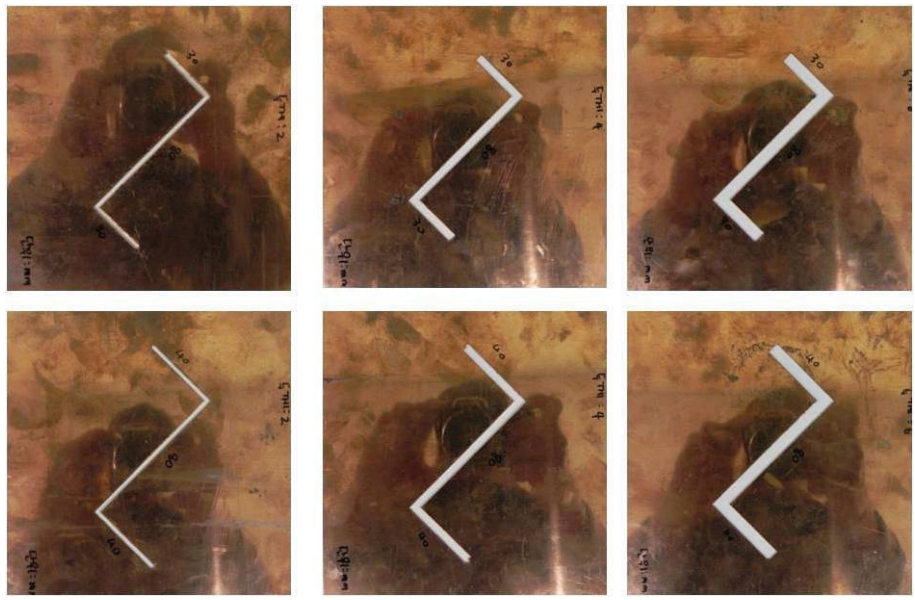


Figure 13: Antenna radiation plates according to the Z-slot parameters.

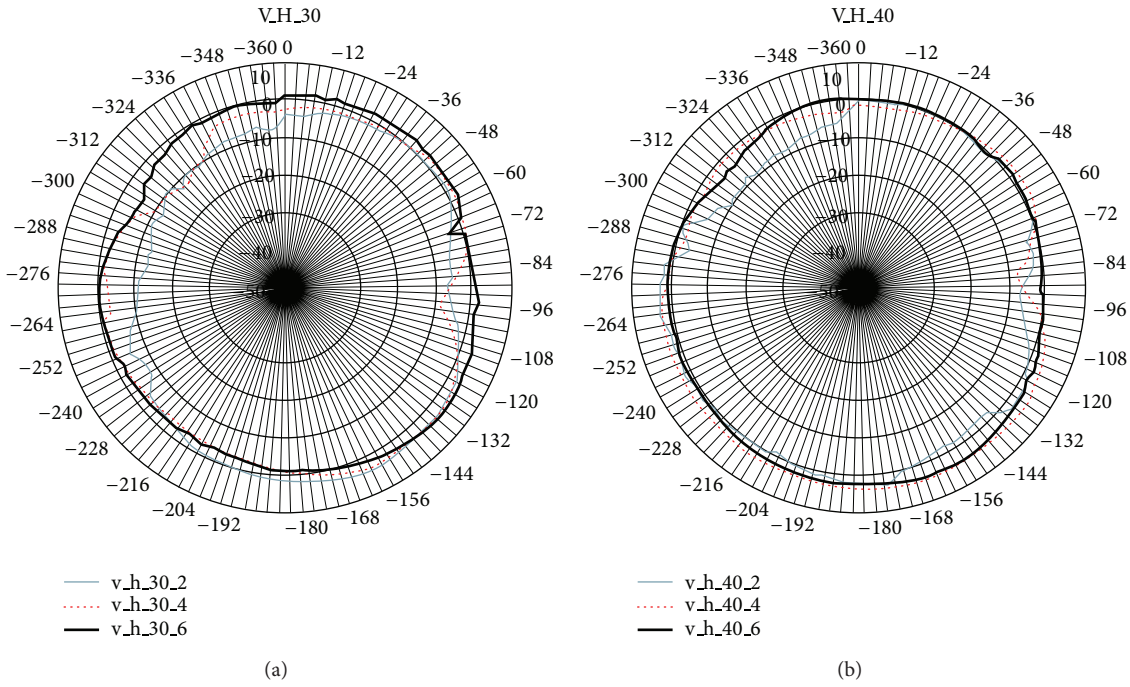


Figure 14: Radiation pattern data according to Z-slot parameters 1 (standard antenna: vertical and the designed antenna: frontal).

Table 6: Performance comparison of the known and the proposed antenna.

Manufacturing company	S Co.	S Co.	S Co.	P Co.	I Co.	I Co.	Manufactured antenna direction (Z-slot)
Antenna size	71.7 × 31.7 cm^2	22.4 × 20.6 cm^2	28.19 × 28.19 cm^2	24.5 × 24.5 cm^2	25.9 × 25.9 cm^2	21.8 × 19.8 cm^2	13.4 × 13.4 cm^2
Gain (dBi)	6.75	5.25	6	6.5 ± 0.5	7	6	5.28~7.36
VSWR	1.25	—	1.22	1.3	1.5	1.5	1.001~1.028
Tag recognition distance	Narrow zone about 3 m	Narrow zone about 3 m	Narrow zone about 3 m	Narrow zone about 3 m	Narrow zone 3 m diffusion	Narrow zone 3 m diffusion	Wide zone about 3 m

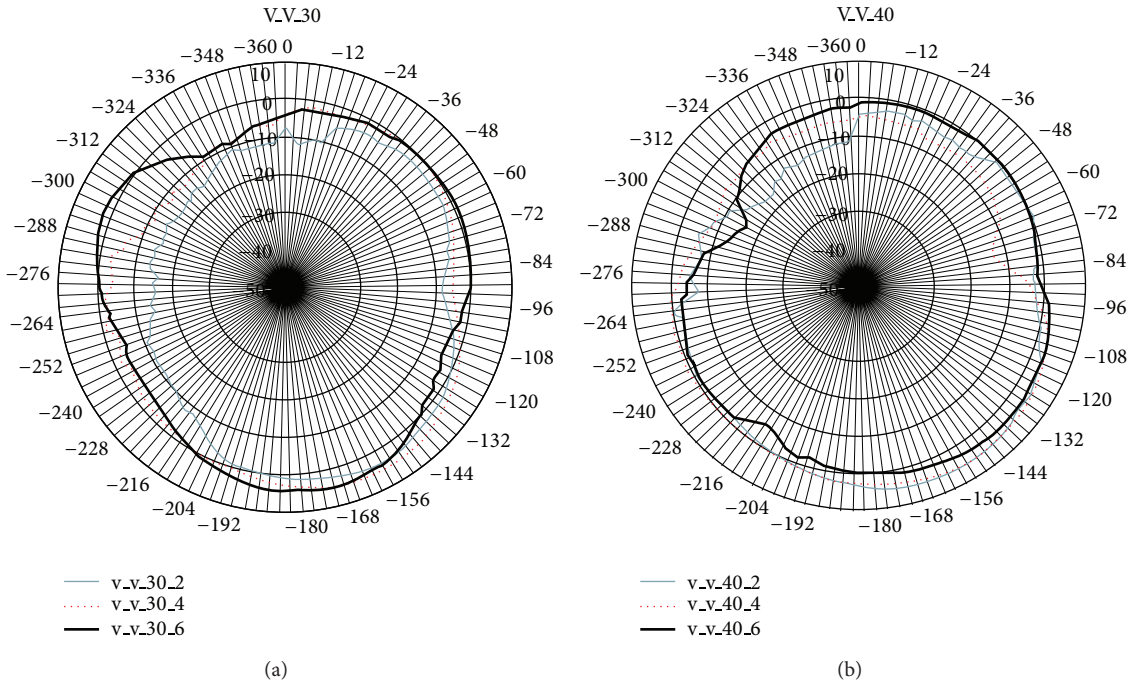


Figure 15: Radiation pattern data according to Z-slot parameters 2 (standard antenna: vertical and the designed antenna: turned 90°).

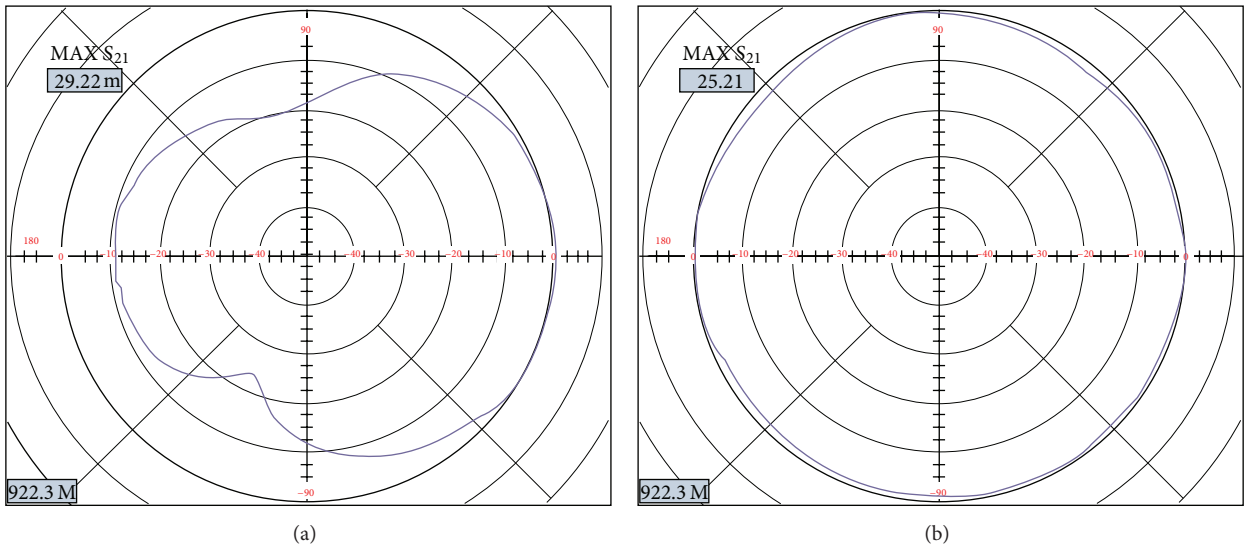


Figure 16: Radiation patterns with and without the Z-slot on the radiation plate (standard antenna: vertical and designed antenna: turned 90°).

the minute amount of power returned from the tags, it was designed to have strip line feeding points of length of $\lambda/2$ and a Z-slot radiation plate with a cross section of 13.4 cm × 13.4 cm. As a result of the test of the designed antenna, the return loss was about −62.213 dB at 925.5 MHz and the antenna gain was 7.36 dBi. Impedance matching was 50.069 Ω at 925.25 MHz and the VSWR was from 1.001 to 1.028, which indicates that it can be used for different RFID readers of WSN.

The designed antenna was half the size of the known antenna, but it had an isotropic radiation pattern as good as and with the same electrical characteristics as the known one and recognized passive tags over a wide range. Such a dual structured and circularly polarized Z-slot antenna with an isotropic pattern can contribute to enhanced receiver sensitivity through maximizing antenna efficiency and to the development of antennas for UHF RFID readers of WSN with respect to compatibility with other systems.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This study was supported by research fund from Chosun University, 2012.

References

[1] J. Landt, "The history of RFID," *IEEE Potentials*, vol. 24, no. 4, pp. 8–11, 2005.

[2] S. H. Rhee, J. A. Park, and J. H. Chun, "Sensitivity improvement of the receiver module in the passive tag based RFID reader," *Ubiquitous Intelligence and Computing*, vol. 4611, pp. 13–22, 2007.

[3] J. H. Chun, H. S. Chung, and J. A. Park, "A circular polarization antenna design for improving the reception of UHF RFID reader," *Wireless Personal Communications*, vol. 60, no. 3, pp. 389–404, 2011.

[4] "ISO/IEC 18000-7, International Standards," 2004.

[5] K. Finkenzeller, *RFID Handbook*, 2nd edition, 2004, http://www.youngjin.com/.

[6] D. M. Pozar and D. H. Schaubert, *Microstrip Antennas: The Analysis and Design of Microstrip Antennas and Arrays*, IEEE Press, New York, NY, USA, 1995.

[7] S. D. Targonski and D. M. Pozar, "Design of wideband circularly polarized aperture-coupled microstrip," *IEEE Transactions on Antennas and Propagation*, vol. 41, no. 2, pp. 214–220, 1993.

[8] W. Hong, N. Behdad, and K. Sarabandi, "Design of tri-band reconfigurable antenna for active RFID," in *Proceeding of the IEEE Antennas and Propagation Society International Symposium*, vol. 1, pp. 117–120, Albuquerque, NM, USA, July 2006.

[9] G. De Vita and G. Iannaccone, "Design criteria for the RF section of UHF and microwave passive RFID transponders," *IEEE Transactions on Microwave Theory and Techniques*, vol. 53, no. 9, pp. 2978–2990, 2005.

[10] A. Adrian and D. H. Schaubert, "Dual aperture-couple microstrip antenna for dual or circular polarization," *Electronics Letters*, vol. 23, no. 23, pp. 1226–1228, 1987.

[11] D. M. Pozar, "Microstrip antenna aperture coupled to a microstripline," *Electronics Letters*, vol. 21, no. 2, pp. 49–50, 1985.

[12] P. Sullivan and D. Schaubert, "Analysis of an aperture coupled micorstrip antenna," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 8, pp. 977–984, 1986.

[13] L. Ukkonen, L. Sydanheimo, and M. Kivikoski, "Read range performance comparison of compact reader antennas for a handheld UHF RFID reader," in *Proceeding of the IEEE International Conference on RFID*, vol. 1, pp. 63–70, Grapevine, Tex, USA, March 2007.

[14] D. Peroulis, K. Sarabandi, and L. P. B. Katehi, "Design of reconfigurable slot antennas," *IEEE Transactions on Antennas and Propagation*, vol. 53, no. 2, pp. 645–654, 2005.

[15] M. Hiyama, E. Kulla, T. Oda, M. Ikeda, and L. Barolli, "Application of a MANET testbed for horizontal and vertical scenarios: performance evaluation using delay and jitter metrics," *Human-Centric Computing and Information Sciences*, vol. 1, no. 3, pp. 1–14, 2011.

[16] W. L. Stutzman and G. A. Thiele, *Antenna Theory and Design*, John Wiley and Sons, New York, NY, USA, 2nd edition, 1998.

[17] Atmel Application Note, "Range Calculation for 300 MHz to 1000 MHz Communication Systems," http://www.atmel.com/Images/doc9144.pdf.

[18] D. C. Hogg, "Fun with the Friis free-space transmission formula," *IEEE Antennas and Propagation Magazine*, vol. 35, no. 4, pp. 33–35, 1993.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 125309, 6 pages
http://dx.doi.org/10.1155/2014/125309

Research Article

Practical RSA-PAKE for Low-Power Device in Imbalanced Wireless Networks

Taek-Young Youn,[1] Sewon Lee,[2] Seok Hie Hong,[2] and Young-Ho Park[3]

[1] *Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea*
[2] *Center for Information Security Technologies, Korea University, Seoul, Republic of Korea*
[3] *Department of Information Security, Sejong Cyber University, Seoul, Republic of Korea*

Correspondence should be addressed to Taek-Young Youn; taekyoung@etri.re.kr

Received 29 November 2013; Accepted 3 March 2014; Published 29 May 2014

Academic Editor: Jongsung Kim

For enhancing the security of ubiquitous communication, we have to consider three keywords: *mobility*, *wireless*, and *low computing capability*. In this paper, we study one of suitable security protocols for the ubiquitous communication environment. We discuss RSA-based password-authenticated key exchange (RSA-PAKE) protocols for imbalanced wireless networks where a party uses a low-power device to communicate with another party equipped with a powerful computing device. For imbalanced wireless network applications, it is important to reduce the cost of communication for a low-power device even though the cost for powerful devices is increasing. The most power-consuming operation in RSA-PAKE protocols is the reliability test of unauthorized RSA public keys. Hence, it is important to design an efficient reliability test method to construct an efficient RSA-PAKE protocol. In this paper, we propose a new reliability test technique and design a provably secure RSA-PAKE protocol using the technique. Our protocol is suitable for securing the communications conducted over imbalanced wireless networks since the operations computed by one communicating party are efficient enough to be implemented on most low-power devices such as mobile phones and PDAs. The cost of a low-power device is reduced by 84.25% compared with CEKEP, the most efficient RSA-PAKE protocol. We prove the security of our protocol under a firmly formalized security model.

1. Introduction

For enhancing the security of ubiquitous communication, we have to consider three keywords: *mobility*, *wireless*, and *low computing capability*. In this paper, we study one of suitable security protocols for the ubiquitous communication environment. We discuss RSA-based password-authenticated key exchange (RSA-PAKE) protocols for imbalanced wireless networks where a party uses a low-power device to communicate with another party equipped with a powerful computing device.

For securing the communications conducted over wireless network, password-authenticated key exchange (PAKE) protocols are needed since two parties can establish a session key without storing any sensitive information in mobile devices. Note that we have an interest in imbalanced wireless networks where two communicating parties have different computational capabilities. Generally, mobile devices have low capabilities. Hence, for imbalanced wireless networks, it is important to reduce the cost of communications for a low-power device even though the cost for a powerful device is increased. For PAKE protocols that have been designed based on the Diffie-Hellman key exchange protocol (DH-PAKE protocol), each communicating party should compute at least two exponentiations with 160-bit exponents (for 80-bit security). Hence, it seems hard to design a DH-PAKE protocol for imbalanced wireless network applications where a party uses a low-power mobile device for communications. For PAKE protocols that have been designed based on the RSA function (RSA-PAKE protocol), one of the two parties can establish a session key by performing a small number of encryptions. Hence, RSA-based PAKE protocols seem to be suitable for imbalanced wireless networks since the cost of the RSA function is imbalanced in the sense that the encryption operation is very efficient while decryption is not. For example, if we use 3 as the public exponent for the RSA function, the encryption operation requires 2 multiplications

while the decryption requires one exponentiation with a full-size (1024-bit) exponent.

For RSA-PAKE protocols, the existence of the public-key infrastructure (PKI) is not assumed, and thus a set of unauthorized public key pairs (n, e) is used without any authorized certificate. Therefore, we have to verify the reliability of the unauthorized RSA keys. Due to the additional cost for verifying the reliability of the keys, it is not easy to design an efficient RSA-PAKE protocol. Note that the most power-consuming operation in RSA-PAKE protocols is the reliability test of unauthorized RSA public keys. Hence, it is important to design an efficient reliability test method to improve the performance of an RSA-PAKE protocol. Until now, several researchers have tried to design efficient RSA-PAKE protocols [1–8]. However, they are not sufficiently efficient enough to be implemented on most of low-power devices. Hence, it will be valuable to design a new RSA-PAKE protocol which provides a more efficient key exchange for low-power devices than existing RSA-PAKE protocols.

The goal of this paper is to design a new RSA-PAKE protocol which is suitable for securing the communications conducted over imbalanced wireless networks. To design an efficient RSA-PAKE, we provide very simple and efficient conditions for testing the reliability of a set of RSA parameters. In our protocol, a low-power device can establish a session key by choosing a 52-bit prime and performing one exponentiation with the prime exponent. According to our experimental results, the cost of a client is reduced by 84.25% compared with the CEKEP, which is the most efficient RSA-PAKE protocol until now. Our protocol can be implemented more efficiently by generating the prime before a key exchange protocol is initiated. In this case, the cost of a client can be reduced by 88.46%. We prove the security of our protocol in the random oracle model under a firmly formalized security model.

2. Preliminary

In this section, we briefly review formal security models for RSA-PAKE protocols and some mathematical backgrounds.

2.1. Security Model. Let A and B be two communicating parties, and let $\mathcal{D}$ be the password space. We assume that A and B share a password $pw \in \mathcal{D}$. Let E be an active adversary who attacks the key exchange between A and B by controlling their messages. The adversary may capture transmitted messages and verify guessed passwords using the collected information until he/she finds the correct password. This type of attack is called an offline password guessing attack. The security goal of a password-authenticated key exchange protocol is to provide password-enabled key exchange which is secure against offline password guessing attacks mounted by the active adversary E. In this paper, we review the main points of well-formalized security models introduced by Bellare et al. [9]. (Refer to [9] for details.)

Adversarial Model. When a protocol is executed, each party behaves as specified in the protocol. For given queries, each instance returns its outputs. Let Π_A^i be the ith instance of A. Note that each instance may be used only once. Each instance has a session key sk, a session id sid, and a partner id pid. In general, the session id of Π_A^i is the ordered concatenation of all messages sent and received by Π_A^i. An adversary can make queries to any instance. When the instance returns its output to the adversary, the internal state of the instance is also updated. E can make the following queries for any instance.

(i) $\mathsf{Send}(A, i, m)$. E sends a message m to the instance Π_A^i. Then, Π_A^i executes as specified by the protocol and returns its response to E. If the instance accepts given m as a valid message, the acceptance of the message, the session id sid, and the partner id pid will be made visible to E. If the message is not accepted as a valid one, the instance terminates the oracle call, and the termination is also made visible to the adversary.

(ii) $\mathsf{Execute}(A, i, B, j)$. By this oracle call, E obtains a transcript of an honest execution between two instances Π_A^i and Π_B^j, where Π_A^i and Π_B^j are unused instances such that $A \neq B$.

(iii) $\mathsf{Reveal}(A, i)$. By this oracle call, sk_A^i is given to the adversary E, where sk_A^i is the session key of Π_A^i.

(iv) $\mathsf{Test}(A, i)$. The instance Π_A^i generates a random bit b. If $b = 1$, real session key of the instance is given to E. If $b = 0$, a random value is given to E as a session key. This query is allowed only once.

In the random oracle model, cryptographic hash functions are treated as random oracles. Hence, the adversary can make queries for random oracles. The queries for random oracles are treated as follows.

(v) $\mathsf{Oracle}(m)$. E obtains a random value for the message m by this oracle call. When the oracle models a hash function, the answer returned by the oracle is the hashed value of m.

Partnering, Freshness, and Correctness. We say that two instances Π_A^i and Π_B^j are partnered if they satisfy the following conditions:

(1) Π_A^i and Π_B^j have accepted given messages;

(2) Π_A^i and Π_B^j have the same session id sid;

(3) the partner id of Π_A^i is B and vice versa.

An instance Π_A^i is called *fresh* if the instance has accepted given messages and E does not ask Reveal oracle queries for Π_A^i or its partner instance Π_B^j. The correctness requires that two instances should have the same session key if they are partnered and they have accepted.

Definitions of Security. Let $\mathcal{P}$ be a password-based protocol and let $\mathcal{A}$ be an adversary who tries to break the security of the protocol $\mathcal{P}$. Let Succ be the event that $\mathcal{A}$ asks a Test query on a fresh instance Π_A^i and correctly guesses the bit b which was selected during the Test query. Then, the advantage of $\mathcal{A}$ is defined as $\mathsf{Adv}(\mathcal{A}, \mathcal{P}) = 2\Pr[\mathsf{Succ}] - 1$. Note that all probabilistic polynomial time adversaries can always test

the validity of a guessed password by performing an online dictionary attack. Hence, the protocol $\mathcal{P}$ is considered to be secure if an online dictionary attack is the best way to break the security of $\mathcal{P}$. Note that an online attack can be mounted by making a Send oracle query. Based on the above observation, the security of an RSA-PAKE protocol can be defined as follows.

Definition 1. Let $|\mathcal{D}|$ be the size of the password space $\mathcal{D}$ and let q_S be the number of Send queries. Then, an RSA-PAKE protocol $\mathcal{P}$ is secure if the following holds:

$$\mathsf{Adv}^{\mathsf{PAKE}}_{\mathcal{P},\mathcal{D}} = \max_{\mathcal{A}\in\mathcal{ADV}} \{\mathsf{Adv}(\mathcal{A},\mathcal{P})\} \leq \frac{q_S}{|\mathcal{D}|} + \epsilon, \quad (1)$$

where $\mathcal{ADV}$ is the set of all probabilistic polynomial time adversaries and ϵ is a negligible value.

2.2. Mathematical Background. We recall a well-known theorem, the prime number theorem [10], and use it for obtaining two theorems, Theorems 2 and 3, that are used to demonstrate the security of our protocol.

Recall that the prime number theorem tells us that the number of primes smaller than a positive integer x is approximately $\pi(x) \approx x/\ln x$ for a large x.

Theorem 2. *Let e be an ℓ_e-bit prime chosen uniformly at random. Then, the probability that someone correctly guesses the prime is about $2\ell_e/2^{\ell_e}$.*

Theorem 3. *Let e be a randomly chosen ℓ_e-bit pseudoprime which is not a prime with the probability $1/2^{\ell_p}$. Then, the probability that someone chooses an ℓ_n-bit integer n such that $e \mid \phi(n)$ is bounded by $2\ell_n/2^{\ell_e} + 1/2^{\ell_p}$.*

It is easy to prove the above theorems, and thus we omit (Omitted proofs will be provided in the full-version of this paper.) them due to lack of space.

3. Our RSA-PAKE Protocol

In this section, we propose an efficient RSA-PAKE protocol which is suitable for imbalanced wireless networks. We assume that two communicating parties A and B share a common password pw for establishing a session key. Let $\mathcal{P}_{\ell_e}$ be the set of all ℓ_e-bit pseudoprimes that are not prime with probability $1/2^{\ell_p} \approx 1/2|\mathcal{D}|$. The size of the prime is determined such that $\ell_e \geq \log_2 \ell_n + \log_2 |\mathcal{D}| + 2$. We use four hash functions $H : \{0,1\}^* \longrightarrow \mathbb{Z}_n$ and $H_i : \{0,1\}^* \longrightarrow \{0,1\}^{\ell_h}$ for $i = 1, 2, 3$. Then, our RSA-PAKE protocol runs as follows.

Step 1. A chooses an ℓ_n-bit RSA modulus n and a random $r_A \in \{0,1\}^{\ell_r}$ and sends them to B.

Step 2. If n is not an ℓ_n-bit odd number, B terminates the protocol. Otherwise, he/she chooses an ℓ_e-bit prime e and a random $r_B \in \{0,1\}^{\ell_r}$ and computes $\alpha = H(pw||\mathsf{info})$, where $\mathsf{info} = id_A||id_B||r_A||r_B||n||e$. If $\gcd(\alpha, n) \neq 1$, B chooses a random $z \in \mathbb{Z}_n^*$. Otherwise, he/she computes $z = \alpha \cdot r^e \bmod n$ for randomly chosen $r \in \mathbb{Z}_n^*$. B gives (id_B, id_A, e, r_B, z) to A.

Step 3. If e is not an ℓ_e-bit prime, A terminates the protocol execution. A computes $\alpha = H(pw||\mathsf{info})$ and tests if two conditions $\gcd(\alpha, n) = 1$ and $\gcd(e, \phi(n)) = 1$ hold. If one of the conditions does not hold, A chooses a random $\hat{r} \in \mathbb{Z}_n^*$. Otherwise, he/she computes $\hat{r} = (z \cdot \alpha^{-1})^d \bmod n$, where $d = e^{-1} \bmod \phi(n)$. Then, A computes $\beta = H_1(\hat{r}||\mathsf{info})$ and sends it to B.

Step 4. If the condition $\beta = H_1(r||\mathsf{info})$ does not hold, B terminates the protocol. Otherwise, B sends $\gamma = H_2(r||\mathsf{info})$ to A. B computes $sk_B = H_3(r||\mathsf{info})$ and uses it as a session key.

Step 5. If $\gamma \neq H_2(\hat{r}||\mathsf{info})$, A terminates the protocol; otherwise, he/she computes $sk_A = H_3(\hat{r}||\mathsf{info})$ and uses it as a session key.

Remark 4. Note that, in our protocol, we use a pseudoprime which is not prime with probability $1/2^{\ell_p}$. Therefore, the number of iterations of the Miller-Rabin primary test is determined so that a pseudoprime is indeed a prime with probability $1 - 1/2^{\ell_p}$.

4. Security Analysis

In this section, we prove the security of our protocol according to the security model described in Section 2.1. Similar to the work by Zhang [7], we define a series of hybrid experiments where the first experiment describes the real adversary attack and each experiment is gradually modified so that the adversary has negligible advantage in the last experiment. We denote these hybrid experiments by Exp_i for $i \in \{0, \ldots, 4\}$. Let $\mathsf{Adv}(\mathcal{A}, i)$ be the advantage of $\mathcal{A}$ in Exp_i.

Experiment Exp_0. The first experiment coincides with the real adversary attack. Therefore, all transmitting messages are computed according to the description of the proposed protocol. Since we prove the security of our protocol in the random oracle model, four hash functions are treated as random oracles and we maintain a list of input-output pairs for each random oracle.

Note that we have $\mathsf{Adv}^{\mathsf{PAKE}}_{\mathsf{RSA\text{-}PAKE},\mathcal{D}}(\mathcal{A}) = \mathrm{Adv}(\mathcal{A}, 0)$ for an adversary $\mathcal{A}$ since the first experiment describes the real adversary attack.

Experiment Exp_1. In this experiment, we modify the Execute oracle. When the Execute oracle is called for two instances Π^i_A and Π^j_B, the session keys sk^i_A and sk^j_B are replaced by an ℓ_h-bit random string rather than an output of the random oracle H_3.

In Lemma 5, we will show that the increment of the advantage of $\mathcal{A}$ that resulted from the modification of the Execute oracle is bounded by a negligible value. Throughout this paper, $\mathsf{Adv}_{\mathsf{RSA}}$ denotes the maximum advantage of adversaries who solve the RSA problem.

Lemma 5. *For any polynomial time adversary $\mathcal{A}$ who asks at most q_E times of* Execute *queries and q_H times of hash queries, one has the following relation:*

$$|\mathsf{Adv}(\mathcal{A}, 0) - \mathsf{Adv}(\mathcal{A}, 1)| \leq q_E \mathsf{Adv}_{\mathsf{RSA}} + \frac{q_E q_H}{\phi(n)}. \quad (2)$$

Proof. We omit (Omitted proofs will be provided in the full-version of this paper.) the proof of Lemma 5, due to lack of space.

The remainder of our security proof is to show that the Send oracle gives negligible advantage to the adversary. We can classify Send oracle into five types as follows.

(i) $\mathsf{Send}_0(A, i)$. The instance Π_A^i generates a random r_A and an RSA modulus n and returns r_A and n.

(ii) $\mathsf{Send}_1(B, j, id_A, r_A, n)$. If n is not an ℓ_n-bit odd number, the protocol is terminated. Otherwise, Π_B^j chooses an ℓ_e-bit prime e and a random $r_B \in \{0,1\}^{\ell_r}$, queries the hash oracle H on $pw||\mathsf{info}$ where $\mathsf{info} = id_A||id_B||r_A||r_B||n||e$, and receives the reply α from the oracle. If $\gcd(\alpha, n) \neq 1$, Π_B^j chooses a random $z \in \mathbb{Z}_n^*$. Otherwise, it computes $z = \alpha \cdot r^e \bmod n$ for randomly chosen $r \in \mathbb{Z}_n^*$. The instance Π_B^j returns (e, r_B, z).

(iii) $\mathsf{Send}_2(A, i, r_B, z, e)$. If e is not an ℓ_e-bit prime, the protocol execution is terminated. Otherwise, Π_A^i queries H on $pw||\mathsf{info}$, receives the answer α from the oracle H, computes $\alpha = H(pw||\mathsf{info})$, and tests if $\gcd(\alpha, n) = 1$ and $\gcd(e, \phi(n)) = 1$. If one of the conditions does not hold, Π_A^i chooses a random $\hat{r} \in \mathbb{Z}_n^*$; otherwise, it computes $\hat{r} = (z \cdot \alpha^{-1})^d \bmod n$. Then, Π_A^i queries the hash oracle H_1 on $\hat{r}||\mathsf{info}$ and returns the reply β received from H_1.

(iv) $\mathsf{Send}_3(B, j, \beta)$. If the answer returned by H_1 on $r||\mathsf{info}$ is not β, Π_B^j rejects the protocol; otherwise, it queries the hash oracles H_2 and H_3 on $r||\mathsf{info}$ and receives the replies γ and sk from H_2 and H_3, respectively. Then, Π_B^j accepts sk as a session key and returns γ.

(v) $\mathsf{Send}_4(A, i, \gamma)$. If the answer returned by H_2 on $\hat{r}||\mathsf{info}$ is not γ, Π_A^i rejects the protocol; otherwise, Π_A^i accepts the answer returned by the oracle H_3 on $\hat{r}||\mathsf{info}$ as a session key.

A message is called oracle generated if it was returned by an instance; otherwise, the message is called adversary generated. If a message was returned by Π_A^i, it is called Π_A^i-oracle-generated message. □

Experiment Exp_2. In this experiment, we modify the following if an instance Π_B^j receives a Π_A^i-oracle-generated message (r_A, n) in a Send oracle call.

(i) If both Π_A^i and Π_B^j accept, we choose a random ℓ_h-bit value and give it to two instances as a session key.

(ii) If Π_A^i does not accept but Π_B^j accepts, we choose a random ℓ_h-bit value and give it to Π_B^j as a session key. In this case, no session key is defined for Π_A^i.

Lemma 6. *For any polynomial time adversary $\mathscr{A}$ who asks at most q_S Send queries, one has*

$$|\mathsf{Adv}(\mathscr{A}, 1) - \mathsf{Adv}(\mathscr{A}, 2)| \leq q_S \mathsf{Adv}_{\mathsf{RSA}} + \frac{q_S q_H}{\phi(n)}. \quad (3)$$

Proof. Let $n = pq$ be an RSA modulus where p and q are distinct primes. Note that since the modulus n is chosen by the instance Π_A^i (not by the adversary $\mathscr{A}$),

$$\Pr[\gcd(\alpha, n) \neq 1] = \frac{n - \phi(n)}{n} = \frac{p + q - 1}{n} \approx \frac{1}{2^{\ell_n/2}} \quad (4)$$

for randomly chosen $\alpha \in \mathbb{Z}_n^*$. Hence, we can assume that $\gcd(\alpha, n) = 1$ for randomly chosen $\alpha \in \mathbb{Z}_n^*$.

Assume that Π_B^j receives a Π_A^i-oracle-generated message (r_A, n) in a Send oracle call and returns (e, r_B, z) where e is ℓ_e-bit prime, $r_B \in \{0,1\}^{\ell_r}$, and $z = \alpha \cdot r^e \bmod n$ for random $r \in \mathbb{Z}_n^*$ and $\alpha \in \mathbb{Z}_n^*$. Note that α is the answer returned by the random oracle H on $pw||id_A||id_B||r_A||r_B||n||e$, since $\gcd(\alpha, n) = 1$ holds with probability $1 \approx 1 - 2^{-\ell_n/2}$. Note that, as proved in Lemma 5, we can show that the probability that $\mathscr{A}$ recovers the random value r is $\mathsf{P}_r \leq \mathsf{Adv}_{\mathsf{RSA}} + q_H/\phi(n)$. Since $\mathscr{A}$ cannot generate valid β and γ without the knowledge of r, two instances Π_A^i and Π_B^j accept the protocol execution only if the instances Π_A^i and Π_B^j receive Π_B^j-oracle-generated message β and Π_A^i-oracle-generated message γ, respectively. Hence, $\mathscr{A}$ can distinguish Exp_1 and Exp_2 only if the adversary can test to see whether or not a session key is the answer returned by the random oracle H_3 on $r||id_A||id_B||r_A||r_B||n||e$. Without the knowledge of r, the session key seems to be a random value in $\mathscr{A}$'s point of view, and thus $\mathscr{A}$ cannot distinguish between two experiments. As a result, we have $|\mathsf{Adv}(\mathscr{A}, 1) - \mathsf{Adv}(\mathscr{A}, 2)| = q_S \mathsf{P}_r \leq q_S \mathsf{Adv}_{\mathsf{RSA}} + q_S q_H/\phi(n)$, since $\mathscr{A}$ asks at most q_S times of Send queries.

Experiment Exp_3. In this experiment, we consider the case where an instance Π_A^i receives a Π_B^j-oracle-generated message (e, r_B, z) in a Send_2 oracle call, while the instance Π_B^j received a Π_A^i-oracle-generated message (r_A, n) in a Send_1 oracle call. If Π_A^i and Π_B^j accept the protocol execution and their session keys are not replaced by a random value in the experiment Exp_2, we give a random ℓ_h-bit value to them as a session key. □

Lemma 7. *For any polynomial time adversary $\mathscr{A}$ who asks at most q_S Send queries, one has* $\mathsf{Adv}(\mathscr{A}, 2) = \mathsf{Adv}(\mathscr{A}, 3)$.

Proof. It is clear that the advantage of $\mathscr{A}$ in Exp_3 is identical with its advantage in Exp_2 since the only way to distinguish two experiments is recovering the random value as discussed in Lemma 6.

Experiment Exp_4. In this experiment, we consider the case where an instance Π_A^i (or Π_B^j) receives an adversary-generated message in a Send_2 (or Send_1) oracle call. If Π_A^i (or Π_B^j) accepts the protocol execution, we stop the experiment and the adversary is said to have succeeded. Note that the modification of the experiment certainly improves the adversary's advantage. □

Lemma 8. *For any polynomial time adversary $\mathscr{A}$, one has the following relation:* $\mathsf{Adv}(\mathscr{A}, 3) \leq \mathsf{Adv}(\mathscr{A}, 4)$.

Table 1: Performance comparison.

(a) Computational cost of client (low-power device holder).

	Parameters		Encryption		Execution time	
	Operation	Time	Operation	Time	Total	Precomp.
SNAPI [3]	C_{PVer}^{1025}	156.89 ms	C_{Exp}^{1025}	49.06 ms	205.95 ms	205.95 ms
PEKEP [7]	—	0 ms	$646C_{Exp}^{|e|}$ $(e = 3)$	188.31 ms	188.31 ms	188.31 ms
CEKEP [7]	—	0 ms	$50C_{Exp}^{|e|}$ $(e = 3)$	29.2 ms	29.2 ms	29.2 ms
Ours	C_{PGen}^{52}	1.23 ms	C_{Exp}^{52}	3.37 ms	4.6 ms	3.37 ms

(b) Computational cost of server (powerful computing equipment holder).

	Parameters		Encryption		Execution time	
	Operation	Time	Operation	Time	Total	Precomp.
SNAPI [3]	C_{MGen}^{1024}	760.58 ms	C_{Exp}^{1024}	48.78 ms	809.36 ms	48.78 ms
PEKEP [7]	C_{MGen}^{1024}	760.58 ms	$2C_{Exp}^{1024}$	97.56 ms	858.14 ms	97.56 ms
CEKEP [7]	C_{MGen}^{1024}	760.58 ms	$3C_{Exp}^{1024}$	146.34 ms	906.92 ms	146.34 ms
Ours	$C_{MGen}^{1024} + C_{PVer}^{52}$	761.13 ms	C_{Exp}^{1024}	48.78 ms	809.91 ms	49.33 ms

(c) Communication overhead

	Transmitting messages		Rounds
	Components	Length	
SNAPI [3]	$2\ell_n + \ell_{e_1} + 2\ell_r + 2\ell_h$	3713 bits	4
PEKEP [7]	$2\ell_n + 2\ell_r + 2\ell_h$	2688 bits	4
CEKEP [7]	$3\ell_n + 4\ell_r + 2\ell_h$	4032 bits	6
Ours	$2\ell_n + \ell_e + 2\ell_r + 2\ell_h$	2740 bits	4

Proof. Note that, in Exp_4, the adversary $\mathcal{A}$ can obtain more information than in Exp_3. Hence, it is obvious that $\mathsf{Adv}(\mathcal{A}, 4)$ is greater than $\mathsf{Adv}(\mathcal{A}, 3)$.

It remains to show that the adversary's advantage in the last experiment is negligible. □

Lemma 9. *For any polynomial time adversary $\mathcal{A}$ who asks at most q_S Send queries and q_H random oracle queries, one has* $\mathsf{Adv}(\mathcal{A}, 4) \leq q_S/|\mathcal{D}| + q_S\mathsf{Adv}_{\mathsf{RSA}} + q_S q_H/\phi(n) + q_S/2^{\ell_h}$.

Proof. We omit (Omitted proofs will be provided in the full-version of this paper.) the proof of Lemma 9, due to lack of space.

By combining Lemmas 5, 6, 7, 8, and 9, we obtain Theorem 10. □

Theorem 10. *Let $\mathcal{A}$ be a probabilistic polynomial time algorithm which asks at most q_E Execute queries, q_S Send queries, and q_H Oracle queries for hash functions. Let $|\mathcal{D}|$ be the size of password space $\mathcal{D}$. Then, one has*

$$\mathsf{Adv}^{\mathsf{PAKE}}_{\mathsf{RSA\text{-}PAKE},\mathcal{D}}(\mathcal{A}) \leq \frac{q_S}{|\mathcal{D}|} + q_1\mathsf{Adv}_{\mathsf{RSA}} + \frac{q_2}{\phi(n)} + \frac{q_S}{2^{\ell_h}}, \quad (5)$$

where $q_1 = q_E + 3q_S$ and $q_2 = (q_E + 3q_S)q_H$.

Proof. Note that, by Lemma 8, it is easy to see that $\mathsf{Adv}(\mathcal{A}, 0) \leq \sum_{i=0}^{2} |\mathsf{Adv}(\mathcal{A}, i) - \mathsf{Adv}(\mathcal{A}, i+1)| + \mathsf{Adv}(\mathcal{A}, 4)$. By Lemmas 5, 6, 7, and 9, we have

$$\mathsf{Adv}(\mathcal{A}, 0) \leq \frac{q_S}{|\mathcal{D}|} + q_1\mathsf{Adv}_{\mathsf{RSA}} + \frac{q_2}{\phi(n)} + \frac{q_S}{2^{\ell_h}}, \quad (6)$$

where $q_1 = q_E + 3q_S$ and $q_2 = (q_E + 3q_S)q_H$. Since $\mathsf{Adv}^{\mathsf{PAKE}}_{\mathsf{RSA\text{-}PAKE},\mathcal{D}}(\mathcal{A}) = \mathrm{Adv}(\mathcal{A}, 0)$, we complete the proof of Theorem 10. □

Note that Theorem 10 tells us that the proposed RSA-PAKE protocol is secure in terms of the security models described in Section 2.1 if the RSA problem is intractable.

5. Performance

In this section, we compare the proposed protocol with existing RSA-PAKE protocols, except the EPAKE protocol since the insecurity of that protocol has been discovered [11]. Let C_{Exp}^{a} be the cost of an exponentiation under a 1024-bit modulo with an a-bit exponent, let C_{MGen}^{a} be the cost of generating an a-bit RSA modulus, let C_{PGen}^{a} be the cost of generating an a-bit prime, and let C_{PVer}^{a} be the cost of verification of an a-bit prime. Note that $\alpha C_{Exp}^{a} = C_{Exp}^{\alpha a}$. For comparison, we set $\ell_n = 1024$, $\ell_r = \ell_h = 160$, $\ell_{e_1} = 1025$, $\ell_{e_2} = 96$, and $\ell_e = 52$. For determining the number of Miller-Rabin primary tests, we refer to [12]. Our protocol uses a 52-bit pseudoprime e which is indeed prime with probability $1 - 1/2^{\ell_p} (= 1 - 2^{-41})$, and so it suffices to perform the primary tests 37 times according to the formula given in [12].

We performed experiments under Windows XP Professional with a 3.4 GHz Pentium 4 processor. As seen in Table 1, our protocol provides very efficient key exchange compared with other RSA-PAKE protocols in terms of the computational complexity. The computational complexity of the party is reduced by 84.25% compared to the CEKEP

protocol. Moreover, the proposed protocol can be implemented more efficiently by generating a prime before a key exchange protocol is initiated. The PEKEP is the most efficient RSA-PAKE protocol in terms of communication overhead. The size of the communicating message of our scheme is almost the same as that of the PEKEP and shorter than other protocols, and thus our protocol is also efficient in terms of the communication overhead.

6. Conclusion

In this paper, we proposed an efficient RSA-PAKE protocol which is suitable for securing the communications conducted over imbalanced wireless networks and proved the security of our protocol in the random oracle model. Compared with other RSA-PAKE protocols, our protocol provides very efficient key exchange. In our protocol, the cost of a low-power device holder can be reduced by 84.25% compared with the CEKEP. Moreover, our protocol can be implemented more efficiently by pregenerating a prime before a key exchange protocol is initiated.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (Grant no. 2011-0029925).

References

[1] F. Bao, "Security analysis of a password authenticated key exchange protocol," in *Proceedings of the 6th International Conference (ISC '03)*, vol. 2851 of *Lecture Notes in Computer Science*, pp. 208–217, Springer, Bristol, UK, 2003.

[2] D. Catalano, D. Pointcheval, and T. Pornin, "Trapdoor hard-to-invert group isomorphisms and their application to password-based authentication," *Journal of Cryptology*, vol. 20, no. 1, pp. 115–149, 2007.

[3] P. MacKenzie, S. Patel, and R. Swaminathan, "Password-authenticated key exchange based on RSA," in *Proceedings of the 6th International Conference on the Theory and Application of Cryptology and Information Security (ASIACRYPT '00)*, vol. 1976 of *Lecture Notes in Computer Science*, pp. 599–613, Springer, 2000.

[4] S. Park, J. Nam, S. Kim, and D. Won, "Efficient password-authenticated key exchange based on RSA," in *Proceedings of the Cryptographers' Track at the RSA Conference*, vol. 4377 of *Lecture Notes in Computer Science*, pp. 309–323, Springer, San Francisco, Calif, USA, 2007.

[5] D. S. Wong, A. H. Chan, and F. Zhu, "More efficient password authenticated key exchange based on RSA," in *Proceedings of the 4th International Conference on Cryptology (INDOCRYPT '03)*, vol. 2904 of *Lecture Notes in Computer Science*, pp. 375–387, Springer, New Delhi, India, 2003.

[6] M. Zhang, "Further analysis of password authenticated key exchange protocol based on RSA for imbalanced wireless networks," in *Proceedings of the 7th International Conference (ISC '04)*, vol. 3225 of *Lecture Notes in Computer Science*, pp. 13–24, Springer, Palo Alto, Calif, USA, 2004.

[7] M. Zhang, "New approaches to password authenticated key exchange based on RSA," in *Proceedings of the 10th International Conference on the Theory and Application of Cryptology and Information Security*, vol. 3329 of *Lecture Notes in Computer Science*, pp. 230–244, Springer, Jeju Island, Republic of Korea, 2004.

[8] F. Zhu, D. S. Wong, A. H. Chan, and R. Ye, "Password authenticated key exchange based on RSA for imbalanced wireless networks," in *Proceedings of the 5th International Conference (ISC '02)*, vol. 2433 of *Lecture Notes in Computer Science*, pp. 150–161, Springer, Sao Paulo, Brazil, 2002.

[9] M. Bellare, D. Pointcheval, and P. Rogaway, "Authenticated key exchange secure against dictionary attack," in *Proceedings of the Eurocrypt*, vol. 1807 of *Lecture Notes in Computer Science*, pp. 139–155, Springer, 2000.

[10] K. H. Rosen, *Elementary Number Theory and Its Applications*, Addison-Wesley, Boston, Mass, USA, 4th edition, 2000.

[11] T.-Y. Youn, Y.-H. Park, C. Kim, and J. Lim, "Weakness in a RSA-based password authenticated key exchange protocol," *Information Processing Letters*, vol. 108, no. 6, pp. 339–342, 2008.

[12] A. Menezes, P. Oorschot, and S. Vanstone, *Handbook of Applied Cryptography*, CRC Press, Boca Raton, Fla, USA, 1997.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 381430, 11 pages
http://dx.doi.org/10.1155/2014/381430

Research Article

A Multifunctional RF Remote Control for Ultralow Standby Power Home Appliances

Kwang-il Hwang and Sung-Hyun Yoon

Department of Embedded Systems Engineering, Incheon National University, Incheon 402-772, Republic of Korea

Correspondence should be addressed to Kwang-il Hwang; hkwangil@incheon.ac.kr

Received 10 December 2013; Accepted 16 April 2014; Published 26 May 2014

Academic Editor: Jongsung Kim

In spite of many benefits, since a target RF should be able to react to real time user commands even during system power-off, RF remote controls generally require more standby energy than IR manner. Therefore, in this paper a multifunctional RF remote control (MRRC), which is capable of providing larger coverage and various services, is introduced, and an ultralow standby power operation method for target RFs, utilizing an extended preamble transmission and a variable length periodic preamble sensing according to system power states, is proposed. In addition, a prototype and implementation details are also described. In order to evaluate the proposed MRRC, several experiments are conducted, and each performance of MRRC is also compared with ZigBee RF4CE no power saving and power saving mode. The experimental results demonstrate that the MRRC system enables not only ultralow standby power operation during system power-off but also low power operation even in system power-on state. In spite of ultralow standby power operation, the experimental result also shows that the MRRC provides reasonable response time to user command.

1. Introduction

After the first wireless remote control, Flash-Matic developed by Eugene Polley in 1955 has been introduced and the IR (infrared) remote control has been used dominantly for TV and other home appliances for about more than 30 years. The IR remote control provides simple and low power connectivity to remotely control several home appliances but also has the following limitations: one way communication, line-of-sight constraint, and one-to-one communication only. In addition, to receive commands from a remote, power of an IR receiver remains active. Recently, as the number of smart appliances is increased, the demand for a new remote control to overcome the limitations of IR remote controls is arising. Furthermore, the rapid advances in wireless systems and user interfaces have accelerated the advent of a new remote control using several wireless systems or various user interfaces [1–8].

Remotes [1, 2] using Bluetooth support simple, familiar association with smart phones. However, due to interference problems with other wireless systems such as Wi-Fi and scalability problems that a slave is allowed to be paired with only one master (target) device, the spread of a remote control based on Bluetooth in various home appliances is limited. In addition, ZigBee RF4CE [5] (Radio Frequency for Consumer Electronics) has been emerged as a representative RF remote control, which enables one-to-many communications based on two-way communications and provides larger coverage due to no line-of-sight constraint.

In spite of many benefits, the RF remote controls, such as RF4CE, generally require more standby energy than IR manner, and thus it might result in increase of standby power consumption in home. In particular, recently as concerns about standby power reduction is being increased much more, some research focuses on reducing standby power in home appliances [9, 10, 16, 17].

Therefore, this paper deals with a problem of how to reduce standby power consumption in home, and, in particular, reduction of standby power consumed in RF remote receivers, which should be able to react to real time user commands even if the system power is off, is focused on. Recently, as home appliances controlled by remote controls are increased, the problem reducing standby power in home appliances is considered to be challenging.

The remainder of this paper is organized as follows. In Section 2 several researches related to our work are investigated. Section 3 introduces RF low power listening and the proposed idea is described in Section 4. Section 5 shows an example of the MRRC operation. Experimental results and performance evaluations are presented in Section 6. Section 7 provides some concluding remarks.

2. Related Work

Over the last decade, there have been a number of researches [1–8] on new remote controls for various home appliances to replace conventional IR remote controls. Kim et al. [3] proposed a new remote control interface using a touch screen and haptic interface. Park and Lee [4] proposed a remote for smart TV using a user pattern recognition technology through a camera placed on TV. Even though these new interfaces for a remote control can provide more convenient environment to users, these might require more complex remote control method and be difficult to apply to various other target systems other than TV.

On the other hand, RF remote controls [5–8, 11] can take advantage of similar user interface to conventional IR remotes and thus enhance its performance and functionality without learning a new remote interface. Han et al. [5] proposed a home appliance control scheme using both IR remote and ZigBee RF communication. Each power outlet and dimming lights are equipped with ZigBee RF module, and they are managed by a ZigBee controller, which is controlled by an IR remote control. Hwang et al. [6] proposed an enhanced version of ZigBee to IR remote [5]. ZigBee to IR remote control performs a remote function for appliances requiring IR remote control, and ZPA (ZigBee Power Adaptor) manages systems requiring power control. Kim et al. [7] proposed a universal remote control having various communication interfaces using a mobile device to overcome limitations of a conventional remote control such as interface complexity and specific application constraint. A mobile device is equipped with Wi-Fi and additional RF module, and RF to relay is used to control power control devices, RF to IR is also used to control conventional appliances, and Wi-Fi is used for Internet access or CCTV.

Finally, the demand for a new remote using low power RF urged to the advent of a new standard ZigBee RF4CE [8] by ZigBee alliance in 2009. The RF4CE provides one-to-many communication and larger coverage over IR remote control, which is limited in line-of-sight. In addition, Hwang [11] proposed an interference which avoided RF4CE in 2.4 GHz ISM band.

In parallel with research on a new remote control, there have been some researches [5, 12–14] on reducing standby power in home. Tsai et al. [12] proposed a standby power reduction method for lighting devices by adaptively controlling lights according to human movement. Han et al. [5] and Han et al. [14] proposed a home energy management scheme in which a power outlet has a function to cutoff standby power if power consumption goes below a predefined threshold and a home server allows users to control home appliances outside the home.

In spite of a lot of efforts to reduce standby power in home, there have been only a few research on standby power minimization of a remote receiver. Kang et al. [13] proposed a remote control and remote receiver based on autonomous power in which transmitter accumulates energy in receiver using laser diode, and an IR receiver is powered by the stored energy. However, this method requires considerable energy in transmitter.

RF standby power minimization has been actively studied in wireless sensor network area. In particular, LPL (low power listening) methods [9, 10, 16, 17] are used to reduce unnecessary listening period and our work is also based on LPL. However, since most existing LPL methods are designed for wireless sensor networks in which sensors are massively deployed and have no intervention of human, it is impossible to apply the LPL itself for sensor networks to remote control applications.

Therefore, in this paper a novel low power RF communication based on LPL for RF remote control, which is capable of minimizing RF standby power by switching the RF listening period adaptively according to the state of target systems, is proposed.

3. Preliminary

Unlike the IR transmission using predetermined frequencies, RF communications are capable of transmitting various formats of data so that it is possible to provide more flexible and smart controls. In particular, the ZigBee RF4CE standard based on IEEE802.15.4 [15] enables low power remote control. However, the IEEE802.15.4 network depends on strict time synchronization between a coordinator and devices to maintain a superframe structure, and thus each node might waist unwanted energy due to idle listening. Furthermore, CSMA/CA based network can provide irregular latency so that users can feel some inconvenience.

Figure 1(a) shows a general RF frame structure. At the beginning of a frame, a preamble, which is a repeated pattern of "1" and "0" and is used as an indicator to notice the start of transmission, is transmitted. At the end of a synchronization word for sampling data, variable length data are transmitted and finally each frame is finished with the CRC (cyclic redundancy check), which is used to detect errors in the frame. As shown in the figure, to receive a frame the receiver should remain active before packet reception, and the active state should be kept until the whole frame is received. Therefore, for low power operation, an RF receiver, normally in sleep state, wakes up and receives a frame at the time when a transmitter sends a frame and then returns to sleep state again. In order to achieve this, the time synchronization between transmitter and receiver is required, and a duty cycle, which includes an active and sleep duration, should be maintained. However, it is difficult to correctly maintain time synchronization due to the effect of clock drift. Moreover, in remote control applications, user

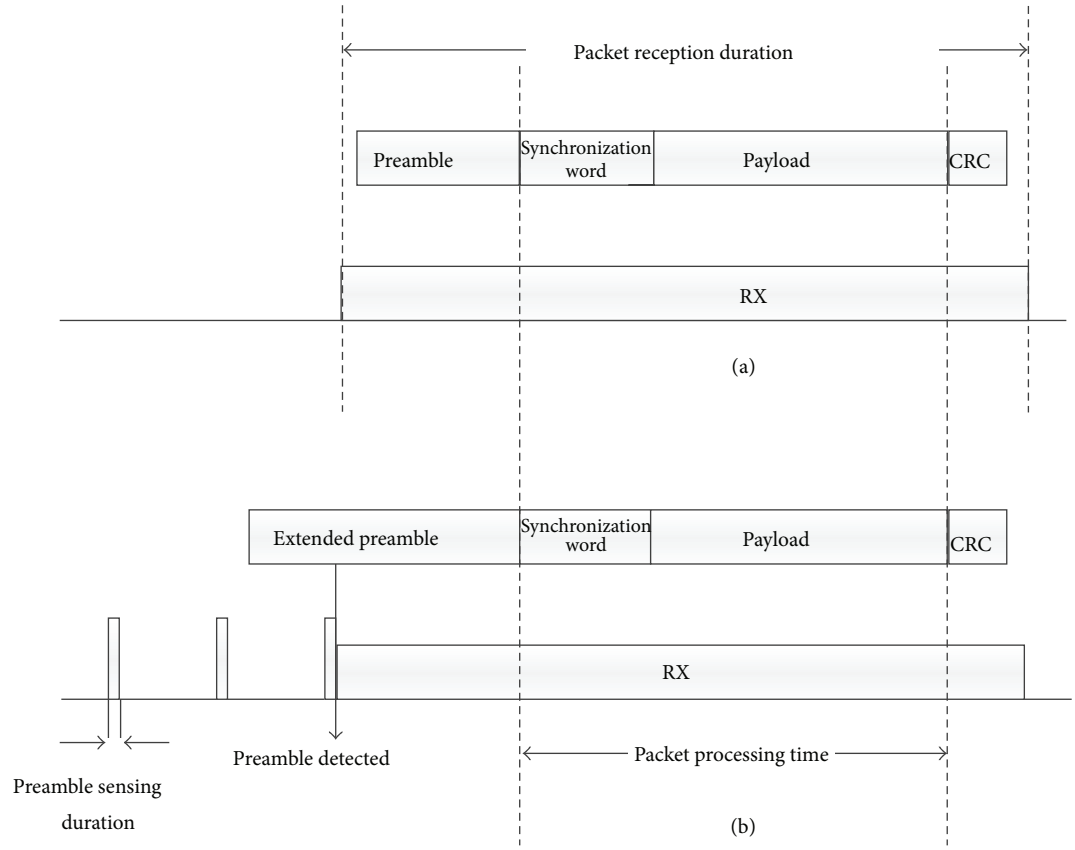


Figure 1: General RF packet structure versus extended preamble packet structure.

commands have unpredictable characteristics. That is, since most listening periods, which are in active state for the packet reception period, might become idle listening period; each target RF wastes considerable energy.

To address idle listening problems, there have been substantial researches [9, 10, 16, 17] on LPL to reduce idle listening periods. As shown in Figure 1(b), the LPL is different from general packet reception processing in that more expanded preamble is transmitted and a receiver is activated repeatedly only for a short duration (preamble sensing duration) to detect preamble transmission and the remainder of the period is in sleep mode.

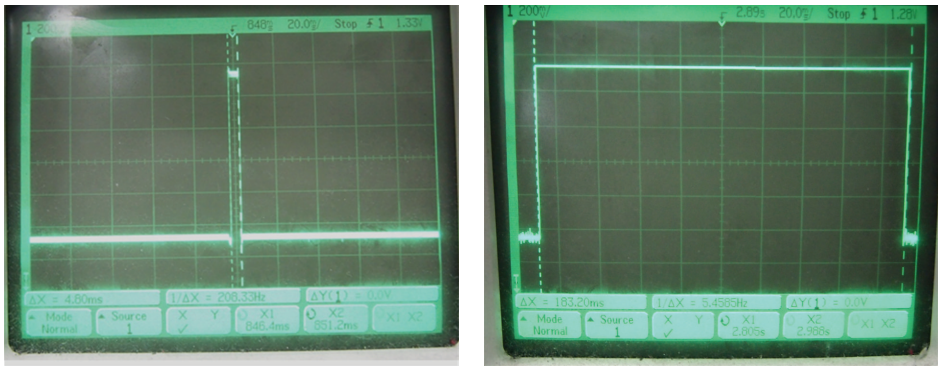
Even though the LPL for sensor networks brings more energy conservation by reducing idle listening, the characteristics of sensor network application, in which sensors are massively deployed and have no intervention of human, are basically different from home remote controls, which are directly controlled by human. Therefore, in the following section, a new RF standby power reduction method based on adaptive low power listening is presented.

4. Multifunctional RF Remote Control for Ultralow Standby Power Home Appliances

In this Section, a multifunctional RF remote control (MRRC), which enables ultralow standby power operation of home appliances, is proposed.

4.1. Ultralow Power Listening with Variable Sensing Interval. In order to minimize power consumption of each target RF associated with MRRC, an extended preamble to trigger a target RF prior to each command is transmitted, and each target RF performs a periodic preamble sensing (PPS) to detect an extended preamble transmission for a very short duration. It is important to note that the length of preamble sensing duration, which is an active duration to perform preamble sensing, should be minimized to maintain a minimum duty cycle. Therefore, a minimum preamble sensing duration, an optimal length satisfying 100% successful preamble detection, is found through experiments in which total 100 trials are conducted. As shown in Figure 2, the obtained minimum preamble sensing duration is 4.8 milliseconds and the length is considerably shorter than minimum packet reception duration (188.8 milliseconds). The result reveals that inactive period of the MRRC system for an idle listening can be extended over 40 times compared to general packet reception schemes within the same period, and it might result in considerable energy saving.

In addition, one of the major differences from other LPL methods is capable of providing minimizing standby power consumption of target RF by applying a variable length preamble sensing interval (PSI) according to the state of each target system, despite providing reasonable response time to user commands. Each target RF performs a PPS to sense an extended preamble for a short period and maintains



(a) (b)

Figure 2: Preamble sensing duration versus packet reception duration.

sleep mode until the next PPS. Since an extended preamble transmission should be able to be sensed by a target RF during a PPS within a PSI, the length of an extended preamble transmitted by MRRC must be longer than a period of a PSI. The MRRC utilizes two different length PPSs: long PPS and short PPS. The long PPS is used to listen to a system power-on command by MRRC, and thus it is maintained for a relatively long period. That is, during system power-off, a target RF performs PPS with a longer interval. The longer the period is, the more energy can be saved, but response time of target RF with respect to user's command request is also longer. Therefore, through experiments a suitable interval value (3 seconds) is determined, which is the maximum waiting time that a user can tolerate. However, the PPS value is also configurable by a user. After system power-on, a target RF performs a short PPS with a shorter interval. Through another experiment, a suitable short PPS time (1 second) to wait for user command during system power-on is also determined.

If a user command is transmitted by MRRC, during PPS each target RF is triggered by an extended preamble and keeps a *wait-for-command* state until following command data are received. If the target ID in the received frame is not matched with my ID, the target RF returns to PPS mode again.

The MRRC also supports consecutive remote commands (i.e., volume up/down, TV channel set, etc.) during system power-on. In particular, to guarantee fast response time to consecutive commands, after the previous command is completed, both MRRC and target RF remain active for the same standby duration to wait for the next command. It is important to note that the commands generated during the standby period are transmitted as a general frame without an extended preamble. This standby duration is used to make a prompt response to consecutive commands by removing redundant delay to process an extended preamble. Furthermore, standby duration is reset whenever a new command is received within the standby duration, so that response time to consecutive commands is considerably reduced. If there is no following command for the standby duration, the standby timer expires so that the target RF returns to PPS state. Figure 3 shows a state transition diagram to perform a variable length PPS according to the states of a target system.

4.2. On-Demand Target Trigger. Another outstanding feature of the MRRC is asynchronous target trigger based on on-demand time synchronization. In order to cope well with on-demand user command, no global time synchronization but on-demand time synchronization is used in a MRRC. As mentioned in the previous subsection, to trigger a target RF which is performing PPS, an MRRC should transmit an extended preamble longer than a period of a PPS. As shown in Figure 4, all the target RFs performing preamble sensing at different time are triggered by an extended preamble of MRRC although preamble detection time of each target RF is different, and they receive simultaneously the command data transmitted at the end of the extended preamble. Therefore, all the target RFs and MRRC can be synchronized with each other. It is important to notice that each device that detects a preamble should be in active state for the *wait-for-command* duration. However, as shown in Figure 4, the length of *wait-for-command* duration of each target RF might be varied according to the point that a preamble is detected. That is, the earlier the preamble detection is, the longer the delay is until user command is received. Therefore, to minimize unnecessary delay, the MRRC also provides an interactive preamble termination method which can be applied when a target system is in user's line-of-sight. Figure 4 illustrates an example of interactive preamble termination. Whereas a normal extended preamble shown in the first command lasts for *maxpreamblelength*, the second preamble is terminated as soon as the button is released. That is, a target system can notice the MRRC user of is the preamble detection using a LED, and thus the system can provide a faster response time by terminating current preamble transmission as soon as the button of a MRRC is released. In particular, the method, which provides a flexible preamble transmission based on a feedback between user and a target system, can compensate relatively long response time in long PPS.

4.3. Bidirectional Communication Capability. Unlike an IR remote control, the MRRC also has bidirectional communication capability. In particular, the MRRC is capable of controlling the length of an extended preamble through a feedback from a target system. Figure 5 illustrates an example of flexible PPS operation based on bidirectional

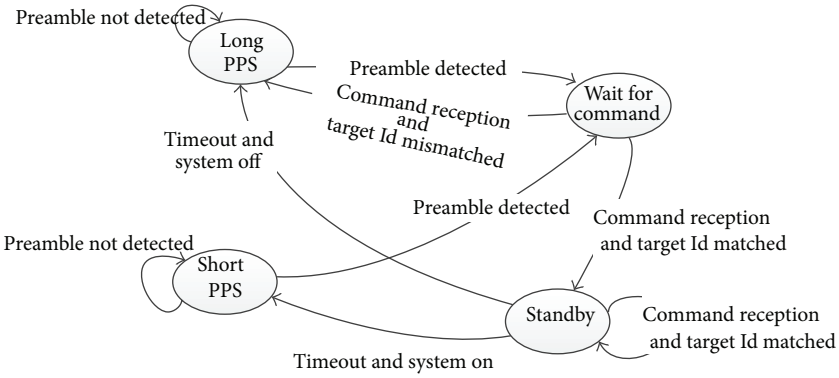


Figure 3: State transition diagram.

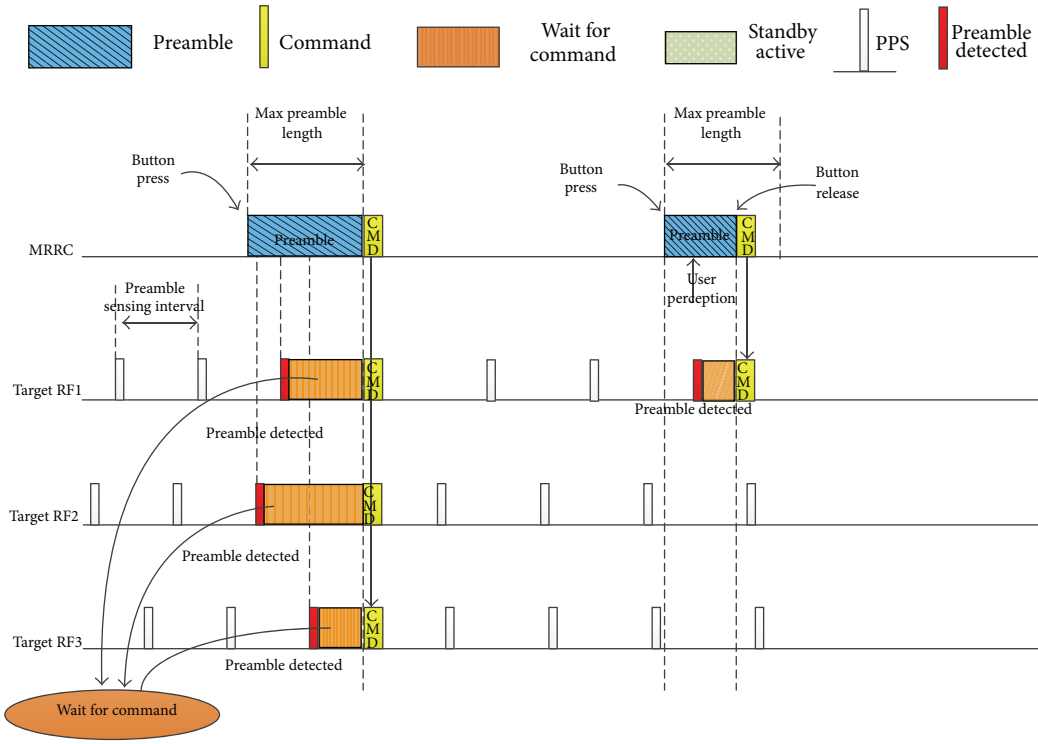


Figure 4: On-demand target trigger by MRRC.

communication between an MRRC and target RF. A target RF is performing PPS at long preamble sensing interval (LPSI) during system power-off, and the MRRC transmits a long preamble to trigger the target RF. At the end of a preamble transmission, user command is transmitted and target system carries out the received command. (As shown in Figure 5, the first command is system-power-on.) On carrying out the corresponding command, the target RF replies to the MRRC. To handle successive user commands, both target RF and MRRC remain active for the standby duration. As mentioned previously, if successive user command is transmitted during this period, the command can be transmitted without an extended preamble so that response time of the command is considerably reduced. Here, since the MRRC and target RF are synchronized by the command frame followed by an extended preamble, the two devices can maintain the same standby time. If there is no command for the standby duration, the target RF performs a PPS at short preamble sensing interval (SPSI) and MRRC returns to sleep mode. After standby duration expires, the MRRC uses a short extended preamble to trigger the target RF, which is performing a short PPS. The short PPS is used to consume RF power as low as possible even during system power-on and also provides the reduced response time over long PPS which is performed during system power-off. In particular, utilizing standby duration associated with short PPS enables fast response time by removing delay in the target RF and MRRC. If a target RF receives a system-off command from MRRC, the target RF turns off the system and responds to the MRRC. After that, to wait for additional consecutive command by user, the two

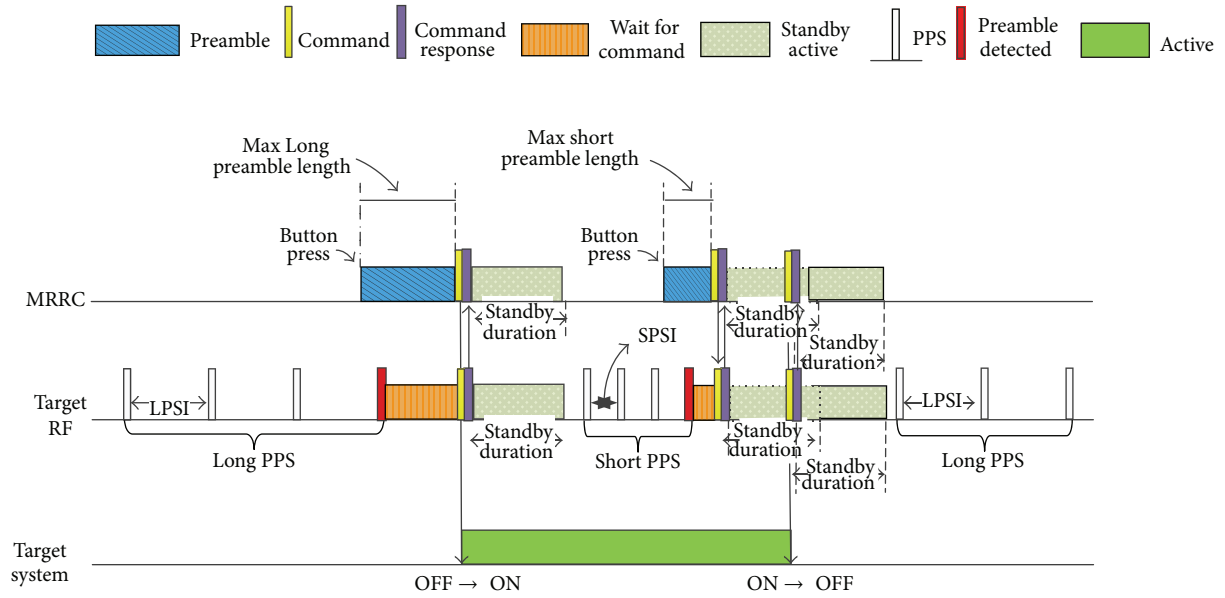


Figure 5: Bidirectional communication between MRRC and target systems.

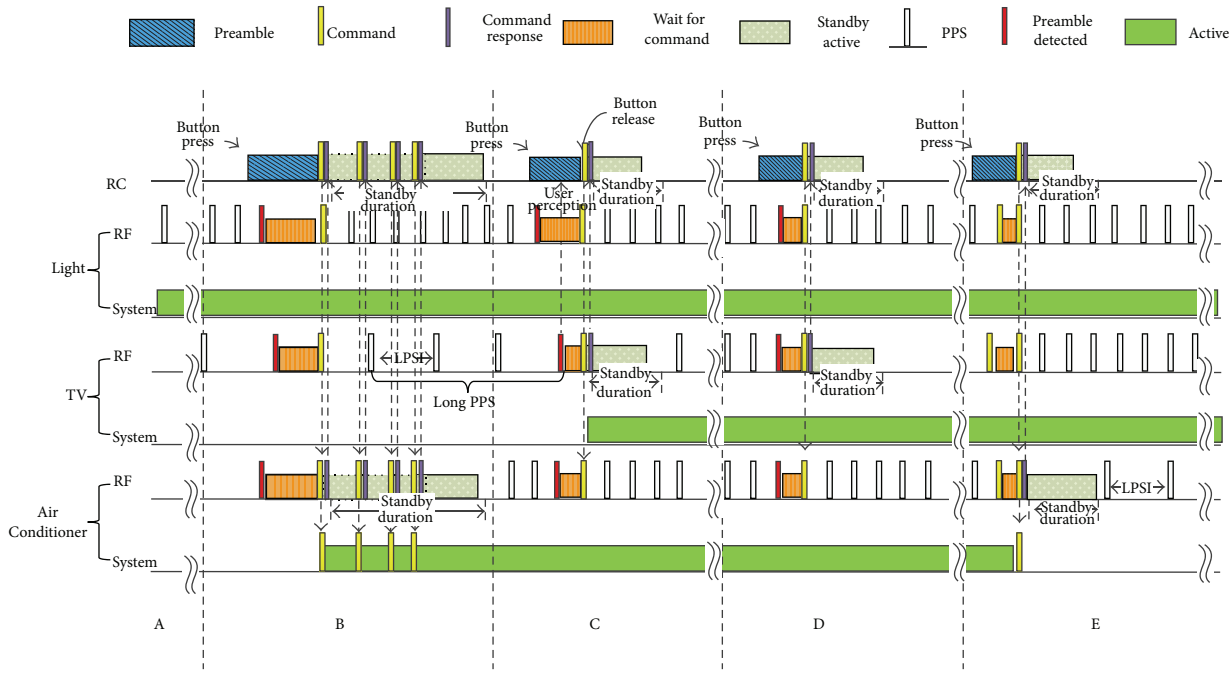


Figure 6: An example of MRRC operation.

devices keep active state for the standby duration, and then if there is no command during the time, the target RF performs long PPS again to save power.

5. Controlling Multiple Target Systems

Figure 6 shows an example of controlling a light, TV, and air conditioner using a single MRRC. First, in section A the light is already turned on, and other systems (TV and air conditioner) are powered off. TV and air conditioner are performing long PPS to reduce standby power, and the light is performing short PPS to react quickly to user commands. In section B, user turns on the air conditioner and immediately controls temperature of the air conditioner successively. As shown in the figure, the successive temperature control commands are processed immediately without extended preamble transmission. During the time, TV that is performing long PPS is triggered by the first command but returns to long PPS again after the command is received due to ID mismatch. Also, the TV is not even triggered

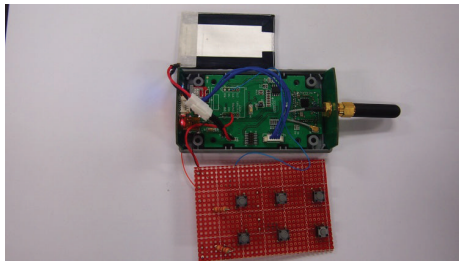


Figure 7: MRRC prototype.

Figure 8: Test bed.

by following short commands for the air conditioner. In section C, the user turns on TV through RF performing long PPS. At this moment, the extended preamble does not last for the *maxlongpreambleduration* but is terminated midway by user that perceives LED signal on TV, informing that a preamble is detected. That is, interactive preamble termination is used in that situation. In section D, the user switches channel. At that moment, since TV is powered on, RF is performing short PPS, and the MRRC sends the command after triggering RF by short extended preamble. This section has no more commands so that after standby duration the RF returns to short PPS. In section E, the user turns up the volume of the TV successively. Since volume controls are generated successively for the short duration, the rest of the commands are transmitted without extended preamble transmission, except for the first command with a short extended preamble. Therefore, the user can experience fast response to successive commands. In the final, user turns off TV and air conditioner. The RF, which is performing short PPS to provide relatively fast response, receives system power-off command and then goes to the long PPS mode after standby duration, to minimize standby power.

6. Experimental Result

In this Section, a MRRC prototype and test bed are introduced, and experimental results are presented. In particular, each performance of the MRRC is compared with ZigBee RF4CE, which is the most representative RF remote control.

6.1. MRRC Prototype. Figure 7 shows an MRRC hardware prototype. The MRRC prototype is composed of 16-bit low power MCU and sub-1-GHz RF, which is capable of transmitting a variable length preamble. In addition, to generate user commands, simple push switches are used in place of key pads. The MRRC is also designed to cope well with several different events (external interrupts, timer interrupts, RF interrupts, etc.) through an event driven lightweight scheduler based on HAL (hardware abstract layer), which manages directly hardware. This development environment might also facilitate various MRRC application developments.

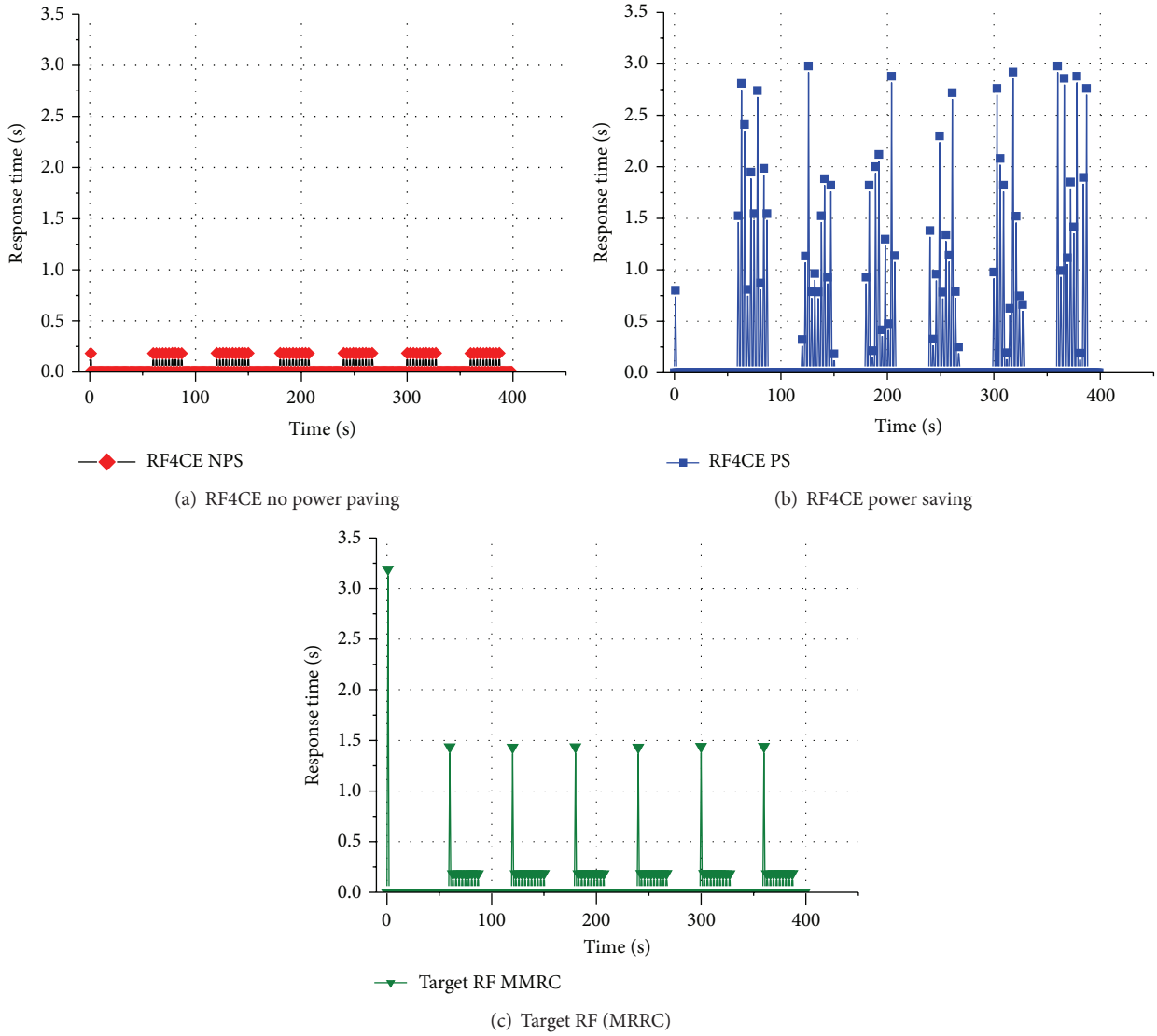
6.2. Experimental Environment. For MRRC experiments, a target appliance emulator, which is a PC application software, is implemented. As shown in Figure 8, the used target emulators accept each command for light, TV and air conditioner, respectively, and individual RF is connected to the corresponding target system via USB. Each RF and a target can communicate with each other and the MRRC can control a designated target by target ID assigned uniquely. Emulator also plays a role in storing and analyzing data received from an MRRC.

Table 1: System parameters.

Parameter	Value	
	MRRC	RF4CE
Supply voltage	3.3 V	
Current consumption		
TX active	26.3 mA	
RX active	22.3 mA	
Sleep	30 uA	
MaxLongPreamble length	3.2 s	—
MaxShortPreamble length	2.2 s	—
nwkDutyCycle	—	3 s
nwkActivePeriod	—	0.183 s
PPS duration	0.0156 s	—
PPS interval		
Long PPS	3 s	—
Short PPS	1 s	—
Standby duration	5 s	—

For comparative analysis of the proposed MRRC, ZigBee RF4CE is also implemented on the same hardware. RF4CE is designed based on IEEE802.15.4 PHY and MAC, and two different modes are implemented, respectively: RF4CE power saving (PS) and no power saving (NPS), which are specified in RF4CE standard. In NPS, all the target RFs wait for commands from an RF4CE remote control in a fully active mode, and in PS each target maintains repeatedly a duty cycle, which includes active state for *nwkActivePeriod* and sleep state for the remainder period. Table 1 summarizes main parameters used in our experiments.

6.3. Performance Evaluations. In this subsection, the MRRC performances compared with ZigBee RF4CE through various experiments are evaluated. First, how fast a target



(a) RF4CE no power paving

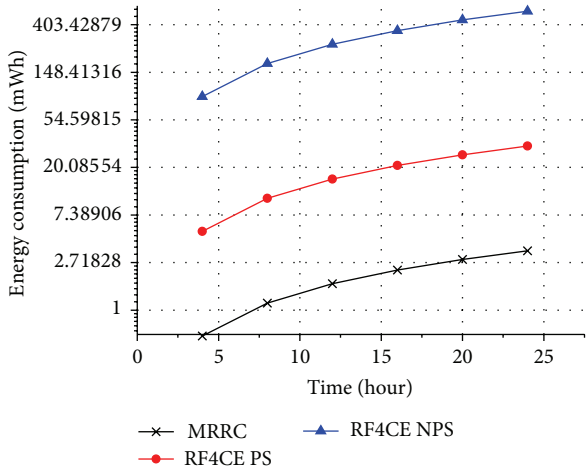
(b) RF4CE power saving

(c) Target RF (MRRC)

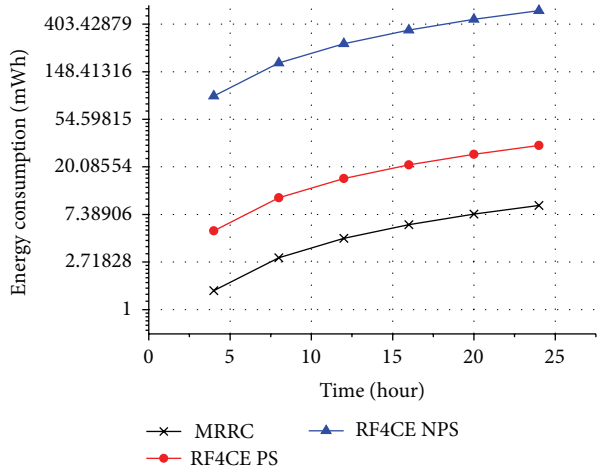
Figure 9: Response time.

system can respond to a user command is evaluated by observing response time. Subsequently, how much energy efficient a target RF and MRRC is evaluated by analyzing energy consumption in various experimental environments. In addition, since PPS duration guaranteeing 100% reception ratio through experiments, as mentioned in Section 4.1, is applied, the experimental result regarding successful data transmission ratio is not presented in this paper.

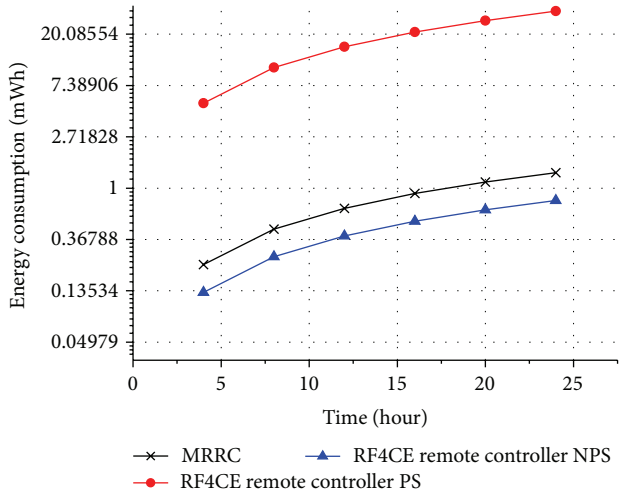
6.4. Response Time Analysis. First, response time of MRRC, RF4CE PS, and RF4CE NPS is observed, respectively. The response time is a round trip time consumed from an instance when user presses a button of an MRRC until the MRRC receives a reply from the target RF. For the experiment, consecutive 10 commands with one second interval at every one minute for total 6 minutes are generated, and each reply time from target RF is measured. Each experiment is repeated 100 times and Figure 9 shows the result that calculates the average of measured value at each experiment. First, the RF4CE NPS presents fast response time less than 250 milliseconds with respect to each command, as shown in Figure 10(a). In the case of RF4CE NPS, since each target RF is always awaken for the packet reception, user command can be processed promptly without any redundant delay. On the other hand, RF4CE PS shows irregular response time distribution as shown in Figure 9(b). The result presents large deviation (300–3,000 milliseconds). In that mode, for power saving a target RF maintains a repeated duty cycle with a period of *nwkdutycycle*, presented in Table 1. Therefore, a target RF should wait to transmit the command until a beacon frame is received from the target RF (in general, the target RF plays a role in a coordinator), and finally the command is transmitted at the active duration of target RF, which is referred to as a superframe duration, in which a beacon frame indicates the beginning of superframe duration. That is, the irregular latency in RF4CE PS results from the fact



(a) Fully standby mode



(b) Fully active mode



(c) Remote control

Figure 10: Energy consumption.

that communications between a remote control and target RF depend on time synchronization based on a beacon transmitted periodically by target RF. Eventually, the irregular response time might bring users inconvenience. Finally, experimental result of MRRC is shown in Figure 9(c). The result shows that response time of the MRRC maintains normally 280 milliseconds, which is similar to the RF4CE NPS. It is noticeable that the initial system-on command shows long response time of 3,300 milliseconds. The MRRC performs long PPS with long interval to save standby energy during system power-off, so that initial power-on command requires longer response time than the following commands. However, since after system power-on each target RF performs short PPS, only the first commands out of consecutive commands at every command interval show response time of 1,400 msec. Furthermore, since the rest of the consecutive commands are transmitted without an extended preamble, they can maintain minimum response time as in RF4CE NPS.

6.5. Energy Consumption. Energy consumption is one of the important performance factors for home appliances and remote control. To evaluate comparative performance, total energy consumed for the test duration under the same test condition is measured for MRRC, RF4CE NPS, and PS, respectively. Energy consumption of a target RF and remote control, respectively, is obtained from power consumed for 24 hours. More specifically, standby energy of target RF in fully standby mode in which a target system is power off but RF is ready to receive user command is observed, and also energy consumption in fully active mode in which various user commands are performed is observed. Since power consumption is different according to each target

system (appliance), for our experiment power consumption of only RF in each target system is considered. All the experimental results present average energy consumption of each RF connected to each appliance emulator.

Figure 10(a) shows energy consumption of a target RF in fully standby mode (system power off) for 24 hours. It is shown that RF4CE NPS having fast response time of 250 milliseconds is considerably inefficient in energy aspect. On the other hand, RF4CE PS shows more efficient energy consumption than NPS. This is because the mode maintains a repeated duty cycle for the standby period. It is noticeable that a target RF in MRRC shows minimum energy consumption over RF4CE. The ultralow standby energy consumption of the MRRC results from maintaining minimum active duration only to detect preamble transmission and utilizing variable length PPS (e.g., long PPS during RF standby period). Furthermore, unlike RF4CE PS, MRRC does not require superframe management based on time synchronization by a periodic beacon frame between a remote control and target RF.

Figure 10(b) shows another experimental result. In contrast with the former experiment (full a RF standby), for this experiment, the target system is turned on and 10 commands are issued at every hour for total 24 hours. The result shows that the MRRC target RF is superior to two RF4CE modes. Two RF4CE modes also show almost similar energy consumption as in fully RF standby mode. This is because the two modes utilize the same power management in both standby mode and active mode. On the other hand, since the MRRC manages different length PPS according to the system power-on/off state, the MRRC can cope well with tradeoff between energy and response time. In particular, even though the target RF consumes more energy over fully standby mode by performing short PPS in system power-on state, the MRRC target RF shows more energy saving than the RF4CE PS.

Figure 10(c) presents experimental result of a remote control. For the experiment, user generates 10 commands using a remote control at every hour for 24 hours. In contrast with the former two experiments in which energy consumption of target RF is only focused, this experiment presents energy consumption of a remote control of MRRC, RF4CE NPS, and RF4CE PS, respectively. The result shows that energy saving of RF4CE NPS remote control is superior to MRRC and RF4CE PS. This is because RF4CE NPS remote control is normally in sleep mode, wakes up only when user command is generated, and returns to sleep mode again. On the other hand, since in the RF4CE PS remote control the generated user command should wait until beacon is received from the target RF, a remote control consumes more energy. In the case of MRRC, the on-demand user command can asynchronously trigger a target RF which is performing PPS. This feature results in minimizing unnecessary energy consumption in a remote control, and thus the MRRC shows similar energy consumption to RF4CE NPS.

7. Conclusion

In this paper a multifunctional RF remote control, which is capable of providing larger coverage and various services, is introduced, and an ultralow standby power operation method for target RFs, utilizing an extended preamble transmission and a variable length PPS according to system power state, is proposed. Furthermore, a target RF can promptly respond to on-demand user command by being asynchronously triggered by an MRRC. In addition, based on bidirectional communication between an MRRC and target RF, user can control multiple target systems with a single remote control.

A prototype and implementation details are also described. To evaluate the proposed MRRC, several experiments are conducted, and each performance of MRRC is also compared with ZigBee RF4CE NPS and PS. The experimental results demonstrated that the MRRC system enables not only ultralow standby power in system power-off state but also low power operation even in system active state. In spite of ultralow standby power operation, the experimental result also shows that the MRRC provides reasonable response time to user command. Finally, it is expected that these outstanding features of MRRC will be able to contribute to constructing ultralow power home network associated with smart appliances.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the Incheon National University Research grant in 2012.

References

[1] C.-H. Hung, Y.-W. Bai, P.-W. Chen, and J.-M. Hsu, "Remote control of home lighting devices with RFID identification and current detection of outlets," in *Proceedings of the IEEE 17th International Symposium on Consumer Electronics (ISCE '13)*, 2013.

[2] T. F. Chueh and Y. Y. Fanjiang, "Universal remote control on smartphone," in *Proceedings of the International Symposium on Computer, Consumer and Control (IS3C '12)*, 2012.

[3] L. Kim, W. Park, H. Cho, and S. Park, "An universal remote controller with haptic interface for home devices," in *Proceedings of the International Conference on Consumer Electronics (ICCE '10)*, pp. 209–210, January 2010.

[4] Y. Park and M. Lee, "Cost effective smart remote controller based on invisible IR-LED using image processing," in *Proceedings of the IEEE International Conference on Consumer Electronics*, January 2013.

[5] J. Han, H. Lee, and K. Park, "Remote-controllable and energy-saving room architecture based on ZigBee communication," in *Proceedings of the International Conference on Consumer Electronics (ICCE '09)*, January 2009.

[6] I.-K. Hwang, D. S. Lee, and J. W. Baek, "Home network configuring scheme for all electric appliances using ZigBee-based integrated remote controller," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1300–1307, 2009.

[7] T. Kim, H. Lee, and Y. Chung, "Advanced universal remote controller for home automation and security," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2537–2542, 2010.

[8] ZigBee RF4CE Standard, *Specifications for ZigBee Radio Frequency for Consumer Electronics*, ZigBee Alliance, San Ramon, Calif, USA, 2009.

[9] J. Polastre, J. Hill, and D. Culler, "Versatile low power media access for wireless sensor networks," in *Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems (SenSys '04)*, pp. 95–107, New York, NY, USA, November 2004.

[10] A. B. Nacef, S. Senouci, Y. Ghamri-Doudane, and A. Beylot, "A cooperative low power Mac protocol for wireless sensor networks," in *Proceedings of the IEEE International Conference on Communications (ICC '11)*, June 2011.

[11] K.-I. Hwang, "Energy efficient channel agility utilizing dynamic multi-channel CCA for ZigBee RF4CE," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, pp. 113–119, 2011.

[12] C.-H. Tsai, Y.-W. Bai, C.-A. Chu, C.-Y. Chung, and M.-B. Lin, "PIR-sensor-based lighting device with ultra-low standby power consumption," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 3, pp. 1157–1164, 2011.

[13] S. Kang, K. Park, S. Shin, K. Chang, and H. Kim, "Zero standby power remote control system using light power transmission," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 4, 2011.

[14] J. Han, S.-C. Choi, and I. Lee, "More efficient home energy management system based on ZigBee communication and infrared remote controls," in *Proceedings of the IEEE International Conference on Consumer Electronics (ICCE '01)1*, pp. 631–632, January 2011.

[15] IEEE Std. 802.15.4, "Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (WPANs)", IEEE Standard for Information Technology.

[16] M. Buettner, G. V. Yee, E. Anderson, and R. Han, "X-MAC: a short preamble MAC protocol for duty-cycled wireless sensor networks," in *Proceedings of the 4th International Conference on Embedded Networked Sensor Systems (SenSys '06)*, pp. 307–320, ACM, New York, NY, USA, November 2006.

[17] A. El-Hoiydi and J.-D. Decotignie, "WiseMAC: an ultra low power MAC protocol for the downlink of infrastructure Wireless Sensor networks," in *Proceedings of the 9th International Symposium on Computers and Communications*, vol. 1, pp. 244–251, July 2004.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 672968, 9 pages
http://dx.doi.org/10.1155/2014/672968

Research Article

Collusion Based Realization of Trust and Reputation Models in Extreme Fraudulent Environment over Static and Dynamic Wireless Sensor Networks

Vinod Kumar Verma,[1] Surinder Singh,[2] and N. P. Pathak[3]

[1] *Department of Computer Science & Engineering, Sant Longowal Institute of Engineering & Technology, Longowal 148106, India*
[2] *Department of Electronics & Communication Engineering, Sant Longowal Institute of Engineering & Technology, Longowal 148106, India*
[3] *Department of Electronics & Communication Engineering, IIT, Roorkee 247667, India*

Correspondence should be addressed to Surinder Singh; surinder_sodhi@rediffmail.com

Received 11 November 2013; Accepted 26 March 2014; Published 26 May 2014

Academic Editor: Ken Choi

We derive a new representation for the collusive sensor nodes when the underlying fraudulent correlated environment has strong influence on wireless sensor networks performance. We have evaluated collusion effect with respect to static (SW) and dynamic (DW) wireless sensor networks to derive the joint resultant. Moreover accuracy, path length, and energy consumption of sensor node operations are also evaluated. Additionally, we emphasized over the satisfaction evaluation for linguistic fuzzy trust and reputation (LFTM) models in the deployed WSN framework. Finally, simulation analysis has been carried out to prove the validity of our proposal. However, collusion for wireless sensor networks seems intractable with the static and dynamic WSNs when varied with specified number of fraudulent nodes in the scenario.

1. Introduction

Rapid development in the area of communications through wireless sensor networks (*WSN*) attracted more attention of scientists and researchers over the last few years [1]. Wireless sensors are small-sized devices equipped with radio transceivers and low power batteries. Typical features of sensor node include power, storage, and low cost computational capability hardware [2, 3]. A wireless sensor network is designed to sense, collect, process, and transmit event specific information in order to accomplish a distributed domain task. Moreover wireless sensor networks [4, 5] are the type of networks where the resultant is fully based on the sensor nodes cooperation. The potential of wireless sensor networks can be equally deployed in the wider area of applications such as defense equipment, ecological and habitat monitoring, industrial process control, home automation, weather forecasting, health care system, traffic control, and civilian applications. Usually wireless sensor networks are deployed in an open informant where the probability of an adversary [6] always remains more than in a closed environment. These malicious nodes may spread wrong information on the entire network which results in overall system performance degradation. Therefore, it is quite mandatory to identify the collusive nodes and punish them in an accord manner. There are numerous proposals to detect an adversary node in the wireless sensor networks. Traditional means to protect a network include cryptography specific techniques and methodologies. Complex computations in the cryptography strategies [7] become its major drawbacks and made these policies unsuitable to be deployed in wireless sensor network which constitutes severe power constraints. Some lightweight cryptographic mechanisms are available in the literature, but they are not serving the goal in entirety. Therefore, there remains a dire need to probe for wireless sensor network reliability aspect and search some complementary means to incorporate more faith in the overall scenario. Trust and reputation models are the solution for the given problem to adhere to reliability in the wireless sensor networks. This is the reason why research on trust and reputation models has

gained considerable momentum in the last few years. Many trust and reputation models have been proposed in the past. Some of them were centered around secure routing, data aggregation, cluster head selection, and synchronized trust management [8–10] but still there is need to address various issues like collusion, scalability, mobility, and computability in the wireless sensor networks. At present, most of the trust evaluation frameworks belong to an algorithm based methodology, over which entire behavior of nodes depends on accuracy, resource usability, and energy consumption. It is therefore necessary to concentrate on these issues in parallel with their performance and some real time aspects like collusion and fraudulent environment. Collusion can be referred to as a specific level of probability with which every malicious server will assign maximum rating to other malicious servers and minimum rating to the benevolent server. As a result, the most obedient node will become unable to contribute its services to the WSN system for most of the times, which further severely affect the overall system performance. Presence of malicious servers in the wireless sensor networks is the major and real root cause behind the collusion parameter. Therefore the specific issue like collusion must be addressed to enhance the capability of the entire WSN framework. So, we selected five popular trust and reputation models for their comparison and evaluation in terms of accuracy, path length, satisfaction, and energy consumption. This research focuses on the collusion issue and presented our analysis in the extreme fraudulent environment.

Section 2 reported surveys of five trust and reputation models with prior work on wireless sensor networks. Section 3 highlights our motivation for research work. Section 4 presented the problem definition and system model. Section 5 describes the detailed design of our experimental setup. Simulation results and validations are presented and discussed in Section 6. Finally, conclusions are made in Section 7.

2. Trust and Reputation Models with Related Prior Work

This section provides the background and related work on trust and reputation models in wireless sensor network with assumptions required for the designed frameworks for the later sections.

2.1. Eigen Trust Model. It is one of the most commonly used trust and reputation models in the wireless sensor network domain. Kamvar et al. [11] evaluated this model on the basis of the peer's history of contributions by assigning a unique global trust value in the peer-to-peer file system for each peer [12, 13]. Further into this model, the authors define S_{ij} as the local trust of peer i about peer j, in the following manner: $S_{ij} = \text{sat}(i, j) - \text{unsat}(i, j)$. It shows the difference between satisfactory and unsatisfactory interaction between peers: (i, j). Further, the authors define normalized local trust value,

$$Cij = \frac{\max(Sij, 0)}{\sum_j \max(Sij, 0)}. \quad (1)$$

It ensures that all the value lies in between 0 and 1. The authors also introduced aggregated local trust value which is defined as $t_{ik} = \sum_j C_{ij}C_{jk}$, where t_{ik} represents the trust that peer i places in peer k based on friends information. This model also incorporates three practical issues like a priori notion of trust, inactive peers, and malicious collectives. First of all, in the presence of malicious peers, $t = (C^T)^n p$ will generally converge faster than $t = (C^T)^n e$, so we use p as our start vector. In the case of inactive peers, C_{ij} refined as

$$Cij = \begin{cases} \dfrac{\max(Sij, 0)}{\sum_j \max(Sij).} & \text{if } \sum_j \max(Sij, 0) \neq 0; \\ \text{otherwise} & Pj. \end{cases} \quad (2)$$

The malicious collectives issue was addressed by the following equation in this model:

$$t^{(k+1)} = (1 - a)\, C^T t^{(k)} + ap \quad \text{where } a < 1. \quad (3)$$

2.2. Peer Trust Model. Xiong and Liu [14] in this model combine many aspects related to the trust and reputation management such as the feedback a peer receives from other peers, the total number of transactions of a peer, the credibility of the recommendations given by a peer, the transaction context factor, and the community context factor. The trust value of peer u, $T(u)$ is represented by the following expression:

$$T(u) = \alpha \sum_{i=1}^{I(u)} S(u, i)\, CR\left(p(u, i)\, TF(u, i) + \beta CF(u)\right), \quad (4)$$

where $I(u)$ represents total number of transactions performed by peer u with all other peers, $p(u, i)$ represents other participating peer in peer u's ith transaction, $S(u, i)$ represents normalized amount of satisfaction peer u receives from $p(u, i)$ in its ith transaction, $CR(v)$ represents credibility of the feedback submitted by v, $TF(u, i)$ represents adaptive transaction context factor for peer u's ith transaction, and $CF(u)$ represents adaptive community context factor for peer u. On the other hand, the credibility of v from w's point of view is computed as

$$Cr\left(p(u, i)\right) = \frac{\text{Sim}\left(p(u, i), w\right)}{\sum_{j=1}^{I(u)} \text{Sim}\left(p(u, j), w\right)}, \quad (5)$$

where

$$\text{Sim}(v, w) = 1 - \left(\sum_{x \epsilon IJS(v,w)} \left(\left(\frac{\sum_{j=1}^{I(x,v)} \text{Sim}(x, i)}{I(x, v)} - \frac{\sum_{j=1}^{I(x,w)} \text{Sim}(x, i)}{I(x, w)} \right)^2 \times (IJS(v, w))^{-1} \right) \right)^{1/2}. \quad (6)$$

$I(u, v)$ represents the total number of transactions performed by peer u with peer v, $IS(v)$ represents a set of peers that have interacted with peer v, and $IJS(v, w)$ denotes a common set of peers which interacted with peers v and w for $IS(v) \cap IS(w)$ computation. The stimulation for the community for the incentive or rewards is done through the context factor, with the following expression: $CF(u) = F(u)/I(u)$, where $F(u)$ represents the total number of feedback peer u gives to others.

2.3. Bioinspired Trust and Reputation Model (BTRM-WSN). This model for wireless sensor networks is based on the bioinspired algorithm of ant colony system [15–19]. In this model, most trustworthy path leads to finding the most reputable service provider in a network. WSN launches a set of artificial agents while searching for a most reputable service provider. In order to carry out a decision about next sensor, a probability is given to each arc by the following expression:

$$pk(r, s) = \begin{cases} \dfrac{[\tau_{rs}]^{\alpha}[\eta_{rs}]^{\beta}}{\sum [\tau_{ru}]^{\alpha}[\eta_{ru}]^{\beta}} & \text{if } s \epsilon Jk(r); \\ \text{otherwise} & 0. \end{cases} \quad (7)$$

The following equation represents modification of the ants [20] pheromone trace:

$$\tau_{s1s2} = (1 - \varphi)\tau_{s1s2} + \varphi\Omega, \quad (8)$$

where $\Omega = (1 + (1 - \varphi)(1 - \tau_{s1s2}\eta_{s1s2}))\tau_{s1s2}$ denotes the convergence value of τ_{s1s2} and φ represents a parameter controlling the amount of pheromone. The best path found by all ants is given by

$$\tau_{rs} = (1 - \rho)\tau_{rs} + \rho\left(1 + \tau_{rs}\eta_{rs}Q\left(S_{\text{Global}_{\text{Best}}}\right)\right)\tau_{rs}, \quad (9)$$

where $Q(S_{\text{Global}_{\text{Best}}})$ denotes path quality. The quality of the S_k paths can be measured as the average of all the edges belonging to that path

$$Q(S_k) = \frac{\tau k}{\sqrt{\text{Length}(S_k)}}\%A_k, \quad (10)$$

where $\%A_k$ denotes the percentage of trustworthy paths. The punishment or rewards of the path leading to the selected peer are given by

$$\tau_{rs} = (\tau_{rs} - \varphi \times df_{rs})\frac{\text{Sat}}{df_{rs}}. \quad (11)$$

The distance factor joining the link between sensor r and s is given by the following equation:

$$df_{rs} = \sqrt{\frac{df_{rs}}{L(S_k)(L(S_k) - d_{rs} + 1}}. \quad (12)$$

2.4. LFTM Model. This linguistic fuzzy trust model [21] uses the concept of fuzzy reasoning. On one hand, it uses the representation power of linguistically labeled as fuzzy sets for the satisfaction of a client or the goodness of a server. On the other hand, it remains affected by the inference power of fuzzy logic, as in the imprecise dependencies between the originally requested service and the actual received one, or the punishment to apply in case of fraud. The expected result will be an easily interpretable system with adequate performance. In this model, a set of linguistic labels describing several levels of a variable or concept could be associated with a fuzzy set. The resultant set constitutes linguistic labels such as "very low," "low," "medium," "high," and "very high." These defined fuzzy sets associated with such labels specify the level of client satisfaction.

2.5. Trust and Reputation Infrastructure Based Proposal (TRIP) Models. This model is based on the environment specific issues like infrastructure, area, density, and so forth, within the specified conditions [22]. Every time a node receives a signal from the other node, it assesses the reputation of the node in order to reject or drop the message based on the trustworthiness of that node. Each message depicts its actual level of importance or risk. Even the harmful message will not affect the system because of the fact that each message constitutes its trust level. The higher the trust level, the better the probability for its selection. Additionally, a reputation score calculation for each message is based on three different aspects, namely, (i) information directly from the targets, (ii) information from neighbor nodes, and (iii) information from the central unit. Informational database from all the three sources can be stored in the central unit. Finally, taking into consideration the entire information the best and appropriate decision can be easily taken.

3. Motivation for Current Work

To choose accurate trust and reputation models remains the top priority for the performance assessment of wireless sensor networks. Optimal trust and reputation models enhance the performance of the overall system about information dissemination, but the wireless sensor network system may not be dependent on the same. A simple trust and reputation modeling strategy may give the best result for a single instance but we have to deploy such efficient trust and reputation modeling strategies that provide optimal results in data dissemination. The improper modeling strategy may overload the entire network and consume more resources both in terms of energy and computation which result in the entire system performance degradation. There always remains dire influence of trust and reputation strategy on the entire operating environment when evaluating a specific wireless sensor network. The goal which remains there is to carefully choose and examine the trust and reputation modeling strategies for information dissemination and present an optimal result without compromising any constraints than the expected outcome. Therefore, a typical realization should be required to access the scope of a particular trust and reputation model strategy for the wireless sensor networks.

4. Problem Definition and System Model

In our analysis, we consider ten networks composed of two hundred sensor nodes, each for twenty scenarios in two-dimensional fields. Sensor nodes in a cluster with a specific radio range transmit the data to the cluster head and then to the base station within the entire network. Network deployment focuses on collusion and fraudulent conditions. Although any trust and reputation sensor node strategy can be used in our model, we used LFTM, BTRM, and peer trust model with static and dynamic wireless sensor network for our proposed framework. Static wireless sensor network can be referred to as a mode of communication where the position of all the nodes remains stationary, whereas in case of dynamic wireless sensor network, the nodes can change their positions in an accord manner. Accordingly, for a given network with static and dynamic wireless sensor network and trust and reputation models node strategy described above, we are interested in finding the following two problems: (i) what is the influence of collusion on static and dynamic communication node operations in the wireless sensor networks and (ii) how collusion affects the accuracy, path length, and energy consumption for different trust and reputation models in wireless sensor network.

5. Detailed Setup

We focused on three parametric aspects, namely: accuracy, path length, and energy consumption for information dissemination in wireless sensor networks. For this, we have developed the unmitigated scenario pinpointing two main targets. Firstly, we are interested in finding the value of three above-mentioned parameters for static wireless sensor network with and without collusion aspect. We want to know the summation of all the node operations with respect to collusion parameter. Lesser path length of node operation always gives due attention as it consumes fewer resources and exhibits more efficiency. Secondly, we want to make an estimation of the mobility effects on communication performance in correlation with the collusion for different trust and reputation models. Finally, we made the comprehensive evaluation of energy consumption with static and dynamic wireless sensor networks in our proposed framework. We designed a wireless sensor network template using the following parameters: 20% of all nodes in a randomly created WSN acted as clients where and the rest 80% of nodes acted as servers. Client nodes refers to the percentage of nodes which want to have or ask for services in a WSN. 5% of the nodes acted as relay servers which do not offer any services and act as relay nodes. The radio range of the nodes set at 10 hops to its neighbors. We consider a scenario where the percentage of fraudulent servers remained 70% which specifies the indispensable condition for our WSN framework evaluation. Fraudulent servers depict the percentage of adversaries in a wireless sensor network. We set the minimum and maximum numbers of nodes that can create a WSN equal to 200. Sensor nodes belonging to our developed networks spread over the area of 100 m × 100 m. A total of ten networks were examined and the final results reflect the average value of all the networks. The process of searching trustworthy server was carried out ten times for each network. Table 1 shows the summary of parameters deployed in our model.

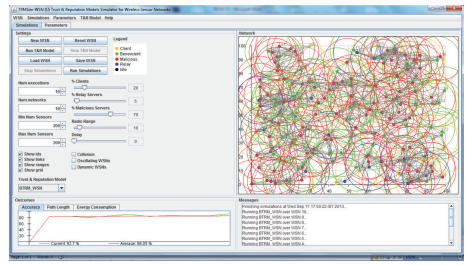


Figure 1: Simulation scenario.

Table 1: Scenario parameters.

Scenario options	Value
% client	20
% relay server	5
% fraudulent server	70
Radio range	10
Delay	0
Number of execution	10
Number of network	10
WSN area	100 m × 100 m
Minimum number of nodes	200
Maximum number of nodes	200
WSN orientation	Static, Dynamic

Figure 1 shows the setup of the simulation. In the simulation window, yellow dots denote client nodes, green dots represent the benevolent nodes, red dots denote malicious node, blue dots represent relay nodes, and black dot denotes idle nodes respectively.

6. Analytical Results and Validations

This section enables us to implement and evaluate trust and reputation models for different wireless sensor network modes. We used Java based event driven TRMSim-WSN simulator [23] version 0.5 forwireless sensor network allowing the researchers to simulate and represent random network distributions and provides statistics of different data dissemination policies including the provision to test the different trust and reputation models' strategies. Many decisions, like static or dynamic or oscillating networks, a combination of dynamic and oscillatory networks, the percentage of fraudulent nodes, the percentage of nodes acting as clients or servers, and so forth, can be implemented and tested over it. The proposed model is tested on five different trust and reputation models with extreme fraudulent conditions. We reported a comprehensive analysis based on collusion with static and dynamic wireless sensor networks. We collected data for four metrics, namely, accuracy, path length, satisfaction, and energy consumption. We investigated the comparative analysis of trust and reputation models

with static WSN and dynamic in contrast with and without collusion parameter. Static node refers to the type of nodes whose position remains fixed and whereas the dynamic node can be mobile in the network. We considered four WSN modes, namely, (i) static WSN (SW), (ii) static WSN with collusion (SWC), (iii) dynamic WSN (DW), and (iv) dynamic WSN with collusion (DWC). We denote Eigen trust model with value 1, peer trust model with 2, BTRM-WSN model with value 3, LFTM with value 4, and TRIP model with value 5. The outcome of the simulations will be subject to the following subsections.

6.1. Accuracy. The term accuracy in the trust and reputation systems may be defined as the selected percentage of trustworthy nodes. We calculated accuracy parameter in terms of their current and average values. Current accuracy denotes the trustworthiness value calculated for the last node, whereas average accuracy presents the value of all nodes available in the mentioned framework. Initially, we calculated average accuracy correspond to different trust and reputation models as reported in Figure 2. The value of current accuracy remains highest in case of static WSN as compared to the rest of the WSN modes because of the fact that static nodes are less prone to failure than the dynamic as well as the combination of static and dynamic WSN with collusion aspect.

Next, we considered the second evaluation for average accuracy with the same WSN framework. According to Figure 3, again average accuracy shows the similar behavior with the current accuracy in Figure 2 above as the value of average accuracy remains highest in case of static WSN than the rest of the WSN modes. For static WSN (SW) and dynamic WSN (DW) modes, the value of current and average accuracy remains highest in LFTM model than other models in most fraudulent conditions, whereas TRIP model depicts the minimum value. In case of static WSN with collusion (SWC) and dynamic WSN with collusion (DWC) mode, the Eigen trust model outperforms the rest of the models in current and average accuracy values, whereas peer trust model shows minimum accuracy value. We have also presented the scalability impact on the wireless sensor network [24]. We enhanced this evaluation towards a bit of an intricate assessment by incorporating collusion, malicious servers, resource utilization, satisfaction, and energy evaluation aspect on a single platform. One common point we have noticed is that there is severe effect of collusion and mobility of nodes on the accuracy of WSN system, as the accuracy declines to a bit of an intricate level when a node changes its state from static to dynamic.

6.2. Path Length. The next parameter of our concern is path length which can be defined as the number of resources a particular network utilizes with a particular trust and reputation model. In the consistent pattern of accuracy evaluation types, we evaluated the current and average path length on the similar pattern of accuracy for all the WSN modes. Current path length depicts the resource utilization value calculated for the last node, whereas average path length exhibits the value of all nodes present in the scenario.

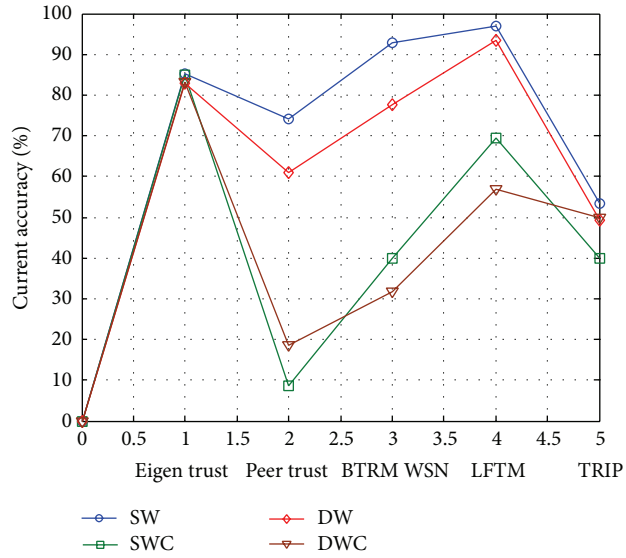


Figure 2: Current accuracy of different WSN modes with trust and reputation models.

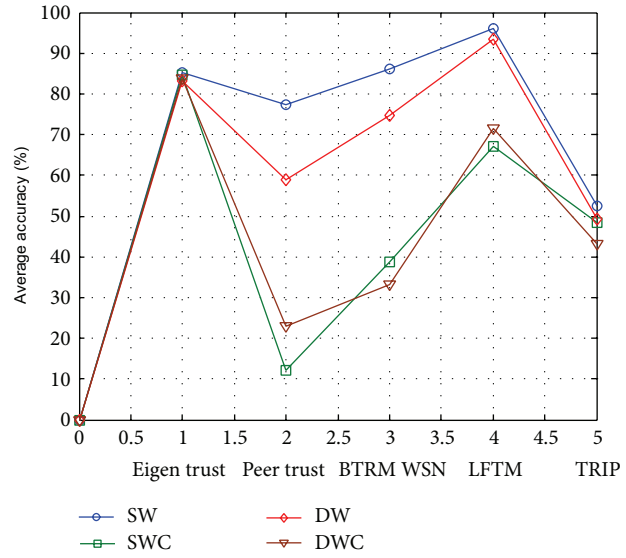


Figure 3: Average accuracy of different WSN modes with trust and reputation models.

Figures 4 and 5 represent the value of current and average path length which remains quiet in case of TRIP model for both the current and average case viewpoints than other models. This is due to the fact that the TRIP model constitutes the fixed infrastructure for its functionality resulting in lesser path length as compared to other models. Among the rest of the models, LFTM model consumes lesser path length than the rest of the models in the case of SW mode and DW mode, whereas the SWC and DWC modes of BTRM utilize the minimum path length.

We also observed that BTRM utilizes the maximum path length of all the SW, SWC, DW, and DWC WSN modes. This shows the excellent agreement with the results reported

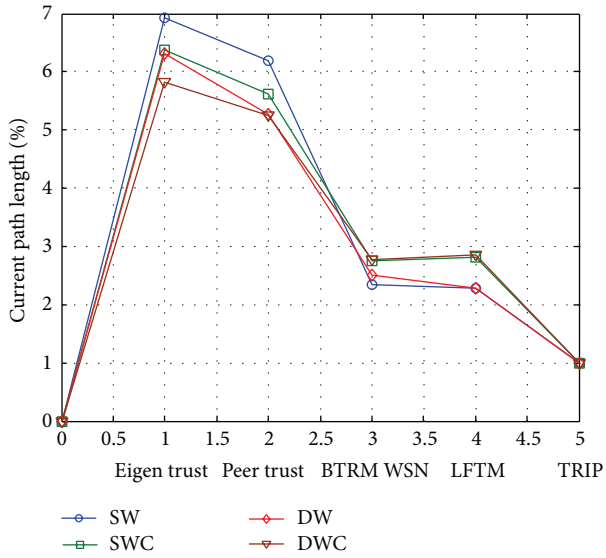


Figure 4: Path length of different WSN modes with trust and reputation models.

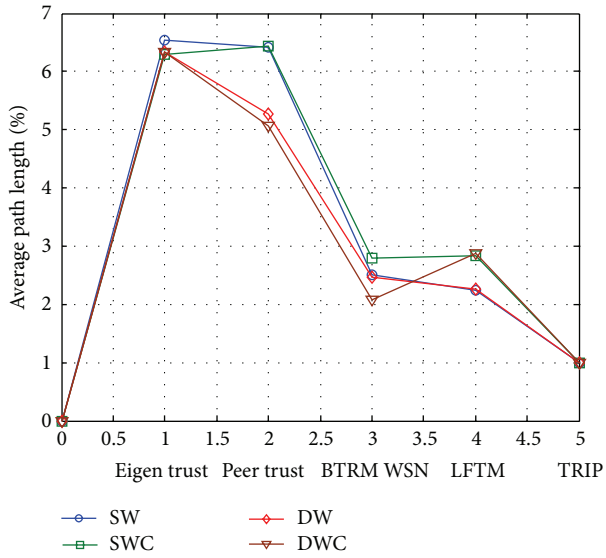


Figure 5: Average path length of different WSN modes with trust and reputation models.

in reference [25]. An initiative towards the description of energy consumption analysis for different trust and reputation models was proposed in reference [25]. We enhanced this evolution towards a bit of a complex assessment by incorporating collusion, satisfaction, and energy evaluation aspect in our scenario. Moreover in the later energy consumption subsection we proposed a mathematical equation for overall energy consumption which adds more robustness in our evaluation.

We proposed a more robust framework subsuming different WSN versus collusion scalability on a single platform. Xiong and Liu [14] reported peer-to-peer trust and reputation based model for structured peer-to-peer networks

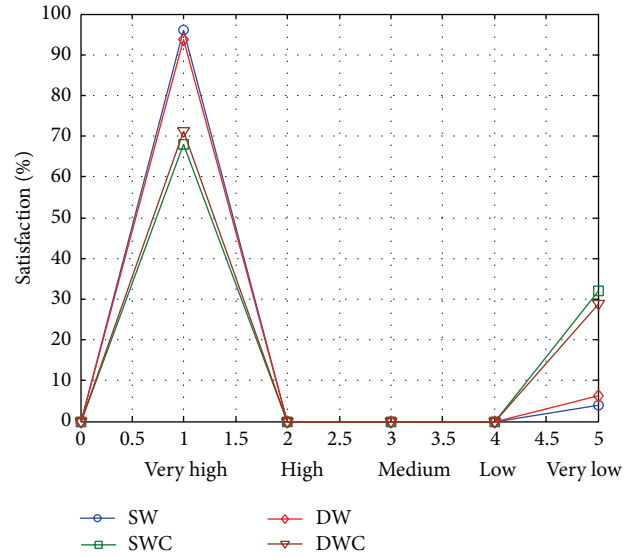


Figure 6: Satisfaction analysis of different WSN modes with LFTM trust and reputation model.

including strategies for their implementation and evaluation in decentralized environmental conditions. Also, Xiong and Liu [14] emphasized over trust metric in order to assess trustworthiness, feedback, and credibility of peer-to-peer networks. Specifically, unstructured peer-to-peer networks based on parameters were proposed by Chen et al. [26]. We enhanced the contribution to a certain extent by incorporating collusion, satisfaction, and energy consumption parameters for wireless sensor network evaluation making our investigation more robust and real time.

6.3. Satisfaction. Additionally, we calculated the satisfaction level of different WSN modes for LFTM model as shown in Figure 6. In the context of trust and reputation models, satisfaction can be defined as a particular level of subjectivity up to a specific degree in which the system can behave as per desired goal for mentioned probability. The observation shows that in the static mode satisfaction level is very high as compared to the rest of WSN modes, whereas incorporation of collusion to SW and DW modes decreases its value.

6.4. Energy Concerns. One of the major issues, when dealing with the wireless sensor network, is energy consumption. So, lastly we emphasized on the average energy consumption by SW, SWC, DW, and DWC modes of five trust and reputation models in wireless sensor networks. The power requirement of a sensor node can be analyzed as a function of distance as per the references [27, 28]. For most of the models, energy consumption E by a message at a distance d is given by [16, 29]

$$E(d) = d^{\alpha} + C, \tag{13}$$

where α represents attenuation factor and C is constant used for radio signal and dimensionless.

Table 2: Energy consumption for trust and reputation models with WSN modes.

Trust and reputation models with WSN modes	Eigen trust (mJ)	Peer trust (mJ)	BTRM (mJ)	LFTM (mJ)	TRIP (mJ)
Static WSN (SW)	8.2×10^{24}	8.7×10^{16}	4×10^{17}	2.5×10^{17}	4×10^{9}
Static WSN with collusion (SWC)	6.4×10^{24}	1.0×10^{17}	8×10^{17}	1.4×10^{18}	4×10^{9}
Dynamic WSN (DW)	5.1×10^{24}	3.2×10^{16}	1.1×10^{17}	1.2×10^{17}	4×10^{9}
Dynamic WSN with collusion (DWC)	8.6×10^{24}	2.2×10^{16}	6.8×10^{17}	7.6×10^{17}	4×10^{9}

Table 2 compares the five trust and reputation models from the energy consumption aspect as shown below. Reference [30] reported a comparative analysis of the energy consumption with respect to sensors value increment.

In our proposal, we extended this concept towards the different WSN modes and simultaneously clubbing these modes with different trust and reputation models. We observed that the Eigen trust model consumes maximum power in all the SW, SWC, DW, and DWC modes, whereas TRIP model reported minimum energy consumption. We observed that more complexity involvement in the Eigen trust model is the reason for utmost energy consumption and in case of TRIP model energy consumption being the minimum because of the simpler computation involved in trust value computation. We also extended the mathematical relation as reported in references [27, 28] for the energy consumption for trust and reputation models in our framework

$$E_o = E_c + E_s + E_F + E_R + E_{\text{Sim}}, \tag{14}$$

where E_o represents overall energy consumption, E_c denotes client nodes energy consumption, E_s depicts server nodes energy consumption, E_F shows energy consumption for fraudulent node, E_R denotes relay node energy consumption, and E_{Sim} denotes energy consumption used by the simulator.

Overall, we investigated the entire framework twenty times for different WSN modes and corresponding five trust and reputation models. One common thing we observed is that more complexity in any trust and reputation model attracts more resources utilization and power consumption. We added a variety of evaluation strategies based on accuracy, path length, satisfaction, and power consumption for sensor node operations in our proposed framework which make over a scenario more robust as compared to the approach reported by Pan et al. [31]. A new trust and reputation model by adding additional constraints to BTRM-WSN adopting an interactive multiple ant colony algorithm was also suggested in reference [31]. We extended the concept by adding more robust constraints like static, dynamic, collusive, and a combination of all these aspects on a single platform for the trust and reputation models investigations in wireless sensor networks. Qureshi et al. [20] presented FIRE trust and reputation model extension to detect and prevent direct interaction and validate interaction collusion attacks in wireless networks. We enhanced this concept of reference [20] for wireless sensor networks with collusion and satisfaction aspect evaluation with five trust and reputation models over wireless sensor networks. Our analysis shows that there remains always significant impact of collusion over static and dynamic mode of WSN, resulting in the performance degradation of the overall system.

7. Conclusions

This paper concluded the impact of collusion on different trust and reputation models in wireless sensor networks. We have observed the effect of collusion for static, dynamic, and collusive sensor nodes in a WSN framework. It is evident from the simulation that there is a strong relationship between collusion and WSN modes in trust and reputation model evaluation. We evaluated a wireless sensor network framework for collusion aspect with reference to four performance metrics, namely: accuracy, path length, satisfaction, and energy consumption viewpoint. We estimated accuracy and path length in terms of overall percentage of the functionality, whereas energy consumption in terms of millijoule specifically for sensor node operations. The performance of the WSN system changes along with the different WSN modes and collusion present in the scenario. We mainly concentrated toward the comparative evaluation of static, dynamic, and collusive WSN modes deployed in our designed model. Our research work presented a comprehensive investigation over collusion parameters with five trust and reputation models. We stressed on three major directions. Firstly, we evaluated accuracy, path length, satisfaction, and energy consumption for collusive and non-collusive modes of wireless sensor networks. Secondly, we investigated the entire framework for comparative evaluation of above-discussed trust and reputation models, and lastly the same model is deployed for the mathematical derivation of the energy equation of a wireless sensor network. We observed that with the collusion adoption in the WSN modes, the result becomes much steeper, that is, performance degradation. In case of static nodes, the collusion affects less to WSN when it is incorporated in dynamic mode. Also, node operations remain more in case of collusion than without it. From this investigation, we can predict that the lesser the collusive nodes the more the probability of accuracy, the better resource utilization, the adequate satisfaction level, and the lesser the energy consumption of the entire WSN will be exhibited by the wireless sensor network system. In the future, we would like to develop further trust and reputation models in our evaluation as well as work towards additions on newer distribution strategies for the wireless sensor network domain. Finally, this work allows us to analytically formulate investigative strategies under specified

scenarios and therefore provides insight for directing the designated model for wireless sensor network evolution.

Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank the Department of Information and Communication Engineering, University of Murcia (Spain) [23], for TRMSim-WSN simulator for wireless sensor network which greatly supports them for their research work. Authors are also thankful to NetSim simulator developers. Additionally, they would like to thank the Department of Electronics and Communication engineering, Sant Longowal Institute of Engineering & Technology, Longowal, India, for providing them with Wireless SignalPro software which helps them in the final result preparation. Last but not least, the authors would like to thank the reviewers for their valuable suggestions which bring the paper in present form.

References

[1] S. Farahani, *ZigBee Wireless Networks and Transceivers*, Elsevier, Oxford, UK, 2008.

[2] A. Alkalbani, T. Mantoro, and A. O. Md. Tap, "Improving the lifetime of wireless sensor networks based on routing power factors," in *Networked Digital Technologies*, vol. 293 of *Communications in Computer and Information Science*, pp. 565–576, Springer, Berlin, Germany, 2012.

[3] H. Chen, H. Wu, X. Zhou, and C. Gao, "Reputation-based trust in wireless sensor networks," in *Proceedings of the International Conference on Multimedia and Ubiquitous Engineering (MUE '07)*, pp. 603–607, Seoul, Republic of Korea, April 2007.

[4] C.-Y. Chong and S. P. Kumar, "Sensor networks: evolution, opportunities, and challenges," *Proceedings of the IEEE*, vol. 91, no. 8, pp. 1247–1256, 2003.

[5] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, 2002.

[6] J. Hurt, Y. Lee, H. Yoont, D. Choi, and S. Jin, "Trust evaluation model for wireless sensor networks," in *Proceedings of the 7th International Conference on Advanced Communication Technology (ICACT '05)*, pp. 491–496, Phoenix Park, Republic of Korea, February 2005.

[7] Q. Jing, L.-Y. Tang, and Z. Chen, "Trust management in wireless sensor networks," *Journal of Software*, vol. 19, no. 7, pp. 1716–1730, 2008.

[8] J. Hur, Y. Lee, S. Hong, and H. Yoon, "Trust-based secure aggregation wireless sensor networks," in *Proceedings of the 3rd International Conference on Computing, Communications and Control Technologies*, vol. 3, pp. 1–6, 2005.

[9] G. V. Crosby, N. Pissinou, and J. Gadze, "A framework for trust-based cluster head election in wireless sensor networks," in *Proceedings of the 2nd IEEE Workshop on Dependability and Security in Sensor Networks and Systems (DSSNS '06)*, pp. 13–22, Columbia, Md, USA, April 2006.

[10] Y. L. Sun, W. Yu, Z. Han, and K. J. R. Liu, "Information theoretic framework of trust modeling and evaluation for ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 2, pp. 305–317, 2006.

[11] S. Kamvar, M. Schlosser, and H. Garcia-Molina, "The Eigen trust algorithm for reputation management in P2P networks," in *Proceedings of the 12th International World Wide Web Conference (WWW '03)*, Budapest, Hungary, May 2003.

[12] "Advogato's trust metric (white paper)," 2000, http://www.advogato.org/trust-metric.html.

[13] J. Douceur, "The Sybil attack," in *Proceedings of the First International Workshop on Peer-to-Peer Systems (IPTPS '02)*, Cambridge, Mass, USA, March 2002.

[14] L. Xiong and L. Liu, "PeerTrust: supporting reputation-based trust for peer-to-peer electronic communities," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 7, pp. 843–857, 2004.

[15] M. Dorigo and L. M. Gambardella, "Ant colony system: a cooperative learning approach to the traveling salesman problem," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 53–66, 1997.

[16] M. Dorigo, L. Gambardella, M. Birattari, A. Martinoli, R. Poli, and T. Stützle, *Ant Colony Optimization and Swarm Intelligence*, vol. 4150 of *Lecture Notes in Computer Science*, Springer, Berlin, Germany, 2006.

[17] O. Cordón, F. Herrera, and T. Stützle, "A review on the ant colony optimization meta heuristic: basis, models and new trends," *Mathware & Soft Computing*, vol. 9, no. 2-3, pp. 141–175, 2002.

[18] M. Dorigo and T. Stützle, *Ant Colony Optimization*, Bradford Book, 2004.

[19] F. G. Mármol and G. M. Pérez, "Providing trust in wireless sensor networks using a bio-inspired technique," *Telecommunication Systems*, vol. 46, no. 2, pp. 163–180, 2011.

[20] B. Qureshi, G. Min, and D. Kouvatsos, "Collusion detection and prevention with FIRE+ trust and reputation model," in *Proceedings of the 10th IEEE International Conference on Computer and Information Technology (CIT '10)*, pp. 2548–2555, July 2010.

[21] F. G. Marmol, J. G. Marin-Blazquez, and G. M. Perez, "Linguistic fuzzy logic enhancement of a trust mechanism for distributed networks," in *Proceedings of the 10th IEEE International Conference on Computer and Information Technology*, pp. 838–845, Washington, DC, USA, July 2010.

[22] F. G. Mármol and G. M. Pérez, "TRIP, a trust and reputation infrastructure-based proposal for vehicular ad hoc networks," *Journal of Network and Computer Applications*, vol. 35, no. 3, pp. 934–941, 2012.

[23] F. G. Mármol and G. M. Pérez, "TRMSim-WSN, trust and reputation models simulator for wireless sensor networks," in *Proceedings of the IEEE International Conference on Communications (ICC '2009)*, pp. 1–5, Dresden, Germany, June 2009.

[24] V. K. Verma, S. Singh, and N. P. Pathak, "Analysis of scalability for AODV routing protocol in wireless sensor networks," *Optik—International Journal for Light and Electron Optics*, vol. 125, no. 2, pp. 748–750, 2014.

[25] A. S. Alkalbani, A. O. Md. Tap, and T. Mantoro, "Energy consumption evaluation in trust and reputation models for wireless sensor networks," in *Proceedings of the 5th International Conference on Information and Communication Technology for the Muslim World*, pp. 1–6, Rabat, Morocco, March 2013.

[26] S. Chen, Y. Zhang, and G. Yang, "Parameter-estimation based trust model for unstructured peer-to-peer networks," *IET Communications*, vol. 5, no. 7, pp. 922–928, 2011.

[27] L. Li and J. Y. Halpern, "Minimum-energy mobile wireless networks revisited," in *Proceedings of the IEEE International Conference on Communications (ICC '01)*, vol. 1, pp. 278–283, Helsinki, Finland, June 2001.

[28] J. A. Sánchez and P. M. Ruiz, "Improving delivery ratio and power efficiency in unicast geographic routing with a realistic physical layer for wireless sensor networks," in *Proceedings of the 9th EUROMICRO Conference on Digital System Design: Architectures, Methods and Tools (DSD '06)*, pp. 591–597, Dubrovnik, Croatia, September 2006.

[29] F. Li and Y. Wang, "Routing in vehicular ad hoc networks: a survey," *IEEE Vehicular Technology Magazine*, vol. 2, no. 2, pp. 12–22, 2007.

[30] F. G. Mármol and G. M. Pérez, "Providing trust in wireless sensor networks using a bio-inspired technique," in *Proceedings of the Networking and Electronic Commerce Research Conference (NAEC '08)*, Lake Garda, Italy, September 2008.

[31] Y. Pan, Y. Yu, and L. Yan, "An improved trust model based on interactive ant algorithms and its applications in wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 2013, Article ID 764064, 9 pages, 2013.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 501534, 8 pages
http://dx.doi.org/10.1155/2014/501534

Research Article

Energy-Efficient Reliable Broadcast Protocol for WSNs Based on IEEE 802.15.5

Juho Lee, Woongsoo Na, and Sungrae Cho

School of Computer Science and Engineering, Chung-Ang University, 221 Heukseok, Dongjak, Seoul 156-070, Republic of Korea

Correspondence should be addressed to Sungrae Cho; srcho@cau.ac.kr

Received 14 December 2013; Accepted 16 April 2014; Published 8 May 2014

Academic Editor: Jongsung Kim

We propose two link-layer reliable broadcast protocols for wireless sensor networks based on IEEE 802.15.5. We compare them in terms of energy consumption. By including both positive and negative acknowledgement, our second proposed scheme can effectively reduce the number of unnecessary error control messages and, thereby, significantly reducing the unnecessary power consumption relative to the first scheme. Also, we provide an analytical framework for the evaluation of different reliable broadcast techniques. Simulation results show that Scheme 2 achieves energy savings of up to about 85% compared to Scheme 1.

1. Introduction

Recently, there have been several proposals to provide reliable transmission in wireless sensor networks (WSNs) between the transport layer and the link layer [1–3]. To keep pace with the rapidly increasing use of WSNs, the IEEE 802 standards association has developed standards for WSNs based on the IEEE 802.15 working group.

IEEE 802.15 Task Group 5 has recently released the IEEE 802.15.5 standard [4]. This standard provides an architectural framework enabling wireless personal area network (WPAN) devices to promote interoperable and stable wireless mesh topologies [5].

The IEEE 802.15.5 standard consists of low-rate and high-rate parts. The low-rate part basically targets a variety of applications in WSNs. Although the applications enable systems based on the low-rate part of IEEE 802.15.5 to utilize a fully distributed MAC without any central coordinator, logical groups are formed around each device to facilitate contention-free exchanges while exploring medium reuse over different spatial regions. The membership of devices to these groups can vary over time due to changes in location or topology.

The distributed MAC mechanism ensures high performance and efficient relaying of a MAC frame from a source to a destination in the network, possibly over several multihop relay devices that form an IEEE 802.15.5-based WSN. Although the applications enabled by WSNs are very attractive, there are many technical challenges to overcome in order to build well-functioning robust systems based on IEEE 802.15.5 technology; the identified challenges include scalability, reliability, and energy efficiency. Reliable broadcast is necessary for the IEEE 802.15.5, since many applications in the network depend on broadcasting, including service discovery, device paging, routing information propagation, and even data transfer. Full (i.e., 100%) reliability for those applications is not possible in the case where any node is not reachable via the mesh and its battery is not operable.

By relaxing the full reliability requirement, we can achieve an energy efficiency gain in WSNs based on IEEE 802.15.5. In this paper, we propose two link-layer reliable broadcast protocols for IEEE 802.15.5: Scheme 1 and Scheme 2 and compare them in terms of energy consumption. Scheme 2 reduces unnecessary error control messages and, thereby, significantly reducing unnecessary power consumption relative to Scheme 1.

The rest of this paper is organized as follows. Section 2 describes prior related work as well as the contributions of this paper. Section 3 describes the proposed reliable broadcasting algorithms. Section 4 presents mathematical analysis of the proposed schemes. Performance comparisons with another legacy technique are given in Section 5. Finally, we draw conclusions in Section 6.

2. Related Work and Contributions

Most of the earlier work in the area of reliable broadcast has focused on TDMA slot assignment [6–8]. These schemes require global time synchronization and a certain level of topology information, which are not easy to implement, especially in dynamically changing environments such as WSNs.

Recent works [9, 10] depended on dense deployment of devices to reduce duplication in the relay of broadcast frames. However, the authors did not consider wireless channel errors or lost frames, and the techniques are unlikely to work well in sparse topologies.

The pump slowly, fetch quickly (PSFQ) transport layer mechanism [11] was proposed for reliable retasking/reprogramming of sensors based on negative acknowledgment (NAK) from receivers. The event-to-sink reliable transport (ESRT) protocol was developed in [1]. ESRT is based on the notion of event-to-sink reliability and provides reliable event detection in WSNs without imposing any intermediate caching requirements and with minimum energy expenditure. Another NAK-based scheme for providing sink-to-sensors reliability in WSNs, called GARUDA, is introduced in [3]. GARUDA incorporates an efficient pulsing-based solution in which sensor nodes are informed of an impending reliable short-message delivery by transmitting a specific series of pulses at a certain amplitude and period. The pulses act as NAKs, if any receiver does not receive a broadcast frame from the transmitter.

The aforementioned techniques are all transport-layer reliable broadcast protocols. The biggest problem with end-to-end recovery has to cope with the link-layer errors which accumulate exponentially over multihop sensor nodes. Recovery mechanisms based on end-to-end reliability might waste considerable amounts of the sensor nodes' energy resources. It would be preferable to cure the errors before they propagate along the path, a solution that necessitates link-layer reliability.

There also exist link-layer approaches [12, 13] for reliable broadcast in wireless sensor networks. Forward error correction (FEC) has been an appealing approach to reduce the feedback implosion that usually occurs when a large-scale reliable broadcast is performed [13]. However, the use of FEC in WSNs requires considerable hardware cost and complexity.

Traditionally, reliable broadcast techniques have been based upon positive acknowledgment (ACK), NAK, or both. In the ACK-based approach, a slot-reservation-based reliable broadcast protocol (SRB) [14] was proposed to add a reliability component to the existing broadcast protocol in the IEEE 802.11 MAC. In the SRB, a transmitter needs an ACK from all receivers to guarantee full reliability. However, since ACKs from the receivers are typically synchronized, it will cause significant contention in the wireless channel. This problem is exacerbated as the number of receivers increases and is referred to as the ACK implosion problem.

In another ACK-based approach, Xie et al. [15] proposed the round-robin acknowledge and retransmit (RRAR) protocol to improve the reliability of broadcasting. In this protocol, after the broadcasting is finished, the sender requires a broadcast acknowledgement (BrACK) from one of its neighbors, and this BrACK scheme is performed in a round-robin fashion for all the neighbors of the sender. Thus, this protocol can reduce ACK implosion problem, but it cannot guarantee the reliability of all nodes as the number of receivers increases.

On the contrary, NAKs are well established as an effective mechanism to advertise losses in multihop wireless networks in particular and group communication in general, as long as the loss probabilities are not high. Cooperative loss recovery for reliable multicast (CoreRM) in ad hoc networks [16] is a NAK-based scheme in which the NAK frames are scheduled by random timers to avoid NAK implosion. Since one NAK is sufficient for the sender to be aware that an error has occurred, retransmission of the original frame informs the receivers with later NAK timers, thus allowing them to cancel their scheduled NAKs.

However, NAKs cannot handle the unique cases in which all frames are lost at a particular node in the network. Because such a node will be unaware that a data frame is expected, it will not advertise a NAK to request retransmission. For short message types, like queries consisting of a few frames, the probability is not negligible that a node will fail to receive any of the packets in a message. Because of the above problems, ACK or NAK has not been utilized for broadcasting services in IEEE 802 families such as IEEE 802.11 or IEEE 802.15.

To tackle the above problems, a hybrid scheme that uses both ACK and NAK has been proposed [17]. In this scheme, a transmitter elects a broadcast group leader. To cope with the ACK implosion problem, nonleader receivers use the NAK-based scheme while the leader uses an ACK-based scheme. However, this scheme still does not work in all cases; it will fail, if the leader receives a frame successfully, while all of the other receivers do not.

In this paper, we propose two reliable broadcast protocols for WSNs based on IEEE 802.15.5: Scheme 1 and Scheme 2. Within the aforementioned taxonomy, Scheme 1 is an ACK-based reliable broadcasting scheme, while Scheme 2 is a hybrid scheme. Compared with Scheme 1, Scheme 2 effectively reduces unnecessary error control messages, avoids unwanted collisions, and thus significantly conserves energy in the network. Simulation results show that Scheme 2 achieves energy savings of up to about 85% compared to Scheme 1 and a legacy technique.

3. Reliable Broadcasting Algorithm

In this paper, we assume that the wireless sensor networks form tree network environments (as in Figure 1) based on the IEEE 802.15.5 mesh formation [4]. In such networks, broadcasting is performed over the meshed tree network. For instance, if node A in the example of Figure 1 has broadcast data to send (A is the transmitter or originator), it transmits that data to its children nodes (B, C, and D). After the children nodes receive the data, they transmit the broadcast data to their own children nodes. Thus, in Figure 1, node C transmits data to nodes E and F. Additionally, the nodes transmit the broadcast data to their associated nodes; that

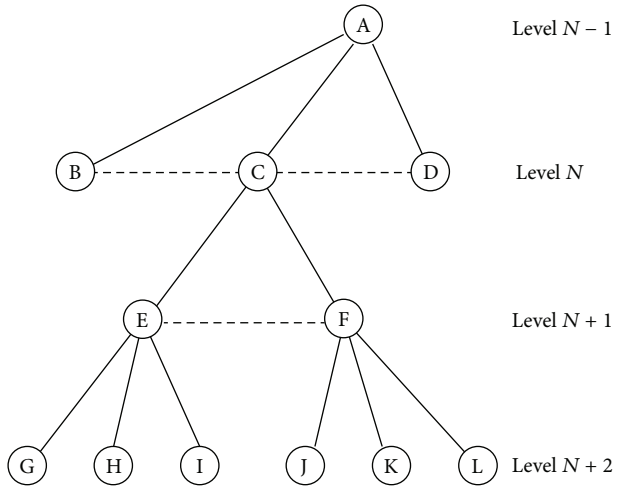


Figure 1: Example meshed tree network (Level = 4) following the IEEE 802.15.5 framework; solid and dotted lines represent the relationships of parents to children and of associated neighbors, respectively.

is, in Figure 1, node C transmits data to nodes B and D (see details on the formation of the meshed tree in [4]).

We first consider simple flooding schemes for IEEE 802.15.5 [18]. Simple flooding starts with a source node broadcasting a frame to all its neighbors. Each of those neighbors in turn forwards the frame to all its neighbors exactly once, and this continues until all reachable network nodes receive the frame. IETF (Internet engineering task force) has proposed the use of this flooding scheme for broadcasting and multicasting in ad hoc networks, which are often characterized by low node densities and/or high mobility. However, it does not guarantee that all nodes receive the broadcast frame and can even cause a broadcast storm problem [9].

Consider a scheme in which, in an attempt to guarantee the reliability of all nodes, all receivers acknowledge the receipt of a frame from the transmitter. We refer to this technique as Scheme 1. Scheme 1 will cause severe network degradation due to serious redundancy, contention, and collisions in the network. The likely cause of this degradation is that the acknowledgments from the receivers are typically synchronized, which will lead to considerable contention in the wireless channel (i.e., an ACK implosion problem). The multiple receiver nodes transmit multiple ACK frames after receiving the data frame from the sender. Typically, the sequence of the ACK transmission is scheduled in a manner that avoids collision among the ACK frames. However, if the sequence of the ACK transmission is not scheduled, ACK frames will collide, wasting additional retransmission time. This problem is exacerbated as the number of receivers increases; we will demonstrate this problem in Section 3.

For the IEEE 802.15.5 wireless sensor network, we propose a new reliable broadcast protocol, referred to as Scheme 2, which we have developed to meet two objectives: reliability and *lower power consumption*. Scheme 2 does not require all receivers to acknowledge a received frame, instead soliciting

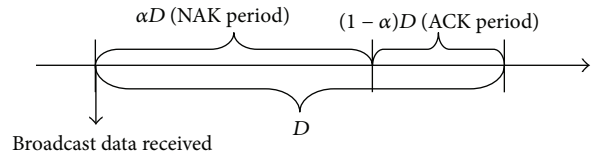


Figure 2: The timer used in the proposed scheme.

only a few of the receivers to acknowledge receipt. The basic idea is to delay each receiver's feedback (ACK or NAK) by a random time, thereby, desynchronizing the feedback to reduce contention. Also, if an earlier feedback transmission is overheard by the other receivers, they can suppress their later feedback when they determine that their feedback is redundant. This will significantly reduce collisions and unnecessary feedback, conserving energy overall.

Separate transmission of the ACK and data frames is considered to be redundant. Hence, in Scheme 2, the receiver that needs to send an ACK simply broadcasts its received data frame without explicitly sending its ACK. Then, when the transmitter receives that data frame, it treats it as an implicit ACK. This results in further energy savings in the network.

Scheme 2 also exploits a random timer set in the range of $[0, D]$ at each node, when it needs to send its feedback. Since feedback can be lost due to errors or collisions, the transmitter also employs a timer D. In the proposed scheme, there are 2 types of timers: a NAK timer and an ACK timer. The NAK timer is set at random in the range $[0, \alpha D]$, while the ACK timer is set at random in the range $[\alpha D, D]$ (see Figure 2). The shorter timer for NAK allows early rebroadcast of the original data to fix errors. The longer timer for ACK is for the case in which all nodes successfully received the data.

3.1. Transmitter Behavior. In the proposed scheme, the transmitter performs the following procedure repeatedly and automatically. When it has received broadcast data, the node checks whether the broadcast data frame was received from its child or its parent. If the data is from its child (see line 5 in Algorithm 1), the node broadcasts the data (piggybacked onto the ACK/NAK) to its children and its parent (not to its siblings) and sets its timer D. On the other hand, if the data is from its parent (see line 9 in Algorithm 1), it broadcasts its data to its children (not to its siblings) and sets its timer D. If the node is the originator of the broadcast data, it also broadcasts the data to its children and its parent (not to its siblings) and sets its timer D. In all cases, a node that has received broadcast data from another node does not unicast back to that other node. If timer D expires before receiving any ACK or NAK feedback (see line 15 in Algorithm 1), it retransmits the original data.

3.2. Receiver Behavior. After receiving broadcast data, the receiver checks whether the broadcast data is erroneous. Based on the result of error detection/correction (see line 5 in Algorithm 2), the receiver uses a NAK/ACK timer and sends the feedback frame to transmitter. When a node receives rebroadcast data (see line 10 in Algorithm 2), the receiver

```
(1)  𝒞 ← the set of children nodes
(2)  𝒫 ← the parent node
(3)  loop
(4)    if Has a data frame then {Transmission case}
(5)      if Data is from child then {Upstream case}
(6)        Piggyback the broadcast data onto the ACK/NAK;
(7)        Broadcast the data to 𝒞, 𝒫;
(8)        Set timer D;
(9)      else if Data is from parent then {Downstream Case}
(10)       Broadcast the data to 𝒞;
(11)       Set timer D;
(12)     else
(13)       Do nothing;
(14)     end if
(15)   else if The timer has expired then {Retransmission case}
(16)     Go to Step 5;
(17)   else if A feedback frame has been received then {Feedback case}
(18)     if An ACK frame has been received then {ACK case}
(19)       Transmit the next broadcast data;
(20)       Set timer D;
(21)     else if A NAK frame has been received then {NAK case}
(22)       Retransmit data;
(23)       Set timer D;
(24)     else
(25)       Do nothing;
(26)     end if
(27)   else
(28)     Do nothing;
(29)   end if
(30) end loop
```

Algorithm 1: Transmitter behavior of Scheme 2.

```
(1)  ℱ_𝒟 ← the received broadcast data
(2)  loop
(3)    if Broadcast data has been received then
(4)      if (ℱ_𝒟 == broadcast data) then
(5)        if (ℱ_𝒟 == erroneous data) then
(6)          Set the NAK timer in the range [0, αD];
(7)        else
(8)          Set the ACK timer in the range [αD, D];
(9)        end if
(10)     else if (ℱ_𝒟 == rebroadcast data) then
(11)       Cancel timer;
(12)       if ℱ_𝒟 already exists then
(13)         Do nothing;
(14)       else
(15)         Go to Step 5;
(16)       end if
(17)     else
(18)       Do nothing;
(19)     end if
(20)   else if Timer has expired then
(21)     Respond with feedback (NAK or ACK) to the transmitter
(22)   else
(23)     Do nothing;
(24)   end if
(25) end loop
```

Algorithm 2: Receiver behavior of Scheme 2.

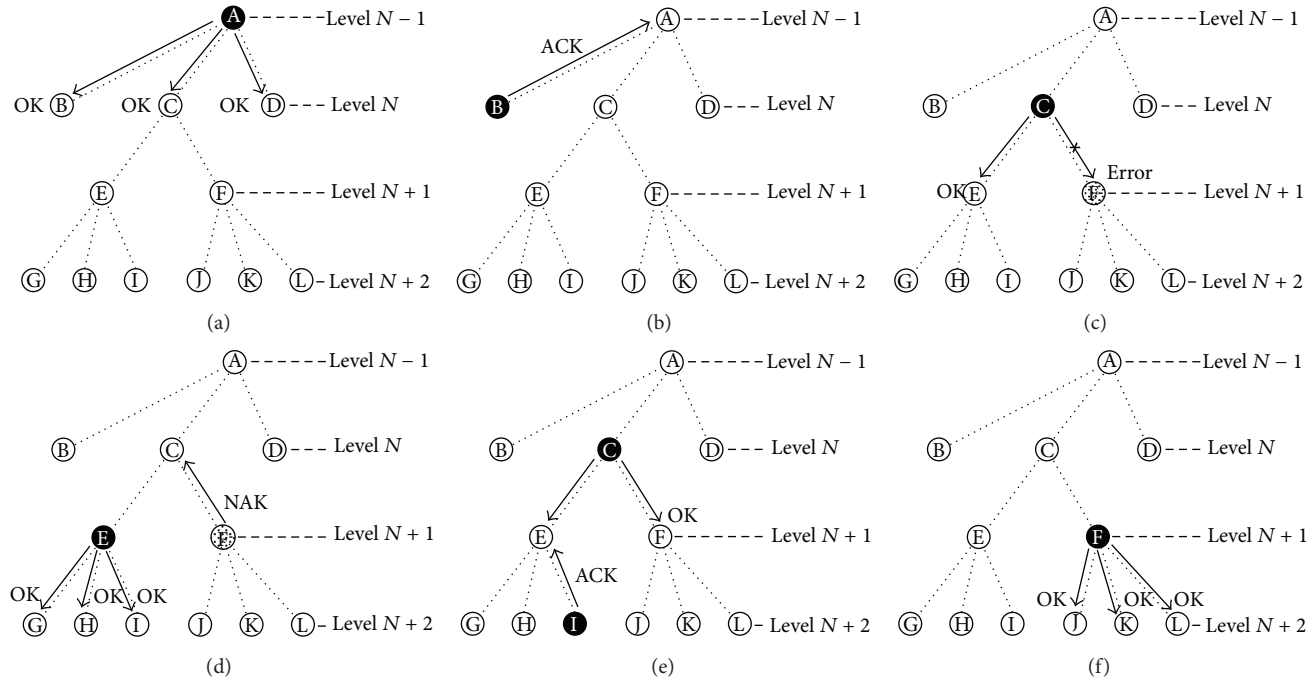


Figure 3: Example of the proposed scheme (downstream case at node C).

node cancels its timer (either NAK or ACK). This can reduce unnecessary feedback to the transmitter.

The following sequence will continue up to a predefined number of times. First, the transmitter broadcasts its data. Upon receiving it, the first and following NAKs are transmitted back to the transmitter from the receivers. Due to the first NAK, the transmitter rebroadcasts the data. Then, each other node suppresses its timer (either NAK or ACK), if the timer is in a waiting state. Suppose that an error is detected at one receiver. Then, the receiver sends its NAK. Upon receipt of the NAK, the transmitter rebroadcasts its data. Finally, all receivers have the error-free data. While one of the receivers sends its ACK, the transmitter can initiate the next broadcast data transmission.

3.3. Example of Scheme 2. Figure 3 shows an example scenario for the proposed scheme (downstream case at node C). As shown in the figure, there are 12 nodes in the network; let node A be the originator. At $T = t_0$, parent node A broadcasts a data frame (Figure 3(a)). Then, node B replies with ACK at $T = t_0$ (Figure 3(b)) and node C may forward the broadcast data to its children (Figure 3(c)). However, if node F receives the data with errors, it replies with NAK to transmitter node C at $T = t_0 + 0.8D$ (Figure 3(d)). At about the same time, node E forwards the data to its children (Figure 3(d)). At $T = t_0 + 0.9D$, node C rebroadcasts the data to its children, and node E suppresses its ACK for the previous broadcast data and also silently ignores the rebroadcast data (Figure 3(e)). After $T = t_0 + 0.9D$, node F receives the rebroadcast data from node C and then forwards the data to its children (Figure 3(f)). Then node F receives ACK replies from its children.

Another example scenario for the proposed scheme is shown in Figure 4; here, node E is the originator node and a child of node C. At $T = t_0$, it broadcasts to its children nodes G, H, and I and its parent node C (Figure 4(a)). Then, node C broadcasts data piggybacked onto the ACK to transmit the data to node A and F at $T = t_0 + 0.6D$ (Figure 4(b)). However, if node A receives the data with errors, it immediately replies with NAK to node C (Figure 4(c)). After correctly receiving broadcast data from node C, node A broadcasts the data, piggybacked onto the ACK, to nodes B and D (Figure 4(e)); also, node F broadcasts the data to its children nodes (Figure 4(f)). Finally, node B replies with ACK and the broadcast transmission is complete.

4. Analysis of the Proposed Scheme

We now provide a mathematical model measuring the energy consumption in the cases of a scheme based on ACK only, one based on NAK only, and the proposed scheme, which uses both ACK and NAK. To measure energy consumption, we calculate the minimum number of transmissions that will guarantee full reliability. Let $E[N]$ be the expected number of nodes that successfully receive a broadcast frame, where N is the number of nodes. For an ACK-based scheme, $E[N]$ is calculated as

$$E[N] = \sigma_{\mathrm{ACK}}\left(1 - \sigma_{\mathrm{ACK}}\right)^{(N-qN-1)}(\gamma), \tag{1}$$

where σ denotes the probability that a node transmits in one time slot, q is the frame error rate, and γ denotes the number of time slots. For a NAK-based scheme, $E[N]$ is calculated as

$$E[N] = \sigma_{\mathrm{NAK}}\left(1 - \sigma_{\mathrm{NAK}}\right)^{(qN-1)}(\gamma)(N - qN). \tag{2}$$

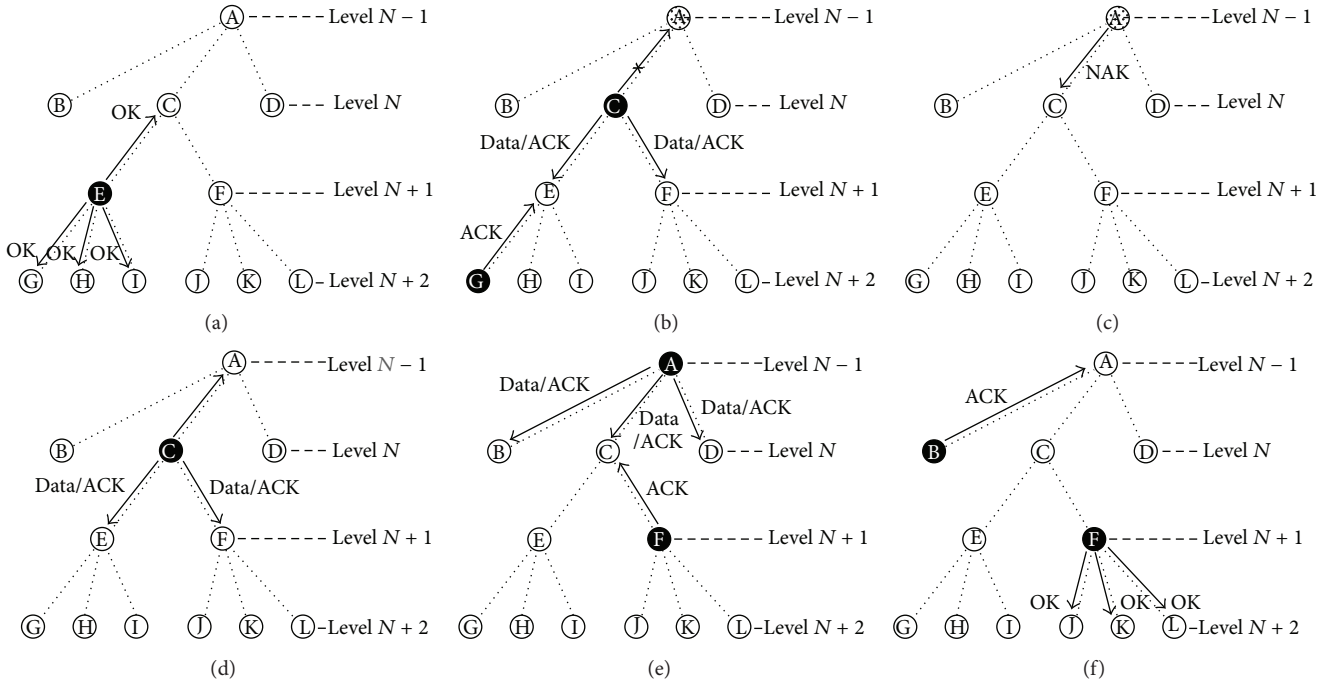


Figure 4: The example of the proposed scheme (upstream case at node E).

Lastly, in the proposed scheme, $E[N]$ is calculated as

$$E[N] = \sigma_{\mathrm{ACK}}(1-\sigma_{\mathrm{ACK}})^{(N-qN-1)}(\gamma)(N-qN) + \sigma_{\mathrm{NAK}}(1-\sigma_{\mathrm{NAK}})^{(qN-1)}(\gamma)(1-qN). \quad (3)$$

The minimum number of the transmissions ($\mathfrak{R}$) can be evaluated as follows:

$$\mathfrak{R} = \arg\min\left(n, \sum_{n=1}^{\infty} E^n[N] > N\right). \quad (4)$$

In an ACK-based scheme and when $q = 0.01$, $\mathfrak{R}$ increases as the number of receiver nodes increases (Figure 5); this is because the probability of collisions increases with the number of nodes. However, $\mathfrak{R}$ is nonincreasing with the number of nodes in the NAK-based and proposed schemes. In the low-error-rate regime, fairly few receivers will fail to receive the broadcast data frame. Therefore, the NAK frame is likely to be successfully delivered in the NAK-based scheme and the proposed scheme. However, if the NAK frame is lost, the sender does not rebroadcast. For this reason, the NAK-based and proposed schemes do not provide the full reliability but consume less energy.

In the proposed scheme, increasing the channel error rate with $N = 10$ has no effect on $\mathfrak{R}$, whereas $\mathfrak{R}$ increases with channel error rate in the schemes based on ACK only and on NAK only (Figure 6); that is to say, our proposed scheme is insensitive to error rate. Especially, $\mathfrak{R}$ is smallest in the proposed scheme among the three schemes tested when the error rate is 0.3 or greater.

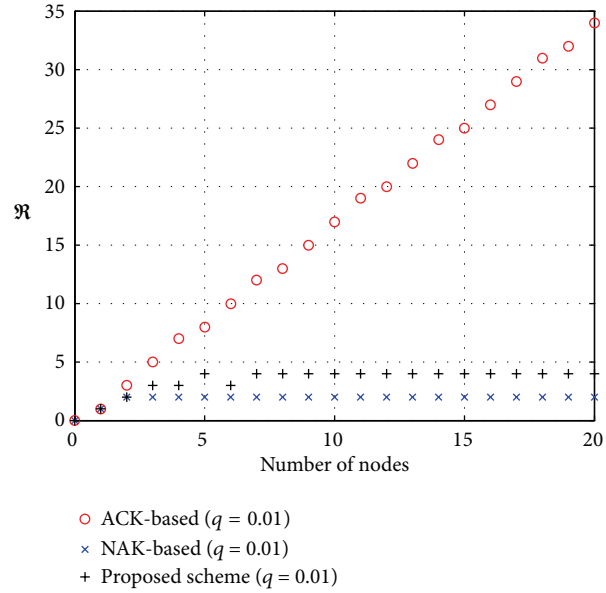


Figure 5: $\mathfrak{R}$ versus number of nodes ($q = 0.01$).

5. Performance Evaluation

To evaluate the efficiency of Scheme 2, we developed a network simulator based on ns-2 [19]. The ns-2 simulator incorporates the IEEE 802.15.5 specification. The IEEE 802.15.5 simulator module includes a unicast routing algorithm, an association procedure, and a disassociation procedure. Additionally, the simulator provides an option of simple flooding, Scheme 1 or Scheme 2. The distance between each pair of

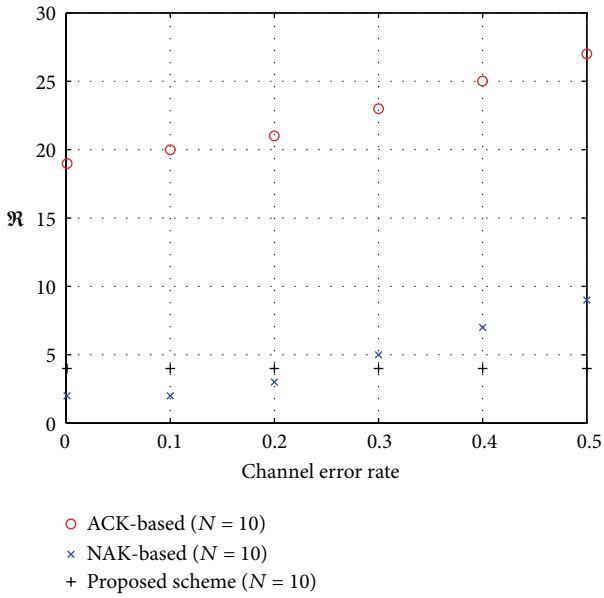


Figure 6: $\mathfrak{R}$ versus channel error rate ($N = 10$).

Table 1: Simulation Parameters.

Parameters	Value
Number of nodes	Variable
Neighbor distance	7–11 m
Tx range	12 m
PAN coordinator (PC)	Bottom node (or any designated node)
Network startup	PC start at 0.0 Any other nodes: random time from 1.0 to 3.0
Packet error rate	10%

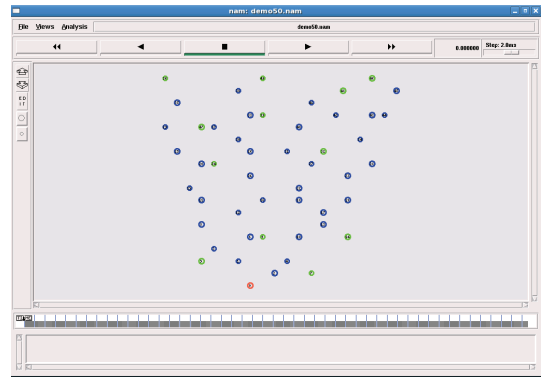


Figure 7: NS-2 WSN simulator.

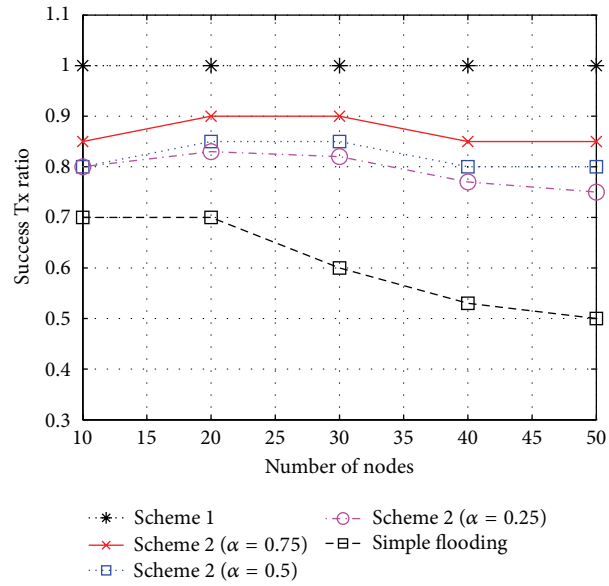


Figure 8: The proportion of nodes that successfully receive the data versus the number of nodes (channel error rate = 10%).

neighbors is less than 11 m. The transmission range is 12 m (see Table 1), and the simulator generates a random topology (e.g., see Figure 7). In this section, we compare Scheme 2 with simple flooding and Scheme 1 through simulations, measuring the proportion of nodes that successfully receive data, as well as $\mathfrak{R}$, in various conditions. Because it is proportional to energy consumption, $\mathfrak{R}$ can be used as an energy budget.

Figure 8 shows the proportion of nodes that successfully received the data to the total number of nodes in each of the transmission schemes: Scheme 1, Scheme 2 ($\alpha = 0.75$, 0.50, 0.25), and simple flooding. As can be seen from the figure, simple flooding does not guarantee 100% reliability. Simple flooding performs especially poorly as the number of nodes increases to near 50, making this scheme unsuitable for applications that require a decent degree of reliability. Hence, simple flooding cannot be used for many applications in IEEE 802.15.5 WSNs; so we exclude it from further performance comparisons. Scheme 2 is more reliable than simple flooding, but also does not provide full reliability because of NAK implosion. As the alpha variable which is related to NAK transmission period is increased, probability of NAK collisions can be decreased significantly. For that reason, Scheme 2, which has a large alpha variable, is more reliable than any other Scheme 2. Among the schemes tested, only Scheme 1 provides full reliability.

Now, to measure the energy conservation that can be realized by using Scheme 2, we compare the number of broadcast data frames and control frames that are generated per broadcast data frame, including rebroadcasting and NAK frames (Figure 9). This number will be directly proportional to the energy consumption, so we use this measure as an energy budget. As observed in the figure, the energy consumption of Scheme 2 is much less than that of Scheme 1. Especially, Scheme 2, which has a small alpha variable, can conserve more energy. This means that the smaller the alpha variable, the lower probability of ACK collisions, and it can reduce the unnecessary retransmissions. Therefore, we can choose either of them (Scheme 1 or Scheme 2) depending on the desired level of reliability or energy efficiency, using Scheme 1 to ensure full reliability and Scheme 2 for low-energy operation.

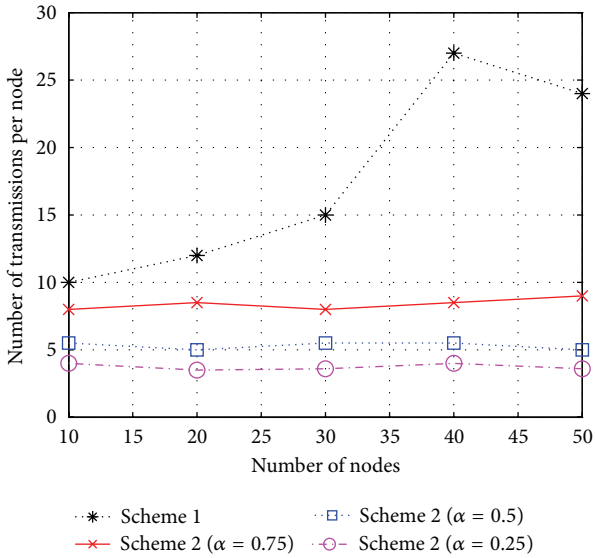


FIGURE 9: Number of transmissions per node versus the number of nodes (channel error rate = 10%).

6. Conclusion

In this paper, we propose two reliable broadcast protocols for wireless sensor networks based on IEEE 802.15.5: Scheme 1 and Scheme 2. Scheme 1 is an ACK-based reliable broadcasting scheme, while Scheme 2 is a hybrid scheme. Compared to Scheme 1, Scheme 2 effectively reduces unnecessary error control messages and avoids unwanted collisions, thereby realizing considerable energy savings in the network overall. Simulation results show that Scheme 2 achieves energy savings of up to about 85% compared to Scheme 1 and the other legacy technique.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by the Chung-Ang University Excellent Student Scholarship and by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2013R1A1A2012443).

References

[1] Ö. B. Akan and I. F. Akyildiz, "Event-to-sink reliable transport in wireless sensor networks," *IEEE/ACM Transactions on Networking*, vol. 13, no. 5, pp. 1003–1016, 2005.

[2] M. J. Lee, R. Zhang, J. Zheng et al., "IEEE 802.15.5 WPAN mesh standard-low rate part: meshing the wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 7, pp. 973–983, 2010.

[3] S.-J. Park, R. Vedantham, R. Sivakumar, and I. F. Akyildiz, "GARUDA: achieving effective reliability for downstream communication in wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 2, pp. 214–230, 2008.

[4] IEEE 802.15.5, "Mesh topology capability in wireless personal area networks (WPANs)," IEEE Std. 802.15.5, 2009.

[5] W. Na, G. Lee, H. Bae, J. Yu, and S. Cho, "Reliable broadcast scheme for IEEE 802.15.5 low-rate WPAN mesh networks," *IEICE Transactions on Communications*, vol. E95-B, no. 9, pp. 2700–2707, 2012.

[6] R. Bar-Yehuda, A. Israeli, and A. Itai, "Multiple communication in multihop radio networks," *SIAM Journal on Computing*, vol. 22, no. 4, pp. 875–887, 1993.

[7] I. Chlamtac and O. Weinstein, "The wave expansion approach to broadcasting in multihop radio networks," *IEEE Transactions on Communications*, vol. 39, no. 3, pp. 426–433, 1991.

[8] C. Lee, J. E. Burns, and M. H. Ammar, "Improved randomized broadcast protocol in multi-hop radio networks," Tech. Rep. GIT-CC-93-14, College of Computing, Georgia Institute of Technology, 1993.

[9] S.-Y. Ni, Y.-C. Tseng, Y.-S. Chen, and J.-P. Sheu, "The broadcast storm problem in a mobile ad hoc network," in *Proceedings of the ACM/IEEE MOBICOM*, Seattle, Wash, USA, 1999.

[10] B. Williams and T. Camp, "Comparison of broadcasting techniques for mobile ad hoc networks," in *Proceedings of the 3rd ACM International Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC '02)*, pp. 194–205, June 2002.

[11] C.-Y. Wan, A. T. Campbell, and L. Krishnamurthy, "Pump-Slowly, Fetch-Quickly (PSFQ): a reliable transport protocol for sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 4, pp. 862–872, 2005.

[12] K. Tang and M. Gerla, "MAC reliable broadcast in ad hoc networks," in *Proceedings of the Communications for Network-Centric: Creating the Information Force (Milcom '01)*, pp. 1008–1013, October 2001.

[13] M. C. Vuran and I. F. Akyildiz, "Error control in wireless sensor networks: a cross layer analysis," *IEEE/ACM Transactions on Networking*, vol. 17, no. 4, pp. 1186–1199, 2009.

[14] V. Srinivas and L. Ruan, "An efficient reliable multicast protocol for 802.11-based wireless LANs," in *Proceedings of the IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks and Workshops (WOWMOM '09)*, June 2009.

[15] J. Xie, A. Das, S. Nandi, and A. K. Gupta, "Improving the reliability of IEEE 802.11 broadcast scheme for multicasting in mobile ad hoc networks," *IEE Proceedings: Communications*, vol. 153, no. 2, pp. 207–212, 2006.

[16] M. Huang, G. Feng, and Y. Zhang, "Cooperative loss recovery for reliable multicast in ad hoc networks," *Internatioanl Journal of Communications, Network, and System Science*, pp. 72–78, 2010.

[17] W.-S. Lim, D.-W. Kim, and Y.-J. Suh, "Design of efficient multicast protocol for IEEE 802.11n WLANs and cross-layer optimization for scalable video streaming," *IEEE Transactions on Mobile Computing*, vol. 11, no. 5, pp. 780–792, 2012.

[18] J. Jetcheva, Y. Hu, D. Maltz, and D. Johnson, "A simple flooding protocol for multicast and broadcast in mobile ad hoc networks," Internet Draft: draft-ietf-manet-simplembcast-01.txt, July 2001.

[19] IEEE 802.15.4 Ns2 Package, http://ees2cy.engr.ccny.cuny.edu/zheng/pub/.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 291295, 4 pages
http://dx.doi.org/10.1155/2014/291295

Research Article

Implementation of Personal Health Device Communication Protocol Applying ISO/IEEE 11073-20601

Deok Seok Seo,[1] Soon Seok Kim,[2] Yong Hee Lee,[2] and Jong Mo Kim[2]

[1] *School of Architecture, Halla University San 66, Heungup-Li, Heungup-myon, Wonju-shi, Gangwon-do 220-712, Republic of Korea*
[2] *Department of Computer Engineering, Halla University, San 66, Heungup-Li, Heungup-myon, Wonju-shi, Gangwon-do 220-712, Republic of Korea*

Correspondence should be addressed to Soon Seok Kim; sskim@halla.ac.kr

Received 9 December 2013; Accepted 16 April 2014; Published 6 May 2014

Academic Editor: Jongsung Kim

In 2010, IEEE and ISO announced the exchange protocol standard (ISO/IEEE 11073-20601), optimized to secure mutual compatibility between all sorts of PHDs and the gateways for collecting bioinformation from the devices and activating related services. This international standard is the first official document that has dealt with communication related to the healthcare device. This paper is about implementing communication protocols between a weight sensor, a kind of personal health devices (PHDs) used in homes, and a gateway collecting a variety of biometric information from multiple sensors, applying international standard ISO/IEEE 11073-20601 for interoperability between medical devices. Moreover, security is enhanced by applying the international symmetric key encryption standard, advanced encryption standard (AES), for secure data transmission from the weight sensor to gateway. When the cipher algorithm was applied, we confirmed that the implementation took about 0.008 second on average than the previous.

1. Introduction

As the core of care has recently moved from treatment through medical practice to prevention or healthcare, owing to the influence of the concept of wellbeing and wellness, devices for letting users undergo an examination and diagnosis at home, such as a hemadynamometer, a blood sugar device, and scales, have been continuously released. Frequently, these are called personal health devices (PHDs), in comparison with point of care (POC) devices, which refer to medical equipment at the point of care in hospitals.

In 2010, IEEE and ISO announced the exchange protocol standard (ISO/IEEE 11073-20601) [1], optimized to secure mutual compatibility between all sorts of PHDs and the gateways for collecting bioinformation from the devices and activating related services. This international standard is the first official document that has dealt with communication related to the healthcare device.

Most of the communication modules for external interface are currently developed by companies based on need, so it is difficult to secure compatibility with other companies' devices, and because they are developed without a standard system, problems arise when linking them to a hospital's information system. As communication between medical devices that support the present network is becoming important, a standardized medical information protocol for sharing and transmitting information is required.

This study uses ISO/IEEE 11073-20601 to realize a communication protocol between the weight sensor and gateway and its purpose is to implement a standard technology for mutual interoperability between medical devices and hospital systems. Moreover, advanced encryption standard (AES), which is an international standard for symmetric key encryption, has been applied to enhance security. As a result, when a cipher algorithm is applied in field of data transmission from PHD to gateway, it takes approximately 0.078 seconds longer on average compared to before.

In Section 2 of this dissertation, the recently revised ISO/IEEE 11073-20601 standard protocol is examined and Section 3 describes the actual implementation method related

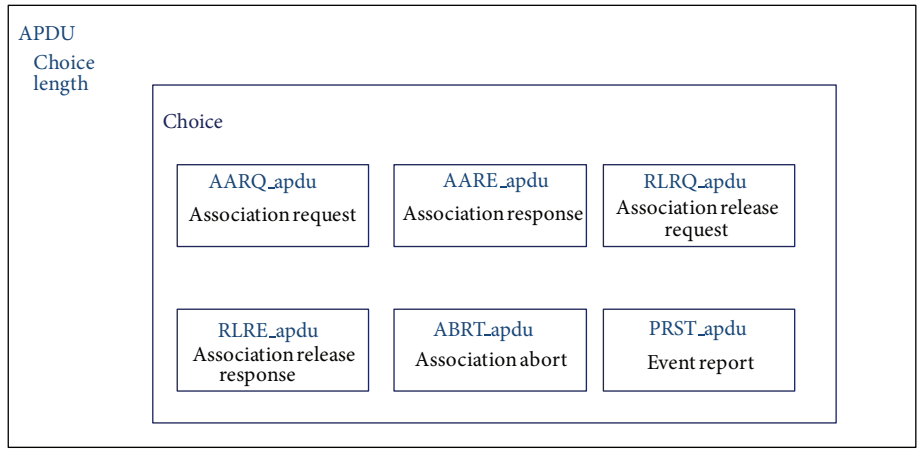


Figure 1: Type of APDU (Application Protocol Data Unit) in ISO/IEEE 11073-20601.

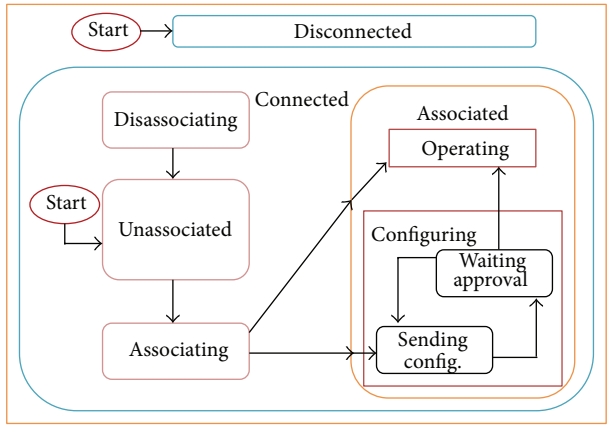


Figure 2: Communication state diagram of weight sensor in ISO/IEEE 11073-20601.

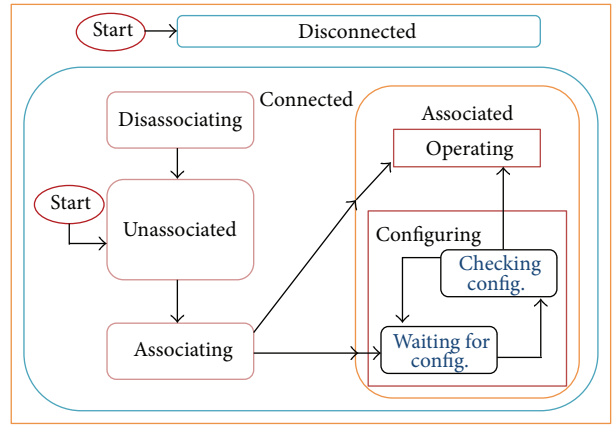


Figure 3: Communication state diagram of gateway in ISO/IEEE 11073-20601.

to the weight sensor. Section 4 proposes the results of implementation and Section 5 draws conclusions.

2. ISO/IEEE 11073-20601 Standard [1]

This standard is a document for defining the standard format for information sent between health devices and data managers, for collecting bioinformation measured by the devices, and for the mutual exchange of information.

A sensor (such as a hemadynamometer, comprising scales, and a blood sugar device, hereafter simply referred to as PHD) collects personal bioinformation and then transmits the information to a gateway (such as a cell phone, a health device, and a personal computer) for the purpose of collection, display, and further transmission. A gateway can transmit data for the purpose of additional analysis to an healthcare service center for teleassistance and utilize information from various domains such as disease control, health and fitness, or an independent age measuring device. The communication path between a PHD and the gateway is assumed to be a logical point-to-point connection. Generally, a PHD communicates with a single gateway at a specific point when necessary. Gateways can communicate with a plurality of PHDs simultaneously using separate point-to-point connections.

Refer to the document for standard [1] for other protocols in further detail.

3. Protocol Implementation Method

In this study, as mentioned previously, a weight sensor was used to apply the ISO/IEEE 11073 standard protocol. ISO/IEEE 11073-20601 (communication protocol standard between PHD and gateway) and ISO/IEEE 11073-10415 (weight sensor communication standard) were used for the application.

The ASN.1 encoding regulation (also known as a medical device encoding rule (MDER)) defined in the standard was used for the exchange of information between the weight sensor and gateway.

According to the definition by the International Telecommunication Union (ITU), ASN.1 is a protocol defining the

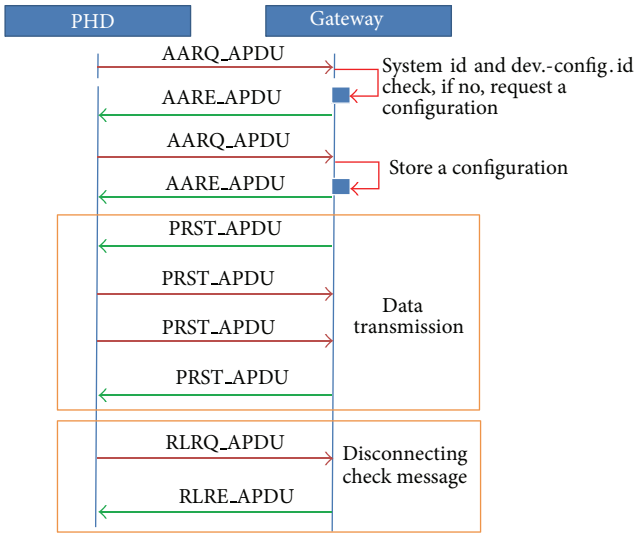


Figure 4: Unknown configuration communication procedure between PHD and gateway in ISO/IEEE 11073-20601.

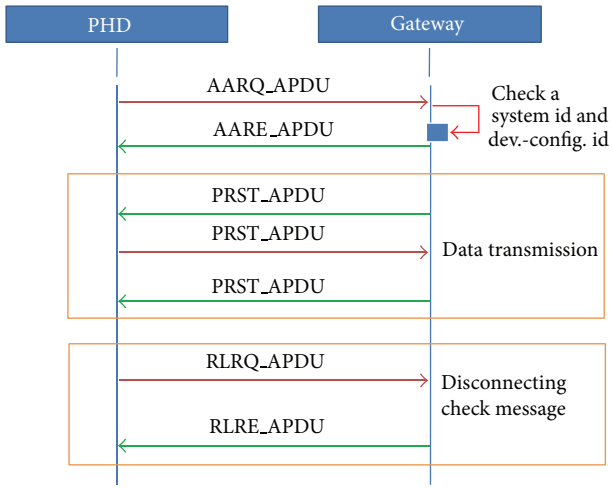


Figure 5: Known configuration communication procedure between PHD and gateway in ISO/IEEE 11073-20601.

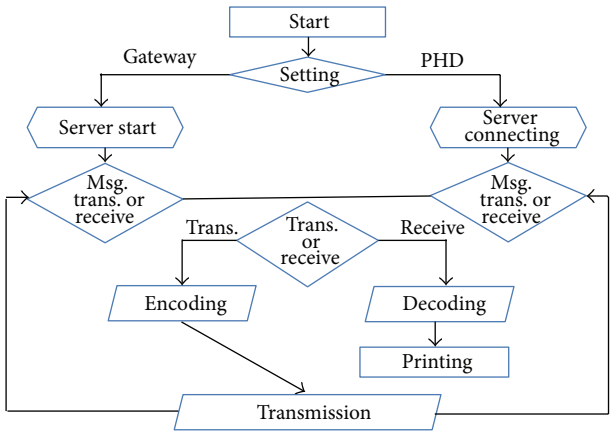


Figure 6: Viewer flow diagram for implementation of communication protocol applying ISO/IEEE 11073-20601.

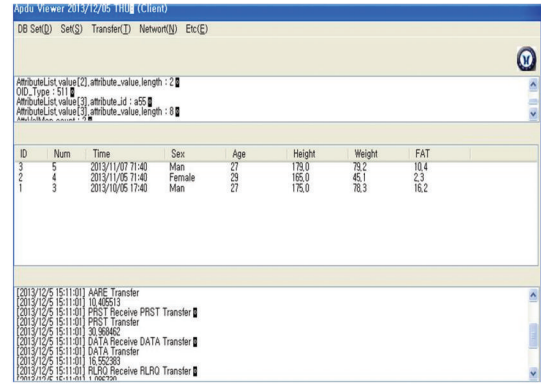


Figure 7: Viewer screen of the weight sensor implementing ISO/IEEE 11073-20601 communication protocol.

data exchange on the network and is a formal language used to exchange abstract messages between different models. It is simply a language that defines the standard and the data created with ASN.1 becomes the standard. If MDER is expressed in C language, it is declared as a strict type that sends basic data using a structure called APDU. In APDU, there are six message formats: AARQ_apdu, AARE_apdu, RLRQ_apdu, RLRE_apdu, ABRE_apdu, and PRST_apdu. According to the circumstances, communication takes place in 1 of the 6 messages (refer to Figure 1).

In order for communication between the weight sensor and the gateway to take place, the two devices must be mutually connected and in each status it can be divided into 10 types as shown below (refer to Figures 2 and 3).

From the weight sensor perspective, first, one's configuration information is sent and the gateway receives this information. The configuration information of the first connection is then saved and if connection is attempted again, only its ID is verified to enable immediate communication. On the other hand, Figure 4 demonstrates the communication process for the initial connection or if there is no configuration information of the weight sensor. Figure 5 demonstrates the communication process in cases where the gateway has configuration information from the weight sensor.

4. Protocol Implementation Result

In this study, the method described in Section 3 is used to apply a mutual communication protocol between the weight sensor and the gateway. The weight sensor (InBody R20 model) of the Biospace company [2] which was being sold on the market was used and the measurements from the weight sensor were received by a Pentium PC 3.0 GHz laptop for sending to the gateway (laptop Pentium PC 3.0 GHz). In order to secure accurate transmission, Visual C++ language was used in the laptop PC to create a viewer interface where features such as saving the data of the weight sensor, sending saved data, and displaying the received data were applied. Figure 6 shows the flow of the viewer program created and Figure 7 shows the weight sensor, while Figure 8 shows the viewer screen of the gateway.

Table 1: Comparison of average transfer time from PHD to gateway (unit: seconds).

PRST_APDU	Encrypted data	Nonencrypted data	Difference
Average transmission time from PHD to gateway in 10,000 attempts	0.026	0.018	**+0.008**

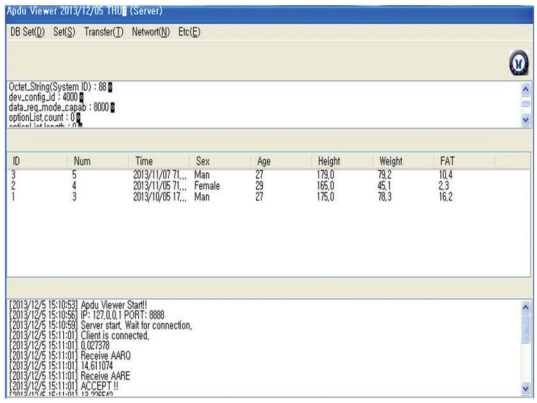


Figure 8: Viewer screen of the gateway implementing ISO/IEEE 11073-20601 communication protocol.

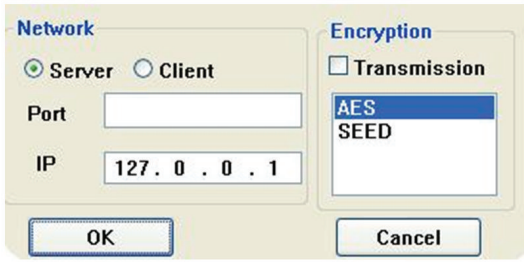


Figure 9: Cryptography algorithm setting in viewer screen implementing ISO/IEEE 11073-20601 communication protocol.

For a secure transmission of PRST_APDU data from PHD to gateway, advance encryption standard (AES) protocol [3], an international standard for symmetric key encryption, was applied and used [4–6] (refer to Figure 9). As a result of this application, when the cipher algorithm was applied, in 10,000 attempts, the average implementation time was approximately 0.078 seconds longer than when it was not applied (refer to Table 1).

It was determined that applying encryption for secured transmission did not significantly influence the entire implementation time.

5. Conclusion

Until now, we have used international standard ISO/IEEE 11073-20601 to apply the communication protocol between a weight sensor and a gateway. The purpose of this dissertation is to realize a standard technology for mutual interoperability between a PHD, a health device used in households, and the hospital systems. The AES protocol, an international standard of symmetric key encryption, was applied to strengthen the security of transmission between devices, which was not available previously. As a result of this realization, when the encryption algorithm was applied, it took approximately 0.078 seconds longer on average than without.

In the future, we intend to expand the range of PHDs for application not only to the proposed weight sensor but also to ECG sensors, blood pressure devices, blood glucose device, and others.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

(1) This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2013R1A1A2006745). (2) This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea (NRF) through the Human Resource Training Project for Regional Innovation.

References

[1] ISO/IEEE, "11073-20601: health informatics-personal health device communication, application profile optimized exchange protocol," http://www.iso.org.

[2] BIOSPACE CO., http://www.e-inbody.com/.

[3] AES (Advanced Encryption Standard), http://en.wikipedia.org/wiki/Advanced_Encryption_Standard.

[4] K. Jeong and S. S. Yeo, "Security analysis of block cipher LED," *Journal of Internet Technology*, vol. 14, no. 2, pp. 281–287, 2013.

[5] I. Lee, S. Jeong, S. Yeo, and J. Moon, "A novel method for SQL injection attack detection based on removing SQL query attribute values," *Mathematical and Computer Modelling*, vol. 55, no. 1-2, pp. 58–68, 2012.

[6] B. Vaidya, J. H. Park, S.-S. Yeo, and J. J. P. C. Rodrigues, "Robust one-time password authentication scheme using smart card for home network environment," *Computer Communications*, vol. 34, no. 3, pp. 326–336, 2011.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 683146, 10 pages
http://dx.doi.org/10.1155/2014/683146

Research Article

Distributed Relay-Assisted Retransmission Scheme for Wireless Home Networks

Seunghyun Park,[1] Hyunhee Park,[2] and Eui-Jik Kim[3]

[1] *Center for Information Security and Technologies, Korea University, 136-713 Seoul, Republic of Korea*
[2] *Institut National de Recherche en Informatique et en Automatique (INRIA), 35042 Rennes, France*
[3] *Department of Ubiquitous Computing, Hallym University, 39 Hallymdaehak-gil, Chuncheon-si, Gangwon-do 200-702, Republic of Korea*

Correspondence should be addressed to Eui-Jik Kim; ejkim32@hallym.ac.kr

Received 28 October 2013; Accepted 30 March 2014; Published 28 April 2014

Academic Editor: Jongsung Kim

A relay transmission is a promising technology to improve network performance in dynamic infrastructure. In this paper, we propose a distributed relay-assisted retransmission (DRR) scheme in multirate wireless home networks. The idea is to exploit overhearing nodes to retransmit on behalf of sender node after receiving the block acknowledgement (B-ACK) from destination node. For the first transmission, a basic relay (BR) node is used by considering the high data rate between source node and BR node. And then, for the retransmission, a retransmission relay (RR) node is used by considering the high data rate between RR node and destination node. The DRR scheme extends a distributed reservation protocol in WiMedia home networks and inquires the candidate relay node as BR nodes and RR nodes during beacon period. In addition, the DRR scheme can minimize control overhead for relay transmission because all nodes should send and listen to the beacon frames of neighbor nodes during beacon period. We also present the relay decision scheme and channel allocation procedure for maximizing the efficiency in the DRR scheme. Extensive simulation results demonstrate that the DRR scheme can improve the overall throughput by 40% and reduce the energy consumption by 47% compared with nonrelay transmission schemes when the number of nodes increases.

1. Introduction

Recently, there has been a growing interest in relay technologies to extend the coverage and improve the reliability of wireless networks by exploiting the spatial diversity gains [1]. In basic relay scheme, packets are transmitted along the relays via a store-and-forward manner, and thus the use of relays does not guarantee perfect transmission (i.e., no error transmission) [2, 3]. Dealing with this, diverse retransmission mechanisms are applied at relays to improve the successful transmission rate over wireless networks. In addition, the multirate transmission mechanism is one of the important relay transmissions to improve the system performance. Most of the wireless networks (i.e., IEEE 802.11 series, 802.15.3 series, and WiMedia MAC) can support multiple transmission rate by adaptively choosing the most appropriate modulation under the current channel conditions [4–6]. For instance, IEEE 802.11a/b wireless local area networks (WLANs) provide diverse transmission rates depending on the distance between source and destination nodes (i.e., 1 Mbps at 100 m, 2 Mbps at 74.7 m, and 11 Mbps at 48.2 m) [7]. The rate adaptive mechanisms by diverse PHY modulations have been studied based on request to send (RTS) and clear to send (CTS) for IEEE 802.11 WLANs. In such multirate wireless networks, relay transmission can improve the overall system performance and reduce the energy consumption, since relay transmission supported that the high transmission rate can reduce the transmission time compared with direct transmission over a lower data link.

Relay communications have been investigated and included in long term evolution-advanced (LTE-A) and IEEE 802.16 m candidates for the international mobile telecommunications-advanced (IMT-A fourth generation) standards [8, 9]. Representatively, the automatic repeat request (ARQ) mechanism can deliver reliable transmission over multicast and broadcast networks [10].

However, the retransmission of failed packets using ARQ schemes may cause a significant delay problem since the transmission failed packets are retransmitted individually and the retransmissions have to be repeated until every destination nodes receive all packets correctly. Furthermore, many studies have been researched to support the relay transmission in various wireless networks. The multihop mechanisms for the relay transmission have been emphasized as the cooperative communications. Liu et al. propose the cooperative MAC by defining the helper ready to send (HTS) using the overheard transmission [11]. Cetinkava and Orsun suggest the cooperative MAC protocol by choosing proper backoff window to be achieved with no compromise in throughput performance for dense wireless networks [12]. Shin et al. provide the use of beacon period to select relay nodes in distributed wireless networks [13]. Wang et al. address the dual communication mode considering different transmission paths with peer to peer path and relay path in ultra-wide band (UWB) WPANs [14]. However, there still exists the problem of compatibility between the relay node and the retransmission problem in distributed wireless networks. The previous work has the control overhead problem; if nodes want to relay packets, they should learn the information of neighbor nodes. In addition, the overall throughput and power consumption of the system suffer from sending the packets through the same path between relay node and retransmission node even though the packets are failed by the relay path.

In this paper, we propose a distributed relay-assisted retransmission (DRR) scheme based on the WiMedia home network standard by using the distributed beacon period. As our knowledge of prior studies, any significant relay mechanism for distributed network to minimize the overheads due to control frames has not been proposed before. Consequently, we focus on the considerations of how to avoid additional control frames when neighbor nodes are collecting information of neighbors and how to maximize the successful transmission rate with the efficient retransmission assisted relay nodes. As a 1st step, we define a distributed relay decision procedure to determine a basic relay (BR) node. Each node manages the modulation support to neighbor (MSN) table which includes the modulations between a node and neighbors. To acquire modulation information of neighbors, we define a new information element for relay decision (RD IE) which includes the request and response frames according to modification of the beacon frame. After all, since all the exchange process of beacon frames is executed during beacon period while RD IE is attached to the default beacon frame, there are no more header frames added except for the minute size of RD IE. Through the MSN table and RD IE, the BR node is selected, then the source node can transmit the packets to the BR node, and the BR node also can transmit the packets to the destination node, respectively. However, this process does not guarantee perfect packet transmission, because the packet error or transmission fail can occur during the source to BR node path or BR node to destination path. Therefore, we also define a relay-assisted retransmission to support the retransmission for the reliable communication as a 2nd step. Specifically, if the intended destination node does not receive the transmitted packets, a retransmission relay (RR) node is elected by investigating the channel condition between the candidate relay node and the destination node. Extensive simulation results demonstrate that the DRR scheme can improve the overall throughput by 47% compared with direct transmission and can reduce the energy consumption up to 40% according to the number of nodes.

The remaining of this paper is organized as follows. In Section 2, we introduce the system model and background of this work. In Section 3, we describe the proposed relay-assisted retransmission scheme. In Section 4, we perform extensive simulations to evaluate the performance of our scheme. Section 5 concludes this paper.

2. System Description

2.1. WiMedia MAC. In the WiMedia MAC, the channel time is divided into fixed-length superframes and each superframe consists of discrete media access slots (MASs), as depicted in Figure 1 [6]. In addition, the superframe consists of a beacon period (BP) and a data transfer period (DTP). During the BP, each node should choose an empty beacon slot in order to transmit its own beacon frame. To occupy an empty beacon slot, all nodes must execute beacon hearing process, in which a new node waits and listens to beacon frames during few superframes. From the received beacon frames, the node can determine an idle beacon slot and transmit its own beacon frame with beacon transmitting rate of 53.3 Mbps during the beacon slot. Furthermore, the DTP defies two distributed channel access mechanisms: a contention-free channel access mechanism as a distributed reservation protocol (DRP) and a contention-based channel access mechanism as a prioritized channel access (PCA). The DRP is a kind of time division multiple access (TDMA) protocol, in which nodes have the exclusive right of transmission during their reserved time slots [15]. On the other hand, the PCA is generally used to send control frames and excessive data of the reservation block as well as to transmit asynchronous traffic with a variable bit rate (VBR) in unreserved periods [16].

WiMedia MAC supports the high speed, short range communications in order to support multimedia transmissions in distributed home networks [17]. Eight traffic classes and four access categories are defined, which can be applied to different QoS and supports both isochronous and asynchronous data types with DRP and PCA. In the proposed DRR scheme, we utilize DRP because it provides robust operations compared with the centralized scheme. In addition, in WiMedia MAC standard supporting UWB PHY, eight kinds of transmission rates are introduced in Table 1. According to the channel condition, PHY modulation algorithm is selected to satisfy the current condition such as bit error rate (BER), signal to noise ratio (SNR), and received signal strength indicator (RSSI). Although the source node should transmit the beacon at the lowest rate (i.e., 53.3 Mbps), the packets are sent at the supported rate as Table 1 received in the PHY capability information element (IE) [6]. The source node may send the data rate information through the link feedback IE attached to the beacon frame. The optimal decision of data

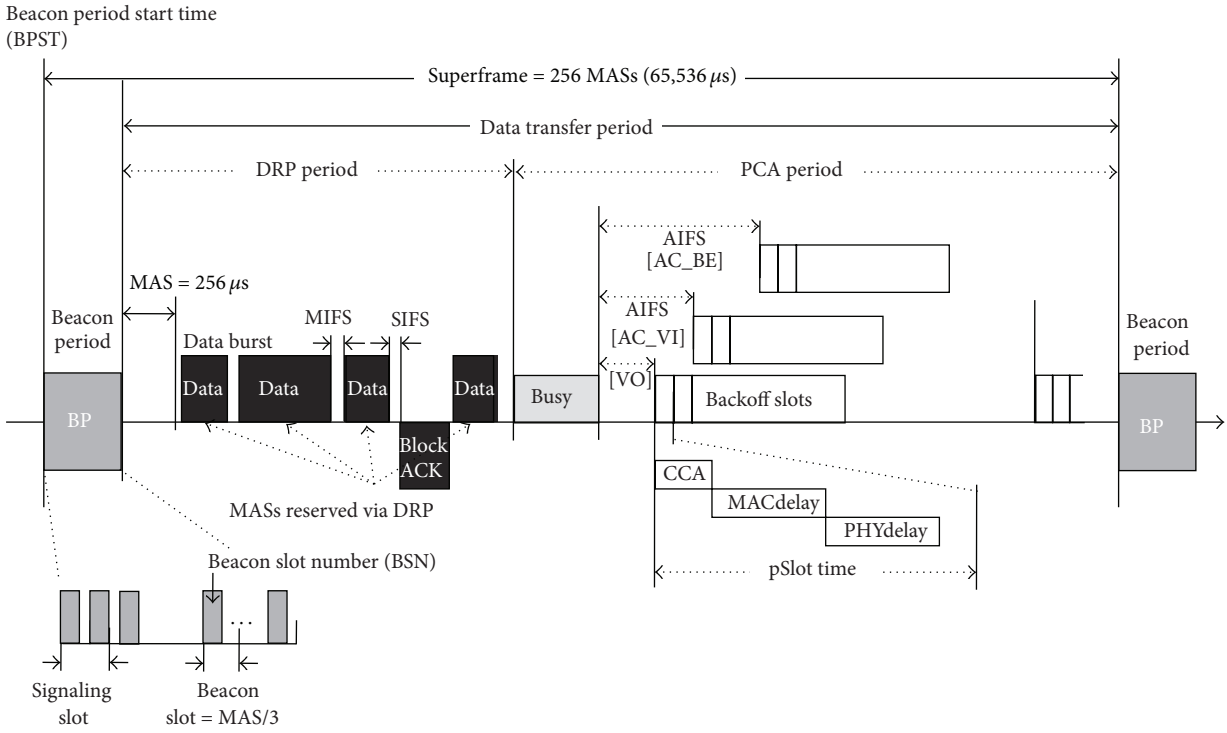


Figure 1: WiMedia superframe structure.

Table 1: Data rate-dependent modulation scheme.

Data rate (Mb/s)	Modulation	Coding rate (R)
53.5	QPSK	1/3
80	QPSK	1/2
106.7	QPSK	1/3
160	QPSK	1/2
200	QPSK	5/8
320	DCM	1/2
400	DCM	5/8
480	DCM	3/4

rate affects the performance of networks and the acceptable BER. Therefore, the source node should consider the recommended rate from the target or should find the optimal rate.

In addition, block-acknowledgement (B-ACK) mode was examined because it helps to save wasteful control space because duration of minimum interframe space (MIFS) between data transmissions is much shorter than short interframe space (SIFS). The superiority of B-ACK has been studied to prove its high throughput achievement [18, 19]. However, if the retransmission policy is used together or the channel condition is bad, B-ACK can deteriorate the performance of networks. In the DRR scheme, as the BER is considered to be up to 10^{-4}, that is, BER is not severe, B-ACK mode is adopted to show the improved throughput.

2.2. Channel Model. In wireless systems, there exist much complicated propagation characteristics of radio waves. Therefore, various channel models were studied to describe the property of wireless channel [20, 21]. Let $G_T(i)$, $G_R(i)$, and $\delta_{i,j}$ be the antenna gains of the transmitter and the receiver of stream i and the distance between the transmitter of stream i and the receiver of stream j, $j \neq i$, respectively. Then the average received signal power of stream i is given by [22]

$$P_R(i) = \kappa_1 G_T(i) G_R(i) P_T(i) \delta_{i,i}^{-\alpha}, \quad (1)$$

where α and $P_T(i)$ are the path loss exponent and the transmission power of stream i, respectively, and κ_1 is a constant depending on wavelength. To obtain $P_T(i)$, $P_R(i)$ is set to the receiver sensitivity as specified in [21]. The receiver sensitivity indicates the threshold of the received power for successful transmission. If the stream i experiences interference by the stream j, then the interference power between them, $I_{j,i}$, is given by

$$I_{j,i} = \kappa_1 G_0 G_T(j) G_R(i) P_T(j) \delta_{j,i}^{-\alpha}, \quad (2)$$

where G_0 is the cross correlation which is assumed to be a constant. The cross correlation is the impact incurred between any two streams caused by using spreading code.

Assuming that the CDMA achieves a higher channel capacity than the TDMA, the interference power should be less than the background noise, that is, $I_{j,i} \leq N_0 W$, where N_0 and W are the one-sided spectral density of white Gaussian noise and the channel bandwidth, respectively. If only one

Table 2: Relation table between the distance and the data rate.

Rate (Mb/s)	480	400	320	200	160	106.7	80	53.3
Distance (m)	2.8	3.2	3.7	4.9	5.5	6.9	8.0	10

flow is allowed to transmit at a time the achievable data rate of the ith flow, R_i [23], can be obtained as

$$R_i = W\log_2\left(\frac{k_1 G_0 G_T(j) G_R(i) P_T(j) \delta_{i,j}^{-\alpha}}{N_0 W} + 1\right). \quad (3)$$

In WiMedia system, data rate determined by PHY modulation is limited to eight discrete values. When a distance $\delta_{i,j}$ between nodes is given, we can calculate the data rate $R_i(\delta_{i,j})$ and then it can be mapped to a discrete rate in the WiMedia standard using the mapping table in Table 2. However, Table 2 does not consider the real world conditions such as BER and SNR conditions. When the source node finds the appropriate data rate that can improve the throughput at the best, it should consider many channel characteristics such as PHY modulation, coding rate, frame size, and acknowledgement mechanism. The coding rate is one of the most important factors to contribute for higher throughput as it determines the data rate. The selection of data rate affects the throughput as well as the BER [24]. As the higher coding rate generally implies the bigger BER, it is a tradeoff to improve the performance of the networks. Therefore, with the desirable BER requirement, the most appropriate coding rate should be found. Therefore, we find the coding rate; we can derive SNR as

$$\mathrm{SNR}_i = \frac{P_R(i)}{N_0 W}. \quad (4)$$

After all, we can find the best coding rate after the calculation of the current SNR in Table 2. When the desirable BER condition is given, that is, BER is 10^{-5}, the coding rate value becomes lower as SNR is the worst. Low coding rate means that many redundant bits are added to the original packet. For example, 1/10 indicates that the source sends the identical ten bits to transmit one bit. In order to decide the appropriate coding rate, the intersection of current SNR and BER is considered. After all, the data rate is calculated according to the PHY modulation. When the PHY operates as BPSK, which sends one bit per one Hertz, data rate can be calculated from the coding rate by multiplying bandwidth.

3. Distributed Relay-Assisted Retransmission Scheme

In this section, we first introduce the distributed relay decision procedure as the 1st step. Then, the relay-assisted retransmission will be followed. In addition, we provide a distributed channel allocation procedure for each path of relay.

The basic idea of distributed relay-assisted retransmission is to have intermediate nodes that overhear a failed packet to retransmit the packet on behalf of the source node. We provide the intuition into the potential benefits of relay-assisted retransmission using the four-node network in Figure 2. We denote P_{xy} as the packet delivery path from x to y. Let us look for the best strategy to deliver a packet from source (S) to destination (D). The simplest strategy is to transmit the packet directly from S to D as P_{SD}. The average of direct communication, P_{SD}, takes

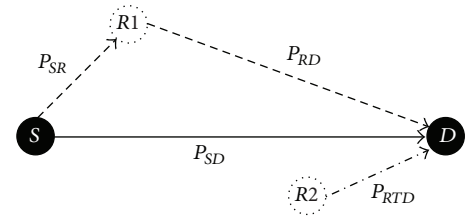


Figure 2: A four-node network containing source (S), destination (D), and relay (R).

$$T_{SD} = \frac{1}{\lambda_{SD}}, \quad (5)$$

transmission to deliver a packet, where λ_{xy} means the packet delivery rate from x to y. On the contrary, an alternative is to exploit relay transmission in this topology for time benefit. The relay strategy is to transmit a packet through two paths such as P_{SR} and P_{RD}, in which P_{RD} means the path of relay and destination. Here, the important assumption is that λ_{SR} has the higher transmission rate than the transmission rate of λ_{SD}. Therefore, in this topology, there are two approaches: S sends a packet to D directly or S sends a packet to $R1$, where $R1$ then forwards it to D. S uses the approach that requires the fewest transmissions. After that, the average time of relay transmission is given by

$$T_{\text{topology}} = \min\left(\frac{1}{\lambda_{SR}} + \frac{1}{\lambda_{RD}}, \frac{1}{\lambda_{SD}}\right), \quad (6)$$

where T_{topology} should choose the minimum transmission rate between the relay path and direct path. In case, if the transmission of relay path is failed from P_{SR} or P_{RD}, then another device, that is, $R2$, can be joined to retransmit a packet. In order to support relay based retransmission, P_{RTD} is defined as the path of retransmission and destination in Figure 2. Based on the distributed relay assisted retransmission, we can derive the expected number of transmission as

$$T_{RTD} = \sum_{k=1}^{\infty} k\lambda(k), \quad (7)$$

where $\lambda(k)$ is the probability of taking k transmission to deliver a packet.

Octets: 1	1	1	2	1	2	1
Element ID	Length	RC IE command	Relay node address	Relay DRP allocation	Destination node address	Data rate

Value	RC IE command
0	Data rate request
1	Data rate response
2	Relay request
3	Relay response
4	Relay retransmission
5	Supported data rate
6-7	Reserved

Figure 3: Relay communication information element (RC IE).

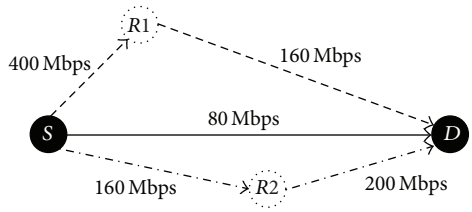


Figure 4: An example for simple topology.

Neighbor node ID	Data rate between S and R (Mb/s)
D	80
$R1$	400
$R2$	160

Figure 5: An example for the MSN table.

3.1. Distributed Relay Decision. A beacon frame consists of a MAC header and a payload. The MAC header follows the WiMedia standard and the payload includes one beacon parameter and several IEs [6]. Every IE has a unique identifier (ID) and is attached to the beacon frame payload in an increasing ID value order. After all, by listening to beacon frames from neighbors, every node can obtain communication types and the neighbor information. Specifically, the beacon frame includes several IEs to configure various functions such as beacon period occupancy IE (BPOIE), DRP IE, and identification of IE. In order to support relay based transmission, we define a new relay communication information element (RC IE) as shown in Figure 3. It shows the elements of RC IE and RC IE command value. The relay node address field notifies the selected relay node to relay communication during the relay DRP allocation. The data rate to destination field is filled with the enumerated values (i.e., 0–7) as mark of data rate, modulation, and coding rate defined in the WiMedia standard to prevent consuming large bits [6].

A node needs to listen to beacon frames from other nodes during a number of superframes. In so doing, the node can reduce the possibility that multiple nodes use the same beacon slots in the current beacon period. Thus, the collision problem of beacon slots can be mitigated. Moreover, by collecting beacon frames during multiple superframes, it is possible to detect the existence of hidden node problems because a beacon frame of neighbor node includes the information of the neighbor's neighbor nodes [25]. Furthermore, during BP, each node listens to the beacon frames of neighbors and then estimates the transmission rate to each neighbor device from the signal strength by channel model.

Through the listening and the exchanging of beacon frames, we prepare the relay decision procedure for P_{SR} and P_{RD}. For better understanding, we describe an example of the relay decision procedure. Let the four nodes be located randomly in the network and assume that they can reach each other. In Figure 4, S forms the MSN table from listening to beacon frames of neighbors during several superframes, as shown in Figure 5. It is noted that the MSN table is controlled between S and Rs, which means S cannot recognize the transmission rate between R and D. Therefore, the relay node is selected by comparing only the transmission rates of P_{SD} and P_{SR} by exchanging RC IE. Following the gathering of the data rate information from RC IE commands of all candidate relay nodes, S can calculate the smallest relay

transmission time (T_{SR}). Firstly, when B-ACK mode is used, direct transmission time, T_{SD}, is given by

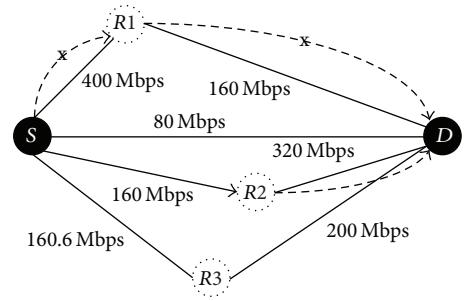
$$T_{SD} = \frac{N_{pk} \cdot l_{MH}}{R_b} + \frac{l_{\text{back}} + l_{MH}}{R_b} + \frac{N_{pk} \cdot l_{pl}}{R_r} + T_M \cdot \left(N_{pk} - 1\right) + T_S, \tag{8}$$

where N_{pk} is the number of packets, and l_{MH}, l_{back}, and l_{pl} denote the size of general MAC header, the payload size of the B-ACK, and the size of packets (bits), respectively. Also, T_M and T_S denote the time of MIFS and SIFS. In general, the MAC headers and control frames are delivered at the base data rate, R_b. On the contrary, the payload of packets is transmitted with supported data rate, R_r, by the flow. In addition, we can derive the time of relay transmission. T_{SRD} can be acquired from the repetition of (8) by summation of T_{SR} and T_{RD}, which is given by

$$T_{SRD} = 2 \cdot \left\{ \frac{N_{pk} \cdot l_{MH}}{R_b} + \frac{l_{\text{back}} + l_{MH}}{R_b} + \frac{N_{pk} \cdot l_{pl}}{R_r} + T_M \cdot \left(N_{pk} - 1\right) \right\} + 2T_S. \tag{9}$$

From the example topology in Figure 4, the relay node is determined $R1$ by exchanging the RC IE command such as data rate request and data rate response, because the transmission rate of P_{SR1} is large among the transmission rates of P_{SR1} and P_{SR2}. And then, S delivers the RC IE with relay response of RC IE command to $R1$ in order to decide the appropriate relay node.

3.2. Relay-Assisted Retransmission. When the channel between S and D is very poor, frequent relaying of ACKs may occur. In that case, it may be more efficient to employ a mesh network based approach (i.e., S sends packets to a relay which forwards them to D) as almost all packets will be relayed anyway. However, relays detect failed transmission through the overheard of B-ACK. Eligible relay nodes which have the packet by overhearing the transmission transmit own RC IE (i.e., relay retransmission and supported data rate). If there are several eligible relay nodes in the topology, R should compare the supported data rate of eligible nodes; then R determines the relay node for retransmission by considering the largest transmission rate among P_{RD}'s. In Figure 6, through the MSN table for retransmission, the best transmission rate between R and D is $R2$ as 320 Mbps. Because the relay node, which has the largest transmission rate, has also the stronger connectivity to the destination node since it has a higher chance of successfully transmitting the packet, if a retransmission is failed with $R2$ (i.e., the highest RSSI), all neighbor nodes can hear the B-ACK frame, and only $R2$ prepares the retransmission again after T_S. That is, non-relay nodes ignore the B-ACK frames, even though they could hear that the retransmission is failed.



Neighbor node ID	Data rate between R and D (Mb/s)
$R1$	160
$R2$	320
$R3$	200

Figure 6: An example for retransmission of DRR scheme and MSN table.

3.3. Medium Access Slot Allocation Procedure. In order to consider the medium access slot (MAS) allocation for relay transmission, the distributed reservation period is used by attaching DRP IE to a beacon frame. We should consider two reserved periods: S to R and R to D, respectively. The DRP is based on a TDMA style reservation [6]. That is, each node reserves its transmission time slots; it needs to reserve the time slots by sending DRP IE. When S broadcasts the beacon frame with the DRP allocation field in DRP IE during BP, all nodes should be aware of the relay node that is selected by receiving DRP IE. And then, the relay node figures out the information of data transfer period. During the reserved MAS block, only the reservation owner can access the channel to deliver packets; that is, DRP will not bring about any channel collisions. First, S should send relay request in RC IE attached to the beacon frame.

4. Performance Evaluation

In this section, the overall throughput and the energy efficiency have been measured when the number of nodes is increasing in order to evaluate the performance improvement of the DRR scheme. The overall throughput means the transmission rate to deliver the packets successfully between S and D considering delay, control space, and frames. Both with no BER and with BER have been evaluated to find the effect of BER on the DRR scheme while the retransmission policy with B-ACK mode is implemented.

4.1. Simulation Parameters. To validate performance of WiMedia MAC and DRR scheme, we develop simulators using Matlab (R2008b) [26, 27]. In order to validate the performance of the DRR scheme, we adopted retransmission policy and B-ACK filling the sending buffer similar to the actual environment. Parameter used in this simulation is shown in Table 3. The PHY model is BPSK described in the channel model. The bandwidth size is set to the closest value

Table 3: Simulation parameters.

Parameter	Value
Max. beacon length	96 slots
Superframe size	65,536 μs
MAS size	256 μs
Beacon slot size	85 μs
Space size	10 m by 10 m
Packet size	512, 1024, 2048, 4096 bytes
PHY header size	5 bytes
MAC header size	10 bytes
B-ACK payload size	7 bytes
MIFS time	1.875 μs
SIFS time	10 μs
Beacon transmission rate	54 Mb/s
Maximum bandwidth	540 MHz
Transmission power	−41.25 dBm/MHz
Receive power	−63 dBm/MHz

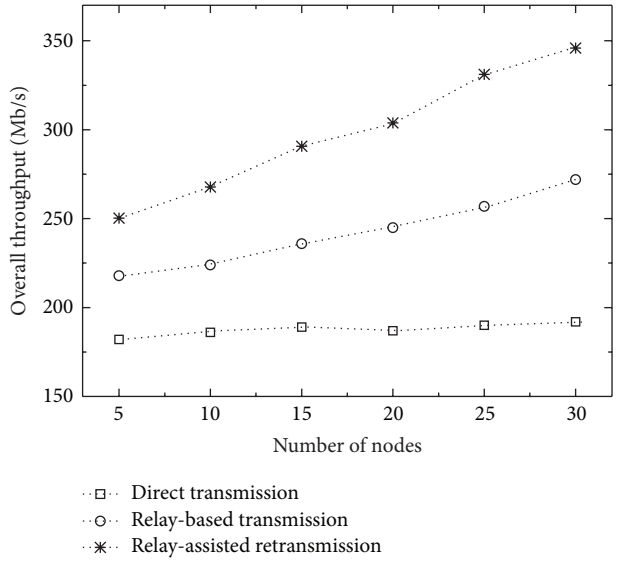


Figure 7: Overall throughput as direct transmission, relay-based transmission, and relay-assisted retransmission with no BER.

to effect similar data rate to the WiMedia MAC standard [6]. In this simulation, since the lowest coding rate is 1/10, basic data rate becomes 54 Mbps. We set free space size as 10 meters by 10 meters considered as home networks. Basically, the transmission power is fixed; however, it is considered to be affected by the Rayleigh fading to deliberate the characteristic of wireless channels. We assume that all nodes in the home networks can hear each other. The packets are generated by constant bit rate. The number of packets is based on the period of 112 MASs (i.e., 1 MAS is 256 μs) because the single node can reserve until 112 MASs at most during one superframe. We also assume that the consecutive collisions in BP do not occur because the mechanism to solve the collision in WiMedia MAC is another research area.

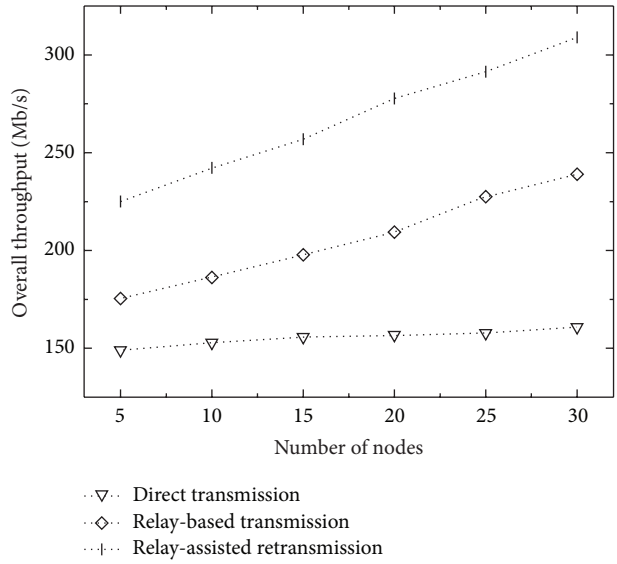


Figure 8: Overall throughput as direct transmission, relay-based transmission, and relay-assisted retransmission with 10^{-5} BER.

4.2. Simulation Results

4.2.1. Overall Throughput. Although the networks support high bandwidth and data rate, transmission rate cannot be equal to them. This is because the transmission rate depends on the control frames, queuing delays, packet losses, and other error environment. Therefore, the overall throughput, S_o, can be obtained as the number of successfully received packets, which is given by

$$S_o = \frac{N_{\text{suc}} \cdot l_{pl}}{\sum T_{tr}}, \tag{10}$$

where N_{suc} is the number of successfully received packets. In addition, the sum of T_{tr} is the total transmission time. This is the equation for overall throughput with no retransmission. When the failed packets occur, the latency per a packet should be considered due to the retransmission

$$S_o^{ret} = \frac{\sum_{k=1}^{N_{tr}} S_{tr,k}}{N_{tr}}, \tag{11}$$

where $S_{tr,k}$ is the throughput for the kth packets and N_{tr} is the number of all sent packets.

The packets are retransmitted up to three times, BER is set to 10^{-5}, and the packet size is 2048 bytes in Figure 8. On the other hand, Figure 7 shows the overall throughput without BER while Figure 8 represents the overall throughput considering BER. We can discover that the more nodes are deployed, the higher overall throughput can be obtained in both Figures 7 and 8. That is practicable because relay candidates are increasing, that is, better choices for the higher throughput. In both Figures 7 and 8, with an adoption of DRR scheme, the throughput improves 26% and 45% with no BER and 27% and 47% with 10^{-5} BER compared to the direct transmission and relay-based transmission, when the

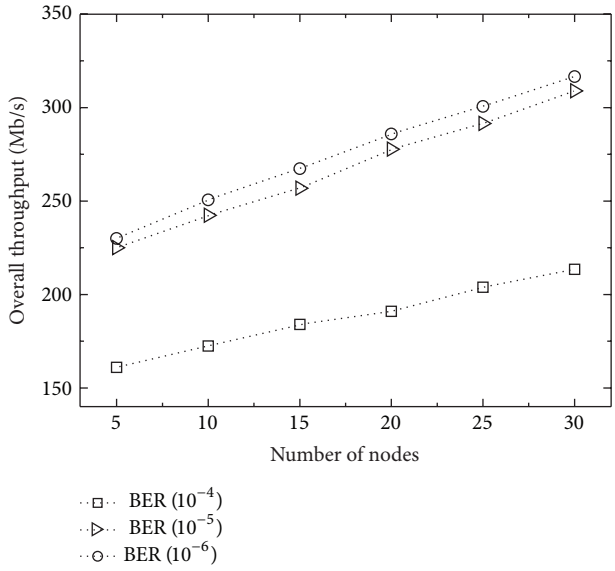


Figure 9: Overall throughput as various BER in case of DRR.

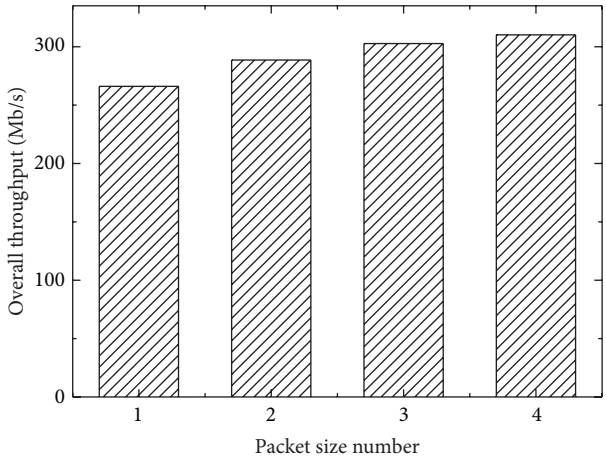


Figure 10: Effect of packet size on BER in case of DRR.

number of nodes is 20. In other words, the consideration of BER does not affect (or barely affects) the improvement by relay. When the packets are relayed, the distance between the source node and the relay node and the distance between the relay node and the destination node become shorter. Because the short distance makes SNR condition higher, the data rate becomes higher than direct transmission.

In addition, BER is the important factor to influence the overall throughput. Also, packet error rate (PER) can be calculated with BER as $\text{PER} = 1-(1-\text{BER})^{l_{pk}}$. Figure 9 shows the overall throughput increase as BER is smaller. The effect of the DRR scheme can be notable in Figure 9.

We compute the overall throughput per various packet sizes. Figure 10 shows that the bigger the packet size, the more improved the overall throughput. This is because the overheads by control frames are decreased with the bigger packet size. However, when we add BER into overall throughput, packet size directly influences the throughput and deteriorates the performance. The reason that the overall throughput decreases after 1024 bytes packet size proves the tradeoff between the burst transmission and the packet error rate. Despite of the tradeoff, the relay-assisted retransmission shows higher overall throughput when the packet size is large. This is because the overhead by control frames influence on the throughput remarkably. In addition, we devise the improvement ratio by

$$\tau\,(\%) = \left(\frac{S_{\text{relay}} - S_{\text{DRR}}}{S_{\text{relay}}} \times 100\right), \tag{12}$$

where S_{relay} and S_{DRR} mean the overall throughput in case of relay-based transmission and relay-assisted retransmission, respectively. When the packet sizes of S_{relay} and S_{DRR} are 512 bytes, 1024 bytes, 2048 bytes, and 4096 bytes, each τ as improvement ratio shows 7.7%, 9.2%, 10%, and 12%, respectively.

4.2.2. Energy Consumption. The energy consumption per bit, E_b, between two nodes, can be computed as

$$E_b = \frac{T_{pk}\cdot(\omega_{tx}+\omega_{rx}) + \omega_{\text{idle}}\cdot T_S}{N_{pk}\cdot l_{pl}}, \tag{13}$$

where ω_{tx} and ω_{rx} mean the power of transmitting and receiving, respectively. Also, T_{pk} is the time to transmit packets except for the control frames. We assume that nodes enter to sleep period during data transfer period for other nodes and there is no energy consumption during the sleep period. The power consumed in idle state ω_{idle} is zero. Despite the activity of the relay, the relay-assisted retransmission time can be shorter than the relay-based transmission as shown in Figure 11. Due to the two separate periods, end node can go to sleep period, which helps to save the energy as well. We define the energy saving as $\omega_{\text{DRR}}/\omega_{\text{relay}}$ for a conspicuous contrast of DRR scheme with the relay-based transmission. Figure 11 describes that the DDR scheme can reduce the energy consumption by 40%–47% according to the number of nodes. This appearance is related to Figure 9. Because the overall throughput improves by increasing nodes, it means that the transmission time becomes short. When the overall throughput improvement ratio is 80% that means the DDR scheme consumes 80% energy comparing that the relay-based transmission consumes 100%. In that case, 40% energy consumption is saved with the DRR scheme.

5. Conclusion Remarks

In this paper, we propose a distributed relay-assisted retransmission scheme that employs a distributed relay path and relay-assisted retransmission to efficient system performance in distributed wireless home networks. The DRR scheme outperforms the direct transmission and relay transmission by examining the appropriate data rate for each path and separating each path for relay transmission and retransmission. In addition, the DRR scheme can shorten the transmission

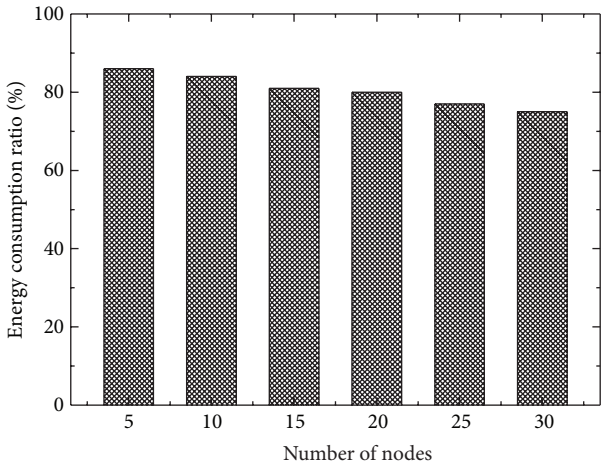


Figure 11: Energy consumption ratio with DRR.

time comparing to the direct transmission and the relay transmission, respectively. The extensive simulation results demonstrate that the performance gain of the DRR scheme is significant when the number of nodes is large. Consequently, it leads to the overall throughput improvement and the energy efficiency. Furthermore, the DRR scheme supports compatibility with the WiMedia standard by keeping the rule of beacon period.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by Hallym University Research Fund, 2014 (HRF-201402-009).

References

[1] H. Wang, S. Ma, and T.-S. Ng, "On performance of cooperative communication systems with spatial random relays," *IEEE Transactions on Communications*, vol. 59, no. 4, pp. 1190–1199, 2011.

[2] P. Kolios, V. Friderikos, and K. Papadaki, "Load balancing via store-carry and forward relaying in cellular networks," in *Proceedings of the 53rd IEEE Global Communications Conference (GLOBECOM '10)*, pp. 1–6, December 2010.

[3] L. Yang and G. B. Giannakis, "Ultra-wideband communications," *IEEE Signal Processing Magazine*, vol. 21, no. 6, pp. 26–54, 2004.

[4] IEEE Standard for Information Technology-Local and Metropolitan Area Networks-Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 2003.

[5] IEEE Standard for IEEE Amendment to Part 15.3: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for High Rate Wireless Personal Area Networks (WPAN): Amendment to MAC Sublayer, 2006.

[6] WiMedia Alliance, *ECMA-368 High Rate Ultra Wide Band PHY and MAC Standard*, ECMA, 3rd edition, 2008.

[7] H. Park, W. Kim, and S. Pack, "A deterministic channel access scheme for multimedia streaming in WiMedia networks," *Wireless Networks*, vol. 18, no. 7, pp. 771–785, 2012.

[8] W. Chung, C. Chang, and L. Wang, "An intelligent priority resource allocation scheme for LTE-A downlink systems," *IEEE Wireless Communications Letters*, vol. 1, no. 3, pp. 241–244, 2012.

[9] IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Broadband Wireless Access Systems Amendment 3: Advanced Air Interface, IEEE Std 802. 16m, 2011.

[10] J. Kao and F. Chen, "On RANC ARQ for wireless relay networks: from the transmission perspective," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2962–2976, 2013.

[11] P. Liu, Z. Tao, S. Narayanan, T. Korakis, and S. S. Panwar, "CoopMAC: a cooperative MAC for wireless LANs," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 2, pp. 340–354, 2007.

[12] C. Cetinkaya and F. Orsun, "Cooperative medium access protocol for dense wireless networks," in *Proceedings of the 3rd Annual Mediterranean Ad Hoc Networking Workshop*, June 2004.

[13] H. Shin, Y. Kim, S. Pack, and C.-H. Kang, "A distributed relay MAC protocol in WiMedia wireless personal area networks," in *Proceedings of the International Symposium on Parallel and Distributed Processing with Applications (ISPA '08)*, pp. 784–789, Sydney, Australia, December 2008.

[14] X. Wang, Y. Ren, J. Zhao, Z. Guo, and R. Yao, "Energy efficient transmission protocol for UWB WPAN," in *Proceedings of the IEEE 60th Vehicular Technology Conference, VTC2004-Fall: Wireless Technologies for Global Security*, pp. 5292–5296, September 2004.

[15] Z.-N. Kong, D. H. K. Tsang, B. Bensaou, and D. Gao, "Performance analysis of IEEE 802.11e contention-based channel access," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 10, pp. 2095–2106, 2004.

[16] X. Shen, W. Zhuang, H. Jiang, and J. Cai, "Medium access control in ultra-wideband wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 5, pp. 1663–1677, 2005.

[17] H. Park, S. Pack, and C.-H. Kang, "Dynamic adaptation of contention window for consumer devices in WiMedia home networks," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, pp. 28–34, 2011.

[18] T. Li, Q. Ni, T. Turletti, and Y. Xiao, "Performance analysis of the IEEE 802.11e block ACK scheme in a noisy channel," in *Proceedings of the 2nd International Conference on Broadband Networks (BROADNETS '05)*, pp. 551–557, October 2005.

[19] H. Chen, Z. Guo, R. Yao, and L. I. Yanda, "Improved performance with adaptive Dly-ACK for IEEE 802.15.3 WPAN over UWB PHY," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E88-A, no. 9, pp. 2364–2372, 2005.

[20] H. Park and C.-H. Kang, "A group-aware multicast scheme in 60GHz WLANs," *KSII Transactions on Internet and Information Systems*, vol. 5, no. 5, pp. 1028–1048, 2011.

[21] H. Park, S. Park, T. Song, and S. Pack, "An incremental multicast grouping scheme for mmWave networks with directional antennas," *IEEE Communications Letter*, vol. 17, no. 1, pp. 616–169, 2003.

[22] H. Park and C.-H. Kang, "Dynamic beam steering using directional antennas in mmwave wireless networks," *IEICE Electronics Express*, vol. 8, no. 6, pp. 378–384, 2011.

[23] L. X. Cai, L. Caai, X. Shen, and J. W. Mark, "Rex: a randomized Exclusive region based scheduling scheme for mmWave WPANs with directional antenna," *IEEE Transactions on Wireless Communications*, vol. 9, no. 1, pp. 113–121, 2010.

[24] F. A. Onat, A. Adinoyi, Y. Fan, H. Yanikomeroglu, J. S. Thompson, and I. D. Marsland, "Threshold selection for SNR-based selective digital relaying in cooperative wireless networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 11, pp. 4226–4237, 2008.

[25] V. M. Vishnevsky, A. I. Lyakhov, A. A. Safonov, S. S. Mo, and A. D. Gelman, "Study of beaconing in multihop wireless PAN with distributed control," *IEEE Transactions on Mobile Computing*, vol. 7, no. 1, pp. 113–126, 2008.

[26] Software Package, MATLAB R2008b, The Mathworks Inc., Natick, Mass, USA, 2008, http://www.mathworks.com.

[27] M. Clark, M. Mulligan, D. Jackson, and D. Linebarger, "Fixed-Point Modeling in an Ultra Wideband (UWB) Wireless Communication System," Matlab Digest, 2004, http://www.mathworks.co.uk/company/newsletters/digest/may04/uwb.html.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 507160, 11 pages
http://dx.doi.org/10.1155/2014/507160

Research Article

RETE-ADH: An Improvement to RETE for Composite Context-Aware Service

Milhan Kim, Kiseong Lee, Youngmin Kim, Taejin Kim, Yunseong Lee, Sungrae Cho, and Chan-Gun Lee

Department of Computer Science and Engineering, Chung-Ang University, Seoul 156-756, Republic of Korea

Correspondence should be addressed to Chan-Gun Lee; cglee@cau.ac.kr

Received 4 December 2013; Accepted 1 April 2014; Published 24 April 2014

Academic Editor: Jongsung Kim

We propose a new pattern matching algorithm for composite context-aware services. The new algorithm, RETE-ADH, extends RETE to enhance systems that are based on the composite context-aware service architecture. RETE-ADH increases the speed of matching by searching only a subset of the rules that can be matched. In addition, RETE-ADH is scalable and suitable for parallelization. We describe the design of the proposed algorithm and present experimental results from a simulated smart office environment to compare the proposed algorithm with other pattern matching algorithms, showing that the proposed algorithm outperforms original RETE by 85%.

1. Introduction

The past decade has seen the dropping costs for wireless personal devices such as smart phones and tablets while the capabilities of those devices continue to evolve rapidly. Interconnecting these computing devices or distributed sensors can be used to provide intelligent services, using a variety of sensing technologies of various sizes, from simple motion sensors to electronic tags or video cameras [1]. Furthermore, recent advances in software technology and computing devices have enabled revolutionary and user-customized digital services [2] such as ubiquitous computing. These differ from conventional communication models, requiring sophisticated integration of huge amounts of information on the fly to enable performing appropriate actions in a timely manner.

Because of these complex aspects, composite context-aware (CCA) architecture, which we describe in detail in Section 2, is promising for ubiquitous computing. Context-aware techniques can be used to provide the user with useful and intelligent applications by taking into account contextual information from the user's environment. The CCA architecture is composed of several elements, which include an *inference control function, interpretation function,* composite context information (CCI) *repository, and inference engine*. Among them, the *inference engine* which uses the *event-condition-action (ECA) engine,* is the key in speeding up CCA services. To construct an efficient *inference engine*, one of the most reasonable solutions is to implement an *ECA engine* by using pattern matching algorithm developed for production systems.

The RETE algorithm [3] is one of the most efficient forward inference algorithms that can be used for constructing an ECA engine; however, it has a few drawbacks [4–6]. While there have been a number of contributions to improving the pattern matching algorithm [7–10] in recent decades, not many of them considered an ECA engine within the CCA architecture as the target environment. In addition, although past contributions have led to notable performance improvements, the improved algorithms have also been more complex. For example, although the RETE' algorithm [10] has shown a performance improvement of more than 80% relative to RETE, it requires a complicated algorithmic structure and introduces time stamp in its data structure.

In this paper, we propose a new pattern matching algorithm called RETE-Alpha network Dual Hashing (RETE-ADH). We developed RETE-ADH for CCA services, and

herein we describe its validation through simulation in a CCA environment. It is a relatively simple algorithm that maximizes the use of referencing rather than comparing or searching all nodes. This characteristic makes RETE-ADH faster than other existing algorithms, as we show in the evaluations in Section 4. Because of its concise network structure, RETE-ADH is scalable, and thus it can handle the varying characteristics of recent complex network services. Furthermore, we predict that RETE-ADH can be implemented in parallel hardware for the same reason. The proposed algorithm efficiently searches only rules that can be fired by reconstructing the alpha network with hash tables. This can increase the speed of pattern matching considerably while providing a full set of matched results like RETE.

2. Background and Related Work

2.1. Composite Context-Aware Service Architecture. Context can be considered as a set of information that includes a user's activity, location, personal preferences, and current status. The most widely accepted formal definition of context is that of Abowd et al. [11]: "Context is any information that can be used to characterize the situation of an entity. An entity can be a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves." In a previous study [12], we have classified context into *unitary* and *composite* contexts. We have defined the *unitary* context as a basic building block that is not further divisible. Also, as shown in Figure 1, we have defined the composite context information as a high-level context abstraction by integrating or composing *unitary* context information with related multientities [12]. In other words, to provide composite contextual service, multiple *unitary* contexts can be combined. For example, suppose that a tourist is planning to visit an attraction near his home. Then, *visiting an attraction near home* would be considered as the composite context. To construct the composite context, the following various *unitary* contexts can be related to the composite context: *weather, allowed time, distance to the site, budget, accompanying people, transportations,* and so forth. Future context-aware services are expected to manipulate such complex composite context.

Figure 2 shows our proposed CCA service architecture. When a user sends composite context information to the CCA service, the service architecture requests/subscribes the CCI through the service layer. The user requests the CCI from the context-aware service, and then the service responds to the user's request from the CCI interface control function if the service is immediately able to find the proper presence information. Otherwise, the user subscribes the CCI to the context-aware service. Then, the composite context-aware service processes the CCI through the underlying CCI interface control function. Once the CCI interface control function finds the proper presence information for the context, the information is sent to the user. After receiving this request/subscribe command, the CCI interface control function saves the CCI to the CCI repository. The CCI repository can be utilized for later requests/subscriptions. In the CCI interpretation function, the request-subscribe command is differentiated through a function-type decision process and passed to the request/response and subscribe/notify processes. After the request/subscribe command is processed, the CCI is extracted by the CCI extraction function and then the corresponding rules are parsed and translated. By using the rule pattern matching algorithm, the inference engine finds proper presence information for the context. We adopt our proposed pattern matching algorithm for this inference engine.

Figure 3 shows an example of the processing of CCI in the proposed composite context-aware service architecture. The context-aware service should increase the capabilities of the user's smart devices in various contexts, such as the home, office, or shopping mall. Composite context-aware applications can monitor such regions and modify their behavior accordingly to help provide comfortable lifestyles. Based on this, the CCI needs to be expanded to consider the user's current environment. For example, the CCI can be expanded by using device ID, user ID, service ID, and location ID. The device can have unique parameters such as device name, type, provider, supported services, connected network status, location, and owner. In a similar fashion, the user ID has unique parameters such as user name, device name, user location, and user's status.

2.2. ECA Architecture and Pattern Matching Algorithms. The ECA pattern is composed of three modules: event, condition, and action. Each rule is expressed as IF <event-condition> THEN <action>. The <event-condition> part of the rule specifies the situation under which the actions are enabled, and it is composed of a logical combination of events. An event models some occurrence of interest in an application or in an environment. The <action> part of the rule is composed of one or more actions that are triggered whenever the <condition> part is satisfied [1].

The ECA architecture can be used to support the composite context-aware service, and one of the most efficient ways to construct an ECA engine is to adopt a pattern matching algorithm. Of course, we can adopt existing rule based pattern matching algorithms; for this reason, we analyze existing algorithms for the inference engine such as RETE [3], TREAT [5], and LEAPS [7]. However, using an algorithm specifically developed for CCA services would be expected to improve the quality of service; accordingly, herein we propose a new pattern matching algorithm.

2.3. Pattern Matching Algorithms. Rule-based systems execute actions based on the rules that are fired by the incoming facts. Each fact is an expression of a certain situation or environment in the real world. Each rule is a predetermined method for how the system should behave in a certain situation.

A typical rule-based system is composed of working memory, a knowledge base, an inference engine, and an action performer. The working memory is a space in which facts are saved; facts are frequently updated and changed in the system. The knowledge base is a space in which rules

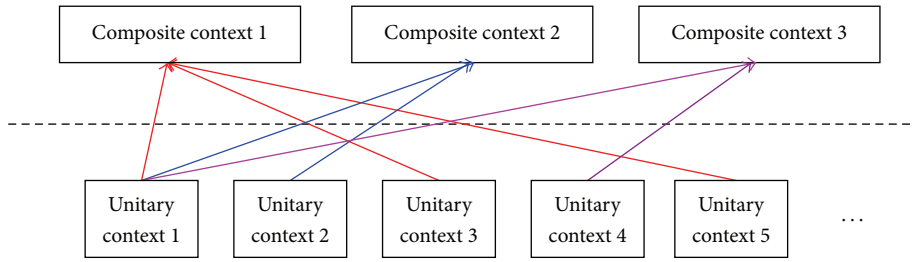


Figure 1: Derivation of composite context.

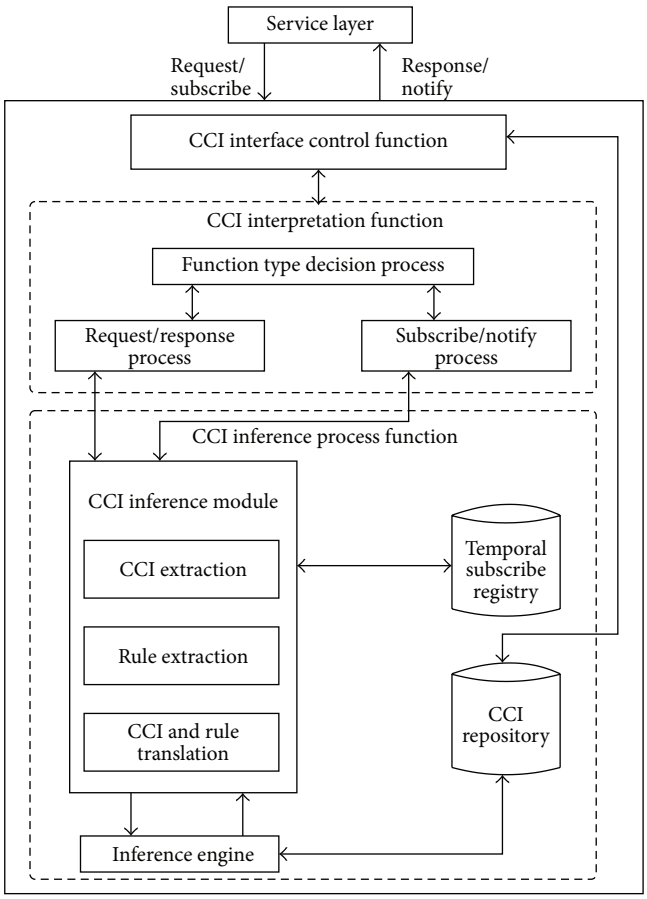


Figure 2: Architecture of composite context-aware service.

are saved. Conditions are tested under the criteria of the rules. The inference engine is a core part of the rule-based system; it checks whether the facts satisfy the rules based on a substitution method during each cycle of the inference engine. This test should be repeated for all the rules; therefore, there are usually a very large number of calculations in each cycle of the inference engine. This process in the calculation is called *pattern matching*. After finally deciding on a rule, the <action> part is executed; this process is called "firing." When a rule is fired, the facts in the working memory can be updated. The pattern matching time represents almost all of the processing time; for this reason, reducing the pattern matching time would directly improve the system's overall performance. An inference engine that uses the RETE or TREAT algorithm has a space called an agenda. The agenda temporally saves rules whose <condition> part matches the facts; this set of rules is called a *conflict* set.

In this section, we will give a brief overview of RETE, TREAT, and LEAPS, which are the most popular pattern matching algorithms that are used in many inference engines. We will analyze these algorithms and discuss their limitations.

2.3.1. RETE. Most pattern matching algorithms save in their working memory any partial matches produced during the previous inference cycle. Thus, they can avoid reevaluating the entire set of facts whenever changes are made. The RETE algorithm [3] is one such pattern matching algorithm; it maintains a network of partial matches to improve run-time efficiency. This network is composed of nodes that contain the facts. Each fact can be mapped to a token, which consists of a tag and a list of data elements. The tag indicates that the corresponding token has been added to or deleted from working memory.

There have been many previous efforts to address RETE's problems [4–6]. Among them, we specifically consider the issue of beta memory explosion. The RETE network has 2-input nodes, which are referred to as beta nodes. They send their outputs to beta memory, and because the RETE retains partial matches for performance reasons, the number of beta nodes increases rapidly, and thus maintaining the beta memory consumes vast memory resources.

2.3.2. TREAT. One of the main goals of the TREAT algorithm is to overcome a major problem of the RETE, namely, to reduce the overhead of network management. This algorithm is motivated by McDermott's hypothesis: "It seems highly likely that for many production systems, the retesting cost will be less than the cost of maintaining the network of sufficient tests" [13]. Unlike RETE, TREAT does not use beta memory; thus, it significantly reduces the overhead of network management. As a result, TREAT exhibits a maximum 50% better performance than RETE [5].

To avoid the use of the beta memory, TREAT recomputes the matches repeatedly [14]. This means that the job of finding the alpha nodes that correspond to the input of beta nodes is done repeatedly, which may render TREAT inappropriate for bounded algorithms.

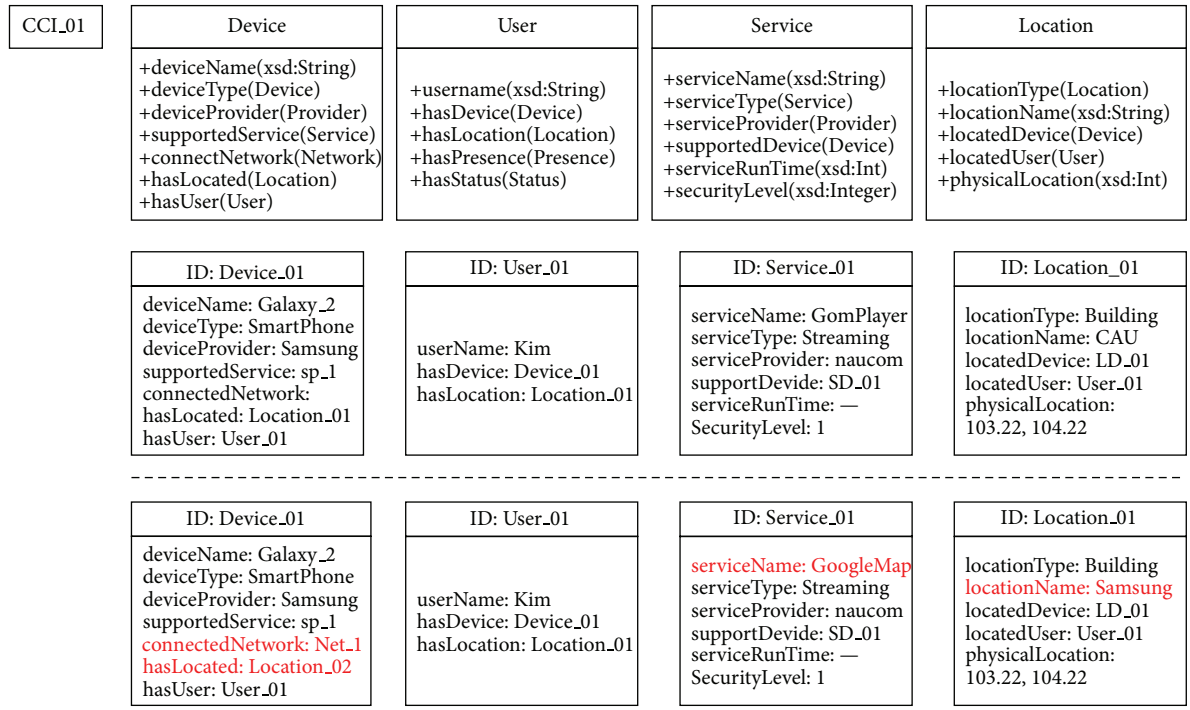


Figure 3: Example of composite context information structure.

2.3.3. LEAPS. Unlike RETE and TREAT, which use eager evaluation techniques, LEAPS uses lazy evaluation. Instead of generating all possible matches, LEAPS computes at most one match per cycle. It discards the conflict set, using a stack structure instead. Thus, it can reduce the computing time relative to the case of RETE and TREAT, which require processes for conflict resolution. Furthermore, LEAPS does not need beta memory. The stack management cost for LEAPS is known to be very low compared to most applications. LEAPS shows better performance than other algorithms, especially when the conditions are complex [15].

In spite of the advantages of LEAPS, there are some obstacles to use it generally. Its characteristic of producing at most one match per cycle makes it unsuitable for some applications [16]. Such applications include the simulator used herein to evaluate the proposed algorithm, and also general ECA engines. Similarly, composite context-aware services require a full set of match instances.

3. Proposed Algorithm

3.1. Overview. As mentioned before, RETE may consume a lot of resources due to its heavy use of beta memory. TREAT requires less memory than RETE, but may undergo an explosive amount of recomputations. LEAPS is more efficient than RETE or TREAT in terms of time and storage, but it may not fit some applications, including our own scenario in which its output of at most one match is not enough. For this reason, we propose a new pattern matching algorithm called RETE-ADH which extends the RETE algorithm with hashing techniques.

One of the primary features of the proposed algorithm is its adoption of double hashing in the alpha network; this increases the matching speed. Double hashing also reduces the number of beta nodes, thereby reducing the volume of the conflict set. This change, consequently, transforms many comparison operations into simple referencing operations; this is the main contributor to performance improvement by this algorithm.

Recently, a few approaches including alpha network hashing have been reported [8, 9]. In the alpha network, the alpha nodes are used to evaluate literal conditions of the facts. The fact data propagates through the next alpha node when it satisfies the current literal condition. The alpha node hashing is effective in the process when the propagation goes from an object-type node to an alpha node [8, 9]. In these approaches, an alpha node is added to a type-node, and the literal value is added as a key to the alpha node.

3.2. Core Algorithm. In our proposed algorithm, we use double hashing as follows.

(i) Each alpha node is hashed to variable nodes.

(ii) Each variable node consists of a variable name and a secondary hash table.

(iii) Each entry in the secondary hash table consists of a pair of fact attributes and a list of the related facts.

Note that in previous approaches using alpha network hashing [8, 9], all the facts in the alpha network have to be searched to build the beta network. In contrast, the proposed algorithm avoids useless alpha nodes by using the secondary

hashing table. For this reason, RETE-ADH can reduce the volume of the beta network and turn some comparison operations into referencing operations.

3.3. Case Study. Assume that we have the set of rules shown in Figure 4(a). In Figure 4, the rule searches for a stack of two blocks to the left of a block with a specific color. This rule has three conditions, which are enclosed by parentheses. Within each condition, let us denote a variable by enclosing it with angle brackets. For example, $\langle x \rangle$ indicates a variable x in the condition. Constants and identifiers are not enclosed by brackets. To fire this rule, we need facts satisfying the three conditions, which we will refer to as c_1, c_2, and c_3.

Assume that we have the fact (b1 ^ on b2), which satisfies condition c_1. Then, the fact that satisfies condition c_2 will be (b2 ^ left-of b3), since b2 in condition c_1 and b2 in condition c_2 should be matched. The fact satisfying condition c_3 will be (b3 ^ color $\langle c \rangle$). The $\langle c \rangle$ in the condition specifies the color that the user wants. Similarly, we can list the matched conditions as shown in Table 1. The RETE pattern matching algorithm constructs the alpha network with Table 1 as shown in Figure 4(b).

Figure 5 shows the alpha network of RETE-ADH based on the rule example shown in Figure 4(a). The matching process of RETE-ADH is similar to that of RETE; the difference is in how to construct the alpha network. In the RETE algorithm, one node is chosen in the alpha network and is attempted to be matched with the nodes of the beta network by using all the facts. In contrast, the RETE-ADH algorithm chooses facts in the alpha network that are highly likely to match with beta nodes. Assume that we try to match c_1 with c_2. In this case, c_1 has two variables $\langle x \rangle$ and $\langle z \rangle$. Because the hashing table is composed of the variables, RETE-ADH tries searching with these variables. Since c_2 has no $\langle x \rangle$, RETE-ADH searches c_2 with $\langle y \rangle$ of c_1. The c_2 has an identifier ^ left-of, and c_2 has a hashing table that sets $\langle y \rangle$ and $\langle z \rangle$ as a key. Each entry in hashing table of c_2 consists of a pair of fact attributes and a list of the related facts. In Figure 5, B2 and B3 are substituted into the $\langle y \rangle$ of c_1. Therefore, we can find the fact (B2 ^ color blue) by using the primary hashing table that sets $\langle y \rangle$ of c_2 as a key and the secondary hashing table that has B2 as a key. In this manner, we can find (B3 ^ color red) with B3. Note that RETE-ADH searches facts using the double hashing table instead of searching all of the facts of the alpha node, as mentioned above.

In the previous example, RETE has to apply 24 combinations of the facts to find condition matches in Figure 4(b); contrastingly, RETE-ADH tries only 4 combinations. In this specific example, the number of beta nodes is the same for both RETE and RETE-ADH; if there are many conditions and facts, there will be a huge number of beta nodes generated in RETE, making the matching execution time of RETE even worse than in Figure 4 example.

3.4. Characteristics. In our study, among RETE, TREAT, and LEAPS, we chose to extend RETE because it best fits our application purpose. If we were to use TREAT for our application, the recursive matching calculation of TREAT could become explosive because there is a huge number of facts in the composite context environment. We also consider LEAPS to be unsuitable because it produces at most one match per cycle, meaning that we cannot choose the most appropriate service by using LEAPS alone.

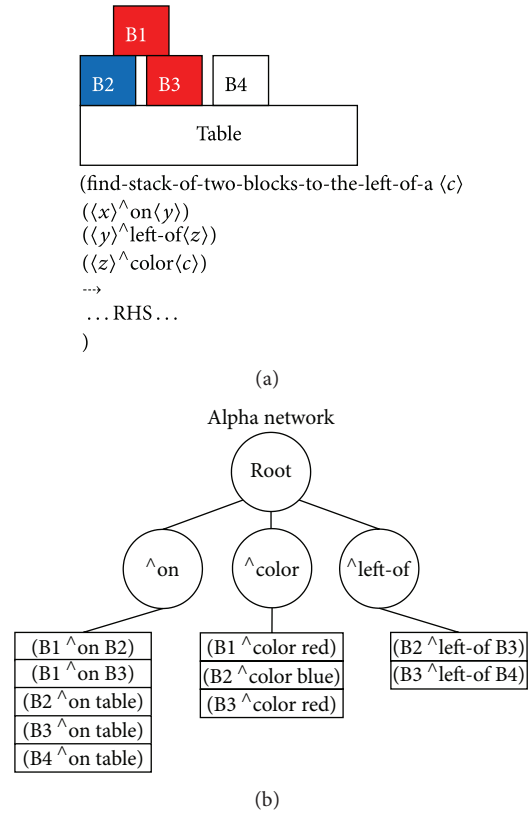


Figure 4: Rule example and its alpha network in RETE.

Table 1: Facts of the ω_n, where $n \in \{1, 2, \ldots, 9\}$.

ω_1:	(B1 ^on B2)
ω_2:	(B1 ^on B3)
ω_3:	(B1 ^color red)
ω_4:	(B2 ^on table)
ω_5:	(B2 ^left-of B3)
ω_6:	(B2 ^color blue)
ω_7:	(B3 ^left-of B4)
ω_8:	(B3 ^on table)
ω_9:	(B3 ^color red)

Figure 6 shows a concise comparison between RETE and our proposed algorithm with a rule (IF A = $\langle x \rangle$^ B = $\langle x \rangle\langle y \rangle$^ C = $\langle y \rangle$ THEN action) and the facts in Table 1. The original RETE tries to trigger the action with two 2-input nodes (Figure 6(a)) accompanying five comparison operations (One A node with three B nodes and one AB node with two C nodes) while RETE-ADH tries to trigger the action with simple referencing (Figure 6(b)) accompanying two operations (node A to B and node B to C). In this example, our proposed algorithm reduces the number of

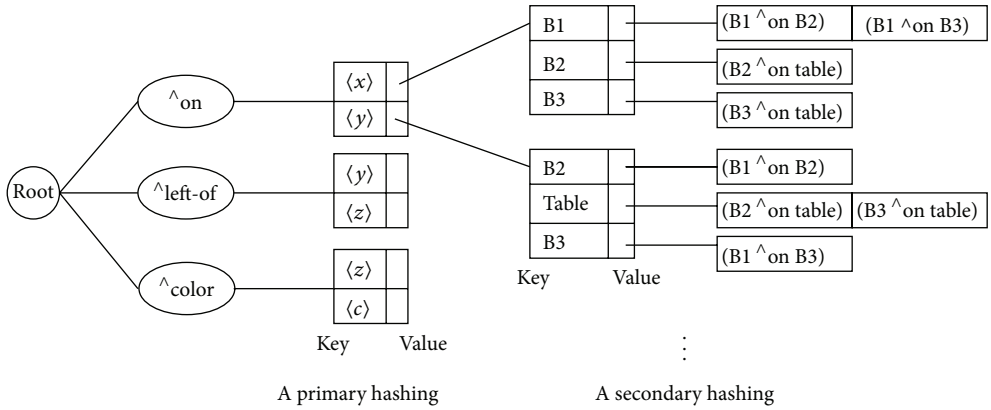


Figure 5: Alpha network of the RETE-ADH.

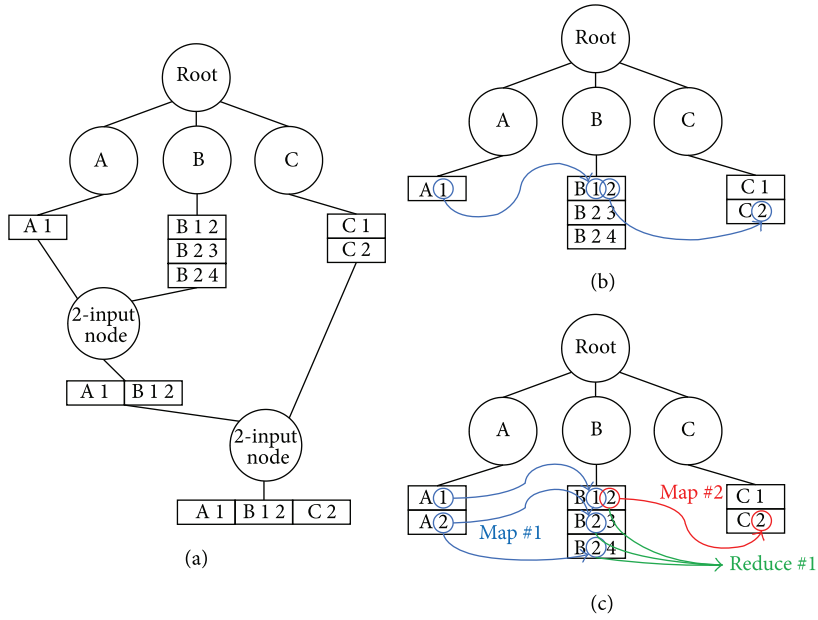


Figure 6: Comparison between RETE and RETE-ADH.

comparison operations by 60%, which roughly matches with the empirical results shown in Section 4.

Figure 6(c) shows another nice property of the proposed algorithm and its suitability for the parallelization. Assume that we use the same rule noted above. The tree in Figure 6(c) has one more node below *node* A compared to those in Figure 6(a) and Figure 6(b). Note that the *nodes* A, B, and C and each of the groups of nodes below them can be stored separately. For example, we can parallelize the algorithm by implementing the *map-reduce* computation. After receiving facts, we can execute a *map* phase in which the nodes that should be searched are selected. Then we have a blue group and a red group of *mappings*. After computations of these groups, we can execute a *reduce* phase, noted in green. Note that we can compute each of the grouped *mappings* concurrently. Although in the case of Figure 6(c), there is some overhead (relative to the sequential algorithm) because each of the computations in grouped *mappings* should be done thoroughly; we expect that for a large set of rules and facts, such a parallelized algorithm can easily outperform sequential ones. In addition, a parallel algorithm can delegate the effort required for the matching process to modern parallel processors such as CUDA GPUs, leaving the main CPU free to perform other tasks.

4. Test Bed: Virtual Simulator for a Smart Office Environment

In order to validate the proposed architecture and algorithm, we prepared a number of test cases targeting application in a smart office environment. Suppose that we are developing the smart office scenario depicted in Figure 7 with a virtual simulator. The imaginary smart office application automatically provides the most appropriate service for the employees by using the pattern matching algorithm. When an employee enters the office, the smart office application

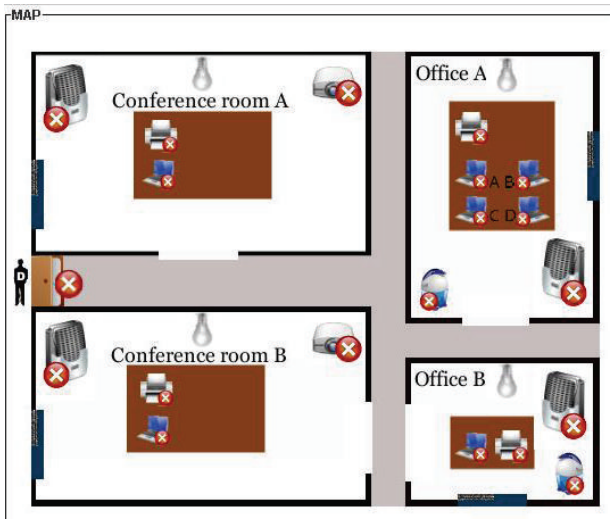


Figure 7: Composition of the smart office.

will automatically provide context-aware services such as turning on his/her computer, light, and printer. In addition, it maintains comfortable temperature, humidity, and illumination for the users by controlling air conditioners, humidifiers, and lamps, allowing employees to work in an optimal environment without bothering with environmental controls.

4.1. Overview. Let us assume that three employees enter office A and one employee enters office B. Before it can yield meaningful context information, the system will need to collect information as follows.

(i) Sensing: gathering context information from sensor devices.

(ii) Aggregating: observing, collecting, and composing context information from various context information processing units.

(iii) Inferring: interpreting context information to derive other types of context information, based on logic rules and the knowledge base, for instance.

(iv) Adopting: projecting context information of given situations.

One smart office scenario for a specific context adaptation case is demonstrated in Figures 8 and 9. After the user enters the smart office building, the user's device sends a request/subscription message to the composite context-aware system. Based on the device ID and user ID, the main system identifies the user and provides services accordingly.

Figure 8 shows a layout of the developed smart office application. Our smart office application mainly consists of three components: the MAP, DATA, and SYSTEM MESSAGE modules. An event controller is also developed, as shown in Figure 8. Each component is described as follows.

(i) MAP module: this module shows the office environment graphically. In this application, we assume that there are four places: conference room A, conference room B, office A, and office B. Also, this module depicts the state of each electronic device, door, and window, as well as each employee's current location.

(ii) DATA module: this module shows the office environment numerically; that is, through context information values. If a scenario occurs, this module shows the resulting changes in each context value.

(iii) SYSTEM MESSAGE module: this module shows the sequential control flow when a scenario operates. It chronologically shows the scenarios that have been occurring, the contexts that have changed, and the values of those contexts.

In this application, we can generate a virtual scenario by using the event controller. Table 2 shows the scenarios included in the smart office application. If we want to simulate any other scenarios in the smart office environment, new scenarios can be added using the event controller.

4.2. Structure. To validate the efficiency of the RETE-ADH algorithm in the composite context-aware system, we compared it with three other pattern matching algorithms in the virtual simulator. For each algorithm, the inference engine executed the algorithm and recorded the execution time for performance comparison. The *control package* integrated each algorithm. By using the *filemanage* function in the *control package*, each algorithm was identified and adopted. The related rules and facts were sent and received through this *filemanage function*. Each algorithm sent and received necessary parameters or call functions using the *AlphaNet.java*, *AlphaNode.java*, and *BetaNode.java* files.

Figure 10 shows a class diagram of RETE-ADH. When the RETE-ADH algorithm receives rules and facts, it differentiates the facts and records them to the alpha memory. After storing the facts, RETE-ADH composes the alpha network while considering their relationships. This process is performed by using the *NewRete*, *AlphaNet*, *AlphaSubNode*, and *AlphaNode* classes. The *AlphaSubNode* class is used only in the RETE-ADH algorithm, to support its secondary hashing. The *BetaNode* class is used for saving interim results. In the cases in which other algorithms are used, the basic structure and process is similar.

Figure 8 shows an empty smart office. In the *datamodule*, we can see the four employees' favorite temperatures, illumination levels, and humidity levels, as well as their respective locations. Note that, their locations are indicated as *Out* because none of them are present. Moreover, we can see that all devices are turned off. Now, we generate an event using the event controller: <*employee* A, B, C. *and* D *enter office* B,A, A, *and* A, *respectively*>.

In the virtual simulator, we follow ECA-DL; accordingly, this event can be expressed as person A goes to work at office B, person B goes to work at office A, person C goes to work at office A, and person D goes to work at office A. Figure 9 shows that all the employees have entered their offices. When the simulation begins with the start button on the event controller, the person icons move to offices

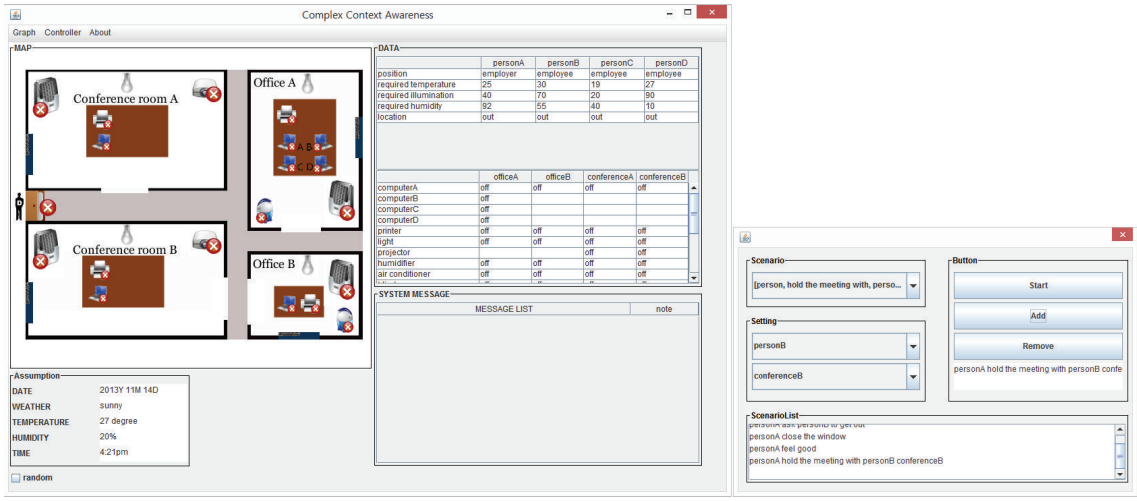


Figure 8: Initial status of the smart office.

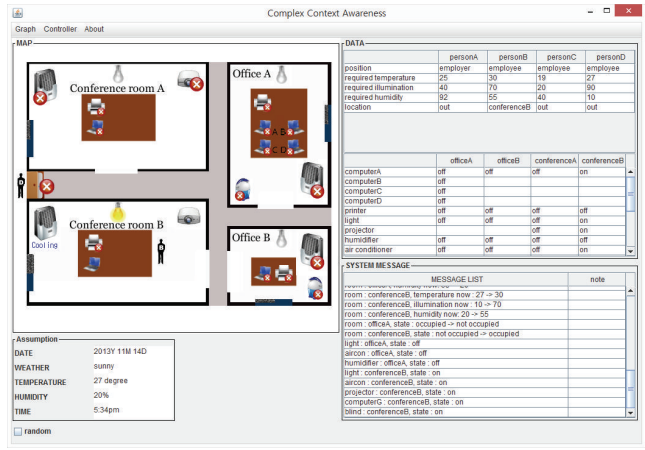


Figure 9: Changed status of the smart office.

A and B. Furthermore, the devices are automatically set to each employee's favorites. In Figure 9, we can observe the changes in the states of each person and electronic devices through the DATA, SYSTEM MESSAGE and MAP modules. Three employees are located in office A and one employee is located in office B. Moreover, personal computers, lights, air conditioners, and humidifiers are turned on in both rooms A and B. In the DATA module, the employees' locations are set up, and other parameters are changed accordingly. However, the humidity, temperature, and illumination are not set to each employee's favorites. This is because the three persons in office A have different favorites.

4.3. Empirical Result. In this section, we compare the proposed scheme with the RETE, TREAT, and LEAPS algorithms in the context of smart office virtual simulator. In the smart office scenario, our smart office application tries to obtain context information by conducting pattern matching between a set of rules and fact information. Then the application picks appropriate services according to the context and provides them to the users.

For this simulation, we randomly generated various scenarios and measured the processing speed of each pattern matching algorithm, which indicates the average processing time to find a successful match.

Figure 11 shows the pattern matching performance for each algorithm. As shown in the figure, the LEAPS algorithm provided the best processing performance in finding a single rule. As noted previously, this is because the LEAPS algorithm does not try to find the full matched set, and fires at most one rule per matching cycle. Our proposed algorithm outperformed TREAT and RETE. As mentioned before, TREAT and RETE spend much time in constructing the network, which makes the corresponding systems access memory frequently whenever the fact information is updated. One interesting finding is that the RETE algorithm had better performance than the TREAT algorithm. In the smart office application, the number of rules that can be fired is relatively limited. It can be deduced that the RETE and RETE-ADH algorithms effectively utilize their beta memories in the given environment. Note that we cannot adopt the LEAPS algorithm for our application, which needs a full set of matched rules

Table 2: Scenarios included in the smart office application.

Scenario rule	Description
Person, go to work at, room	Commutes to the office
Person, go home, room	Get off work
Person, hold a meeting with, person, at room	Hold a meeting (two people)
Person, hold a meeting with, person, person, at room	Hold a meeting (three people)
Person, hold a meeting with person, person, person, at room	Hold a meeting (four people)
Person, adjourn the meeting with, person	Adjourn a meeting (two people)
Person, adjourn the meeting with, person, person	Adjourn a meeting (three people)
Person, adjourn the meeting with, person, person, person	Adjourn a meeting (four people)
Person, call, person	Call person
Person, call, person, person	Call two persons
Person, call, person, person, person	Call three persons
Person, ask, person, to leave	Ask person to leave
Person, ask, person, person, to leave	Ask two persons to leave
Person, ask, person, person, person, to leave	Ask three persons to leave
Person, open the window	Open the window
Person, close the window	Close the window
Person, open the blinds	Open the blinds
Person, close the blinds	Close the blinds
Person, use printer	Use a printer
Person, get tired	Get tired (abstract scenario)
Person, feel bad	Feel bad (abstract scenario)
Person, feel good	Feel good (abstract scenario)
Person, happy	Happy (abstract scenario)
Person, unhappy	Unhappy (abstract scenario)

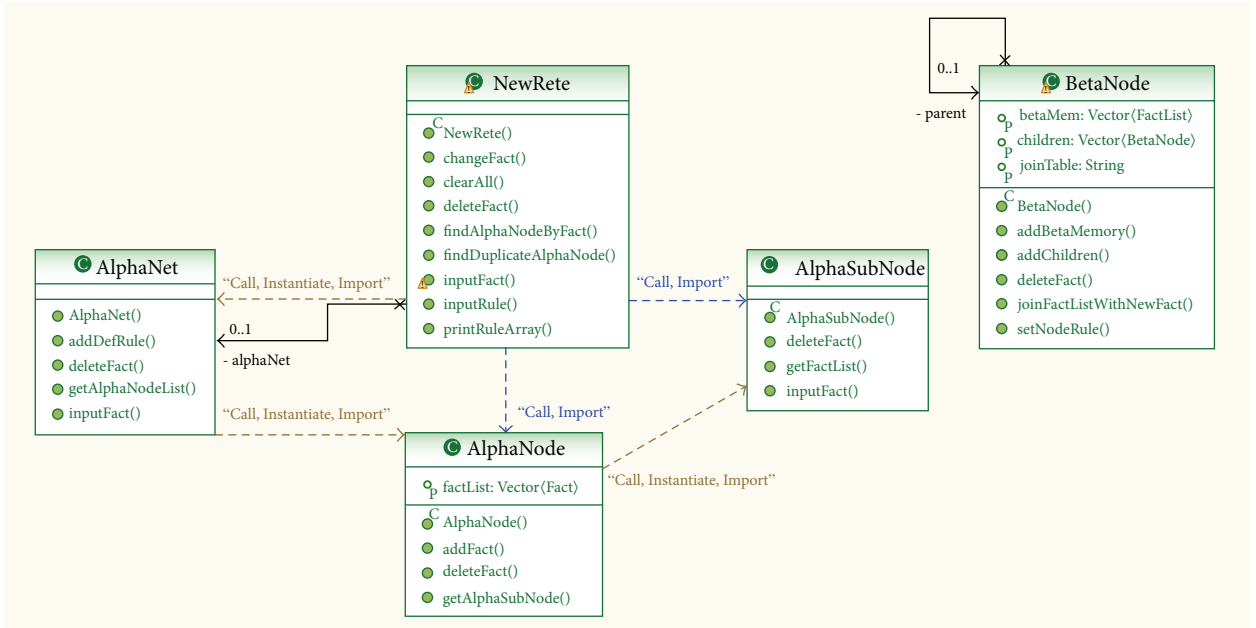


Figure 10: Class diagram of the RETE-ADH. (NewRete class represents RETE-ADH class).

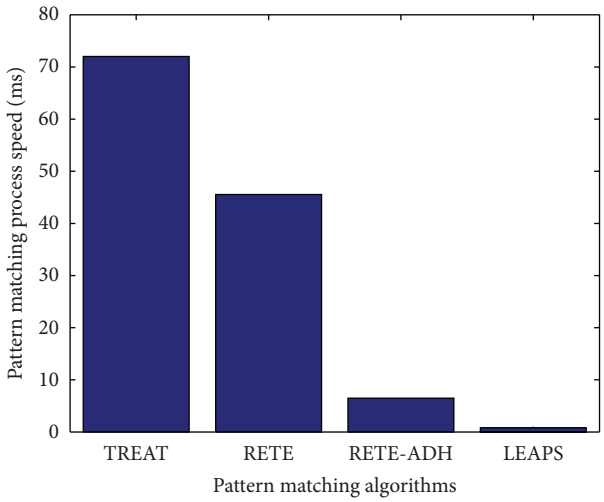


Figure 11: Experimental results.

to signify accurate context, and to enable the system to provide the most appropriate services. Rejecting the LEAPS algorithm leaves us with the options of TREAT, RETE, and our proposed algorithm. The experimental results suggest that our proposed algorithm is the best fit among these options for the targeted composite context-aware service.

5. Conclusions and Future Work

In this paper, we proposed a composite context-aware service architecture and a new pattern matching algorithm. In addition, we implemented a virtual simulator to validate our architecture and algorithm. The proposed algorithm provides enhanced matching performance by searching only a subset of the rules that can be matched. This improvement was made possible by the adoption of double hashing in the alpha network. We compared the proposed algorithm with the well-known pattern matching algorithms RETE, TREAT, and LEAPS by using our virtual simulator. The simulation results show that our proposed algorithm outperforms the TREAT and RETE algorithms. In addition, LEAPS was rejected due to its unique behavior of firing at most one rule per matching cycle, which is insufficient for context aware services. It was observed that the matching performance of the proposed algorithm was improved by 85% compared to that of RETE. We presented a practical scenario set in a smart office to show the applicability and validity of our composite context-aware service architecture.

In the future work, we will extend the proposed algorithm to exploit a parallel hardware architecture such as that of a CUDA GPU. In addition, we plan to carry out experiments using actual sensor nodes in various real-world scenarios.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by the Chung-Ang University Excellent Student Scholarship, the National Research Foundation (NRF-2011-0013924) and the MSIP (Ministry of Science, ICT and Future Planning) under the ITRC (Information Technology Research Center) support Program (NIPA-2013-H0301-14-1023) supervised by the NIPA.

References

[1] L. M. Daniele, P. D. Costa, and L. F. Pires, "Towards a rule-based approach for context-aware applications," in *Proceedings of the 13th Open European Summer School and IFIP TC6.6 Workshop on Dependable and Adaptable Networks and Services*, Lecture Notes in Computer Science, Springer, Enschede, The Netherlands, 2007.

[2] A. Zaslavsky, "Mobile agents: can they assist with context awareness?" in *Proceedings of the IEEE International Conference on Mobile Data Management (MDM '04)*, pp. 304–305, January 2004.

[3] C. L. Forgy, "Rete: a fast algorithm for the many pattern/many object pattern match problem," *Artificial Intelligence*, vol. 19, no. 1, pp. 17–37, 1982.

[4] M. Matsushita, M. Umano, I. Hatono, and H. Tamura, "A fast pattern-matching algorithm using matching candidates for production systems," in *Proceedings of the 4th Pacific Rim International Conference on Artificial Intelligence*, Lecture Notes in Computer Science, Springer, Cairns, Australia, 1996.

[5] D. P. Miranker, "TREAT: a better match algorithm for AI production systems," in *Proceedings of the 6th National Conference on Artificial Intelligence*, pp. 42–47, 1987.

[6] D. P. Miranker and B. J. Lofaso, "The organization and performance of a TREAT-based production system compiler," *IEEE Transactions on Knowledge and Data Engineering*, vol. 3, no. 1, pp. 3–10, 1991.

[7] D. Batory, "The LEAPS algorithms," Tech. Rep., University of Texas at Austin, Austin, Tex, USA, 1994.

[8] D. Liu, T. Gu, and J.-P. Xue, "Rule engine based on improvement rete algorithm," in *Proceedings of the International Conference on Apperceiving Computing and Intelligence Analysis (ICACIA '10)*, pp. 346–349, Chengdu, China, December 2010.

[9] D. Xiao and X. Zhong, "Improving rete algorithm to enhance performance of rule engine systems," in *Proceedings of the International Conference on Computer Design and Applications (ICCDA '10)*, pp. V3572–V3575, Qinhuangdao, China, June 2010.

[10] D. Zhou, Y. Fu, S. Zhong, and R. Zhao, "The Rete algorithm improvement and implementation," in *Proceedings of the International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII '08)*, pp. 426–429, December 2008.

[11] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles, "Towards a better understanding of context and context-awareness," in *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing*, Lecture Notes in Computer Science, Springer, London, UK, 1999.

[12] W. Na, S. Cho, E. Kim, and Y. Choi, "Event detection in composite context aware-service," in *Proceedings of the 3rd International Conference on Ubiquitous and Future Networks (ICUFN '11)*, pp. 342–345, Dalian, China, June 2011.

[13] J. McDermott, A. Newell, and J. Moore, "The efficiency of certain production system implementations," *ACM SIGART Bulletin*, no. 63, pp. 38–38, 1977.

[14] K. Oflazer, "Highly parallel execution of production systems: a model, algorithms and architecture," *New Generation Computing*, vol. 10, no. 3, pp. 287–313, 1992.

[15] D. A. Brant, T. Grose, B. Lofaso, and D. P. Miranker, "Effects of database size on rule system performance: five case studies," in *Proceedings of the 17th International Conference on Very Large Data Bases*, pp. 287–296, San Francisco, Calif, USA, 1991.

[16] J. Yoon and K. Chung, "An efficient pattern matching alogorithm for AI production system," in *Proceedings of the Korean Institute of Information Scientists and Engineers (KIISE '94)*, 1994.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 920534, 11 pages
http://dx.doi.org/10.1155/2014/920534

Research Article

An Obstacle Avoidance Scheme Maintaining Connectivity for Micro-Unmanned Aerial Vehicles

Hyo Hyun Choi,[1] HyunSoo Choi,[2] Myungwhan Choi,[2] Taeshik Shon,[3] and ByoungSeob Park[4]

[1] *Department of Computer Science, Inha Technical College, Incheon 402-752, Republic of Korea*
[2] *Department of Computer Science and Engineering, Sogang University, Seoul 121-742, Republic of Korea*
[3] *Division of Information and Computer Engineering, Ajou University, Suwon, Gyeonggi 443-749, Republic of Korea*
[4] *Department of Computer Systems & Engineering, Inha Technical College, Incheon 402-752, Republic of Korea*

Correspondence should be addressed to ByoungSeob Park; bspark@inhatc.ac.kr

Received 27 November 2013; Accepted 25 January 2014; Published 24 April 2014

Academic Editor: Jongsung Kim

This paper suggests an obstacle avoidance scheme that enables a group of micro-UAV (Unmanned Aerial Vehicles) to avoid colliding with an obstacle that is found in the course of the flight. This scheme considers a method for UAVs to avoid colliding with such obstacles and a plan of action for when the UAVs in the group lose connectivity during flight. The main goal is for UAVs to reach the target without colliding with any obstacles. To achieve this goal, directly after the UAVs avoid the obstacle at a raised altitude above it, they fly at a higher altitude for a while and then descend to their original height. When this approach is judged to be inefficient because the height of the obstacle is too high, the UAVs divide into two groups and move to either side of it while continuing to fly at the same altitude, thereby avoiding the obstacle. Afterwards, they gather into one group again. We verify the proposed scheme and do the performance evaluation by ns-2.

1. Introduction

UAVs, which have been studied for the military purpose of detecting enemies, act as spies and can be used to communicate between friendly forces; they are expected to be used in various places and for many purposes; for example, searching for patients under circumstances in which a communication infrastructure has been destroyed or monitoring the scene of a disaster. The usefulness of UAVs is greater when several UAVs that maintain communications with each other are used compared with when only one UAV is used. In this paper, we assume that UAV is so small that the people can carry it with their hands or the general car can carry several ones. We define it as micro-UAV and we use the terms UAV with the same meaning as the micro-UAV in this paper.

When several micro-UAVs are used, we can obtain the following effects. First, when UAVs process detections, they can make the boundary of the detection wider and the time of the detection shorter. Second, when several UAVs form a communication network with one another, a rapid decision on a situation can be made because the information on the target is sent to the base station through communication between the UAVs without any communication infrastructure.

To use these advantages, this paper suggests an effective scheme in which several UAVs maintain communication connections with each other in a group, reach a target, and escape rapidly when they find a fixed obstacle. This capability can be used, for example, while investigating the scene of a disaster. We assume that the communication among UAVs can be done through short range wireless technology such as Wi-Fi or ZigBee. It will be useful that the proposed scheme is applied to the disaster environment or the wartime.

The remainder of this paper is arranged as follows. In Section 2, we describe the related works and show the differences from our work. In Section 3, we describe the process of finding obstacles, the method for transmitting the information on found obstacles, and the method of

UAVs' moving and splitting. The experiments are described and the results are discussed in Section 4. In Section 5, the conclusions are drawn, and finally, we discuss the questions that are considered in this study.

2. Related Work

There have been studies on UAV formations, on UAVs' communication with each other, and on designs for finding obstacles with various sensors. The obstacle avoidance scheme in this paper has two methods: planning and reaction. Planning is used to plan the flight course of the UAVs, and Reaction is used to avoid an obstacle while UAVs are flying. Some examples of studies of a whole group of UAVs are mentioned below.

By means of the algorithm in [1], which ensures there is no collision, the flight course is fixed; otherwise, the UAVs would collide with each other during the flight. Richards and How [2] show a strategy with which UAVs will avoid colliding not only with obstacles but also with each other when they are flying in a group. Although the scenario in the studies of [1, 2], which involves swarms of UAVs that avoid colliding with obstacles that they detect, is similar to the scenarios addressed in this paper, these studies do not consider the communication between UAVs and they are conducted in a 2D environment. Scherer et al. [3] provide a scheme in which a micro-unmanned helicopter avoids fixed obstacles such as buildings or trees when it has a map of an area in advance. Similar to the studies in [1–3], which use the planning method of obstacle avoidance, the UAVs' operators already had information about the obstacles' positions, and they planned a flight course to avoid colliding with these obstacles. Therefore, these studies are different from the present paper, which suggests an immediate reaction to obstacles and an obstacle avoidance flight method.

This study is preceded by a study in which micro-unmanned flights installed with various sensors perceive obstacles and avoid them. Kwag and Chung [4] is a study in which UAVs find obstacles and avoid them by using a radar sensor. Watanabe et al. [5] uses the strategy that micro-unmanned flights equipped with 2D cameras detect obstacles and avoid them. Both papers focus more on the scheme that processes the information obtained using sensors, and they studied the plans of the flight courses in accordance with that information. Additionally, the communication between UAVs is not considered in these works, which makes these references different from the scheme that this paper proposes. Bethke et al. [6] studies a scheme that contends with a breakdown, such as running out of fuel while several UAVs are flying in a group. Although [6] also utilizes several UAVs, it focuses on methods that address an accident that occurs in the middle of a flight. Thus, its theme is different from this paper. Xu et al. [7] is a study that secures the communication connection between manned vehicles and UAVs; it proposes how to secure the communication connections between several manned vehicles and a UAV and several manned vehicles and several UAVs. Though the study in [7] may be similar to the scheme that we suggest, and it focuses on maintaining the connection between vehicles and UAVs while they are moving. However, it does not consider how to address the situation in case UAVs encounter obstacles while they are moving.

3. Proposed Method

This paper considers a scenario in which several micro-unmanned aerial vehicles that compose a group search for a site whose communication facilities were destroyed by war or disaster. It is our goal that when these UAVs discover obstacles in the middle of the searching process where there is a limited area for communicating with each other in a group, they return to the former formation after avoiding the obstacles; note that the avoidance can cause the disconnection of connection between the UAVs. The scheme that this paper proposes is reactive: when UAVs discover obstacles during their flight, they avoid the obstacles as a reaction to them.

This paper proposes two types of avoidance as follows.

(1) One type of avoidance is the scheme that UAVs in a group climb above the obstacle and avoid it as quickly as possible while their intercommunication is maintained.

(2) The other type of avoidance is the scheme that after UAVs are divided into two parts and avoid the obstacle by moving to the two sides of it, they gather together at an appointed location and restore the communication connection between themselves. However, it is possible that their intercommunication is not maintained.

The UAVs assumed in this paper are Quadcopter types whose free flight is possible, and the fuel that the UAVs use during the flight is enough, and the effects that the obstacle has on the communications between the UAVs are not considered. Each UAV has its own index, and it can prevent the UAVs in a group from colliding with each other. And the communication range between two UAVs is 150 m in the present study. UAV has an attached device to detect an obstacle such as camera or ultrasonic sensor. Actually, we assume that UAV has the camera in this present study.

UAVs that form a group begin to search at the maximum velocity at the same searching altitude. The search is undertaken along the Y-axis, and the X-axis remains constant. The obstacles arranged in this paper have their own indexes and are considered to be in the shape of a cube, which is similar to buildings in a city. Additionally, the characteristics of these obstacles can be reflected in a simulated experiment. All of the trajectories of discovering and avoiding obstacles are recorded in the computer simulation. And an obstacle is assumed as the shape of a rectangular. The example of scenario before the obstacle is discovered and is shown in Figure 1.

3.1. The Message Type and How to Send a Message. In this paper, the information sent in the form of a packet (i.e., the message) can be divided into two types. First, a beacon

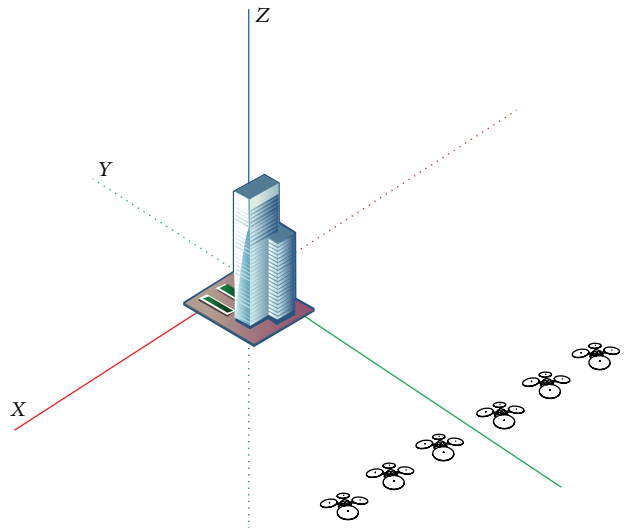


Figure 1: The Initial State of the UAVs' formation.

Table 1: The Messages.

Type of the Message	Role of the Message
Beacon Message	Send the UAV's current location and identify a neighboring UAV's location
Forwarding Message	Check and send the corresponding value in a packet at the time of the event among the UAVs

message is periodically transmitted to other UAVs to identify the neighbors (broadcast) (see Table 1 and Algorithm 1). Second, a forwarding message is used to indicate the discovery of obstacles and exchange the information. Additionally, an acknowledgement message (ACK) can be sent to check the success of forward message transmission. A forwarding message is different from a beacon message in that a forwarding message is generated and transmitted when an event occurs (see Algorithm 2).

3.2. Discovering Obstacles and Selecting and Canceling a Leader. By transmitting and receiving a beacon message periodically while the UAVs in a group are flying, each UAV recognizes the index of the neighboring UAV and its position. All of the UAVs in the group move while making a 180° observation of the direction of movement, and whenever the UAVs send a beacon message to each other, they move to the destination and confirm the existence and nonexistence of obstacles. When the UAVs and an obstacle are at a certain distance while the UAVs are moving, they discover the existence of the obstacle. The UAVs continue to move to the destination even at the time when the obstacle is discovered and before a leader has been selected.

There are three cases when a UAV discovers an obstacle as follows.

(1) The case in which a UAV in the group discovers one obstacle.

(2) The case in which each UAV in the group discovers the same obstacle and sends the information.

(3) The case in which each UAV in the group sends information on different obstacles.

To avoid the obstacle or obstacles under conditions such as the above, the UAV that discovered the obstacle continues to inform the other UAVs in the group that it will be their leader until it receives their answers, and the task of the UAV that discovered the obstacle is as follows.

(1) Maintain the information on the place and shape of the discovered obstacle.

(2) Decide on a suitable obstacle avoidance scheme by employing the available data.

(3) Calculate the avoidance point and the moving speed of the neighboring UAV and send the information to it.

The avoidance point is defined as a location that the UAV should fly towards to avoid an obstacle. A leader UAV calculates the avoidance point and sends it to neighbor UAVs, then neighbor UAVs calculates their avoidance point to fly. For example, in this present study, the avoidance point of the leader is calculated simply by adding 10 m to the height of the obstacle in the case that UAVs fly higher to avoid an obstacle. It is explained more latter in this paper.

The following is the process in which the UAV that discovered an obstacle while flying is selected to be the leader.

(1) The UAV that discovered an obstacle takes the information on the UAV's position and the width and height of the obstacle and performs a calculation. This calculation is made to choose which one of the two avoidance schemes that this paper proposes will be used.

(2) After choosing an avoidance scheme, the UAV calculates the position and the speed of the other UAVs that are flying at either side of it with the avoidance algorithm that this paper proposes.

(3) After recording its identification number on the calculated value, the avoidance scheme and the leader's identification number show who the leader is. And the UAV sends this information to the UAVs on either side of it.

(4) When all of the UAVs at both ends of the group receive the information smoothly, they send their confirmation messages to the UAV that discovered the obstacle. When the UAV that discovered the obstacle receives the confirmation messages from the other UAVs in the group, it becomes the leader of the group.

When the leader is selected, all of the UAVs in the group begin to move to avoid the obstacle, and they retain the information on the leader and the obstacle until they each reach their avoidance points. When the UAVs in the group reach their avoidance points, they remove the information on

Beacon Message: The type of beacon message in this paper

The UAVs encapsulate their IDs and (x, y, z) coordinates in a packet
The UAVs periodically broadcast the Beacon Message
A UAV receives a Beacon Message
This UAV can identify its nearest neighbor and its location.

Algorithm 1: The Algorithm of the beacon message.

LN: ID of the Leader UAV
Oposx, Oposy: Coordinate of the obstacle

Leader UAV
(1) Upon discovering the obstacle,
(2) Transmit LN, Oposx, Oposy, Altitude and Velocity to the neighboring UAVs
(3) When it receives a message that confirms information reception, It transmits the command for movement

Other UAVs
(1) Receive LN, Oposx, Oposy, Altitude and Velocity information.
(2) If self's Altitude, and Velocity value is NULL, Save Altitude and Velocity
(3) If self's Altitude and Velocity values are not NULL
Compare the stored and the new Altitude and Velocity
(4) Update LN, Oposx, Oposy, Altitude and Velocity values for the data with a higher Altitude
(5) Calculate the Altitude and Velocity for the neighboring UAVs
(6) Transmit the information to neighboring UAVs
(7) If the UAV does not have neighbors to its left or right, Transmit a message that confirms the information's reception.

Algorithm 2: The Algorithm for the forward message.

the leader and the obstacle, and they manage the conditions in the same way if they discover new obstacles. If any UAV in the group discovers a new obstacle in the midst of moving to the avoidance point, this UAV does not inform the other UAVs until it reaches the avoidance point; then, this UAV informs them that it discovered a new obstacle after removing the information on the leader and the obstacle. The removal of the information on the selected leader and the obstacle are performed differently according to the scheme that this paper proposes. In the case of avoidance at a higher altitude, the UAVs keep the information until they pass the obstacle completely. This strategy is intended to prevent the UAVs from colliding with the obstacle in case they perform the avoidance at a lower altitude and then discover a new obstacle that is shorter than the former one. Additionally, if the UAVs avoid dividing their group into two parts, as soon as they reach the avoidance points they initialize the information on both the obstacle and the leader. Afterward, they contend with the next obstacle. Figure 2 shows the abstract scheme of how to decide between two methods. α means the weight value of the possibility of the communication disconnection. In SPLIT method, dividing their group into two parts, there is the possibility that UAVs cannot maintain the communication after avoiding an obstacle.

3.3. The JUMP Method. The JUMP Method is designed to avoid an obstacle by climbing to a higher altitude that is at least as high as the obstacle's height, and this method ensures that the communication connections continue between the UAVs. This approach involves the action that the UAVs in the group avoid an obstacle by rapidly moving as minimally as possible when they perform the avoidance, and they perform three types of action, which are the following (see Algorithm 3).

(1) To avoid an obstacle, they climb to a higher altitude and move to an avoidance point.

(2) When UAVs that have reached the avoidance point find a new obstacle at the moving speed, they maintain the current obstacle information before avoiding the obstacle.

(3) The UAVs use the information on new obstacle that they had when avoiding the obstacle.

First, in the process of climbing to a higher altitude and moving to the avoidance point to avoid the obstacle, the action algorithm that the leader UAV performs is similar to the following.

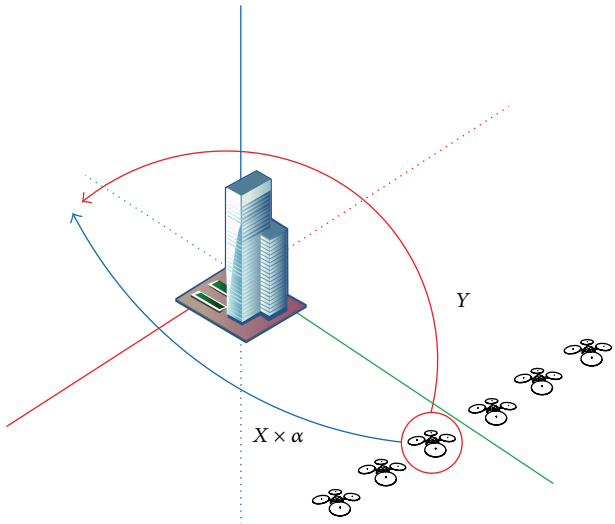
The UAV that discovered the obstacle decides the location at which it will be situated after avoiding the obstacle while considering a safety margin distance of $\sqrt{D_0}$, to have more altitude than the obstacle's height. After deciding its avoidance point, this UAV should decide the positions of the UAVs on both sides of it. The extent of the communication that is possible between UAVs is 200 m, and the UAVs are 150 m away from the X-axis; thus, by making use of this arrangement, the leader UAV plays a role in deciding the

ON: ID of the obstacle
Oposx, Oposy, Oheight: Coordinate of the obstacle and height

Leader UAV
(1) Upon discovering the obstacle,
(2) Get ON, Oposx, Oposy and Oheight
(3) Calculate own Altitude and Velocity
(4) Calculate the Altitude and Velocity by formulas (1) and (2)
(5) Encapsulate the Altitude and Velocity information into a packet
(6) Transmit the packets to neighboring UAVs

Other UAVs
(1) Receive the Altitude and Velocity information.
(2) If self's Altitude and Velocity value is NULL, Save Altitude and Velocity
(3) If self's Altitude and Velocity value is not NULL
Compare the stored and the new Altitude and Velocity
Calculate the Altitude and Velocity by formulas (1) and (2) for the neighboring UAVs
(4) Transmit the information to the neighboring UAVs
(5) If the UAV does not have neighbors to its left or right, Transmit a message confirming the information's reception.

Algorithm 3: The Algorithm of the JUMP method.



X: Leader's distance of SPLIT method
Y: Leader's distance of JUMP method
α: Weight of connection loss

Figure 2: Deciding the avoidance method.

UAVs' positions to make their flight distance as short as possible from the neighboring UAVs as they avoid colliding with the obstacle.

The calculation of the location of the UAVs on both sides of the leader follows formula (1) as follows:

$$\Big(\big(\text{avoidance point of the leader UAV} - \text{avoidance points of the neighboring UAVs}\big)^2\Big)^{1/2} \le D_0. \quad (1)$$

If the neighboring UAVs on both sides move using the value calculated through formula (1), when the process of avoiding the obstacle is completed, the flight distance of the UAVs on both sides becomes as minimal as possible, which ensures continuous communication between the leader UAV and the other UAVs. Although formula (1) ensures the communication connection between the UAVs in the group as much as possible, in the midst of moving for avoidance, it is difficult to ensure the communication connections between the UAVs. The reason is that their traveling distance is relatively short compared with the leader's distance, but if they move at the same speed as the leader UAV, then the distance between the leader UAV and the neighboring UAVs on both sides of it become farther compared with the time at which they start to move.

The leader decides the speeds of the neighboring UAVs on both sides of it to ensure the communication connection between them in the course of moving. This decision can prevent the communication connection from being cut, which can occur during the course of avoidance if the leader and the neighboring UAVs reach the avoidance spot at the same time.

The method of deciding the speed of the neighboring UAVs is satisfied with the following formula (θ is an angle between the avoidance point and the current position of UAV):

$$\text{The speed of the neighboring UAVs} = \frac{\sin\theta_{\text{leader UAV}}}{\sin\theta_{\text{UAVs at both sides}}} \times \text{Velocity}_{\text{leader UAV}}. \quad (2)$$

Because the leader moves the farthest, it rises to a higher altitude at the maximum speed and avoids the obstacle. As the leader calculates what the positions of the UAVs would be on both of its sides through formula (1), it decides its position, the neighboring UAVs' positions, and the speed at which the neighboring UAVs will move using its speed and formula (2).

Finally, the leader transmits the positions and the speeds of the neighboring UAVs, as decided through formula (1) and formula (2), using the Unicast method. The action algorithm of the UAVs that are not the leader is similar to the following: because the UAVs that are not the leader did not discover the obstacle, they move to the target spot and continually check whether there is an obstacle or not until the leader is decided. When the other UAVs receive the information about their avoidance points and moving speeds, in a process that is similar to what the leader did, they decide the avoidance points and moving speeds of their neighboring UAVs.

The UAVs at both sides of the leader each calculate the positions of the UAVs beside them. They can each decide the positions of the UAVs beside them by using formula (1); at this time, instead of the leader's avoidance point of formula (1), they use the heights that they received from the leader to determine where they will be situated.

The received moving speed decides the neighboring UAVs' speeds by using formula (2), and the process is repeated until the last UAVs in the group receive their moving speeds. If all of the UAVs in the group receive their moving positions and speeds and perform their avoidance schemes until the moment that the process of avoiding the obstacle and reaching the destination begins, it is assured that all of the UAVs in the group avoid the obstacle and maintain the communication connection.

The UAVs of the group that arrived at the avoidance point continue moving at the same speed as before the avoidance actions and maintain the same (high) altitude. When they move to the end of the obstacle, each of the UAVs initializes the information about the leader and the obstacle that it has avoided. Lastly, the UAVs restore their formation to what it was before the avoidance. All of the UAVs descend to the avoidance points at the same speed, such as the distance of the Y-axis to which they climbed to avoid the obstacle; then, they restore the formation that was present before the obstacle avoidance action. If there is a new obstacle to avoid in the course of restoring the group's shape to its formation before the obstacle avoidance action and at the place this obstacle was discovered, the UAVs in the group perform another avoidance action through the same method after choosing the leader in the group and deciding on an avoidance scheme. The path that each UAV in the group takes to avoid the obstacle and reach the destination is similar to Figure 3.

3.4. The SPLIT Method. The SPLIT method can retain the original group's shape and ensure a continuous mutual communication connection because the distance between the neighboring UAVs that are moving in the same direction as the leader is the same as the distance to the leader. The other group moving in the different direction from that of the group containing the leader maintains the communication connection between the UAVs in that group but the communication connection with the leader may be broken.

The SPLIT method performs three types of movement, which are described as follows (see Figure 4 and Algorithm 4).

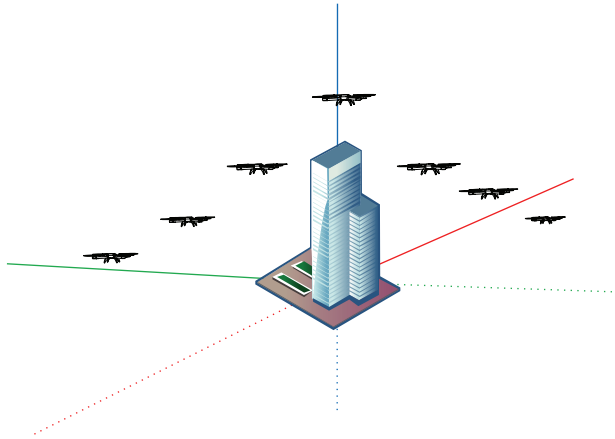


Figure 3: An example of the JUMP method.

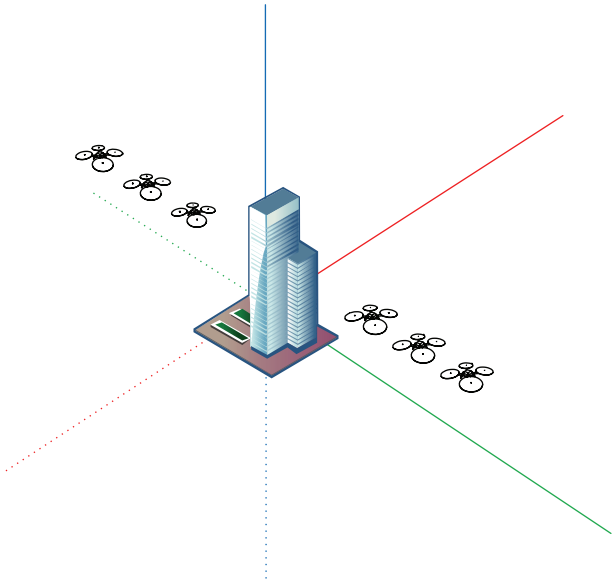


Figure 4: An example of the SPLIT method.

(1) The UAVs move to the appointed avoidance point, at which they divide into two groups.

(2) UAVs that have reached the avoidance point move to avoid the obstacle, but the altitude and moving speed are retained.

(3) The UAVs move to restore the former formation that was in place before avoiding the obstacle.

First, when the UAVs move to the avoidance point and are divided into two groups to avoid an obstacle, the role of the leader UAV is as follows: the leader decides the shorter direction to avoid the obstacle from the two possible directions, and the leader considers the safety margin for preventing a collision with the ends of the left side and the right side of the obstacle. The leader that decided the avoidance direction lets the neighboring UAVs on both sides of it know their avoidance points. The avoidance points of the neighboring UAVs moving in the same direction as the

ON: ID of the obstacle
Oposx, Oposy, Oheight: Coordinate of obstacle and height
myX, myY, myZ: UAV's coordinate
UAV_Interval: Interval of each UAVs

Leader UAV
(1) Upon discovering the obstacle,
(2) Get ON, Oposx, Oposy and Oheight
(3) Calculate own coordinates for avoidance
(4) Calculate neighbor UAV's coordinates By myX + UAV_Interval
(5) Encapsulate the coordinate information into a packet
(6) Transmit the packets to neighboring UAVs

Other UAVs
(1) Receive coordinate information.
(2) If coordinate information value is NULL, Save the coordinates
(3) If self's coordinate value is not NULL
    Compare the stored and the new coordinates
    Calculate a new coordinate by myX + UAV_Interval for neighboring UAVs
    Transmit information to neighboring UAVs
(4) If the UAV does not have neighbors to its left or right, Transmit a message confirming the information's reception.

Algorithm 4: The Algorithm of the SPLIT method.

leader are decided, including the distance between them to prevent them from colliding with each other at the leader's avoidance point. Because the leader knows the information about the positions of the neighboring UAVs, which was obtained through a beacon-message, the avoidance points of the neighboring UAVs are in the opposite direction from the direction the leader decided upon, which adds to the distance that the leader moves relative to the positions of the other UAVs and prevents the leader from going in the opposite direction. As a result, because the UAVs all travel the same distance, they can restore their original formation after avoiding the obstacle. If the information about the end of the left side of the obstacle is updated by the leader, then the leader informs the others of the value that was decided after considering the safety margin in the information with respect to the end of the left side of the obstacle.

The action algorithm of the UAVs that did not discover the obstacle is the following: the UAVs on both sides that received the information about the place to travel to from the UAV that discovered the obstacle inform the next neighboring UAVs about the avoidance points. To do so, they use the same method used by the UAV that discovered the obstacle, which is based on their positions. When all of the UAVs know the information about the positions that they will avoid, then they perform their avoidance at the appointed positions. The group of UAVs that has reached the avoidance points moves while observing the obstacle at the same speed as before; thus, they avoid the obstacle but maintain the same altitude. When they move to the end of the obstacle after observing it, each UAV initializes the information about the leader and the obstacle that it has. Lastly, before beginning the avoidance scheme, the UAVs return to the positions that they memorized, and they restore the formation that the group had originally had. In case the distances of the two groups that moved are the same when they performed these actions, they can restore the original formation and fly to the destination again. However, if the distances that they traveled are very different, then the condition occurs that one group goes before the other, and at worst, they reach the destination with the communication connection between the two groups having been cut.

If a new obstacle is discovered at the designated meeting place, for the UAVs to return to an original formation after avoidance, each UAV performs the same course from the action of discovering the obstacle. If the communication connection between the two groups has been cut, then the leaderless UAVs elect a new leader for the divided group and avoid the additionally observed obstacle.

Furthermore, the UAVs that have avoided the obstacle compare their positions with the positions stored when they started from the base station. At this time, if their positions are different from their original positions, to remain in the original formation, they move to the original destination by moving to the X-axis position that was stored at the beginning. The proposed method to avoid the obstacle by dividing into groups is performed when site is not efficient to adjust the altitude to avoid the obstacle because the obstacle is too tall. Additionally, the communication connection between UAVs in the group could be cut, especially when the group is divided. Thus, this method does not assume that the optimized action coincides with the core goal that the UAVs in the group avoid the obstacle promptly and maintain the communication connection with each other.

4. Experiments

4.1. Simulation Model. To evaluate the proposed method, an experiment was performed by using NS-2 (Network Simulator v.2.34) of the simulator on a Linux. We modify NS-2 objects to enable them to decide the position by themselves.

Table 2: The configuration and parameters of the simulation.

Configuration and parameters	Value
Channel type	Wireless channel
Network interface type	Wireless physical
MAC type	802.11
Link layer type	LL
Antenna model	Omni antenna
Distance of UAVs	150 m
Detection Range	300 m
V_{max}	10 m/s
Altitude	50 m
α	1, 1.05, 1.1, 1.15, 1.2
Width	100, 60, 20

Table 3: Information on an obstacle's position.

Index of obstacles	X, Y, Z	Width (m)
1	700, 800, 100	100
2	1050, 1500, 200	100
3	1500, 2100, 150	100

And we collect the position of objects periodically and then draw the graph as a 3-demension because original NS-2 simulation does not support 3-dimension in its animation. In the simulation, 8 UAVs numbered from 0 to 7 move at 10 m/s at a height of 50 meters over a layout of land whose size is 3000 m × 3000 m. We perform the simulation to verify the proposed scheme and to do performance evaluation. We investigate to what extent the value of α has an effect on choosing an avoidance method, the height of the obstacle to avoid, and the time required to avoid the obstacle through the two methods. The specific configuration and parameters of the simulation follow Table 2, and the information on the arrangement of the obstacle follows Table 3. We arranged the obstacles as shown in Table 3 and observed the result of the movement in the first simulation. Then, we experimented with changing the value of α, the width in Table 2, and the value of Z of the first obstacle in Table 3.

4.2. Analysis of the Result of the Simulation. Figure 5 shows the result of the simulation of the SPLIT Method using NS-2. Through the result, we can observe a scenario in which the proposed method works smoothly. The UAVs restore their communication connection as they restore their prior formation from before the avoidance action. The restoration occurs while the UAVs are moving to the rendezvous point that was appointed in advance, and even though the communication connection between them is cut during the course of the avoidance action, they can avoid this problem when they discover another obstacle. Figure 6 extends the simulator's actions beyond two dimensions, as motion in three dimensions is appropriate and shows the result as a trajectory. The X-axis of the graph shows the positions in which the UAVs were arranged, the Y-axis is the direction to move for detecting obstacles, and the Z-axis shows the altitude of the flight of the UAVs and the height of the fixed obstacles.

UAV number 1 discovers a fixed obstacle and avoids it by rising to an altitude of 110 m, and at this time, the UAVs on both sides of UAV number 1 keep moving. At this time, only the speed is controlled because the UAVs can preserve their communication connection despite maintaining only the altitude of their flight. We can see that UAV number 1 avoids the first obstacle and controls the altitude to restore the original formation. Afterwards, when UAV number 3 discovers a second obstacle, if number 3 avoids it alone, it is impossible to sustain the communication connection with the other UAVs; as a result, the UAVs on both sides of number 3 also adjust their altitude and avoid the obstacle with number 3. Figure 7 is the resulting graph of the experiment, which indicates how high the obstacle avoided by the JUMP method is and what effect the change of the obstacle's width has when the value of α is applied. The X-axis of the graph shows the change of the value α, and the Y-axis shows the height that the UAVs should pass over to avoid the obstacle. A larger value of α implies a larger width of the obstacle, and thus, the height that is required for the UAVs to pass over and avoid the obstacle becomes higher. When the UAVs perform avoidance by dividing into two groups, the total distance that the leader moves is multiplied by as much as the value of α. Therefore, the larger the value of α becomes, the higher the height is for the UAVs to pass over and avoid the obstacle by performing the JUMP method. Additionally, a shorter width of the obstacle implies that there is a stronger tendency for the UAVs to divide into groups; thus, the larger the width is, the higher the height is to perform the JUMP method. Furthermore, as the value of α becomes larger, the width of the obstacle tends to become smaller. When the value of α is not considered because the moving distance is immediately affected by the width of the obstacle, the height for the avoidance action changes. However, when the width of the obstacle is small, the distance that the leader moves to avoid the obstacle is not long and the value of α is also small. Therefore, though the value of α becomes larger, the height for the avoidance makes little difference.

Next, we observed the operating time through two methods. Figure 8 is the graph of the time taken to reach the avoidance point through the two methods when the value of α is 1 and the height of the obstacle becomes higher. In addition, Figure 9 is the graph when the value of α is changed to 1.2. The X-axis of the graph shows the height of the obstacle, and the Y-axis shows the time taken to reach the avoidance point. Figure 8 is the graph of the case in which the penalty of the loss of the communication connection was not considered. Furthermore, because a higher height of the obstacle implies a longer flight time even though the height of the obstacle becomes only slightly higher, the UAVs show the pattern of avoiding the obstacle using the SPLIT method. Afterward, they move to the destination. Figure 9 shows that when the communication connection is maintained, the UAVs also avoid an extremely high obstacle compared with Figure 8. The two graphs show that the UAVs avoid this obstacle without any change in the avoidance time, though the height of the obstacle becomes higher when it is over a

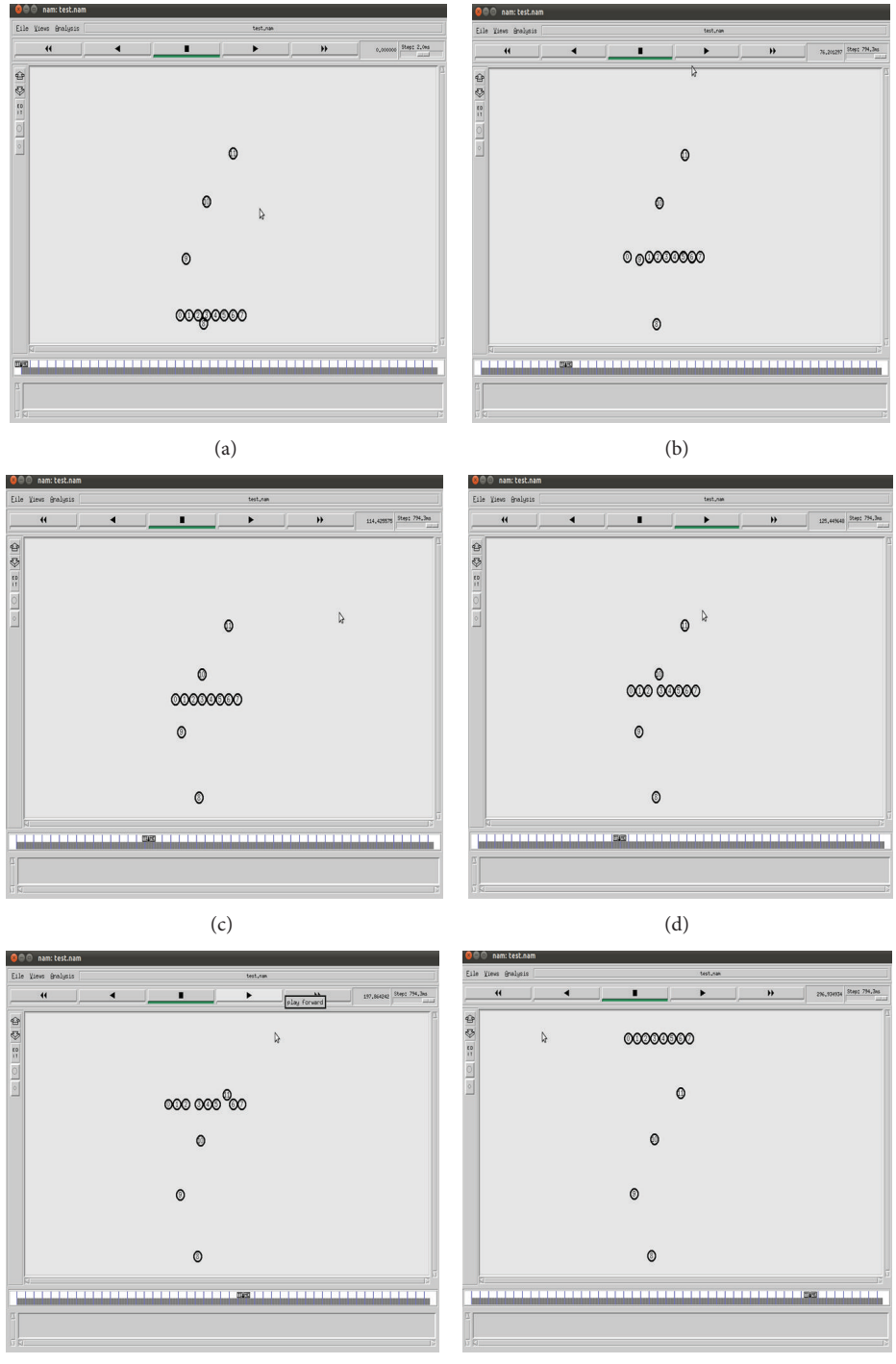


Figure 5: The simulation result of the SPLIT method.

specific constant height. This scenario shows that the UAV that discovered an obstacle avoids it by choosing the SPLIT method, and at this time, the SPLIT method addresses the problem. This choice can impair a key aim, which is to move to the destination right after avoiding the obstacles because of the long avoidance time that is required to avoid all of the obstacles by using the JUMP method.

5. Conclusions

This paper proposed a method for the process in which UAVs avoid three-dimensional, fixed obstacles. Moreover, this paper proposed a specific method to control the altitude and the velocity and to avoid obstacles while retaining the multihop connection between the UAVs during the obstacle

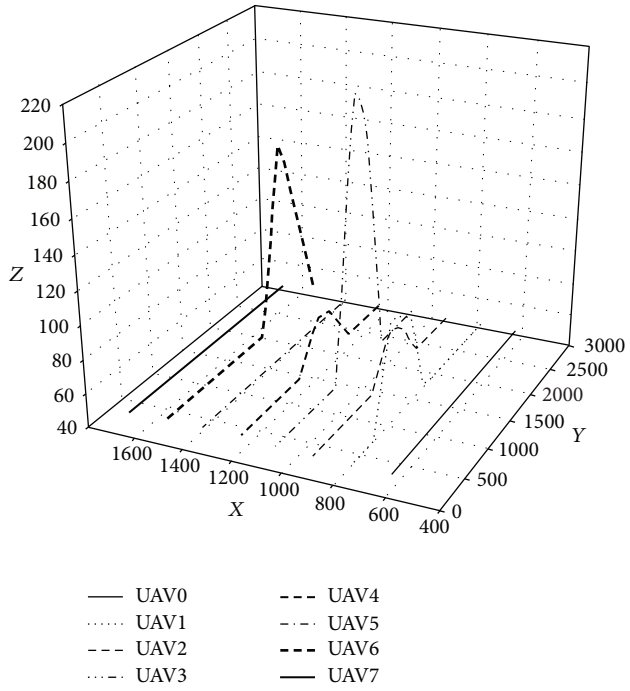


Figure 6: The trajectory of the JUMP method.

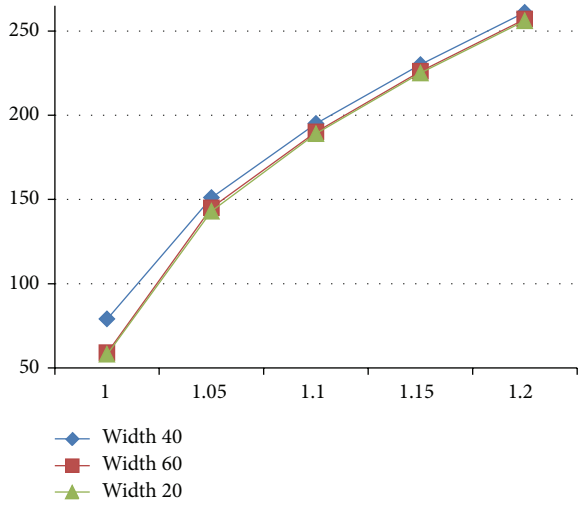


Figure 7: Resulting graph of the experiment that observes how high the obstacle avoided by the JUMP Method is and what effect the change in the obstacle's width has when the value of α is applied.

avoidance process. Another method is proposed to restore the original formation after it has been divided into two groups that were formed to avoid both sides of an obstacle to address the problem that was in the above method. In addition, the proposed method was inspected and evaluated by using a simulation. If the proposed method is applied, it ensures that several UAVs efficiently search without colliding with any obstacle while they are searching an area about which there is little information. When we perform the SPLIT method as a research task in future work, we will

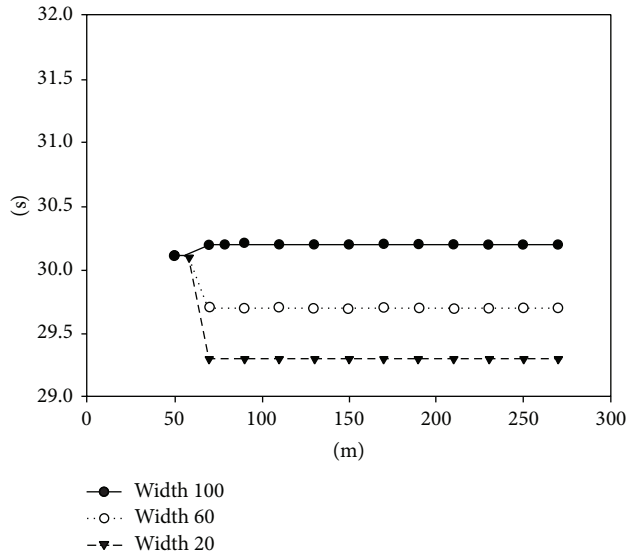


Figure 8: The resulting graph of the time taken to reach the avoidance point through the two methods.

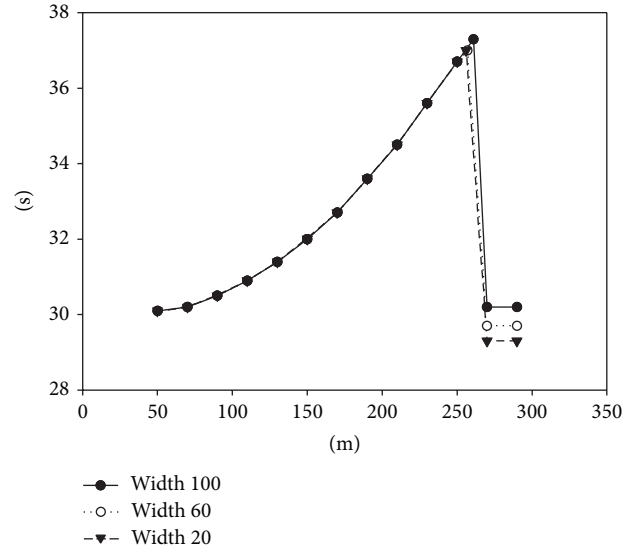


Figure 9: Result graph when the value α is changed into 1.2.

research the method for avoiding obstacles and securing the communication connection between UAVs in a group.

Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT, and Future Planning (no. NRF-2011-0014740).

References

[1] L. F. Bertuccelli, H.-L. Choi, P. Cho, and J. P. How, "Real-time multi-UAV task assignment in dynamic and uncertain environments," in *Proceedings of the AIAA Guidance, Navigation, and Control Conference and Exhibit*, August 2009.

[2] A. Richards and J. P. How, "Aircraft trajectory planning with collision avoidance using mixed integer linear programming," in *Proceedings of the American Control Conference*, pp. 1936–1941, May 2002.

[3] S. Scherer, S. Singh, L. Chamberlain, and S. Saripalli, "Flying fast and low among obstacles," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '07)*, pp. 2023–2029, April 2007.

[4] Y. K. Kwag and C. H. Chung, "UAV based collision avoidance radar sensor," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS '07)*, pp. 639–642, June 2007.

[5] Y. Watanabe, A. J. Calise, and E. N. Johnson, "Vision-based obstacle avoidance for UAVs," in *Proceedings of the AIAA Guidance, Navigation, and Control Conference*, pp. 4862–4872, August 2007.

[6] B. Bethke, J. P. How, and J. Vian, "Group health management of UAV teams with applications to persistent surveillance," in *Proceedings of the American Control Conference (ACC '08)*, pp. 3145–3150, June 2008.

[7] Z. Xu, J. Huo, Y. Wang, J. Yuan, X. Shan, and Z. Feng, "Analyzing two connectivities in UAV-ground mobile ad hoc networks," in *Proceedings of the IEEE International Conference on Computer Science and Automation Engineering (CSAE '11)*, pp. 158–162, June 2011.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 507942, 6 pages
http://dx.doi.org/10.1155/2014/507942

Research Article

Autonomous Position Estimation of a Mobile Node Based on Landmark and Localization Sensor

Se-Jun Park,[1] Jeong-Sik Park,[2] Yong-Ho Seo,[1] and Tae-Kyu Yang[1]

[1] *Department of Intelligent Robot Engineering, Mokwon University, 88 Doanbuk-ro, Seo-gu, Daejeon 302-318, Republic of Korea*
[2] *Department of Information and Communication Engineering, Yeungnam University, 280 Daehak-ro, Gyeongsan, Gyeongbuk 712-749, Republic of Korea*

Correspondence should be addressed to Jeong-Sik Park; parkjs@yu.ac.kr

Received 7 December 2013; Accepted 1 April 2014; Published 16 April 2014

Academic Editor: Jongsung Kim

This study proposes an efficient position estimation method for localizing a mobile node in indoor environment. Although several conventional methods have been successfully applied for position estimation, they have some drawbacks such as low extendibility in an indoor space, intensive computation, and estimation errors. We propose a precise estimation approach based on a localization sensor and artificial landmarks. In our approach, a mobile node autonomously measures the location of landmarks attached to the ceiling with a localization sensor while moving across the landmarks and building a landmark map. And then, the node estimates its location under the ceiling using the map. In this process, we use a landmark histogram and a Kalman filter to reduce estimation errors. Several experiments performed using a mobile robot successfully demonstrated the feasibility of our proposed approach.

1. Introduction

A sensor node in wireless sensor network performs some processes inside its given location, gathers sensory information, and communicates with other connected nodes in the network. On the other hand, a mobile sensor node moves around its given space to perform duties along with above tasks. Hence, it should be capable of measuring a global map of their moving area and estimating their physical location anytime. In recent years, many efforts have been made to deal with such a localization problem. In particular, the localization issue has been directly towards mobile robotics as a fundamental research topic [1–4].

There are efficient localization algorithms for indoor and outdoor environments. Common global positioning system (GPS) is representatively applied for precise outdoor localization. Although GPS is a standard localization system, it often fails in indoor environments due to fading and multipath of GPS signals. The main line of research on indoor localization is to use odometer information obtained from infrared and ultrasonic sensors [5, 6]. However, this approach requires high cost and also gives high risk because its performance is directly affected by the correctness of odometer information. A study proposed a sensor network based localization that uses a multiple sensor nodes adopting low-priced sensors [7]. This approach provides very efficient performance in installed sensor network environments, but its scalability is restricted due to a limited number of sensors.

A few methods take advantage of landmarks and a distance measurement sensor [8, 9]. First, several artificial landmarks are attached to the ceiling. Each landmark possesses its discriminate form to be distinguished from other landmarks. Mobile nodes identify their location by recognizing the landmarks based on distance information obtained from a distance measurement sensor while they move under the ceiling. This approach provides stable localization performance with low cost and it can cope with larger areas for localization by extending the number of landmarks. In this study, we address several drawbacks of conventional landmark-based approaches and propose a better approach.

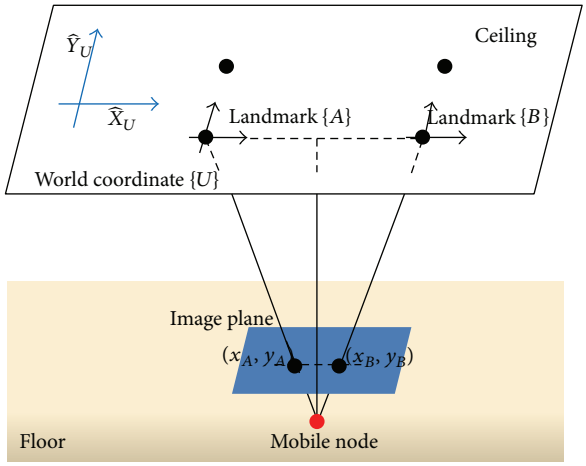


Figure 1: Coordinate system for a node's self-creation of a global map.

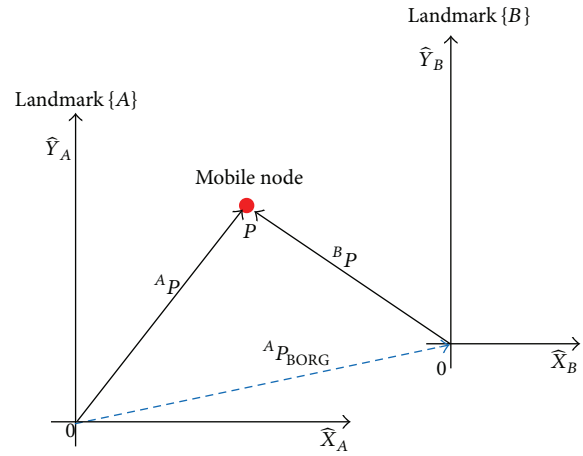


Figure 2: Mobile node's estimation of landmark position.

The remainder of this paper is organized as follows. Section 2 introduces several drawbacks of conventional landmark-based localization approaches. Section 3 and Section 4 explain our proposed method for autonomous position estimation of a mobile node. Section 5 describes experimental setup and results. Finally, conclusions are presented in Section 6.

2. Drawbacks of the Conventional Landmark-Based Localization Approaches

Conventional landmark-based approaches have several drawbacks in real-world environments. The localization performance is greatly affected by the correctness of landmark recognition. The most ideal arrangement of landmarks is to form a grid, but this is not an easy task for a human. Incorrect location of landmarks produces a dead zone or an overlap zone that negatively affects the landmark recognition. Misrecognition of landmarks may induce a mobile node to make a localization error. In addition, the distance sensor might generate distance measurement error, when a node sways or vibrates during movement.

To solve these problems, this paper proposes a map creation procedure with which a mobile sensor node autonomously creates a global map of landmarks for itself. We also propose a precise position estimation method to remove position errors and distance measurement errors.

3. Self-Creation of a Global Landmark Map

A global landmark map means the position of each landmark on the ceiling. A mobile sensor node's self-creation of a global landmark map can assist the node in correctly estimating its position. To introduce our proposed procedure for the self-creation of the map, we make several environmental assumptions. First, the ceiling that landmarks are attached to is a flat surface. Second, the node moves around a flat ground that is parallel with the ceiling. Finally, the mobile node recognizes one or two landmarks in real time while moving under the ceiling and estimates its absolute position in the two-dimensional global map with the coordinates of the recognized landmark(s). Figure 1 represents a coordinate system in which a mobile node estimates its position according to the assumptions.

Figure 2 describes a location of a mobile node in landmark coordinate system. As shown in this figure, the coordinate system of a landmark $\{B\}$ is translated from the system of $\{A\}$ while the two systems have a same bearing. The translation can be described as a vector, ${}^AP_{BORG}$. The position of the mobile node is defined as a vector $({}^AP)$ in the coordinate system of $\{A\}$ and a vector $({}^BP)$ in the system of $\{B\}$. Hence, the vector ${}^AP_{BORG}$ can be represented as given in

$$ {}^AP_{BORG} = {}^AP - {}^BP. \quad (1) $$

This vector means a relative coordinate of the origin of $\{B\}$ to the origin of $\{A\}$. From this point of view, the node is capable of estimating relative coordinates of respective landmarks while searching for the landmarks.

The node should firstly detect and recognize the landmarks before estimating the relative coordinates of landmarks. In general, the node detects the same pair of landmarks several times. For this reason, an average value of relative coordinates that were estimated in each time provides more accurate position of corresponding landmarks. The average value $\text{Mark}_{\text{local}}$ is described as

$$ \text{Mark}_{\text{local}} = \frac{1}{D}\sum_{i=1}^{D} {}^AP_{BORG}(i), \quad (2) $$

where D refers to the number of detection of same landmarks and ${}^AP_{BORG}(i)$ is the relative coordinate estimated from the ith detection.

Finally, relative coordinates of respective landmarks located on the ceiling can be described as

$$\text{Mark}_{\text{world}}(m) = \sum_{i=1}^{m} \sum_{k=1}^{i} \text{Mark}_{\text{local}}(i), \quad (3)$$

where m means the number of all landmarks on the ceiling and $\text{Mark}_{\text{world}}(m)$ denotes a relative coordinate of mth landmark. The coordinate of mth landmark is obtained by $m \times (m + 1)/2$ times in movement of relative coordinates from the origin. The number of movements is defined by i and k values in (3). From relative coordinates calculated by (3), a global map is created.

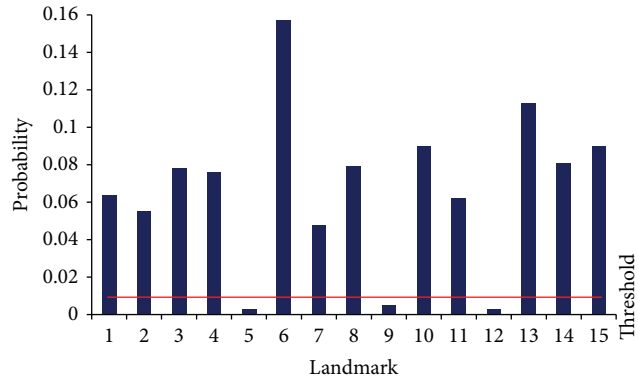


Figure 3: Histogram of consecutive landmark.

4. Precise Position Estimation of a Mobile Node

A mobile node can estimate its position based on the global landmark map. For more precise position estimation of a mobile node, we apply two approaches.

4.1. Removal of Position Errors. While moving under the ceiling, a node first should recognize the landmarks through image processing to estimate its position. A lot of position estimation errors occur due to misrecognition of landmarks. We propose a method to remove position errors caused by incorrectly recognized landmarks. The node estimates its position by recognizing the nearest ceiling landmark while it navigates. At this time, an incorrect recognition of landmarks occurs by dead zone or overlap zone. An observation probability ($p(r_k)$) for consecutive landmarks (r_k) is defined as

$$p(r_k) = \frac{n_k}{n} \quad k = 1, 2, 3, \ldots, L, \quad (4)$$

where n is the total number of landmarks, n_k is the number of kth landmarks, and L is the number of consecutive landmarks.

Figure 3 shows the histogram of consecutive landmarks detected during a node's navigation. A threshold obtained from this result can be used to remove the position errors caused by incorrectly recognized landmarks by removing landmarks (r_k) indicating probability smaller than the threshold.

4.2. Removal of Distance Measurement Errors. Distance measurement errors occur when a node sways or vibrates during movement. Thus, we estimate the state variable value of the dynamic system as input with incoming noise through the sensor using a Kalman filter [10] and estimate the precise position of the mobile node. The state variable value is expressed as

$$\begin{aligned} x_k &= Ax_{k-1} + Bu_{k-1} + w_{k-1}, \\ z_k &= Hx_k + v_k. \end{aligned} \quad (5)$$

In this equation, w_k is a process noise and v_k is a measurement noise. The covariance (Q) of a process noise and the covariance (R) of a measurement noise follow a normal distribution, and Q and R are expressed as

$$p(w) \sim N(0, Q), \qquad p(v) \sim N(0, R). \quad (6)$$

Given the state variable value x_k, the system model can be described as

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k, \\ z_k &= Hx_k + v_k, \end{aligned} \quad (7)$$

where $A = 1, B = 0.2, H = 1, Q = 0$, and $R = 4$. Then, this equation can be expressed as

$$\begin{aligned} x_{k+1} &= x_k + 0.2u_k, \\ z_k &= x_k + v_k. \end{aligned} \quad (8)$$

In this equation, x_{k+1} represents the present position of the mobile node. z_k represents the measured position of the node including noise ($v_k = N(0.2^2)$).

5. Experimental Results

In order to evaluate the proposed position estimation method, we set up experimental environment by attaching 12 landmarks on a ceiling (7.45 m × 6.07 m × 2.4 m) as in Figure 4.

To simulate a mobile node, we designed a type of a mobile robot equipped with four wheels. The body size of the node is W500 × D600 × H500. It can move by itself with a high speed up to 1.5 m/sec. At the top of the node, a distance measurement sensor, StarGazer, is mounted. StarGazer recognizes one or two landmarks and calculates the absolute distance between the node and landmarks at 10 times in every second. Table 1 shows the specification of the mobile node, and Table 2 describes the specifications of StarGazer.

Table 3 shows the performance of self-creation of a global landmark map using actual values and estimated

Figure 4: Experimental environment.

Table 1: Specification of the mobile node.

Item	Specification
Body	
Size	W500 × D600 × H600 (mm)
Weight	30 Kg
Driving	
Driving method	2-wheel differential drive
Speed	Max. 1.5 m/sec

Table 2: Specification of StarGazer.

Item	Specification
Product	HAGISONIC CO., LTD.
Model	StarGazer (HSG-A-02)
Landmark recognition	Max. two
Measurable time	10 times/sec

values of respective landmarks and the distance error for each landmark. The average distance error is relatively low at less than 10 cm (0.0993 m), demonstrating the reliable performance of the map creation method.

Figure 5 illustrates a global map of the ceiling obtained from the estimation values. Each number represents the landmark ID.

Figure 6 shows a moving trajectory of the mobile node, which was estimated when it moves from landmark ID 544 to landmark ID 66 during 68 seconds.

Figure 7 shows a histogram of consecutive landmarks investigated while the node moves. In this figure, landmarks indicating small numbers are due to incorrectly recognized ID.

Figure 8 represents the position of the mobile node when a threshold obtained from the histogram of Figure 7 was applied. In this figure, the measurement noise still remains due to shaking or vibration of the node. Thus, we attempted

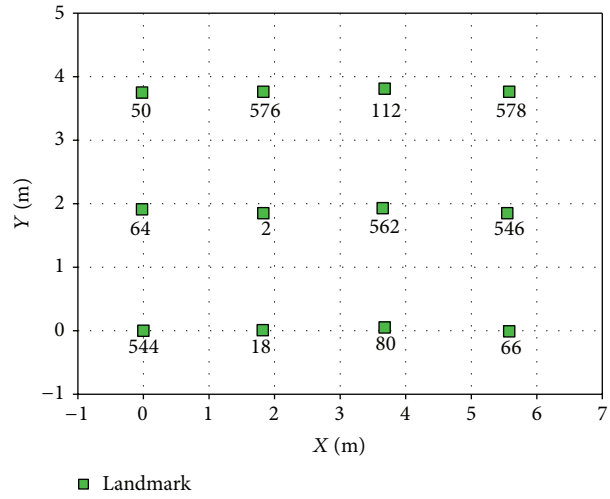


Figure 5: Global landmark map of the ceiling.

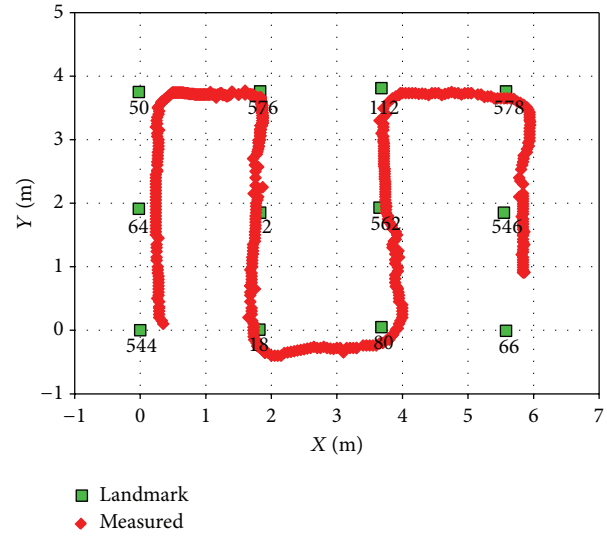


Figure 6: Moving trajectory of a mobile node.

to remove the measurement noise using a Kalman filter and estimated the average and standard deviation of the position noise and distance noise of the mobile node. As shown in Table 4, more accurate position was estimated by removing distance noise (approximately 0.08 m) using the proposed algorithm.

Figure 9 shows experimental results of autonomous navigation to a goal point. The red circles represent the starting point and the goal point, and the black rectangle represents obstacles used for this experiment. The mobile node successfully navigates toward the goal point, avoiding obstacles with accurate position estimation.

Table 3: The performance of the proposed map creation method.

Landmark ID	Actual		Estimation		Distance error (m)
	X (m)	Y (m)	X (m)	Y (m)	
544	0.0	0.0	0.0	0.0	0.0
64	0.0	1.825	−0.0153	1.9059	0.0823
50	0.0	3.645	−0.0090	3.7420	0.0974
576	1.820	3.635	1.8440	3.7689	0.1360
2	1.820	1.820	1.8343	1.8809	0.0626
18	1.820	0.0	1.8199	0.0188	0.0188
80	3.640	0.0	3.6839	0.0750	0.0869
562	3.640	1.815	3.6719	1.9468	0.1356
112	3.640	3.640	3.6844	3.8100	0.1757
578	5.460	3.645	5.5932	3.7748	0.1860
546	5.460	1.825	5.5659	1.8644	0.1130
66	5.460	0.0	5.5571	−0.0113	0.0978

Table 4: Position and distance noise of a mobile node.

	X (m)	Y (m)	Distance (m)
Average	0.0301	0.0599	0.0794
Standard deviation	0.0490	0.0727	0.0876

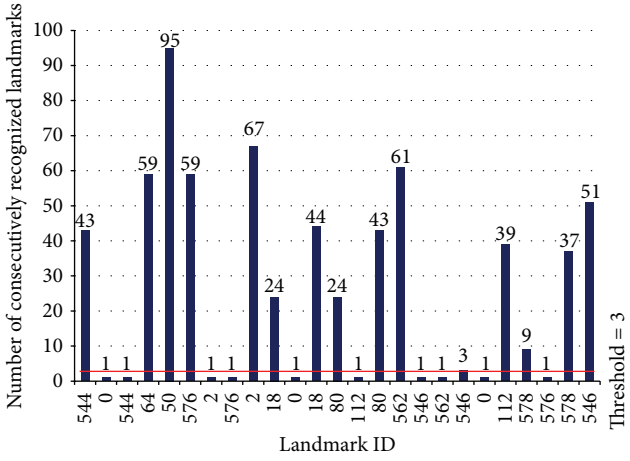


Figure 7: Histogram of consecutive landmarks.

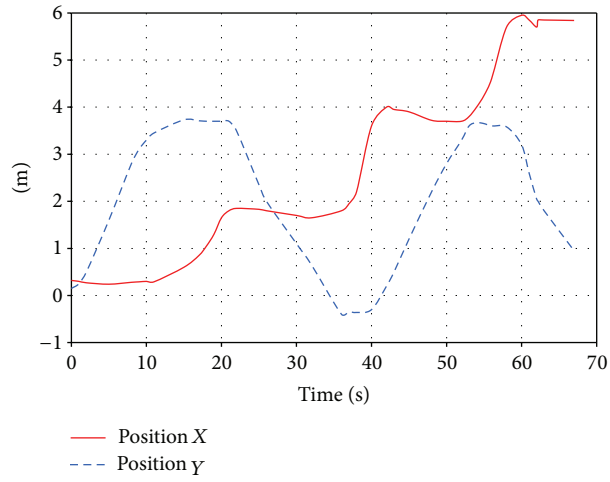


Figure 8: Measured position of mobile node (processing of incorrectly recognized ID).

6. Conclusion

In this study, we proposed an efficient position estimation method for a mobile node's localization based on landmark approach. The method supports the mobile node to autonomously create a global landmark map, while the node moves around the ceiling and searches for every landmark. To cope with position estimation errors, misrecognized landmarks are removed based on a threshold determined by a landmark histogram. In addition, we remove distance measurement errors using a Kalman filter.

We simulated an experimental environment for the verification of the proposed position estimation approach. The method for a mobile node's autonomous creation of a global map was successfully verified, indicating a small difference between actual landmark position and estimated position. Estimation errors were significantly reduced by the histogram and Kalman filter-based approach. We also observed that the mobile node successfully navigates toward any goal points while avoiding obstacles.

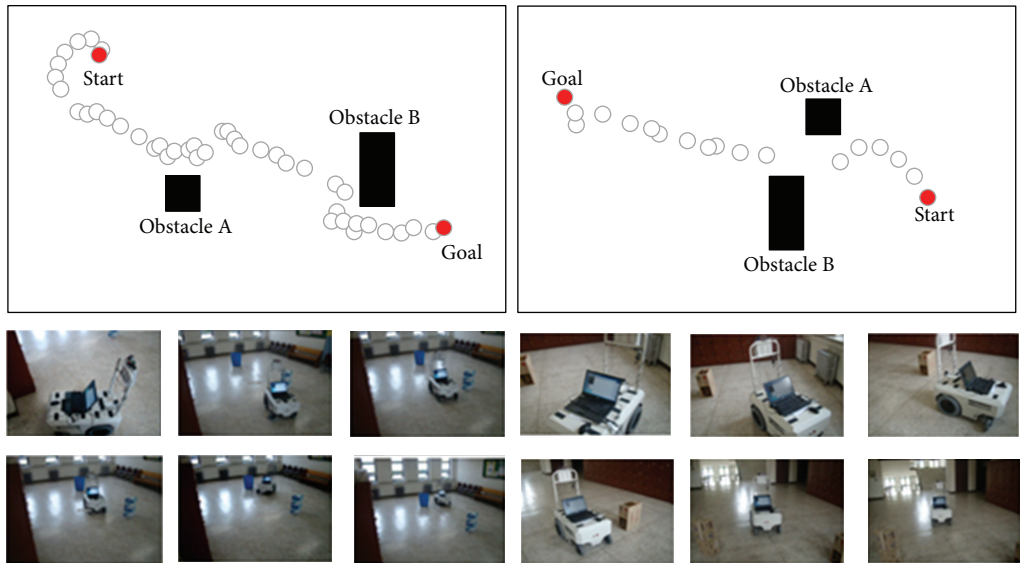


Figure 9: Experiment of goal point tracking.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by the Converging Research Center Program through the Ministry of Science, ICT, and Future Planning, Republic of Korea (2013K000358) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (MEST) (2011-0013776).

References

[1] N. Karlsson, E. di Bernardo, J. Ostrowski, L. Goncalves, P. Pirjanian, and M. E. Munich, "The vSLAM algorithm for robust localization and mapping," in *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 24–29, Barcelona, Spain, April 2005.

[2] M. Montemerlo, *Fast SLAM : a factored solution to the simultaneous localization and mapping problem with unknown data association [Ph.D. thesis]*, Robotics Institute, Carnegie Mellon University, Pittsburgh, Pa, USA, 2003.

[3] M. Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte, and M. Csorba, "A solution to the simultaneous localization and map building (SLAM) problem," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 3, pp. 229–241, 2001.

[4] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*, MIT Press, 2005.

[5] J. Borenstein and L. Feng, "Measurement and correction of systematic odometry errors in mobile robots," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 6, pp. 869–880, 1996.

[6] K. Pahlavan, X. Li, and J.-P. Mäkelä, "Indoor geolocation science and technology," *IEEE Communications Magazine*, vol. 40, no. 2, pp. 112–118, 2002.

[7] F. Dressler, "Sensor-based localization-assistance for mobile nodes," in *Proceedings of the 4th GI/ITG KuVS Fachgespräch Drahtlose Sensornetze*, pp. 102–106, Zurich, Switzerland, March 2005.

[8] S. Lee and J.-B. Song, "Mobile robot localization using infrared light reflecting landmarks," in *Proceedings of the International Conference on Control, Automation and Systems (ICCAS '07)*, pp. 674–677, Seoul, Republic of Korea, October 2007.

[9] H. Wang, H. Yu, and L. Kong, "Ceiling light landmarks based localization and motion control for a mobile robot," in *Proceedings of the IEEE International Conference on Networking, Sensing and Control (ICNSC '07)*, pp. 285–290, London, UK, April 2007.

[10] G. Welch and G. Bishop, "An introduction to the Kalman filter," Tech. Rep. 95-041, University of North Carolina, Chapel Hill, NC, USA, 2006.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 297689, 6 pages
http://dx.doi.org/10.1155/2014/297689

Research Article

A Novel 3D Indoor Localization Scheme Using Virtual Access Point

Taekook Kim[1] and Eui-Jik Kim[2]

[1] *Department of Electrical Engineering, Korea University, Anam-dong 5-ga, Seongbuk-gu, Seoul 136-713, Republic of Korea*
[2] *Department of Ubiquitous Computing, Hallym University, 39 Hallymdaehak-gil, Chuncheon-si, Gangwon-do 200-702, Republic of Korea*

Correspondence should be addressed to Eui-Jik Kim; ejkim32@hallym.ac.kr

Received 12 December 2013; Accepted 27 March 2014; Published 14 April 2014

Academic Editor: Ken Choi

The global positioning system (GPS) is a popular choice for accurate location sensing and is often used in navigation systems. However, GPS fails to operate inside buildings because satellite signals are blocked. Thus, GPS cannot be used for localization indoors. Moreover, it is difficult to distinguish between the different floors. Research to overcome this problem using the ultrawideband (UWB), radio frequency identification (RFID), and infrared ray (IR) has been conducted. These methods, however, have drawbacks, such as the need for an additional communication module and database. Wireless local area network (WLAN) access point (AP) is widely installed at various locations, and using these WLAN APs can be a viable alternative. The proposed system in this paper indicates the location information in the service set identifier (SSID) of virtual AP. Therefore, merely by scanning the WLAN signals, mobile users can detect their location indoors.

1. Introduction

The advancement of information and communications technology and mobile radio communication network has rendered it possible to obtain the information regarding the location of a user. Thus, location based services (LBS) are gaining ground. GPS is commonly used to estimate a user's location. However, weak reception of signals indoors makes it difficult to pinpoint the location indoors. Further, it is difficult to distinguish between the floors within a building [1].

The proposed system in this paper indicates the floor information in the service set identifier (SSID) of AP, besides indicating the GPS coordinates including the latitude and longitude. The SSID is an important scanning parameter, along with the name of the network. While scanning, a mobile node (MN) searches for an SSID and may build a list of SSIDs for presentation to the user. Therefore, by simply scanning the WLAN signals, an MN can detect its own location indoors. The mobile user's latitude (Y-coordinate), longitude (X-coordinate), and floor (Z-coordinate) information is known, enabling 3D localization. Furthermore, the proposed system can be applied to all wireless access methods including the UWB, RFID, IR, and WLAN.

There have been researches to estimate the location indoors using various methods like the UWB, RFID, IR, and WLAN. Methods which use the UWB, RFID, and IR have a weakness; they require an additional communication module [2–7]. To tackle this problem, a widely installed AP is used to develop an alternative indoor localization scheme. Currently, a fingerprint method using an AP is widely in use for indoor localization [8, 9].

However, the conventional fingerprint method using AP is disadvantageous, as it requires a database (DB) server that stores all the location information [10–12]. Furthermore, several reference points should be set in advance to gather received signal strength indication (RSSI) that forms a fingerprint DB (a signal pattern map) [13]. Therefore, when the position of an AP or the neighboring environment changes, the RSSI information in each location has to be measured again and modified, to be updated in the DB server. Further, an MN that wants to estimate its own location must access the DB server or download the fingerprint DB beforehand,

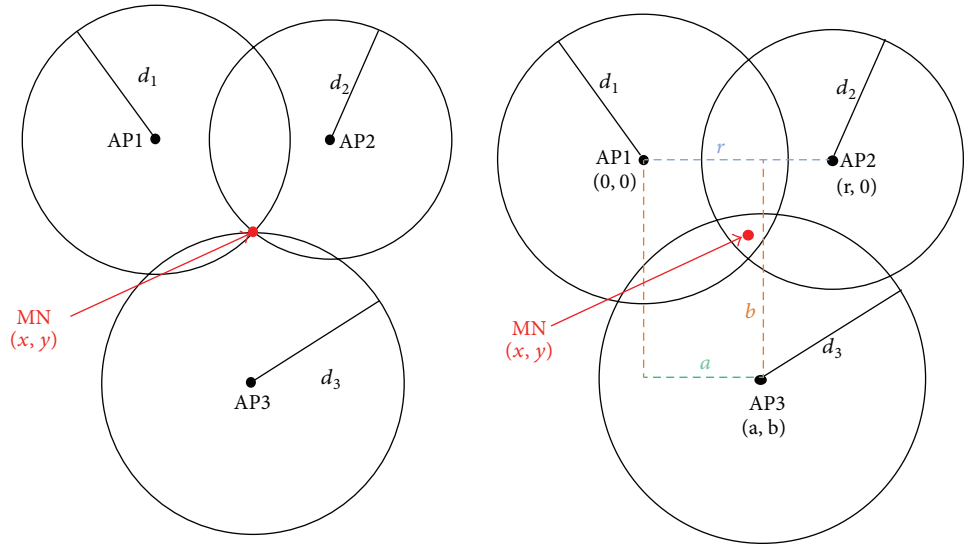


Figure 1: Trilateration (MN's ideal location versus MN's real location).

to compare the DB information with the RSSI and AP information [14, 15].

2. Background

2.1. Distance Estimation Using RSSI. A localization scheme uses both distance estimation using RSSI and position estimation using trilateration. To plot the location of a mobile user, the MN sends a "probe request message" to WLAN AP, requesting information about SSID. "Probe response message," which is the response by an AP, includes an SSID [16]. RSSI is an indication of the power level being received by the antenna. Distance between the AP and the MN can be estimated using RSSI [17–19]. Distance estimation using the path loss model is as follows. Distance estimation method using WLAN RSSI utilizes the "free space loss" formula that uses the distance between antennas to calculate the path loss. Pass loss L is computed as shown in the following:

$$L = 20\log_{10}\left(\frac{4\pi d}{\lambda}\right)[\text{dB}]. \quad (1)$$

Using the L calculated in (1), the distance between antennas, d, can be determined using the following [17]:

$$d = \frac{\lambda}{4\pi}10^{L/20} = \frac{c}{4\pi f}10^{L/20} = 10^{(L-40)/20}[\text{m}]. \quad (2)$$

Here, f refers to the "frequency (2.4×10^9 Hz)," and c is the "ray velocity (3.0×10^8 m/s)." Position is estimated using the distances between the three APs and the MN, all of which are calculated using (2).

2.2. Trilateration. A trilateration method is used to estimate the location, as shown in Figure 1.

Trilateration is a method that uses the distances between MN and three known reference points to estimate MN's position. We know the distances (d_1, d_2, and d_3) between three APs (AP1, AP2, and AP3) and MN; the position of MN (x, y) can be calculated using (3). The path loss measured, as shown in Figure 2, is greater than the theoretical path loss because of noise. Therefore, the actual location of MN is as shown on the right side of Figure 1. The x and y values in (3) were derived from equations of circles:

$$d_1^2 = x^2 + y^2,$$
$$d_2^2 = (x - r)^2 + y^2,$$
$$d_3^2 = (x - a)^2 + (y - b)^2,$$

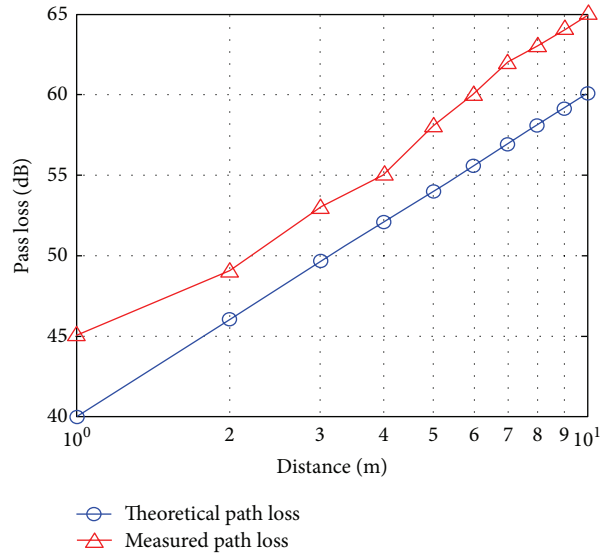


Figure 2: Distance error of free space loss model.

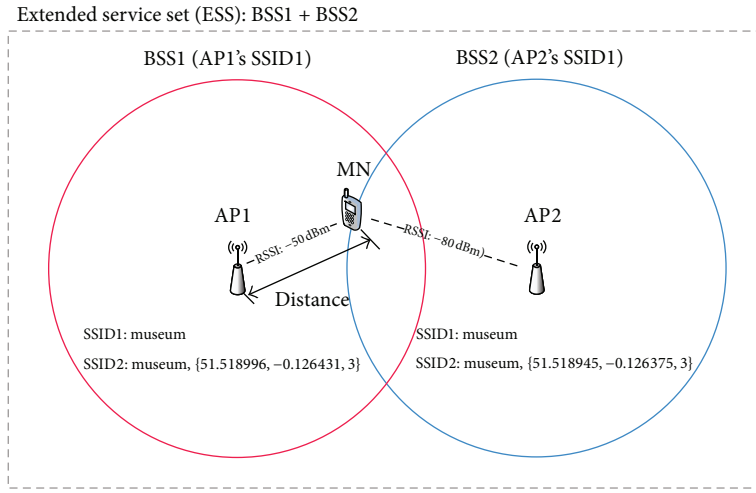


Figure 3: Configuration of the proposed system.

$$x = \frac{d_1^2 - d_2^2 + r^2}{2r},$$
$$y = \frac{d_1^2 - d_3^2 - x^2 + (x-a)^2 + b^2}{2b} = \frac{d_1^2 - d_3^2 + a^2 + b^2}{2b} - \frac{a}{b}x. \quad (3)$$

Table 1: AP's SSID under the proposed system.

Symbol	Description
{, }	Brace: indicator of location information in SSID 2 under the proposed system.
SSID 1	Network name. For wireless internet.
SSID 2	Network name, {latitude, longitude, floor}. For localization.

3. Proposed System

The proposed system in this paper indicates the location information (latitude, longitude, and floor) in the SSID of virtual AP. Generally, APs are fixed in a particular position without considerable movement; the location of APs can be plotted with latitude, longitude, and floor information.

The position (latitude and longitude coordinates) of APs installed indoors can be obtained externally using GPS coordinates and floor plan. Once we determine the GPS coordinates of the face of the exterior wall and the floor plan, we can calculate the GPS coordinates of a specific location. The circumference of Earth along the equator is 40,075.017 km in length; hence, the distance between two consecutive longitudes (1 degree) is 109.8 km. This paper used GPS coordinates of six-digit precision. GPS coordinates of one-digit decimal point have a margin error of 11 km, while those of six-digit decimal points have a margin error of only 11 cm. GPS is a satellite navigation system that provides the user's current location by receiving signals from the satellite. Therefore, commercial GPS has an error margin of about 10 m while plotting GPS coordinates on the map, because of reception errors. This paper, however, uses the absolute value of the GPS coordinates on the map and not the measured GPS signals; therefore, there is an 11 cm error margin only.

In this paper, a brace "{, }" is proposed and used as a location information indicator in an AP's SSID.

AP's SSID should retain the same name to provide the mobility in the extended service set (ESS), which is composed of several basic service sets (BSSs) [16]. Therefore, under the proposed system, each AP, which has multiple SSIDs (SSID 1, SSID 2), operates as a virtual AP. Hence, SSID 1 of an AP, which forms one ESS, should be set with the same name. Further, SSID 2 is used to display the location information of each AP.

The configuration of the proposed system is shown in Figure 3. Here, an AP has two SSIDs, operating like two virtual APs.

4. Experiment Results

4.1. Location Error. The proposed system uses SSID for localization in AP. As is evident from Table 1, the SSID indicates the name and the location information including the latitude, longitude, and floor under the proposed system. The brace "{, }" is introduced as an indicator to show the location information in SSID.

Location estimation was tested with one to three APs, with 10 m interval between the APs. As demonstrated in Figure 4, the average distance error of localization was 3.8 m with one AP, but it reduced to 2.1 m with the trilateration method using three APs.

4.2. Demonstration of 3D Indoor Localization. Google Mobile Maps provides indoor maps for several major buildings in the world, and the number of buildings for which such maps are provided is increasing [20]. The proposed method was

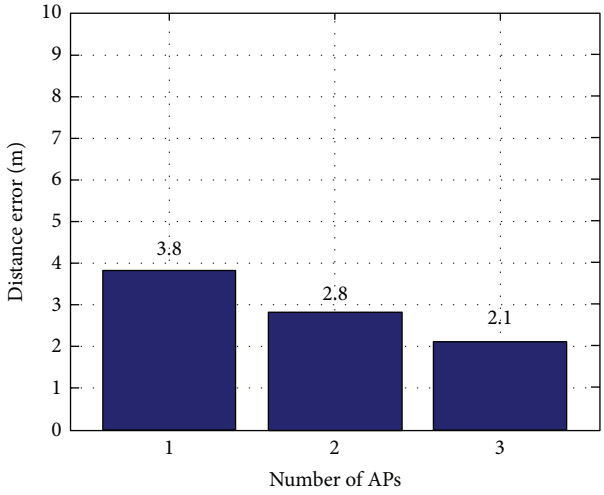


Figure 4: Distance error of localization.

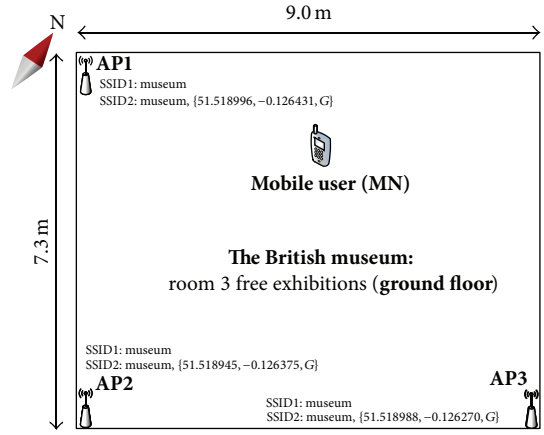


Figure 5: Experimental setup (the British Museum: ground floor).

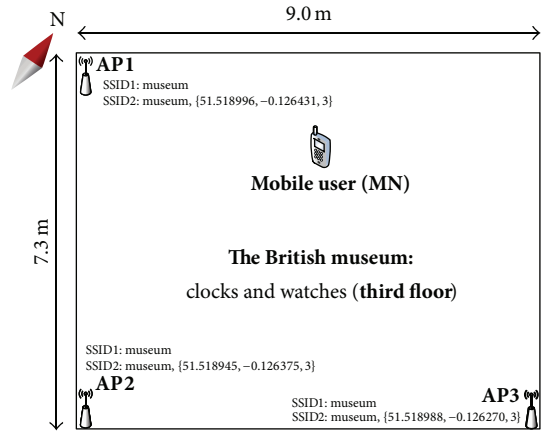


Figure 6: Experimental setup (the British Museum: third floor).

tested for localization within the British Museum, for which Google provides indoor maps. We opted for a room with the same size as "room 3 free exhibitions (or clocks and watches)" of the British Museum ground floor (or third floor) and assumed this to be the British Museum. Figures 5 and 6 display the experimental setup of the British Museum. The location information (latitude, longitude, and floor) of "room 3 free exhibitions (or clocks and watches)" was set in the SSID of AP1, AP2, and AP3, and the 3D indoor localization was tested.

Figure 7 shows the SSID of the AP that was extracted by scanning the WLAN signals near the MN. MN extracts position information from three strongest signals among the received SSIDs with location information. This information is used to estimate the distances between AP1 and MN, AP2 and MN, and AP3 and MN. For example, the figure of latitude 51.518996, longitude −0.126431, and the third floor can be detected in SSID "museum, {51.518996, −0.126431, 3}." Therefore, the MN knows the location (latitude, longitude, and floor) of AP1 using the received SSID from AP1. Furthermore, RSSI is estimated in signal strength of received SSID from AP1. As shown in Figure 1, d_1, the distance between MN and AP1, can be estimated. Further, by employing the same method, the location (latitude, longitude, and floor) of AP2 and AP3 can be estimated; d_2, which is the distance between MN and AP2, and d_3, which is the distance between MN and AP3, can be computed as well. That is, the distances between APs and MN are calculated using the information received from the three SSIDs, and the position is estimated using the three distances. MN's position is calculated using trilateration using the three estimated distances.

The calculated location information can be plotted on a map. Figures 8 and 9 are examples of the calculated location information (latitude, longitude, and floor) marked on Google Mobile Indoor Maps. When the floor information is set as "G" in the SSID of the AP, MN recognizes its current location as ground floor, as in Figure 8. Likewise, when the floor information is set as "3," MN recognizes its current location as third floor, as in Figure 9.

5. Conclusion

The proposed system in the paper displays the GPS coordinates and floor information in the SSID of virtual AP, without any additional equipment or communications module to detect MN's own location. A mobile user scans the nearby WLAN signals to collect the SSID of AP and the RSSI information. The distances are estimated using the received SSIDs, and MN's own position is estimated from the distances from three sides (APs). By simply scanning the WLAN signals, MN can detect its own location. Furthermore, the proposed method transmits signals that include the position information in a periodic beacon, which can be applied to all wireless access methods like UWB, RFID, IR, and WLAN. Thus, the proposed method can be applied in real life without modifying the AP's software and hardware.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

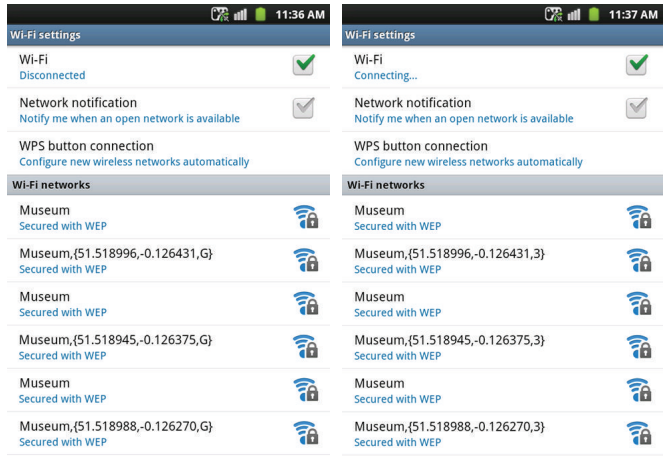


Figure 7: Extracting the location information in SSID by scanning the WLAN signals (the British Museum: ground floor versus third floor).

Figure 8: 3D localization in Google Mobile Indoor Maps (the British Museum: ground floor, room 3 free exhibitions).

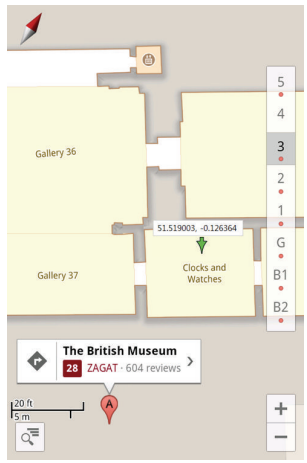


Figure 9: 3D localization in Google Mobile Indoor Maps (the British Museum: third floor, clocks and watches).

Acknowledgment

This research was supported by Hallym University Research Fund, 2014 (HRF-201402-009).

References

[1] K. Pahlavan, X. Li, and J. Mäkelä, "Indoor geolocation science and technology," *IEEE Communications Magazine*, vol. 40, no. 2, pp. 112–118, 2002.

[2] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 37, no. 6, pp. 1067–1080, 2007.

[3] H. Koyuncu and S. H. Yang, "A survey of indoor positioning and object locating systems," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 10, pp. 121–127, 2010.

[4] Z. Li, W. Dehaene, and G. Gielen, "A 3-tier UWB-based indoor localization system for ultra-low-power sensor networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 6, pp. 2813–2818, 2009.

[5] L. M. Ni, D. Zhang, and M. R. Souryal, "RFID-based localization and tracking technologies," *IEEE Wireless Communications*, vol. 18, no. 2, pp. 45–51, 2011.

[6] S. S. Saad and Z. S. Nakad, "A standalone RFID indoor positioning system using passive tags," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 5, pp. 1961–1970, 2011.

[7] S. A. Golden and S. S. Bateman, "Sensor measurements for Wi-Fi location with emphasis on time-of-arrival ranging," *IEEE Transactions on Mobile Computing*, vol. 6, no. 10, pp. 1185–1198, 2007.

[8] S.-H. Fang and T.-N. Lin, "Principal component localization in indoor wlan environments," *IEEE Transactions on Mobile Computing*, vol. 11, no. 1, pp. 100–110, 2012.

[9] S.-H. Fang, T.-N. Lin, and K. Lee, "A novel algorithm for multipath fingerprinting in indoor WLAN environments," *IEEE Transactions on Wireless Communications*, vol. 7, no. 9, pp. 3579–3588, 2008.

[10] S.-H. Fang, T.-N. Lin, and P.-C. Lin, "Location fingerprinting in a decorrelated space," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 5, pp. 685–691, 2008.

[11] K. Kaemarungsi and P. Krishnamurthy, "Modeling of indoor positioning systems based on location fingerprinting," in *Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE INFOCOM '04)*, vol. 2, pp. 1012–1022, March 2004.

[12] C. Nerguizian, C. Despins, and S. Affès, "Geolocation in mines with an impulse response fingerprinting technique and neural networks," *IEEE Transactions on Wireless Communications*, vol. 5, no. 2, pp. 603–611, 2006.

[13] P. Bahl and V. N. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," in *Proceedings of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE INFOCOM '00)*, vol. 2, pp. 775–784, March 2000.

[14] A. S. Paul and E. A. Wan, "RSSI-based indoor localization and tracking using sigma-point Kalman smoothers," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 5, pp. 860–873, 2009.

[15] A. S. Krishnakumar and P. Krishnan, "On the accuracy of signal strength-based location estimation techniques," in *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '05)*, vol. 1, pp. 642–650, March 2005.

[16] M. Gast, *802.11 Wireless Networks: The Definitive Guide*, O'Reilly, Sebastopol, Calif, USA edition, 2005.

[17] D. B. Green and M. S. Obaidat, "An accurate line of sight propagation performance model for Ad-Hoc 802.11 wireless LAN (WLAN) devices," in *Proceedings of the IEEE International Conference on Communications (ICC '02)*, vol. 5, pp. 3424–3428, May 2002.

[18] S. Siddiqi, G. S. Sukhatme, and A. Howard, "Experiment in Monte-Carlo localization using WiFi signal strength," in *Proceedings of the International Conference on Advanced Robotics*, pp. 210–223, 2003.

[19] J. Yin, Q. Yang, and L. M. Ni, "Learning adaptive temporal radio maps for signal-strength-based location estimation," *IEEE Transactions on Mobile Computing*, vol. 7, no. 7, pp. 869–883, 2008.

[20] http://support.google.com/gmm/bin/answer.py?hl=en&answer=1685827/.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 578213, 8 pages
http://dx.doi.org/10.1155/2014/578213

Research Article

Side Information Generation for Distributed Video Coding Using Spatiotemporal Joint Bilinear Upsampling

Wenhui Liu, Krishna Rao Vijayanagar, and Joohee Kim

Illinois Institute of Technology, Chicago, IL 60616, USA

Correspondence should be addressed to Joohee Kim; joohee@ece.iit.edu

Received 26 December 2013; Accepted 18 March 2014; Published 14 April 2014

Academic Editor: Thomas Wook Choi

Distributed video coding presents a viable solution for power-constrained multimedia communication. However, its relatively low coding efficiency compared to the conventional video coding schemes remains a challenging issue. The rate-distortion performance of distributed video coding is highly dependent on the quality of side information generated at the decoder and various techniques have been proposed to improve the side information quality in block-based and frame-based distributed video coding architectures. In this paper, a robust spatiotemporal joint bilateral upsampling based side information generation method is proposed. The proposed side information generation method is based on a block-based low-complexity distributed video coding architecture with adaptive block coding mode classification. A partially reconstructed Wyner-Ziv (WZ) frame with skip and key blocks is downsampled and spatiotemporal error concealment and joint bilateral upsampling are used to generate the side information. Simulation results show that the proposed method improves the quality of side information significantly while keeping low computational complexity.

1. Introduction

Video coding technology has played a key role in the explosion of the current multimedia society. The success of the widespread deployment of digital video applications and services is largely built on the predictive video coding paradigm where the encoder exploits the video redundancy and irrelevancy. This type of video coding is well suited for broadcasting or one-to-many video transmission systems where video is encoded once and decoded many times. However, in resource-constrained environments, a low-complexity encoder is necessary at the expense of a high-complexity decoder while still maintaining a high coding efficiency.

Distributed video coding (DVC) has emerged as a new video compression paradigm for video applications with resource-constrained devices because it enables low-complexity encoding and is naturally robust against transmission errors. Over the past decade, several practical implementations of DVC have been proposed including the Stanford codec [1], PRISM codec [2], and DISCOVER codec [3]. However, current DVC architectures still have several technical limitations that prevent their widespread use in real-world applications. In particular, there is still a significant gap in terms of compression efficiency between the current DVC solutions and conventional predictive video coding techniques.

Since the coding efficiency is highly affected by the quality of SI in DVC, lots of efforts have been made to improve the quality of SI [4–18]. Popular SI generation techniques exploit spatial correlation within the same frame and/or temporal correlation between the consecutive frames [4–6]. Recently, optical flow based methods [7, 8], hash information generated at the encoder [9–11], or multiresolution based techniques [12–18] have been introduced to improve the SI quality. However, most of these methods have high complexity and long decoding time due to a feedback-based architecture.

In this paper, we propose a novel SI generation scheme based on spatiotemporal joint bilateral upsampling (STJBU), which is simple and applicable to any block-based DVC architecture. The proposed method consists of three steps: (1) downsampling of a partially reconstructed WZ frame, (2) SI

generation for the WZ blocks in the downsampled WZ frame, and (3) upsampling of the WZ frame using the proposed STJBU algorithm.

The rest of the paper is organized as follows. We review related work on SI generation in DVC in Section 2. In Section 3, the low-complexity DVC (LC-DVC) [19] architecture is briefly introduced which is the basis for the proposed SI generation technique. Then the proposed STJBU based SI generation method is explained in Section 4. Simulation results are presented in Section 5 and the conclusion of the paper is given in Section 6.

2. Related Work

In the past few years, various approaches have been proposed to improve the performance of DVC. The main issues restricting the use of current DVC architectures in practical applications are its low coding efficiency, high decoding latency, and the presence of a feedback channel. In particular, since the coding efficiency is highly affected by the SI quality in DVC, extensive research has been performed to improve the quality of SI.

A multiple motion hypotheses pixel-based temporal interpolation method is proposed in [4], where global and local motion estimation is incorporated. This work has been extended to an adaptive pixel-based temporal interpolation scheme [5] which can adaptively switch between spatial interpolation and forward/backward temporal extrapolation for SI generation. Similarly, a mode decision scheme is presented in [6] to determine the interpolation mode for each block by combining forward and backward motion vectors. Recently, the optical flow algorithm has been exploited for SI generation to compensate for the weaknesses of block-based methods. An optical flow based SI generation algorithm is proposed in [7] which improves the SI quality by obtaining more accurate motion vectors. A similar method proposed in [8] uses optical flow to improve the SI quality and block clustering to increase local adaptivity in the noise modeling. In general, the complex motion estimation process used in these methods incurs high computational complexity and long decoding time.

In the SI generation method proposed in [9], seed blocks are selected first and these blocks are used for motion estimation of the other blocks. Extra information for WZ blocks was transmitted in [10] to help the block matching process at the decoder. Another method called frame-hash uses a highly compressed WZ frame with zero motion vectors to improve the quality of SI [11]. However, the performance of the hash-based DVC schemes is highly dependent on the accuracy of the rate allocation mechanism. An alternative method is to use multiple resolutions in encoding WZ frames. Recently, several SI generation methods based on a mixed-resolution (MR) DVC architecture have been proposed [12–18]. In the MR-DVC architecture, the SI quality is improved by exploiting the spatial relationship between the original frames and the scaled ones.

Spatial low-pass filtering is used in image processing to replace a pixel by a uniform or weighted average of its neighboring pixels. An edge preserving bilateral filter was originally proposed in [20] to alleviate the drawback of spatial low-pass filtering when it is performed over discontinuous regions. It takes into account both the geometric closeness of pixels and their photometric similarity. This noniterative filter smooths images while preserving edges by means of a nonlinear combination of nearby pixel values. Joint bilateral filter proposed in [21] extends the bilateral filter to two correlated images. It filters one image with weights generated using the other image. An alternative joint edge-preserving filter, the guided filter, has been proposed in [22] where the guided filter is derived from a local linear model and can perform filtering in constant time. In [23], the joint bilateral filter has been further extended on image pairs with different resolutions, namely, the joint bilateral upsampling. In [24], a multiresolution bilateral filtering is proposed where the bilateral filter is combined with wavelet thresholding to provide an image denoising framework. The joint bilateral filtering has been successfully applied in a variety of image processing and computer vision applications such as photo enhancement and stereo matching [25].

3. Architecture of Low-Complexity DVC

A simple and unidirectional LC-DVC architecture is proposed in [19]. In the encoder of LC-DVC, an incoming frame is adaptively classified as a key or a WZ frame. The key frame is encoded using the H.264/AVC encoder in intramode. The WZ frames are divided into 4×4 nonoverlapping blocks and the blocks are further classified into skip, key, and WZ blocks. The classification map resulting from the block classification process is compressed using arithmetic coding and sent to the decoder. The skip blocks are not transmitted and can be reconstructed at the decoder with help of the previous frame. The key blocks are encoded using H.264/AVC in intramode. The WZ blocks are transformed, quantized, and the bit planes are extracted and encoded using BCH codes.

At the decoder, the key frames are decoded using the H.264/AVC decoder. For a WZ frame, the key blocks are decoded first and then the skip blocks are copied from colocated blocks in the previous frame according to the classification map. As it is shown in Figure 1, a partially reconstructed WZ frame which contains the key and skips blocks is generated. Then, the SI for the WZ blocks is generated by using the proposed method which can be applied to any block-based DVC architecture.

4. Proposed SI Generation Algorithm

The procedure of the proposed SI generation method is shown in Figure 2 and can be divided into 3 steps: (1) downsampling of a partially reconstructed WZ frame, (2) SI generation in the downsampled partially reconstructed WZ frame, and (3) upsampling of the error-concealed WZ frame using the proposed STJBU algorithm.

Figure 1: An example of a partially reconstructed WZ frame.

4.1. Downsampling of the Partially Reconstructed WZ Frame. In order to reduce the computational complexity in spatiotemporal SI generation methods, the partially reconstructed WZ frame is first downsampled. Downsampling has been used in various image or video compression applications to improve the compression efficiency while reducing the computational complexity [26–30]. The simplest downsampling method is to retain only every Mth sample to create a lower resolution signal in downsampling by a factor of M. However, this simple downsampling method causes aliasing in the resulting downsampled signal. In this paper, four different downsampling methods are used.

4.1.1. Nearest Neighbor Downsampling. The intensity of a pixel in the downsampled image is the intensity of the nearest pixel in the original image as shown in (1):

$$K_{NN}(x) = \begin{cases} 1; & \text{if } |x| < 0.5 \\ 0; & \text{otherwise.} \end{cases} \tag{1}$$

4.1.2. Bilinear Downsampling. Bilinear downsampling considers the closest 2×2 neighborhood of known pixel values surrounding the unknown pixel. It can be implemented by the triangle kernel given in the following:

$$K_{BL}(x) = \begin{cases} 1 - |x|; & |x| < 1 \\ 0; & \text{otherwise.} \end{cases} \tag{2}$$

4.1.3. Bicubic Downsampling. The output pixel value after bicubic downsampling is a weighted sum of the pixels in the nearest 4×4 neighborhood as shown in the following:

$$K_{BC}(x) = \begin{cases} 1.5|x|^3 - 2.5|x|^2 + 1; & \text{if } x \le 1 \\ -0.5|x|^3 + 2.5|x|^2 - 4|x| + 2; & \text{if } 1 < x \le 2 \\ 0; & \text{otherwise.} \end{cases} \tag{3}$$

4.1.4. Lanczos Downsampling. The output pixel value of the downsampled image is obtained by using a convolution kernel given in the following:

$$K_{LZ}(x) = \begin{cases} \sin(x) \sin c\left(\dfrac{x}{a}\right); & \text{if } |x| < 3 \\ 0; & \text{otherwise.} \end{cases} \tag{4}$$

4.2. SI Generation at a Lower Resolution. After the partially reconstructed WZ frame is downsampled, SI is generated for the WZ blocks by exploiting the spatial and temporal correlation. Within a low-delay DVC, the decoder cannot wait for the future frame to arrive before starting the SI generation process and so it must use only the previously reconstructed frame for temporal information. Since the proposed DVC method is block-based and it uses a unique block classification scheme, the decoder is ensured that every WZ block is surrounded by either a key or a skip block in its adjacent 4 neighbors. In this paper, we consider two different methods which are bilinear error concealment and inpainting for SI generation at a lower resolution.

4.2.1. Bilinear Interpolation. SI generation at decoder can be regarded as error concealment (EC) process where the WZ blocks have to be estimated using EC techniques. Among various spatial error concealment techniques [31–34], bilinear error concealment [31] is chosen to estimate the WZ blocks because it is simple but highly efficient.

Bilinear interpolation is a spatial error concealment method which uses the spatially adjacent blocks to recreate the missing pixels by a weighted averaging procedure. Let x and y represent the vertical and horizontal coordinates of the WZ block, where $0 \le x \le Q-1$ and $0 \le y \le Q-1$ (Q is the WZ block size). Let $T(y)$ and $B(y)$ be the pixels to the top and bottom of the WZ block and let $L(x)$ and $R(x)$ be the pixels to the left and right of the WZ block. If P is the estimated pixel, it can be calculated by (5). The weights are defined in (6) so that they are inversely proportional to the distance of the neighboring pixels from the estimated pixel:

$$p = \frac{T(y)\, w_T(x) + B(y)\, w_B(x) + L(x)\, w_L(y) + R(x)\, w_R(y)}{w_T(x) + w_B(x) + w_L(y) + w_R(y)} \tag{5}$$

$$\begin{aligned} w_T(x) &= Q - x \\ w_B(x) &= x + 1 \\ w_L(y) &= Q - y \\ w_R(y) &= y + 1. \end{aligned} \tag{6}$$

4.2.2. Region-Filling Inpainting. EC at the lower resolution frame can also be regarded as a hole-filling problem. Region-filling inpainting technique proposed in [35–37] fills holes within the image by propagating linear structure (also called isophotes) into the target region by diffusion. This interactive processing includes 3 steps, namely, patch priorities

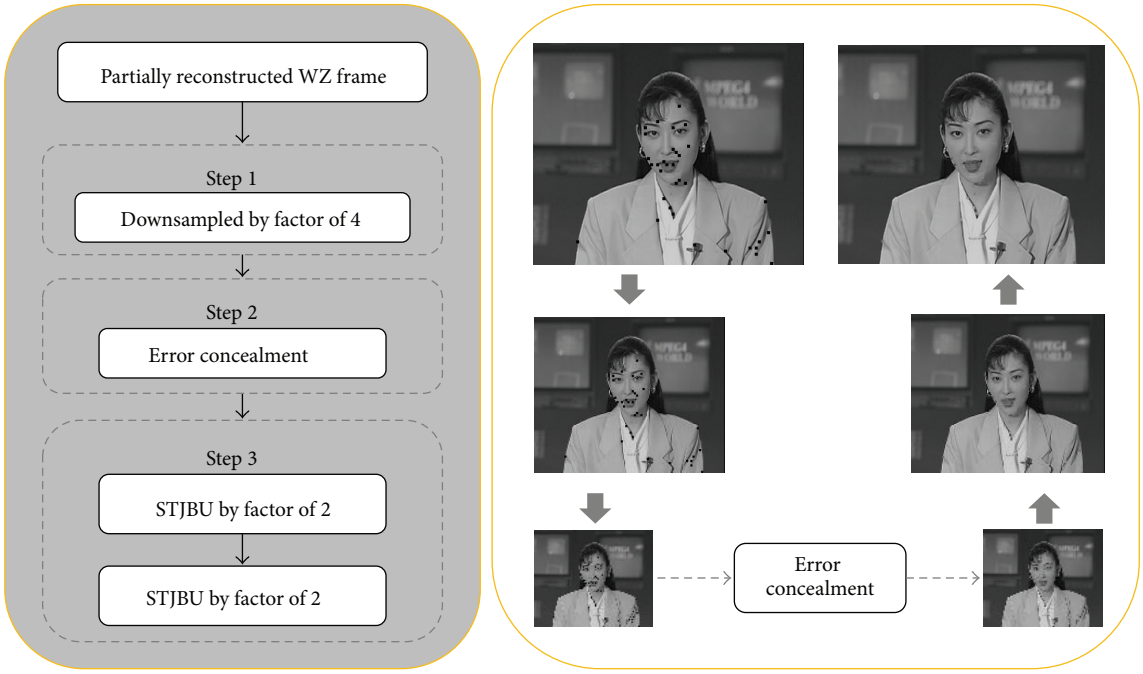


Figure 2: Flowchart of the proposed SI generation scheme.

computation, texture and structure information propagation, and confidence value updating. The initial setting includes target region (Ω) specification, source region (Φ) definition by subtracting the target region from the entire image, and the specification of template window size (Ψ) which is usually set to be slightly larger than the largest distinguishable texture element in the region Φ. Once the parameters are determined, the iterative inpainting process starts automatically until all pixels have been filled. In general, region-filling inpainting incurs high computational complexity, but the processing time can be reduced in the proposed method since inpainting is performed at the lower-resolution WZ frame.

4.3. Spatiotemporal Joint Bilateral Upsampling. After applying EC, the error concealed frame is upsampled using the proposed STJBU method. STJBU is an extension of joint bilateral upsampling (JBU) [24]. JBU is an extension of bilateral filtering [23] and it uses both a domain filter and a range filter to adaptively combine pixels based on both their geometric closeness and their photometric similarity. The difference between JBU and bilateral filtering is that the range filter in JBU is applied to a second guidance image.

In the proposed method, JBU cannot be applied directly because the target reference pixels used for the range filter are not available. In order to solve this problem, the temporal correlation between the consecutive frames is considered. The information in the previous frame is exploited to be used as the second guide image for the range filter. The collocated block in the previous frame is found by boundary matching and it is used as the reference block for the range filter. The scheme of STJBU is shown in Figure 3.

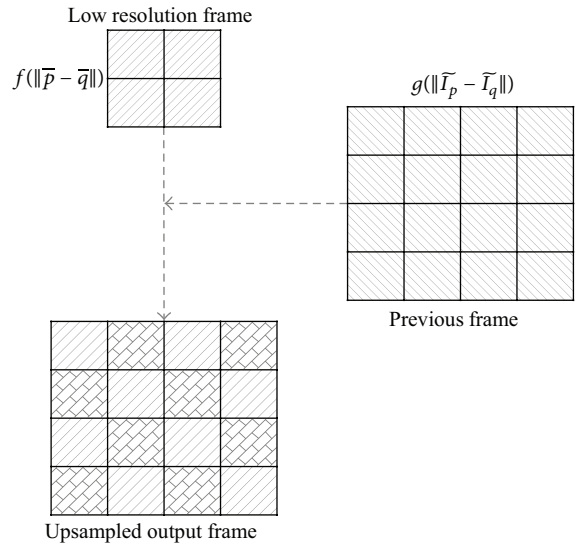


Figure 3: Proposed spatiotemporal joint bilateral upsampling (STJUB) scheme.

Given a previously decoded frame at high resolution $\widetilde{I}_p$ and a low resolution input $S_{\overline{q}}$, which is the error concealed downsampled WZ frame, a spatial filter is applied to the low resolution input $S_{\overline{q}}$, while the range filter is jointly applied on the previous high resolution frame $\widetilde{I}_p$. The upsampled WZ frame $\widetilde{S}_p$ is obtained using the following:

$$\widetilde{S}_p = \frac{1}{k_p} \sum_{\overline{q}} S_{\overline{q}} f\left(\|\overline{p} - \overline{q}\|\right) g\left(\left\|\widetilde{I}_p - \widetilde{I}_q\right\|\right), \tag{7}$$

(a) (b) (c) (d)

Figure 4: Comparison of visual quality of the SI generated by different methods: (a) partially reconstructed 12th frame (WZ frame) of Akiyo sequence, (b) hybrid EC; PSNR = 43.259 dB. (c) Proposed 2; PSNR = 42.726 dB. (d) Proposed 1; PSNR = 44.602 dB.

where p and q denote integer positions in the high resolution frame grid. $\overline{p}$ and $\overline{q}$ denote the corresponding coordinates in the lower resolution frame grid, f is the domain filter centered over $\overline{p}$, g is the range filter centered at the image value at p, and the normalization term k_p is the sum of the $f \cdot g$ filter weights which ensures that the weights for all the pixels add up to one.

5. Simulation Results

To evaluate the performance of the proposed SI generation technique, we conducted experiments using four standard test sequences, Hall Monitor, Akiyo, Mother and Daughter, and Foreman of QCIF size (176 × 144) sampled at 15 frames per second. The luminance component of key and WZ frames and the classification map are taken into consideration for the bitrate computation. The GOP size is adaptive and the maximum GOP size is 5. Only the DC band and first two AC bands of the WZ blocks are refined using the BCH code.

5.1. Comparison of Different Downsampling Methods. First, we compare the performance of four different downsampling methods introduced in Section 4.1. For each test sequence, the first 50 frames are used for simulation. For the experiments, the frames are downsampled using different downsampling methods and then upsampled using the proposed STJBU. The resulting frames are compared to the original frame to calculate the peak-signal-to-noise ratio (PSNR). Table 1 shows the average PSNR value of the four test sequences when different downsampling methods are applied.

Table 1: Comparison of different downsampling methods.

	Bilinear (dB)	Bicubic (dB)	Nearest (dB)	Lanczos (dB)
Hall Monitor	34.495	34.807	33.290	35.362
Akiyo	42.910	43.313	41.591	44.057
Mother Daughter	38.703	39.984	37.256	40.178
Foreman	32.795	33.194	31.137	33.826

As shown in Table 1, the nearest neighbor downsampling algorithm has the lowest computational complexity but it produces the lowest quality. The Lanczos algorithm is much more complex than the other methods but gives the best quality. The processing time of the Lanczos algorithm is almost 10 times higher than the other methods. Bilinear and bicubic downsampling algorithms have lower computational complexity with acceptable output quality. By considering the trade-off between the performance and the processing speed, bilinear downsampling is chosen to downsample the partially reconstructed WZ frames.

5.2. Comparison of the Reconstructed WZ Frame Quality. This section compares the visual quality of the SI generated by the proposed method with that of the hybrid spatiotemporal

Table 2: Different SI generation methods being compared.

Proposed 1	Inpainting as EC and upsampled by STJBU
Proposed 2	BI as EC and upsampled by STJBU
Hybrid EC [19]	Hybrid spatiotemporal EC

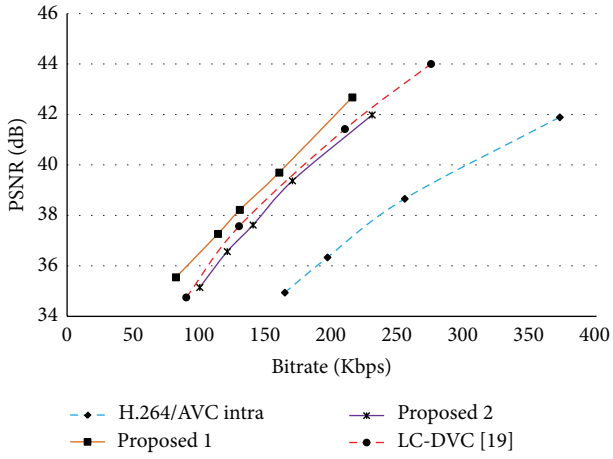


Figure 5: RD performance comparison for Akiyo sequence.

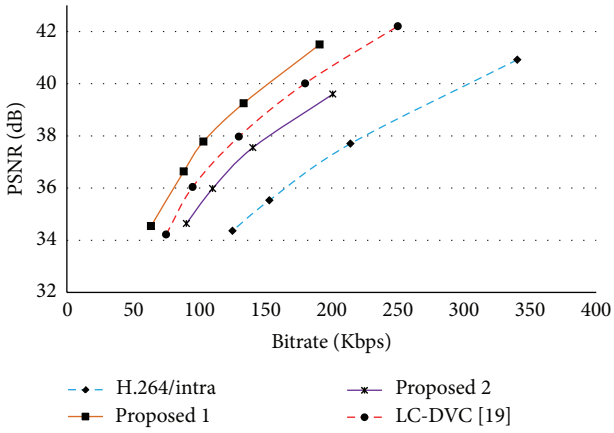


Figure 6: RD performance comparison for Mother and Daughter sequence.

error concealment [19]. Akiyo and Hall Monitor sequences were encoded and decoded using the LC-DVC architecture setting the QPISlice value to 30. For the experiments, we use two different EC techniques along with STJBU. In the following sections, we refer to different techniques as defined in Table 2.

The simulation results shown in Figure 4 illustrate the visual quality of WZ frames obtained by different methods for the Akiyo sequence. As can be seen in Figure 4, the proposed methods produce WZ frames with higher PSNR compared to the ones obtained by the hybrid EC [19]. Specifically, Proposed 1 (inpainting + STJBU) achieves better performance than Proposed 2 (BI + STJBU) because image inpainting is more effective than simple BI in error concealment, while

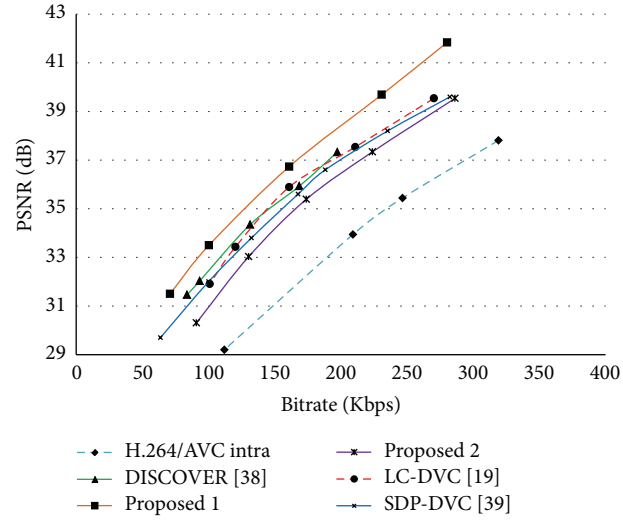


Figure 7: RD performance comparison for Hall Monitor sequence.

it increases the computational complexity. However, since image inpainting is applied to a lower resolution image, the proposed method maintains low computational complexity.

5.3. Comparison of the Rate-Distortion Performance in SI Generation. Next, we encode the first 150 frames of Akiyo, Mother and Daughter, Hall Monitor, and Foreman sequences at various bitrates and compare the RD performances of the proposed methods with that of DISCOVER [38], H.264/AVC intramode, LC-DVC [19], and a recent proposed selective data pruning (SDP)-DVC [39]. It should be noted that DISCOVER uses bidirectional motion estimation for SI generation and uses a feedback channel. Therefore, DISCOVER achieves higher rate-distortion performance than block-based DVC without a feedback channel for high motion sequence but incurs prohibitively long delay.

Rate-distortion (RD) performances of four test sequences are shown in Figures 5, 6, 7, and 8, respectively. Taking the Hall Monitor sequence as an example, it can be seen in Figure 7 that Proposed 1 gives the best RD performance, even better than the SDP-DVC [39]. However, the RD performance of Proposed 2 is lower than that of LC-DVC and DISCOVER. Both Proposed 1 and Proposed 2 perform better than H.264/AVC in intramode with an extremely simple encoder. BI based error concealment used in Proposed 2 enables a very simple encoder, but it reduces SI quality and RD performance.

As shown in Figure 8, the proposed method performs worse for higher motion sequences such as the Foreman sequence. However, it should be noted that DISCOVER uses bidirectional motion estimation for SI generation and uses a feedback channel. Therefore, the DISCOVER codec incurs extremely long decoding time and system delay. Since the proposed method is very simple, has low system delay, and does not require a feedback channel while producing

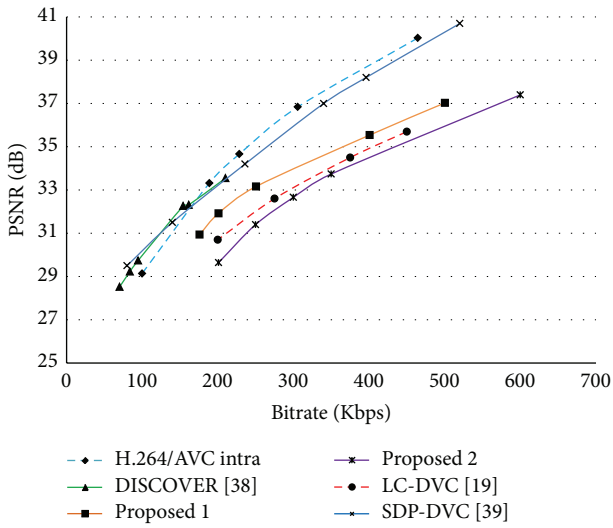


Figure 8: RD performance comparison of Foreman sequence.

a comparable rate-distortion performance to state-of-the-art DVC methods, it can be a promising solution for video applications in resource-limited environments.

6. Conclusion

In this paper, we present a robust STJBU-based SI generation method. The proposed method consists of 3 steps: (1) downsampling of a partially reconstructed WZ frame, (2) SI generation for the WZ blocks in the downsampled WZ frame, and (3) upsampling of the WZ frame using the proposed STJBU algorithm. Results show that the proposed method improves the visual quality of the SI by preserving the edges and improves the RD performance by more than 1 dB in comparison to other DVC architectures. The proposed SI generation method is simple and can be implemented into any exiting block-based DVC architecture. Moreover, with its low complexity and low latency, the proposed method can be a promising solution for video applications in resource-limited environments with a tight delay bound.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the Technology Development Program for Commercializing System Semiconductor funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea). (No. 10041126, title: International Collaborative R&BD Project for System Semiconductor).

References

[1] A. Aaron, S. Rane, E. Setton, and B. Girod, "Transform-domain wyner-ziv codec for video," in *Visual Communications and Image Processing*, vol. 5308 of *Proceedings of SPIE*, pp. 520–528, January 2004.

[2] R. Puri and K. Ramchandran, "PRISM: a new robust video coding architecture based on distributed compression principles," in *Proceedings of the Allerton Conference on Communication, Control, and Computing*, 2002.

[3] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, "The DISCOVER codec: architecture, techniques and evaluation," in *Proceedings of the Picture Coding Symposium (PCS '07)*, 2007.

[4] R. Hänsel, H. Richter, and E. Müller, "Incorporating feature point-based motion hypotheses in distributed video coding," in *Proceedings of the 3rd International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT '11)*, October 2011.

[5] R. Hansel and E. Muller, "Improved adaptive temporal inter/extrapolation schemes for distributed video coding," in *Proceedings of the International Conference on Picture Coding Symposium (PCS '12)*, pp. 213–216, 2012.

[6] S. U. Park, Y. Y. Lee, C. S. Kim, and S. U. Lee, "Efficient side information generation using assistant pixels for distributed video coding," in *Proceedings of the International Conference on Picture Coding Symposium (PCS '12)*, pp. 161–164, 2012.

[7] P. Ren, P. Shi, C. Luo, and Q. Liu, "A new scheme for side information generation in DVC by using optical flow algorithm," in *Proceedings of the 2nd International Conference on Multimedia Technology (ICMT '11)*, pp. 2852–2856, July 2011.

[8] H. V. Luong, L. L. Raket, X. Huang, and S. Forchhammer, "Side information and noise learning for distributed video coding using optical flow and clustering," *Transactions on Image Processing*, vol. 21, no. 12, pp. 4782–4796, 2012.

[9] D. Y. Kim, D. S. Jun, and H. W. Park, "An efficient side information generation using seed blocks for distributed video coding," in *Proceedings of the 28th Picture Coding Symposium (PCS '10)*, pp. 86–89, December 2010.

[10] A. Aaron, S. Rane, and B. Girod, "Wyner-Ziv video coding with hash-based motion compensation at the receiver," in *Proceedings of the International Conference on Image Processing (ICIP '04)*, pp. 3097–3100, October 2004.

[11] E. Martinian, A. Vetro, J. S. Yedidia, J. Ascenso, A. Khisti, and D. Malioutov, "Hybrid distributed video coding using SCA codes," in *Proceedings of the IEEE 8th Workshop on Multimedia Signal Processing (MMSP '06)*, pp. 258–261, October 2006.

[12] B. Macchiavello, R. L. De Queiroz, and D. Mukherjee, "Motion-based side-information generation for a scalable Wyner-Ziv video coder," in *Proceedings of the 14th IEEE International Conference on Image Processing (ICIP '07)*, pp. 413–416, September 2007.

[13] B. MacChiavello, F. Brandi, E. Peixoto, R. L. De Queiroz, and D. Mukherjee, "Side-information generation for temporally and spatially scalable Wyner-Ziv codecs," *Eurasip Journal on Image and Video Processing*, vol. 2009, Article ID 171257, 2009.

[14] B. Macchiavello, D. Mukherjee, and R. L. de Queiroz, "Iterative side-information generation in a mixed resolution wyner-ziv framework," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 10, pp. 1409–1423, 2009.

[15] D. Mukherjee, "A robust reversed-complexity Wyner-Ziv video codec introducing sign-modulated codes," Tech. Rep. HPL-2006-80, HP Lab., 2006.

[16] D. Mukherjee, B. Macchiavello, and R. L. De Queiroz, "A simple reversed-complexity Wyner-Ziv video coding mode based on a spatial reduction framework," in *Visual Communications and Image Processing*, vol. 6508 of *Proceedings of SPIE*, pp. 65081Y1–65081Y12, February 2007.

[17] T. T. Phan, Y. Tanaka, M. Hasegawa, and S. Kato, "Mixed-resolution Wyner-Ziv video coding based on selective data pruning," in *Proceedings of the 3rd IEEE International Workshop on Multimedia Signal Processing (MMSP '11)*, pp. 1–5, November 2011.

[18] Y. Zhang, D. Zhao, J. Zhang, R. Xiong, and W. Gao, "Interpolation-dependent image downsampling," *IEEE Transactions on Image Processing*, vol. 20, no. 11, pp. 3291–3296, 2011.

[19] K. R. Vijayanagar and J. Kim, "Dynamic GOP size control for low-delay distributed video coding," in *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP '11)*, pp. 157–160, September 2011.

[20] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the 1998 IEEE 6th International Conference on Computer Vision*, pp. 839–846, January 1998.

[21] K. J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 650–656, 2006.

[22] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proceedings of the European Conference on Computer Vision (ECCV '11)*, pp. 1–14, 2010.

[23] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *IEEE Transactions on Graphics SIGGRAPH*, vol. 26, pp. 96–100, 2007.

[24] M. Zhang and B. K. Gunturk, "Multiresolution bilateral filtering for image denoising," *IEEE Transactions on Image Processing*, vol. 17, no. 12, pp. 2324–2333, 2008.

[25] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama, "Digital photography with flash and no-flash image pairs," in *Proceedings of the ACM Transactions on Graphics (SIGGRAPH '04)*, pp. 664–672, August 2004.

[26] A. M. Bruckstein, M. Elad, and R. Kimmel, "Down-scaling for better transform compression," *IEEE Transactions on Image Processing*, vol. 12, no. 9, pp. 1132–1144, 2003.

[27] Y. Tsaig, M. Elad, P. Milanfar, and G. H. Golub, "Variable projection for near-optimal filtering in low bit-rate block coders," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 154–160, 2005.

[28] W. Lin and L. Dong, "Adaptive downsampling to improve image compression at low bit rates," *IEEE Transactions on Image Processing*, vol. 15, no. 9, pp. 2513–2521, 2006.

[29] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, 2007.

[30] J. T. Shangguan, Y. L. Li, Y. G. Wang, and H. L. Li, "Fast algorithm of modified cubic convolution interpolation," in *Proceedings of the 4th International Congress on Image and Signal Processing (CISP '11)*, pp. 1072–1075, October 2011.

[31] L.-J. Liu, H. Zhang, and L. Chen, "Bilinear interpolation of geomagnetic field," in *Proceedings of the International Conference on Computer Application and System Modeling (ICCASM '10)*, pp. V2665–V2668, October 2010.

[32] Su Ben-yue and S. Min, "Adaptive algorithm for image interpolation based on blending osculatory rational interpolants," *Computer Engineering and Applications*, vol. 46, no. 1, pp. 196–199, 2010.

[33] J. K. Han and H. M. Kim, "Modified cubic convolution scaler for multiformat conversion in a transcoder," *Optical Engineering*, vol. 43, no. 7, pp. 1596–1608, 2004.

[34] J. Shi and S. E. Reichenbach, "Image interpolation by two-dimensional parametric cubic convolution," *IEEE Transactions on Image Processing*, vol. 15, no. 7, pp. 1857–1870, 2006.

[35] A. Criminisi, P. Perez, and K. Toyama, "Object removal by exemplar-based inpainting," in *Proceedings of the International Conference on Computer Vision and Pattern Recog (CVPR '03)*, vol. 2, pp. II-721–II-728, 2003.

[36] T. C. W. Lei and S. J. Chern, "Partial boundary matching algorithm and spatio-temporal texture synthesis in distributed video coding," in *Proceedings of the 4th International Conference on Innovative Computing, Information and Control (ICICIC '09)*, pp. 353–356, December 2009.

[37] R. Hänsel and E. Müller, "Error locating for plausible Wyner-Ziv video coding using turbo codes," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP '09)*, pp. 1–6, October 2009.

[38] http://www.img.lx.it.pt/~discover/rd_performance.html.

[39] P. T. Tuan, Y. Tanaka, M. Hasegawa, and S. Kato, "Mixed-resolution Wyner-Ziv video coding based on selective data pruning," in *Proceedings of the 3rd IEEE International Workshop on Multimedia Signal Processing (MMSP '11)*, pp. 1–5, November 2011.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 795153, 6 pages
http://dx.doi.org/10.1155/2014/795153

Research Article

Design and Implementation of Software-Based Simulator for Performance Evaluation of Transmission Protocol

Chang-Su Kim, Jong-Il Park, and Hoe-Kyung Jung

Department of Computer Engineering, Pai Chai University, Daejeon 302-735, Republic of Korea

Correspondence should be addressed to Hoe-Kyung Jung; hkjung@pcu.ac.kr

Received 24 December 2013; Revised 6 February 2014; Accepted 25 February 2014; Published 30 March 2014

Academic Editor: Jongsung Kim

We propose a method of software-based transmission protocol simulation for the establishment of ubiquitous infrastructure as a method of securing reliability of a ubiquitous sensor network (USN), which is an important element of ubiquitous infrastructure establishment. For this, we designed a transmission protocol network performance measurement simulator. The network simulator was simplified and, thus, could be used for all types of protocol tests. It enabled transmission control protocol (TCP) and user datagram protocol (UDP) transmission and reception regardless of the number of times of transmission and reception. In addition, we implemented an independent network simulator on a platform that enabled transmission at predetermined intervals according to the number of times of transmission and provided the function of randomly transmitting a defined message.

1. Introduction

Recently, as networking description has been improved; implementing ubiquitous environment is being studied which makes easily obtain information anytime and anywhere you want. Especially, each heterogeneous device has sensor consisting of ubiquitous sensor network (USN) to collect information in ubiquitous environment. For this device, networking function is very important which is reliable between heterogeneous sensor devices.

For a reliable USN, the technology to measure and improve the networking performance between dissimilar sensor devices is very important [1, 2]. However, the network performance evaluation method using smart bits, although it renders a reliable performance evaluation to provide the information needed to improve the software and the hardware, may not be easily introduced to a company having a disadvantaged development environment because the technology requires expensive hardware for the special purpose. In addition, there is an urgent need for a transmission protocol simulator, which can be readily applied to actual work in order to improve the development speed and can be easily used [3].

The most software used in real work-site needs transmission control protocol (TCP) and user datagram protocol (UDP) necessarily performing network communication; so extra test program has been made for all cases to test developing program. Therefore, transfer protocol simulator is highly required to utilize easily and fastening development applying for field work.

The simulator should not be limited to a specific protocol such as session initiation protocol (SIP) but should be applicable to various protocols in which software development is required, such as simple network management protocol (SNMP) [4], in order to maximize its utilization [5].

In this study, a software-based transmission protocol simulator was designed and implemented. The network simulator was simplified and, thus, could be used for all types of protocol tests. It enabled TCP and UDP transmission and reception regardless of the number of times of transmission and reception. Further, an independent network simulator was realized on a platform that provided the function of randomly transmitting a defined message. The software-based network simulator can be generally used in a graphical user interface (GUI) mode without the need for expensive hardware.

2. Analysis of Requirements

As most of the conventional network simulators are specially designed for a particular protocol, they may be used only when the protocol has been accurately analyzed. Moreover, the use of conventional network simulators is limited because of their high price, which mostly runs into tens of thousands of US dollars [6–9].

Therefore, in this study, we analyzed the operating methods of various conventional simulators and designed a simulator that can be used for testing all the different protocols, allowing TCP and UDP transmission and reception regardless of the number of times of transmission and reception, enabling transmission at predetermined intervals according to the number of times of transmission, and providing the function of randomly transmitting a defined message. The designed network simulator has the following functions:

(i) functioning independently on the platform,

(ii) allowing for random transmission of a defined message,

(iii) not being limited to a specific protocol,

(iv) allowing for choosing either server or client.

3. Simulator Design

This section describes the design and realization of the proposed simulator. The basic block diagram of the simulator is shown in Figure 1. The SIP Tester module comprises three modules, namely, the TCP module, the UDP module, and the Test Msg module.

First, the SIP Tester module is the core module controlling the TCP module, the UDP module, and the Test Msg module. Second, the TCP module controls TCP transmission and reception, and the UDP module controls UDP transmission and reception. Third, the Test Msg module provides the text transmitted by the TCP module and the UDP module.

3.1. Simulator Environment Design. Table 1 shows the basic environment for the transmission protocol simulator.

3.2. Simulator Default Setting. For the simulator to test network traffic, the CUI was designed to reflect the following requirements. In the default setting, a server or a client that plays the main function in the network can be set (Table 2). The records of the transmitted and the received data can be viewed in order for the simulator to use the data transmission and reception information. A message to be transmitted can be conveniently prepared and tested. With respect to the performance information, the transmission interval, the number of times of transmission, and the random message generation can be set.

4. Simulator Implementation

The implemented simulator largely comprises a UDP server, a UDP client, a TCP server, and a TCP client. Among them, the UDP server and the TCP server perform the echo server functions, while the UDP client and the TCP client perform reception functions.

4.1. UDP Server. Figure 2 shows the execution screen of the UDP server. The UDP server sets its own IP and port number and determines whether to function as an echo server by using a radio button. When the function is executed, the Start

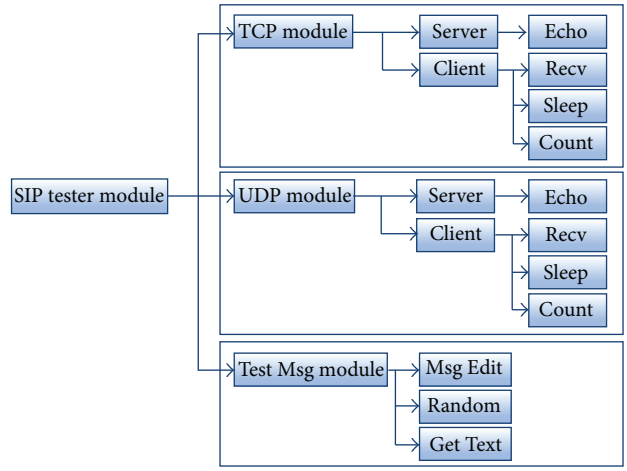


Figure 1: Simulator block diagram.

Table 1: Simulator design environment.

Item	Environment
Development platform	Windows XP Professional Service Pack 3
Development tool	Microsoft Visual Studio 2008
Application range	Windows, Linux compatible
Protocol	TCP, UDP
SC model	Allows for the selection of server-client
Message	Allows for the determination of transmission message by selecting file
Transmission speed	Allows for millisecond-unit speed control and MAX speed
Transmission amount	Allows for the generation of 1 to 2^{32} data
Random data	Allows for the generation and transmission of random data

Table 2: List of simulator default settings.

Item	Content	Note
Server, client	Server, client	Radio button
Server	Echo server function selection	Check box
Client	Reception function selection	Check box
IP, port	IP, port setting	Text box
Start, stop	Intuitive start and stop	Button, activated/inactivated

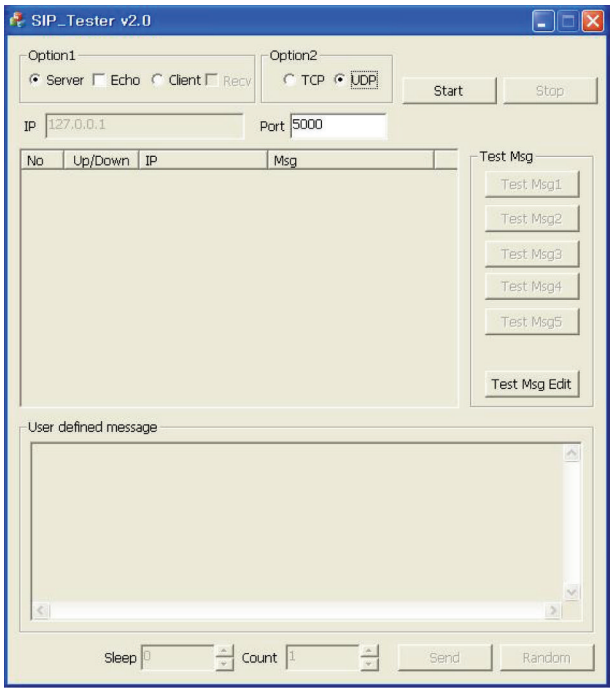


Figure 2: UDP server execution screen.

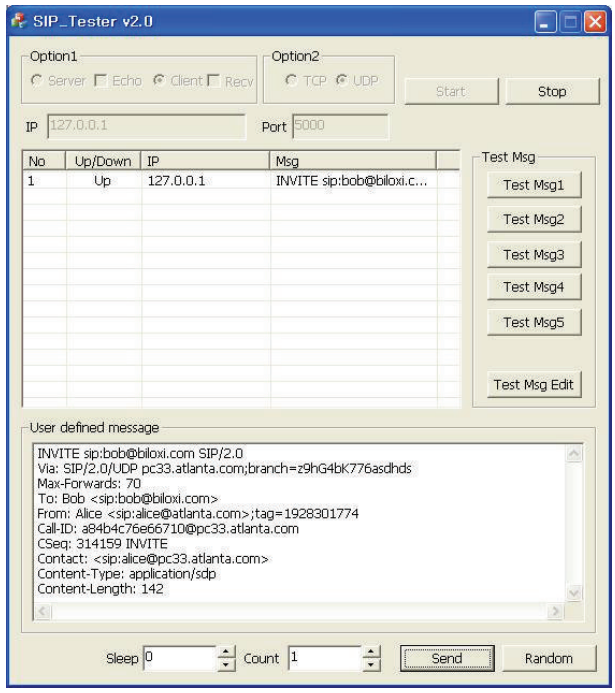


Figure 3: SIP INVITE message transmission.

button is inactivated and data reception starts by using the IP and the port that have been set in advance. This implies that the UDP server functions as a UDP socket server by using the set values.

As the UDP server can intuitively recognize whether it is normally functioning, the log for data transmission and reception is managed as table-type record information at the center of the screen in order to inform the user regarding the normal function of the simulator. As the UDP server has the immediate processing function enabling the optimum reception, the UDP server has the immediate processing mechanism without an additional setting for the Sleep or Count Information.

4.2. UDP Client. The UDP client is a simulator function that is the opposite of the UDP server. The UDP client transmits a message to the server waiting for data reception. The UDP client execution screen includes a setting window in which the IP and the port of the server waiting for reception are set.

The UDP client provides a convenient interface to the user as it has a screen composition similar to that of the UDP server. However, different from the server execution screen, there are five message selection buttons on the right, and the "Test Msg Edit" button is additionally activated. These are necessary for the client to create the data to be transmitted to the server. Although the user can create and transmit the required message each time the user needs it by using the "User defined message," the user may create a data file in advance and then test the data conveniently. An SIP message test is performed as follows.

4.2.1. Direct Input by User. When a user directly creates an SIP message and transmits the message, the user creates "User defined message" and transmits it, as shown in Figure 3. This method is useful when the transmission of a short message is tested, but the method is inefficient when the test needs to be performed repeatedly.

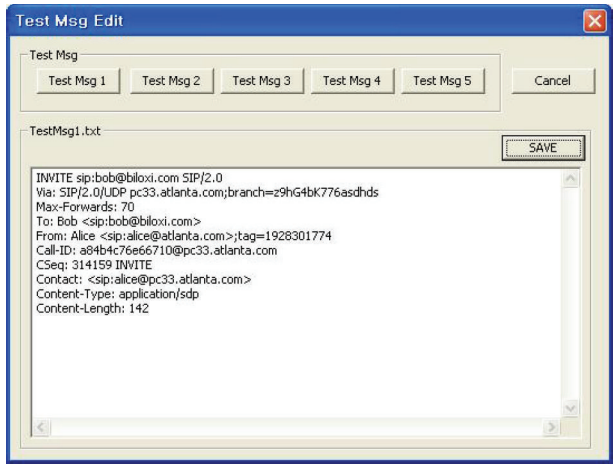


Figure 4: Defining transmitted message in advance.

4.2.2. Test Using File. It may not be efficient for a user to input an input message each time. To resolve the inconvenience, a test message can be prepared in advance. For this, the user has to click the "Test Msg Edit" button, and, then, a dialog box appears. Figure 4 shows the procedure for saving a message. The message to save is put into the text window. When the user clicks the "SAVE" button, the notification window "Save OK" appears.

Figure 5 shows an example of generating a message 15 times by using a predefined file.

4.2.3. Random Transmission. The function to randomly transmit various types of messages is essential in a software development process. Figure 6 shows the screen after the data have been transmitted. A hundred messages are randomly generated from selected objects to enable the simulation of various types of data.

4.2.4. Transmission Speed. The simulator may transmit data in a millisecond-unit interval. To prevent another transmission while transmitting data at a predetermined interval, the "Send" button is inactivated and the data are transmitted at the defined interval.

4.2.5. Load Test Using Transmission Data Amount. For the load test, which is an essential test item for a network simulator, the amount of transmission data can be set. The number of amount of transmission data that may be selected is 2^{32}.

4.3. TCP Server. As TCP is a connection-oriented transmission protocol, a server needs a procedure for waiting for a socket connection, which includes a socket connection from a client or finishing off a socket connection when a client is finished. For this, it is necessary to express the connection status information of a client. As the TCP server was realized by using the same GUI as that of the UDP server, a user may intuitively understand the information. The method of using the TCP server is the same regardless of the TCP/UDP socket. As shown in Figure 7, the biggest difference between the TCP server and the UDP server is the method of expressing the connection information. The TCP server expresses the status information by means of the "notice" value in the Up/Down domain.

4.4. TCP Client. In contrast to UDP, the TCP client firstly attempts the connection for the communication with a server, preferably with the TCP server. As in the case of the UDP client, the connection and the finish of the connection are performed simply by using the "Start" and "Stop" buttons, respectively. Figure 8 shows the normal data transmission by the TCP client without a problem in the connection between the servers.

4.5. Verification between Dissimilar Platforms. Thus far, the simulator has been used as a server and a client, and the simulator functioning was verified by using the same system for convenience. However, for the actual functioning of a simulator, the data transmission between dissimilar platforms should be tested to check out whether the simulator can be used in an actual development process.

For this, a UDP server was created on Linux and the simulator suggested in this study was used for verifying the functioning. Each of the UDP and TCP transmission protocols was tested as a server and as a client. Herein, only the test performed with UDP is described.

To test the UDP server, the simulator suggested in this study was used as the client. Figure 9 shows the data transmission procedure using the simulator as a client in the process of testing the Linux-Windows network.

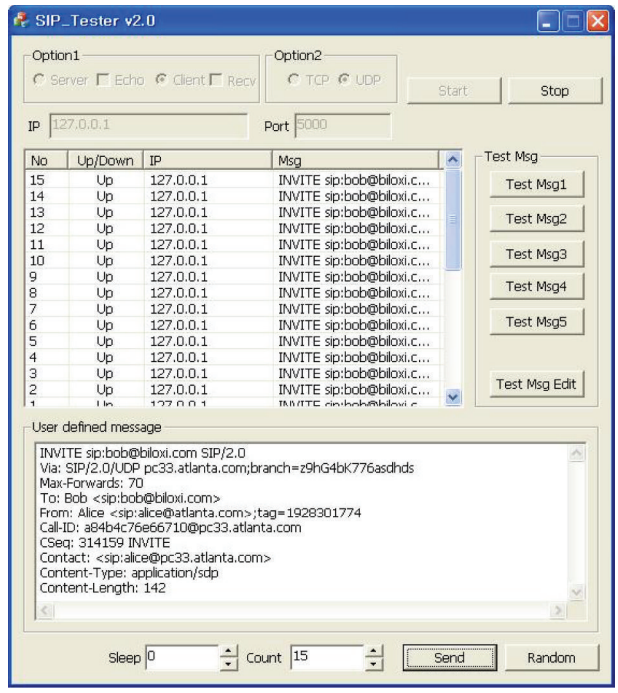


Figure 5: Mass transmission of message by using predefined file.

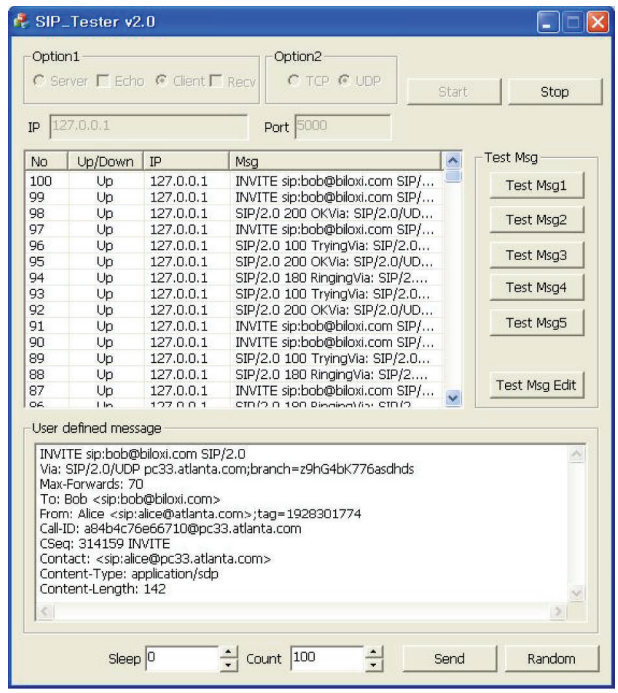


Figure 6: Random data transmission screen.

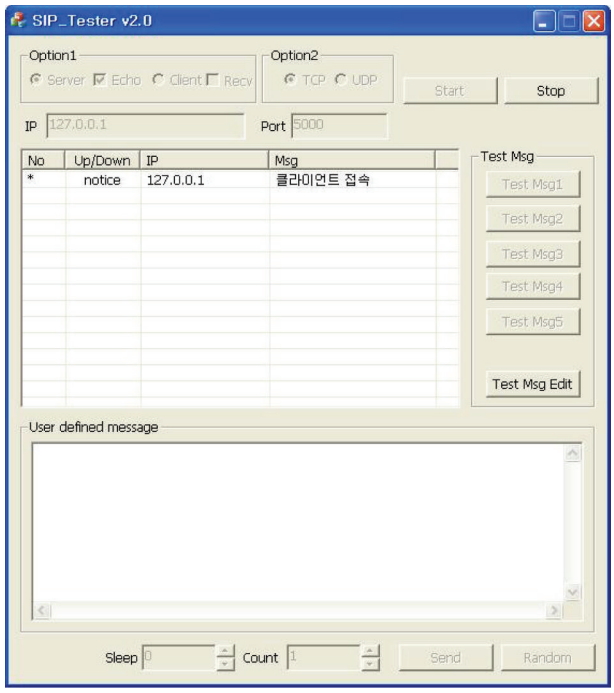


FIGURE 7: TCP server client connection screen.

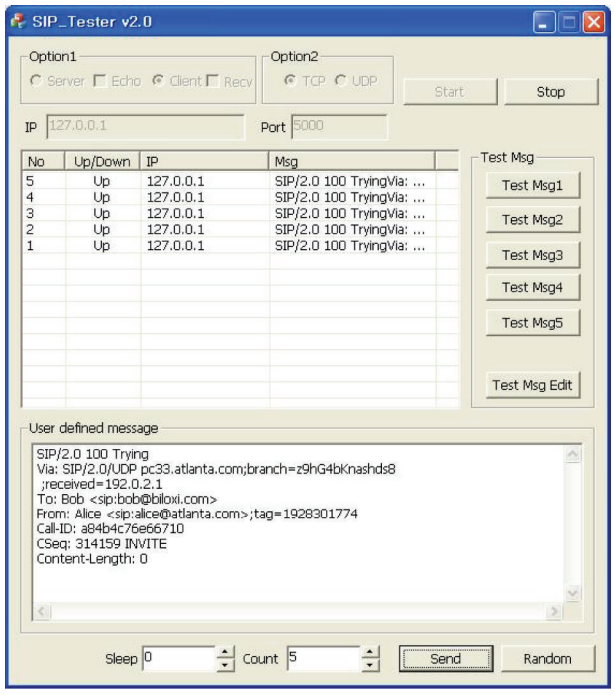


FIGURE 8: TCP client data transmission.

4.6. Measurement of Network Performance. A software developer often confronts a situation where the network performance should be measured. In such a situation, data may be normally processed by adding a basic algorithm for performance measurement to the code under development.

Such a procedure was performed repeatedly with respect to the same data; the data were transmitted by the simulator

FIGURE 9: Simulator for the test of data transmission between dissimilar platforms (client).

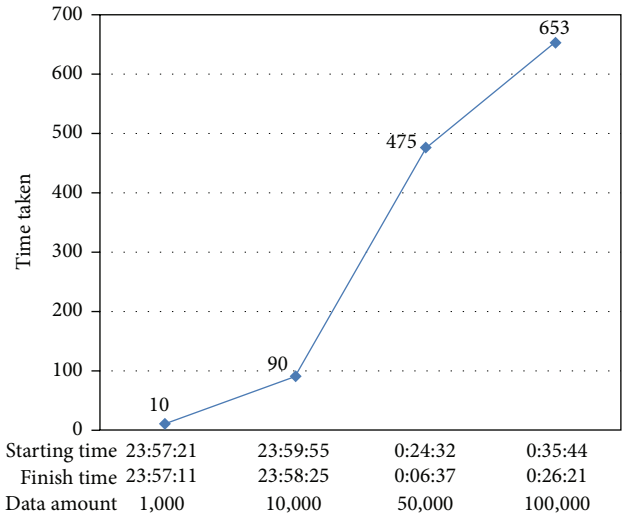


FIGURE 10: Measurement of time taken for transmitting the same data.

for 1000, 10000, 50000, and 100000 times and processed by a Linux server. Figures 10 and 11 show the correlation between the number of times of the data transmission and the time taken for the transmission.

The two figures given above indicate that the transmission of random data has a delay of approximately several percentage points in comparison with the transmission of the same data. The result is consistent with the fact that, during a network software development, the transmission performance is excellent when the same data are processed or the data of the same length are repeatedly processed. The

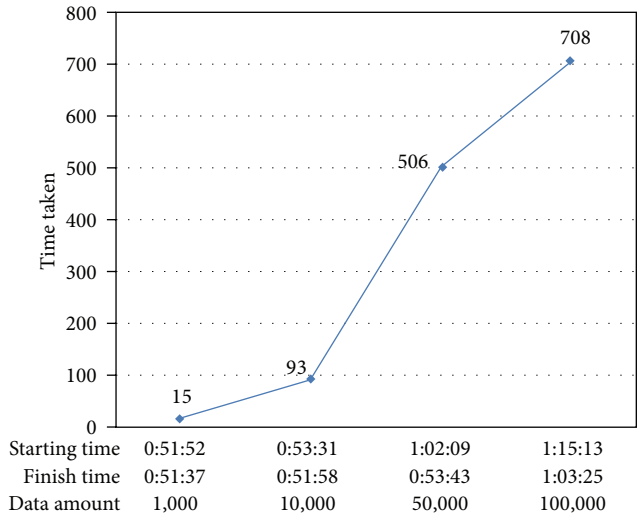


Figure 11: Measurement of time taken for transmitting random data.

delay can be attributed to the data processing delay by the simulator or by the Linux server. Irrespective of what the cause of the delay may be, the result indicates that the simulator functions normally. The verification of the simulator showed that the simulator can be used as a more convenient tool for a developer to develop the network software.

5. Discussion and Conclusions

In this study, a software-based network simulator was designed, realized, and verified. The transport layer simulator suggested in this paper supports both the TCP and the UDP protocols, functions as a server and a client for each of the protocols, and allows for the testing of the server and the client. Further, the simulator was designed and realized so that an echo server, which is the most basic element of socket communication, is built-in for the simulator to be generally used by a beginner who studies socket communication or even by a high-ranking engineer who develops advanced software.

Approximately 4 billion data may be transmitted for a test. Data transmission with a similar platform such as Linux was also verified. Further, the simulator proposed in this paper provides the advantage of convenient use as it is developed on the basis of GUI. The simulator also provides a functional advantage in the fact that a user can separately define an SIP message to measure the network performance. In particular, the performance of all text-based UDP and TCP protocols can be measured.

Functions to evaluate the performance of various protocols such as SCTP may need to be added to the simulator in future studies.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] G. Chen, J. Branch, M. Pflug, L. Zhu, and B. Szymanski, "SENSE: a wireless sensor network simulator," *Advances in Pervasive Computing and Networking-*, pp. 249–267, 2005.

[2] A. Howard, M. J. Mataric, and G. S. Sukhatme, "Mobile sensor network deployment using potential fields: a distributed, scalable solution to the area coverage problem," in *Distributed Autonomous Robotic Systems 5*, pp. 299–308, Springer, Tokyo, Japan, 2002.

[3] N. Baldo, F. Maguolo, M. Miozzo, M. Rossi, and M. Zorzi, *Ns2-MIRACLE: A Modular Framework form Multi-Technology and Cross-Layer Support in Network simulator 2*, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, Belgium, 2007.

[4] M. Schoffstall, J. Case, M. Fedor, M. Schoffstall, and J. Davin, A simple network management protocol (SNMP). RFC, 1157, 1990.

[5] S. Floyd and Yokohama IETF TSVWG, HighSpeed TCP for Large Congestion Windows, 2002.

[6] W. S. Lee and H. W. Lee, "Performance evaluation of coordinated multi-point transmission and reception in indoor mobile communication systems," *Journal of Information and Communication Convergence Engineering*, vol. 11, no. 3, pp. 167–172, 2013.

[7] E. M. Nahum, J. Tracey, and C. P. Wright, "Evaluating SIP server performance," *ACM SIGMETRICS Performance Evaluation Review*, vol. 35, no. 1, pp. 349–350, 2007.

[8] D. Pesch, M. I. Pous, and G. Foster, "Performance evaluation of SIP-based multimedia services in UMTS," *Computer Networks*, vol. 49, no. 3, pp. 385–403, 2005.

[9] Z. S. Sensinode, K. Hartke, and C. Bormann, Constrained Application Protocol(CoAP), 2013.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 142610, 7 pages
http://dx.doi.org/10.1155/2014/142610

Research Article

Practical Electromagnetic Disturbance Analysis on Commercial Contactless Smartcards

Jaedeok Ji,[1] Dong-Guk Han,[2] Seokwon Jung,[3] Sangjin Lee,[4] and Jongsub Moon[4]

[1] *Information Technology Team, Korea Testing Certification, 22 Heungan-daero 27 beon-gil, Gunpo-si, Gyeonggi-do 435-823, Republic of Korea*
[2] *Department of Mathematics, Kookmin University, Jeongneung-Ro 77, SeongBuk-Gu, Seoul 136-702, Republic of Korea*
[3] *Department of Information Security, Mokpo National University, 1666 Youngsan-ro, Chenggye-myeon, Muan-gun, Jeollanam-do 534-729, Republic of Korea*
[4] *Graduate School of Information Security, Korea University, Anam-dong 5, Seongbuk-Gu, Seoul 136-701, Republic of Korea*

Correspondence should be addressed to Dong-Guk Han; christa@kookmin.ac.kr

Received 25 November 2013; Accepted 26 February 2014; Published 27 March 2014

Academic Editor: Jongsung Kim

Contactless smart cards are being widely employed in electronic passports, monetary payments, access control systems, and so forth, because of their advantages such as convenience and ease of maintenance. In this paper, we present a new side-channel attack method for contactless smart cards. This method exploits the information leakage stemming from electromagnetic disturbances (EMD). We also made a convenient and low-cost EMD reader board that performs side-channel attacks on contactless smart cards. In order to demonstrate that EMDs can become another information-leakage side channel, we have carried out side-channel analysis on a commercial contactless smart card that performs 128-bit ARIA encryptions, and we have been able to successfully find all 16 bytes of the ARIA key from the target device. From our experimental results, we conclude that the proposed EMD analysis yields better results than the conventional power analysis.

1. Introduction

Side-channel attacks exploit information leakage stemming from variations in physical quantities such as timing, power consumption, and electromagnetic (EM) radiation [1–3]. By measuring the variations in processing time, power consumption, and EM radiation during the execution of the target algorithm and correlating these variations with the data being manipulated, the secret key can be obtained from the target device. From among all of the information leakage channels, the information leakage from the electromagnetic radiation is the best for use in side-channel attacks on contactless smart cards because the other information leakage channels are very limited. An EM side-channel analysis exploiting the information leakage stemming from EM radiation has several advantages such as higher signal-to-noise ratio and the ability to bypass the power analysis countermeasures. However, considerable effort is required to measure and analyze the electromagnetic radiation, such as separating the smart card chip radiation from the antenna radiation and filtering the carrier signal [2, 4].

In this paper, we present a new side-channel attack method that exploits the information leakage due to electromagnetic disturbances (EMD) in contactless smart cards. EMDs are an unwanted form of load modulation caused by dynamic load changes. These dynamic load changes are caused by the switching operations of the contactless smart card's internal digital circuits, especially during cryptographic operations [5]. Under ISO 14443, a contactless smart card is also referred to as a Proximity Integrated Coupling Circuit (PICC), whereas the card reader is called a Proximity Coupling Device (PCD). The PICC consumes the energy generated by the PCD in order to gather the energy needed to operate; this energy consumption has the same reactive effect on the PCD as a voltage amplitude modulation. This effect is used to transfer data from the PICC to the PCD by changing the resistive load in the PICC according to the transmitted data bit (0 or 1). This resistive load change in the PICC is also

generated by the energy consumption required for processing cryptographic operations or EEPROM programming. This load change (variations in the internal power consumption of the PICC) causes an unwanted form of load modulation, and the EMD becomes another information leakage channel vulnerable to side-channel attacks on contactless smart cards.

Compared to conventional EM radiation, (1) EMDs can be measured more easily and observed in real time through a PCD demodulation process and (2) EMD information leakage can be used to successfully attack a contactless smart card. In order to demonstrate that EMDs are a viable side channel, we performed side-channel analysis on a commercial contactless smart card, which utilizes 128-bit ARIA, a block cipher designed in 2003 by South Korean researchers. In 2004, the Korean Agency for Technology and Standards selected it as the standard cryptographic technique [6]. We could successfully obtain the entire 16-byte key used in the first round of the ARIA encryption by using the EMD side-channel analysis with 50,000 traces. These results show that the EMD side-channel analysis outperformed the more conventional power analysis technique.

2. Previous Studies on Contactless Smart Cards

There are three basic types of contactless smart cards: close-coupling cards defined in ISO 10536, proximity-coupling cards defined in ISO 14443, and vicinity-coupling cards defined in ISO 15693. Most commercial contactless smart cards are ISO 14443 compliant. In this study, we focused mainly on a PICC proximity coupling card working at a carrier frequency of 13.56 MHz. Because of several advantages such as convenience and easy maintenance, an increasing number of contactless smart cards are being deployed for various applications such as in electronic passports, for monetary payments, and in access control systems. With this increased use of contactless smart cards, side-channel attacks on contactless smart cards have increased and so these attacks have become an area of intense study; several important studies related to this topic have been published. Carluccio et al. [2] performed an EM attack on a contactless smart card. In order to minimize the adverse influence of the field of RFID readers on the measurements, they separated the chip and antenna radiations and measured the EM radiation by using a near-field magnetic probe, which was placed perpendicular to the chip surface; however, the attack was not successful. Hutter et al. [3] published the first reported results of successful EM attacks on hardware and software AES implementation in RFID tag prototypes. In Hutter's experiment, the target device was not a commercial product but a self-made RFID prototype in which the analog front end, that is, the antenna and the rectifier circuit, was separated from the digital circuit. In commercial products, the analog and digital components are fixed together. The configuration of the target device in Hutter's experiment leads to more EM emissions than in the commercial products and the effect of the field of RFID readers on the measurements could be easily eliminated. The main approach used to perform side-channel attacks in these experiments was indirectly measuring the power consumption via the EM field of the device by using a magnetic near-field probe. However, this was not easy because the carrier signal of the reader was much stronger than the field of the contactless device. In order to minimize the adverse influence of the carrier frequency on the measurements, it is necessary to use complex active and passive analog filters. Therefore, in order to ameliorate these EM measurement setup difficulties, it is necessary to develop more convenient and efficient side-channel analysis methods.

Kasper et al. [4, 7] proposed an analog demodulator specifically designed for filtering the signal measured by an EM probe. In Carluccio and Hutter's works, the EM radiation from the chip surfaces was directly measured using an EM probe. However, the EM radiation from the chip surfaces was much weaker than that of the field signal of the reader. Thus, the attack was not successful. Hutter's approach was successful, but the target device was not a commercial contactless product. To successfully attack contactless cards, Kasper et al. exploited the information leakage that stemmed from amplitude modulation (load modulation) of the 13.56 MHz field of the reader. In [4, 6], the authors assumed that the power consumption of the inductive-coupled smart card leads to very weak amplitude modulation of the field generated by the reader. To extract the weak information leakage from the amplitude modulation signal, the authors used an incoherent demodulation approach. In Kasper's experiment, the EM radiation was measured by an EM probe, while the target was executing an operation. The measured raw signals were processed using an analog demodulator and filter. For this purpose, the authors designed a custom analog circuit for amplification, rectification, and filtering of the raw analog signal. The authors demonstrated the side-channel vulnerability of the Mifare DESFire MF3ICD40 contactless smart card using this approach.

3. EMD Analysis on Contactless Smart Cards

When the proximity cards inductive coupled with readers execute their operations, the variations in the internal power consumption of the proximity cards cause physical changes such as EM radiation from the chip surface and amplitude modulation of a carrier field of the reader inductive coupled with the proximity card. These changes are exploitable in the side-channel analysis.

As described above, in Carluccio and Hutter's experiments, the information leakage from the EM radiations was measured using an EM probe. However, the isolation of very weak EM radiation from a strong carrier field is difficult. Hence, performing the side-channel analysis on proximity smart cards with EM radiation emitted from the chip is challenging. On the other hand, in Kasper's experiment, the authors have studied information leakage caused by amplitude modulation of the carrier field. Their incoherent demodulation approach apparently improved the efficiency of SCA on proximity card but there is no major enhancement in the measurement setup; it is still necessary to use additional complex hardware analogue filter.

Similar to Kasper's study, we have also adopted a demodulation approach but have focused more on the reader demodulation process. The extent of information leakage caused by amplitude modulation could be measured and processed more easily through an ISO 14443 compliant reader demodulation process without any EM probe or additional hardware filter. Proximity cards communicate through intentional load modulation, but variations in the power consumption of the card also create unintentional load modulation [8]. This kind of unintentional load modulation is defined as EMD according to ISO standards [5]. The variations in some cards are so large that some readers detect false card responses. The allowable EMD levels and handling method for the ISO 14443 compliant reader are standardized by ISO [5, 9]. This means that the amplitude modulation that contains the information leakage of proximity cards can also be measured and processed in the reader.

Because the levels of load modulation can be measured through the demodulated signal after IQ demodulation inside the reader [10], the unwanted load modulation EMD can be measured from the demodulated analog signal after the IQ demodulation. This demodulated signal can be also measured easily using the debugging functionality of the commercial contactless smart card reader IC chip. In our experiment, the amplitude load modulation was measured and processed during the demodulation process of the reader. For this purpose, we have designed an EMD measurement board, which is a slightly modified version of the common ISO/IEC 14443 reader. In Kasper's study, the analog demodulated signals were measured to maximize the vertical resolution of the measurements and capture all the relevant information [4, 7]. In our study, we have also measured the amplified demodulated analog signals by configuring the internal register setting of the chip in order to minimize the loss of information. This approach makes it possible to ameliorate the EM measurement setup difficulties as well as to improve the efficiency of SCA on proximity card. In this study, we have used a commercially available smart card, which provides dual communication interfaces (ISO 7816 contact and ISO 14443 contactless) and performs 128-bit ARIA encryption implemented in software without side-channel countermeasures.

4. EMD Measurement Setup

In a side-channel attack on a contactless smart card, the data- and operation-dependent power consumption is generally measured indirectly via the EM field of the target device by using an EM probe [3, 4]. In our approach, instead of measuring the EM field of the target device, the data- and operation-dependent power consumption is measured indirectly via the EMD level of the target device by using an EMD board.

Figure 1 shows the schematic of the EMD measurement setup. The PICC consumes energy generated by the PCD in order to gather the energy required for its operation. In general, the power consumption of the PICC (P) depends on the Hamming weight (the number of $1'$s in a binary sequence) of the data being processed; it has been confirmed that this model is suitable for smart cards. The power consumption and the Hamming weight of the data being manipulated at a given instance show a linear relationship in the Hamming weight model. P can be expressed as $P = \varepsilon \cdot \mathrm{Hw}(x) + L + N$, where $\mathrm{Hw}(x)$ is the Hamming weight of the intermediate data x; ε is the incremental amount of power for each extra 1 in the Hamming weight; L is the additive constant portion of the total power; and N is the noise [11]. Note that N is assumed to be independent and have a zero mean. Therefore, P is proportional to the Hamming weight of the intermediate processed data x:

$$P \propto \mathrm{Hw}(x). \tag{1}$$

As described in [12], the load resistance R_L is an expression for P in the PICC and the transformed impedance Z_t in the antenna coil of the PCD is proportional to R_L. The voltage U_L at the antenna is also proportional to Z_t. Therefore,

$$U_L \propto \mathrm{Hw}(x). \tag{2}$$

The change in U_L caused by the dynamic change in R_L induces an arbitrary amplitude modulation at the PCD antenna. The dynamic change in R_L is caused by the variations in P during the PICC operation. This arbitrary load modulation is defined as the EMD [5, 13]. Therefore, using (2) and the fact that the change in U_L induces the EMD, we can conclude that the EMD is proportional to the Hamming weight of the intermediate processed data x:

$$\mathrm{EMD} \propto \mathrm{Hw}(x). \tag{3}$$

From (1)–(3), we conclude that the data- and operation-dependent power consumption can be measured indirectly via the EMD level of the target card.

The methods for measuring the EMD level of a PICC can be implemented using a spectrum analyzer or a vector signal [10]. In order to demodulate the received RF signal from the card, most commercial readers first perform quadrature demodulation of the 13.56 MHz carrier signal using an IQ demodulator [10]. The IQ demodulator converts an RF input into two intermediate outputs with a 90° phase difference [14]. The demodulated signal is amplified by an amplifier and digitized by the digitizer circuit.

Because the EMD is a kind of load modulation whose levels can be measured through the demodulated signal after IQ demodulation inside a reader, the EMD can also be measured from the demodulated analog signal after IQ demodulation. This demodulated signal can be also measured by using the debugging functionality of commercial contactless smart card reader IC chips. In our experiment, instead of using the complex measurement setups, we designed an EMD reader board, which is a slightly modified version of the common ISO/IEC 14443 reader. The EMD reader board consists of a commercial ISO/IEC 14443 reader IC chip, an RF antenna, and a CPU.

The board is controlled by an ST STM32F103 microcontroller and provides an ISO 14443 compliant RF front end with an NXP MFRC531 reader IC chip and an RF antenna. The board is equipped with a USB interface for

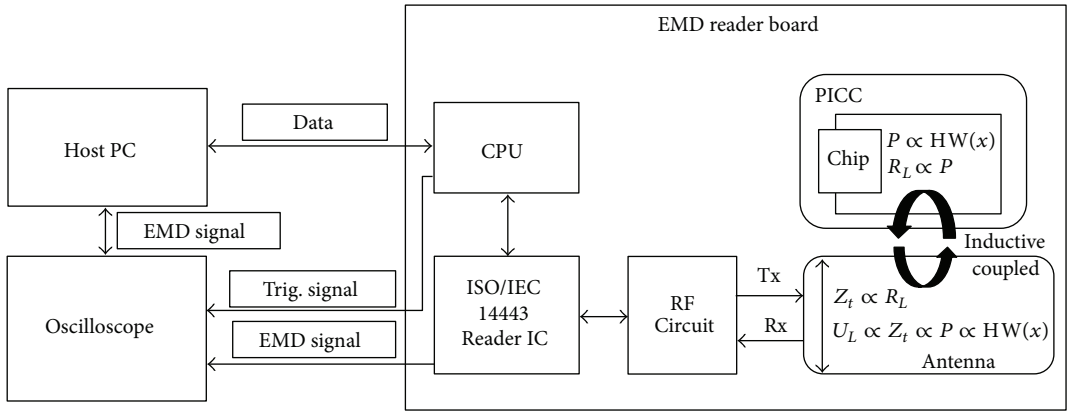


Figure 1: Schematic of the EMD measurement setup.

communication with a host PC. A USB connector and an external 5.0 V DC power supply provide power for operation of the board. The power supply is physically separated to reduce the power supply noise. Two BNC connectors are placed on the reader board for measuring the EMD signal: an EMD port for the "EMD signal" and a TRG port for the "TRIG signal," as shown in Figure 1. The microcontroller in the reader board communicates with a host PC and controls the reader IC chip to perform ISO 14443 compliant communication with the card. A trigger-timing signal, which signals the measurement equipment about when to begin acquiring EMD traces, is generated by the microcontroller through the TRG port (Figure 1).

The reader IC chip used in the EMD board basically performs modulation and demodulation for passive contactless communication. The signal received through the antenna of the reader board is forwarded to the reader IC chip through the RF circuit. Quadrature demodulation of the 13.56 MHz carrier signal is performed using the IQ demodulator. The demodulated signal is filtered and amplified within the reader IC chip and digitized in the digitizer circuit.

The reader IC chip is also responsible for measuring the EMD signal of the card. As described above, the EMD can be measured from the demodulated analog signal after IQ demodulation in the reader. Because the reader IC chip used in the EMD reader board has a built-in monitoring functionality, that is, an internal reference voltage and an amplified demodulated analog signal, the EMD signal can be easily measured using the amplified demodulated signal monitoring functionality of the reader IC chip. The amplified demodulated analog signal after the IQ demodulation in the reader is routed to the auxiliary output pin of the reader IC chip using the internal register configuration settings of the chip. The measured EMD signal is output from the reader board through the EMD port placed on the auxiliary output lines of the reader IC chip. The EMD signal, shown in Figure 1, is the demodulated analog signal, which is stored on the host PC after passing through the oscilloscope.

In order to increase the quality of the EM measurements, we have adjusted and optimized the external factor that influences the measurement of the load modulation. From our experiments, it was observed that the load modulation level is inversely proportional to the field strength of the reader.

To change the field strength of the reader, the conductance of the reader antenna was adjusted. In general, the field strength of the reader depends on the conductance of the antenna that can be adjusted using the internal register configuration setting of the chip. By configuring the internal register of the chip, the field strength of the reader was adjusted to a value that was as low as possible yet sufficient to operate the proximity card. This approach was also the most effective and inexpensive for avoiding the adverse influence of the reader field.

Using those approaches, we can measure the EMD during the cryptographic operation of the PICC without additional specific signal processing. Figure 2 shows a picture of the actual experiment configuration. A host PC controls the EMD reader board and the oscilloscope. From the host PC, a smart card application protocol data unit (APDU) and board control commands are transferred into the EMD board through a USB port. The board control commands are used to initialize the connection between the host PC and the EMD board, to set or adjust the timing of trigger signal, and configure the parameters such as the alternative magnetic field strength generated by the reader.

The EMD reader board is responsible for measuring the EMD signal of the card during its operation. The received signal, including the unwanted load modulation generated by the PICC at its antenna, is forwarded to the reader IC chip of the board through the RF circuit. The received signal is demodulated and amplified within the reader IC chip of the PCD. This amplified demodulated analog signal is routed to the auxiliary output pin of the ISO/IEC 14443 reader IC chip by using the internal register configuration settings of the chip. Then, the measured EMD signal is passed through the oscilloscope and stored on the host PC.

5. Attacks on Dual-Type Smart Cards with the S/W ARIA Implementation

We performed power and EMD side-channel analyses on a commercial dual-type smart card by using the Correlation Power Analysis (CPA) method [15]. CPA exploits the

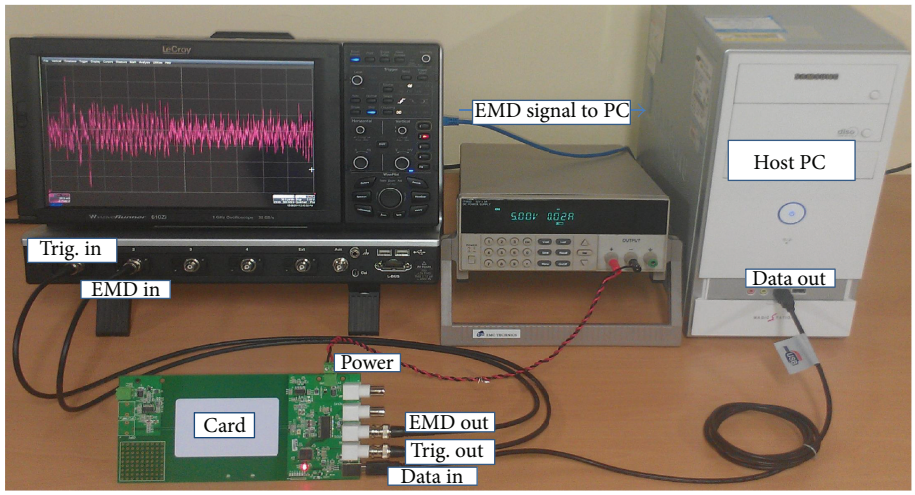


Figure 2: Image showing the experimental configuration.

correlation between the power consumption of the target device and its power consumption model. To perform CPA, the attacker measures the power consumption of the target device and then calculates the power estimates of the target from different ciphertexts using a predicted partial key and a power model. Assuming that the power model is valid, the power estimates and the measured power traces are correlated if the partial key prediction is correct. From the key hypothesis, the attackers can find the correction key with the highest correlation coefficient [16]. The target card supports both contact and contactless communication interfaces and performs 128-bit ARIA encryptions in its software.

A power analysis attack was performed as the reference for the EMD analysis attack. The goal of the first EMD analysis attack was to verify that the EMD of the target device indeed leaked side-channel information. The attacks that we discuss in this section used the output of the entire S-box operation in the first ARIA round to reveal all 16 bytes of the secret key. For each attack, 50,000 traces were recorded and the Hamming weight model was employed. Approximately, one day was needed to reveal the keys for performing the EMD analysis using CPA; a half day was required for measuring the EMD traces and the remaining half was required for the analysis.

5.1. The Power Analysis Attack Results. The power consumption of the target card was measured during the power analysis attacks using a power acquisition board designed for a contact smart card. Figure 3 shows the plot of the power trace (amplitude versus time in μs) captured by an oscilloscope from the target device as it performed the first round of the ARIA encryption. From Figure 3, we cannot clearly identify the times of ARIA's S-box and diffusion operations.

When performing the CPA attack, we could find all the 16 bytes of the first round key from the target card using 50,000 power traces.

5.2. The EMD Attack Analysis Results. The next experiment focused on the EMD analysis. The EMD signal of the target card was measured using the proposed EMD reader board, described in Section 3. Figure 4 shows the plot of an EMD trace from the target device as it performed the first ARIA encryption round. Compared to the power trace shown in Figure 3, many details of the ARIA operation are more clearly visible in Figure 4. The ARIA S-box operations are repeated 16 times (16 dotted lines in Figure 4), and then the diffusion operations are carried out. Because of the distinct EMD signal, the S-box operations in the first round of the ARIA were identifiable without needing any additional signal filtering process.

Figure 5 shows the result of the EMD analysis attack on the contactless target card. Figure 5(a) shows the maximum correlation coefficient of each candidate key for the first byte round key. In this result, the correlation coefficients for the incorrect key candidates are significantly smaller than those found for the correct key ($0 \times$ D4). The right-hand side of Figure 5 shows the plots for all of the key candidates. $0 \times$ D4 is plotted in black, whereas all the other keys are plotted in gray. There are no significant peaks in gray; only the plot for $0 \times$ D4 contains high peaks.

When performing a standard CPA on the EMD signals, we also found all of the 16 bytes round keys from the target contactless card using 50,000 power traces. In Table 1 and Figure 6, we compare the efficiency of the CPA on the EMD signals to that of the CPA on the power signals. To illustrate the comparison results clearly, we define SNR to be the ratio between the correlation of the CPA peak corresponding to the correct key and the highest correlation of the CPA peak resulting from the other wrong keys. In other words, SNR is the maximum correlation of the right key/the maximum correlation of the wrong keys. In this manner, we find that SNR is greater than 1; that is, we can guess the correct key. The greater the SNR is, the higher the accuracy of the round key hypothesis is. As can be seen in Table 1 and Figure 6, all of the 16 round keys could be correctly found with high SNR

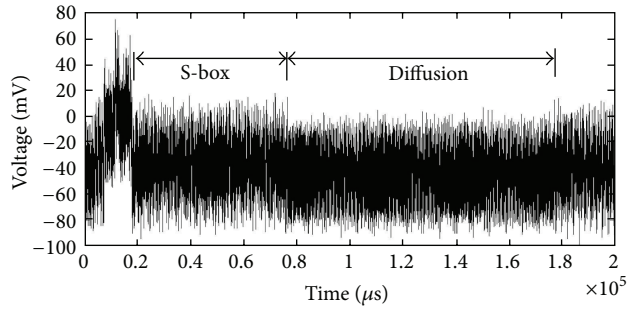


Figure 3: Power trace showing the first ARIA round.

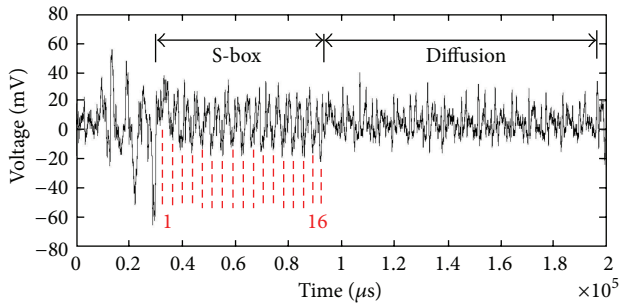


Figure 4: Plot of the EMD trace showing the first ARIA round.

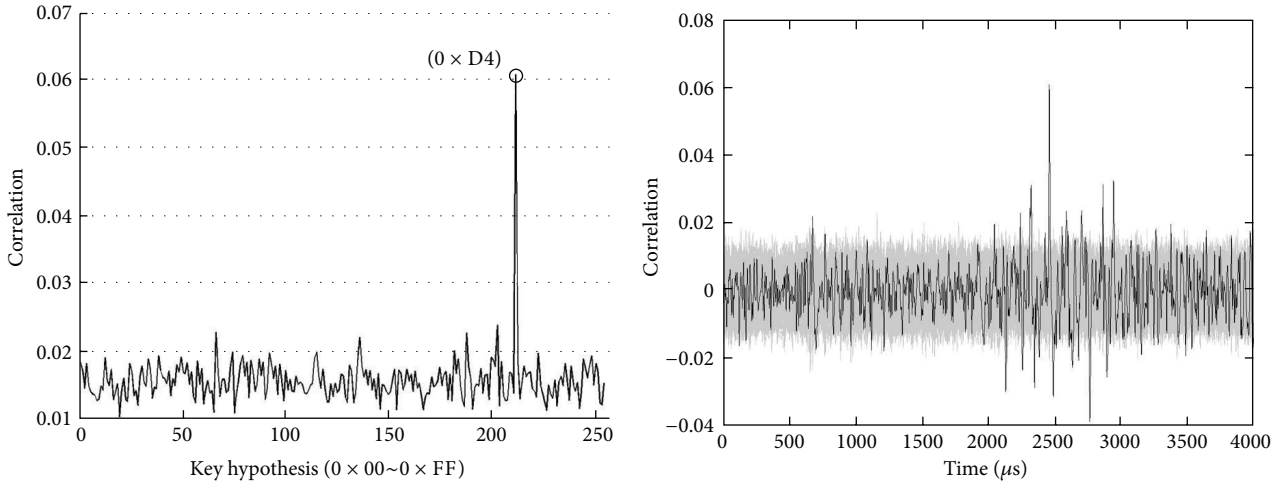


Figure 5: Result of EMD analysis for the first S-box.

ratios from both the EMD and the power analyses. With the exception of the 6th and 9th round keys, the experiments show that the EMD analysis gave better results than did the power analysis.

6. Conclusions

In this paper, we presented a new side-channel attack that exploited the information leakage from the EMD of a contactless smart card and showed that the EMD can be used as another information leakage channel exploitable in side-channel analysis on contactless smart cards. This novel EMD side-channel analysis allowed for a simple and efficient measurement setup in order to perform side-channel analysis on contactless smart cards.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

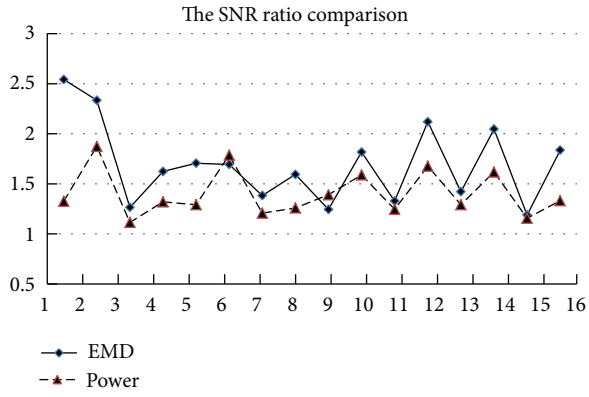


Figure 6: SNR comparison.

Table 1: SNR Comparison.

Round key	SNR ratio	
	EMD analysis	Power analysis
1st	2.542	1.326
2nd	2.335	1.874
3rd	1.264	1.115
4th	1.624	1.320
5th	1.706	1.291
6th	1.694	1.787
7th	1.383	1.207
8th	1.594	1.258
9th	1.244	1.391
10th	1.818	1.589
11th	1.328	1.246
12th	2.120	1.676
13th	1.421	1.292
14th	2.047	1.618
15th	1.189	1.156
16th	1.836	1.332

Acknowledgments

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the IT/SW Creative Research Program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2013-H0502-13-1074) and partly supported by Korea University Grant.

References

[1] P. Kocher, J. Jaffe, and B. Jun, "Differential power analysis," in *Proceedings of the 19th Annual International Cryptology Conference on Advances in Cryptology (CRYPTO '99)*, vol. 1666 of *Lecture Notes in Computer Science*, pp. 388–397, Springer, 1999.

[2] D. Carluccio, K. Lemke, and C. Parr, "Electromagnetic side channel analysis of a contactless smart card: first results," in *Cryptographic Hardware and Embedded Systems - CHES 2007*, vol. 4727 of *Lecture Notes in Computer Science*, pp. 320–333, Springer, 2007.

[3] M. Hutter, S. Mangard, and M. Feldhofer, "Power and EM attacks on passive 13. 56 MHz RFID devices," in *Cryptographic Hardware and Embedded Systems - CHES 2007*, vol. 4727 of *Lecture Notes in Computer Science*, pp. 320–333, Springer, 2007.

[4] T. Kasper, D. Oswald, and C. Paar, "EM side-channel attacks on commercial contactless smartcards using low-cost equipment," in *Information Security Applications*, pp. 79–93, Springer, 2009.

[5] ISO/IEC, "14443-2/CD Amd 3, Identification cards—contactless integrated circuit cards—proximity cards—part 2: radio frequency power and signal interface—amendment 3: limits of electromagnetic disturbance levels," JTC10, 2010.

[6] D. Kwon, J. Kim, S. Park et al., "New block cipher: ARIA," in *Information Security and Cryptology - ICISC 2003*, vol. 2971 of *Lecture Notes in Computer Science*, pp. 432–445, Springer, 2004.

[7] T. Kasper, D. Oswald, and C. Paar, "Side-channel analysis of cryptographic RFIDs with analog demodulation," in *RFID Security and Privacy*, pp. 61–77, Springer, 2012.

[8] F. Peters, "Physical interface of contactless cards past and future evolution," in *Proceedings of the APTA ITS Best Practices Workshop: Electronic Payment Systems*, American Public Transportation Association, 2010.

[9] C. Ziomek and P. Corredoura, "Digital I/Q demodulator," in *Proceedings of the 16th Particle Accelerator Conference*, pp. 2663–2665, May 1995.

[10] ROHDE&SCHWARZ, *Measuring Electro Magnetic Disturbance During ISO/IEC, 14443 Chipcard Data Transmission, Application Note*, ROHDE&SCHWARZ, 2007.

[11] T. S. Messerges, *Power analysis attacks and countermeasures for cryptographic algorithms [Ph.D. dissertation]*, University of Illinois at Chicago, 2000.

[12] K. Finkenzeller, *RFID Handbook: Fundamentals and Applications in Contactless Smart Cards*, Wiley & Sons, 2nd edition, 2004.

[13] ISO/IEC, "10373-6 Ed. 2. 0 en, Identification cards&test methods&part 6: proximity cards," JTC1, 2010.

[14] H. Zangl, M. J. Moser, T. Bretterklieber, and A. Fuchs, "Passive Wireless Devices Using Extremely Low to High Frequency Load Modulation," http://cdn.intechweb.org/pdfs/8991.pdf.

[15] E. Brier, C. Clavier, and F. Olivier, "Correlation power analysis with a leakage model," in *Cryptographic Hardware and Embedded Systems - CHES 2004*, vol. 3156 of *Lecture Notes in Computer Science*, pp. 16–29, Springer, 2004.

[16] T. Sugawara, Y. Hayashi, N. Homma et al., "Spectrum analysis on cryptographic modules to counteract side-channel attacks," in *Proceedings of the International Symposium on Electromagnetic Compatibility (EMC '09)*, pp. 21–24, July 2009.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 706287, 15 pages
http://dx.doi.org/10.1155/2014/706287

Research Article

User-Independent Activity Recognition via Three-Stage GA-Based Feature Selection

Theresia Ratih Dewi Saputri, Adil Mehmood Khan, and Seok-Won Lee

Division of Information and Computer Engineering, Ajou University, San 5 Woncheon-Dong, Suwon 443-749, Republic of Korea

Correspondence should be addressed to Seok-Won Lee; leesw@ajou.ac.kr

Received 29 November 2013; Accepted 19 February 2014; Published 26 March 2014

Academic Editor: Jongsung Kim

Advancement in wireless sensor networks gave birth to applications that can provide friendly and intelligent services based on the recognition of human activities. Although the technology supports monitoring activity patterns, enabling applications to recognize activities user-independently is still a main concern. Achieving this goal is tough for two reasons: firstly, different people exhibit different physical patterns for the same activity due to their different behavior. Secondly, different activities performed by the same person could have different underlying models. Therefore, it is unwise to recognize different activities using the same features. This work presents a solution to this problem. The proposed system uses simple time domain features with a single neural network and a three-stage genetic algorithm-based feature selection method for accurate user-independent activity recognition. System evaluation is carried out for six activities in a user-independent setting using 27 subjects. Recognition performance is also compared with well-known existing methods. Average accuracy of 93% in these experiments shows the feasibility of using our method for subject-independent human activity recognition.

1. Introduction

The advancement in technology and the widespread of smart devices, such as smart phone, over the past few years provided a computational model that makes it possible to recognize human user's context anywhere and anytime. One area under the umbrella of automatic context recognition, which has been extensively studied over the past decade, is human activity recognition (HAR). HAR deals with the automatic recognition of activities of daily living using computers. These activities include both high-level activities, such as cooking and taking a shower, and low-level physical ones, such as walking and running. Physical activity patterns can provide significant support in various system (e.g., health care system).

In order to gather the information about physical activities, various sensing technologies have been introduced. One such technology is an accelerometer sensor. Due to high improvement in their sensing technology, it is now possible to use accelerometers to gather acceleration information about physical movement to recognize physical activities of a person in a more pervasive fashion. Although the technology supports the monitoring of activity patterns using accelerometers, the effectiveness of the recognition algorithm is still the main concern to interpret the accelerometer data based on different subjects and different activities as HAR requires an objective and reliable technique that can be used under the condition of daily living [1].

Even though there exist a number of research studies that have investigated the area of HAR via accelerometer (a-HAR) at length [2–5], there are two important aspects that have stayed unobserved. The first aspect is the fact that different people exhibit different physical activities for the same activity pattern due to their different behavior. For example, some people walk fast, whereas others walk at a slower pace. This phenomenon could result in misclassification of walking as running activity. The second aspect is that different activities performed by the same person could have different underlying models which makes it unwise to recognize them using the same feature. For example, walking is very different than cycling because in walking activity the whole body plays its role, whereas in cycling its mainly the

legs that are involved. In order to overcome this problem, the stronger analytical method must be carried out to understand the behavior of different subjects regarding their physical activities for selecting any features. Therefore, this work proposes a feature selection method that is able to extract the most appropriate features of accelerometer data by analyzing a vast set of features based on subject and activity behavior.

This work makes several contributions in the area of a-HAR. Firstly, we have collected a significant amount of activity data from a large number of subjects using accelerometer-enabled smart phones. We have analyzed these data to demonstrate for the fist time that different people perform the same activities with different behaviors, and different activities performed by the same subject could follow different models. Secondly, based on our findings, we implemented a three-stage genetic algorithm-based feature selection method. This method produces a feature set that is both subject-independent and is capable of representing multiple activities effectively in the feature space. Thirdly, we used the selected features set with neural network, as the classifier, and compared its performance with seven existing works to show the feasibility of using our method for a-HAR via smart phone accelerometers.

The rest of the paper is organized into the following sections. In Section 2, we explain the background and related work of this research in the area of HAR (in general) and HAR (in particular). Section 3 explains in detail the proposed approach for subject-independent activity recognition. Section 4 talks about experiments and presents the experimental and comparison results for our approach and some existing a-HAR algorithms. Finally, in Section 5, we conclude our work and briefly talk about the future directions.

2. Background and Related Work

In this section, we briefly discuss the related work. Firstly, we explain the motivation behind context aware system. Next, we talk about one of the examples of context aware systems, that is, activity recognition and the existing activity recognition research. Lastly, this section discusses accelerometer, a low-cost wearable sensor, along with some related work in the field of a-HAR.

2.1. Context Awareness System. Ubiquitous computing, a computing paradigm that emerged about two decades ago, introduced the idea of making computing devices available everywhere in the physical world, while keeping these devices effectively invisible to the user at the same time. With the use of ubiquitous computing people can receive and process information anytime and anywhere through a device which can connect to the internet. This would result in reducing complexity of using devices and making people live easier and more efficiently [6].

Ubiquitous computing uses context as its core resource to provide proper service and information. Context is any information that can be used to characterize the situation of entities that are considered relevant to the interaction between users and application themselves [7]. One of the recent applications of ubiquitous computing is context aware system.

A context aware system is one that actively and autonomously adapts and provides the appropriate service or context to users, using the advantages of contextual information [8]. Though context comes in different types, one such type is the activity being performed by a user at any given time.

2.2. Human Activity Recognition (HAR). HAR requires an objective and reliable technique that can be used under the condition of daily living [1]. In order to achieve this goal, HAR system should be equipped with sensing ability. Two approaches have been mainly used for this purpose [9]. The first approach is external sensor, fixedly placed in a particular location at the predetermined point of interest. On the other hand, the second approach, which is a wearable sensor approach, is a dynamic device attached to a user. Based on [9], wearable sensor is better than external sensor because the external sensor is only able to capture human activity when users are in the coverage range of the sensor which makes it lacking of pervasiveness. Due to its capability to capture human activity without position boundary, the wearable sensor approach became the most accepted approach. One of widely used wearable sensors for HAR is the triaxial accelerometer. First research in the area of a-HAR was conducted in the late 90's yet convincing challenge still exists within this field [3].

In [5] they conducted an activity recognition using single triaxial accelerometer worn near the pelvic area. They focused on eight activities including standing, walking, running, climbing upstairs, climbing downstairs, sit-ups, vacuuming, and brushing teeth. In order to recognize those activities, a particular algorithm is used to recognize the accelerometer signal pattern corresponding to each activity. Using a set of simple time domain features, which include mean, standard deviation, energy and correlation, they evaluated the performance of several classifiers such as Decision Tree, K-Nearest Neighbors, SVM, and Naïve Bayes.

In [2], the authors compared accuracy for different features across a number of different lower limb placements. In this research they investigated eight different dynamic activities including walking, walking up stairs and down stairs, jogging, running, hopping, on the left and right leg, and jumping. Seven sets with different number of features were evaluated using K-Nearest Neighbor classifier. This research found that it reaches a good level of classification accuracy when using simple time domain features.

In [4] three features were extracted from each axis of the accelerometer including peak-to-peak amplitude, standard deviation, and correlation between axes. In order to preserve the accuracy, they selected the significant features and eliminated the ineffective ones. Fuzzy inference system was used to classify four activities including moving forward, going down stairs, going up stairs, and jumping.

Researchers in [10] recognized a group of daily activities using evolutionary fuzzy models. Seven common dynamic activities were selected as the basic activities of daily life to be

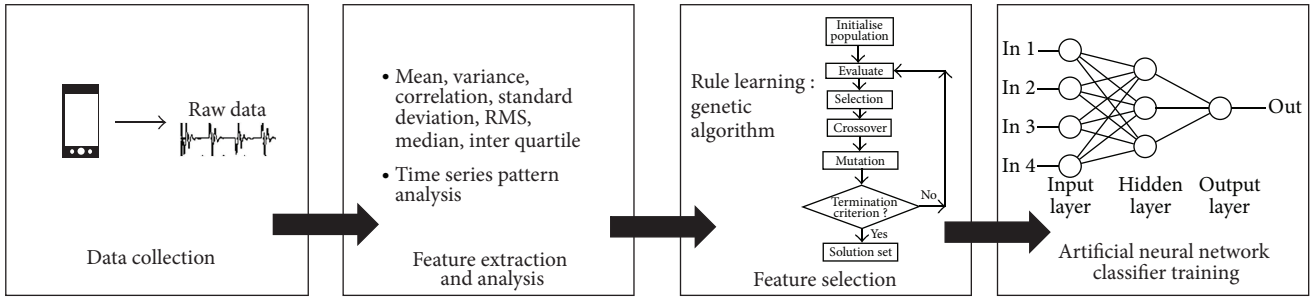


Figure 1: The adopted methodology for subject-independent activity recognition on accelerometer data using neural network classifier with genetic algorithm feature selection.

recognized, that is, walking, jogging, running, cycling, going up stairs, going down stairs, and hopping. Their evolutionary fuzzy model was able to estimate the membership functions through a statistical method and fuzzy rules using genetic algorithm optimization.

In [11] the research focused on five daily activities including walking, cycling, running, idling (sitting or standing), and driving a car. The research aimed at providing real-time activity recognition. In this research, 21 features including standard deviation, mean, and percentile were extracted from the accelerometer. Those features were used to classify selected activities using k-nearest neighbor and quadratic discriminant analysis classifiers. This research was able to show both classifiers are reliable for real-time activity recognition.

Those previous researches show a remarkable result in the activity recognition area. Various activities have been classified using several classifier algorithms based on numerous selected features. The previous studies achieved a good performance for recognizing the human activity. However, they failed to achieve good performance for subject-independent activity recognition, as we show in the Section 4.2. The previous research overlooked two important aspects, that is, different people perform the same activity differently and different activities performed by the same person could have different models. In other words, every person has a different behavior such as gesture to perform a certain activity. And, it is important to understand these different behaviors, as by understanding this behavior we will able to provide a more reliable activity recognizer. In order to overcome this matter, we analyze influential features from different subjects for different activities using a three-staged genetic algorithm based feature selection method. Those selected features are then used to classify activities using neural networks. Our proposed model is able to understand dynamic activities from different subjects using their accelerometer data and is capable of providing high accuracy for subject-independent activity recognition.

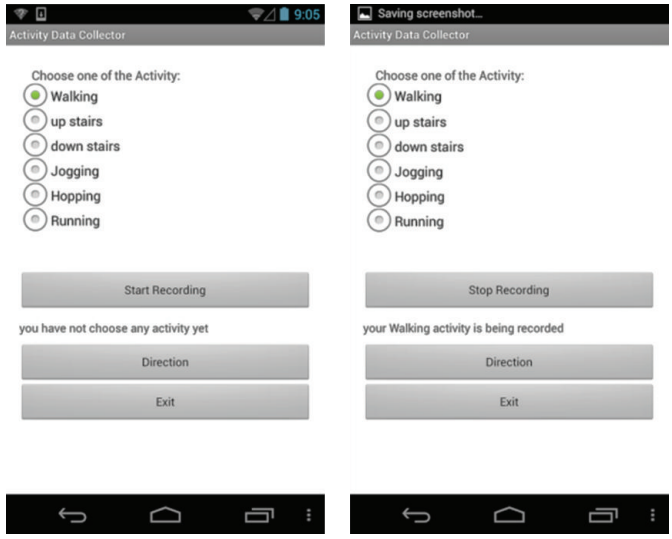
3. Approach

The adopted methodology of this research for dynamic activity recognition is illustrated in Figure 1. The first step in our proposed model is data collection using accelerometer enabled smart phones. The second step is feature extraction and analysis based on time domain feature analysis. The third step is the feature selection method for subject-independent human activity recognition using genetic algorithm. The learning process for activity classification is done in the fourth step using neural network, based on selected influential features.

3.1. Data Collection. As we can see in Figure 1, there are four major steps in our research methodology. The first step is data collection, that is, a collection of raw signals from accelerometer sensor, as people perform daily activities. In this research, we are focusing on recognizing dynamic activities. These activities include walking, jogging, running, going upstairs, going downstairs, and hopping. Those activities are selected based on the conducted research [10]. In order to get common position, subjects were asked to place their smart phone at front right pocket of a pant. This location is designed to capture user activity based on their leg movement due to our focus on dynamic activities. The accelerometer captures the activity by measuring the orientation of the device. Therefore, it could result in different patterns when the device is put in the different positions. The work of [12] shows that the accelerometer that is put on the thigh gives a powerful performance to differentiate the activities.

The android smartphone accelerometer is used to collect the activity data set. Each subject was asked to collect the data activity using our custom build application that can be seen in Figure 2. As we know, different devices have different sampling rates based on the smart phone model, so in order to control the data collection process, we did not use the highest number of sampling rate because it may differ for various android devices making the method less device model dependent. Based on [10] it shows that 50 Hz is a suitable sampling rate for recognizing dynamic activities with acceptable accuracy, which is used in this work as well.

In this study, the data sets were collected from 27 healthy subjects (12 females and 15 males) between the ages of 18 and 29 years old. The criteria of selecting the subjects are based on their gender and age. We considered the gender and age because we assumed that different age and gender could perform different behaviors for the same activity. Those subjects were asked to perform more than one activity each day and each activity should be performed more than twice. We collected those data for more than one month. Therefore,



(a) (b)

Figure 2: Accelerometer-enabled activity data collector application before the user starts recording their activities (a) and while the user is recording their activities (b).

we are able to collect data from the same subject and the same activity but performed on different days.

3.2. Feature Extraction. In order to recognize the activity, each behavior of activity should be represented with simple and general features [13]. The second step in our methodology is feature extraction which extracts the representative feature to recognize the activities. An accelerometer sensor generates time series signals that are highly fluctuating and oscillatory in nature [10]. Those accelerometer signal characteristics make activity recognition more difficult if we directly use raw signal data. Therefore, feature extraction is needed to gather the nontrivial data from such signals.

In order to extract information from those data, we divided signal data in several equal-sized windows. The windowing process reduces the flow rate and sends less data to system to recognize the activity performed by a certain subject [14]. Given a sampling rate of 50 Hz, we chose a window size of 100 samples, meaning two seconds, as such a window provides enough data for quality feature extraction while ensuring a fast response at the same time. Each window contains 100 numbers of samples as shown in Figure 3.

As for the feature extraction, there are several types of features that can be extracted from raw signal data such as time domain features and frequency domain features. The work done in [9] showed that time domain features are able to effectively represent the data that can be used for activity recognition. The research found simple statistical feature and coefficient of time series analysis to be highly suitable for smartphone based activity recognition, as these features are capable of providing high recognition rates at lower sampling rates. Based on this finding, we chose the same features for our work.

As for simple time domain features, several features including mean, root mean square, variance, correlation, and standard deviation were used. The mean feature helps to characterize each window. The root mean square feature measures the tendency of data [15]. Also, variance feature is used to measure the data spread among different activities. Meanwhile, correlations between axes are also considered as a feature to represent the interrelationship of among triaxial accelerometer data. Standard deviation helps in capturing the range of acceleration.

In order to understand the individual behavior in subjects' physical pattern, we also analyzed each activity data using time series modeling techniques, as time series analysis can reveal the unusual observation and particular patterns of data [16]. There are several models that are commonly used to perform time series analysis, such as a moving average model, autoregressive model, and combination of both models. Autoregressive model is useful for describing situations in which the present value of time depends on its preceding value and its random shock which represents the phenomena of data behavior [17]. While moving average is useful in describing phenomena in which event produce and immediate effect that only last for a short period of time [18].

In order to identify the model in our data, partial autocorrelation function (PACF) and autocorrelation function (ACF) coefficients were used as the characteristics of those models. Those coefficients reveal the pattern of each datum and indicate the possible model of the data. Determination of the model for the data is done based on the characteristic of theoretical ACF and PACF that can be seen in Table 1 [16] and sample of PACF and ACF of the activity, which can be seen in Figure 4.

The fitting process of time series model to any data means estimating the parameter values for that model based on a selected model order. The parameter estimation process of autoregressive and moving average requires an iteration procedure [19]. Among other iteration procedures, we adopt

Table 1: Characteristic of theoretical ACF and PACF for determining process model.

Process	ACF	PACF
AR (p)	Tails off as exponential decay or damped sine wave	Cuts off after lag p
MA (q)	Cuts off after lag q	Tails off as exponential decay or damped sine wave
ARMA (p, q)	Tails off after lag ($q - p$)	Tails off after lag ($p - q$)

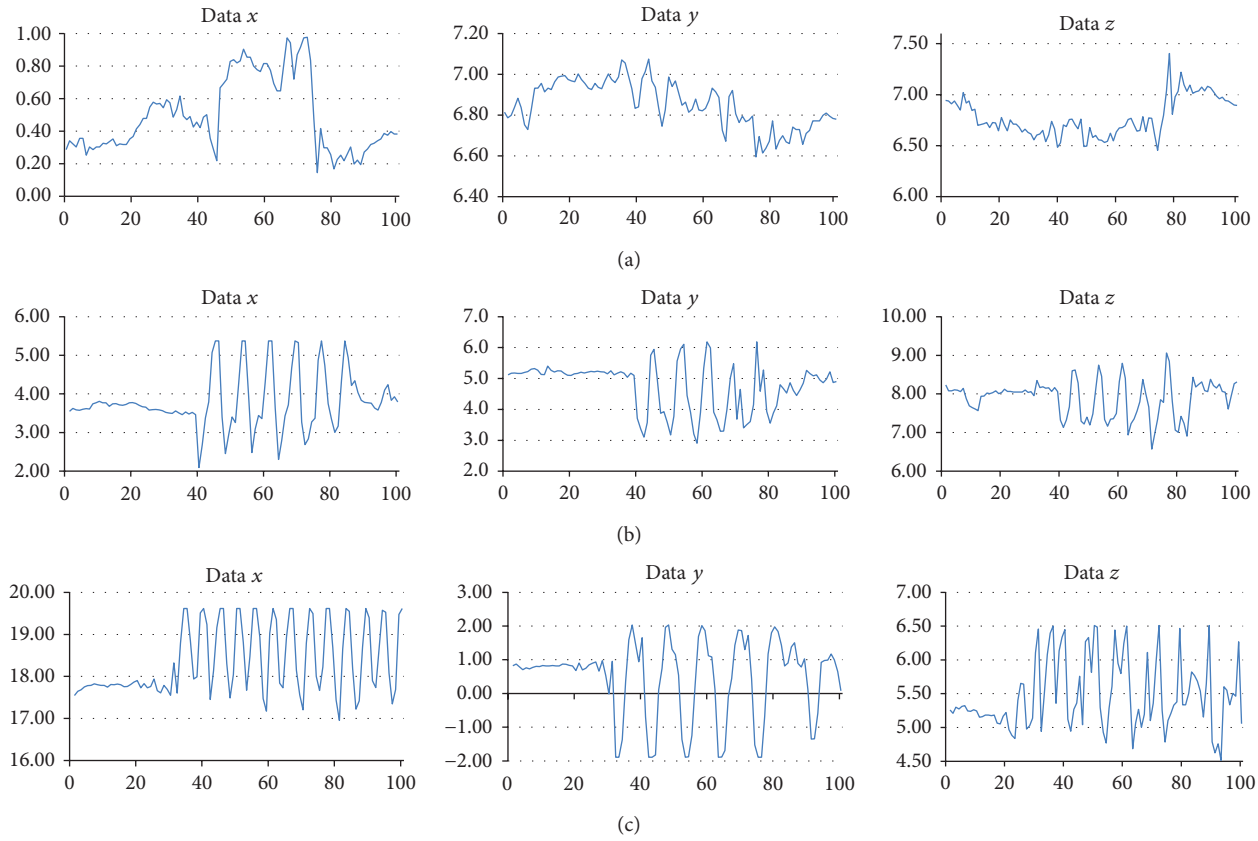


Figure 3: Representative raw signals of going downstairs (a), hopping (b) and running (c) activities which have different signal patterns.

Box-Jenkins model estimation due to its flexibility to the inclusion of both autoregressive and moving average model [20]. Determined model and parameter have to be verified to ensure that estimated parameters are statistically significant [21]. In this research, we used likelihood ratio to test model specification [22].

After the feature extraction, the selected features were analyzed. Figure 5 shows running activity and walking activity from the same subject, and it can be seen that they have different process models. For example, based on the characteristic of the model presented in Table 1, we are able to see that x-axis data of walking activity is an autoregressive (AR) model. On the other hand x-axis data of running activity is autoregressive moving average (ARMA) model. Same phenomena were witnessed for many other activities. Based on this difference, we can conclude that different activities could exhibit different data behaviors.

That phenomenon does not only happen in different activities; even the same activity could exhibit different behaviors. This phenomenon could happen when an activity is performed by different subjects. This is very distinguishable because every person shows a different behavior while performing different activities. Figure 6 shows that different subjects (a), (b), (c), and (d) show different behaviors while performing the same activity, which is the running activity in this case. As we can see, the subjects (a) and (b) have ARMA model in every axis of their data but in different order. Compared to them, subject (c) has an AR model for every axis of their data. On the other hand, subject (d) has different models for every axis of their data. Due to these differences, we can see that every subject has different behavior, even performing the same activity. Therefore, it is important to understand those behaviors in order to get common features for every subject to support subject independent activity recognition.

As we can see in Figures 5 and 6, every activity performed by different subject could fall in the different underlying models. Those differences also result in different features for each activity. Therefore, a single feature is not able to represent the entire activity. In order to solve this problem, we

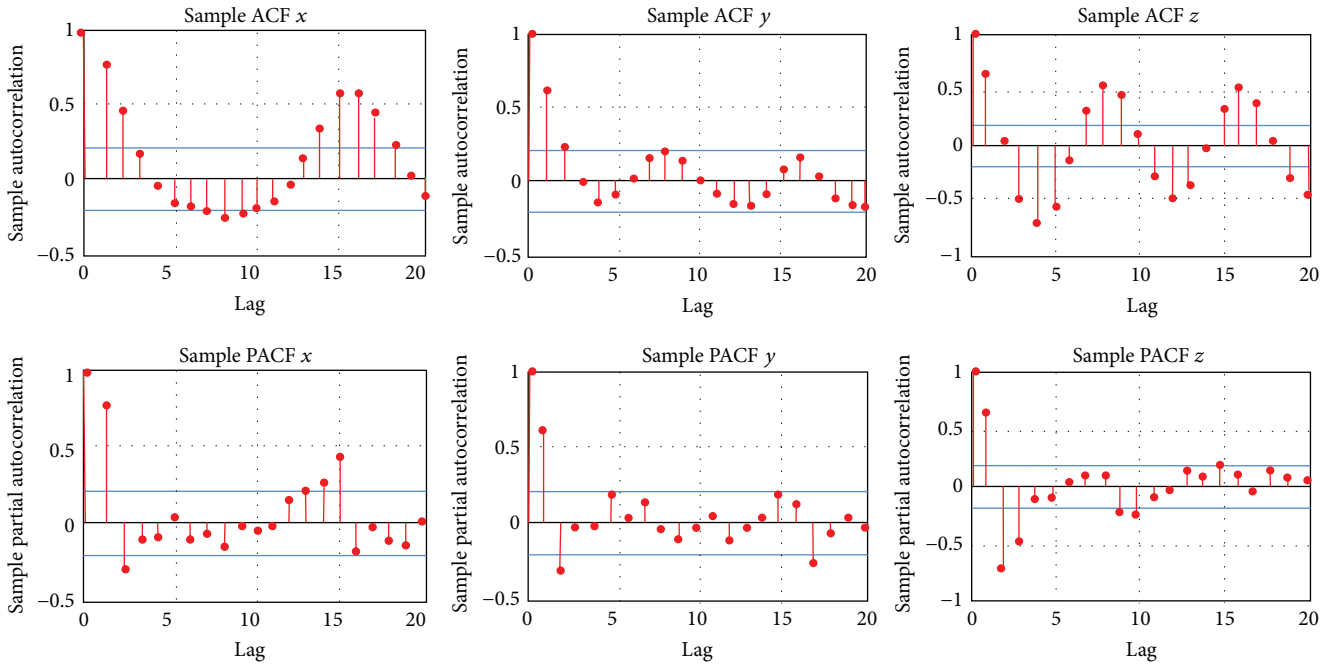


Figure 4: Sample of ACF and PACF from a certain activity that can be used to determine the time series model of the activity data.

decided to create a big set of features and implement feature selection method.

3.3. Feature Selection. The third process is feature selection which is the selection of features that have high impact on the intended activities. Ladha and Deepa [23] define feature selection as a process commonly used in machine learning wherein subsets of the features available from the data are selected for application of learning algorithm. There are several advantages of feature selection. Feature selection is able to reduce dimensionality of feature space which can avoid the curse of dimensionality [23]. The main purpose of feature selection is to increase the accuracy of the resulting model. Feature selection also helps to reduce the abundant, irrelevant, misleading, and noisy features. Also, the use of feature selection is able to reduce the cost of the system in most applications [24]. As we showed in the previous section, there is no single feature set that is able to consistently perform better for all activities. Therefore, it is important to determine the features which have high impact.

Several algorithms have been presented as a computational solution for the feature selection problem [23]. The first approach is filter method which selects the feature based on discriminating criteria that are relatively independent of classification. The method uses minimum redundancy-maximum relevance feature selection. This method is fast and scalable. This method also provides good computational complexity. Unfortunately, it ignores the interaction with the classifiers. Some examples of algorithms for this method are Euclidean Distance and Correlation-based Feature Selection. The second method is the wrapper method which is a feature selection method that utilizes the classifier as a black box to score the subset feature based on their predictive power. The wrapper method uses simple and less computational feature selection. This method also interacts with the classifier to optimize the feature subset. The disadvantage of this method is its dependency on the classifier that makes classifier selection become an important process in this method. The algorithms that use this method are sequential forward selection, simulated annealing, and genetic algorithm.

The learning algorithm that we are going to use for feature selection is genetic algorithm. This algorithm gained a lot of attention due to its ability to reduce the likelihood of getting trapped in local optimum which inevitably is present in many practical optimization problems [25]. Genetic algorithm is parallel, iterative, optimized, and has been successfully applied to a broad spectrum of optimization problems [26]. The genetic algorithm evaluates the features by finding the maximum fitness of population by selecting feasible individuals from the population and uses its genetic information to produce the new optimal population of solution.

There are two basic operations in a genetic algorithm to produce new generation: crossover and mutation [27] process from the chromosomes. This chromosome which represents the set of selected feature is composed of several genes. Each feature is treated as a single gene. This gene is mapped into a chromosome by given a certain index, which is as follows:

$$\text{gene index} = \begin{cases} 1 & \text{if the feature is selected} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The encoding result of the features based on (1) can be seen in Figure 7 where F_x is the xth feature in the system.

Since our aim is to find feature set which is both appropriate for different activities for single person and effective

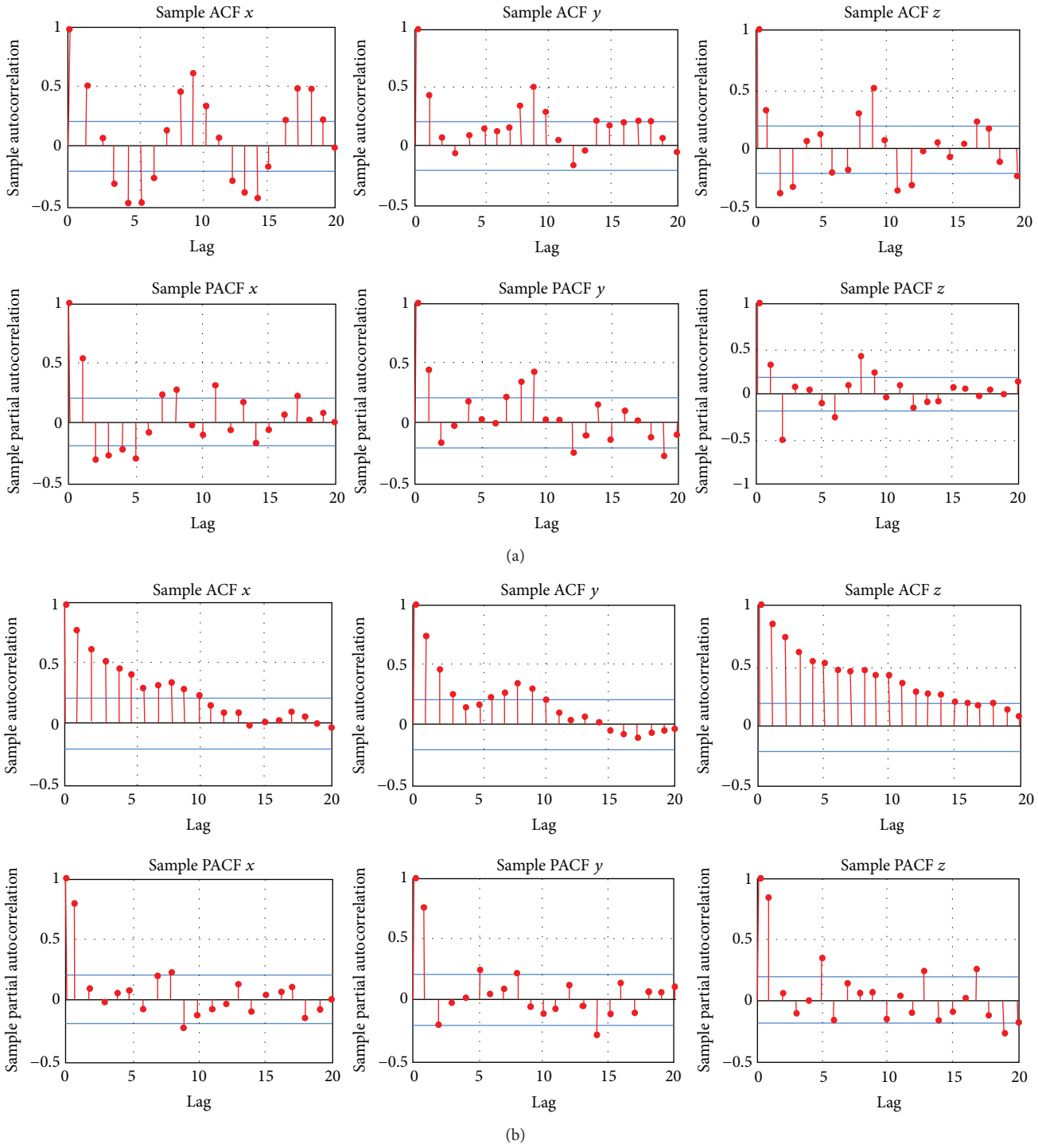
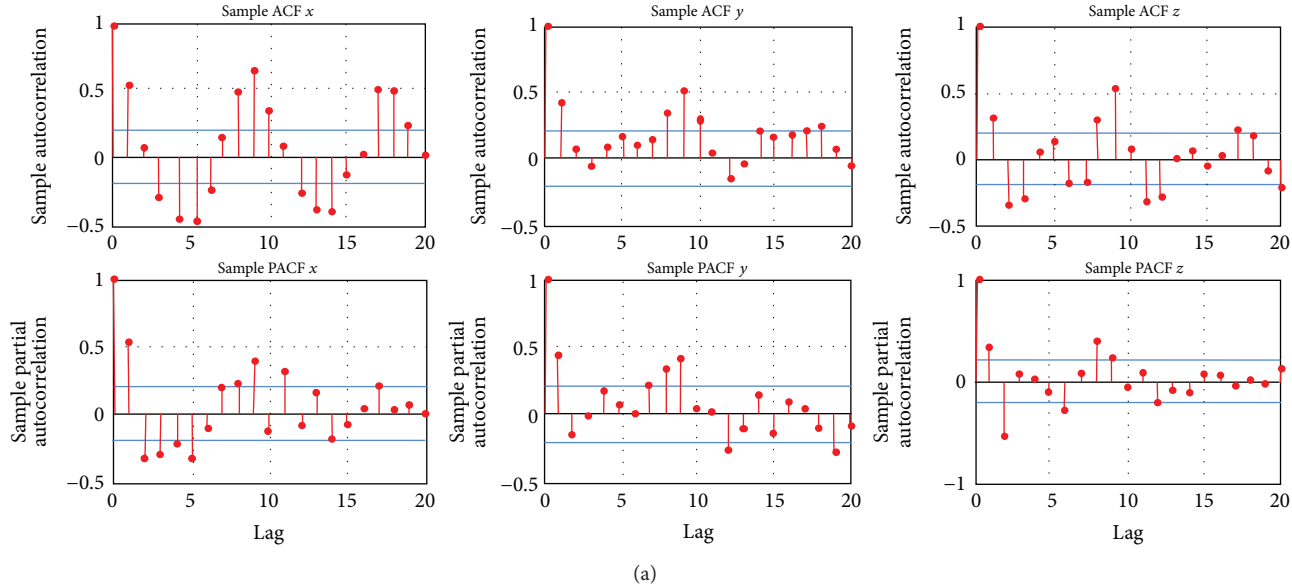


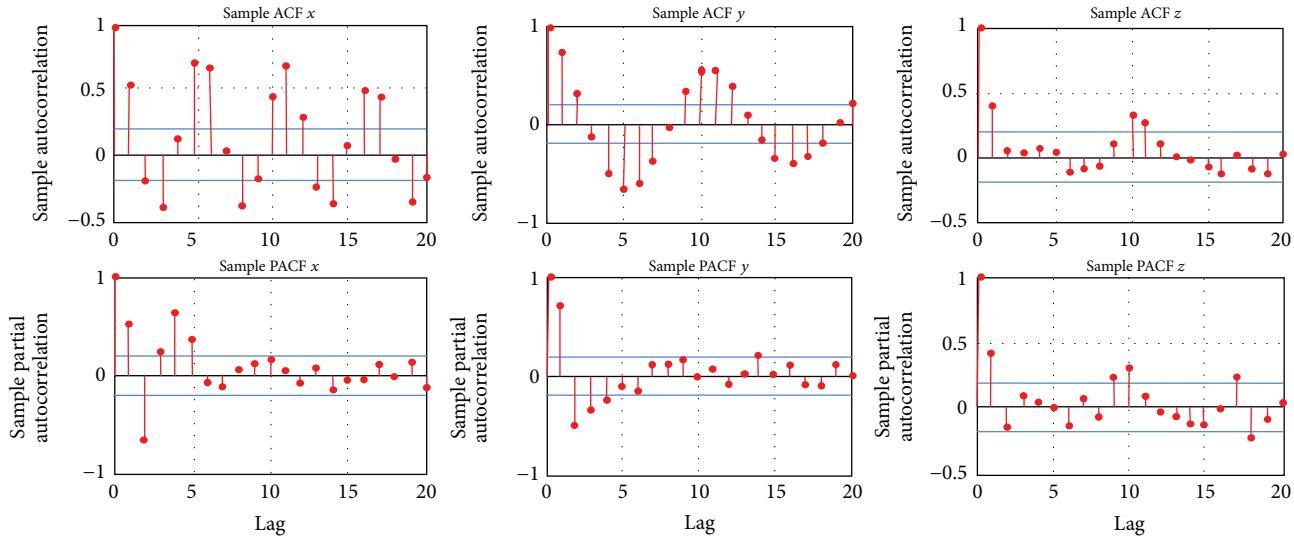
Figure 5: ACF and PACF graph of running (a) and walking (b) activities from the same subject showing different patterns.

in representing these activities across multiple subjects, therefore, we have devised a three-staged genetic algorithm-based feature selection method as seen in Figure 8. Therefore, different number of x is used based on the number of selected features in each stage.

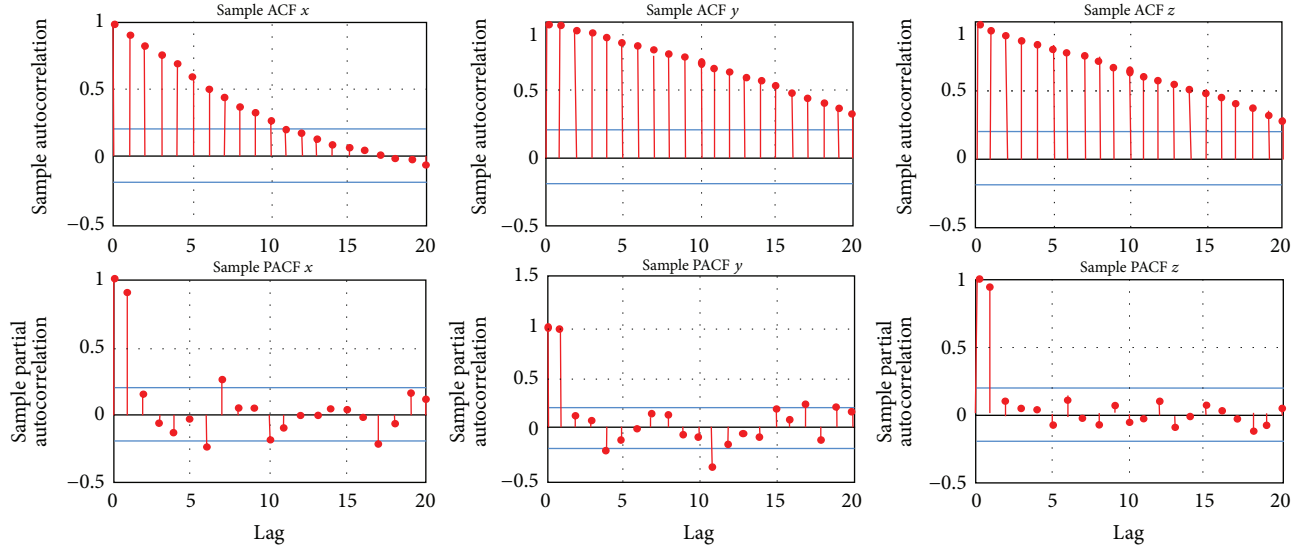
The first stage of our proposed method analyzes the feature from each activity of each user. Based on this step we are able to determine which features have high impact in every user physical activity. As we can see in Figure 8, every A activity from m number of subjects is evaluated. This stage is aimed to analyze the behavior of the same activity performed by the same subject in different time frames. For example, subject m performs A_1 in different time frames represented as $A_{1\text{-}1}$, $A_{1\text{-}2}$, $A_{1\text{-}3}$ using the entire extracted features. Using this



(a)



(b)



(c)

Figure 6: Continued.

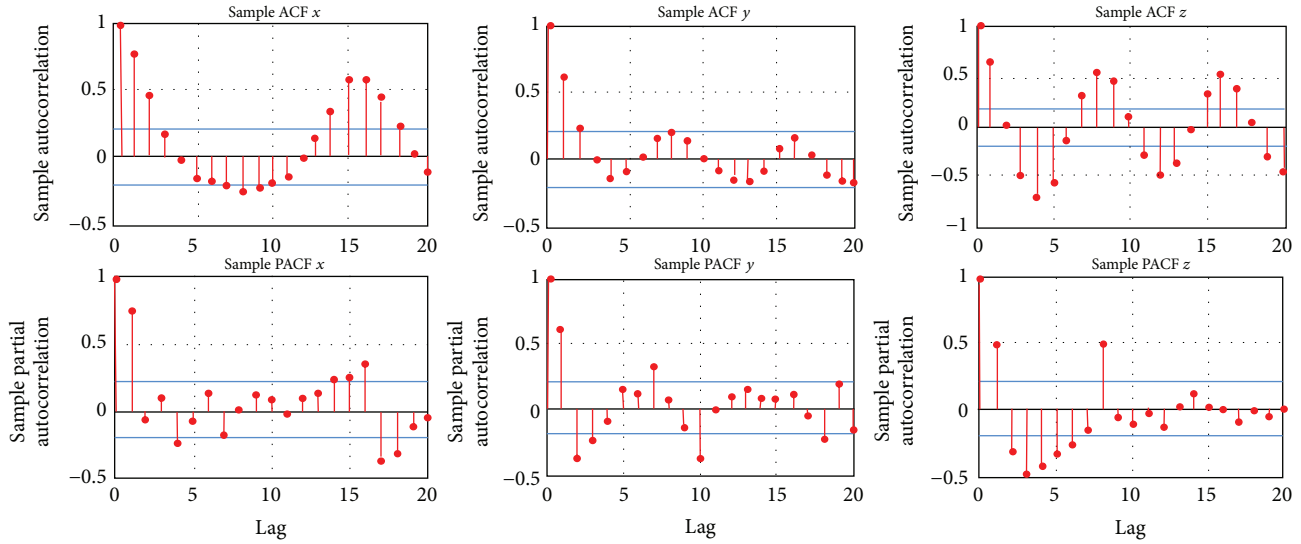


Figure 6: ACF and PACF graphs of running activity gathered from different subjects show that every subject has a different behavior in performing a certain activity.

F_1	F_2	F_3	F_4	F_5	...	F_x
1	0	1	1	0	1	1

Figure 7: Chromosome encodes of features.

stage, we are able to learn how a particular person performs a certain activity.

The first stage result is $m \times n$ number of sets of features, where m is the number of subjects and n is the number of activities (which is six activities). In this study, we use union set theory to combine the selected feature. It is possible that the new combination backs to the original features. Therefore, a particular rule should be applied to avoid that circumstance. Once a feature is selected a counter number is assigned to this feature. Not all of the selected features will be used for the next stage. There is a threshold for number of counters that should be fulfilled. Therefore, the relation of number of features and selected feature is $y \leq x$, where x is the original feature length and y is the number of selected features. This rule is also applied in the entire stage of the feature selection process. The result of this stage is used as the input of the second stage of feature selection.

The second stage of feature selection step is aimed to analyze the feature based on each subject. In this stage genetic algorithm is run based on the number of subjects using their entire sample. The input for this stage is the selected feature of each activity from every subject. This stage is aimed to determine the different behavior of each activity performed by the same subject. Using the second stage we can determine the important features for each subject. As we can see in Figure 8, each subject gives a different set of features. Therefore, set of feature from each subject is combined using the same rule used in the first stage. The second stage feature set is structured from the sets of each subject based on its counter. For example mean feature appears on set of subject 1 and set of subject 2 then its counter is 2. In order to be selected in second stage features each feature should have more than 50% of number of subjects. In order to get global feature selection, the third stage genetic algorithm is used.

The selected features from each subject as the result from the second stage are used as the input for the third stage of feature selection. The third stage analyzes every feature from each subject. This process is used to determine the common features of physical activity for every subject and every activity. Those common features are the features that are used as the set of features for activity recognition step.

3.4. Activity Classification. After we get the common important feature from the feature selection process, the next process that should be done is activity learner and recognizer. The learning activities during the training process and recognizing the activity in the testing process will be done using artificial neural network (ANN) classifier. This classifier is chosen due to its adaptive characteristic and able to provide accurate classification result. ANN is able to classify a certain pattern in which data have not been trained. The characteristics of ANN are inspired by the work performance of a biological brain system which has nonlinear characteristic, robustness, fault tolerance, and fuzzy information [28].

There are several algorithms that can be used as a classifier in human activity recognition area, for example, Bayesian rule, decision tree, regression, and neural network. One of the widely known algorithms for activity recognition is decision tree. The decision tree is a classification algorithm based on a hierarchical data structure that is composed of internal decision nodes and terminal leaves [29]. In [30], decision tree was trained based on mean acceleration to recognize the activity. Based on the accuracy, decision tree is able to provide good performance to recognize the posture

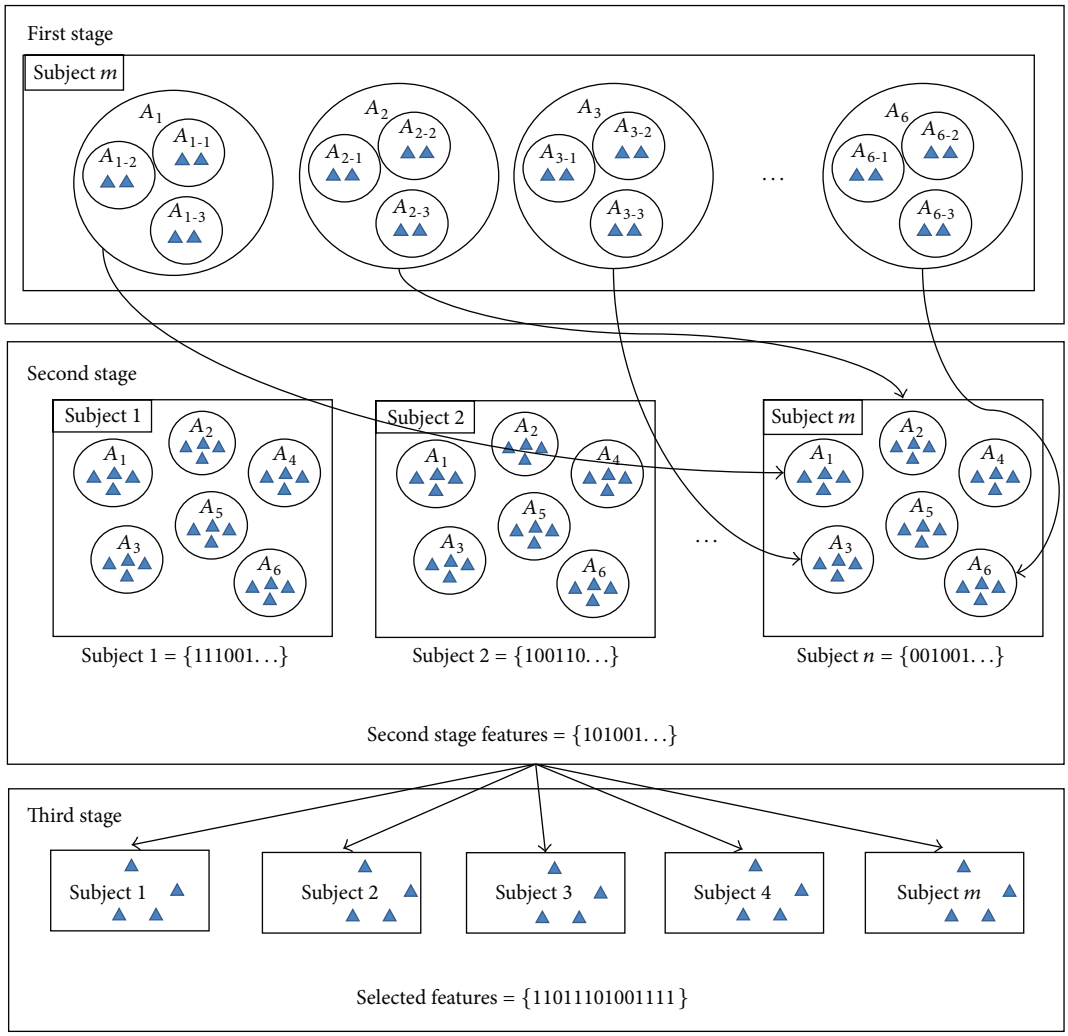


FIGURE 8: The feature selection process based on the extracted features using a three-stage genetic algorithm. The selected feature (represented as triangles) as the result of each stage is used as the input of the next stages (represented as arrows).

such as sitting and lying down. Unfortunately, it gave lower accuracy for activities such as stretching. The other algorithm is Bayesian rule. This algorithm classifies the activity by calculating the probability of each class [29]. The result in [30] shows that in order to provide accurate classification process using bayesian rule tends to need more data. It also shows that bayesian rule shows weaker performance due to its characteristic that is unable to precisely model the independence of features.

ANN is more robust and has better performance compared with other computational tools. One of the widely known learning algorithms in the neural network is back propagation neural network. Back propagation learns by iteratively processing a data set of training tuples by comparing the predicted value and actual target value (also called a class label) [31]. The network structure used in this research showed in Figure 9 consists of three stages which are input stage, hidden stage, and output stage.

The input for input stage that will be used in this network is the features gathered from the feature selection process. Therefore the x number of nodes in input stage is based on the length of the selected features. The z number of the output stage is calculated based on the number of activities as the target class. Activation of each node in hidden stage and the output stage is done using the log sigmoid function. A sigmoid is the frequently used activation function. This function is easy to distinguish, so it can minimize the computational capacity of the training process [32]. The network learns about the process by adjusting weight based on the error value. Adjusting the weight is done based on the error and learning rate. In order to evaluate the entire method, some scenarios which are sample based activity recognition and subject based activity recognition will be executed.

4. Performance Evaluation and Comparison

In this section, we present the experimental design to evaluate our proposed method. Case study designed methodology which has been explained in [33] is used to validate the

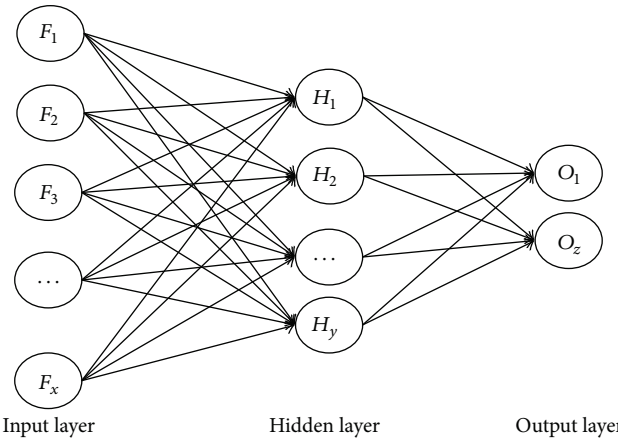


Figure 9: Neural network classifier structure for subject-independent activity recognition using selected feature $(F_1, F_2, F_3, \ldots, F_x)$ as the input criteria.

effectiveness of our method. We also present our evaluation process and the results in this section.

4.1. Experimental Design. In [33] Lee and Rine explained that it is necessary to have an empirical research methodology that is able to validate the effectiveness of a particular method. It provides a conceptual framework to support the collection of evidence as a set of conclusions to support our research hypothesis. Based on that study, the following components were defined.

4.1.1. Study Question. In order to validate the stated hypothesis, study questions need to be clarified precisely [33]. The "how and why" type questions are derived from research goal. Generated study questions for this study are as follows.

(1) Why is subject-independent activity recognition difficult to perform?

(2) How can feature selection improve the accuracy of subject-independent activity recognition?

(3) How does the proposed activity recognition approach perform in contrast to the previous approaches for subject-independent activity recognition?

4.1.2. Study Proposition. The study proposition is derived from study questions [33]. It is composed of a set of facts related to a research hypothesis that should be examined thought a certain measurement. Using study propositions, we are able to point out the goal of study, give a certain scope of experiments, and suggest possible links between phenomena (e.g., different behaviors of same activity performed by different subjects) during the evidence collection process. The derived study propositions related to our study question are as follows.

(1) Why is subject-independent activity recognition difficult to perform?

(1.1) Every subject has different statistical process models from other subject to perform the same activity.

(1.2) Every activity has different statistical process models from other activity.

(2) How can feature selection improve the accuracy of subject-independent activity recognition?

(2.1) The feature selection process is able to reduce the number of features.

(2.2) Feature selection learns characteristic of each activity behavior.

(2.3) Feature selection provides common features for every subject.

(3) How does the proposed activity recognition approach perform in contrast to the previous approaches for subject-independent activity recognition?

(3.1) When applied as a three-stage process analyzing the performance of different activities for each subject, it can help us identify the most suitable feature set for subject-independent activity recognition.

4.1.3. Unit of Analysis. A set of selected resources to be examined during the experiment process. This unit of analysis is used as the evidence to support our research hypothesis. Unit of analysis is the actual source of information that measures the achievement of study proposition. The units of analysis that we use in this study are as follows:

(i) process model of activity;

(ii) number of feature;

(iii) the accuracy rate for subject independent activity recognition without feature selection;

(iv) the accuracy rate for subject independent activity recognition using feature selection;

(v) the accuracy rate of the existing activity recognition approaches for subject-independent activity recognition under the exact same setting.

4.1.4. Linking Data. In order to connect the generated unit of analysis and study proportion, linking both of those entities is important. Table 2 shows the relation of unit analysis used as the validity evidence of study proposition.

4.1.5. Criteria to Interpret the Finding. This process corresponds to the measures used in evaluating the result of the experiments. This process is the iteration between propositions and data. Therefore, these criteria could help to support the study proposition. This criteria is also used as the proof of our research hypothesis. Using this criteria, we are able to determine whether our research hypothesis is accepted or not. The interpretation of our experiments is explained in detail in the following section.

Table 2: Evidence Collection.

Code	Study Proposition description	Evidence
1.1	Every subject has different statistical process model from other subjects to perform the same activity.	Process model of subject
1.2	Every activity has different statistical process model from other activities	Process model of activity
2.1	The feature selection process reduces the number of features	Number of feature
2.2	Feature selection learns characteristic of each activity behavior	(i) The accuracy rate for subject-independent activity recognition without feature selection (ii) The accuracy rate for subject independent activity recognition using feature selection
2.3	Feature selection provides common features for every subject	(i) The accuracy rate for subject-independent activity recognition without feature selection (ii) The accuracy rate for subject-independent activity recognition using feature selection
3.1	When applied as a three-stage process analyzing the performance of different activities for each subject, it can help us identify the most suitable feature set for subject-independent activity recognition	The accuracy rate of the existing activity recognition approaches for subject-independent activity recognition under the exact same setting

Table 3: Result of classification process based on activity using all extracted features.

Activity	Running	Walking	Down Stair	Up Stair	Hopping	Jogging
Running	68%	0.4%	1.5%	0.1%	—	30%
Walking	5%	68%	—	—	—	27%
Down stair	15%	—	65%	—	—	20%
Up stair	7%	11%	2%	69%	3%	6%
Hopping	—	2%	—	25%	65%	8%
Jogging	25%	4%	—	—	—	71%

4.2. Experiment Result. In order to gather mentioned evidences from the experimental design process, several experiment settings are conducted. For every experiment, the same set of data is used. The sample data from all the subjects was divided into training data and testing data, based on subjects. As we mentioned in the data collection section, we have sample data from 27 subjects. For the set of training data, we used the entire data from 20 subjects. The data from the rest of the 7 subjects were used as the testing data.

4.2.1. First Experiment. The first experiment in our research is aimed to evaluate the performance of subject independent activity recognition without feature selection. The features used in this experiment were gathered from the feature extraction process. The feature extraction process results in 66 numbers of features gathered from 20 numbers of subjects.

The result of first experiment which is based on sample data dividing by activity using all extracted feature is shown in Table 3. Based on this result, we can see that using entire 61 features for classification is not effective to understand each different behavior from the same activity performed by different subject. Due to a big number of features it is possible that some features are against each other. From Table 3, we can see that different subjects could have different behaviors to perform the same activity. This difference has been analyzed in Section 3.2 to prove that same activity could fall in different models as the evidences of study Propositions (1.1) and (1.2). Therefore, some activities are misclassified into different activities. For example, due to its different behavior of subject, running activity from 30% subjects was classified as jogging activity.

Same with previous analyses, the first experiment which is based on the entire activities divided by each subject using the entire features of the feature extraction process gives the lower amount of accuracy as seen in Table 4. From this table we can conclude that even performed by same subject, each activity could have different types of statistical model that should be determined.

4.2.2. Second Experiment. In this experiment, we studied the effectiveness of using selected features for subject-independent activity recognition process. From this experiment we would like to see the influence of using the selected features to understand different behavior from the same activity performed by different subject. Compared to the previous experiment, in this experiment we also want to figure out the effectiveness of the selected features.

As we mentioned in Section 3.3, three-stage feature selection process was run to get the influential features for every activity from each subject. By evaluating each feature based on subject and activity in the first stage, 61 numbers of features are selected for the next step of feature selection.

Table 4: Result of classification process based on subject using all extracted features.

Subjects	Running	Walking	Downstairs	Upstairs	Hopping	Jogging
Subject 1	40%	44%	41%	68%	63%	43%
Subject 2	56%	62%	69%	41%	67%	60%
Subject 3	69%	60%	42%	62%	62%	53%
Subject 4	54%	49%	48%	52%	62%	67%
Subject 5	62%	48%	46%	43%	56%	63%
Subject 6	42%	50%	59%	55%	62%	41%
Subject 7	62%	41%	65%	64%	55%	51%

Table 5: Result of classification process based on activity using selected features.

Activity	Running	Walking	Downstairs	Upstairs	Hopping	Jogging
Running	92%	1%	2%	1%	—	4%
Walking	2%	91%	—	—	—	7%
Downstairs	2%	—	93%	—	—	5%
Upstairs	—	8%	—	92%	—	—
Hopping	—	1%	—	7%	92%	—
Jogging	6%	3%	—	—	—	91%

Those features are the features that influence every subject for each activity. By gathering entire selected features, second stage of feature selection which is feature selection based on each subject was run. Every subject that appears in more than ten subjects is selected. That number is chosen based on the number of evaluated subjects which is 50% of the total number of subjects. From second stage of the feature selection process, 35 numbers of features are selected for the next process. The third stage of the feature selection process results in 21 numbers of features including mean, correlation, and process model. Based on the last stage of the feature selection process, not all of the axes are selected. for example from standard deviation features only (standard deviation feature) x-axis is selected. From this result, we are able to get the evidence of study Proposition (2.1) which means that our feature selection process is able to reduce the number of features for activity recognition.

In order to evaluate those selected features, the second experiment was conducted. The selected features based on three-stage feature selection process are used for the second experiment. The result of second experiment which is based on sample data dividing by activity using the selected features from the feature selection process is shown in Table 5. Compared to the result in Table 3, we can see that there is a big improvement of accuracy when we used selected features. From this table we can see that the feature selection process is able to determine the common feature from each activity based on different subjects which can improve the performance of a neural network classifier. From this result, we are able to get the evidence of study Propositions (2.2) and (2.3) which means that the proposed feature selection approach is able to learn about the characteristic of each activity behavior by providing common features for the entire activity of each subject.

The result presented subject-wise, shown in Table 6, shows the same results. From Tables 5 and 6, we can conclude that three-stage feature selection process is able to learn behavior from each sample. Also, the feature selection process is able to give common feature for every subject and activity. This finding supports the fact that understanding each activity behavior from each subject is able to improve the learning process of activity recognition.

4.2.3. Third Experiment. The third experiment is aimed to evaluate the effectiveness of our adopted methodology compared to other previous works. In this experiment, different sets of features and classifier are used based on their own proposed methodology [3, 5]. Different numbers and types of features are used under the same setting used in those related works. In order to compare our approach to the existing approach, we used the same setting for our approach which is subject-independent activity recognition with three-stage feature selection as we mentioned in the second scenario. This experiment setting is conducted to evaluate whether our proposed approach using a certain set of features from feature selection is able to give better performance in subject-independent activity recognition. The comparison result for classification using different classifier is shown in Table 7.

From this result, we can see that our adopted methodology used neural network classifier with 21 numbers of features is able to represent the behavior of each activity from each subject. Table 7 also shows that our proposed approach for subject-independent activity recognition outperforms the existing works. Based on this result, we can conclude that our proposed recognition scheme is able to classify the activity accurately. It shows that our proposed approach is able to learn the data from the subject even though we do not include those subjects not only in the training process of classification but also in feature selection process. This result came out due to the ability of our proposed recognition scheme to learn from new data and its ability to handle the noise of the data. From this result, we are also able to conclude

Table 6: Result of classification process based on subject using selected features.

Subjects	Running	Walking	Down Stair	Up Stair	Hopping	Jogging
Subject 1	93%	93%	91%	93%	93%	91%
Subject 2	94%	91%	91%	91%	93%	94%
Subject 3	91%	92%	91%	94%	94%	93%
Subject 4	90%	89%	91%	91%	93%	94%
Subject 5	93%	91%	92%	92%	93%	94%
Subject 6	93%	89%	94%	93%	92%	91%
Subject 7	89%	91%	92%	93%	94%	92%

Table 7: Accuracy rate of the existing activity recognition approaches and proposed approach for subject-independent activity recognition.

Related Work	Classifier	Number of features	Accuracy rate
Lara and Labrador [3]	Decision Tree	24	77.33%
Ravi et al. [5]	Naïve Bayes	9	73%
	K-Nearest Neighbors (1)	9	58%
	K-Nearest Neighbors (5)	9	66.52%
	K-Nearest Neighbors (10)	9	61.27%
	Decision Tree	9	67%
Adopted method	Neural Network	21	93%

that our adopted methodology is able to perform subject-independent activity recognition. As we can see in Table 7, the previous methods failed to perform subject-independent activity due to its lack of capability to represent the behavior of the entire activity. Moreover, using a big number of features does not mean a wise decision. Using the smallest number of features does not make the classification process give worse performance. Therefore, the problem related to features in subject-independent activity recognition is not only about the number of features but also the effectiveness of the selected features to represent each behavior of entire activities.

Finally, based on the gathered evidence that we need to evaluate the study proposition as we mentioned in Section 4.1, we can conclude that our adopted methodology is able to provide better performance for recognizing subject-independent activity. This goal has been achieved by determining particular model for each activity and understanding different behavior of activity performed by different subjects through analysing the extracted feature and three-stage feature selection process, respectively.

5. Conclusion

Our proposed method uses accelerometer to capture dynamic activity from each subject. Android accelerometer is chosen due to its effective ability to capture movement. Gathered data of activity are performed on raw signal data. In order to classify activity, time-domain features are extracted from those raw signal data. The classification process of dynamic activity is not a trivial problem. It is because every subject has different behavior to perform activity. Our proposed method is able to overcome this problem using three-stage feature selection process using genetic algorithm. Demonstrated experiment shows that feature selection process is able to increase the overall accuracy of activity classification process. The experiment result shows that our proposed approach for subject-independent activity recognition outperforms the existing works.

In this study, we run the entire process of our method using Matlab. Our aim for handling subject-independent activity recognition has been achieved successfully using offline process. Therefore, our future plan includes online activity recognition by using the proposed method.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF), Grant funded by the Korea government (MSIP) (no. 2010-0028631).

References

[1] A. M. Khan, Y.-K. Lee, S. Y. Lee, and T.-S. Kim, "A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 5, pp. 1166–1172, 2010.

[2] S. J. Preece, J. Y. Goulermas, L. P. J. Kenney, and D. Howard, "A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 3, pp. 871–879, 2009.

[3] O. D. Lara and M. A. Labrador, "A mobile platform for real-time human activity recognition," in *Proceedings of the Consumer Communications and Networking Conference (CCNC '12)*, pp. 667–671, IEEE, January 2012.

[4] M. Helmi and S. M. T. AlModarresi, "Human activity recognition using a fuzzy inference system," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, pp. 1897–1902, August 2009.

[5] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence (IAAI '05)*, pp. 1541–1546, July 2005.

[6] W. Liu, X. Li, and D. Huang, "A survey on context awareness," in *Proceedings of the International Conference on Computer Science and Service System (CSSS '11)*, pp. 144–147, June 2011.

[7] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles, "Towards a better understanding of context and context-awareness," in *Handheld and Ubiquitous Computing*, pp. 304–307, Springer, Berlin, Germany, 1999.

[8] S. Lee, S. Park, and S.-G. Lee, "A study on issues in context-aware systems based on a survey and service scenarios," in *Proceedings of the 10th ACIS Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD '09)*, pp. 8–13, May 2009.

[9] A. M. Khan, M. H. Siddiqi, and S. W. Lee, "Exploratory data analysis of acceleration signals to select light-weight and accurate features for real-time activity recognition on smartphones," *Sensors*, vol. 13, no. 10, pp. 13099–13122, 2013.

[10] M. Fahim, I. Fatima, S. Lee, and Y. T. Park, "EFM: evolutionary fuzzy model for dynamic activities recognition using a smartphone accelerometer," *Applied Intelligence*, vol. 39, no. 3, pp. 475–488, 2013.

[11] P. Siirtola and J. Röning, "Recognizing human activities user-independently on smartphones based on accelerometer data," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 1, no. 5, pp. 38–45, 2012.

[12] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[13] M. Luštrek and B. Kaluža, "Fall detection and activity recognition with machine learning," *Informatica*, vol. 33, no. 2, pp. 197–204, 2009.

[14] A. Palaniappan, R. Bhargavi, and V. Vaidehi, "Abnormal human activity recognition using SVM based approach," in *International Conference on Recent Trends In Information Technology (ICRTIT '12)*, pp. 97–102, IEEE, April 2012.

[15] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to Time Series Analysis and Forecasting*, vol. 526, John Wiley & Sons, 2011.

[16] W. S. William and S. Wei, *Time Series Analysis: Univariate and Multivariate Methods*, 1990.

[17] V. Cuomo, V. Lapenna, M. Macchiato, and C. Serio, "Autoregressive models as a tool to discriminate chaos from randomness in geoelectrical time series: an application to earthquake prediction," *Annali di Geofisica*, vol. 40, no. 2, pp. 385–400, 1997.

[18] R. P. Haining, "The moving average model for spatial interaction," in *Transactions of the Institute of British Geographers*, pp. 202–225, 1978.

[19] C. Chatfield, *The Analysis of Time Series: An Introduction*, CRC Press, 2003.

[20] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, John Wiley & Sons, 2013.

[21] O. D. Anderson, "Time series analysis and forecasting: another look at the Box-Jenkins approach," in *The Statistician*, pp. 285–303, 1977.

[22] J. D. Hamilton, *Time Series Analysis*, vol. 2, Princeton University Press, Princeton, NJ, USA, 1994.

[23] L. Ladha and T. Deepa, "Feature selection methods and algorithms," *International Journal on Computer Science and Engineering*, vol. 3, no. 5, pp. 1787–1797, 2011.

[24] B. Yu and B. Yuan, "A more efficient branch and bound algorithm for feature selection," *Pattern Recognition*, vol. 26, no. 6, pp. 883–889, 1993.

[25] A. A. Javadi, R. Farmani, and T. P. Tan, "A hybrid intelligent genetic algorithm," *Advanced Engineering Informatics*, vol. 19, no. 4, pp. 255–262, 2005.

[26] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 2, pp. 164–171, 2000.

[27] J. T. Alander, "On optimal population size of genetic algorithms," in *Proceedings of the Computer Systems and Software Engineering (CompEuro '92)*, pp. 65–70, IEEE, May 1992.

[28] P. J. Drew and J. R. T. Monson, "Artificial neural networks," *Surgery*, vol. 127, no. 1, pp. 3–11, 2000.

[29] E. Alpaydin, *Introduction to Machine Learning*, 2nd edition, 2010.

[30] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Pervasive Computing*, pp. 1–17, Springer, Berlin, Germany, 2004.

[31] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[32] S. Ishak and C. Alecsandru, "Optimizing traffic prediction performance of neural networks under various topological, input, and traffic condition settings," *Journal of Transportation Engineering*, vol. 130, no. 4, pp. 452–465, 2004.

[33] S. W. Lee and D. C. Rine, "Case study methodology designed research in software engineering methodology validation," in *Proceedings of the 16th International Conference on Software Engineering and Knowledge Engineering (SEKE '04)*, pp. 117–122, 2004.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 842675, 10 pages
http://dx.doi.org/10.1155/2014/842675

Research Article

The Security Weakness of Block Cipher Piccolo against Fault Analysis

Junghwan Song, Kwanhyung Lee, and Younghoon Jung

Department of Mathematics, Hanyang University, Seoul 133-791, Republic of Korea

Correspondence should be addressed to Younghoon Jung; sky1236@hanyang.ac.kr

Received 23 December 2013; Accepted 21 January 2014; Published 13 March 2014

Academic Editor: Jongsung Kim

Piccolo is a 64-bit lightweight block cipher which is able to be implemented in constrained hardware environments such as a wireless sensor network. Fault analysis is a type of side channel attack and cube attack is an algebraic attack finding sufficiently low-degree polynomials in a cipher. In this paper, we show a fault analysis on the Piccolo by using cube attack. We find 16 linear equations corresponding to a round function F by cube attack, which are used to fault analysis. Our attack has the complexity of $2^{8.49}$ and $2^{9.21}$ encryptions with fault injections of target bit positions into Piccolo-80 and Piccolo-128, respectively. And our attack needs $2^{20.86}$ and $2^{21.60}$ encryptions with *random* 4-bit fault injections for Piccolo-80 and Piccolo-128, respectively.

1. Introduction

Fault analysis is a type of side channel attack. This analysis was introduced by Boneh et al. in [1]. Differential fault analysis (DFA), which is an improved method of fault analysis, was introduced by Biham and Shamir in [2]. DFA is applied to various block ciphers such as AES [3, 4], ARIA [5], SEED [6], CLEFIA [7], LED [8], Piccolo [9–12], PRESENT [13], and KATAN32 [14]. Cube attack was introduced by Dinur and Shamir in [15]. This attack is an algebraic attack by finding sufficiently low-degree polynomials in a cipher. Cube attack is applied to various cryptosystems such as block cipher [16] and stream cipher [15, 17].

In CHES 2011, Piccolo was introduced by Shibutani et al. in [18]. Piccolo is a block cipher which supports 80-bit and 128-bit secret key size. In this paper, we analyze two versions of Piccolo [18] with fault analysis by using cube attack. In ISPEC 2012, fault analysis using cube attack was introduced by Abdul-Latip et al. in [16]. In this paper, we apply this method on Piccolo-80 and Piccolo-128. As a result, we find 16 linear equations corresponding to a round function F by cube attack, which are used to fault analysis.

Piccolo is analyzed by various techniques. In ISPEC 2012, Wang et al. suggest biclique cryptanalysis of reduced round Piccolo in [19]. They analyze reduced version of Piccolo-80 without postwhitening keys XOR and reduced 28-round Piccolo-128 without prewhitening keys XOR. In 2013, Song et al. suggest biclique cryptanalysis of full rounds of Piccolo [20]. And also Jeong suggests a differential fault analysis of full rounds of Piccolo [9].

In this paper, we show a fault analysis on the Piccolo by using cube attack. We find 16 linear equations corresponding to a round function F of Piccolo by using cube attack. These equations are used to our attack. In this paper, we describe the case that an adversary injects random 4-bit faults. Our attack has the complexity of $2^{20.86}$ and $2^{21.60}$ encryptions for Piccolo-80 and Piccolo-128, respectively, while the assumption of [9] is an adversary that injects random byte faults. Reference [9] has the complexity of 2^{24} and 2^{40} encryptions for Piccolo-80 and Piccolo-128, respectively. Our attack has a lower computational complexity than [9], even though the assumption of fault injection in our attack differs from [9].

In Section 2, we briefly describe the procedures of cube attack and cube tester. And then we describe the brief specifications of Piccolo in Section 3. In Section 4, a method of fault analysis of Piccolo by using cube attack is presented. Finally, our conclusions are in Section 5.

2. Cube Attack and Cube Tester

Algebraic attack is to find a solution, which is the key, of a system of equations that represent target cipher with

given plaintext and the corresponding ciphertext, that is representing cipher as a system of equations with multiple variables defined over finite field where each key bit is represented as a variable in the system. Solving the system is equivalent to finding the secret key of the target cipher. Cube attack is an algebraic attack finding sufficiently low-degree polynomials in cipher.

2.1. Cube Attack. Cube attack was introduced by Dinur and Shamir in [15]. Cube attack is a chosen plaintext attack. The main idea of cube attack is to find linear equations consisting of secret variables by using cube sum. Let $p(v_1, \ldots, v_n, k_1, \ldots, k_m)$ be a polynomial derived from a cipher, where $v_1, \ldots, v_n$ are public variables and $k_1, \ldots, k_m$ are secret variables. In other words, each secret variable is considered a bit in secret key and each public variable is considered a bit in plaintext or internal state. Let $I = \{I_1, \ldots, I_s\} \subseteq \{1, \ldots, n\}$ be a set and let t_I be the monomial $x_{I_1} x_{I_2} \ldots x_{I_s}$. Note that the set in terms of I is called cube index. Then the polynomial p is represented by three polynomials t_I, $p_{S(I)}$, and q as the following form:

$$\begin{aligned} &p(v_1, \ldots, v_n, k_1, \ldots, k_m) \\ &\quad = t_I \cdot p_{S(I)} + q(v_1, \ldots, v_n, k_1, \ldots, k_m), \end{aligned} \tag{1}$$

where q is not consisting of a monomial which has a factor t_I.

Cube attack is required to check the linearity of $p_{S(I)}$ which is called superpoly. A superpoly $p_{S(I)}$ is called a maxterm if $p_{S(I)}$ is linear. We use the following cube sum to find a $p_{S(I)}$:

$$p_{S(I)} = \sum_{(v_{I_1}, \ldots, v_{I_s}) \in \mathrm{GF}(2)^s} p(v_1, \ldots, v_n, k_1, \ldots, k_m), \tag{2}$$

where plaintext bits except cube index $(v_i, i \in \{1, \ldots, n\} - I)$ are fixed as constants.

As the above representation, cube is completed with the sum total 2^S pairs of plaintext and ciphertext for a cube index $I = \{I_1, \ldots, I_s\}$. To check whether $p_{S(I)}$ is a maxterm, linearity test is required. Let $p_{S(I)}(k_1, \ldots, k_m)$ be a polynomial of m variables over GF(2). Let t be the number of tests. The following is a procedure of linearity test.

Step 1. Choose 2 random vectors $x, y \in \mathrm{GF}(2)^m$.

Step 2. If $p_{S(I)}(x) \oplus p_{S(I)}(y) \oplus p_{S(I)}(0) \neq p_{S(I)}(x \oplus y)$, then $p_{S(I)}$ is not linear. Stop the test.

Step 3. Repeat Steps 1 and 2, t times.

Step 4. $p_{S(I)}$ is linear. Stop the test, where $0 = (0, \ldots, 0) \in \mathrm{GF}(2)^m$.

If $p_{S(I)}(k_1, \ldots, k_m)$ is linear, the above equation in Step 2 is always correct for all inputs $x, y \in \mathrm{GF}(2)^m$. Because checking all inputs is impossible, an upper bound of number of linearity tests has to be set. If there are at most d_1 elements in a cube index for testing linearity, at most $2^{d_1} \times (3 \times t + 1)$ pairs of plaintext and ciphertext are needed. Cube attack consists of preprocessing phase and online phase. Preprocessing phase is to find a system of linear equations by using cube sum and linearity test. Online phase is recovering the master key stored by using an encryption oracle. The following are details for the two phases.

Preprocessing Phase. After finding a polynomial from a cipher, find a cube, that is, a maxterm, by using linearity test. Since we know output after all plaintext bits are entered in encryption oracle, fix plaintext except cube index as a constant. Fixed constants of every cube do not have to be equal. Let f_i be a maxterm which consists of only secret variables $k_1, \ldots, k_m$ and let b_i be the value of the maxterm f_i which is found from online phase. We consider the following system of equations:

$$\begin{aligned} f_1(k_1, \ldots, k_m) &= b_1 \\ &\vdots \\ f_l(k_1, \ldots, k_m) &= b_l. \end{aligned} \tag{3}$$

In the preprocessing phase, find enough maxterms to recover the master key and precalculate this system of equations by using Gaussian elimination. If we find m linear independent maxterms, then recover all the master keys with m^3 operations for recovery by using Gaussian elimination. In general, it is lower than complexity $l \times 2^{d_1} \times (3 \times t + 1)$ for finding l maxterms. Let $f_1, \ldots, f_m$ be linearly independent. Then the master keys are represented as the following system:

$$\begin{aligned} k_1 &= \sum_{i=1}^{m} a_{1,i} \cdot b_i \\ &\vdots \\ k_m &= \sum_{i=1}^{m} a_{m,i} \cdot b_i, \end{aligned} \tag{4}$$

where $a_{i,j} \in \mathrm{GF}(2)$.

Online Phase. In online phase, calculate the value of cube sum from an encryption oracle by using the cube that has been found in the preprocessing phase. Let each plaintext bit not in the cube be constant. The calculated value is b_i, that is, the value of the maxterm. By substituting the value b_i into (4), we recover the master key. Let cube index found at preprocessing phase have at most d_2 elements. Then the complexity of online phase is $m \times 2^{d_2}$.

2.2. Cube Tester. Cube attack finds a maxterm by testing linearity of $p_{S(I)}$ of a given polynomial p and cube index I. Cube tester distinguishes a polynomial from a random polynomial by many tests including linearity test. There are some other tests using cube sum in [21]. In cube attack, a plaintext bit not in the cube is fixed as a constant. However all bits not in the cube have to be considered variables in the cube tester. Since the purpose of using the cube tester is getting information, which are properties of polynomial, we use the

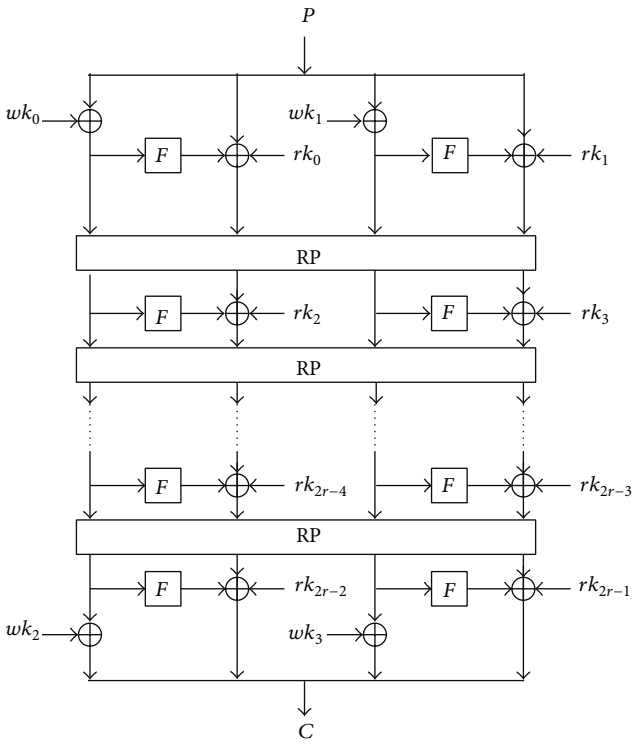


Figure 1: Encryption process of Piccolo.

low-degree test that is in [21]. The degree N is determined by low-degree test. Let I be a cube index, let J be the number of bits not in the cube index (i.e., $J = n + m - S$; $p_{S(I)}$ consists of J variables), and t is the number of tests. Since low-degree test is valid only when $p(0) = 0$ for the given polynomial p, we define $p^*_{S(I)}(x) = p_{S(I)}(x) + p_{S(I)}(0)$. Then low-degree test for the polynomial $p^*_{S(I)}(x)$ is as follows.

Step 1. Choose $N + 1$ random vectors $y_1, \ldots, y_{N+1} \in \mathrm{GF}(2)^J$.

Step 2. If $\sum_{\phi \neq S \subset \{y_1, \ldots, y_{N+1}\}} p^*_{S(I)}(\sum_{y_i \in S} y_i) \neq 0$, then degree of $p_{S(I)} > N$. Stop the test.

Step 3. Repeat Steps 1 and 2, t times.

Step 4. Degree of $p_{S(I)} \leq N$. Stop the test.

If $N = 1$, then the low-degree test is similar to the linearity test. We use the idea of the cube tester which uses every bit not in the cube index (consisting of plaintext and the master key) as a variable.

3. Description of Piccolo

Piccolo is a 64-bit block cipher with 80- and 128-bit key size. The structure of Piccolo is a Feistel network. Piccolo-80 consists of 25 rounds and Piccolo-128 consists of 31 rounds. Figure 1 illustrates the working processing of Piccolo. Each round consists of two functions, round function F and round permutation RP. The round functions F and RP are as follows.

Table 1: S-box of Piccolo.

x	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
$S(x)$	e	4	b	2	3	8	0	9	1	a	7	f	6	c	5	d

Round Function F. The round function F is defined by

$$F(x_0, x_1, x_2, x_3) = (S(x_0), S(x_1), S(x_2), S(x_3)) \cdot M \cdot (S(x_0), S(x_1), S(x_2), S(x_3))^t, \quad (5)$$

where X^t is the transposition of X.

$S(x)$ is the 4-bit S-box and M is the diffusion matrix as follows (see Table 1):

$$M = \begin{pmatrix} 2 & 3 & 1 & 1 \\ 1 & 2 & 3 & 1 \\ 1 & 1 & 2 & 3 \\ 3 & 1 & 1 & 2 \end{pmatrix}. \quad (6)$$

The multiplications between M and vectors are defined by an irreducible polynomial $x^4 + x + 1$ over $\mathrm{GF}(2^4)$.

Round Permutation RP. The round permutation RP is defined by

$$\mathrm{RP}(x_0, x_1, \ldots, x_7) = (x_2, x_7, x_4, x_1, x_6, x_3, x_0, x_5), \quad (7)$$

where x_i is byte.

For description of Piccolo and our attack, we denote intermediate variables before r-round as $X^r = X^r_0 \mid X^r_1 \mid X^r_2 \mid X^r_3 = (x^r_0, \ldots, x^r_{63})$ and intermediate variables before r-round's RP function as $Y^r = Y^r_0 \mid Y^r_1 \mid Y^r_2 \mid Y^r_3$ (i.e., $\mathrm{RP}(Y^r) = X^{r+1}$). Let the round function F in r-round be $F(X^r_i) = (F_0(X^r_i), \ldots, F_{15}(X^r_i))$ and let the round key be $rk_r = (k^r_0, \ldots, k^r_{15})$. The other notations are as follows:

$(X^r_i)^L$: left 8 bits of X^r_i and $(X^r_i)^R$: right 8 bits of X^r_i;

K^L_i: left 8 bits of K_i and K^R_i: right 8 bits of K_i;

$A \mid B$: concatenation of A and B.

Let the 64-bit plaintext and ciphertext be P and C, respectively. Encryption of Piccolo is defined as follows:

(1) $P_0 \mid P_1 \mid P_2 \mid P_3 \leftarrow P$ (P_i is the 16-bit plaintext);

(2) $X^1_0 \leftarrow P_0 \oplus wk_0$, $X^1_1 = P_1$
$X^1_2 \leftarrow P_2 \oplus wk_1$, $X^1_3 = P_3$;

(3) for $i = 1$ to $r - 1$

$$Y^i_0 \leftarrow X^i_0, Y^i_1 \leftarrow X^i_1 \oplus F(X^i_0) \oplus rk_{2i-2}$$
$$Y^i_2 \leftarrow X^i_2, Y^i_3 \leftarrow X^i_3 \oplus F(X^i_2) \oplus rk_{2i-1}$$
$$X^{i+1} \leftarrow \mathrm{RP}(Y^i);$$

(4) $Y^r_0 \leftarrow X^r_0 \oplus wk_2, Y^r_1 \leftarrow X^r_1 \oplus F(X^r_0) \oplus rk_{2r-2}$
$Y^r_2 \leftarrow X^r_2 \oplus wk_3, Y^r_3 \leftarrow X^r_3 \oplus F(X^r_2) \oplus rk_{2r-1}$;

(5) $C \leftarrow Y^r_0 \mid Y^r_1 \mid Y^r_2 \mid Y^r_3$ (Y^r_i is the 16-bit ciphertext).

Key schedule of Piccolo consists of the following.

Piccolo-80:

$$wk_0 \longleftarrow K_0^L \mid K_1^R, wk_1 \longleftarrow K_1^L \mid K_0^R,$$
$$wk_2 \longleftarrow K_4^L \mid K_3^R, wk_3 \longleftarrow K_3^L \mid K_4^R, \quad (8)$$

for $i \leftarrow 0$ to 24 do

if $i \mod 5 = 0$ or 2, then

$$(rk_{2i}, rk_{2i+1}) \longleftarrow \left(\mathrm{con}_{2i}^{80}, \mathrm{con}_{2i+1}^{80}\right) \oplus (K_2, K_3) \quad (9)$$

if $i \mod 5 = 1$ or 4, then

$$(rk_{2i}, rk_{2i+1}) \longleftarrow \left(\mathrm{con}_{2i}^{80}, \mathrm{con}_{2i+1}^{80}\right) \oplus (K_0, K_1) \quad (10)$$

if $i \mod 5 = 3$, then

$$(rk_{2i}, rk_{2i+1}) \longleftarrow \left(\mathrm{con}_{2i}^{80}, \mathrm{con}_{2i+1}^{80}\right) \oplus (K_4, K_4), \quad (11)$$

where con_i^{80} is the round constant.

Piccolo-128:

$$wk_0 \longleftarrow K_0^L \mid K_1^R, wk_1 \longleftarrow K_1^L \mid K_0^R,$$
$$wk_2 \longleftarrow K_4^L \mid K_7^R, wk_3 \longleftarrow K_7^L \mid K_4^R, \quad (12)$$

for $i \leftarrow 0$ to 61 do

if $(i + 2) \mod 8 = 0$, then

$$(K_0, K_2, K_6, K_4) \longleftarrow (K_2, K_6, K_4, K_0)$$
$$(K_3, K_7, K_5) \longleftarrow (K_7, K_5, K_3) \quad (13)$$
$$rk_i \longleftarrow rk_{(i+2) \bmod 8} \oplus \mathrm{con}_i^{128},$$

where con_i^{128} is the round constant.

Since key schedule of Piccolo is just performing XOR determined constants to the master key, recovering the round key and recovering the master key are the same. Table 2 is showing the master key used for the round key of Piccolo. Detailed descriptions of Piccolo are in [18].

4. Fault Analysis on the Piccolo

In this section, we show the fault analysis for Piccolo-80 and Piccolo-128. We assume that an adversary is able to make 4-bit errors in a maximum at a time on a round during an encryption process. By using cube sum, find system of linear equations in the common F of Piccolo-80 and Piccolo-128. And use the system to represent the phase recovering the master key of Piccolo-80 and Piccolo-128. Analysis of a round function F in Section 4.1 is corresponding to the preprocessing phase of cube attack. The attack in Sections 4.2 and 4.3 is the case of an encryption oracle that is given and is corresponding to online phase.

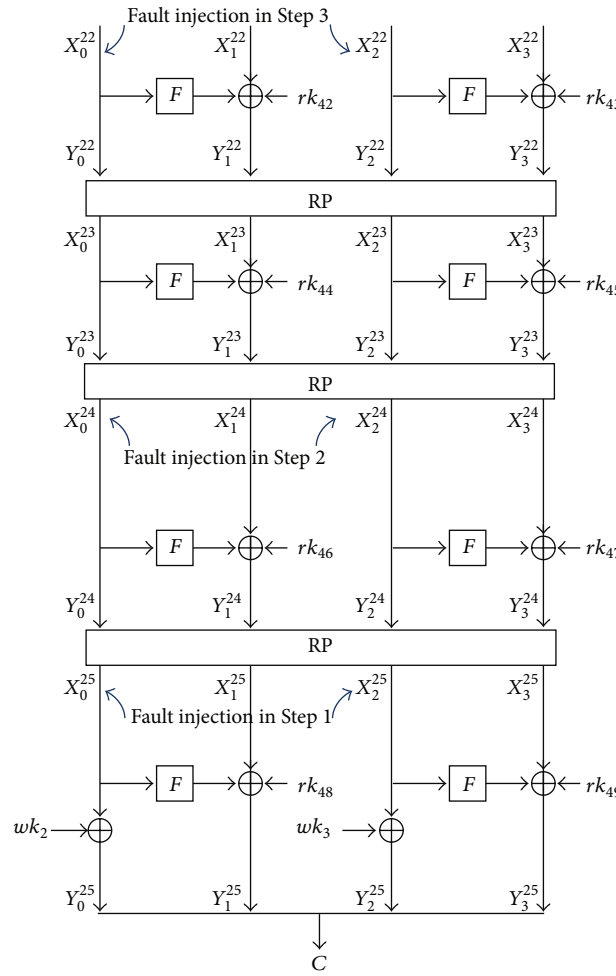


Figure 2: Fault analysis of Piccolo-80.

4.1. Equations of Round Function F. Since round function F is the same for Piccolo-80 and Piccolo-128, the results of fault injection attack on F are the same in the both algorithms. Let $F(X) = (F_0(X), \ldots, F_{15}(X))$, where $X = (x_0, x_1, \ldots, x_{15})$ is an 16-bit intermediate value and each $F_j(x)$ is a bit ($F : \mathrm{GF}(2)^{16} \rightarrow \mathrm{GF}(2)^{16}$). We test all possible cubes of degree 1 to degree 4 and all possible inputs for each cube. We get many linear polynomials and choose 16 appropriate polynomials for recovering the master key. Table 3 shows our selected 16 polynomials, cube index, and output bit (F_i).

4.2. Analysis on the Piccolo-80. We explain how to recover all the master keys of Piccolo-80. By key schedule of Piccolo-80, recovering wk_2, wk_3, rk_{44}, rk_{48}, and rk_{49} is equal to recovering all the master keys of Piccolo-80. Let plaintext P be given and let X_i^j, Y_i^j be intermediate values for plaintext P. In this paper, we recover the master key of Piccolo-80 by recovering some X_i^js. Figure 2 is for the last 4 rounds of Piccolo-80. The following is the attack on Piccolo-80.

Step 1. First, we analyze the last round (i.e., round 25). Perform cube sum by using the cube in Table 3 for F which

Table 2: Round key of Piccolo.

Piccolo-80			Piccolo-128		
Round	Round key	Master key	Round	Round key	Master key
First	wk_0, wk_1	$K_0^L \mid K_1^R, K_1^L \mid K_0^R$	First	wk_0, wk_1	$K_0^L \mid K_1^R, K_1^L \mid K_0^R$
1	rk_0, rk_1	K_2, K_3	1	rk_0, rk_1	K_2, K_3
2	rk_2, rk_3	K_0, K_1	2	rk_2, rk_3	K_4, K_5
3	rk_4, rk_5	K_2, K_3	3	rk_4, rk_5	K_6, K_7
4	rk_6, rk_7	K_4, K_4	4	rk_6, rk_7	K_2, K_1
5	rk_8, rk_9	K_0, K_1	5	rk_8, rk_9	K_6, K_7
6	rk_{10}, rk_{11}	K_2, K_3	6	rk_{10}, rk_{11}	K_0, K_3
7	rk_{12}, rk_{13}	K_0, K_1	7	rk_{12}, rk_{13}	K_4, K_5
8	rk_{14}, rk_{15}	K_2, K_3	8	rk_{14}, rk_{15}	K_6, K_1
9	rk_{16}, rk_{17}	K_4, K_4	9	rk_{16}, rk_{17}	K_4, K_5
10	rk_{18}, rk_{19}	K_0, K_1	10	rk_{18}, rk_{19}	K_2, K_7
11	rk_{20}, rk_{21}	K_2, K_3	11	rk_{20}, rk_{21}	K_0, K_3
12	rk_{22}, rk_{23}	K_0, K_1	12	rk_{22}, rk_{23}	K_4, K_1
13	rk_{24}, rk_{25}	K_2, K_3	13	rk_{24}, rk_{25}	K_0, K_3
14	rk_{26}, rk_{27}	K_4, K_4	14	rk_{26}, rk_{27}	K_6, K_5
15	rk_{28}, rk_{29}	K_0, K_1	15	rk_{28}, rk_{29}	K_2, K_7
16	rk_{30}, rk_{31}	K_2, K_3	16	rk_{30}, rk_{31}	K_0, K_1
17	rk_{32}, rk_{33}	K_0, K_1	17	rk_{32}, rk_{33}	K_2, K_7
18	rk_{34}, rk_{35}	K_2, K_3	18	rk_{34}, rk_{35}	K_4, K_3
19	rk_{36}, rk_{37}	K_4, K_4	19	rk_{36}, rk_{37}	K_6, K_5
20	rk_{38}, rk_{39}	K_0, K_1	20	rk_{38}, rk_{39}	K_2, K_1
21	rk_{40}, rk_{41}	K_2, K_3	21	rk_{40}, rk_{41}	K_6, K_5
22	rk_{42}, rk_{43}	K_0, K_1	22	rk_{42}, rk_{43}	K_0, K_7
23	rk_{44}, rk_{45}	K_2, K_3	23	rk_{44}, rk_{45}	K_4, K_3
24	rk_{46}, rk_{47}	K_4, K_4	24	rk_{46}, rk_{47}	K_6, K_1
25	rk_{48}, rk_{49}	K_0, K_1	25	rk_{48}, rk_{49}	K_4, K_3
Final	wk_2, wk_3	$K_4^L \mid K_3^R, K_3^L \mid K_4^R$	26	rk_{50}, rk_{51}	K_2, K_5
			27	rk_{52}, rk_{53}	K_0, K_7
			28	rk_{54}, rk_{55}	K_4, K_1
			29	rk_{56}, rk_{57}	K_0, K_7
			30	rk_{58}, rk_{59}	K_6, K_3
			31	rk_{60}, rk_{61}	K_2, K_5
			Final	wk_2, wk_3	$K_4^L \mid K_7^R, K_7^L \mid K_4^R$

takes X_0^{25}. For example, consider 6th equation of Table 3. Suppose that inject fault into x_4^{25} to x_8^{25}. Then, since fault is injected into only X_0^{25}, value of X_1^{25} or rk_{48} is not changed. We notate the following to explain our attack:

$X_0^{25} = (x_0^{25}, \ldots, x_{15}^{25}), X_1^{25} = (x_{16}^{25}, \ldots, x_{31}^{25})$;

$Y_0^{25} = (y_0^{25}, \ldots, y_{15}^{25})$;

$y_{12}^{25}[x_4^{25}]$: y_{12}^{25} when fault is injected into x_4^{25};

$y_{12}^{25}[x_8^{25}]$: y_{12}^{25} when fault is injected into x_8^{25};

$y_{12}^{25}[x_4^{25}, x_8^{25}]$: y_{12}^{25} when fault is injected into both x_4^{25}, x_8^{25}.

We calculate cube sum for cube index $\{4, 8\}$ like the following:

$$\begin{aligned}\text{Cube sum} &= \sum_{x_4, x_8 \in \{0,1\}} F_{12}\left(x_0^{25}, \ldots, x_{15}^{25}\right) \\ &= \sum_{x_4, x_8 \in \{0,1\}} \left[F_{12}\left(x_0^{25}, \ldots, x_{15}^{25}\right) \oplus x_{28}^{25} \oplus k_{12}^{48}\right] \\ &= y_{12}^{25} \oplus y_{12}^{25}\left[x_4^{25}\right] \oplus y_{12}^{25}\left[x_8^{25}\right] \oplus y_{12}^{25}\left[x_4^{25}, x_8^{25}\right].\end{aligned} \tag{14}$$

y_{12}^{25} is not the output of F. But since cube sum does XOR even times, x_{28}^{25} and rk_{12}^{48} are offset. That is, we know value of cube sum cause of $Y_0^{25} \mid Y_1^{25} \mid Y_2^{25} \mid Y_3^{25} = C$. In the same way, cube sum using fault injection in this paper is performed. By performing cube sum for every cube in Table 3, we get 16 systems of equations. Recover input X_0^{25}.

Table 3: Cube sum result of $F(x_0, \ldots, x_{15})$.

Cube index	Outbit (F_i)	Polyequation
1, 5, 6	8	$x_0 + 1$
0, 8, 9, 11	10	$x_1 + 1$
1, 5, 6	12	$x_0 + x_2$
0, 8, 9, 11	7	x_3
0, 1, 5, 6	4	$x_4 + 1$
4, 8	12	$x_5 + x_9$
4, 5, 8, 9	4	$x_6 + 1$
4, 8, 9, 11	7	$x_5 + x_7 + 1$
5, 6, 9	12	$x_8 + 1$
4, 5, 7, 8	6	x_9
5, 6, 9	8	$x_{10} + 1$
4, 5, 7, 8	3	x_{11}
5, 6, 13	8	$x_{12} + x_{14}$
0, 12	4	$x_1 + x_{13}$
5, 6, 13	12	$x_{14} + 1$
0, 8, 11, 12	7	$x_3 + x_{15}$

Similarly, we recover input X_2^{25} using F which takes X_2^{25}. Since $X_0^{25} \oplus wk_2 = Y_0^{25}$, $X_2^{25} \oplus wk_3 = Y_2^{25}$, we recover wk_2, wk_3 (i.e., K_3, K_4).

Step 2. Since we know wk_2 and wk_3, calculate intermediate value X_0^{25}, X_2^{25} for given ciphertext. Round permutation RP in round 24 is as follows:

$$\begin{aligned} Y_1^{24} &= \left(X_0^{25}\right)^L \mid \left(X_2^{25}\right)^R, \qquad Y_3^{24} = \left(X_2^{25}\right)^L \mid \left(X_0^{25}\right)^R \\ Y_0^{24} &= \left(X_3^{25}\right)^L \mid \left(X_1^{25}\right)^R, \qquad Y_2^{24} = \left(X_1^{25}\right)^L \mid \left(X_3^{25}\right)^R. \end{aligned} \tag{15}$$

Therefore, we calculate Y_1^{24}, Y_3^{24} for given ciphertext. By using this, analyze round 24. In a similar way with Step 1, recover X_0^{24}, X_2^{24} by using the cube in Table 3 for F which takes X_0^{24}, X_2^{24}. Since $X_0^{24} = Y_0^{24}$, $X_2^{24} = Y_2^{24}$, $Y_0^{24} = (X_3^{25})^L \mid (X_1^{25})^R$, and $Y_2^{24} = (X_1^{25})^L \mid (X_3^{25})^R$, we recover X_1^{25}, X_3^{25}. Then we recover rk_{48}, rk_{49} (i.e., K_0, K_1) since $X_1^{25} \oplus F(X_0^{25}) \oplus rk_{48} = Y_1^{25}$, $X_3^{25} \oplus F(X_2^{25}) \oplus rk_{49} = Y_3^{25}$.

Step 3. We recover K_0, K_1, K_3, and K_4 so far. Given ciphertext C, we calculate X_0^{24}, X_1^{24}, X_2^{24}, and X_3^{24}. That is, we recover X_0^{23}, X_2^{23}, Y_1^{23}, and Y_3^{23}. We want to recover X_1^{23}. Since $X_0^{22} = Y_0^{22} = (X_3^{23})^L \mid (X_1^{23})^R$, $X_2^{22} = Y_2^{22} = (X_1^{23})^L \mid (X_3^{23})^R$, if we recover right 8 bits of X_0^{22} and left 8 bits of X_2^{22}, then we recover X_1^{23}. To recover right 8 bits of X_0^{22}, inject fault into bits corresponding to cube index of last 8 equations in Table 3. For recovering left 8 bits of X_2^{22}, inject fault into bits corresponding to cube index of first 8 equations in Table 3. Then we recover X_1^{23}. Since $X_1^{23} \oplus F(X_0^{23}) \oplus rk_{44} = Y_1^{23}$,we recover rk_{44} (i.e., K_2).

Use the above 3 steps to recover all the master keys used for Piccolo-80. This needs the assumption that we inject fault into at most 4 bits in the same time. Thus in this paper we consider the following as an analyzing way.

Table 4: Attack complexity of Piccolo-80.

Assumption	Required fault	Complexity
Assumption 1	$132 \approx 2^{7.04}$	2^{48}
Assumption 2	$264 \approx 2^{8.04}$	$2^{16.01}$
Assumption 3	$347 \approx 2^{8.44}$	$359.1 \approx 2^{8.49}$

Assumption 1. We inject fault into at most 4 bits in the same time. But, apply this to only last round.

Assumption 2. We inject fault into at most 4 bits in the same time. But, apply this to only rounds 24 and 25.

Assumption 3. We inject fault into at most 4 bits in the same time.

That is, suppose that we analyze only Step 1 or Steps 1 and 2. Even though we analyze only Steps 1 and 2, since at least 32 bits of 80-bit master key are recovered, we recover all the master keys with less operations than brute-force attack. To recover X_i^r, inject fault into 11 bits among 16 bits of internal state X_i^r by using injections 66 times. To recover left 8 bits and right 8 bits of X_i^r (i.e., $(X_i^r)^L, (X_i^r)^R$), we inject fault into 8 bits and 10 bits among 16 bits of X_i^r, respectively. Then $(X_i^r)^L, (X_i^r)^R$ are recovered by using injections 44 and 39 times, respectively.

Under Assumption 1, we need 133 encryptions for recovering 32-bit master key (wk_2, wk_3). We exclusively search to recover remaining 48-bit master key. Therefore, we need total $133 + 2^{48} \approx 2^{48}$ encryptions for recovering the master key under Assumption 1.

Under Assumption 2, we need 133 encryptions for recovering 32-bit master key (wk_2, wk_3). For recovering 32-bit round keys rk_{48} and rk_{49}, we have to calculate Y_1^{24}, Y_3^{24}. Given ciphertext, calculating Y_1^{24} is equivalent to 0.5 round encryption. So is Y_3^{24}. Hence, we need $132 + (132 \times 0.5 + 1)/25$ encryptions for recovering 32-bit round keys rk_{48} and rk_{49}. We exclusively search to recover remaining 16-bit master key. Therefore, we need total $133 + (132 + 67/25) + 2^{16} \approx 2^{16.01}$ encryptions for recovering the master key under Assumption 2.

Under Assumption 3, we need $133+(132+67/25)$ encryptions for recovering wk_2, wk_3, rk_{48}, and rk_{49}. Given ciphertext, calculating Y_1^{22} is equivalent to 2.5 round encryption. We need $44 + 39 + \{(44 + 39) \times 2.5 + 1 \times 3\}/25$ encryptions for recovering 32-bit round keys rk_{44}, rk_{45}. Therefore, we need total $133+(132+67/25)+(83+210.5/25) \approx 2^{8.49}$ encryptions for recovering the master key under Assumption 3. Table 4 is showing encryption complexity needed for recovering the master key of each assumption.

4.3. Analysis on the Piccolo-128. Piccolo-128 recovers the master key in a similar way to Piccolo-80. Figure 3 is for the last 5 rounds of Piccolo-128. The following is how Piccolo-128 recovers the master key.

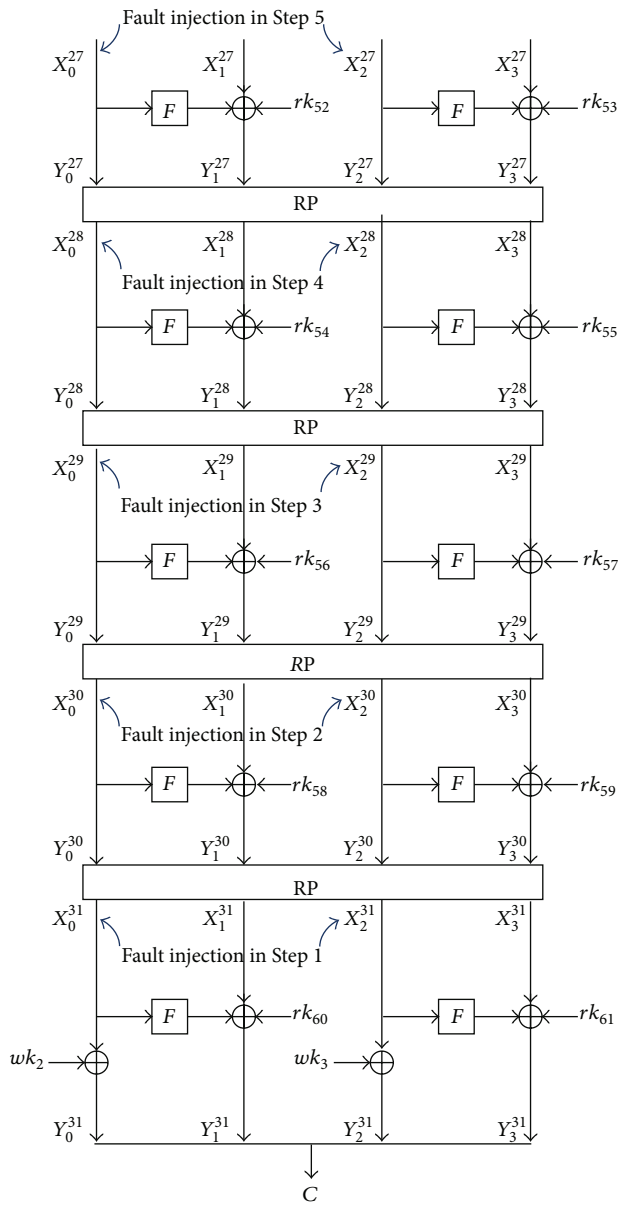


Figure 3: Fault analysis of Piccolo-128.

Step 1. First, we analyze the last round (i.e., round 25). In a similar way with Step 1 in analysis of Piccolo-80, recover X_0^{31}, X_2^{31} by using the cube in Table 3 for F which takes X_0^{31}, X_2^{31}. Since $X_0^{31} \oplus wk_2 = Y_0^{31}$, $X_2^{31} \oplus wk_3 = Y_2^{31}$, we recover wk_2, wk_3 (K_4, K_7).

Step 2. Since we know wk_2 and wk_3, calculate intermediate values X_0^{31}, X_2^{31} for given ciphertext. Since $Y_1^{30} = (X_0^{31})^L \mid (X_2^{31})^R$, $Y_3^{30} = (X_2^{31})^L \mid (X_0^{31})^R$, we calculate Y_1^{30}, Y_3^{30} for given ciphertext. In a similar way with Step 1 in analysis of Piccolo-80, recover X_0^{30}, X_2^{30} by using the cube in Table 3 for F which takes X_0^{30}, X_2^{30}. We recover X_1^{31}, X_3^{31}, since $X_0^{30} = Y_0^{30}$, $X_2^{30} = Y_2^{30}$, and $Y_0^{30} = (X_3^{31})^L \mid (X_1^{31})^R$, $Y_2^{30} = (X_1^{31})^L \mid (X_3^{31})^R$. Since $X_1^{31} \oplus F(X_0^{31}) \oplus rk_{60} = Y_1^{31}$, $X_3^{31} \oplus F(X_2^{31}) \oplus rk_{61} = Y_3^{31}$, we recover rk_{60}, rk_{61} (K_2, K_5).

Step 3. Since we know wk_2, wk_3, rk_{60}, and rk_{61}, calculate intermediate values X_0^{30}, X_2^{30}, Y_1^{30}, and Y_3^{30} for given ciphertext. Since $Y_1^{29} = (X_0^{30})^L \mid (X_2^{30})^R$, $Y_3^{29} = (X_2^{30})^L \mid (X_0^{30})^R$, we calculate Y_1^{29}, Y_3^{29} for given ciphertext. In a similar way with Step 1 in analysis of Piccolo-80, recover X_0^{29}, X_2^{29} by using the cube in Table 3 for F which takes X_0^{29}, X_2^{29}. We recover X_1^{30}, X_3^{30} since $X_0^{29} = Y_0^{29}$, $X_2^{29} = Y_2^{29}$, and $Y_0^{29} = (X_3^{30})^L \mid (X_1^{30})^R$, $Y_2^{29} = (X_1^{30})^L \mid (X_3^{30})^R$. Since $X_1^{30} \oplus F(X_0^{30}) \oplus rk_{58} = Y_1^{30}$and $X_3^{30} \oplus F(X_2^{30}) \oplus rk_{59} = Y_3^{30}$, we recover rk_{58}, rk_{59} (K_6, K_3).

Step 4. This step is similar to Step 3 in analysis of Piccolo-80. Given ciphertext C, we calculate X_0^{29}, X_2^{29}, Y_1^{29}, and Y_3^{29}. We want to recover X_1^{29}. Since $X_0^{28} = Y_0^{28} = (X_3^{29})^L \mid (X_1^{29})^R$, $X_2^{28} = Y_2^{28} = (X_1^{29})^L \mid (X_3^{29})^R$, if we recover right 8 bits of X_0^{28} and left 8 bits of X_2^{28}, then we recover X_1^{29}. To recover right 8 bits of X_0^{28}, inject fault into bits corresponding to cube index of last 8 equations in Table 3. For recovering left 8 bits of X_2^{28}, inject fault into bits corresponding to cube index of first 8 equations in Table 3. Then X_1^{29} is recovered. And then we recover rk_{56} (K_0), because $X_1^{29} \oplus F(X_0^{29}) \oplus rk_{56} = Y_1^{29}$.

Step 5. This step is similar to Step 3 in analysis of Piccolo-80. We want to recover X_3^{28}. Given ciphertext C, we calculate X_0^{28}, X_2^{28}, Y_1^{28}, and Y_3^{28}. Since $X_0^{27} = Y_0^{27} = (X_3^{28})^L \mid (X_1^{28})^R$, $X_2^{27} = Y_2^{27} = (X_1^{28})^L \mid (X_3^{28})^R$, we recover left 8 bits of X_0^{27} and right 8 bits of X_2^{27}. To recover left 8 bits of X_0^{27}, inject fault into bits corresponding to cube index of first 8 equations in Table 3. For recovering right 8 bits of X_2^{27}, inject fault into bits corresponding to cube index of last 8 equations in Table 3. Then we recover X_3^{28}. And we recover rk_{55} (K_1) since $X_3^{28} \oplus F(X_2^{28}) \oplus rk_{55} = Y_3^{28}$.

Use the above 5 steps to recover all the master keys used for Piccolo-128. This needs the assumption that we inject fault into at most 4 bits in the same time. Assume that we analyze only Steps 1, 2, 3, and 4 such as Piccolo-80. Even though we analyze only Steps 1, 2, 3, and 4, since at least 32 bits of 128-bit master key are recovered, we recover all the master keys with less operations than brute-force attack. Table 5 is showing encryption complexity needed for recovering the master key of each assumption.

4.4. Improved Attack. The attacks described in Sections 4.2 and 4.3 are valid under the assumption that an adversary is able to control the injecting time and bit positions to inject fault and the number of bits of fault injection. However it is difficult to control bit positions where faults are injected. Therefore, in this section, we explain the way of attack under the assumption that an adversary is not able to control bit positions to inject faults. That is, an adversary is able to inject i bits of random faults into Piccolo-80 and Piccolo-128. The attack of Section 4.4 is similar to the attack of Sections 4.2 and 4.3. But this attack adds the process that determines fault position before each Step of Sections 4.2 and 4.3.

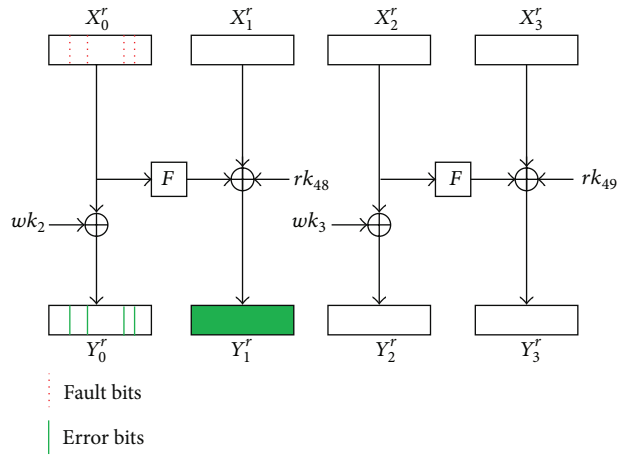


Figure 4: Determine position of fault injection for Case 1 ($i = 4$).

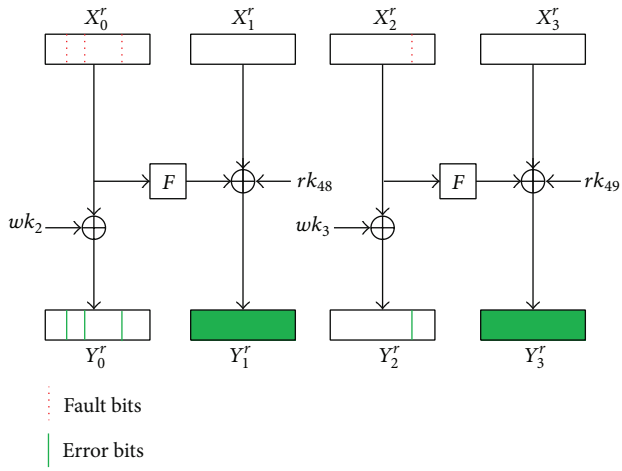


Figure 5: Determine position of fault injection for Case 2 ($i = 4$, $j = 1$).

Assume that we inject i bits of random fault into last round. In some cases, we determine fault position. Then, the following is how to determine position of fault injection for 3 cases. (Each case with 4 fault bits is described in Figures 4, 5, and 6).

Case 1. There are i bits error in $Y_0^r(Y_2^r)$ after fault injection. Since $X_0^r \bigoplus wk_2 = Y_0^r$ ($X_2^r \bigoplus wk_3 = Y_2^r$) and i fault bits, if Y_0^r (Y_2^r) is different i bits from original ciphertext, then fault injection position in $X_0^r(X_2^r)$ must be same with changed i bits after fault injection in Y_0^r (Y_2^r).

Case 2. There are j bits error in Y_0^r and $i - j$ bits error in Y_2^r after fault injection. In a similar way to Case 1, fault injection position in X_0^r must be same with changed j bits after fault injection in Y_0^r. And fault injection position in X_2^r must be same with changed i-j bits after fault injection in Y_2^r.

Case 3. There are $i - j$ bits error in Y_0^r (Y_2^r) and j bits error in $Y_3^r(Y_1^r)$. And there is no error in $Y_2^r(Y_0^r)$. If there are j bits error

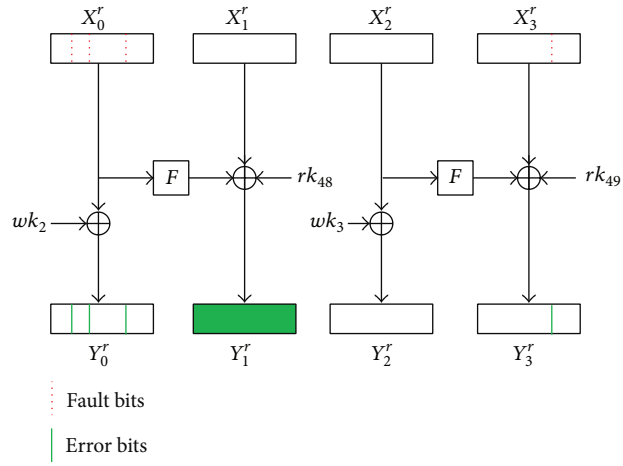


Figure 6: Determine position of fault injection for Case 3 ($i = 4$, $j = 1$).

Table 5: Attack complexity of Piccolo-128.

Assumption	Required fault	complexity
Step 1	$132 \approx 2^{7.04}$	2^{96}
Step 2	$264 \approx 2^{8.04}$	2^{64}
Step 3	$396 \approx 2^{8.63}$	2^{32}
Step 4	$479 \approx 2^{8.90}$	$2^{16.01}$
Step 5	$562 \approx 2^{9.13}$	$593.64 \approx 2^{9.21}$

in $Y_3^r(Y_1^r)$, then $X_2^r(X_0^r)$ or $X_3^r(X_1^r)$ is fault injected. Since there is no error in $Y_2^r(Y_0^r)$, fault injection position in $X_3^r(X_1^r)$ must be same with changed j bits after fault injection in $Y_3^r(Y_1^r)$. Furthermore, $X_0^r(X_2^r)$must be same with changed $i - j$ bits after fault injection in Y_0^r (Y_2^r) since there are $i - j$ bits error in Y_0^r (Y_2^r).

Therefore, we determine the position of fault injection for these 3 cases. Table 3 is the table of optimal cubes that are used for the attack in Sections 4.2 and 4.3. Therefore, there are many cubes except cubes in Table 3. Because there are too many cubes, we do not describe all cubes in this paper. For example, suppose that we get 16 ciphertexts for cube sum of {0, 1, 3, 4}. Then we calculate all of subcubes of {0, 1, 3, 4}. So, we use cubes in Table 6 and recover 1st, 2nd, 5th, and 7th bit of whitening key.

By the same method, we use sufficient cubes for recovering wk_2 and wk_3 of Piccolo-80 and Piccolo-128 and recover wk_2 and wk_3. Since we know wk_2 and wk_3, calculate intermediate values X_0^r, X_2^r for given ciphertext. Therefore, we inject faults into $(r - 1)$th round, determining position of fault injection by 3 cases that described this section. And we recover $(r - 1)$th round key. This process is the same with Step 2 in Section 4.2 and Step 2 in Section 4.3 except determining position of fault injection. By the same way, we recover all the master keys of Piccolo-80 and Piccolo-128 using Steps in Sections 4.2 and 4.3 with determining position of fault injection, respectively.

Table 6: Subcube of {0, 1, 3, 4}.

Cube index	Outbit (F_i)	Polyequation
0, 3	0	$x_2 + 1$
0, 4	8	$x_1 + x_5 + 1$
0, 1, 3, 4	2	x_5
0, 1, 3, 4	15	x_7

Table 7: Necessary positions of fault injections for Table 3.

Number of fault bits (Number of faults)	Necessary positions of fault injection
1 (11)	(0), (1), (4), (5), (6), (7), (8), (9), (11), (12), (13)
2 (27)	(0, 1), (0, 5), (0, 6), (0, 8), (0, 9), (0, 11), (0, 12), (1, 5), (1, 6), (4, 5), (4, 7), (4, 8), (4, 9), (4, 11), (5, 6), (5, 7), (5, 8), (5, 9), (5, 13), (6, 9), (6, 13), (7, 8), (8, 9), (8, 11), (8, 12), (9, 11), (11, 12)
3 (22)	(0, 1, 5), (0, 1, 6), (0, 5, 6), (0, 8, 9), (0, 8, 11), (0, 8, 12), (0, 9, 11), (0, 11, 12), (1, 5, 6), (4, 5, 7), (4, 5, 8), (4, 5, 9), (4, 7, 8), (4, 8, 9), (4, 8, 11), (4, 9, 11), (5, 6, 9), (5, 6, 13), (5, 7, 8), (5, 8, 9), (8, 9, 11), (8, 11, 12)
4 (6)	(0, 1, 5, 6), (0, 8, 9, 11), (0, 8, 11, 12), (4, 5, 7, 8), (4, 5, 8, 9), (4, 8, 9, 11)

The complexity of this attack depends on the complexity for finding ciphertext to recover round key. For Table 3, we need 66 ciphertexts. Table 7 is showing necessary positions of fault injection for cubes of Table 3.

For calculating complexity, we assume that an adversary always injects 4-bit random fault into the last three and five rounds for Piccolo-80 and Piccolo-128, respectively. Since an adversary injects exactly 4-bit fault, position of all fault bits has to match position that we want. This probability is $1/\binom{64}{4} \approx 2^{-19.277}$. That is, an adversary injects $2^{-19.277}$ times for each round and gets all ciphertexts that correspond to Table 3 for each round. Therefore, this attack on Piccolo-80 needs $3 \cdot 2^{19.277} \approx 2^{20.862}$ fault injections. Similarly, this attack on Piccolo-128 needs $5 \cdot 2^{19.277} \approx 2^{21.599}$ fault injections. The number of additional encryptions to recover the master key is negligible. Hence, our attack has the complexity of $2^{20.86}$ and $2^{21.60}$ encryptions with four bits of random fault injections for Piccolo-80 and Piccolo-128, respectively.

5. Conclusions

In this paper, we present the security weakness of Piccolo against fault analysis. Our attack fully exploits the structure of Piccolo, which is a Feistel network. We describe an attack for fault injection of target bit positions on Piccolo-80 and Piccolo-128. The master key of Piccolo-80 and Piccolo-128 is recovered by fault analysis by using cube attack with injecting faults $2^{8.44}$ and $2^{9.14}$, respectively. Our attack has the complexity of $2^{8.49}$ and $2^{9.21}$ encryptions for Piccolo-80 and Piccolo-128, respectively, which are practical complexities. And finally, an attack for *random* four bits fault injection for Piccolo-80 and Piccolo-128 is presented. This attack needs $2^{20.86}$ and $2^{21.60}$ encryptions with four bits of random fault injections for Piccolo-80 and Piccolo-128, respectively.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] D. Boneh, R. A. DeMillo, and R. J. Lipton, "On the importance of checking cryptographic protocols for faults," in *Advances in Cryptology—EUROCRYPT'97*, vol. 1233 of *Lecture Notes in Computer Science*, pp. 37–51, Springer, 1997.

[2] E. Biham and A. Shamir, "Differential fault analysis of secret key cryptosystems," in *Advances in Cryptology—CRYPTO'97*, vol. 1294 of *Lecture Notes in Computer Science*, pp. 513–525, Springer, 1997.

[3] C. H. Kim and J.-J. Quisquater, "New differential fault analysis on AES key schedule: two faults are enough," in *Smart Card Research and Advanced Applications*, vol. 5189 of *Lecture Notes in Computer Science*, pp. 48–60, Springer, 2008.

[4] M. Tunstall, D. Mukhopadhyay, and S. Ali, "Differential fault analysis of the advanced encryption standard using a single fault," in *Information Security Theory and Practice. Security and Privacy of Mobile Devices in Wireless Communication*, vol. 6633 of *Lecture Notes in Computer Science*, pp. 224–233, Springer, 2011.

[5] W. Li, D. Gu, and J. Li, "Differential fault analysis on the ARIA algorithm," *Information Sciences*, vol. 178, no. 19, pp. 3727–3737, 2008.

[6] K. Jeong, Y. Lee, J. Sung, and S. Hong, "Differential fault analysis on block cipher SEED," *Mathematical and Computer Modelling*, vol. 55, no. 1-2, pp. 26–34, 2012.

[7] H. Chen, W. Wu, and D. Feng, "Differential fault analysis on CLEFIA," in *Information and Communications Security*, vol. 4861 of *Lecture Notes in Computer Science*, pp. 284–295, Springer, 2007.

[8] K. Jeong and C. Lee, "Differential fault analysis on block cipher LED-64," in *Future Information Technology, Application, and Service*, vol. 1, pp. 747–755, Springer, 2012.

[9] K. Jeong, "Security analysis of block cipher Piccolo suitable for wireless sensor networks," *Peer-to-Peer Networking and Applications*, 2013.

[10] S. Li, D. Gu, Z. Ma, and Z. Liu, "Fault analysis of the Piccolo block cipher," in *Proceedings of the 8th International Conference on Computational Intelligence and Security (CIS '12)*, pp. 482–486, 2012.

[11] H. Yoshikawa, M. Kaminaga, A. Shikoda, and T. Suzuki, "Round addition DFA on 80-bit Piccolo and TWINE," *IEICE Transactions on Information and Systems*, vol. E96-D, no. 9, pp. 2031–2035, 2013.

[12] F. Zhang, X. Zhao, S. Guo, T. Wang, and Z. Shi, "Improved algebraic fault analysis: a case study on Piccolo and applications to other lightweight block ciphers," in *Proceedings of the 4th International Workshop on Constructive Side-Channel Analysis and Secure Design (COSADE '13)*, vol. 7864 of *Lecture Notes in Computer Science*, pp. 62–79, Springer, 2013.

[13] N. Bagheri, R. Ebrahimpour, and N. Ghaedi, "New differential fault analysis on PRESENT," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, article 145, 10 pages, 2013.

[14] W. Y. Zhang, F. Liu, X. Liu, and S. Meng, "Differential fault analysis and meet-in-the-middle attack on the block cipher KATAN32," *Journal of Shanghai Jiaotong University (Science)*, vol. 18, no. 2, pp. 147–152, 2013.

[15] I. Dinur and A. Shamir, "Cube attacks on tweakable black boxp polynomials," in *Advances in Cryptology—EUROCRYPT 2009*, vol. 5479 of *Lecture Notes in Computer Science*, pp. 278–299, Springer, 2009.

[16] S. F. Abdul-Latip, M. R. Reyhanitabar, W. Susilo, and J. Seberry, "Fault analysis of the KATAN family of block ciphers," in *Information Security Practice and Experience*, vol. 7232 of *Lecture Notes in Computer Science*, pp. 319–336, Springer, 2012.

[17] I. Dinur and A. Shamir, "Applying cube attacks to stream ciphers in realistic scenarios," *Cryptography and Communications*, vol. 4, no. 3-4, pp. 217–232, 2012.

[18] K. Shibutani, T. Isobe, H. Hiwatari, A. Mitsuda, T. Akishita, and T. Shirai, "Piccolo: an ultra-lightweight blockcipher," in *Proceedings of the 13th International Workshop on Cryptographic Hardware and Embedded Systems (CHES '11)*, vol. 6917 of *Lecture Notes in Computer Science*, pp. 342–357, 2011.

[19] Y. Wang, W. Wu, and X. Yu, "Biclique cryptanalysis of reduced-round Piccolo block cipher," in *Information Security Practice and Experience*, vol. 7232 of *Lecture Notes in Computer Science*, pp. 337–352, Springer, 2012.

[20] J. Song, K. Lee, and H. Lee, "Biclique cryptanalysis on lightweight block cipher: HIGHT and Piccolo," *International Journal of Computer Mathematics*, vol. 90, no. 12, pp. 1–17, 2013.

[21] J.-P. Aumasson, I. Dinur, W. Meier, and A. Shamir, "Cube testers and key recovery attacks on reduced-round MD6 and trivium," in *Fast Software Encryption*, vol. 5665 of *Lecture Notes in Computer Science*, pp. 1–22, Springer, 2009.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 237125, 7 pages
http://dx.doi.org/10.1155/2014/237125

Research Article

Spyware Resistant Smartphone User Authentication Scheme

Taejin Kim,[1] Jeong Hyun Yi,[1] and Changho Seo[2]

[1] *School of Computer Science and Engineering, Soongsil University, Seoul 156-743, Republic of Korea*
[2] *Department of Applied Mathematics, Kongju National University, Kongju 314-701, Republic of Korea*

Correspondence should be addressed to Jeong Hyun Yi; jhyi@ssu.ac.kr

Received 7 November 2013; Accepted 3 February 2014; Published 9 March 2014

Academic Editor: Jongsung Kim

As smart phones are becoming widely used, a variety of services to store and use important information such as photos and financial information are now provided. User authentication to protect this information is increasingly important. The commonly used 4-digit PIN, however, is vulnerable to the Brute Force Attack, Shoulder-Surfing Attack, and Recording Attack. Various authentication techniques are being developed in order to solve these problems. However, the technique that provides perfect protection, even from the Recording Attack, is not yet known, and in most cases, a password can be exposed by multiple Recording Attacks. This paper proposes a new user authentication method that protects against a Recording Attack from spyware on the user's smart phone. The proposed method prevents password exposure by multiple Recording Attacks, is implemented on a real Android phone, and has been evaluated for usability.

1. Introduction

A smart phone is different from feature phones in that it has a mobile OS that makes it possible to freely install and remove applications just as for personal computers. Because of this, a variety of services can be provided in addition to basic functions such as calling and messaging. However, malware also occurs in smart phones, just as for personal computers. This malware could expose information in the smart phone, and consequently invade privacy or cause financial damage.

Among these types of malware, spyware could exist which leaks the authentication screen and touch coordinates of the user. When the authentication screen and touch coordinates are exposed, it is effectively a Recording Attack [1]. The Recording Attack is a type of Shoulder-Surfing Attack [2] where the attacker records the entire user authentication process including ID and password input for a service. If a spyware records and sends this process to an attacker's server, the password of the user can be easily taken. Although many authentication methods have been developed in order to prevent Shoulder-Surfing Attacks, these are only for those attacks where an attacker simply views the authentication process over a shoulder and remembers the password. In order to resist such attacks, these methods only momentarily expose the screen information that is provided during the authentication process or increase the amount of data that has to be remembered by the attacker. However, these methods are limited because they cannot perfectly protect against a Recording Attack, which records the entire authentication screen.

Therefore, this paper proposes an authentication method devised such that the password cannot be easily taken even when the entire authentication screen is recorded and exposed. This method generates and authenticates a one-time password that is changed each time according to prior knowledge of the user without the need for separate hardware and includes wrong information in the inputted password, making it possible to prevent the exposure of the correct password to the attacker. The proposed method has been developed for the smart phone environment and ensures safety from the Shoulder-Surfing Attack, Brute Force Attack [3], Smudge Attack [4], and Recording Attack that threaten user authentication.

2. Related Work

Since a variety of attack methods such as the Brute Force Attack, Smudge Attack, and Shoulder-Surfing Attack have

been known to threaten user authentication, many studies [5–10] have attempted to solve these problems. Among them, we briefly introduce some techniques that can be applied to mobile devices.

2.1. DAS (Dynamic Authentication System). DAS [11] is a scheme for bank ATMs to prevent the Shoulder-Surfing Attack. This scheme uses a common PIN as the password, and during user authentication shows the digits of a number keypad that are randomly positioned only when the "view" button is pressed in order to reduce the exposure time to the attacker. After the user checks the position of digits by pressing the "view" button, the digits disappear when the finger is lifted and the user inputs the remembered positions of digits. After the input, the digits are repositioned randomly. This method is safe from the temporary Shoulder-Surfing Attack due to the short exposure time, but unsafe from multiple Shoulder-Surfing Attacks or a Recording Attack.

2.2. Passfaces. Passfaces [12] uses the image of a human face as the password instead of numbers or characters. It has been developed based on the theory that it is easier to use images than numbers or characters for memory and that human faces can be easily remembered. The user selects a face image to be used as the password among those provided by the system. Then the user familiarizes him/herself with the images by looking at each face image for a few seconds. During user authentication, nine images are shown, of which eight are dummies and one is the user-set password image. The user selects the password image that has been set by him/her and repeats the process for the length of the password. If all the images of the password are identical to the selected images, the authentication will be successful. This scheme prevents the Shoulder-Surfing Attack as human faces look similar from a distance. In addition, dummy images are shown which are similar in gender or appearance to the password face images in order to confuse the attacker. However, this method has a weakness in that the password images can be clearly seen at close range or by using devices such as a telescope and is vulnerable to the Recording Attack.

2.3. M.TransKey. M.TransKey [13] is a virtual keypad scheme for protecting the touch log in the smart phone environment. Random spaces are added to the QWERTY keyboard and keys are differently positioned for each use. As the user inputs different positions for each key input, the attacker cannot know the inputted key even when the coordinates are exposed through the touch log. It can also be applied to existing systems as it is able to input alphanumeric values. If the Shoulder-Surfing Attack is used to view the authentication screen, however, the password can be easily exposed, and the method is also unsafe from spyware that leaks the screen as well as the touch log.

2.4. Dementor-SGP. Dementor-SGP [14] is an authentication scheme using the relative path between images. The password consists of one hole image and three user images. The user images are inserted into the hole image for user authentication. Images are randomly positioned and the relative path is changed for each authentication. The security levels are organized according to security and usability, and the user can select the level. At the lowest level, the password image is easily exposed as user images are dragged and inserted into the hole image. From the second level upwards, however, user images are moved to the hole image by arrow keys instead of a mouse and inputted with the input button. The hole image and user images are not exposed as they are not directly selected, and it is safe even from the Shoulder-Surfing Attack as images are randomly positioned and consequently arrow keys are inputted differently for each authentication. In the case of the Recording Attack, it is possible to check which user image is inserted into each image, but it is not possible to know which one is the hole image among them. Nonetheless, multiple Recording Attacks can determine the image into which the same user image is inserted and steal the password.

3. Proposed Scheme

The existing password schemes were devised to protect from the Cognitive Shoulder-Surfing Attack, which depends upon human memory. They are, however, vulnerable to the Recording Attack, which records the entire authentication process by using a camera or other recording devices. In particular, the password is easily exposed to spyware, which leaks the touch information and screen. This paper proposes an authentication method to solve these problems.

The password of the proposed scheme consists of a sequential selection of grid cells and a number of errors to include. During user authentication, up, down, left, and right arrows are shown on each cell and the message "Start" is displayed on a random cell. The user moves from the "Start"-displayed position according to the arrow directions that are shown on the cells selected as the password by him/her. When moving according to arrow directions, the user intentionally inputs the indicated number of wrong authentication values by moving in a direction different from the arrow shown on the password cell. The number of included errors must be equal to the number set when setting up the password. The left image of Figure 1 indicates how to set the password.

Using (x, y) coordinates for the grid, the password cells are, sequentially, (1, 1), (2, 1), (3, 1), (4, 1), and (5, 1). If pressing the "OK" button after password cells have been selected, a window pops up which requests the number of errors to complete setting the password. In the example in Figure 1, the number of errors is set to one. The center of Figure 1 shows an example of a typical user authentication. The user should move, sequentially, in the "left," "up," "left," "left," and "down" directions as the password is (1, 1), (2, 1), (3, 1), (4, 1), and (5, 1). If the user moves from the starting position, (5, 7) according to the arrows shown on the password cells, the coordinates to be inputted are (5, 7), (4, 7), (4, 6), (3, 6), (2, 6), and (2, 7). As the number of errors is one, however, the last password moves to (2, 5) instead of (2, 7), to include an error. Once the selection has been finished, the user authentication will be complete.

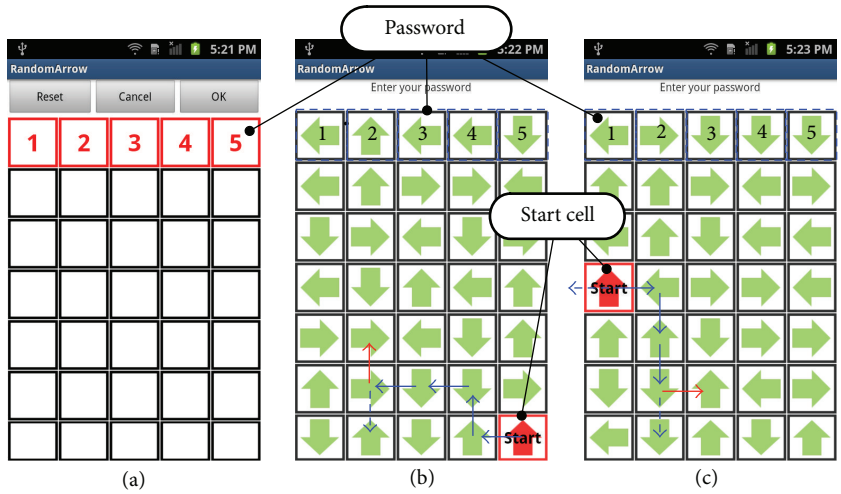


Figure 1: Password setting (a), common user authentication (b), and an example where an authentication value is omitted during user authentication (c).

Table 1: Parameter description.

Parameter	Description
X	Number of horizontal cells in the grid
Y	Number of vertical cells in the grid
N	Password length
S	Number of exposed authentication processes
E	Number of included errors when passwords are inputted
T	Number of correct authentication values out of all exposed authentication values
I	Number of authentication values for intersection

In the proposed scheme, arrow directions could go beyond the scope of the grid, as shown on the right of Figure 1, as the starting position is selected randomly. As shown, the directions to move to are "left," "right," "down," "down," and "down," the starting position is (1, 4), and the password is (1, 1), (2, 1), (3, 1), (4, 1), and (5, 1). If the user moves according to the password, the coordinates should be (1, 4), (0, 4), (1, 4), (1, 5), (1, 6), and (1, 7). However, a value less than 0 on the x-axis of the grid cannot be inputted. In these cases, the arrow that goes out of the grid is omitted and the next password direction is inputted. In the above example, as the first "left" arrow cannot be inputted, it is omitted, which makes the coordinates (1, 4), (2, 4), (2, 5), (2, 6), and (2, 7). If the user moves "right" in order to make an error for the last authentication value, the actual moved coordinates become (1, 4), (2, 4), (2, 5), (2, 6), and (3, 6).

4. Security Analysis

The safety of the proposed scheme is analyzed in this section. The parameters used for safety analysis are as shown in Table 1.

Table 2: Number of possible password combinations according to the password length.

Password length	1 digit	2 digits	3 digits	4 digits	5 digits
PIN	10	100	1,000	10,000	100,000
Proposed scheme	35	1,190	39,270	1,256,640	38,955,840

4.1. Brute Force Attacks. The Brute Force Attack is a method that finds the password by inputting possible password combinations one by one. The number of possible password combinations of the proposed scheme is influenced by the grid size and the password length. This can be expressed as ${}_{XY}P_N$. As the grid size of the implemented prototype is 5×7, the results are as shown in Table 2 for comparison with the commonly used PIN.

The number of possible password combinations is about 125 times more than the commonly used 4-digit PIN, which significantly increases safety.

4.2. Smudge Attacks. The Smudge Attack is a method that finds the password by using smudges left on the input interface by a user. As for the 4-digit PIN password, it is possible to conjecture the password by using the fingerprints left on the keypad to determine which buttons have been pressed. In the commonly used Android pattern lock, it is difficult to find out the password by Brute Force Attacks as more than 380,000 password combinations are possible. Smudge Attacks, however, can easily determine the path of the user. Besides the naked eye, a microscope or other fingerprinting tool can be also used to easily obtain the input smudges. For the proposed scheme, the information that could be obtained from the Smudge Attack is the starting position and movement directions. However, the starting position is changed randomly each time and as for the movement directions, the arrow shapes created on the password cells are changed for each authentication. Therefore, the Smudge Attack cannot find the password from the proposed scheme.

4.3. Spyware-Based Recording Attacks. The spyware-based Recording Attack is a much more threatening attack than existing Shoulder-Surfing Attacks and general Recording Attacks. Therefore, if safety can be confirmed from a spyware-based Recording Attack, it is also safe from the Shoulder-Surfing Attack or other types of Recording Attack. The spyware-based Recording Attack mentioned here is assumed to record and leak to the attacker all the authentication process.

The proposed scheme has three elements for protection against spyware-based Recording Attacks. The first element is arrows in the same direction, the second is the omission of authentication values, and the last is included errors.

4.3.1. Arrows in the Same Direction. Even if the attacker has recorded the entire authentication process to steal a password, it is not possible to determine password cells. The reason is that the attacker cannot know which cell is the password as a quarter of the total cells have the same arrow as the movement direction used for authentication. The attacker, however, could get a candidate password cell set, consisting of a quarter of the total cells, which indicate the same directions as the recorded directions moved for authentication. The attacker could get another candidate password cell set, also amounting to a quarter of the total cells using the same method in another authentication process, and conjecture the password cells from an intersection of the two sets. The following expression shows the average number of candidate password cell sets that could be conjectured from s recorded authentication processes:

$$f(s) = XY\left(\frac{1}{4}\right)^{s}. \quad (1)$$

If $f(s)$ is less than or equal to 1, the password cell is exposed. So, for a grid size of 5×7, $f(s)$ becomes less than 1 when the value of s is 3. That is, the password is exposed when making an intersection of three candidate password cell sets.

4.3.2. Omission of Authentication Values. Some user authentication cases could occur where input is not possible as the starting point is located at an edge and the arrow goes beyond the grid. The arrow that cannot be inputted is omitted and the next authentication value is inputted in order to solve this problem. As a part of the inputted directions are omitted, the potentially exposed authentication values are decreased and safety becomes significantly enhanced. The probability that input values are not to be omitted in user authentication can be obtained by dividing the number of paths for which the authentication values are not omitted by the number of total paths which could be inputted for a password. The number of total paths is the number of movable directions for all (x, y) coordinates raised to the number of movements N. It is simply expressed by $4^N XY$. The following expression shows the probability that the authentication values are not omitted:

$$\frac{R(N, X, Y)}{4^N XY}. \quad (2)$$

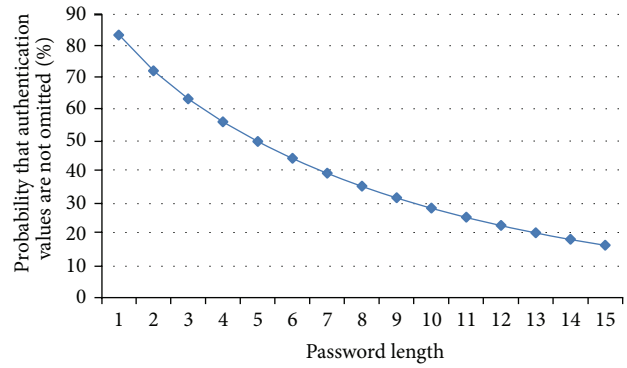


Figure 2: Probability that authentication values are not omitted according to the password length.

The function R in Expression (2) is an algorithm that enumerates the number of paths in which authentication values are not omitted. Algorithm 1 shows the algorithm for R.

The path can be moved by as much as N and the scope cannot go beyond (X, Y) in the function R. The recursive function r used in R finds a path where it is possible to move to up, down, left and right from the current position (x, y), moves to that position, and finds the next allowable path. The recursive function repeats itself until N becomes 0, when it returns to 1 or 0 by checking if the current position is valid. It also returns to 0 when N is not 0 and the (x, y) value goes beyond the scope of the grid. In short, the recursive function r obtains the number of paths in which authentication values are not omitted and it is possible to move N lengths from the start coordinate (x, y). The function R executes the recursive function r at all possible (x, y) positions in the grid and sums the results. That is, it obtains the number of paths in which authentication values are not omitted for all possible (x, y). It is then possible to obtain the probability that authentication values are not omitted and the authentication process is exposed by using Expression (2) for a grid size of 5×7 (as in the implemented prototype). Figure 2 shows the probability that authentication values are not omitted versus password length.

Therefore, multiple Recording Attacks are required to obtain an authentication process where authentication values are not omitted. Using Expression (1), the number of authentication processes required to determine the password is three for a grid size of 5×7. The probability that authentication values are not omitted and exposed is about 48.95% when the password length is five. That is, it is possible to obtain three authentication processes where authentication values are not omitted and determine the password, only after more than about seven Recording Attacks.

4.3.3. Errors Included in Authentication Values. It has been shown that it is possible to be safe from about seven Recording Attacks due to the inclusion of arrows in the same direction and the omission of authentication values. However, the authentication process can be continually leaked

```
Function R(N, X, Y)
    VAR tot
    tot = 0
    FOR y = 0 to Y
        FOR x = 1 to X
            tot = tot + r(x, y, N, X, Y)
        ENDFOR
    ENDFOR
    RETURN tot
Function r(x, y, N, X, Y)
    IF N == 0 THEN
        IF x <= X And 0 < x And y <= Y And 0 < y
            RETURN 1
        ELSE
            RETURN 0
        ENDIF
    ELSE
        VAR right, left, top, bottom
        right = 0
        left = 0
        top = 0
        bottom = 0
        IF x <= X And 0 < x And y <= Y And 0 < y THEN
            right = r(x + 1, y, N − 1, X, Y)
            left = r(x − 1, y, N − 1, X, Y)
            top = r(x, y + 1, N − 1, X, Y)
            bottom = r(x, y − 1, N − 1, X, Y)
        ENDIF
    ENDIF
    RETURN right + left + top + bottom
```

Algorithm 1: An algorithm to obtain the number of paths for which authentication values are not omitted.

and the password eventually exposed if the Recording Attack is based upon spyware. Errors are intentionally included in authentication values to prevent the correct password cells from being exposed. The attacker has to make an intersection of many candidate password cell sets in order to determine the password. The attacker gets the wrong password if there are incorrect authentication values among them. Therefore, the attacker has to check if the password is correct. When errors are not included, it is possible to determine the password cell by using the same authentication processes and making an intersection of them for each password digit. However, the attacker has to use different authentication processes for making the intersection for each password digit, as one of the password digits includes an error. In addition, it is necessary to check if the password is correct, independently, for each password digit. The probability that one password digit has an error is e/N. Using this, the number of correct authentication values for each password digit among the total exposed authentication processes becomes the following:

$$T = \frac{s\,(N - e)}{N}. \tag{3}$$

The attacker has to select candidate password cells to be used for intersections from the T set in order to find out the password. Assuming that the number of candidate password cells used for intersections is i, the number of combinations is ${}_TC_i$. It is possible to obtain the probability that the attacker could get a correct digit of password cells by dividing this value by the total number of combinations. It is also possible to obtain the probability that the attacker gets the password by raising it to the power of the password length, as password digits are independent of one another. The following expression shows the probability that the attacker gets the correct password among the total intersections:

$$\left(\frac{{}_TC_i}{{}_sC_i}\right)^N. \tag{4}$$

The probability that the attacker gets the password has been obtained according to the number of exposed authentication processes for $N = 5$, $e = 1$, and $i = 3$, using Expression (4). Figure 3 shows the probability of obtaining the password according to the number of exposed authentication processes.

From Figure 3, we can see that the probability of obtaining the password converges to a certain value as the number of exposed authentication processes increases. It is possible to get the average password-exposure probability as the number of exposed authentication processes increases to infinity. The

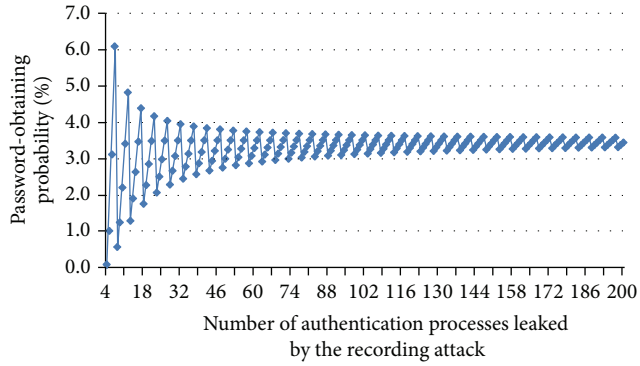


Figure 3: Probability that the password will be obtained according to the number of leaked authentication processes.

following expression shows the average password-exposure probability:

$$\left(\lim_{s\to\infty} \frac{{}_{T}C_i}{{}_{s}C_i}\right)^{N}. \quad (5)$$

By Expression (5), the password is exposed with the probability of about 3.51% when $N = 5$, $e = 1$, and $i = 3$. Table 3 compares the probability of success of a spyware-based Recording Attack for existing schemes and the proposed scheme.

The password is 100% exposed to a spyware-based Recording Attack for all methods other than the proposed scheme. DAS randomly positions the keypad and reduces the amount of exposure time. It is possible, however, to find out the password with one Recording Attack as it is possible to recheck the screen where the numbers have been exposed. The password is easily exposed to a Recording Attack as the password is directly touched and selected in Passfaces and M.TransKey. Dementor-SGP protects the password for one Recording Attack as input is executed with arrow keys instead of direct selection of the hole and user images. However, the single Recording Attack can see which user image is inserted into each image displayed on the screen, and two Recording Attacks make it possible to find out the password if the attacker can obtain the image into which the same user image is inserted. The proposed scheme is safe even from multiple Recording Attacks as errors are included in the values used for authentication, and consequently, the correct password is not exposed.

5. Experimental Results

In this section, the existing and proposed schemes were implemented and their usability was experimentally quantified. The development tools used for the implementations were Eclipse Indigo, Android SDK 2.3, and JAVA 1.7.0. The test equipment consisted of a smart phone SHW-M250S (1.2 GHz Dual-Core CPU, 1 GB RAM) running Android 2.3.

5.1. GOMS Model Test. The GOMS model [15] was used for the objective usability test. The GOMS model is one of the human information processor models used for observing the interaction between humans and computers and divides human behavior into four rules: Goals, Operators, Methods, and Selection, and analyzes them. Evaluation tools such as CogTool [16], GOMSED [17], and QGoms [18] are based upon this model, and this paper used CogTool 1.2.2 for evaluation. Table 4 shows the comparison of estimated authentication times of the existing and proposed schemes as measured by CogTool.

Table 3: Success probability of the spyware-based Recording Attack.

Authentication method	Success probability for a spyware-based Recording Attack
DAS	100%
Passfaces	100%
M.TransKey	100%
Dementor-SGP	100%
Proposed scheme	$\left(\lim_{s\to\infty} \frac{{}_{T}C_i}{{}_{s}C_i}\right)^{N}$

Table 4: CogTool measurement results.

Authentication method	Authentication time
DAS	11.27
Passfaces	8.64
M.TransKey	12.19
Dementor-SGP	11.91
Proposed scheme	10.72

The test results show that the proposed scheme has a slightly faster authentication time than existing methods. The fastest scheme is Passfaces, as it is only necessary to select five user-set images to authenticate the password. M.TransKey showed a slow result as it has smaller buttons than Passfaces and the password length is at least six digits. DAS was slower than the proposed scheme even though it uses a 4-digit PIN as the password, because the user also has to press the "view" button to check numbers. Dementor-SGP was also slow as many inputs are necessary to move user images to the hole image using the up, down, left, and right buttons. The proposed scheme showed a comparatively fast result as the password is inputted with one drag from the starting position.

5.2. User Test. Authentication time and error rate were measured for ten test participants in order to check if the existing and proposed schemes are convenient to use. Test participants were given detailed explanations and exercise time to become familiar with each authentication method. The authentication time and error rate were measured after executing each authentication method five times. Table 5 shows the authentication time and error rate for each scheme.

A test log was saved to collect the information about password authentication time and error rate. Test results show that DAS and M.TransKey have fast authentication times. The reason is that they use typical 4-digit PIN and alphanumeric symbols that facilitate easy use. The other methods use graphical passwords, and among these, the proposed scheme

Table 5: User test results.

Authentication method	Authentication time	Error rate
DAS	9.87	12.0
Passfaces	14.55	14.0
M.TransKey	9.97	18.0
Dementor-SGP	16.60	22.0
Proposed scheme	11.47	18.0

showed the fastest result. Unlike the other methods, the proposed scheme has a fast authentication as it uses one instance of drag for authentication. Passfaces took a long time for the user to remember the password, and Dementor-SGP also needed a long time to move user images to the hole image with just the arrow keys.

The error rate of the proposed scheme is 18.0%, which is not better than the existing methods, but it is at the same allowable level as in the commonly used M.TransKey. There were many cases where the user did not input errors as they should have and the authentication failed. As for these cases, the error rate can be reduced by classifying security levels and making authentication successful even when errors have not been inputted. As this increases usability but decreases security, it is possible to divide security levels so that the user selects the desired level. Further, the proposed scheme can enhance memorability by adding a background image to the grid or images to each cell because a user can remember images longer than simple cell positions.

6. Conclusion

This paper proposed a new authentication method that is safe from spyware-based Recording Attacks. The proposed scheme uses the grid cells and the number of errors as the password, moves according to arrow directions shown on the password cells, and also moves to incorrect positions according to a predefined number of errors for user authentication. The proposed scheme generates and authenticates the one-time password without the need to separate hardware and is safe from the Brute Force Attack, Smudge Attack, Shoulder-Surfing Attack, and Recording Attack. Usability test results also show that its error rate is an allowable level, though it is a little higher than existing methods, and that its authentication speed is slightly faster. Therefore, the proposed scheme is a practical new password authentication method that has similar usability to existing methods and also can protect from the spyware-based Recording Attacks that threaten the mobile environment.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported in part by the IT R&D program of MOTIE/KEIT (10039180, Intuitive, convenient, and secure HCI-based usable security technologies for mobile authentication and security enhancement in mobile computing environments) and in part by the National Research Foundation of Korea (NRF) Grant funded by the Ministry of Education (2013R1A1A2010382).

References

[1] R. Biddle, S. Chiasson, and P. C. van Oorschot, "Graphical passwords: learning from the first twelve years," *ACM Computing Surveys*, vol. 44, no. 4, pp. 1–41, 2012.

[2] A. Stevenson, *Shorter Oxford English Dictionary*, Oxford University Press, 2007.

[3] C. Paar, J. Pelzl, and B. Preneel, *Understanding Cryptography: A Textbook for Students and Practitioners*, Springer, 2010.

[4] A. Mylonas, A. Kastania, and D. Gritzalis, "Delegate the smartphone user? Security awareness in smartphone platforms," *Computers & Security*, vol. 34, pp. 47–66, 2013.

[5] S. Wiedenbecka, J. Watersa, J. C. Birgetb, A. Brodskiyc, and N. Memonc, "PassPoints: design and longitudinal evaluation of a graphical password system," *International Journal of Human-Computer Studies*, vol. 63, no. 1-2, pp. 102–127, 2005.

[6] R. Biddle, M. Mannan, P. C. van Oorschot, and T. Whalen, "User study, analysis, and usable security of passwords based on digital objects," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 970–979, 2011.

[7] J.-C. Birget, D. Hong, and N. Memon, "Graphical passwords based on robust discretization," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 3, pp. 395–399, 2006.

[8] E. Hayashi, R. Dhamija, N. Christin, and A. Perrig, "Use your illusion: secure authentication usable anywhere," in *Proceedings of the 4th Symposium on Usable Privacy and Security (SOUPS '08)*, pp. 35–43, July 2008.

[9] Y. Ma, J. Feng, L. Kumin, and J. Lazar, "Investigating user behavior for authentication methods: a comparison between individuals with down syndrome and neurotypical users," *ACM Transactions on Accessible Computing*, vol. 4, no. 4, pp. 1–27, 2013.

[10] V. Roth, K. Richter, and R. Freidinger, "A PIN-entry method resilient against shoulder surfing," in *Proceedings of the 11th ACM Conference on Computer and Communications Security (CCS '04)*, pp. 236–245, October 2004.

[11] S. B. Park, "A method for preventing input information from exposing to observers," Patent App. No. 10-2004-0039209, Korea, 2004.

[12] Passfaces Co., http://www.passfaces.com/.

[13] Raonsecure Co., http://www.raonsecure.com/.

[14] Dementor Co., http://www.dementor.co.kr.

[15] S. K. Card, T. P. Moran, and A. Newell, *The Psychology of Human-Computer Interaction*, Lawrence Erlbaum Associates, 1983.

[16] CogTool, http://cogtool.hcii.cs.cmu.edu.

[17] GOMSED, http://www-cgi.psychologie.tu-darmstadt.de/kogpsy/indexgoms.htm

[18] D. V. Beard, D. K. Smith, and K. M. Denelsbeck, "QGOMS: a direct-manipulation tool for simple GOMS models," in *Proceedings of the International Conference on Human Factors in Computing Systems (CHI '96)*, pp. 25–26, April 1996.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 530416, 13 pages
http://dx.doi.org/10.1155/2014/530416

Research Article

Mobility Aware Energy Efficient Congestion Control in Mobile Wireless Sensor Network

Awais Ahmad,[1] Sohail Jabbar,[2] Anand Paul,[1] and Seungmin Rho[3]

[1] *School of Electrical Engineering and Computer Science, Kyungpook National University, Daegu 702-701, Republic of Korea*
[2] *Department of Computer Science, COMSATS Institute of Information Technology, Sahiwal 57000, Pakistan*
[3] *Department of Multimedia, Sungkyul University, Anyang-si 430-742, Republic of Korea*

Correspondence should be addressed to Seungmin Rho; smrho@sungkyul.ac.kr

Received 29 November 2013; Accepted 1 January 2014; Published 4 March 2014

Academic Editor: Jongsung Kim

In this paper, we introduce a mobility aware and energy efficient congestion control protocol "time sharing energy efficient congestion control (TSEEC) for mobile wireless sensor network." TSEEC is based on hybrid scheme of time division multiple access protocol (TDMA) and statistical time division multiple access (STDMA) protocol that inform the sensor nodes when to wake up and to go to listening state so as to save energy. This management helps in minimizing congestion and improving network energy conservation through its load based allocation (LBA) and time allocation leister (TAL) techniques. LBA is typically based on STDMA that uses sensor node information for assignment of dynamic timeslots to the sensor, nodes, whereas TAL targets the mobility management of sensor nodes that further comprises three main strategies of cluster member node that is, joining cluster, leaving cluster, and absence of data/redundant data with extricated time allocation (ETA), shift back time allocation (SBTA), and eScaped Time Allocation (STA) subtechniques. In addition, TSEEC protocol introduces the mobility pattern to control the mobile sensor node (MSN) movement to enable the protocol to effectively adapt itself to change the traffic environments and mobility. Mathematical analysis and NS2 simulation show that TSEEC outperforms SMAC in terms of energy consumption and packet deliver ratio. Furthermore, a comparative analysis of TSEEC with various well-known related MAC protocols is also given.

1. Introduction

Wireless sensor network consists of autonomous sensors and uses a shared medium for monitoring physical environment like temperature, sound, pressure, or vibration and shares this data with a base station. These are surveillance devices that are of three basic elements: sensing subsystem, processing subsystem, and storage.

The advancement in static WSN along with distributed robotics [1] has led to a new set of Mobile WSN (MWSN). MWSNs have same network architecture as WSNs have; however they are provided with explicit or implicit mechanisms that provides mobility to these sensor nodes to move in space (e.g., terrestrial robotic car, underwater, or air current) over time [2]. In addition, MWSNs are able to derive their coordinates through relative means (localization techniques [3]) or absolute means (e.g., geographic positioning system (GPS)). There are quite a few classes of MWSN that can be plainly categorized into the following classes: (i) highly mobile; in this scenario, devices can move at a velocity such as human, cars, and airplanes, (ii) mostly static; in this scenario, the devices can move at a very low velocity such as moving robots, and (iii) hybrid, in this scenario, we have both classes, that is, highly mobile and mostly static, such as moving cars having sensors installed in it [2].

MWSN application has perceptibly diverse characteristics as well as requirements from traditional wireless network applications that pose some exceptional challenges, that is, energy efficiency and congestion. An MSN in MWSN environment is equipped with battery and recharging or changing the battery frequently is very difficult, and in some scenarios, it is almost not possible. Therefore, energy conservation is measured as the vital design confront in MWSN, which is essential for prolonging network lifetime than such other performance parameters such as latency, throughput, and bandwidth. Consequently, majority of the ongoing research

on MWSN aims to minimize the energy consumption ratio. Apparently, radio functions such as sending, receiving, or idle listening to the channel drain the energy a lot. On the other hand, congestion is another key factor that affects the energy conservation. Congestion usually arises when there is simultaneous transmission of data from multiple nodes to the sink node resulting in packet loss. The lost packets require retransmission of the same data packet that minimizes the overall energy efficiency of the network. In order to minimize the energy consumption ratio, an efficient MAC protocol that allocates medium resources shared by different sensor node in MWSN is required.

Due to aforementioned constraints, this research study is to implement a novel mobility aware energy efficient MAC protocol for MWSN. The scheduling (awake/listening states) algorithm developed for the proposed MAC is aiming to achieve relative better energy efficient for time-critical application traffics. Next is to accomplish the minimization of lower energy consumption. TSEEC uses mobility pattern that enables protocol to robustly adjust the mobility pattern, making it suitable for sensor environments suitable for both high and low mobilities. TSEEC assumes that MSNs know their location information by using any localization technique [2]. This location information is used to predict the mobility pattern by using [4].

In the proposed scheme, we have classified cause of congestion into two classes, that is, (i) node level congestion (NLC) and (ii) link level congestion (LLC). The NLC is caused by the buffer overflow whereas LLC is caused by large data being pumped into the channel by various neighboring nodes at the same time. Therefore, in this paper, a protocol is designed based on STDMA that improves performance in both NLC and LLC and ultimately results in energy efficiency of the network. With the help of STDMA, each MSN shares its statistical information in *Hello Packet* with cluster head (CH). This *Hello Packet* contains *unique ID*, *battery information*, and *location information* in the cluster. The CH then uses MSN feedback and uses a modified TDMA technique of timeslots allocation to its MSNs in specific cluster. This innovative statistical technique in TDMA provides energy efficiency to network and also minimizes congestion at sink node in MWSN. The technique works in a scenario wherein the mobile node may or may not sense any environmental physical quantity and leaves or joins new cluster and so on, while CH remains as the coordinator throughout this span of communication. Deficiencies in the static WSN [5, 6] are also handled by mobile sensor nodes.

The rest of the paper is organized as follows. Section 2 briefly describes review of some related work. Section 3 presents the proposed protocol (i.e., TSEEC). Section 4 shows the energy model for TSEEC. Finally, we present simulation results and discussions in Section 5, followed by conclusion in Section 6.

2. Related Work

Existing protocols that controls energy conservation, end-to-end network delay, and congestion in MWSNs are few among many of the issues. Congestion at the cluster head can lead to more overheads that consume a considerable amount of energy as well as packet loss and delay in the network.

In [5], randomized coordination algorithm that improves sensor networks efficiency and energy is described. This algorithm makes decision to sleep and wake up. The sensor node that wakes up is responsible to serve as a coordinator. The preference is given to hop-to-hop communication (short range) over direct communication (long range), which saves the network capacity. The number of coordinators should, however, be kept as minimum as possible to achieve maximum network life time. In [7] GAF scheme is proposed that uses geographical location information related to division of area into fixed square grids with constant size has been provided. All sensor nodes should follow the scheme of sleep and awake state to minimize the battery usage, but one sensor node should always be in awake state in order to route the data.

An algorithm PCAP is a MAC protocol based on CSMA which uses existing handshake mechanism of RTS and CTS which helps in avoiding collision in the network [8]. While waiting for CTS, PCAP permits the nodes to perform other tasks. The maximum throughput of PCAP is about 20%. CSMA technique however increases delay in the network since it uses RTS and CTS packets. To avoid congestion, various techniques and algorithms are used in WSN like CSMA/CA in IEEE 802.15.4 for CD-WPAN. This technique is used to introduce the beacon enable mode [9]. Sharing of information is done through cross-layer communication. This technique may however prove not so useful to avoid congestion in the network and loss of source packet. Key reason of dropping of packets is many-to-one communication under heavy traffic load.

One major concern of congestion on sink node is due to transmission of packets to common receiving node. This congestion is controlled by robust routing algorithm with fair congestion control (RRA-FCC) [10]. Typical wireless sensor network has 1 or 2 sink nodes, but there are numerous research scenarios where tens to hundreds of sink nodes are deployed in WSN region.

A scheme is proposed using mobile sink or minisink in WSN to avoid congestion [11]. Minisink is responsible for retrieving data from the sensor nodes. After retrieving data from nodes, aggregation function is applied to reduce the number of node's packets carried out on the network. Reduction in number of packets transmitted and limiting the number of hops in conservation of energy during transmission. WSN runs on batteries and if deployed in a region where reviving these batteries is impossible, then there is a great chance of failing of WSN. For such problems, TRAMA is suited protocol which can be used for energy conservation and for avoiding congestion [12]. It is a TDMA technique that divides the timeslots into two parts. One is random access protocol and the other is scheduled access period. The main drawback of this technique is cutting off the long communication process which results in dropping of packets.

A scheme is proposed based on zone based routing based on modification of adhoc on demand distance vector routing protocol (AODV) [13]. This scheme is capable of

reducing network segment error as well as reducing the amount of control information. This can be obtained by the effect of path finding and also addresses various problems of reliability, improved error control mechanism, and link repair with minimum overhead in mobile sensor networks. In this protocol, the member node transmits data to zone head in three phases, that is, mobility factor and zone head selection, route maintenance and node mobility.

A location based scheme is proposed based on routing protocol in which the concept of greedy forwarding on the basis of cost function for each node has been used [14]. This cost is close to the Euclidean length of the shortest path from sensor node to the base station. After a few hops when the greedy routing is not possible, the packet uses the high-cost-to-low-cost rule and forwards it to coordinator/sink node.

The original AODV protocol suffers from performance degradation in mobile environment, but it performs well in static environment [15]. Moreover, there is no mechanism in AODV protocol to derive the cost and willingness of a node to be a part of data transmission.

DMAC considers congestion on the sink node that consumes extra energy [16]. Robust routing algorithm with fair congestion control (RRA-FCC) [17] is proposed along with time-driven or event-driven models. RRA-FCC is not only to minimize the congestion and maximize energy efficiency in the network.

In WSN environment, the energy conservation depends on the distance between sensor nodes, The greater the distance between two nodes, the greater the energy consumption and vice-versa. In addition, another vital cause of sensor node's energy wastage is the state of the sensor node, that is, idle listening, collision, over hearing, central packet overhead, and overmitting [17]. To cope with the said problem, there are other schemes that generally minimize the energy consumption of a sensor node, that is, duty cycling and data-driven technique. But these schemes still needed better time allocation technique that helps in minimizing congestion in the WSN [17], CCP [18], SPAN [19], HEED [20], ECODA [21], and LCM [22] well-known protocols that are used to minimize congestion. However, the said schemes do not cope with the delay arise in the network. Therefore, based on these reservations, the proposed hybrid protocol "TSEEC" has been used to effectively circumvent all these constraints. (MICA2 Mote) is a basic example whose energy consumption is given in Table 1 [23]. Different techniques to compress data for energy conservation are discussed [24–27]. Due to compression, WSNs omit important data, which can be a loss for data manipulation.

In order to collect data information in an efficient manner, several algorithms that are used for efficient collection of data from WSN environment are proposed. These schemes use clustering approach that is based on residual energy. The cluster head rotating mechanism is also in introduced that enhance network life time. However, most of the algorithms have not yet considered the expected residual energy that is predicted the residual energy for being selected the cluster head. To avoid such problems, a fuzzy-logic based clustering approach along with extension of energy prediction is proposed. This scheme is used for prolonging network life by means of evenly distribution of workload amongst various sensor nodes [28].

Table 1: MICA sensor specification.

State	Energy consumption (mW)
Transmission	80
Reception/idle	30
Sleep	0.003

However, given brittle conditions of a mobile WSN, where link breakage is quite high among various mobile sensor nodes due to their abrupt movement, TSEEC technique works well by introducing mobility factor along with the scheduling technique. Furthermore, it also helps in minimizing the time required for mobile sensor node long communication period with CH which results in dropping of packets.

3. Proposed Solution

Application Scenario. Due to broadcast nature of WSN, it is a difficult job to incorporate congestion control at sink node. Therefore, a new MAC protocol is required to design focusing on energy efficiency and congestion control. The designed protocol results in prolonging network lifetime and enhancing network efficiency. In mobile network, there are two phases of a node: (i) mobility phase: node moves from one location to another location in the network and (ii) stagnant phase: between two mobility phases, there is stagnant phase where node stays at same location for some time interval. In the scenarios, where stagnant phase does not exist or at least exists for very short time, there the time sharing based algorithm is not possible to implement. In our underlying scenario for the designing of TSEEC, stagnant phase is long enough for a node to stay in some cluster that time sharing based algorithm can be used without any objection. Moreover, node location information is used for decision making of mobility model in our algorithm as shown in Figure 1. Hence, time sharing based algorithm can be implemented over the underlying scenario.

3.1. Overview. Consider an MWSN composed of n number of MSNs, that are deployed in x and y coordinate plane. The deployed nodes grouped together to form their respective cluster. In each cluster, they have their static cluster head (CH) as shown in Figure 2.

Formation of cluster is described in Section 3.1.1. CH assigns timeslots to its joined MSNs with the help of STDMA. The STDMA acquires MSN's statistical information, that is, *unique ID*, *battery information*, and *location information*. After getting statistical information from each MSN, the CH evaluates this statistical information and assigns dynamic timeslots to each MSN.

TSEEC comprises two main strategies which deal with delay in a network. This delay is caused by freeing up of assigned timeslots and other issues which have been mentioned above. These strategies are load based allocation (LBA)

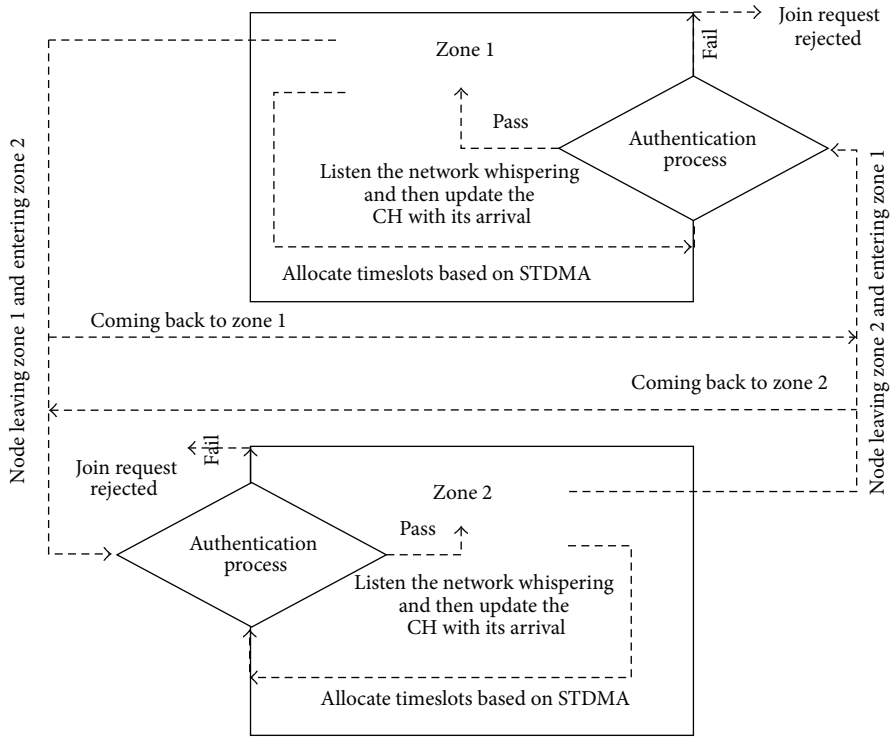


Figure 1: Typical working of TSEEC.

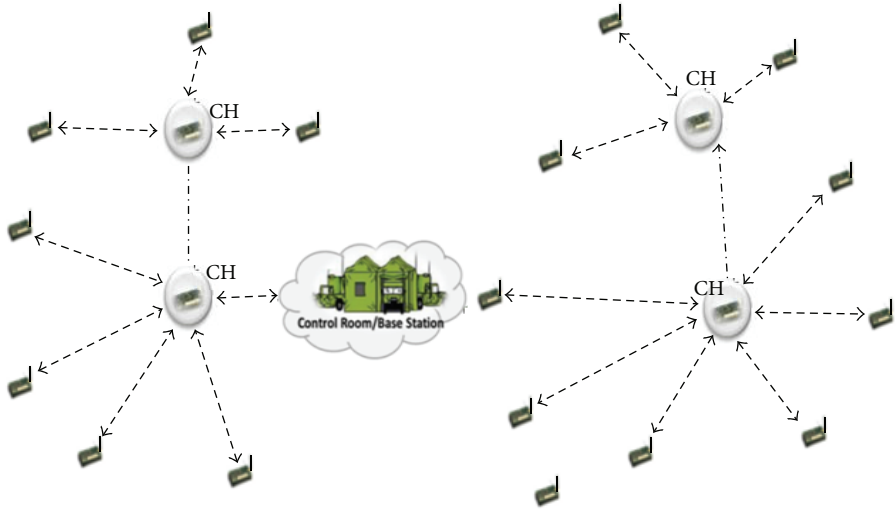


Figure 2: Deployment of sensor nodes along with cluster head.

and time allocation leister (TAL). Figure 3 summarizes the process of TSEEC.

3.1.1. Cluster Head Formation. Cluster Head (CH) is based on self-capability of organizing sensor nodes. Assume that all the MSNs know there location as well as their neighbors in its vicinity head (VH). We use the term vicinity for neighboring area of a sensor node, that is, signal range of a sensor node. However, the terminology vicinity head (VH) can be relevant to CH. The deployment of sensor nodes is settled on; thus, at the beginning, selection of CH is on hand. Since the nature of MSN is homogenous, therefore, the MSN with maximum neighboring nodes is selected as CH. After the selection of CH, each MSN attaches to the CH on the basis of received signal strength (RSSI). In case an MSN receives request from neighboring CH (CHs which are nearer to this node), then the following algorithm is followed:

$$Sweight_i > Sweight_j \quad \forall j, \tag{1}$$

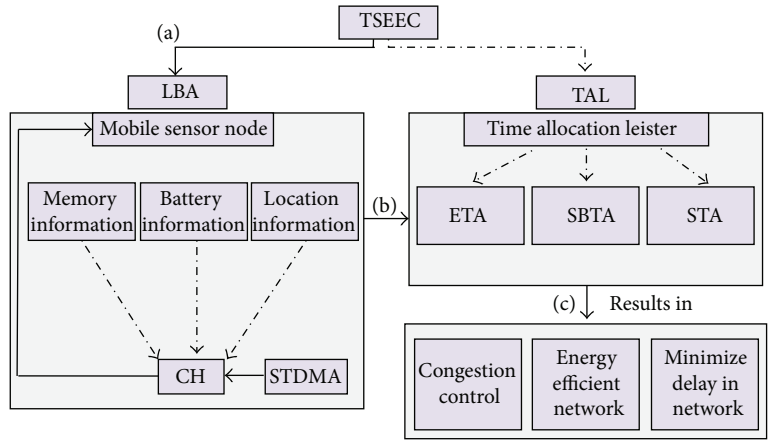


Figure 3: Working of TSEEC.

where *Sweight* is weight or strength of received signal of the invitee CH.

If

$$Sweight_i = Sweight_j \quad \forall j. \tag{2}$$

Then, the selection is on the basis of

$$Eweight_i > Eweight_j \quad \forall j, \tag{3}$$

where *Eweight* is the weight of energy level of invitee CH.

If

$$Sweight_i = Sweight_j \quad \forall j, \tag{4}$$

random selection is made.

3.1.2. Load Based Allocation. In this phase, CH assigns timeslots to each mobile cluster member node. The process of assigning timeslots works as follows: each MSN send a *Type-I Hello Packet* to the CH. The format of the *Hello Packet* consists of different information, that is, *Type-I_Broadcast_msg*{*node ID*, *battery information*, *location information*, *propagation time*}.

The CH computes the distance between MSN and itself using time of arrival (TOA). After calculating distance, the CH assigns weight to each MSN on the basis of distance and the information available in *Type-1_Broadcast_msg*. Similarly, TSEEC uses the weighted value and *broadcast_msg* as a metric for assigning timeslots to each MSN in the network. The weighted value is calculated by using (5)

$$\omega = \left(\frac{1 - (ToA)}{C} \times 100\right) + K, \tag{5}$$

where ω represents the weight of the particular node, K represents priority value of the candidate node, and C represents the cost, that is, statistical information of a node. C is calculated by taking the mean of the statistical information of the ith node. Since TSEEC uses the weighted value and *broadcast_msg* as a metric for assigning timeslots to each MSN in the network,

$$\text{Cost}(C_i) = (\text{battery life time of a node} + \text{Location information of a node})(2)^{-1}. \tag{6}$$

The algorithm for assignment of timeslots is shown in Algorithm 1.

After calculating the weighted value for each MSN, this information is broadcasted in the cluster in *Type-II_broadcast_msg*. The format of the *Type_II* broadcast message is as follows: *Type-I_Broadcast_msg*{*node id's, starting time of cycle, scheduling information, dynamic timeslots*}. Upon receiving *Type_II* broadcast message, each node becomes aware of each MSN's information that is in same cluster.

To elaborate the assignment of dynamic timeslots based on their statistical information, an example scenario is illustrated in Figure 4. In this scenario, we have considered a single cluster having five MSNs (n_1, n_2, n_3, n_4, and n_5) and one cluster head (CH_1). Assume that all MSNs broadcast *Type-I* message for the CH_1. Upon receiving *Type-I* message by CH_1, CH_1 calculates weight by using two types of information, that is, (i) the information of all nodes based on information inside the header of the broadcast message and (ii) the distance calculated by using ToA mechanism. The following metric Table 2 is assumed after calculating *Type-I* broadcast message information.

The information available in Table 2 is broadcasted by CH_1 in the cluster in *Type-II* broadcast message. Upon receiving *Type-II* broadcast message by all five MSNs, all the MSNs get the information of starting time of the data communication cycle and scheduling information along with other node's information. This mechanism helps in providing efficiency in time synchronization among all the MSNs that is, the (n_4) which is in communication with the CH_1, and the rest of MSNs (n_2, n_5, n_3 and n_1) turn their radio to listening state. In listening state, the MSNs can only listen to the channel but cannot communicate with other nodes

```
Mobile_sensor_node
MSN node_ID, Battery_life, Location_Info, time_of_sending
        Hello_Packet(){
node_ID = MSN → node_ID
                Battery_life = MSN → Battery_life
                Location_info = MSN → Location_Info
                Location_info = MSN → time_of_sending}
while(!total_mobile_sensor_nodes)
        MSN Hello_ Packet() to CH
        TOA(){
        distance_MSN = mobile_sensor_node → arrival_tim
        e-MSN → time_of_sending}
        CH assign Weights to MSN
        End while
```

Algorithm 1

Table 2: Example scenario of calculated weighted value.

Node ID	ToA (μsec)	Battery information (J)	Location information (x, y, and z coordinate plane)	Weighted value	Assigned timeslots (sec)
n_4	150	20J	i	25	20
n_2	135	18J	j	21	15
n_5	100	14J	k	18	13
n_3	85	12J	l	14	10
n_1	70	10J	m	13	5

or CH_1. After completing its timeslots by n_4, n_2 (which has the second highest weighted value), it turns on its radio and starts communicating with the CH, while n_4 turns its radio to listening state. This mechanism continues till the last node of the cluster accomplishes its communication.

After completing one cycle, the same mechanism continues for next cycle.

The strategy LBA, which is based on STDMA, helps in acquiring statistical information from different MSNs and assigns timeslots after the evaluation of statistical information. However, mobility aware TAL technique deals with the exploitation of free timeslots that is done after node's joining or leaving the cluster or absence of data in that particular sensing region. TAL mainly comprises three basic loops of joining, leaving and absence of data in the scenarios: extricated time allocation (ETA), shift back time allocation (SBTA), and eScaped time allocation (STA) as shown in Figure 5.

3.1.3. Time Allocation Leister. The rapid change in the location of sensor node results in modification in the network topology. The abrupt change in topology as well as location of different sensor nodes influences the MWSN's scheduling process and affects the allocation of timeslots in the network. The problem of rapid change in topology as well as location (joining and leaving of the sensor node from cluster to another cluster) is handled by TAL mobility model. The three prongs of TAL are extricated time allocation (ETA), shift back time allocation (SBTA) and escaped time allocation (STA).

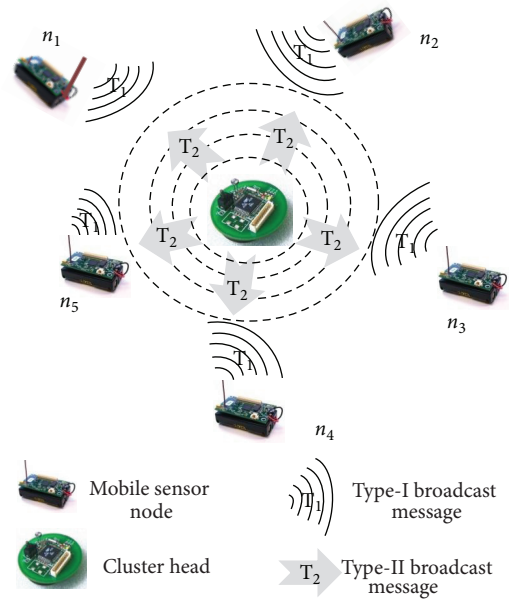


Figure 4: Scenario illustrating the Type-I and Type-II broadcast messages to calculate weight for each MSN and assign dynamic timeslots to all mobile sensor nodes.

(i) *ETA.* In this scenario, when there is absence of data, the allotted timeslots should be unutilized by that sensor node. In this technique, the MSN senses the

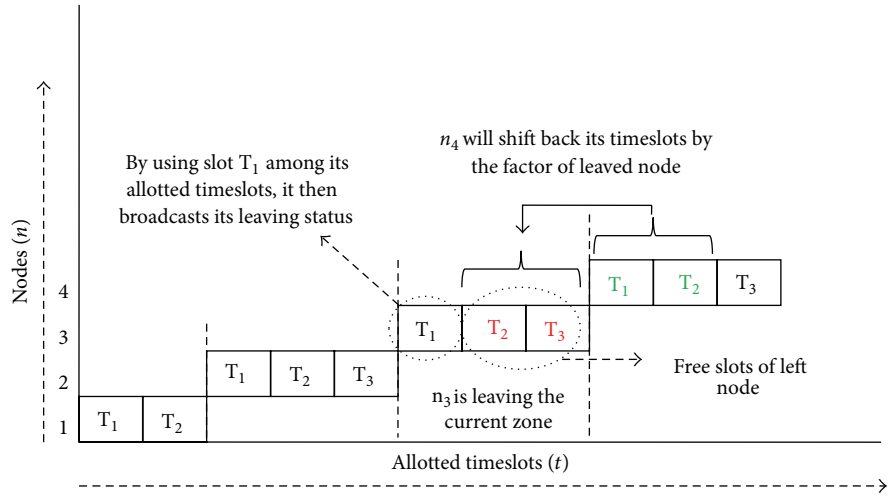


Figure 5: Sharing of timeslots among different mobile nodes.

same data which we call it "redundant data" or, sometimes, it does not sense data at all. After that, MSN broadcasts the free timeslot message for its neighboring MSN to shift back there timeslots as per free time timeslots kept over. Moreover, when new MSN enters in the cluster, the CH assigns that free timeslots to the new entrant MSN. This free timeslot must mull over two scenarios for its further utilization in order to provide efficiency to the mobile WSN as shown in Figure 5.

(a) In first scenario, CH assigns free timeslot to those MSNs that recently enter into new cluster.

(b) Secondly, assignment of free timeslots to the neighboring in that cluster by CH takes place (neighboring node will shift back its timeslot).

(ii) *SBTA*. This is the case in which MSN leaves its current cluster. Initially, it broadcasts its timeslots as well as leaving status. The neighboring nodes in the cycle shift back its timeslot as per free timeslots which are kept over by the departed MSN.

(iii) *STA*. STA comprises both ETA as well as SBTA. In this technique, MSN has no data to sense or has redundant data and other MSN for leaving the cluster.

Mobility Model for TAL. TSEEC uses location information to predict the performance of a mobile sensor node. However, localization technique is well studied in WSN [4]. However, most of the work assumes that the physical location of a sensor node is known. Therefore, TSEEC assumes that every MSN in the network knows their location information by using [29].

Let $\Gamma(\alpha + F_1) \rightarrow$ current mean (x, y) of α in F_i, where $x = x$ coordinates and $y = y$ coordinates.

In the proposed scheme, we used AR-1 model [4] that is used for MSN mobility pattern prediction. The state of the MSN at time t is defined by a column vector

$$S_t = \left[x_t, \dot{x}_t, \ddot{x}_t, y_t, \dot{y}_t, \ddot{y}_t\right]^{\tilde{}}. \tag{7}$$

In (7), S_t is the state of the mobility, (x_t, y_t) represents the position of a node, $(\dot{x}_t, \dot{y}_t)$ represents the velocity of a node, and symbol $(\tilde{})$ represents the metric transpose operator. However, $(\ddot{x}_t, \ddot{y}_t)$ represents the acceleration of a node in x and y direction, respectively.

Hence, AR-1 model [4] provides

$$S_{t+1} = As_t + \omega_t. \tag{8}$$

In this equation, A represents a 6×6 transformation metric and ω_t is the vector, that is, 6×1 discrete-time zero mean, White-Gaussian process with autocorrelation function $R_\omega(k) = \delta_k Q$, where $\delta_0 = 1$ and $\delta_k = 0$ when $k \neq 0$. Since the Q metric is the covariance metric of ω_t, the metric A values and covariance Q values are likely based on training data by using Yule-walker equation [30]. Interested readers can refer to [4, 31, 32] for more details.

Hence, at any given time t, the mobility state information of MSN $\dot{\tilde{s}}_t$ is used for prediction at any $t + 1$. However, the optimized predictable state $\dot{\tilde{s}}t+1$ of any MSN in the minimum mean square error sense is given by

$$\dot{\tilde{s}}t + 1 = A^i \dot{\tilde{s}}_t. \tag{9}$$

After estimating mobility of MSN, it is now required to broadcast the mobility information in the cluster. For this procedure, modification in the MAC header that contains the information about mobility of MSN has been done. Let F_i denote frame, and node α calculate the expected mean (x, y) of ith in a frame F_{i+1} as $\Gamma(\alpha + F_1) = \text{average}\ (\dot{\tilde{s}}_{t+0}, \dot{\tilde{s}}_{t+1}, \dot{\tilde{s}}_{t+2}, \ldots, \dot{\tilde{s}}_{t+\max})$.

After calculating the value of $\Gamma(\alpha + F_1)$, this information is added to header of every frame of ith node. The format of

```
N = node
Number of nodes = 50
Z = Zone (z1 , z2)
Cluster Head = CH1 , CH2
        Loop i ← 0 to 24
                z1 = node[i]
                Z1 = z1 + CH1
        Loop i = ← 25 to 50
                z2 = node[i]
                Z2 = z2 + CH2
        Loop j = 0 to 24
                Node[i] ← t[j]
        Loop j = 25 to 50
                Node[j] ← t[i]
Proc TAL
Loop i ← 0 to 24
          If (payload_node = = 0) // condition 1
                CH1 ← message ← node[i]
                        then
                                New_arr_node[i] ← t[i] ← CH1
else      // condition 2
        CH1 ← leave_msg ← node[i]
              Node[i + 1] ← t[i]
        CH1 → shiftback T of other nodes by a factor of ti + 1, ti + 2
else
combine (condition 1 + condition 2)
Repeat procTAL for zone 2
```

Algorithm 2

data packet is Format $\{\Gamma(\alpha + F_1)$, Header, Data$\}$. Rest of the nodes in the cluster keep the radio on to listening state for receiving of $\Gamma(\alpha + F_1)$. Each MSN in the cluster follows the same procedure and update information each time they get any data packet.

Algorithm 2 represents the working of TSEEC.

4. Hop-by-Hop Performance Evaluation in Mobile WSN

We have used the probabilistic approach for evaluating energy efficiency between two MSNs. We have classified our evaluation into three broad categories, that is, *Probability of packet reception* ($\text{Prob}_{(\text{PR})}$), *Probability of delay in delivery of packets* ($\text{Prob}_{(D)}$), and *Probability of retransmission* ($\text{Prob}_{(R)}$). At the end, we have calculated overall energy consumption by all nodes in the network. In MWSN, the probability of packet reception (between sender and receiver) can be calculated as

$$\text{Prob}_{\text{PR}} = \frac{\sum_{i=1}^{R} R}{\sum_{i=1}^{N} N}. \tag{10}$$

In (10), R is the number of packets which are received by a node and N is the number of data packets which are propagated.

For the measurement of average delay between two nodes deployed in wireless sensor network, the following equation is used:

$$\text{Prob}_{D} = \frac{\sum_{t=1}^{K} K}{\sum_{t=1}^{N} N}. \tag{11}$$

In (11), t is the time required to deliver K amount of packets successfully, N is the total number of packets which are transmitted and to be delivered in time t.

Retransmission consumes much energy. When packet is lost or delivered erroneously, that packet must be retransmitted. The node sends negative acknowledgment (NACK) to sender in its existing timeslots. While receiving NACK, the sender will retransmit that packet again. The probability of retransmission (by sender) can be found through the below equation:

$$\text{Prob}_{R} = \frac{\sum_{d=1}^{n} D - d}{\sum_{j=1}^{D} D}. \tag{12}$$

In (12), d is the sum of all packets, which are received by destination node, D is number of delivered packets, and d is the number of lost packets.

We next investigate how TSEEC reduces energy consumption. TSEEC plays a dominant role in consuming less energy due to time sharing TDMA based on STDMA. Let E_{ab} denote the energy consumption of the ath node in the bth

communication path. The overall energy consumption can be expressed as

$$E = \sum_{a=1}^{n} E_{ab}. \tag{13}$$

5. Energy Model

In TSEEC, an assumption has been made, that is, each MSN sends and receives data packets with an equal battery power, that is, sending and receiving. Therefore, the energy used by an MSN is autonomously depending on distance between adjacent MSNs. Consequently, we assume the below energy model to calculate power consumption for MSN in MWSN

$$P = E_{\mathrm{BR}} + E_{\mathrm{BT}}, \tag{14}$$

where P represents the power consumption of a sensor node, and E_{BR} represents the received bits, E_{BT} represents the transmitted bits.

Let S_D be the total amount of data sensed by each MSN per cycle "c"

$$D_T^c = D_R^c + S_D, \tag{15}$$

where D_T^c and D_R^c are the amount of data received and transmitted by a node "z" per cycle "c."

In this paper, we assume that all MSNs forward data in multihop to a CH. Equation (16) describes the relationship between the total amount of data received by all nodes and the sum of hops

$$\sum_{n=1}^{k} D_r^c = \sum_{n}^{k} H_n * Z, \tag{16}$$

where H_n represents the path from MSN "n" to its destination CH, if n itself is a gateway node, $H_n = 0$. (A gateway node is a node which is near to the CH.)

Based on (16), the total energy consumption can be expressed in terms of the total sum of all hops from all member nodes to their CH as follows:

$$\text{Power} = \sum_{n=1}^{k} P_n = \sum_{n=1}^{k} E\left(D_T^c + D_r^c\right), \tag{17}$$

where P_n represents the energy consumption of node "n" per cycle.

6. Experimental Results and Discussions

To study the evaluation performance and comparison of TSEEC protocol with other protocols and scheduling techniques, we have used NS2 simulator.

6.1. Simulation Environment. Mobile wireless sensor nodes were randomly deployed within an area of 1000 m × 1000 m. Simulations were performed with different number of MSNs, that is, 50 and 100.

The distances between MSNs are not kept constant variable and hence totally depend on the mobility pattern of each MSN in the given cluster. In the scenario of 50 MSNs, we have two clusters, each having a static CH. However, in the 100 MSNs scenario, there are 3 static clusters instead with the same ratio of one static CH in each cluster. The initial energy of each MSN was set to 15 Joules. In addition, we have used MAC type 802.11n in order to minimize network delay. 802.11n is used only by CH that avails frame aggregation technique [33] to deliver data packets at surface station. Each MSN generates data packets using constant bit rate (CBR), while *hello packet* is used only once at the start of the communication.

6.1.1. Message Interval Time (MIT). Figures 6(a) and 6(b) show the MIT of TSEEC with and without using periodic active/listening scheduling technique. In Figure 6(a), we have 50 MSNs with two clusters. Each cluster has 1 CH, that is, CH_1 and CH_2; both CH_1 and CH_2 have 25 members each. The scheduling technique with periodic active/listening scheduling technique consumes less energy as compared to the one without it. When periodic active/listening scheduling is being used, MSNs, waiting for their turn to start communication with the respective CH, switch their radios on to a listening mode. This technique helps consume least energy, that is, 35 Joules, only for listening without the need to transmit any data packets. However, without active/listening scheduling technique, all MSNs turn their radios to full active mode and sense the environment. For instance, when there are 50 sensor nodes, TSEEC without periodic active/listening technique consumes 68 Joule of energy. To reduce this energy consumption, TSEEC with periodic active/listening scheduling technique is preferred as it inactivates the extra sensing of an MSN. Hence, the active/listening scheduling technique consumes 35 Joule of energy throughout the communication period. On the other hand with an increase in the number of MSNs to 100, as shown in Figure 6(b), the amount of energy consumed by MSN also increases in without periodic active/listening, that is, 118 Joule of energy consumption. This is the result of additional packet losses due to poor allotment of timeslots to all MSNs. However, TSEEC gives us satisfactory results while used with periodic active/listening scheduling technique if the number of MSNs increases, that is, 43 Joules of energy consumption.

6.2. Energy Level of Mobile Sensor Nodes. Figure 7 shows the energy consumption of both of the protocols, that is, TSEEC and SMAC. For comparison purpose, we implement SMAC in MWSN and compare its results with TSEEC. In SMAC, there is a long but same schedule during sleep/listen period that drops extra amount of data packets during sleep state. The protocol like SMAC is not suitable for MWSN, as the latter has abrupt movement of sensor nodes that affects the scheduling process. In mobility aware WSNs, the data frame advertising a mobile node's location information could be dropped by a node which is in sleep state. As a result, the node in sleep state would not have the information about the mobility pattern of all nodes in a given cluster. In terms

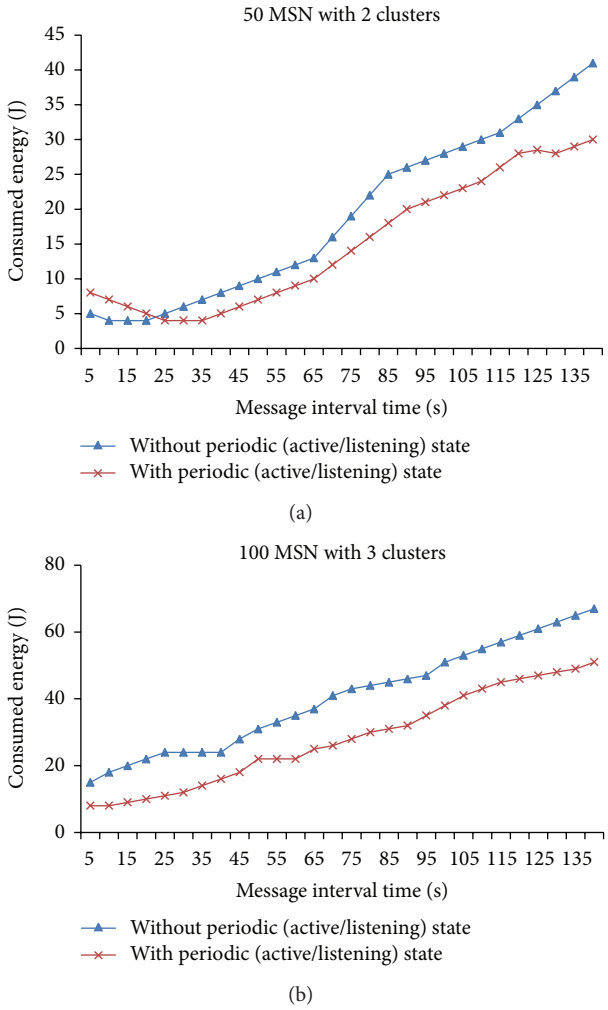


Figure 6: Interarrival time in seconds under low traffic load.

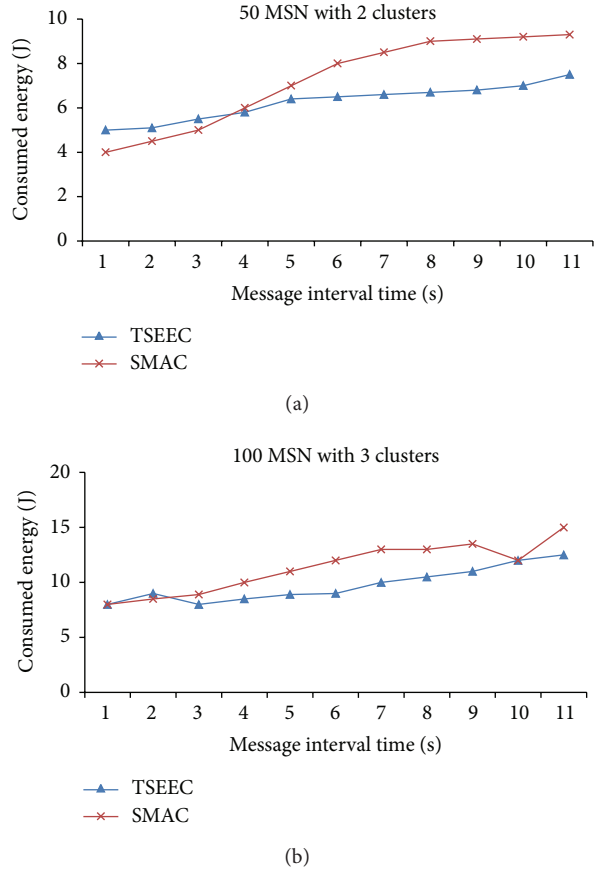


Figure 7: Calculating energy level at every mobile sensor node.

of the overall effect on energy consumption of the MWSN, whenever there is a fixed sleep/listen schedule, a network having mobile nodes is affected considerably. As shown in Figure 7(a), the SMAC consumes maximum energy of 9 Joules per single cycle (if we have 50 nodes in a network, that is, 25 nodes in each cluster). On the other hand, TSEEC consumes considerably lower energy as it uses a periodic with active/listening scheduling technique. For instance, if we increase the number of MSNs as shown in Figure 7(b), the performance of SMAC is considered worse (consumes 14.5 Joules of energy) as compared to TSEEC (consumes 11.5 Joules of energy).

6.3. Delivery Ratio. Figures 8(a) and 8(b) show the packet delivery ratio of both protocols. In SMAC, the long listen/sleep period results in energy saving. But, in this case, traffic load and the mobility pattern of MSNs result in dropped data packets. Since MSNs are unable to receive data packets (mobility information and timeslot allotment of other nodes), that leads to deficient information about the cluster environment.

This mechanism affects the overall data transmissions; hence, maximum amount of data packets is dropped due to lack of cluster information. As shown in Figure 7(a), increase in the speed of MSNs affects the scheduling process of SMAC; hence, delivery ratio decreases to 0.5%. However, TSEEC has 0.69% delivery ratio. For instance, if we increase the number of MSNs to 100 nodes as shown in Figure 7(b), the delivery ratio drastically decreases to 0.4%. However, TSEEC is not affected by increase in number of nodes in the cluster and has delivery ratio of 0.3%.

6.4. 802.11n Energy Level. Figure 9 shows the implementation of 802.11n at back end, that is, at the cluster head. There are various tests by employing scheduling techniques on CH, that is, idle listening, transmitting state, listening state and receiving state. It is noticed that CH in receiving state consumes less energy, that is, 2 Joules per one complete cycle of communication. In transmitting state, CH consumes 4.2 Joules of energy, since it uses the 802.11n aggregation technique (see [33] for more details) to deliver data to surface station. In listening state, the CH consumes negligible energy, as CH is always in listening and transmitting state. At the start of a communication period, it listens for member MSNs to get statistical information and assign timeslots. Since the CH cannot be kept in an idle state, it uses its radio throughout the

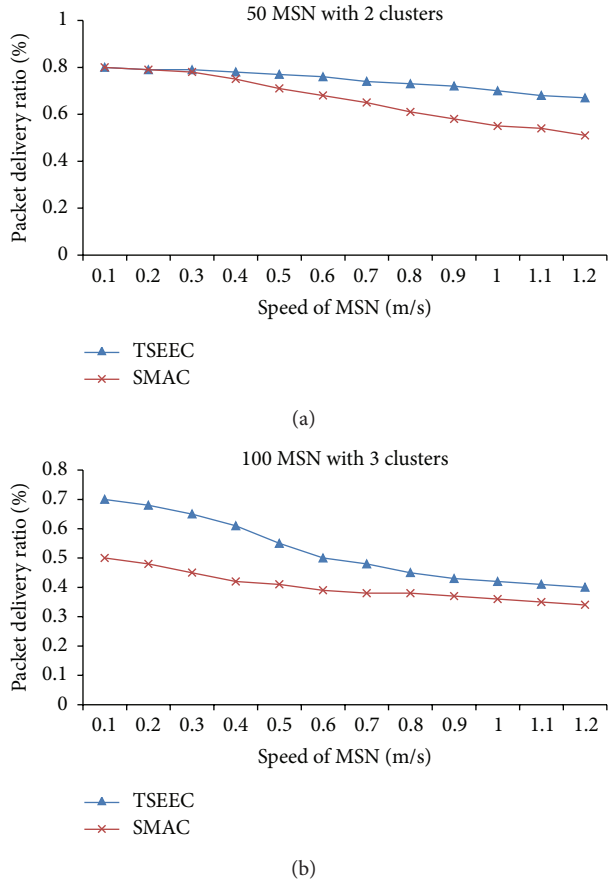


Figure 8: Calculated packet delivery ratio under low traffic Load (LTR).

communication period consuming 0.2 Joules of energy. The total consumed energy can be calculated as

$$\text{Consumed energy}\,(C_e) = \text{Total Energy}\,(T_E) - \text{Sum of energy in all states}\,(T_S). \quad (18)$$

C_e is the level of energy consumption of our entire scenario.

7. Comparative Analysis with Existing MAC Protocols

In Table 3, we compared our protocol with the existing protocol in terms of synchronization, communication pattern, protocol type, and adaptively changes to atmosphere. In the given table, two S-MAC protocols, namely, T-MAC and DSMAC have same features with S-MAC. Cross-layer MAC protocols are not considered in this comparison. However, WiseMAC and TRAMA have different characteristics. It is a TDMA based protocol which increases the use of TDMA in the context of energy efficiency. This protocol follows random allotment of time to different nodes. It uses high ratio of sleep time due to which less collisions take place in the network vis-à-vis the CSMA protocol. However, wireless

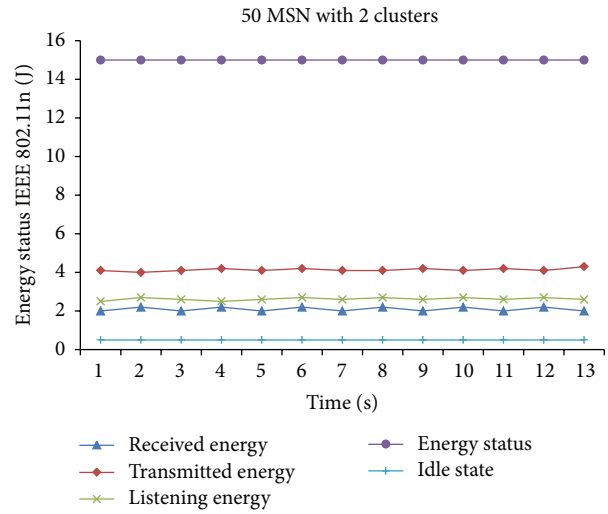


Figure 9: Energy status of IEEE 802.11n.

sensor MAC (WiseMAC) for the downlink of infrastructure WSN which is based on synchronized preamble sampling focuses on low traffic. WiseMAC is not a useful remedy for high traffic scenarios. SIFT protocol is used for event-driven sensor network and is used in the network when an event has been sensed in the environment. The crucial part of this protocol is the R and N reports based on low latency. If we compare 802.11 MAC protocol with SIFT, the latter clearly stands out in effectively decreasing latency in the network. DMAC protocol achieves very low latency in the network. Secondly, it is an energy efficient protocol. Low latency is achieved due to assignment of different timeslots to different nodes at leaf which helps in successive data transmission. While discussing features and challenges faced by WSN due to MAC protocols, the newly designed protocol TSEEC does not consider the challenging features like synchronization amongst mobile nodes, computational power, timeslots, and insertion of new nodes in the network. However, insertion of new nodes and timeslot sharing are welcomed by TSEEC to provide better throughput, avoid collision minimizes energy consumption, and efficiently sharing of timeslots.

8. Conclusion

Available diverse techniques are studied so far for congestion control, delay, and energy wastage in the network. We have also implemented our novel technique for congestion control, providing efficiency, and minimizing delay for message interval and interarrival time in WSN. Furthermore, allocation of timeslots with the help STDMA and sharing of timeslots in existence TDMA technique have come up with minimizing memory wastage, energy conservation, efficient allocation of timeslots, and minimizing communication delay. These techniques are implemented on mobile sensor nodes with different subareas along with static CH. From result and discussion, it is concluded that TSEEC has proved itself the most suited solution for congestion control, energy

Table 3: Comparison of TSEEC with existing MAC protocols.

	Synchronization of time	Communication pattern	Type	Adaptively to changes
S-MAC T-MAC DSMAC	No	All	CSMA	Average
WiseMAC	No	All	Np-CSMA	Average
TRAMA	Yes	All	TDMA/CSMA	Average
SIFT	No	All	CSMA/CA	Average
DMAC	Yes	ConvergeCast	TDMA/Slotted Aloha	Average
TSEEC	Yes	All	TDMA/STDMA	Good

reservation, minimizing delay, and providing efficiency to the network.

In our future work, we will work on finding the neighboring mobile sensor nodes. Mobile sensor nodes move freely causing rapid changes in topology of the WSN network. Rapid changes in topology influence TDMA scheduling that causes delay in the network. Moreover, further studies reveal that the scheme can be useful in various ubiquitous networks [26, 27, 29–34]. Hence, a mobility model that overwhelms the absence of the neighboring node is required.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2061978) and this project is supported by the research collaboration of COMSATS Institute of Information Technology, Sahiwal Pakistan, and Kyungpook National University 2013 research fund.

References

[1] S. Bergbreiter and K. S. J. Pister, "CotsBots: an off-the-shelf platform for distributed robotics," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1632–1637, October 2003.

[2] D. Zeinalipour-Yazti, H. Papadakis, C. Georgiou, and M. D. Dikaiakos, "Metadata ranking and pruning for failure detection in grids," *Parallel Processing Letters*, vol. 18, no. 3, pp. 371–390, 2008.

[3] L. Hu and D. Evans, "Localization for mobile sensor networks," in *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking (MobiCom '04)*, pp. 45–57, October 2004.

[4] Z. R. Zaidi and B. L. Mark, "Mobility estimation for wireless networks based on an autoregressive model," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '04)*, pp. 3405–3409, December 2004.

[5] D. Zorbas and C. Douligeris, "Connected coverage in WSNs based on critical targets," *Computer Networks*, vol. 55, no. 6, pp. 1412–1425, 2011.

[6] D. Zorbas, T. Razafindralambo, and I. Lilli, "Wireless sensor network redeployment under target coverage constraint," in *Proceedings of the 5th International Conference on Digital object IdentifierNew Technology, Mobility and Security (NTMS '12)*, pp. 1–5, 2012.

[7] Y. Xu, J. Hiedemann, and D. Estrin, "Geography informed energy conservation for Ad Hoc routing," in *proceeding of the 7th annual ACM lst IEEE International Conference on Mobile computing and Networking (MobiCom '01)*, Rome, Italy, July 2001.

[8] X. Guo, M. R. Frater, and M. J. Ryan, "A propagation-delay-tolerant collision avoidance protocol for underwater acoustic sensor networks," in *Proceedinsg of the Asia Pacific (OCEANS '06)*, May 2007.

[9] R. Sokullu and C. Donertas, "Combined effects of mobility, congestion and contention on network performance for IEEE 802.15.4 Based Networks," in *Proceedings of the 23rd International Symposium on Digital Object Identifier Computer and Information Science (ISCIS '08)*, 2008.

[10] Y. Liu, Y. Liu, J. Pu, and Z. Xiong, "A robust routing algorithm with fair congestion control in WSN," in *Proceedings of 17th International Conference on Digital Object Identifier Computer Communication and Networks (ICCCN '08)*, pp. 1–4, 2008.

[11] M. Khan, W. Gansterer, and G. Haring, "Congestion avoidance and energy efficient routing protocol for wireless sensor networks with a mobile sink," *Journal of Networks*, vol. 2, pp. 42–49, 2007.

[12] I. Chih-Lin and G. P. Pollini, "Tree-search resource auction multiple access (TRAMA) protocol for wireless personal communications," in *Proceedings of the IEEE 44th Vehicular Technology Conference*, pp. 1170–1174, June 1994.

[13] U. Ahmed and F. B. Hussain, "Energy efficient routing protocol for zone based mobile sensor networks," in *Proceedings of the 7th International Wireless Communications and Mobile Computing Conference (IWCMC '11)*, pp. 1081–1086, July 2011.

[14] L. Zou, M. Lu, and Z. Xiong, "A distributed algorithm for the dead end problem of location based routing in sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 4, pp. 1509–1522, 2005.

[15] S. R. Das, C. E. Perkins, and E. M. Royer, "Performance comparison of two on-demand routing protocols for ad hoc networks," in *Proceedings of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '00)*, vol. 1, pp. 3–12, 2000.

[16] K. Oh, S. Woo, S. Sung, and K. Kim, "Improved energy efficiency of DMAC with periodic full sleep cycle for wireless sensor networks with heavy traffics," in *Proceedings of the International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP '07)*, pp. 85–89, December 2007.

[17] G. Anastasi, M. Conti, M. Di Francesco, and A. Passarella, "Energy conservation in wireless sensor networks: a survey," *Ad Hoc Networks*, vol. 7, no. 3, pp. 537–568, 2009.

[18] X. Wang, G. Xing, Y. Zhang, C. Lu, R. Pless, and C. Gill, "Integrated coverage and connectivity configuration in wireless sensor networks," in *Proceedings of the 1st International Conference on Embedded Networked Sensor Systems (SenSys '03)*, pp. 28–39, November 2003.

[19] B. Chen, K. Jamieson, H. Balakrishnan, and R. Morris, "Span: an energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks," *Wireless Networks*, vol. 8, no. 5, pp. 481–494, 2002.

[20] O. Younis and S. Fahmy, "HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks," *IEEE Transactions on Mobile Computing*, vol. 3, no. 4, pp. 366–379, 2004.

[21] L. Tao and F. Yu, "ECODA: enhanced congestion detection and avoidance for multiple class of traffic in sensor networks," in *Proceedings of the 15th Asia-Pacific Conference on Communications (APCC '09)*, pp. 726–730, October 2009.

[22] S. S. Wang and Z. P. Chen, "LCM: a link-aware clustering mechanism for energy-efficient routing in wireless sensor networks," *IEEE Sensor Journal*, vol. 13, no. 2, 2013.

[23] MICA2 Mote Datasheet, http://www.xbow.com/Products/Product_pdf_files/Wireless.

[24] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): design and construction," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 626–643, 2003.

[25] C. Tang and C. S. Raghavendra, *Compression Techniques For Wireless Sensor Networks, Book Wireless Sensor Networks Kluwer Academic*, 2004.

[26] M. Wu and C. W. Chen, *Multiple bit stream image transmission over wireless sensor networks*, Book Sensor Network Operations IEEE & Wiley Interscience, 2006.

[27] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 80–94, 2004.

[28] J.-S. Lee and W.-L. Cheng, "Fuzzy-logic-based clustering approach for wireless sensor networks using energy predication," *IEEE Sensor Journal*, vol. 12, no. 9, 2012.

[29] M. Ali, T. Suleman, and Z. A. Uzmi, "MMAC: a mobility-adaptive, collision-free MAC protocol for wireless sensor networks," in *Proceedings of the 24th IEEE International Performance, Computing, and Communications Conference (IPCCC '05)*, pp. 401–407, April 2005.

[30] J. S. Lim and A. V. Oppenheim, *Advanced Topics in Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, USA, 1987.

[31] Z. R. Zaidi, B. L. Mark, and R. K. Thomas, "A two-tier representation of node mobility in ad hoc networks," in *Proceedings of the 1st Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks (IEEE SECON '04)*, pp. 153–161, October 2004.

[32] J. S. Lim and A. V. Oppenheim, *Advanced Topics in Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, USA, 1987.

[33] K.-T. Feng, Y.-Z. Huang, and J.-S. Lin, *Design of MAC-Defined Aggregated ARQ Schemes For IEEE 802.11n Networks*, Springer, Wireless Network, 2011.

[34] A.-C. Tsai, A. Paul, J.-C. Wang, and J.-F. Wang, "Efficient intra prediction in H.264 based on intensity gradient approach," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '07)*, pp. 3952–3955, May 2007.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 736312, 12 pages
http://dx.doi.org/10.1155/2014/736312

Research Article

High Performance and Low Power Hardware Implementation for Cryptographic Hash Functions

Yunlong Zhang,[1] Joohee Kim,[1] Ken Choi,[1] and Taeshik Shon[2]

[1] *Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616, USA*
[2] *Division of Information and Computer Engineering, College of Information Technology, Ajou University, San 5, Woncheon-Dong, Yeongtong-Gu, Suwon 443-749, Republic of Korea*

Correspondence should be addressed to Ken Choi; kchoi@ece.iit.edu

Received 12 September 2013; Accepted 4 January 2014; Published 2 March 2014

Academic Editor: Jongsung Kim

Since hash functions are cryptography's most widely used primitives, efficient hardware implementation of hash functions is of critical importance. The proposed high performance hardware implementation of the hash functions used sponge construction which generates desired length digest, considering two key design metrics: throughput and power consumption. Firstly, this paper introduces unfolding transformation which increases the throughput of hash function and pipelining and parallelism design techniques which reduce the delay. Secondly, we propose a frequency trade-off technique which can give us a scope of frequency value for making a trade-off between low dynamic power consumption and high throughput. Finally, we use load-enable based clock gating scheme to eliminate wasted toggle rate of signals in the idle mode of hash encryption system. We demonstrated the proposed design techniques by using 45 nm CMOS technology at 10 MHz. The results show that we can achieve up to 47.97 times higher throughput, 6.31% delay reduction, and 13.65% dynamic power reduction.

1. Introduction

The explosion of e-commerce nowadays boosts the transaction over the internet; thus we have to prevent intruders from accessing the sensitive information. According to this circumstance, we call for higher security level protection. There are many types of modern cryptography, for example, symmetric-key cryptography, public-key cryptography, and cryptographic hash function. Cryptographic hash function is used in almost every modern application, especially in a multitude of protocols, be it as digital signatures for achieving message authentication and integrity protection. For example, hash-based message authentication codes (HMACs) are used in IP security protocol and also in secure sockets layer (SSL) protocol [1].

As we know, some hash functions, such as message-digest algorithm (MD) series (MD4 and its strengthened variant MD5) and secure hash algorithm (SHA) series (SHA-0 and SHA-1), were widely used, however, broken in practice. Considering the potential danger of being attacked for SHA-2, in 2008, the National Institute of Standards and Technology (NIST) has started the NIST hash competition to develop the future hash standard SHA-3 [2].

Although software encryption is becoming more prevalent today, hardware design is the embodiment of choice for many commercial applications and military [3]. Firstly hardware design is much faster than the corresponding software implementation [4]. Secondly, hardware implementation provides physical protection as high level of security [5]. However, higher security level hash function means more complicated gates, and much more information needs higher frequency to improve the efficiency (or throughput). As a result, the power dissipation of hardware design would increase tremendously. This will cause serious problems in hardware systems, such as less reliability, higher energy consumption, and higher device costs. Thus, low power techniques are highly appreciated in nowadays hardware design.

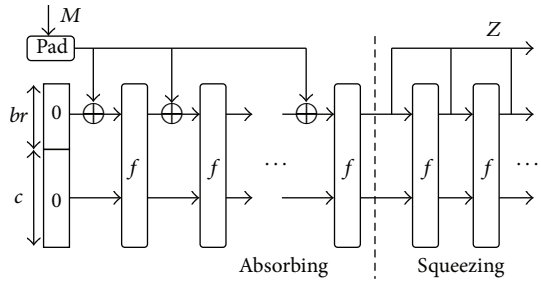


Figure 1: Sponge construction [6].

The rest of this paper is organized as follows. Sponge construction and low power methods which are used in this paper will be introduced in Section 2. In Section 3, we analyze the hash function designed by sponge construction and its original hardware implementation, and then unfolding transformation and pipelining and parallelism design techniques used to improve the throughput and delay of hash function are presented. In Section 4, we construct the hash encryption system and introduce two low power techniques, the frequency trade-off technique and load-enable based clock gating scheme. This paper is concluded in Section 5.

2. Background of the Research

In this section, first, sponge construction will be explained. Next, we will introduce two dynamic power reduction methods which are used in this paper.

2.1. Sponge Construction. The idea of sponge construction came from the design of RadioGatún, and its final definition was given at the Ecrypt Hash Workshop in Barcelona [6]. As shown in Figure 1, sponge construction takes arbitrary length input with finite internal state and gives an output of any desired length.

There are three components in sponge construction [7]:

(i) a state memory;

(ii) a function of fixed length that permutes or transforms the state memory;

(iii) a padding function.

The state memory in Figure 1 is divided into two parts: the top section called bitrate of br bits and the bottom section called capacity of c bits. And the input message (M in Figure 1) will be padded as a whole multiple of the bitrate. Thus this padded input message could be broken into many br-bit blocks.

Sponge construction consists of two processes: absorbing and squeezing. Considering the left part of dash line in Figure 1, called absorbing, firstly, the input message is padded and the state memory will be initialized; secondly, the first br-bit block of padded input will be XORed with the initial br bit of state memory; thirdly, the fixed length function (block f in Figure 1) updates the state memory. Then steps two and three will be repeated until all the padded br-bit blocks are used up. Considering the right section which is squeezing, firstly, the br bit of the latest state memory is the first br-bit output; secondly, if we need more output bits, the fixed length function is used to update the state memory and the br bit of new state memory is the second br-bit output. This process is repeated until the desired number of output bits (Z in Figure 1) is produced.

The extent c-bit part which is altered by the input message depends on the fixed length function [7]. The security of hash function, for example, resistance to collision or preimage attacks, relies on this c-bit part. Because of its arbitrarily long input and output sizes, the sponge construction allows building various primitives such as hash function. Keccak hash function, known as the new SHA-3, uses this sponge construction.

2.2. Dynamic Power Reduction Methods. Digital circuits will consume dynamic power in the active mode. There are two sources of dynamic power consumption [8]:

(i) charging and discharging processes of output capacitance;

(ii) short-circuit current when PMOS and NMOS networks are all ON.

Because the short circuit power is usually less than 10% of total dynamic power [9], the dynamic power consumption which we try to reduce in this paper is referred to as switching power for the rest of this paper. Dynamic power can be explained in (1). Note that f is the clock frequency and TR is the toggle rate of gate output:

$$P_{\text{dynamic}} = \frac{1}{2} C_L V_{\text{DD}}^2 f \cdot \text{TR}. \quad (1)$$

Since the power optimization at RTL has significant impact with reasonable accuracy, RTL is considered as the optimal stage for low power techniques [8]. According to (1), four parameters, such as voltage, clock frequency, load capacitance, and the toggle rate of gate output, determine the dynamic power consumption. Because reducing supply voltage will increase critical path delay and changing the capacitance of gate output needs to redesign the load logic, it is more efficient to focus on clock frequency and toggle rate at RTL.

2.2.1. Dynamic Voltage/Frequency Scaling. Figure 2 gives us a basic dynamic voltage/frequency scaling (DVFS) system. The DVFS controller will determine the clock frequency, which is sufficient to finish work and gives the best performance without overheating by collecting information about the workload and the temperature. Then this variable clock frequency scheme will lead to dynamic power reduction by choosing proper clock frequency.

2.2.2. Load-Enable Based Clock Gating. As we all know, combinational clock gating technique is widely used to solve dynamic power issue for single level register. And sequential clock gating method considers multiple level (pipeline) registers. In this research, we focus on the combinational clock

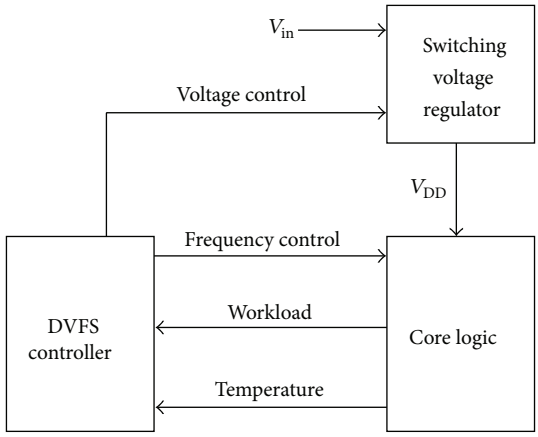


Figure 2: DVFS system [9].

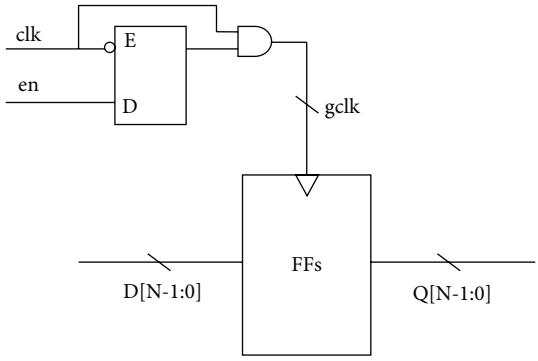


Figure 3: Load-enable based clock gating.

gating technique; particularly, we use load-enable based clock gating scheme [10].

Figure 3 shows a normal structure of load-enable based clock gating scheme. As we know, if the data do not change during some consecutive clock periods or the enable signal is kept low, those clock periods are wasted. This technique can be applied to a circuit with *mux* in which an enable signal is a selection signal or a pipeline construction circuit, such as hash encryption system in this research.

3. Proposed High-Speed Hashing Module in Hardware

Cryptographic hash function provides powerful protection for data; it has been utilized in the security layer of every communication protocol. However, as protocols evolve, data sizes and communication speeds are dramatically increasing; low throughput of hash function seems to be a bottleneck in these digital communications systems. A promising solution is the hardware implementation on reconfigurable devices which combines high flexibility with the speed and physical security.

Various techniques have been proposed to speed up or to improve the throughput of hash function, for example, unfolding transformation and pipeline and parallelism techniques. In this section, the characteristics which are relevant to the hardware implementation of the hash algorithm will be presented. Then the high-speed hashing methodology module will be introduced based on the delay bound analysis. Then two techniques, such as unfolding transformation and pipeline and parallelism, will be used to optimize the inner logic of transformation rounds.

Table 1: The parameters of SHAT.

SHAT	Hash value	Number of steps
SHAT-128	128 bits	48
SHAT-256	256 bits	48
SHAT-384	384 bits	48

3.1. Hash Algorithm Specification. In this section, we introduce a cryptographic hash algorithm with sponge construction, called sponge hash algorithm (SHAT). SHAT is a hash function generating 128-/256-/384-bit hash values. According to the hash value length, SHAT can be denoted by SHAT-$(128 \cdot i)$ $(i = 1, 2, 3)$. The parameters of SHAT are shown in Table 1.

3.1.1. G Function. G function of SHAT consists of an S-box and a diffusion layer. S-box is a substitution function that satisfies the confusion property on each 4-bit word. A 32-bit input word W, for example, is divided into eight 4-bit words $(w_0, \ldots, w_7)$. Each 4-bit word needs to go through this S-box. The definition of the S-box is $sw_i = S\text{box}(w_i)$ $(i = 0, \ldots, 7)$. This S-box is specified in Table 2. The diffusion layer is a permutation that satisfies the diffusion property (the same as the P function of Camellia [11]). Considering computational efficiency, this diffusion layer should be represented using only bit-wise exclusive ORs. The branch number of diffusion layer

$$\begin{pmatrix} w'_0 \\ w'_1 \\ w'_2 \\ w'_3 \\ w'_4 \\ w'_5 \\ w'_6 \\ w'_7 \end{pmatrix} = \begin{pmatrix} 01111001 \\ 10111100 \\ 11010110 \\ 11100011 \\ 01111110 \\ 10110111 \\ 11011011 \\ 11101101 \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ w_7 \end{pmatrix} \quad (2)$$

should be optimal against differential and linear cryptanalysis for security [11]. When we get all eight 4-bit outputs of S-box $(sw_0, \ldots, sw_7)$, this diffusion layer mixes them. Diffusion layer is defined as (2).

3.1.2. Hash Function of SHAT. SHAT uses the hermetic sponge construction as shown in Figure 4. As we mentioned in Section 2, br is called bitrate and c is called capacity. And

Table 2: S-box of the G function.

sw	$Sbox(w)$	sw	$Sbox(w)$
0×0	0×1	0×8	$0 \times F$
0×1	0×2	0×9	0×8
0×2	0×4	$0 \times A$	0×9
0×3	$0 \times B$	$0 \times B$	0×7
0×4	$0 \times D$	$0 \times C$	0×6
0×5	$0 \times E$	$0 \times D$	0×3
0×6	$0 \times A$	$0 \times E$	0×0
0×7	0×5	$0 \times F$	$0 \times C$

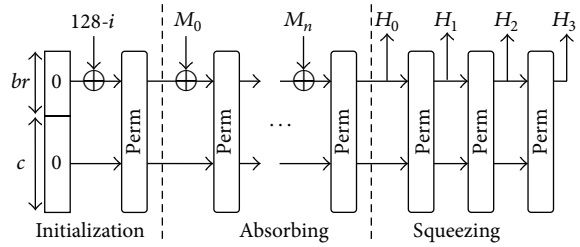


Figure 4: Sponge construction of SHAT.

the bitrate (br) and the capacity (c) of SHAT-$(128 \cdot i)$ ($i = 1, 2, 3$) are $32 \cdot i$ and $96 \cdot i$, respectively. The internal state, S, is divided into $4 \cdot i$ ($i = 1, 2, 3$) sections as $S = (S_0, \ldots, S_{4i-1})$ ($i = 1, 2, 3$).

In the absorbing phase, the input message $M = (M_0, M_1, \ldots, M_{n-1})$ shown in Figure 4 is padded as a whole multiple of bitrate (br). Then we will explain our padding method; l is the total length of input message (we assume that l is whole multiple of four as integer multiples of hexadecimal number), and then we append 1 to the end of the message, followed by k bits zero where k is the smallest nonnegative integer to set up the following formulation:

$$(l + 1 + k) \bmod (32 \cdot i) = 0. \tag{3}$$

Then, we set S_{4i-1} as the bitrate that used to be XORed with the padded br-bit message block. Then the result goes through that one-way compression function, Perm. Perm is a permutation process which has 48 steps. Each STEP is defined in Algorithm 1. In Algorithm 1, the left circular rotations rot_k are $\text{rot}_0 = 19$, $\text{rot}_1 = 1$, and $\text{rot}_2 = 14$. In the squeezing phase, SHAT was defined in (4). This SHAT-$(128 \cdot i)$ ($i = 1, 2, 3$) is specified in Algorithm 2:

$$\text{SQUEEZE}(S, i) = \begin{cases} S_3, & i = 1; \\ S_3 \parallel S_7, & i = 2; \\ S_3 \parallel S_7 \parallel S_{11}, & i = 3. \end{cases} \tag{4}$$

3.2. Hardware Implementation. Following the guidelines of SHAT-$(128 \cdot i)$ ($i = 1, 2, 3$) as shown in Algorithm 2, the architecture of SHAT is illustrated in Figure 5.

S-box of G function is designed from Karnaugh map. According to Table 2, we get the logic functions of S-box as shown in (5). We set A_i ($i = 0, 1, 2, 3$) as the input bit of S-box and Q_i ($i = 0, 1, 2, 3$) as the output bit:

$$\begin{aligned} Q_3 &= \overline{A}_3\overline{A}_2A_1A_0 + A_3A_2A_1A_0 + \overline{A}_3A_2\overline{A}_1 \\ &\quad + \overline{A}_3A_2\overline{A}_0 + A_3\overline{A}_2\overline{A}_1 + A_3\overline{A}_2\overline{A}_0, \\ Q_2 &= A_3\overline{A}_1\overline{A}_0 + \overline{A}_3A_2\overline{A}_1 + A_3A_1A_0 \\ &\quad + \overline{A}_3A_2A_0 + \overline{A}_3\overline{A}_2A_1\overline{A}_0, \\ Q_1 &= A_3\overline{A}_1\overline{A}_0 + A_2\overline{A}_1A_0 + \overline{A}_3\overline{A}_2A_0 \\ &\quad + \overline{A}_2A_1A_0 + \overline{A}_3A_2A_1\overline{A}_0, \\ Q_0 &= \overline{A}_3\overline{A}_1\overline{A}_0 + \overline{A}_3A_1A_0 + A_3A_2\overline{A}_1A_0 \\ &\quad + A_3\overline{A}_2\overline{A}_0 + A_3\overline{A}_2A_1. \end{aligned} \tag{5}$$

There are 48 iteration rounds in the basic architecture of Perm function. Then we use rolling loop technique to reduce area requirement. Our design is a single operation block which is reused 48 times as shown in Figure 6. Here r_i ($i = 1$ to 47) is a counter for the number of iteration rounds from 0 to 47. The critical path is highlighted by bold line. Since the delay of circular shift is negligible in hardware implementation, the critical path delay of this architecture is shown as

$$\widehat{T}_n = 4 \cdot \text{Delay}(\oplus) + \text{Delay}(g). \tag{6}$$

3.3. Proposed High-Speed Module. In the previous section, we introduce rolling loop technique to construct Perm function. Although this approach considers area efficiency, throughput is kept low due to the requirement of 48 clock cycles to generate the result. There are many architectures that can be made by varying the Perm function to solve this problem. We performed the unfolding transformation technique. This high-speed module combines STEP blocks into a single round and even can take advantage of architectures with complete round-unrolled circuit. By unfolding, the hidden concurrencies can be parallelized [12]. Also in [13], the pipeline and parallelism technique was explained to improve the unfolding construction of hash function. This technique is related to precomputing by analysing the inner logic and architecture of hash function.

3.3.1. Unfolding Transformation. According to Figure 6, the mathematical expression of one iteration round is described as

$$\begin{aligned} S_3' &= \text{ROT}(S_1') \oplus (S_0' \oplus S_2), \\ S_2' &= S_1, \\ S_1' &= G(S_3 \oplus r \oplus S_2) \oplus (S_0 \oplus S_1), \\ S_0' &= S_3 \oplus r. \end{aligned} \tag{7}$$

Step(S)

(i) For $k = 0$ to $i - 1$
  (a) $S_{4k+3} = S_{4k+3} \oplus r$;
  (b) $S_{4k} = S_{4k} \oplus S_{4k+1}$;
  (c) $S_{4k+2} = S_{4k+2} \oplus S_{4k+3}$;
  (d) $S_{4k} = S_{4k} \oplus G(S_{4k+2})$;
  (e) $S_{4k+2} = S_{4k+2} \oplus (S_{4k} <<< \text{rot}_k)$;
(ii) Temp $= S_{4i-1}$;
(iii) For $k = 4i - 1$ to 1
  $S_k = S_{k-1}$;
(iv) $S_0 =$ Temp;

Algorithm 1: Typical one step algorithm.

SHAT-$(128 \cdot i)(M)$

Inputs: n padded message blocks $M = (M_0, M_1, \ldots, M_{n-1})$
Outputs: $(128 \cdot i)$-bit hash value (H_0, H_1, H_2, H_3)

(1) $S = (S_0, \ldots, S_{4i-1}) = (0, 0, \ldots, 0, 128 \cdot i)$; // initialization
(2) Perm(S)
(3) For $j = 0$ to $n - 1$ // absorbing phase
  (i) For $k = 0$ to $i - 1$
    $S_{4k+3} = S_{4k+3} \oplus M_{j,k}$;
  (ii) Perm(S);
(4) $H_0 =$ SQUEEZE(S, i); // squeezing phase
(5) For $k = 1$ to 3
  (i) Perm(S);
  (ii) $H_k =$ SQUEEZE(S, i);

Algorithm 2: SHAT-$(128 \cdot i)$.

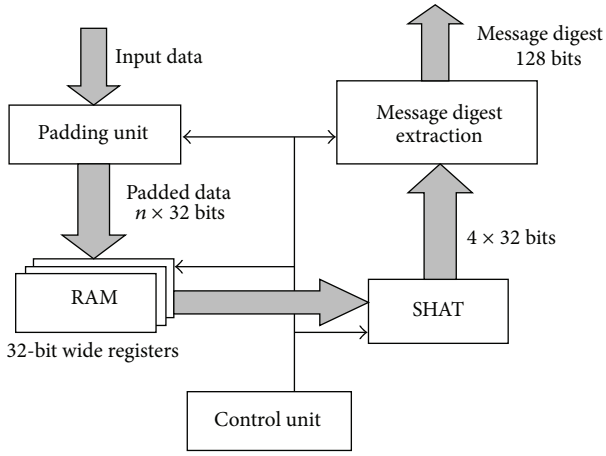


Figure 5: A typical SHAT core.

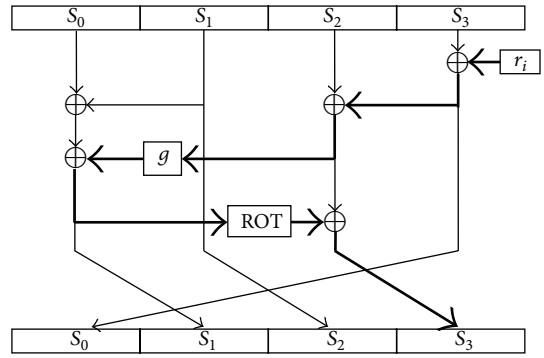


Figure 6: Typical architecture of one STEP round.

Here S_i $(i = 0, 1, 2, 3)$ is the input of current round and S_i' $(i = 0, 1, 2, 3)$ is the output of this round (or input of next round). In order to distribute 48 operations equally over each round, the possible values for unfolding factors are divisors of 48, that is, 1, 2, 3, 4, 6, 8, 12, 16, 24, and 48. For example, we can unfold two STEP operations in each round; then we get 24 rounds in one permutation process. The expression of throughput is given as

$$\text{Throughput} = (\#\text{ of bits}) \cdot \frac{f_{\text{round}}}{\#\text{ of rounds}}. \quad (8)$$

Considering (7), although this unfolding transformation reduces the maximum operation frequency, the throughput is increased significantly due to the fact that the operation

numbers are reduced from 48 to 24. The mathematical expression of one iteration round is replaced by

$$\begin{aligned}
\text{temp}_3 &= \text{ROT}\left(\text{temp}_1\right) \oplus \left(\text{temp}_0 \oplus S_2\right),\\
\text{temp}_2 &= S_1,\\
\text{temp}_1 &= G\left(S_3 \oplus r \oplus S_2\right) \oplus \left(S_0 \oplus S_1\right),\\
\text{temp}_0 &= S_3 \oplus r,\\
S_3' &= \text{ROT}\left(S_1'\right) \oplus \left(S_0' \oplus \text{temp}_2\right),\\
S_2' &= \text{temp}_1,\\
S_1' &= G\left(\text{temp}_3 \oplus r \oplus \text{temp}_2\right) \oplus \left(\text{temp}_0 \oplus \text{temp}_1\right),\\
S_0' &= \text{temp}_3 \oplus r.
\end{aligned} \tag{9}$$

3.3.2. Pipeline and Parallelism. We assume to unroll two STEP operations in each round; for sure it will reduce the frequency to increase the throughput. However, the increased area is introduced as penalty. If some logics can be done in parallel, and this parallelism happens in critical path, then the delay of each round could be decreased, so that the frequency of each operation will be increased. According to (8), when the number of operations is kept as constant (the number of bits is also kept as constant), the throughput will increase with its frequency. This method could be used in any other hardware implementation of hash function.

For example, Figure 7 shows the architecture of unfolding two STEP operations in one round, which has the minimum critical path delay. The critical path is composed of seven XOR gates and two G functions. By unfolding two STEPs in one round, we have a gain of three 32-bit XOR gates and one G function in critical path comparing with the architecture of one STEP block. The critical path is highlighted by bold line.

In Figure 7, cycle counter r_{i+1} can be calculated with temp_2 first, and then XORed with temp_3 in second STEP part. Comparing with the first STEP part where r_i XORed with S_3 and then XORed with S_2, we can figure out that there is another additional component which used to make a calculation with temp_3 and r_{i+1}. Because of the mandatory output generation necessity, this area penalty cannot be avoided.

Thus, when we increase the number of unfolding STEP operations, for example, three, four, five, . . ., each round delay will increase by three 32-bit XOR gates and one G function. Therefore, the normalized delay with unfolding factor n ($n = 1, 2, 3, \ldots$) is shown as

$$\widehat{T}_n = \frac{4 \cdot \text{Delay}\left(\oplus\right) + \text{Delay}\left(g\right) + (n-1) \cdot \left(3 \cdot \text{Delay}(\oplus) + \text{Delay}(g)\right)}{n}. \tag{10}$$

When we have a limit of n, (10) could be changed into

$$\lim_{n \to \infty} \widehat{T}_n = 3 \cdot \text{Delay}\left(\oplus\right) + \text{Delay}\left(g\right). \tag{11}$$

This is the delay bound of SHAT, which means that a delay of one SHAT operation round cannot be less than this bound.

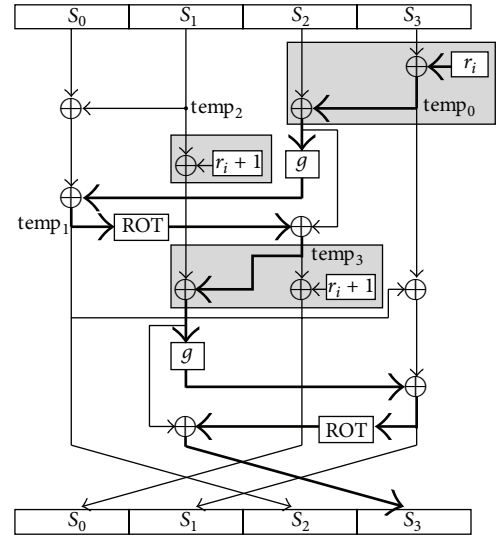


Figure 7: Proposed architecture of two STEPs round.

3.4. Experimental Results. We introduce a measurement of hardware efficiency in (12) [14]. This is the improvement of normal figure of merit (FOM). We assume that the power is proportional to the gate count; then we could divide the metric by another GE instead of power dissipation when we want to trade off throughput for power. Note that one gate equivalent (GE) is equal to the area of two-input NAND gate in 45 nm CMOS technology:

$$\text{FOM} = \frac{\text{Throughput}}{\text{GE}^2}. \tag{12}$$

Table 3 shows the hardware implementation results of some 128-bit hash functions by using 100 kHz clock frequency and 45 nm CMOS technique. Firstly, the throughput of SHAT-128 (66.67 kbps) is less than that of other 5 hash algorithms, such as MD4 (112.28 kbps), MD5 (83.66 kbps), H-Present-128-32-round (200 kbps), and ARMADILLO2-B (250 kbps and 1000 kbps). However the area of SHAT-128 is only 28.42% of that of hash functions in average. This results in having the hardware efficiency of SHAT-128 to be 13.12 times higher in average. Secondly, the area of SHAT-128 (1605 GE) is larger than that of 3 hash algorithms, for example, U-QUARK-544-round (1379 GE), PHOTON-128-996-round (1122 GE), and SPONGENT-128-8-bit-2380-round (1060 GE); however, the throughput of SHAT-128 is 94.27 times higher. Thus the FOM of SHAT-128 is 46.75 times higher in average. Finally, the area of SHAT-128 (1605 GE) is less than that of other 4 hash algorithms, for example, H-Present-128-559-round (2330 GE), U-QUARK-68-round (2392 GE), PHOTON-128-156-round (1708 GE), and SPONGENT-128-70-round (1687 GE). And the throughput of SHAT-128 is also 5.95 times higher than that of 4 hash algorithms in average. This results in having the FOM of SHAT-128 to be 9.66 times higher in average.

In Table 4, firstly, the throughput of SHAT-256 is 51.05% of that of Grostl; however, the area of SHAT-256 is only 21.84% of that of Grostl; this results in having 84.47

Table 3: Hardware implementation results of some 128-bit hash functions.

Hash function	Block size (bits)	Number of operations	Throughput at 100 kHz (kbps)	Area (GE)	FOM
SHAT-128	32	48	66.67	1605	258.80
H-Present-128 [15]	128	559	11.45	2330	21.09
	128	32	200	4256	110.41
MD4 [15]	512	456	112.28	7350	20.78
MD5 [15]	512	612	83.66	8400	11.86
ARMADILLO2-B [15]	64	256	250	4353	13.19
	64	64	1000	6025	27.55
U-QUARK [15]	8	544	1.47	1379	7.73
	8	68	11.76	2392	20.56
PHOTON-128 [15]	16	996	1.61	1122	12.78
	16	156	10.26	1708	35.15
SPONGENT-128 [15]	8	2380	0.34	1060	2.99
	16	70	11.43	1687	40.16

Table 4: Hardware implementation results of some 256-bit hash functions.

Hash function	Block size (bits)	Number of operations	Throughput at 100 kHz (kbps)	Area (GE)	FOM
SHAT-256	64	48	133.33	3193	130.78
SHA-256 [14]	512	490	104.48	8588	14.17
ARMADILLO2-E [14]	128	512	25	8653	3.34
	128	128	100	11914	7.05
BLAKE [14]	32	816	72.79	13575	0.21
Grostl [14]	64	196	261.14	14622	1.53
PHOTON-256 [14]	32	156	3.21	2177	6.78
	32	156	20.51	4362	10.17
SPONGENT-256 [14]	16	9520	0.17	1950	0.44
	16	140	11.43	3281	10.62

Table 5: Hardware implementation results of some 384-bit hash functions.

Hash function	Block size (bits)	Number of operations	Throughput at 100 kHz (kbps)	Area (GE)	FOM
SHAT-384	96	48	200	4753	88.53
SHA-384 [14]	1024	84	1219.04	43330	6.49

Table 6: Performance results of hash function using pipeline and parallelism.

Number of iteration rounds	Area		Delay		Power	
	(GE)	Increase (%)	(ns)	Reduction (%)	(μW)	Increase (%)
48	965	0.00	0.94	0.00	27.27	0.00
24	2010	4.15	1.87	2.60	79.05	0.91
16	3055	5.53	2.81	3.44	136.42	3.52
12	4100	6.22	3.74	4.35	193.71	4.67
8	6190	6.91	5.61	4.92	308.48	5.73
6	8280	7.25	7.47	5.32	423.18	6.21
4	12460	7.60	11.20	5.64	652.62	6.68
3	16640	7.77	14.93	5.74	882.00	6.90
2	25000	7.94	22.40	5.88	1340.80	7.12
1	50080	8.12	44.70	6.31	2695.40	7.40

Table 7: Performance results of unrolling steps constructions.

Number of iteration rounds	Area (GE)	Delay (ns)	Power (μW)	Throughput at 10 MHz (Mbps)
48	965	0.94	27.27	6.67
24	1930	1.92	78.34	13.33
16	2895	2.91	131.78	20.00
12	3860	3.91	185.06	26.67
8	5790	5.90	291.77	40.00
6	7720	7.89	398.42	53.33
4	11580	11.87	611.78	80.00
3	15440	15.84	825.06	106.67
2	23160	23.80	1251.70	160.00
1	46320	47.71	2509.70	320.00

times higher hardware efficiency of SHAT-256. Secondly, the throughput of SHAT-256 (3193 GE) is 412.91 times higher than that of 2 hash algorithms, such as PHOTON-256-156-round (2177 GE) and SPONGENT-256-9520-round (1950 GE), in average; although the area of SHAT-256 is larger, the FOM of SHAT-256 is still 158.25 times higher than that of 2 hash algorithms. Thirdly, comparing with SHA-256, ARMADILL02-E, BLAKE, PHOTON-256-156-round, and SPONGENT-256-140-round, the throughput of SHAT-256 is 4.65 times higher in average, and the area of SHAT-256 is only 49.15% of that of hash algorithms, in average. Therefore, the FOM of SHAT-256 is 119.14 times higher in average.

In Table 5, the throughput of SHA-384 is 6.09 times higher than that of SHAT-384; however, the area of SHA-384 is 9.11 times higher; this results in having the hardware efficiency of SHAT-384 to be 13.64 times higher than that of SHA-384.

Then we implement unfolding transformation technique with 10 different numbers of unrolling loops $(1, 2, \ldots, 48)$ by using 45 nm CMOS technology at 10 MHz to evaluate the performances of SHAT-128; the results are shown in Table 7. As we can see in Table 7, the throughput of PERM function can be achieved up to 47.97 times higher than original one which is 6.67 Mbps. However, area, delay, and power will increase dramatically as penalty.

Finally we implement pipeline and parallelism technique to reconstruct STEP block, as shown in Table 6; comparing with the performances of original circuit, the critical path delay reduces to 6.31% at most, while the power and area will increase in 8%.

4. Low Power Design for Hash Function

Low power design is a significant consideration in hardware implementation. How much the power consumption is will determine a device's life, reliability, and energy cost. Thus low power technique is applied normally to every application nowadays. There are many methods to reduce power consumption such as clock gating and power gating related to dynamic power and leakage power. Frequency decreasing technique will pull down the power dissipation dramatically as well.

Firstly, we will propose the frequency trade-off technique. By using this method we could achieve a range of frequency values for making a trade-off between low power consumption and high throughput of hash function. Secondly, we construct a hash encryption system which includes input data padding unit, RAM registers, main hash computing construction, message digest extraction component, and main control unit. Thirdly, by analyzing the idle mode and control signals of this hash encryption system, load-enable based clock gating scheme is applied to reduce the dynamic power consumption.

4.1. Frequency Trade-Off Technique. According to (1), reducing clock frequency is an effective method to decrease dynamic power dissipation linearly. In Section 2.2, we talked about the DVFS technique. By collecting the information about workload and temperature, DVFS will determine the sufficient clock frequency for the proper performance. However, modifying the clock frequency at RTL is not easy. Normally, we treat the clock frequency as constant. Also as we know, dynamic frequency scaling reduces the number of operations a system can issue in a given amount of time, thus reducing performance. Therefore, there is an issue we need to consider: high clock frequency brings high level throughput; however dramatically increased dynamic power consumption is the critical drawback. Low clock frequency minimizes the dynamic power dissipation; however it decreases the throughput as well.

However, according to the unfolding transformation technique which is introduced in Section 3.3, the maximum frequency of Perm function will decrease, while the number of unrolling loops increases. It means that we can decrease the clock frequency while increasing throughput of the hash algorithm. Thus, this unrolling transformation technique compromises high performance without high clock frequency. According to this advantage, by choosing proper clock frequency, we can make a trade-off between high performance and low power consumption.

Next, we explain how to get this scope of frequency value from the two performance bounds. For example, first we achieve two values of rolling Perm circuit: dynamic power consumption P_1 and clock frequency f_1 which is defined by the necessity of circuit design (the clock period computed from f_1 needs to be not less than the critical path delay). Then, according to (8), we can get the throughput T_1 at this frequency. Thus, those two performance bounds are defined in (13), where n is the number of iteration rounds in one Perm function with rolling STEPs:

$$P_{\max} = P_1 \cdot n, \qquad T_{\min} = T_1. \tag{13}$$

This method can be defined as the following: referring to the performance of original folding circuit (we assume that this circuit is the one with 48 iteration rounds in one

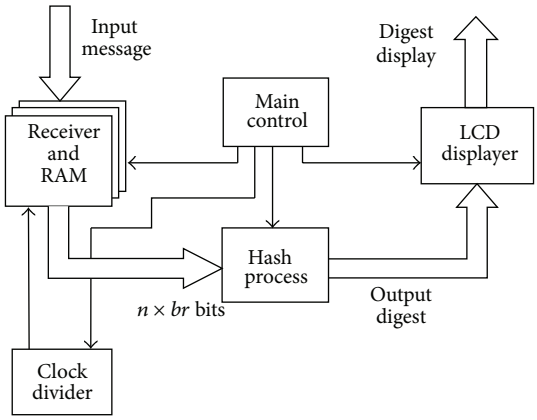


Figure 8: Hash encryption system.

Perm function), each unfolding transformation design with different numbers of unrolling STEPs $(2, 3, \ldots, 48)$ has two performance bounds: one is maximum dynamic power and the other is minimum throughput of the circuit. These two performance bounds are used to determine the boundary of proper frequency range for each unfolding transformation circuit. It means that when we choose one specific clock frequency in this value scope, the total dynamic power consumption of that PERM function will be not more than defined maximum dynamic power $P_{\max}$ and its throughput will be not less than that fixed minimum throughput $T_{\min}$.

This clock frequency scope gives us many different choices for different circuit designs by using unfolding transformation technique. The results of this frequency trade-off technique are shown in Table 9 in Section 4.4.

4.2. Hash Encryption System Design. The hash encryption system is divided into 5 main parts as shown in Figure 8.

Firstly, the receiver and RAM section is actually our padding unit. We use serial communication technique to connect PC and the hash encryption system. Thus, we need clock divider to generate proper clock cycle to be synchronous with Baud rate of serial communication. We choose 4800 Baud/s as our transmission Baud rate which is not a quick speed for low error rate (less than 3%). In this case, one Baud represents 1 bit. Our rule of transmission is a one start bit "0", then 8-bit message, and one finish bit "1". This start bit and finish bit will be added into the transmission message bits automatically; the sampling rate of receiver is 16 and FPGA board provides 100 MHz clock frequency. Thus, the clock period used in sampling is 1302 times provided 100 MHz clock period as shown in (14). This error is 0.0064% less than 3%:

$$\begin{aligned} \text{Sampling Clock Cycles} &= \frac{\text{Clock Frequency}}{\text{Baud rate} \cdot \text{Sampling Rate}} \\ &= \frac{100\,\text{MHz}}{16 \times 4800\,\text{B/s}} \\ &\approx 1302. \end{aligned} \tag{14}$$

Because the liquid crystal display (LCD) limits the number of characters we can display which are 32 characters in hexadecimal, this number is suitable for the number of digest bits of SHAT-128. Thus, our *br* for each padded block is determined to be 32 bits which consist of eight 4-bit hexadecimal numbers.

Secondly, hash function which we introduced in Section 3 is designed as sponge construction as shown in Figure 4. Absorbing n 32-bit message blocks, there are 128 bits digest that will be squeezed out.

Finally, the main control unit is designed for managing the working order between receiver, hash process, and LCD display. Figure 9 shows the pipeline working of system.

Because we use serial communication technique, the speed will be slow. We apply 4800 Baud/s as our Baud rate for low error rate; thus each 32-bit block needs roughly 7 ms. For example, there are seven 32-bit blocks that need to be transmitted; roughly 50 ms needs to be dissipated for data receiving and padding. Although the hash function that we used in this system is one STEP each round, this means that there are 48 iteration rounds for a complete Perm function. However, hash processing just needs roughly 6 μs. It also costs much time in LCD displaying period. Even though we can finish LCD initialization before we get hash digest, we still need roughly 1.5 ms to completely display all data.

4.3. Load-Enable Based Clock Gating. In this section, we introduce the load-enable based clock gating technique for the hash encryption system.

Clock gating is the most widely used low power technique at RTL. It is more reasonable to determine the toggle rate of gate output at RTL than any other three components, such as V_{DD}, clock frequency, and gate output capacitance. According to Figure 9, the hash encryption system is composed of a pipeline construction. Finishing signal of each process can be treated as enable signal in load-enable based clock gating as shown in Figure 3. On the other hand, XOR-based clock gating technique needs to specify the outputs of single level flip-flops which is not easily determined in our encryption system; thus the load-enable based clock gating is our best option for low power method.

As shown in Figure 10, there are three signal pairs to realize this load-enable based clock gating: *en_div* and *fsh_r*, *en_h* and *fsh_h*, and *en_lcd* and *fsh_lcd*. Because receiver is implemented in a specific clock frequency which is corresponding to the serial communication, the main control unit will not gate the clock signal of receiver directly; by controlling the clock signal of clock divider with *en_div*, receiver can be properly managed.

Figure 9 gives us three operation phases of the encryption system. In first phrase, *en_div* and *en_lcd* signals are asserted to logic one and *en_h* is asserted to logic zero; thus receiver starts receiving input messages and padding them into RAM. At the meantime, system will begin the initialization process for LCD displayer. However, the hash processing unit is waiting for the padded input message. Considering the serial communication takes long time due to the low Baud rate and its characteristic which is transmitting message bit one by

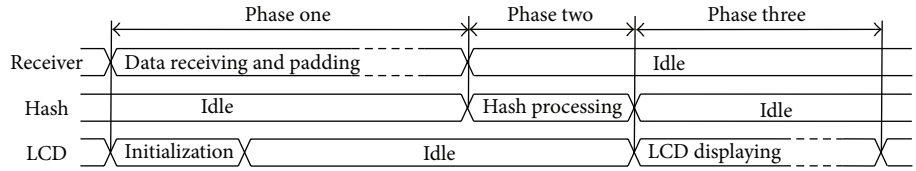


Figure 9: Three phases of hash encryption system.

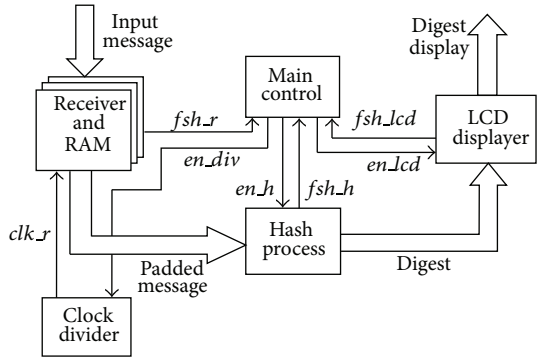


Figure 10: Control signals of hash encryption system.

Table 8: Hardware implementation with/without load-enable based clock gating.

System type	Area		Delay		Power	
	(GE)	Increase (%)	(ns)	Increase (%)	(μW)	Reduction (%)
Original	14053	n/a	1.63	n/a	1830.20	n/a
Clock gated	14565	3.64	1.72	5.52	1580.36	13.65

Table 9: Area and delay performances of frequency trade-off technique.

Number of iteration rounds	Area (GE)	Delay (ns)	Frequency (MHz)
48	965	0.94	10.00
24	1930	1.92	$5.00 < f_{24} < 6.96$
16	2895	2.91	$3.33 < f_{16} < 6.20$
12	3860	3.91	$2.50 < f_{12} < 5.89$
8	5790	5.90	$1.67 < f_{8} < 5.60$
6	7720	7.89	$1.25 < f_{6} < 5.47$
4	11580	11.87	$0.83 < f_{4} < 5.34$
3	15440	15.84	$0.63 < f_{3} < 5.28$
2	23160	23.80	$0.42 < f_{2} < 5.22$
1	46320	47.71	$0.21 < f_{1} < 5.21$

one, LCD displayer initialization can be finished before the padded message is ready. Thus, *en_lcd* can be asserted to logic zero by main control unit when *fsh_lcd* is switching to logic one.

During the second phase, because the padded message is ready, then *fsh_r* switches to logic one. Then *en_div* is asserted to zero which means that clock divider is turned off; then no specific clock frequency is produced; thus the receiver will stop working. In this phase, *en_h* is asserted to logic one for hash encryption which is our core function. *en_lcd* is still zero waiting for the hash digest generated by hash processing.

This system will enter the third phase when the *fsh_h* signal switches to logic one. In this phase, hash digest is ready; thus both receiver and hash processes are in idle mode which means that *en_div* and *en_h* are all asserted to logic zero. Signal *en_lcd* will be asserted to logic one to start LCD displaying. *en_lcd* will be asserted back to zero when the displaying process is finished. This is the end of the whole system; then the device will be turned off or repeats these three phases for another input message.

By analyzing the construction and process of hash encryption system, we can figure out the idle time for each component. Then applying the load-enable based clock gating to each component, the dynamic power dissipation of this system can be properly reduced as shown in Table 8 in Section 4.4.

4.4. Experimental Results. By using 10 MHz clock frequency and 45 nm CMOS technology, the results of frequency trade-off technique are shown in Tables 9, 10, and 11. Table 9 shows that the area and critical path delay are not changed comparing with the unfolding transformation technique. Tables 10 and 11 give us the variation of dynamic power consumption and throughput with frequency trade-off method. Note that f_i stands for frequency, T_i stands for throughput, and $T_{i\text{pct}}$ is the percentage of increasing comparing with the minimum throughput ($T_{\min}$) which is 6.67 Mbps. P_i means the total dynamic power consumption by finishing a complete Perm function and $P_{i\text{pct}}$ is the percentage of power reduction comparing with the maximum power consumption ($P_{\max}$) defined as 1308.96 μW which is calculated from the product of 48 (number of iteration rounds) and 27.27 μW (as shown in Table 7). Note that *i* stands for the number of iteration rounds.

Then we apply load-enable based clock gating scheme to hash encryption system by using 100 MHz clock frequency, which can be provided on FPGA board, and 45 nm CMOS technology. As shown in Table 8, the dynamic power decreases 13.65%. However, 3.64% increased area and 5.52% increased critical path delay are sacrificed.

Table 10: Dynamic power consumption of frequency trade-off technique.

Number of iteration rounds	Power (μW)	Power Reduction (%)	Frequency (MHz)
48	1308.96	n/a	10.00
24	$940.08 < P_{24} < 1308.48$	$28.18 < P_{24pct} < 0.04$	$5.00 < f_{24} < 6.96$
16	$702.88 < P_{16} < 1307.20$	$46.30 < P_{16pct} < 0.13$	$3.33 < f_{16} < 6.20$
12	$555.12 < P_{12} < 1308.00$	$57.59 < P_{12pct} < 0.07$	$2.50 < f_{12} < 5.89$
8	$389.04 < P_{8} < 1307.12$	$70.28 < P_{8pct} < 0.14$	$1.67 < f_{8} < 5.60$
6	$298.80 < P_{6} < 1307.64$	$77.17 < P_{6pct} < 0.10$	$1.25 < f_{6} < 5.47$
4	$203.92 < P_{4} < 1306.76$	$84.42 < P_{4pct} < 0.17$	$0.83 < f_{4} < 5.34$
3	$154.71 < P_{3} < 1306.89$	$88.18 < P_{3pct} < 0.16$	$0.63 < f_{3} < 5.28$
2	$104.30 < P_{2} < 1306.76$	$92.03 < P_{2pct} < 0.17$	$0.42 < f_{2} < 5.22$
1	$52.29 < P_{1} < 1307.60$	$96.01 < P_{1pct} < 0.10$	$0.21 < f_{1} < 5.21$

Table 11: Throughput performances of frequency trade-off technique.

Number of iteration rounds	Throughput (Mbps)	Throughput Improvement (%)	Frequency (MHz)
48	6.67	n/a	10.00
24	$6.67 < T_{24} < 9.28$	$0.00 < T_{24pct} < 39.13$	$5.00 < f_{24} < 6.96$
16	$6.67 < T_{16} < 12.4$	$0.00 < T_{16pct} < 85.91$	$3.33 < f_{16} < 6.20$
12	$6.67 < T_{12} < 15.71$	$0.00 < T_{12pct} < 135.53$	$2.50 < f_{12} < 5.89$
8	$6.67 < T_{8} < 22.40$	$0.00 < T_{8pct} < 235.83$	$1.67 < f_{8} < 5.60$
6	$6.67 < T_{6} < 29.17$	$0.00 < T_{6pct} < 337.33$	$1.25 < f_{6} < 5.47$
4	$6.67 < T_{4} < 42.72$	$0.00 < T_{4pct} < 540.48$	$0.83 < f_{4} < 5.34$
3	$6.67 < T_{3} < 56.32$	$0.00 < T_{3pct} < 744.38$	$0.63 < f_{3} < 5.28$
2	$6.67 < T_{2} < 83.52$	$0.00 < T_{2pct} < 1152.17$	$0.42 < f_{2} < 5.22$
1	$6.67 < T_{1} < 166.72$	$0.00 < T_{1pct} < 2399.55$	$0.21 < f_{1} < 5.21$

5. Conclusion

In order to achieve high performance and low power hardware implementation for cryptographic hash function which uses sponge construction, firstly, we use unfolding transformation technique to improve the throughput of hash function; secondly, pipeline and parallelism design techniques are implemented to reduce the critical path delay by modifying the structure of permutation function; thirdly, frequency trade-off technique is proposed to calculate a frequency scope which can be used to make a trade-off between low dynamic power consumption and high throughput of hash function; finally, load-enable based clock gating scheme is applied in hash encryption system to eliminate wasted toggle rate of signals in the idle mode.

The experimental results have shown that unfolding transformation technique can achieve up to 47.97 times higher throughput, pipeline and parallelism methods give 6.31% delay reduction, load-enable based clock gating scheme decreases 13.65% dynamic power consumption, and frequency trade-off technique shows how to decide the clock frequency of the hash function to achieve low power consumption and high throughput.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2012-H0301-12-3007) supervised by the NIPA (National IT Industry Promotion Agency).

References

[1] H. Michail and C. Goutis, "Holistic methodology for designing ultra high-speed SHA-1 hashing cryptographic module in hardware," in *Proceedings of the IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC '08)*, pp. 1–4, Hong Kong, December 2008.

[2] "Cryptographic hash algorithm competition," NIST Computer Security Resource Center, http://csrc.nist.gov/groups/ST/hash/sha-3/index.html.

[3] B. Schneier, *Applied Cryptography: Protocols, Algorithms, and Source Code in C*, John Wiley & Sons, New York, NY, USA, 2nd edition, 1996.

[4] J. Nakajima and M. Mitsuru, "Performance analysis and parallel implementation of dedicated hash function," in *Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT '02)*, vol. 2332, pp. 165–180, Amsterdam, The Netherlands, 2002.

[5] P. C. van Oorschot, A. Somayaji, and G. Wurster, "Hardware-assisted circumvention of self-hashing software tamper resistance," *IEEE Transactions on Dependable and Secure Computing*, vol. 2, no. 2, pp. 82–92, 2005.

[6] G. Bertoni, J. Daemen, M. Peeters, and G. van Assche, "Cryptog-raphic sponge functions," The Sponge Functions Corner, http://sponge.noekeon.org/index.html.

[7] "Sponge function," WIKIPEDIA, http://en.wikipedia.org/wiki/Sponge_function.

[8] L. Li, *Power optimization from register transfer level to transistor level in deeply scaled CMOS technology [Ph.D. thesis]*, Illinois Institute of Technology, Chicago, Ill, USA, 2012.

[9] N. Weste and D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*, Addison-Wesley, Reading, Mass, USA, 2010.

[10] Y. Zhang, Q. Tong, L. Li et al., "Automatic register transfer level CAD tool design for advanced clock gating and low power schemes," in *Proceeding of the International SoC Design Conference (ISOCC '12)*, pp. 21–24, Jeju Island, Republic of Korea, 2012.

[11] K. Aoki, T. Ichikawa, and M. Kanda, "Specification of *Camellia*—a 128-bit block cipher," Nippon Telegraphy and Telephone Corporation, Mitsubishi Electric Corporation, 2000.

[12] Y. K. Lee, H. Chan, and I. Verbauwhede, "Throughput optimized SHA-1 architecture using unfolding transformation," in *Proceedings of the 17th IEEE International Conference on Application-Specific Systems, Architectures and Processors (ASAP '06)*, pp. 354–359, Steamboat Springs, Colo, USA, September 2006.

[13] H. Michail, A. P. Kakarountas, O. Koufopavlou, and C. E. Goutis, "A low-power and high-throughput implementation of the SHA-1 hash function," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '05)*, vol. 4, pp. 4086–4089, Kobe, Japan, May 2005.

[14] S. Badel, N. Dağtekin, J. Nakahara Jr. et al., "ARMADILLO: a multi-purpose cryptographic primitive dedicated to hardware," in *Cryptographic Hardware and Embedded Systems, CHES 2010*, vol. 6225 of *Lecture Notes in Computer Science*, pp. 398–412, 2010.

[15] K. Lin, Y. Zhang, K. Choi, J. Kang, and S. Hong, "Multi-purpose bimodal cryptographic algorithm and its hardware implementation," in *Proceedings of the FTRA International Conference on Advanced IT, Engineering and Management (FTRA AIM '13)*, Seoul, Korea, 2013.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 389476, 10 pages
http://dx.doi.org/10.1155/2014/389476

Research Article

Related-Key Cryptanalysis on the Full PRINTcipher Suitable for IC-Printing

Yuseop Lee,[1] Kitae Jeong,[1] Changhoon Lee,[2] Jaechul Sung,[3] and Seokhie Hong[1]

[1] *Center for Information Security Technologies (CIST), Korea University, Seoul 136-701, Republic of Korea*
[2] *Department of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul 139-743, Republic of Korea*
[3] *Department of Mathematics, University of Seoul, Seoul 130-743, Republic of Korea*

Correspondence should be addressed to Changhoon Lee; chlee@seoultech.ac.kr

Received 28 August 2013; Accepted 22 October 2013; Published 16 January 2014

Academic Editor: Jongsung Kim

PRINTcipher-48/96 are 48/96-bit block ciphers proposed in CHES 2010 which support the 80/160-bit secret keys, respectively. In this paper, we propose related-key cryptanalysis of PRINTcipher. To recover the 80-bit secret key of PRINTcipher-48, our attack requires 2^{47} related-key chosen plaintexts with a computational complexity of $2^{60.62}$. In the case of PRINTcipher-96, we require 2^{95} related-key chosen plaintexts with a computational complexity of 2^{107}. These results are the first known related-key cryptanalytic results on them.

1. Introduction

Recently, the security of constrained hardware environments such as RFID tags and sensor nodes is major topic in cryptography [1, 2]. The research on lightweight block ciphers suitable for the efficient implementation in constrained hardware environments such as RFID tags and sensor nodes has been studied. As a result, PRESENT [3], LED [4], HIGHT [5], PRINTcipher [6], and Piccolo [7] were proposed.

PRINTcipher is a 48/96-bit block cipher proposed in CHES 2010 thst supports the 80/160-bit secret keys. According to the block size, they are denoted by PRINTcipher-48 and PRINTcipher-96, respectively. The attractive properties of PRINTcipher are that all rounds use the same round key and differ only by a round counter and that the linear layer is partially key-dependent. Because of these properties, most known cryptanalytic results on PRINTcipher are based on weak keys (see Table 1). The best attack results on PRINTcipher are invariance subspace attacks on the full PRINTcipher-48/96 [8]. In detail, the attack on the full PRINTcipher-48 is applicable to 2^{52} weak keys and requires 5 chosen plaintexts with a negligible computational complexity. In the case of the full PRINTcipher-96, it is applicable to 2^{102} weak keys and requires 5 chosen plaintexts with a negligible computational complexity.

In this paper, we find weakness of PRINTcipher-48/96 on related-key attacks. To construct related-key differential characteristics, we focus on related keys that have different values in the part related to a key-dependent permutation. Thus, we can construct t-round related-key differential characteristics with a probability of 2^{-t}. By using these characteristics, we can recover the secret keys of PRINTcipher-48/96. Our results are summarized in Table 1. In detail, to recover the 80-bit secret key of PRINTcipher-48, our attack requires 4 related keys, 2^{47} related-key chosen plaintexts, and a computational complexity of $2^{60.62}$. In the case of PRINTcipher-96, we require 4 related keys, 2^{95} related-key chosen plaintexts, and a computational complexity of 2^{107}. These results are the first known related-key cryptanalytic results on them.

This paper is organized as follows. In Section 2, we describe briefly the structure of PRINTcipher. In Section 3, we explain how to construct related-key differential characteristics on PRINTcipher. Related-key attacks on PRINTcipher-48 and PRINTcipher-96 are proposed in Sections 4 and 5, respectively. Finally, we give our conclusion in Section 6.

Table 1: Summary of cryptanalytic results on PRINTcipher.

Target algorithm	Attack method	Attack rounds	Number of weak keys	Complexity		Reference
				Data	Computation	
PRINTcipher-48	DC	22	·	2^{48}	2^{48}	[9]
	LC *	29	2^{75}	2^{48}	2^{73}	[10]
	CDLC *	31	$2^{68.56}$	$2^{46.92}$	$2^{25.92}$	[11]
	ISA *	Full (48)	2^{52}	5	Negligible	[8]
	RKDC	Full (48)	·	2^{47}	$2^{60.62}$	This paper
PRINTcipher-96	ISA *	Full (96)	2^{102}	5	Negligible	[8]
	RKDC	Full (96)	·	2^{95}	2^{107}	This paper

*Attack results based on weak keys.
CDLC: combined differential and linear cryptanalysis.
ISA: invariant subspace attack.
RKDC: related-key differential cryptanalysis.

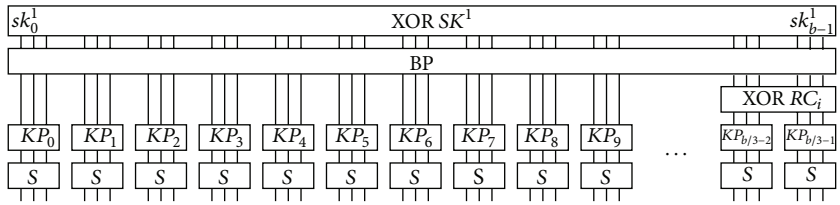


Figure 1: A single round of PRINTcipher-b.

2. Description of PRINTcipher

PRINTcipher-48/96 are 48/96-bit block ciphers supporting the 80/160-bit secret keys. Note that PRINTcipher uses the same round key (SK^1, SK^2) for all rounds. In detail, the secret key K is divided into (SK^1, SK^2). SK^1 is used to XORing with an intermediate value, and SK^2 controls a key-dependent permutation.

Figure 1 presents a single round of PRINTcipher-b, where $b \in \{48, 96\}$. The encryption process of PRINTcipher-b is as follows. Here, the number of rounds is b. A b-bit round key $SK^1 = (sk_0^1, \ldots, sk_{b-1}^1)$ and a $(2b/3)$-bit round key $SK^2 = (SK_0^2, \ldots, SK_{b/3-1}^2)$; here, sk_0^1 is the most significant bit of SK^1.

(1) A b-bit plaintext $P = (p_0, p_1, \ldots, p_{b-1})$ is loaded to a b-bit intermediate value $X = (x_0, x_1, \ldots, x_{b-1})$.

(2) For $i = 1, \ldots, b$, do the following steps.

(a) X is XORed with a b-bit round key SK^1. Consider

$$x_j \longleftarrow x_j \bigoplus sk_j^1 \quad (j = 0, 1, \ldots, b-1). \tag{1}$$

(b) Consider $X \leftarrow BP(X)$, where BP is a bit permutation.

(c) Consider $X \leftarrow X \oplus RC_i$, where RC_i is a round constant.

(d) For $l = 0, \ldots, b/3 - 1$, $(x_{3l}, x_{3l+1}, x_{3l+2}) \leftarrow KP_l((x_{3l}, x_{3l+1}, x_{3l+2}))$, where KP_l is the lth key-dependent permutation based on a 2-bit $SK_l^2 = (sk_{l,0}^2, sk_{l,1}^2)$ (see Figure 1).

(e) Consider that X is mixed by using $b/3$ identical 3×3 S-boxes.

(3) $C = (c_0, \ldots, c_{b-1}) \leftarrow X$.

With a bit permutation BP, a value of the bit position i is moved to the bit position j, where

$$j = \begin{cases} 3i \bmod (b-1), & \text{for } 0 \le i \le b-2, \\ b-1, & \text{for } i = b-1. \end{cases} \tag{2}$$

In a key-dependent permutation KP, an intermediate value X is arranged in $b/3$ blocks of 3 bits each, which are permuted individually according to a 2-bit SK_l^2. Out of 6 possible permutations on 3 bits, only four permutations are valid for PRINTcipher. In detail, as shown in Table 2, a 3-bit input value (y_0, y_1, y_2) is permuted to the corresponding output value according to a 2-bit $(sk_{l,0}^2, sk_{l,1}^2)$. Here, KP_l^{mn} means KP_l in the case that $(sk_{l,0}^2, sk_{l,1}^2) = (m, n)$.

3. Construction of Related-Key Differential Characteristics on PRINTcipher

In this section, we present how to construct t-round related-key differential characteristics on PRINTcipher-48/96 by using properties of a key-dependent permutation KP and S-box.

3.1. Related-Key Properties on Key-Dependent Permutation and S-Box. We consider a 3-bit input value (y_0, y_1, y_2) of

Table 2: Key-dependent permutation KP_l.

(a)

Notation	$(sk^2_{l,0}, sk^2_{l,1})$	$KP_l(y_0, y_1, y_2)$
KP_l^{00}	$(0,0)$	(y_0, y_1, y_2)
KP_l^{01}	$(0,1)$	(y_1, y_0, y_2)
KP_l^{10}	$(1,0)$	(y_0, y_2, y_1)
KP_l^{11}	$(1,1)$	(y_2, y_1, y_0)

(b)

x	0	1	2	3	4	5	6	7
$S(x)$	0	1	3	6	7	4	5	2

a key-dependent permutation KP_l ($l = 0, \ldots, 15$). If a 2-bit round key $(sk^2_{l,0}, sk^2_{l,1})$ is equal to $(0,0)$ or $(0,1)$, from Table 2, the corresponding output value is computed as follows:

(i) $(0,0)$: $(y_0, y_1, y_2) \xrightarrow{KP_l^{00}} (y_0, y_1, y_2)$;

(ii) $(0,1)$: $(y_0, y_1, y_2) \xrightarrow{KP_l^{01}} (y_1, y_0, y_2)$.

In the above relations, if y_0 is equal to y_1, each permutation outputs the same value and vice versa. That is, the following equation holds:

$$y_0 = y_1 \Longleftrightarrow KP_l^{00}\left((y_0, y_1, y_2)\right) = KP_l^{01}\left((y_0, y_1, y_2)\right). \quad (3)$$

In total, we can obtain the following six properties of KP.

Property 1. Consider $y_0 = y_1 \Leftrightarrow KP_l^{00}((y_0, y_1, y_2)) = KP_l^{01}((y_0, y_1, y_2))$.

Property 2. Consider $y_1 = y_2 \Leftrightarrow KP_l^{00}((y_0, y_1, y_2)) = KP_l^{10}((y_0, y_1, y_2))$.

Property 3. Consider $y_0 = y_2 \Leftrightarrow KP_l^{00}((y_0, y_1, y_2)) = KP_l^{11}((y_0, y_1, y_2))$.

Property 4. Consider $y_0 = y_1 = y_2 \Leftrightarrow KP_l^{01}((y_0, y_1, y_2)) = KP_l^{10}((y_0, y_1, y_2))$.

Property 5. Consider $y_0 = y_1 = y_2 \Leftrightarrow KP_l^{01}((y_0, y_1, y_2)) = KP_l^{11}((y_0, y_1, y_2))$.

Property 6. Consider $y_0 = y_1 = y_2 \Leftrightarrow KP_l^{10}((y_0, y_1, y_2)) = KP_l^{11}((y_0, y_1, y_2))$.

Furthermore, from the definition of S-box S, we can obtain the following property.

Property 7. If KP_l^{00} and KP_l^{01} have the same input value Y, the output difference of S-box, $S((KP_l^{00}(Y))) \oplus S(KP_l^{01}(Y))$, should be included in a set $\{0, 2, 4\}$.

3.2. Related-Key Differential Characteristics on PRINTcipher-48. Among the above seven properties, we focus on Properties 1, 2 and 3. To apply these properties on the proposed attack, we first consider the following related-key pairs ($K = (SK^1, SK^2), K^* = (SK^{1*}, SK^{2*})$). Here, $l = 0, \ldots, 15$.

Case 1 (l). Consider the following:

(i) $SK^1 = SK^{1*}$;

(ii) $SK_l^2 = (0,0), SK_l^{2*} = (0,1)$;

(iii) $SK_i^2 = SK_i^{2*}$ where $i \neq l$,

Case 2 (l). Consider the following:

(i) $SK^1 = SK^{1*}$;

(ii) $SK_l^2 = (0,0), SK_l^{2*} = (1,0)$;

(iii) $SK_i^2 = SK_i^{2*}$, where $i \neq l$.

Case 3 (l). Consider the following:

(i) $SK^1 = SK^{1*}$;

(ii) $SK_l^2 = (0,0), SK_l^{2*} = (1,1)$;

(iii) $SK_i^2 = SK_i^{2*}$, where $i \neq l$.

We assume that the input difference of the target round is zero. If a related-key pair (K, K^*) satisfies Case 1 (0), KP_0 has a nonzero related-key difference. Here, from Property 1, it can be easily shown that the output difference of KP_0 is zero with a probability of 2^{-1} (i.e., the probability that y_0 is equal to y_1). Since the output differences of other KP_l's except KP_0 are zero, as shown in Figure 2, we can construct a 1-round related-key differential characteristic $0 \xrightarrow{\text{Case 1 (0)}} 0$ with a probability of 2^{-1} under Case 1 (0). Since PRINTcipher-48 uses the same round key for all rounds, we can easily extend this result to a t-round related-key differential characteristic. That is, we can construct a t-round related-key differential characteristic $0 \xrightarrow{\text{Case 1 (0)}} 0$ with a probability of 2^{-t}. The other cases of Case 1 (l) are explained in a similar fashion. Moreover, in Case 2 (l) and Case 3 (l), we can construct t-round related-key differential characteristics $0 \xrightarrow{\text{Case 2 } (l)} 0$ and $0 \xrightarrow{\text{Case 3 } (l)} 0$ with a probability of 2^{-t}, respectively.

Note that we cannot control the exact values of the related-key pair (K, K^*) under a related-key attack scenario. In other words, we cannot apply the above related-key differential characteristics to our attack directly. To solve this problem, for each KP_l, we consider the following four related keys simultaneously. Here, K means the right secret key of PRINTcipher-48.

(i) Consider $K_l^{(0,0)} = [SK^1, (SK_0^2, SK_1^2, \ldots, SK_l^2 \oplus (0,0), \ldots, SK_{15}^2)] = K$.

(ii) Consider $K_l^{(0,1)} = [SK^1, (SK_0^2, SK_1^2, \ldots, SK_l^2 \oplus (0,1), \ldots, SK_{15}^2)]$.

(iii) Consider $K_l^{(1,0)} = [SK^1, (SK_0^2, SK_1^2, \ldots, SK_l^2 \oplus (1,0), \ldots, SK_{15}^2)]$.

(iv) Consider $K_l^{(1,1)} = [SK^1, (SK_0^2, SK_1^2, \ldots, SK_l^2 \oplus (1,1), \ldots, SK_{15}^2)]$.

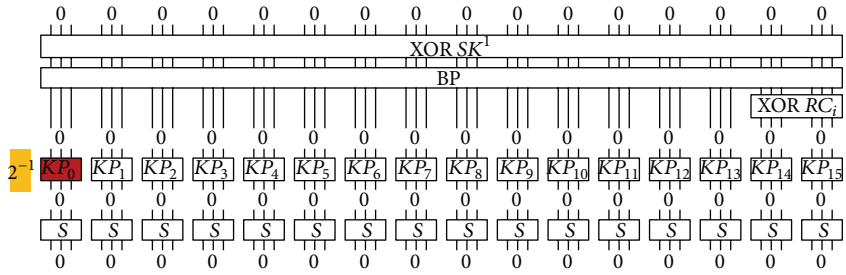


Figure 2: 1-round related-key differential characteristic under Case 1 (0).

From these four related keys, we can combine six related-key pairs. For each value of SK_l^2, a related-key pair is satisfied one among three cases, Case 1 (l), Case 2 (l), and Case 3 (l) (see Table 3). For example, we assume that SK_l^2 is equal to $(1,0)$. Then, four related keys are computed as follows.

(i) Consider $K_l^{(0,0)} = [SK^1, (SK_0^2, SK_1^2, \ldots, ((1,0) \oplus (0,0) = (1,0)), \ldots, SK_{15}^2)]$.

(ii) Consider $K_l^{(0,1)} = [SK^1, (SK_0^2, SK_1^2, \ldots, ((1,0) \oplus (0,1) = (1,1)), \ldots, SK_{15}^2)]$.

(iii) Consider $K_l^{(1,0)} = [SK^1, (SK_0^2, SK_1^2, \ldots, ((1,0) \oplus (1,0) = (0,0)), \ldots, SK_{15}^2)]$.

(iv) Consider $K_l^{(1,1)} = [SK^1, (SK_0^2, SK_1^2, \ldots, ((1,0) \oplus (1,1) = (0,1)), \ldots, SK_{15}^2)]$.

Recall that the condition of Case 1 (l) is that $SK_l^2 = (0,0)$ and $SK_l^{2*} = (0,1)$. Thus, $(K_l^{(1,0)}, K_l^{(1,1)})$ satisfies this condition. Similarly, the condition of Case 2 (l) is that $SK_l^2 = (0,0)$ and $SK_l^{2*} = (1,0)$. Thus, $(K_l^{(1,0)}, K_l^{(0,0)})$ satisfies this condition.

If SK_l^2 is equal to $(0,0)$, three related-key pairs $(K_l^{(0,0)}, K_l^{(0,1)})$, $(K_l^{(0,0)}, K_l^{(1,0)})$, and $(K_l^{(0,0)}, K_l^{(1,1)})$ satisfy Case 1 (l), Case 2 (l), and Case 3 (l), respectively (see Table 3). It means that t-round related-key differential characteristics $0 \xrightarrow{\text{Case 1 } (l)} 0$, $0 \xrightarrow{\text{Case 2 } (l)} 0$, and $0 \xrightarrow{\text{Case 3 } (l)} 0$ are satisfied with a probability of 2^{-t}, respectively. However, if SK_l^2 is not equal to $(0,0)$, only one of related-key pairs satisfies the corresponding condition. For example, it is assumed that the right SK_l^2 is $(1,0)$. Only one key pair, $(K_l^{(0,0)}, K_l^{(1,0)})$, among three related-key pairs is satisfied Case 2 (l) from Table 3. It means that the corresponding differential characteristic $0 \xrightarrow{\text{Case 2 } (l)} 0$ is satisfied with a probability of 2^{-t} and that the other differential characteristics $0 \xrightarrow{\text{Case 1 } (l)} 0$ and $0 \xrightarrow{\text{Case 3 } (l)} 0$ are satisfied with a random probability ($= 2^{-2t}$), respectively.

3.3. Related-Key Differential Characteristics on PRINTcipher-96. In the case of PRINTcipher-96, we can construct $96 (= 3 \cdot 32)$ t-round related-key differential characteristics with a probability 2^{-t} by using the similar method on PRINTcipher-48. In detail, for each KP_l, we can obtain the same six properties and three cases, Case 1 (l), Case 2 (l), and Case 3 (l). Thus, we get three t-round related-key differential characteristics for each KP_l ($l = 0, \ldots, 31$).

Table 3: The corresponding related-key pair according to the value of SK_l^2.

SK_l^2	Case 1 (l)	Case 2 (l)	Case 3 (l)
$(0,0)$	$(K_l^{(0,0)}, K_l^{(0,1)})$	$(K_l^{(0,0)}, K_l^{(1,0)})$	$(K_l^{(0,0)}, K_l^{(1,1)})$
$(0,1)$	$(K_l^{(0,1)}, K_l^{(0,0)})$	$(K_l^{(0,1)}, K_l^{(1,1)})$	$(K_l^{(0,1)}, K_l^{(1,0)})$
$(1,0)$	$(K_l^{(1,0)}, K_l^{(1,1)})$	$(K_l^{(1,0)}, K_l^{(0,0)})$	$(K_l^{(1,0)}, K_l^{(0,1)})$
$(1,1)$	$(K_l^{(1,1)}, K_l^{(1,0)})$	$(K_l^{(1,1)}, K_l^{(0,1)})$	$(K_l^{(1,1)}, K_l^{(0,0)})$

4. Related-Key Cryptanalysis on PRINTcipher-48

We are ready to propose related-key cryptanalysis on PRINTcipher-48. Recall that we can construct t-round related-key differential characteristics on PRINTcipher-48 with a probability of 2^{-t}. These related-key differential characteristics depend on the concrete key value. However, the attacker did not control the exact key value under a related-key attack scenario.

To solve this, we use 4 related keys $K_0^{(0,0)}, K_0^{(0,1)}, K_0^{(1,0)}$, and $K_0^{(1,1)}$. In detail, two key pairs $(K_0^{(0,0)}, K_0^{(0,1)})$ and $(K_0^{(1,0)}, K_0^{(1,1)})$ are considered. Then, among two key pairs, only one should satisfy Case 1(0). Thus we first apply the related-key pair $(K_0^{(0,0)}, K_0^{(0,1)})$ to our attack. After that, we repeat the same procedure using the other related-key pair $(K_0^{(1,0)}, K_0^{(1,1)})$. For the convenience of description, we assume that the key pair $(K_0^{(0,0)}, K_0^{(0,1)})$ satisfies Case 1(0).

4.1. Basic Related-Key Attack on PRINTcipher-48. To attack the full PRINTcipher-48, we consider 44-round related-key differential characteristic $0 \xrightarrow{\text{Case 1 } (0)} 0$ (from round 2 to round 45) with a probability of 2^{-44}.

Our attack procedure mainly consists of the following steps. First, we consider plaintext structures which are composed of 4 plaintexts each. Second, we discard the wrong ciphertext pairs from the difference between ciphertexts. Finally, we guess the partial secret key and determine related-key pairs satisfying Case 1 (0).

4.1.1. Collection of Ciphertexts. First, we consider the following plaintext structures $\mathcal{S}^{x_0,x_1}$ which are composed of 4 plaintexts each (see Figure 3).

(i) Consider $\mathcal{S}^{x_0,x_1} = \{s_{i,j}^{x_0,x_1}\}$, where

(a) $s_{i,j}^{x_0,x_1} = (i \| x_0 \| j \| x_1)$;
(b) $i, j \in \{0, 1\}$;
(c) x_0: arbitrary 15-bit value;
(d) x_1: arbitrary 31-bit value.

Among all possible sixteen plaintext pairs for each plaintext structure, we consider only the following 8 plaintext pairs:

$$\begin{aligned} &\left(s_{0,0}^{x_0,x_1}, s_{0,0}^{x_0,x_1}\right), \quad \left(s_{0,1}^{x_0,x_1}, s_{0,1}^{x_0,x_1}\right), \\ &\left(s_{1,0}^{x_0,x_1}, s_{1,0}^{x_0,x_1}\right), \quad \left(s_{1,1}^{x_0,x_1}, s_{1,1}^{x_0,x_1}\right), \\ &\left(s_{0,0}^{x_0,x_1}, s_{1,1}^{x_0,x_1}\right), \quad \left(s_{0,1}^{x_0,x_1}, s_{1,0}^{x_0,x_1}\right), \\ &\left(s_{1,0}^{x_0,x_1}, s_{0,1}^{x_0,x_1}\right), \quad \left(s_{1,1}^{x_0,x_1}, s_{0,0}^{x_0,x_1}\right). \end{aligned} \tag{4}$$

Recall that $(K_0^{(0,0)}, K_0^{(0,1)})$ is assumed to satisfy Case 1 (0). Thus, for only four plaintext pairs in each plaintext structure, the input difference of round 2 is zero. Here, when we use 2^{44} plaintext structures, the expected number of right pairs is $4(= 4 \cdot 2^{44} \cdot 2^{-44})$. Note that our related-key differential characteristics hold with a probability of 2^{-44}.

4.1.2. Filtering the Wrong Pairs. We discard the wrong ciphertext pairs by checking the difference between ciphertexts. For the right ciphertext pair, the output difference of round 45 should be zero as shown in Figure 3.

In round 46, from Property 7, the possible output difference of the first S-box is 0 or 2 or 4. The differential propagation in round 47~48 is shown in Figure 3. Then, we discard the wrong ciphertext pairs by checking the following three checkPoints.

(i) $CheckPoint_1$. The difference between the rightmost 30 bits of ciphertext is zero.

(ii) $CheckPoint_2$. The output difference of $KP_1, KP_2, \ldots, KP_5$ in round 48 should be included in a set $\{0, 1, 2, 4\}$.

(iii) $CheckPoint_3$. The input difference of KP_0 should be included in a set $\{0, 4\}$.

Since the filtering probability of this step is $2^{-37}(= 2^{-30} \cdot 2^{-1.5} \cdot 2^{-2})$, $2^{10}(= 8 \cdot 2^{44} \cdot 2^{-37})$ ciphertext pairs are survived.

4.1.3. Recovery of the Secret Key of PRINTcipher-48. For each ciphertext pair passing the above steps, we guess the following 16-bit key:

$$\left(SK_1^2, SK_2^2, SK_3^2, SK_4^2, SK_5^2, sk_0^1, sk_1^1, sk_2^1, sk_3^1, sk_4^1, sk_5^1\right). \tag{5}$$

Then we recover 16-bit key by checking the following checkPoints (see Figure 3).

(i) $CheckPoint_4$. Input differences of KP_1, KP_2, KP_3, KP_4, and KP_5 in round 48 should be 0 or 4.

(ii) $CheckPoint_5$. Input differences of KP_0 and KP_1 in round 47 should be 0 or 4.

(iii) $CheckPoint_6$. Output difference of the first S-box in round 46 should be 0 or 2 or 4.

First, we check $CheckPoint_4$. In this step, we guess a 10-bit key $(SK_1^2, SK_2^2, SK_3^2, SK_4^2, SK_5^2)$. Since the filtering probability of this step is 2^{-5}, $2^5(= 2^{10} \cdot 2^{-5})$ ciphertext pairs remained for each guessed key. Then, we guess an additional 6-bit key $(sk_0^1, sk_1^1, sk_2^1, sk_3^1, sk_4^1, sk_5^1)$ and check $CheckPoint_5$ and $CheckPoint_6$. Since the filtering probabilities of this step are 2^{-4} ($CheckPoint_5$) and 0.75 ($CheckPoint_6$), the expected number of the remaining ciphertext pairs is 1.5 for each guessed key. Finally, we determine the guessed key with the maximal number of remaining ciphertext pairs as the right key.

Until now, we introduced the attack procedure with a related-key pair $(K_0^{(0,0)}, K_0^{(0,1)})$. In case of the related-key pair $(K_0^{(1,0)}, K_0^{(1,1)})$, the attack procedure can be explained in a similar fashion. Our attack procedure on the full PRINTcipher-48 is summarized as follows.

(1) Select 2^{44} plaintext structures which are composed of 4 plaintexts each and obtain the corresponding ciphertexts under four related keys $K_0^{(0,0)}$, $K_0^{(0,1)}$, $K_0^{(1,0)}$, and $K_0^{(1,1)}$, respectively.

(2) Considering the related-key pair $(K_0^{(0,0)}, K_0^{(0,1)})$, we have the following.

(a) Discard wrong pairs which do not satisfy $CheckPoint_1$, $CheckPoint_2$, and $CheckPoint_3$.
(b) Guess 16-bit key $(SK_1^2, SK_2^2, SK_3^2, SK_4^2, SK_5^2, sk_0^1, sk_1^1, sk_2^1, sk_3^1, sk_4^1, sk_5^1)$ and count the ciphertext pairs satisfying $CheckPoint_4$, $CheckPoint_5$, and $CheckPoint_6$.

(3) Considering the related-key pair $(K_0^{(1,0)}, K_0^{(1,1)})$, we have the following.

(a) Discard wrong pairs which do not satisfy $CheckPoint_1$, $CheckPoint_2$, and $CheckPoint_3$.
(b) Guess 16-bit key $(SK_1^2, SK_2^2, SK_3^2, SK_4^2, SK_5^2, sk_0^1, sk_1^1, sk_2^1, sk_3^1, sk_4^1, sk_5^1)$ and count the ciphertext pairs satisfying $CheckPoint_4$, $CheckPoint_5$, and $CheckPoint_6$.

(4) Output the guessed key with the maximal count as the right key.

(5) Do an exhaustive search for the remaining secret key by using trial encryption.

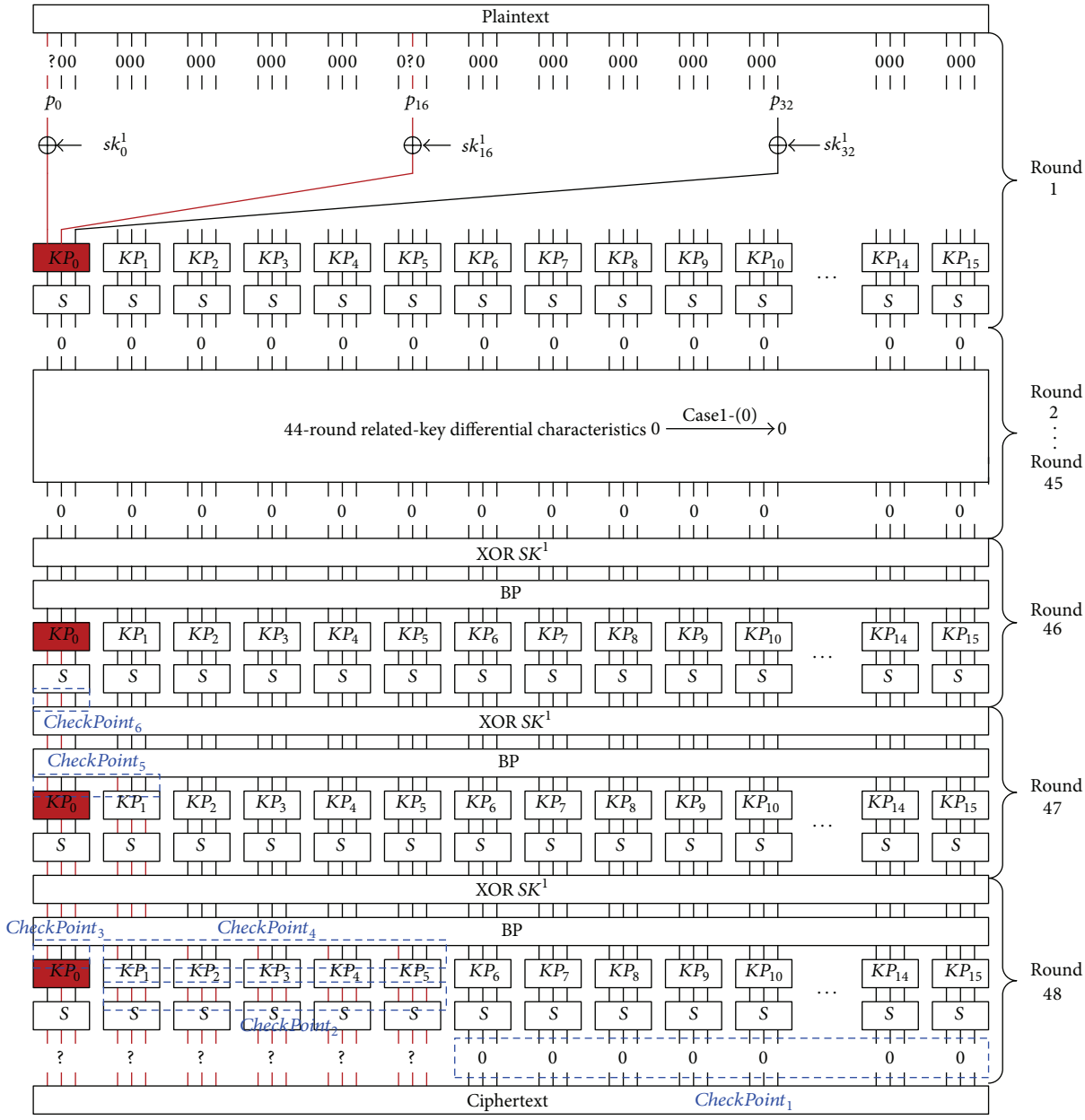


Figure 3: Basic related-key differential attack on PRINTcipher-48.

4.2. Complexities of Basic Related-Key Attack on PRINTcipher-48. This attack considers 4 related keys ($K_0^{(0,0)}, K_0^{(0,1)}, K_0^{(1,0)}$, and $K_0^{(0,0)}$). And 2^{46} plaintexts are used for each related-key. Hence, a data complexity of our attack is 2^{48} related-key chosen plaintexts.

A computational complexity of Step 1 is 2^{48} PRINTcipher-48 encryptions. In Step 2(a) and Step 3(a), $2^{10}(= 8 \cdot 2^{44} \cdot 2^{-37})$ ciphertext pairs are survived. Thus computational complexities of Step 2(b) and Step 3(b) do not exceed 2^{26} PRINTcipher-48 encryptions, respectively. From Step 2 and Step 3, we can recover the 17-bit information on the 80-bit secret key of PRINTcipher-48. Thus, a computational complexity of Step 5 is 2^{63} PRINTcipher-48 encryptions. Hence, a total computational complexity of our attack is 2^{63}.

4.3. Improved Related-Key Attack on PRINTcipher-48. In this subsection, we improve the basic related-key attack on the full PRINTcipher-48. To improve the basic attack, we consider 43-round related-key differential characteristic $0 \xrightarrow{\text{Case 1 (0)}} 0$ (from round 2 to round 44) with a probability of 2^{-43}. The overall attack procedure is similar to the basic attack procedure.

(1) Select 2^{43} plaintext structures which are composed of 4 plaintexts each and obtain the corresponding ciphertexts under four related keys $K_0^{(0,0)}$, $K_0^{(0,1)}$, $K_0^{(1,0)}$, and $K_0^{(1,1)}$, respectively.

(2) Considering the related-key pair $(K_0^{(0,0)}, K_0^{(0,1)})$, we have the following.

(a) Discard wrong pairs which do not satisfy $CheckPoint_7$, $CheckPoint_8$, and $CheckPoint_9$.
(b) Guess 30-bit key $(SK_1^2, \ldots, SK_{15}^2)$ and remain the ciphertext pairs satisfying $CheckPoint_{10}$ and $CheckPoint_{11}$.
(c) Guess 18-bit key $(sk_0^1, \ldots, sk_{17}^1)$ and count the ciphertext pairs satisfying $CheckPoint_{12}$, $CheckPoint_{13}$, and $CheckPoint_{14}$.

(3) Considering the related-key pair $(K_0^{(1,0)}, K_0^{(1,1)})$, we have the following.

(a) Discard wrong pairs which do not satisfy $CheckPoint_7$, $CheckPoint_8$ and $CheckPoint_9$.
(b) Guess 30-bit key $(SK_1^2, \ldots, SK_{15}^2)$ and remain the ciphertext pairs satisfying $CheckPoint_{10}$ and $CheckPoint_{11}$.
(c) Guess 18-bit key $(sk_0^1, \ldots, sk_{17}^1)$ and count the ciphertext pairs satisfying $CheckPoint_{12}$, $CheckPoint_{13}$ and $CheckPoint_{14}$.

(4) Output the guessed key with the maximal count as the right key.
(5) Do an exhaustive search for the remaining secret key using trial encryption.

4.4. Complexities of Improved Related-Key Attack on PRINTcipher-48. The improved attack uses 4 related keys $(K_l^{(0,0)}, K_l^{(0,1)}, K_l^{(1,0)}$, and $K_l^{(0,0)})$ and 2^{47} related-key chosen plaintexts. Hence, a data complexity of our attack is 2^{47} related-key chosen plaintexts.

A computational complexity of Step 1 is 2^{47} PRINTcipher-48 encryptions.

In Step 2(a) and Step 3(a), similar to the basic attack, we discard wrong ciphertext pairs by checking the following three checkPoints (see Figure 4).

(i) $CheckPoint_7$. The output difference of $KP_2, KP_3, \ldots, KP_{15}$ in round 48 should be included in a set $\{0, 1, 2, 4\}$.
(ii) $CheckPoint_8$. The input difference of KP_0 in round 48 should be included in a set $\{0, 2, 4, 6\}$.
(iii) $CheckPoint_9$. The output difference of KP_1 in round 48 should be included in a set $\{0, 1, 2, 3, 4, 5, 6\}$.

The filtering probability of this step is computed as follows:

$$\frac{7}{2^{18}}\left(= 2^{-1\cdot 14}\cdot 2^{-1}\cdot \frac{7}{8}\right). \tag{6}$$

Thus, $7 \cdot 2^{28} (= 8 \cdot 2^{43} \cdot 7/2^{18})$ ciphertext pairs remained on average. So, computational complexities of Step 2(b) and 3(b) are computed as follows:

$$2^{55.22}\left(\approx 7\cdot 2^{28}\cdot 2^{30}\cdot \frac{1}{48}\right). \tag{7}$$

For each remaining ciphertext pair, we guess total 48-bit key $(SK_1^2, \ldots, SK_{15}^2, sk_0^1, \ldots, sk_{17}^1)$. Then we recover 16-bit key by checking the following checkPoints.

(i) $CheckPoint_{10}$. Input difference of $KP_2, KP_3, \ldots, KP_{15}$ in round 48 should be included in a set $\{0, 4\}$.
(ii) $CheckPoint_{11}$. Input difference of KP_1 in round 48 should be included in a set $\{0, 2, 4, 6\}$.
(iii) $CheckPoint_{12}$. Input difference of $KP_0, \ldots, KP_5$ in round 47 should be included in a set $\{0, 4\}$.
(iv) $CheckPoint_{13}$. Input difference of KP_0 and KP_1 in round 46 should be included in a set $\{0, 4\}$.
(v) $CheckPoint_{14}$.Output difference of the first S-box in round 45 should be included in a set $\{0, 2, 4\}$.

In Step 2(b) and Step 3(b), we guess 30-bit key $(SK_1^2, \ldots, SK_{15}^2$ and check $CheckPoint_{10}$ and $CheckPoint_{11}$. The filtering probability of this step is computed as follows:

$$\frac{6}{7\cdot 2^{14}}\left(= 2^{-1\cdot 14}\cdot \frac{6}{7}\right). \tag{8}$$

Thus, $3 \cdot 2^{15}$ $(= 7 \cdot 2^{28} \cdot 6/(7 \cdot 2^{14}))$ ciphertext pairs remained for each guessed key. Since $3 \cdot 2^{15}$ ciphertext pairs remained in Step 2(b) and Step 3(b) for each guessed key, computational complexities of Step 2(c) and Step 3(c) are $3 \cdot 2^{59} (= 3^{15} \cdot 2^{48} \cdot 3/48)$ PRINTcipher-48 encryptions.

In Step 2(c) and Step 3(c), we guess 15-bit key $(sk_0^1, \ldots, sk_{14}^1)$ in order to check $CheckPoint_{12}$, $CheckPoint_{13}$, and $CheckPoint_{14}$. Similar to the basic attack, the expected number of the remaining ciphertext pairs is 1.5 for each guessed key. Note that the expected number of right pairs is 4 similar to the basic attack.

Since we can recover the 49-bit information on the 80-bit secret key of PRINTcipher-48, a computational complexity of Step 5 is 2^{31} PRINTcipher-48 encryptions. Hence, a total computational complexity of our attack is computed as follows:

$$2^{60.62}\left(\approx 2^{47} + 2^{55.22} + \left(3\cdot 2^{59}\right) + 2^{31}\right). \tag{9}$$

5. Related-Key Cryptanalysis on PRINTcipher-96

The overall attack procedure on the full PRINTcipher-96 is similar to that on the full PRINTcipher-48. Thus, we explain the attack procedure on the full PRINTcipher-96 briefly. In this attack, we consider 91-round related-key differential characteristics $0 \xrightarrow{\text{Case 1 (0)}} 0$ (from round 2 to round 92) with a probability of 2^{-91}. The checkPoints used in this attack are as follows (see Figure 5).

(i) $CheckPoint_{15}$. The rightmost 40-bit of ciphertext does not have difference.
(ii) $CheckPoint_{16}$. The output difference of $KP_1, KP_2, \ldots, KP_{17}$ in round 48 is included in a set $\{0, 1, 2, 4\}$.
(iii) $CheckPoint_{17}$. The input difference of KP_0 is included in a set $\{0, 4\}$.

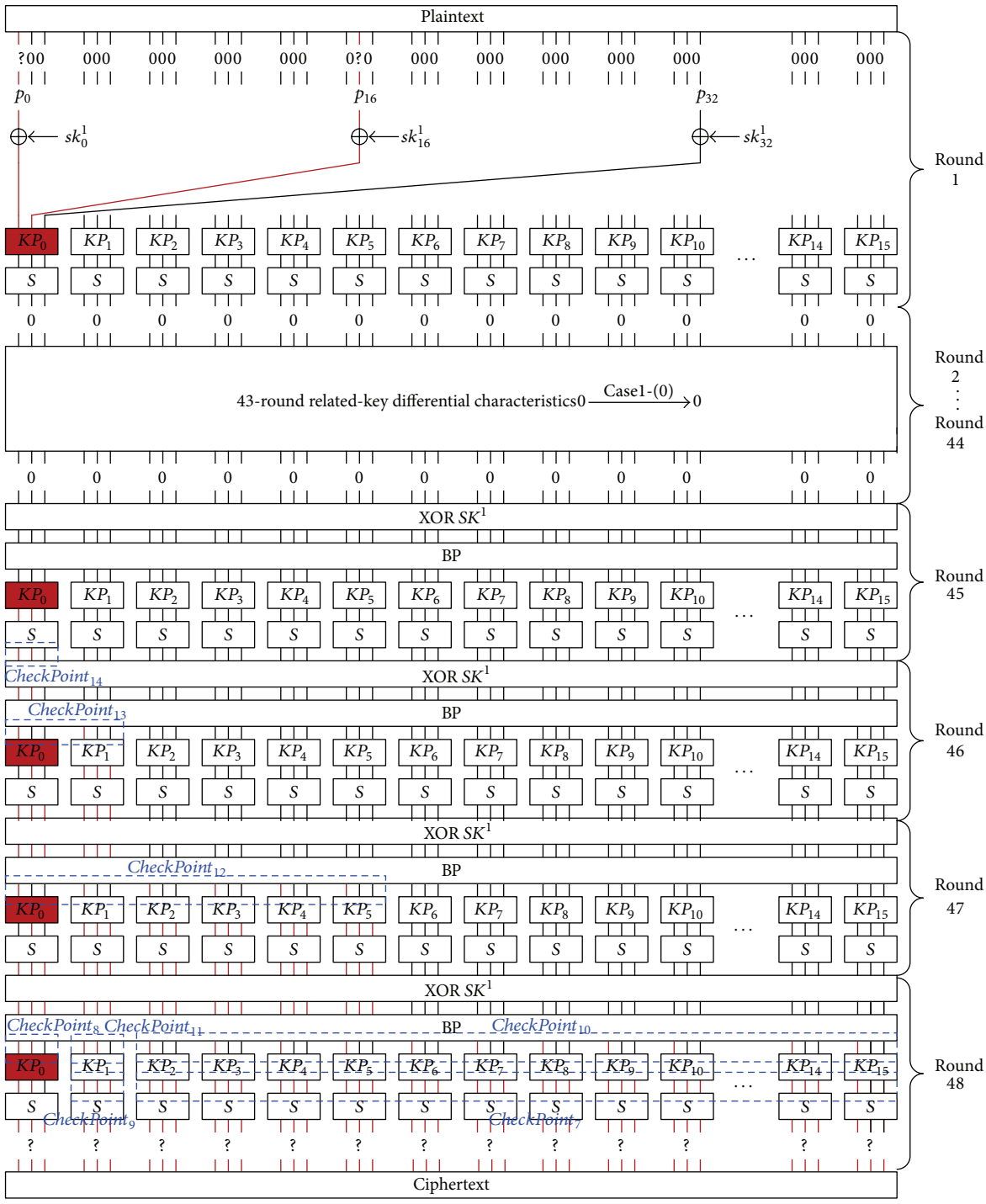


Figure 4: Improved related-key differential attack on PRINTcipher-48.

(iv) $CheckPoint_{18}$. Input difference of $KP_1, KP_2, \ldots, KP_{17}$ in round 96 should be included in a set $\{0, 4\}$.

(v) $CheckPoint_{19}$. Input difference of $KP_0, \ldots, KP_5$ in round 95 should be included in a set $\{0, 4\}$.

(vi) $CheckPoint_{20}$. Input difference of KP_0 and KP_1 in round 94 should be included in a set $\{0, 4\}$.

(vii) $CheckPoint_{21}$. Output difference of the first S-box in round 93 should be included in a set $\{0, 2, 4\}$.

The attack procedure on the PRINTcipher-96 is summarized as follows.

(1) Select 2^{91} plaintext structures which are composed of 4 plaintexts each and obtain the corresponding

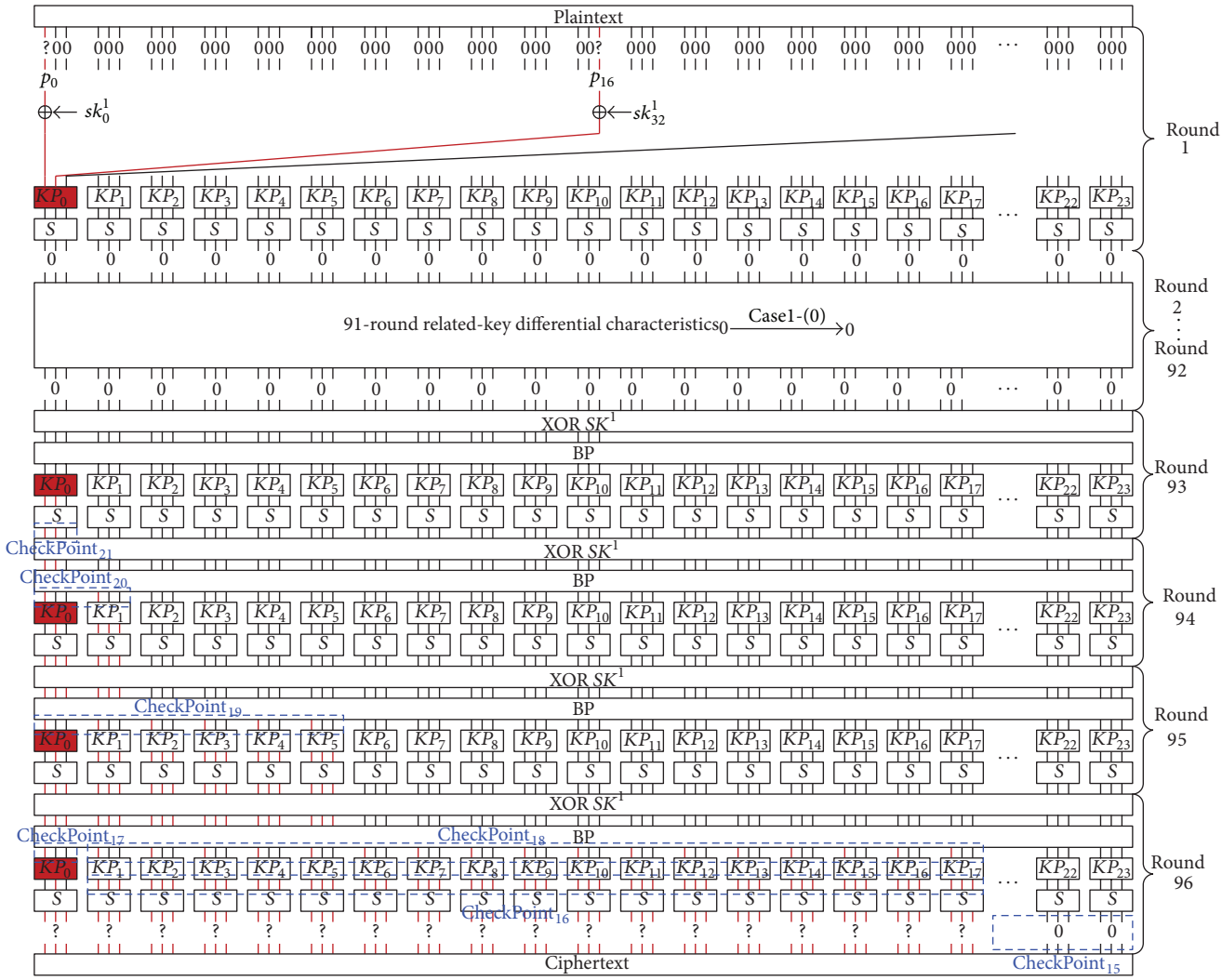


Figure 5: Related-key differential attack on PRINTcipher-96.

ciphertexts under four related keys $K_0^{(0,0)}$, $K_0^{(0,1)}$, $K_0^{(1,0)}$, and $K_0^{(1,1)}$, respectively.

(2) Considering the related-key pair $(K_0^{(0,0)}, K_0^{(0,1)})$, we have the following.

(a) Discard wrong pairs which do not satisfy $CheckPoint_{15}$, $CheckPoint_{16}$, and $CheckPoint_{17}$.

(b) Guess 34-bit key $(SK_1^2, \ldots, SK_{17}^2)$ and remain the ciphertext pairs satisfying $CheckPoint_{18}$.

(c) Guess 18-bit key $(sk_0^1, \ldots, sk_{17}^1)$ and count the ciphertext pairs satisfying $CheckPoint_{19}$, $CheckPoint_{20}$ and $CheckPoint_{21}$.

(3) Considering the related-key pair $(K_0^{(1,0)}, K_0^{(1,1)})$, we have the following.

(a) Discard wrong pairs which do not satisfy $CheckPoint_{15}$, $CheckPoint_{16}$, and $CheckPoint_{17}$.

(b) Guess 34-bit key $(SK_1^2, \ldots, SK_{17}^2)$ and remain the ciphertext pairs satisfying $CheckPoint_{18}$.

(c) Guess 18-bit key $(sk_0^1, \ldots, sk_{17}^1)$ and count the ciphertext pairs satisfying $CheckPoint_{19}$, $CheckPoint_{20}$, and $CheckPoint_{21}$.

(4) Output the guessed key with the maximal count as the right key.

(5) Do an exhaustive search for the remaining secret key using trial encryption.

Since 4 related keys $(K_0^{(0,0)}, K_0^{(0,1)}, K_0^{(1,0)}$, and $K_0^{(0,0)})$ are used in our attack, the data complexity of our attack is 2^{95} related-key chosen plaintexts. And a total computational complexity of our attack is about 2^{107} PRINTcipher-96 encryptions.

6. Conclusion

In this paper, we proposed related-key cryptanalysis of PRINTcipher. Our attack results are summarized in Table 1. To recover the 80-bit secret key of PRINTcipher-48, our

attack required 2^{47} related-key chosen plaintexts with a computational complexity of $2^{60.62}$. In the case of PRINTcipher-96, 2^{95} related-key chosen plaintexts with a computational complexity of 2^{107} are required. These results are the first known related-key cryptanalytic results on them.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the C-ITRC (Convergence Information Technology Research Center) Support Program (NIPA-2013-H0301-13-3007).

References

[1] A. Grover and H. Berghel, "A survey of RFID deployment and security issues," *Journal of Information Processing Systems*, vol. 7, no. 4, pp. 561–580, 2011.

[2] D. Seo and I. Lee, "A study on RFID system with secure service availability for ubiquitous computing," *Journal of Information Processing Systems*, vol. 1, no. 1, pp. 96–101, 2005.

[3] A. Bogdanov, L. R. Knudsen, G. Leander et al., "PRESENT: an ultra-lightweight block cipher," in *Cryptographic Hardware and Embedded Systems—CHES 2007*, vol. 4727 of *Lecture Notes in Computer Science*, pp. 450–466, Springer, Berlin, Germany, 2007.

[4] J. Guo, T. Peyrin, A. Poschmann, and M. Robshaw, "The LED block cipher," in *Cryptographic Hardware and Embedded Systems—CHES 2011*, vol. 6917 of *Lecture Notes in Computer Science*, pp. 326–341, Springer, Berlin, Germany, 2011.

[5] D. Hong, J. Sung, S. Hong et al., "HIGHT: a new block cipher suitable for low-resource device," in *Cryptographic Hardware and Embedded Systems—CHES 2006*, vol. 4249 of *Lecture Notes in Computer Science*, pp. 46–59, Springer, Berlin, Germany, 2006.

[6] L. Knudsen, G. Leander, A. Poschmann, and M. J. B. Robshaw, "PRINTcipher: a block cipher for IC-printing," in *Cryptographic Hardware and Embedded Systems, CHES 2010*, vol. 6225 of *Lecture Notes in Computer Science*, pp. 16–32, Springer, Berlin, Germany, 2010.

[7] K. Shibutani, T. Isobe, H. Hiwatari, A. Mitsuda, T. Akishita, and T. Shirai, "Piccolo: an ultra-lightweight blockcipher," in *Cryptographic Hardware and Embedded Systems—CHES 2011*, vol. 6917 of *Lecture Notes in Computer Science*, pp. 342–357, Springer, Berlin, Germany, 2011.

[8] G. Leander, M. A. Abdelraheem, H. Alkhzaimi, and E. Zenner, "A cryptanalysis of PRINTcipher: the invariant subspace attack," in *Advances in Cryptology—CRYPTO 2011*, vol. 6841 of *Lecture Notes in Computer Science*, pp. 206–221, Springer, Berlin, Germany, 2011.

[9] M. A. Abdelraheem, G. Leander, and E. Zenner, "Differential cryptanalysis of round-reduced PRINTcipher: computing roots of permutations," in *Fast Software Encryption*, vol. 6733 of *Lecture Notes in Computer Science*, pp. 1–17, Springer, Berlin, Germany, 2011.

[10] M. Ågren and T. Johansson, "Linear cryptanalysis of PRINTcipher—trails and samples everywhere," in *Progress in Cryptology—INDOCRYPT 2011*, vol. 7107 of *Lecture Notes in Computer Science*, pp. 114–133, Springer, Berlin, Germany, 2011.

[11] F. Karakoç, H. Demirci, and A. E. Harmancı, "Combined differential and linear cryptanalysis of reduced-round PRINTcipher," in *Selected Areas in Cryptography*, vol. 7118 of *Lecture Notes in Computer Science*, pp. 169–184, Springer, Berlin, Germany, 2012.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2013, Article ID 281895, 8 pages
http://dx.doi.org/10.1155/2013/281895

Research Article

Tone-Independent Orthogonalizing Lattice Equalization for Insufficient Cyclic-Prefix OFDM Transmissions

Dong Kyoo Kim[1] and Yang Sun Lee[2]

[1] *Electronics and Telecommunications Research Institute, Daejeon 305-700, Republic of Korea*
[2] *Division of Computer Engineering, Mokwon University, Daejeon 302-729, Republic of Korea*

Correspondence should be addressed to Yang Sun Lee; yslee48@gmail.com

Received 24 September 2013; Revised 1 November 2013; Accepted 4 November 2013

Academic Editor: Jongsung Kim

Tone-independent orthogonalizing lattice per tone equalizer (TOL-PTEQ) is introduced and its convergence is analyzed. Cyclic-prefix redundancy, one of the major drawbacks of orthogonal frequency division multiplexing (OFDM), can be reduced by TOL-PTEQ. Fast convergence and low computational complexity of TOL-PTEQ are also suitable properties for packet-based wireless communications and detections in which OFDM is widely deployed for their modulation technique.

1. Introduction

PTEQ was originally proposed for optimizing bit rate of discrete-multitone (DMT) modem in wired communications such as digital subscriber lines (DSLs), where SNR of each tone can be independently maximized [1–4]. In these literatures, computational complexity is a major issue because they should cover a large number of tones, for example, DMTs with 512, 1024, or 2048 tones. Several types of stochastic-gradient algorithms for PTEQ have been proposed to reduce the computational complexity [1, 2]. But convergence rate is considered as a minor issue for DSLs because various DSLs have long training sequences in their initial set-up process.

Wireless broadband communication technology is becoming more important for pervasive healthcare solutions as healthcare applications [5–7] are extending their coverage up to global scale as shown in Figure 1 [8]. According to [8], researches on more fast and reliable wireless infrastructures are conducted in order to improve the healthcare services in remote location. Candidate wireless technologies are as follows: IEEE 802.11x, IEEE 802.16x, ETSI HiperLAN, ETSI HiperMAN, and so on. A common feature of the candidates is that they use OFDM as their modulation method which is the promising technology because it is easy to handle the multipath channel problem by using fast Fourier transform. It is also widely utilized for multiple-access method, say OFDMA. It gives multiuser diversity taking advantage of channel frequency selectivity and good scalability over wide range of bandwidth that is achieved just by adjusting FFT size, where FFT stands for fast Fourier transform [9]. OFDM can also be utilized in the field of radar technologies as shown in Figure 2, where the multitone technique can be applied to enhance the radar scanning performance [10]. In this case, various OFDM technologies are essential to the multitone based radar systems. As shown in Figure 2, the radar transmits and receives the radar signal through the antennas. The received signal contains various reflection signals generated by the interfaces between two different layers. To obtain the high-resolution reflection signals, the radar system should use the ultrawideband signal, which can be created by composing several narrowband signals as shown in Figure 2. The multitone based radar system uses the multitone signal as the narrowband signal, in which we can utilize various existing advanced technologies of OFDM such as the channel estimation and the computationally efficient and fast implementation architectures.

Two major drawbacks of OFDM are the peak-to-average power ratio and cyclic-prefix (CP) redundancy. To cope with the CP redundancy problem that decreases spectral efficiency, several approaches have been proposed [11, 12]. In [11], iterative cancellation method was used to cancel interferences due to the insufficient CP, where terrestrial HDTV

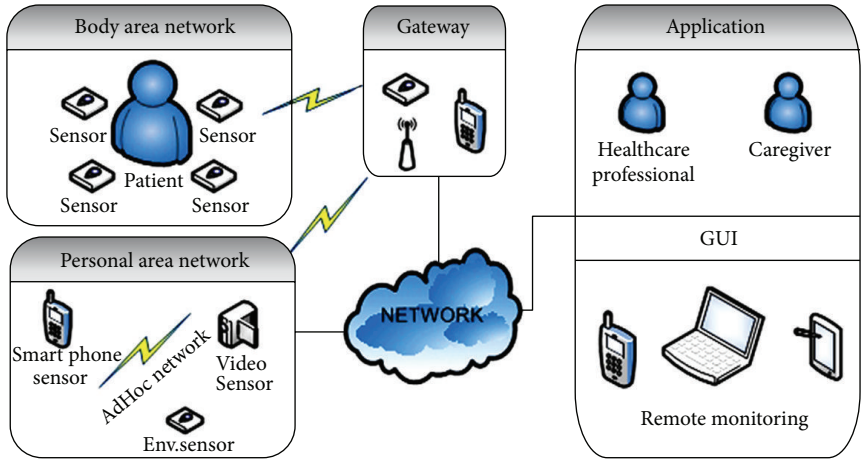


Figure 1: Overview of a simple WSN application scenario for healthcare (this figure is quoted from Figure 1 in [8]).

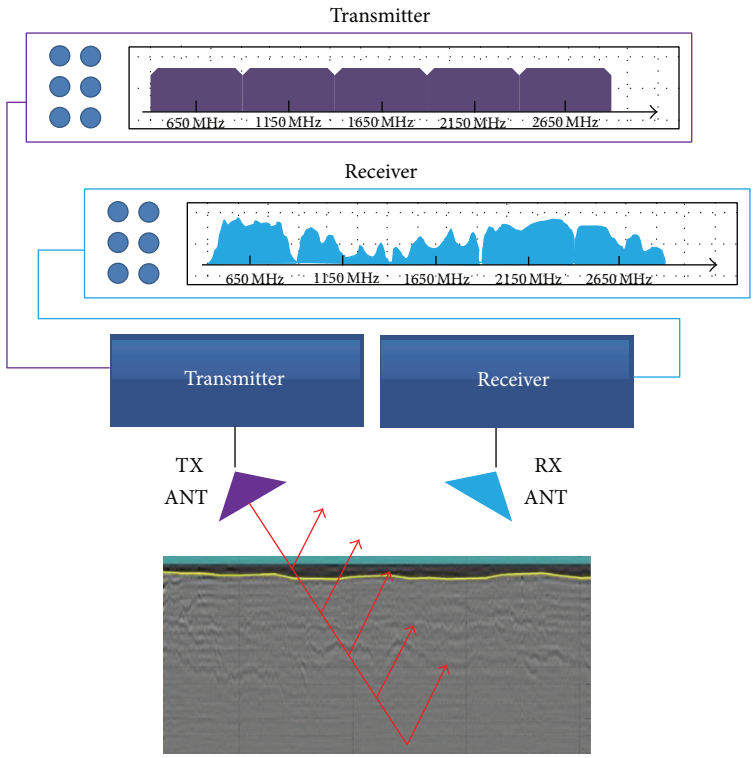


Figure 2: Multitone based radar technology.

broadcasting scenario was considered as its application. As mentioned in page 1598 of [11], computational burdens of iterations are very big because $2I + 2$ FFT operations are needed for I iterations. In [12], precompensation was used to avoid differential group delay of optical OFDM systems. This precompensation of delay can ensure zero intersymbol interference while reducing CP length. To do this, channel state information is fed back from receiver to transmitter.

PTEQ technique can be applied to the CP reduction problem of the OFDM-based wireless communication systems because PTEQ is the frequency-domain version of the time-domain channel-shortening equalizer (TEQ) which shortens channel impulse response length [13, 14]. Thus, this shortening property can be applied to the problem that makes CP length less than the channel impulse response length. Almost all wireless OFDM systems are based on packet-based

transmissions and a quasistatic environment is assumed, where a packet consists of a preamble and a payload. A new channel estimation should be performed in every packet; thus fast convergence of PTEQ is needed not to slow down the overall transmission throughput.

In this paper, we analyze convergence properties of the existing PTEQ algorithms. It will be shown that their convergence properties are related to eigenvalue spread of PTEQ-input signals especially in the initial training period.

We then show a stochastic-gradient lattice-typed PTEQ algorithm, TOL-PTEQ, with fast convergence rate while maintaining a low computational complexity, where its convergence rate is fast enough for packet-based wireless communications. As opposed to the existing transversal filter-based PTEQs, TOL-PTEQ has lattice feature to orthogonalize a portion of its input signals with low computational complexity. To the best of our knowledge, there is no approach in which the lattice-typed PTEQ is applied to the OFDM technology before. We also provide the convergence model of TOL-PTEQ, where the accuracy of the model is shown by comparing the model with numerical simulation results.

The rest of this paper is organized as follows. In Section 2, problem setup for PTEQ is described and one existing PTEQ algorithm, RLSLMS-PTEQ, is analyzed. In Section 3, TOL-PTEQ is shown in detail. In Section 4, simulation results and comparisons of the various PTEQ algorithms are provided. Finally, conclusion is provided in Section 5.

2. Problem Setup and RLSLMS-PTEQ

2.1. Problem Setup. Signal vector $\mathbf{x}^{(k)}$ at block time k is transmitted and propagated through the $(L+k+1)$th-order channel $\mathbf{h}$. Then, the received signal vector $\mathbf{y}^{(k)}$ is written as

$$\mathbf{y}^{(k)} = \left[\mathbf{O} \mid \mathrm{T}\left(\left[h_L \;\; \mathbf{0}_{1X(L+K)}\right]^T, \left[\mathbf{h} \;\; \mathbf{0}_{1X(L+K)}\right]\right) \mid \mathbf{O}\right] \cdot \mathbf{x}^{(k)} + \mathbf{n}^{(k)}, \tag{1}$$

where the superscript T is the transpose operator, $\mathrm{T}(\mathbf{a}, \mathbf{b})$ is the Toeplitz matrix of which first column and row vector are $\mathbf{a}$ and $\mathbf{b}$, L and K are the length of post- and precursors, and $\mathbf{n}^{(k)}$ is the white Gaussian noise vector.

According to [1], (1) is equalized by the Tth-order PTEQ $\overline{\mathbf{v}}_i$ of the ith subcarrier and its output can be written as

$$Z_i^{(k)} = \overline{\mathbf{v}}_i^T \begin{bmatrix} \mathbf{I}_{T-1} & \mathbf{O}_{(T-1)X(N-T-1)} & -\mathbf{I}_{T-1} \\ \mathbf{0}_{1X(T-1)} & F_N(i,:) & \end{bmatrix} \mathbf{y}^{(k)} = \overline{\mathbf{v}}_i^T \mathbf{u}^{(k)}, \tag{2}$$

where $F_N(i,:)$ is the ith row vector of N-point FFT matrix, $\mathbf{u}^{(k)} = \left[\mathbf{d}^{(k)T} \;\; Y_i^{(k)}\right]$ is the input vector, $Y_i^{(k)}$ is the Fourier transform of $\mathbf{y}^{(k)}$, and $\mathbf{d}^{(k)} \equiv \left[d_0^{(k)} \;\; \cdots \;\; d_{T-2}^{(k)}\right]$ is the difference vector of which element is written as $d_l^{(k)} \equiv -y_{(k+1)s-l} + y_{ks+\nu-l}$. $\widehat{\overline{\mathbf{v}}}_i^{\mathrm{opt}}$ is hereby the estimated optimal vector that minimizes the expectation of $|Z_i^{(k)} - X_i^{(k)}|^2$, where $X_i^{(k)}$ is the training or pilot subcarrier signal.

The most powerful algorithm to obtain $\widehat{\overline{\mathbf{v}}}_i^{\mathrm{opt}}$ is the recursive least squares (RLS). The performance of RLS-PTEQ [15] will be shown in Section 3, in which it outperforms other PTEQs. In case that the number of subcarriers is large, stochastic-gradient-based algorithms such as NLMS-PTEQ and RLSLMS-PTEQ [2] are preferred because the computational complexity of RLS-PTEQ is too big to be implemented. However, convergence rates of the stochastic-gradient-based PTEQs are so slow that they may slow down the transmission throughput of packet-based wireless communications. The convergence of these stochastic-gradient PTEQs is highly related to the eigenvalue spread of their input autocorrelation matrix $R_{\mathbf{uu}} = E[\mathbf{u}^{(k)}\mathbf{u}^{(k)H}]$, where superscript H denotes the Hermitian. $\mathbf{u}^{(k)}$ is highly correlated because its elements consist of the combination of received signals. With high eigenvalue spread, it is well known that NLMS-PTEQ has poor performance in the sense of convergence rate and steady-state misadjustment; this can be also seen through simulations in Section 3.

RLSLMS-PTEQ which combines NLMS-PTEQ with RLS-PTEQ has effectively low computational complexity compared with that of RLS-PTEQ, where RLS is not used for $\overline{\mathbf{v}}_i$ update but for only $\mathbf{d}^{(k)}$ calculations. Main purpose of RLSLMS-PTEQ is the decorrelation of $\mathbf{d}^{(k)}$ with computationally complex RLS, and this decorrelation result is commonly utilized to all subcarriers' PTEQ in which the equalization of each subcarrier is performed by simple NLMS. The steady-state misadjustment of RLSLMS-PTEQ is close to that of RLS-PTEQ. However, its convergence rate is much slower than that of RLS-PTEQ as shown in [2], and it will be further discussed in the next section.

2.2. Analysis of RLSLMS-PTEQ. The autocorrelation $R_{\mathbf{lu}}$ of RLSLMS-PTEQ has, approximately, $(T-2)$ eigenvalues of $1/W$ and two eigenvalues of $(1 \pm d)/W$ with d as

$$d = \sqrt{\frac{W^2}{e_i^{(k+1)}} E\left[Y_i^* \mathbf{d}^{(k+1)T}\right] \overline{\mathbf{S}}^{(k+1)T} \overline{\mathbf{S}}^{(k+1)} E\left[\mathbf{d}^{(k+1)} Y_i\right]}, \tag{3}$$

where d, W, $\mathbf{l}$, $\overline{\mathbf{S}}_i$, and $e^{(k+1)}$ are defined in [2]. Equation (3) can be written as

$$\begin{aligned} d &= \sqrt{\frac{W^2}{e_i^{(k+1)}} E\left[Y_i^* \mathbf{d}^{(k+1)T}\right] E\left[\overline{\mathbf{k}}^{(k+1)} Y_i\right]} \\ &= \sqrt{\frac{W^2}{e_i^{(k+1)}} E\left[Y_i^* Y_i \mathbf{d}^{(k+1)T} \overline{\mathbf{k}}^{(k+1)}\right]} \\ &\approx \sqrt{WE\left[\mathbf{d}^{(k+1)T} \overline{\mathbf{k}}^{(k+1)}\right]}, \end{aligned} \tag{4}$$

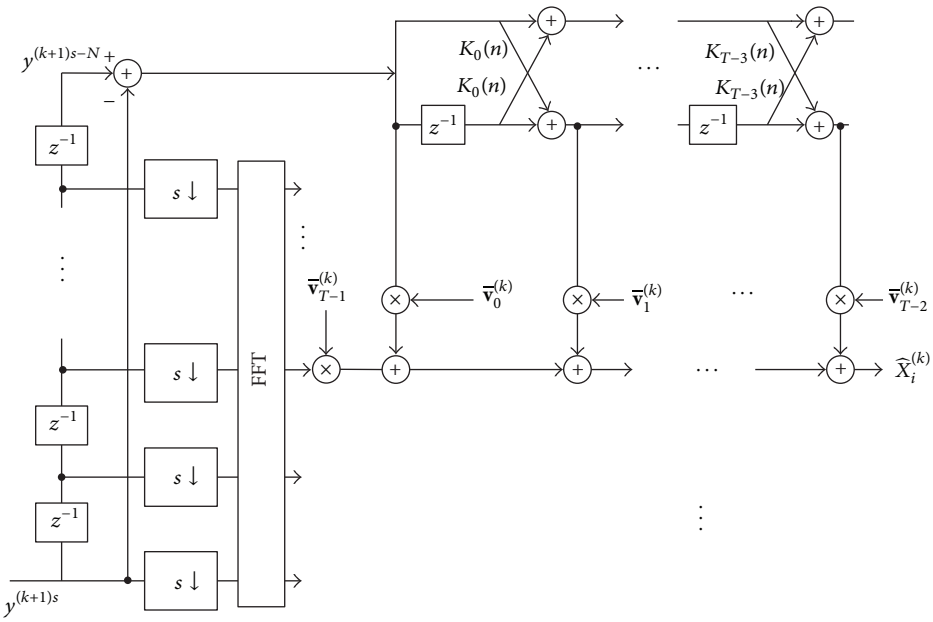


Figure 3: Structure of TOL-PTEQ.

where $\overline{\mathbf{k}}^{(k+1)} = \overline{\mathbf{S}}^T \overline{\mathbf{S}} \mathbf{d}^{(k+1)}$. Conversion factor is defined as

$$\overline{\gamma}^{(k+1)} = 1 - \mathbf{d}^{(k+1)T} \overline{\mathbf{k}}^{(k+1)}. \tag{5}$$

Using (4) and (5), we can write

$$d \approx \sqrt{W\left(1 - E\left[\gamma^{(k+1)}\right]\right)}. \tag{6}$$

On initial transient phase in training period of RLS, $E[\gamma^{(k)}]$ starts with very small value, and then it converges to 1 according to the inverse law: $E[\gamma^{(k)}] \approx 1 - ((T-1)/k)$ [16]. Hence, in case of RLSLMS-PTEQ, the large eigenvalue spread of $R_{\mathbf{lu}}$ induced by the small value of $E[\gamma^{(k)}]$ on the initial transient phase causes the convergence rate to be slow. Thus, it can be said that RLS property of RLSLMS-PTEQ is helpful in the steady-state phase but is not too much effective in the initial transition phase.

3. TOL-PTEQ

3.1. Algorithm and Architecture. We show the TOL-PTEQ of which structure is depicted in Figure 3. There is no RLS in TOL-PTEQ. As mentioned in Section 2.2, RLSLMS-PTEQ suffers from slow convergence in its initial transition phase. Two different optimization criteria in one recursive algorithm may work poorly in the initial transition phase. Therefore, decorrelation of $\mathbf{d}^{(k)}$ in TOL-PTEQ is performed by stochastic-gradient algorithm. In general, convergence performance of adaptive algorithm using stochastic-gradient decorrelation method lies in between those of LMS and RLS, which is also applied to TOL-PTEQ as shown in simulation results in Section 3. As TOL-PTEQ operates entirely in a stochastic-gradient criterion, its misadjustment in a steady state is worse than that of RLS-PTEQ. But, as in our simulation results, TOL-PTEQ reach a certain steady-state error level faster than RLSLMS-PTEQ. TOL-PTEQ also inherits lattice features such as easily implemented modular structure, computational efficiency, and numerical stability. Complete algorithm of TOL-PTEQ and complexity analysis is written in the rest of this subsection.

Algorithm: TOL-PTEQ. The algorithm is as follows.

Tone independent:

For $n = 0, 1, \ldots, T-2$

$$f_0^{(k)}(n) = b_0^{(k)}(n) = d_0^{(k)}. \tag{7}$$

For $m = 0, 1, \ldots, \min(n, T-3)$

$$\begin{aligned}
f_{m+1}^{(k)}(n) &= f_m^{(k)}(n) + K_m^{(k)}(n)\, b_m^{(k)}(n-1), \\
b_{m+1}^{(k)}(n) &= b_m^{(k)}(n-1) + K_m^{(k)}(n)\, f_m^{(k)}(n), \\
E_m^{(k)}(n) &= (1-\beta) E_m^{(k)}(n-1) \\
&\quad + \beta\left(f_m^{(k)2}(n) + b_m^{(k)2}(n-1)\right), \\
K_m^{(k)}(n) &= K_m^{(k)}(n-1) + \mu\left(a + E_m^{(k)}(n)\right)^{-1} \\
&\quad \cdot \left(f_m^{(k)}(n)\, b_{m+1}^{(k)}(n) \right. \\
&\quad \left. + b_m^{(k)}(n-1)\, f_{m+1}^{(k)}(n)\right).
\end{aligned} \tag{8}$$

Tone dependent:

$$\begin{aligned}
C_i^{(k)} &= (1-\beta) C_i^{(k)} + \beta \left|Y_i^{(k)}\right|^2, \\
\overline{\mathbf{v}}_i^{(k+1)} &= \overline{\mathbf{v}}_i^{(k)} + \mu \mathbf{c}_i^{(k)} e_i^{(k)*},
\end{aligned} \tag{9}$$

Table 1: Computational complexity for PTEQs.

	Multiplications	Square-root operations	Divisions
TOL-PTEQ	$3T^2 + (4N-2)T + 5$	0	$T-2$
RLS-PTEQ	$3T^2 + (20N+1)T + 10N - 4$	$T+N-1$	0
RLSLMS-PTEQ	$3T^2 + (4N+1)T + 12N - 4$	$T-1$	0
NLMS-PTEQ	$(T-1) + (4T+8)N$	0	1

where $f_m^{(k)}(n)$ and $b_m^{(k)}(n)$ are mth-order forward and backward prediction error, $K_m^{(k)}(n)$ is the partial correlation coefficient, $e_i^{(k)} \equiv X_i^{(k)} - \overline{\mathbf{v}}_i^{(k)H}\mathbf{c}_i^{(k)}$, and $\mathbf{c}_i^{(k)} \equiv \left[\mathbf{b}^{(k)T} \;\; Y_i^{(k)}/C_i^{(k)}\right]^T$. Tone-independent part of the algorithm consists of lattice modules, in which the update equations for partial correlation coefficients in (8) are based on gradient adaptive method [17]. Equation (9) in the tone dependent part is the computationally efficient LMS algorithm

The whole algorithm requires $3T^2 + (4N-2)T + 5$ multiplications and $T-2$ divisions. Computation complexities of PTEQs are written in Table 1. RLS-PTEQ and RLSLMS-PTEQ require $3T^2 + (20N+1)T + 10N - 4$ and $3T^2 + (4N+1)T + 12N - 4$ multiplications, respectively. They also require $T+N-1$ and $T-1$ square-root operations. Compared with the two PTEQs, TOL-PTEQ reduces $(16N+1)T + 5N - 4$ and $T + 12N + 1$ multipliers, respectively. In the sense that square-root operations and division are performed by Newton's method, it is considered that they require the same order of multiplications. Thus, it can be said that TOL-PTEQ saves $N+1$ multiplications and one square-root operation compared with RLS-PTEQ and RLSLMS-PTEQ, respectively. NLMS-PTEQ has the smallest number of multiplications; however the performance of this algorithm is lower than that of the other algorithms.

3.2. Convergence Analysis. We extend convergence model described in [17] to TOL-PTEQ. Convergence model for TOL-PTEQ consists of three parts; one is the model for forward and backward predictors as described in [17], and the other two parts are the learning-curve models of $E[\overline{\mathbf{v}}_i^{(k)}]$ and $E[|e_i^{(k)}|^2]$ which are written here as follows. Equation (9) can be written as

$$\overline{\mathbf{v}}_i^{(k+1)} = \left(\mathbf{I} - \mu\mathbf{c}_i^{(k)}\mathbf{c}_i^{(k)H}\right)\overline{\mathbf{v}}_i^{(k)} + \mu\mathbf{c}_i^{(k)}X_i^{(k)*}. \tag{10}$$

By taking expectation of both sides of (10), we can write

$$\begin{aligned} E\left[\overline{\mathbf{v}}_i^{(k+1)}\right] &= E\left[\left(\mathbf{I} - \mu\mathbf{c}_i^{(k)}\mathbf{c}_i^{(k)H}\right)\overline{\mathbf{v}}_i^{(k)}\right] + \mu E\left[\mathbf{c}_i^{(k)}X_i^{(k)*}\right] \\ &\approx \left(\mathbf{I} - \mu E\left[\mathbf{c}_i^{(k)}\mathbf{c}_i^{(k)H}\right]\right)E\left[\overline{\mathbf{v}}_i^{(k)}\right] + \mu E\left[\mathbf{c}_i^{(k)}X_i^{(k)*}\right]. \end{aligned} \tag{11}$$

The (m,n)th element of $E[\mathbf{c}_i^{(k)}\mathbf{c}_i^{(k)H}]$ in (11) can be computed as

$$\begin{aligned} E\left[\mathbf{c}_i^{(k)}\mathbf{c}_i^{(k)H}\right]_{(1,1)} &= E\left[|Y_i|^2\right], \\ E\left[\mathbf{c}_i^{(k)}\mathbf{c}_i^{(k)H}\right]_{(1,n\neq 1)} &= E\left[Y_i\left(d_n^{(k)} - \sum_{j=n+1}^{T-2} f_{b,j}^{(k)}d_j^{(k)}\right)\right] \\ &\approx E\left[Y_i d_n^{(k)}\right] - \sum_{j=n+1}^{T-2} f_{b,j}^{(k)}E\left[Y_i d_j^{(k)}\right], \\ E\left[\mathbf{c}_i^{(k)}\mathbf{c}_i^{(k)H}\right]_{(m\neq 1,1)} &= E\left[\mathbf{c}_i^{(k)}\mathbf{c}_i^{(k)H}\right]^*_{(1,n\neq 1)}, \\ E\left[\mathbf{c}_i^{(k)}\mathbf{c}_i^{(k)H}\right]_{(m\neq 1,n\neq 1)} &= E\left[\left(d_m^{(k)} - \sum_{j=m+1}^{T-2} f_{b,j}^{(k)}d_j^{(k)}\right)\left(d_n^{(k)} - \sum_{p=n+1}^{T-2} f_{b,p}^{(k)}d_p^{(k)}\right)\right] \\ &\approx E\left[d_m^{(k)}d_n^{(k)}\right] - \sum_{p=n+1}^{T-2} f_{b,p}^{(k)}E\left[d_m^{(k)}d_p^{(k)}\right] \\ &\quad - \sum_{j=m+1}^{T-2} f_{b,j}^{(k)}E\left[d_j^{(k)}d_n^{(k)}\right] \\ &\quad + \sum_{j=n+1}^{T-2}\sum_{p=n+1}^{T-2} f_{b,j}^{(k)}f_{b,p}^{(k)}E\left[d_j^{(k)}d_p^{(k)}\right], \end{aligned} \tag{12}$$

and mth element of $E[\mathbf{c}_i^{(k)}X_i^{(k)*}]$ can be computed as

$$\begin{aligned} E\left[\mathbf{c}_i^{(k)}X_i^{(k)*}\right]_{m=1} &= E\left[Y_i^{(k)}X_i^{(k)*}\right], \\ E\left[\mathbf{c}_i^{(k)}X_i^{(k)*}\right]_{m\neq 1} &= E\left[\left(d_m^{(k)} - \sum_{j=m+1}^{T-2} f_{b,j}^{(k)}d_j^{(k)}\right)X_i^{(k)*}\right] \\ &\approx E\left[d_m^{(k)}X_i^{(k)*}\right] - \sum_{j=m+1}^{T-2} f_{b,j}^{(k)}E\left[d_j^{(k)}X_i^{(k)*}\right], \end{aligned} \tag{13}$$

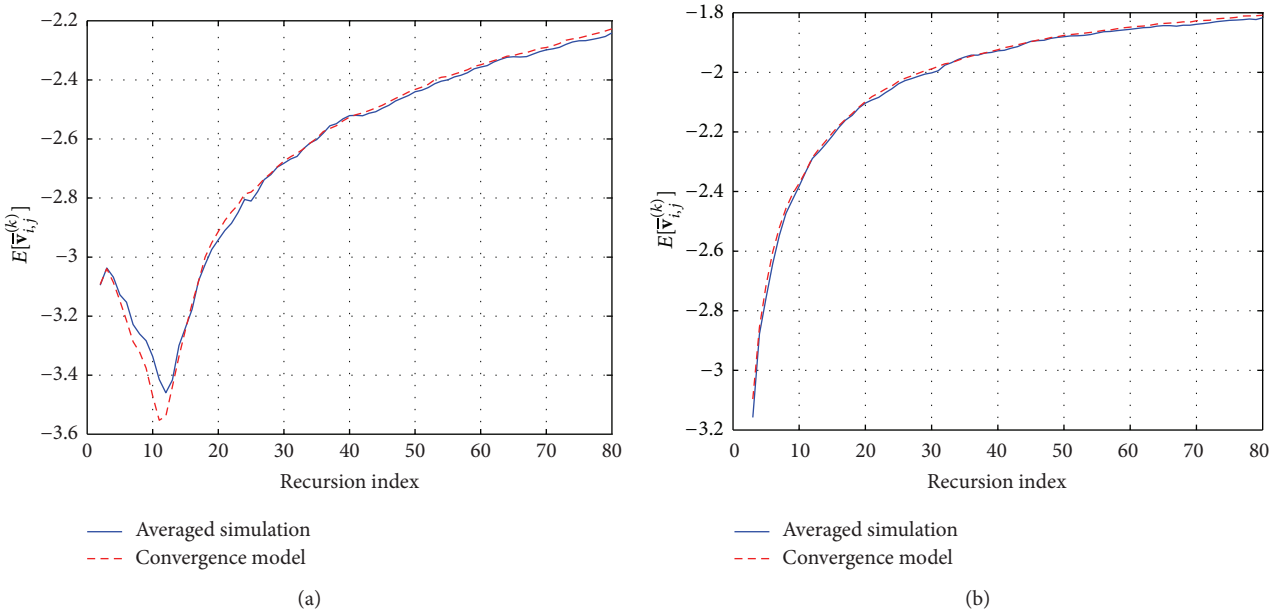


FIGURE 4: Trajectories of $E[\bar{v}_{i,j}^{(k)}]$; dashed line is for convergence model, and solid line is for averaged simulation result. (a) $j = 3$; (b) $j = 12$.

where $m, n \in \{1, 2, \ldots, T-2\}$ and the computation of $E[d_m^{(k)} d_n^{(k)}]$ and $f_{b,m}^{(k)}$ in the above equations are provided in [17]. Mean square error, $E[|e_i^{(k)}|^2]$, can be computed as

$$\begin{aligned}
\text{M.S.E} &= E\left[\left|X_i^{(k)}\right|^2 - \bar{\mathbf{v}}_i^{(k)T} X_i^{(k)} \mathbf{c}_i^{(k)*} \right. \\
&\qquad \left. - \bar{\mathbf{v}}_i^{(k)H} X_i^{(k)} \mathbf{c}_i^{(k)} + \bar{\mathbf{v}}_i^{(k)H} \mathbf{c}_i^{(k)} \mathbf{c}_i^{(k)H} \bar{\mathbf{v}}_i^{(k)}\right] \\
&\approx E\left[\left|X_i^{(k)}\right|^2\right] - \bar{\mathbf{v}}_i^{(k)T} E\left[X_i^{(k)} \mathbf{c}_i^{(k)*}\right] \\
&\quad - \bar{\mathbf{v}}_i^{(k)H} E\left[X_i^{(k)} \mathbf{c}_i^{(k)}\right] + \bar{\mathbf{v}}_i^{(k)H} E\left[\mathbf{c}_i^{(k)} \mathbf{c}_i^{(k)H}\right] \bar{\mathbf{v}}_i^{(k)},
\end{aligned} \tag{14}$$

where $E[X_i^{(k)} \mathbf{c}_i^{(k)}]$ and $E[\mathbf{c}_i^{(k)} \mathbf{c}_i^{(k)H}]$ can be obtained by the same manner as in (11). We obtained two learning curve models as the equalizer tap-weight model (11) and the mean square error model (14).

For example, trajectories of $E[\bar{v}_{i,j}^{(k)}]$ for $j = 3$ and $j = 12$ at $i = 100$ are depicted in Figures 4(a) and 4(b). In the figures, it is also shown that the equalizer tap-weight model (11) coincides with that of simulation results, where simulation results are obtained by averaging over 1000 independent trials. Detailed simulation setup will be seen in the first paragraph of the next section. Figure 5 shows the trajectories of the mean square error, $E[|e_i^{(k)}|^2]$ ($i = 100$). The mean square error (M.S.E) model (14) also accurately tracks the mean square error of the simulation results. From Figure 4, it can be said that the convergence model of TOL-PTEQ is accurate with small β; it is also mentioned in [17] that small β results in accurate convergence model.

Cumulative distribution function of eigenvalues for $\mathbf{u}^{(k)}$ and $\mathbf{c}^{(k)}$ is depicted in Figure 6. This shows that TOL-PTEQ effectively decorrelates $\mathbf{u}^{(k)}$ to $\mathbf{c}^{(k)}$ that helps TOL-PTEQ speed up the convergence.

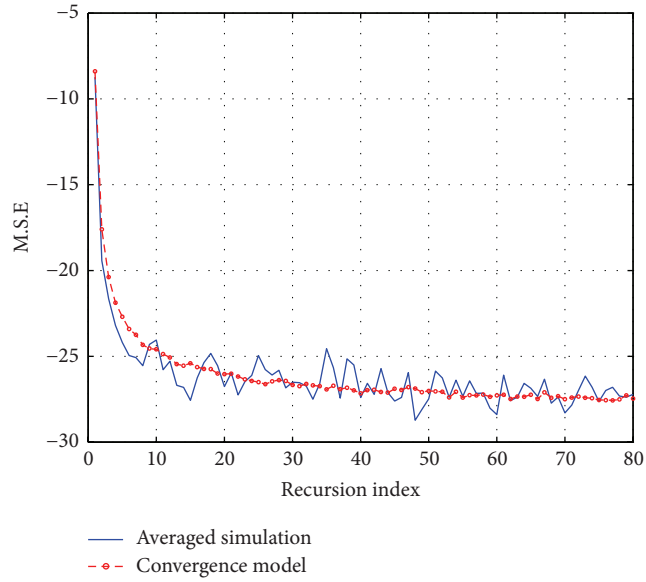


FIGURE 5: M.S.E trajectories with $T = 16$; dashed line is for convergence model, and solid line is for averaged simulation result.

4. Simulation Results

This section provides simulation results of learning curves of four PTEQs and further detailed simulations for analysis of convergence rate according to SNR and order of PTEQ tap weights.

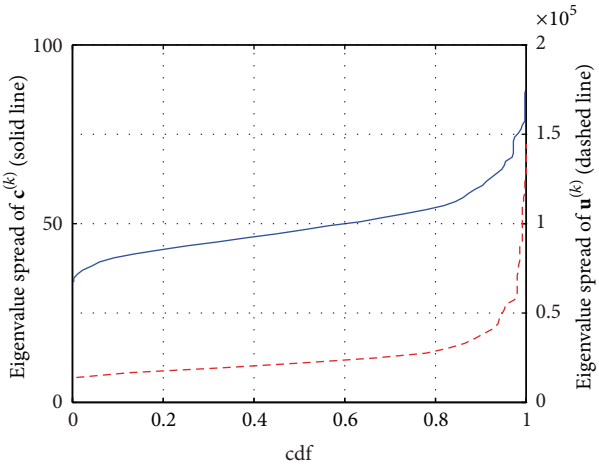


Figure 6: Cumulative distribution function of eigenvalue spread (solid line: $\mathbf{u}^{(k)}$, dashed line: $\mathbf{c}^{(k)}$).

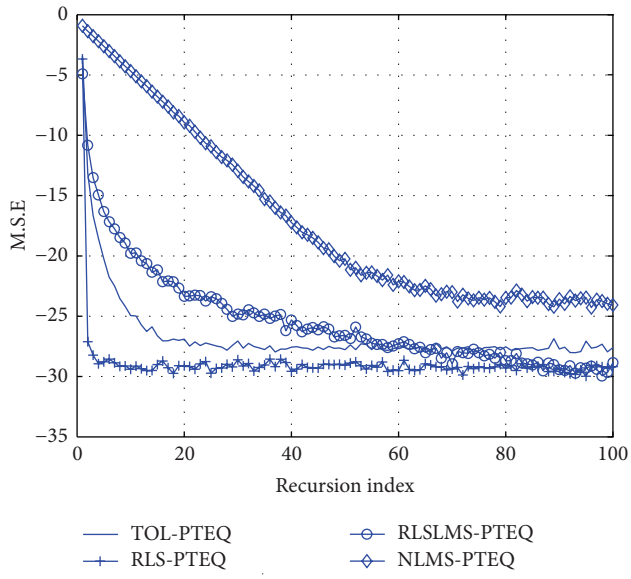


Figure 7: Comparison of M.S.E trajectories for PTEQs.

Maximum scalable OFDMA downlink of mobile WiMAX [18] is considered for this simulation, where signal bandwidth is 20 MHz, the FFT size is 2048, the number of pilot subcarriers is 240, and the number of data subcarriers N_u is 1440. Extended ITU Vehicular A channel model [19] with additive white Gaussian noise is considered. The number of cyclic prefix is set to 64, where it is set to 256 in the standard [18]. This enhances spectral efficiency of about 10 percent compared with the standard in [18].

TOL-PTEQ is compared with NLMS-PTEQ, RLS-PTEQ, and RLSLMS-PTEQ. All simulations are performed over 1000 independent trials. All PTEQs have their order, T, as 16, step-size parameter μ in NLMS part is 1, forgetting factor in RLS part is 0.998, and β of TOL-PTEQ is 0.1.

Learning curves of the four PTEQs are shown in Figure 7. Among the four PTEQs, the convergence rate of RLS-PTEQ

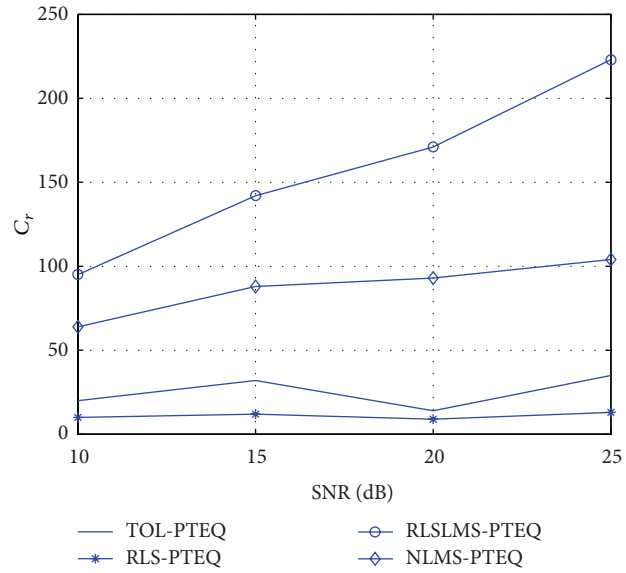


Figure 8: Comparison of C_r versus SNR for PTEQs.

is the fastest and its steady-state M.S.E is the smallest at the expense of the largest computational complexity as described in Section 3.1. NLMS-PTEQ has the smallest computational complexity, but its performance is the worst. RLSLMS-PTEQ converges to the same M.S.E as RLS-PTEQ; however its convergence rate is very slow which is undesirable to the packet-based wireless communications. It can be seen that TOL-PTEQ, in spite of its slightly higher steady-state M.S.E compared with those of the two RLS-typed PTEQs, has very fast convergence rate with the smaller computational complexity than the two RLS-typed PTEQs.

To see the more specific behavior of the convergence, we define the convergence rate C_r as the number of symbols that is required until the M.S.E reaches 98 percent of its steady-state M.S.E. In this analysis, NLMS-PTEQ is excluded because its convergence rate and steady-state M.S.E are of too low performance to be compared with other PTEQs.

Two factors, SNR and order of PTEQ tap-weights, are considered in this simulation analysis. It is desirable that C_r is maintained as small as possible regardless of these factors. Figure 8 shows the C_r depending on SNR, where SNR is ranged from 10 dB to 25 dB. C_r of TOL-PTEQ and RLS-PTEQ are small and independent of the variation of SNR, while C_r of RLSLMS-PTEQ increases as SNR increases. This shows that convergence rate of TOL-PTEQ does not deteriorate at low SNR. In Figure 9, C_r according to the variation of tap-weight order T are shown, where $2 \leq T \leq 30$. It is shown that TOL-PTEQ keeps C_r small with little dependency on T. Hence TOL-PTEQ maintains small values of C_r over simulated SNR and T which is a desirable feature of PTEQ designs.

5. Conclusions and Future Works

We analyzed convergence of several existing PTEQs. Then, we showed the TOL-PTEQ that has enhanced convergence rate with small computational complexity. We also provided

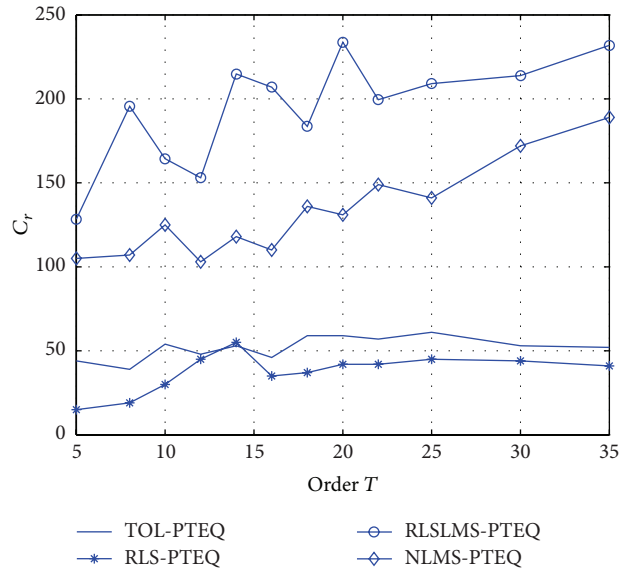


FIGURE 9: Comparison of C_r versus T for PTEQs.

the convergence model of TOL-PTEQ, and it was shown to be accurate by the comparison between the model and simulation results. In the comparison of M.S.E learning curves of the three PTEQs for the mobile WiMAX with the extended ITU Vehicular A channel model, it was shown that TOL-PTEQ has the slower convergence rate than RLS-PTEQ, but its convergence was stabilized within the size of cyclic prefix. It was also shown that TOL-PTEQ had desirable features that its convergence rate was rarely dependent on either SNR or tap-weight order T. TOL-PTEQ can be applied to the radar channel estimation such as long impulse responses with short cyclic prefix. In our future work, we will extend TOL-PTEQ to the multitone based radar systems.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by MKE/ISDK (Development of Mobile Safety-Inspection Systems Using High Resolution Penetration Imaging Technology for Transportation Infrastructure).

References

[1] K. van Acker, G. Leus, M. Moonen, O. van de Wiel, and T. Pollet, "Per tone equalization for DMT-based systems," *IEEE Transactions on Communications*, vol. 49, no. 1, pp. 109–119, 2001.

[2] K. van Acker, G. Leus, M. Moonen, and T. Pollet, "RLS-based initialization for per-tone equalizers in DMT receivers," *IEEE Transactions on Communications*, vol. 51, no. 6, pp. 885–889, 2003.

[3] S. Sitjongsataporn and P. Yuvapoositanon, "Bit rate maximising per-tone equalisation with adaptive implementation for dmt-based systems," *Eurasip Journal on Advances in Signal Processing*, vol. 2009, Article ID 380560, 2009.

[4] G. Arslan, B. L. Evans, and S. Kiaei, "Equalization for discrete multitone transceivers to maximize bit rate," *IEEE Transactions on Signal Processing*, vol. 49, no. 12, pp. 3123–3135, 2001.

[5] J. Zhang, C.-D. Wu, Y.-Z. Zhang, and P. Ji, "Energy-efficient adaptive dynamic sensor scheduling for target monitoring in wireless sensor networks," *ETRI Journal*, vol. 33, no. 6, pp. 857–863, 2011.

[6] S. -M. Yoo and P. H. Chou, "MHP: Master-Handoff protocol for fast and energy-efficient data transfer over SPI in wireless sensing systems," *ETRI Journal*, vol. 34, no. 4, pp. 553–563, 2012.

[7] Y. Geum, C. Kim, S. Lee, and M.-S Kim, "Technological convergence of IT and BT: evidence from patent analysis," *ETRI Journal*, vol. 34, no. 3, pp. 439–449, 2012.

[8] H. Alemdar and C. Ersoy, "Wireless sensor networks for healthcare: a survey," *Computer Networks*, vol. 54, no. 15, pp. 2688–2710, 2010.

[9] W. Han, A. T. Erdogan, T. Arslan, and M. Hasan, "High-performance low-power FFT cores," *ETRI Journal*, vol. 30, no. 3, pp. 451–460, 2008.

[10] D. K. Kim, Y. W. Choi, and D. W. Kang, "Feasibility study of multi-carrier ground penetrating radar technology," in *Proceedings of the 4th International Conference on Next Generation Information Technologies*, pp. 715–720, June 2013.

[11] D. Kim and G. L. Stüber, "Residual ISI cancellation for OFDM with applications to HDTV broadcasting," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1590–1599, 1998.

[12] A. J. Lowery, "Reducing cyclic prefix overhead in optical OFDM systems," in *Proceedings of the 35th European Conference on Optical Communication (ECOC '09)*, Vienna, Austria, September 2009.

[13] N. Al-Dhahir, "Optimum finite-length equalization for multicamer transceivers," *IEEE Transactions on Communications*, vol. 44, no. 1, pp. 56–64, 1996.

[14] R. Baldemair and P. Frenger, "A time-domain equalizer minimizing intersymbol and intercarrier interference in DMT systems," in *Proceedings of the IEEE Global Telecommunicatins Conference (GLOBECOM '01)*, vol. 1, pp. 381–385, San Antonio, Tex, USA, November 2001.

[15] K. van Acker, G. Leus, M. Moonen, and T. Pollet, "RLS-based initialization for per-tone equalizers in DMT receivers," *IEEE Transactions on Communications*, vol. 51, no. 6, pp. 885–889, 2003.

[16] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 3rd edition, 1996.

[17] M. L. Honig and D. G. Messerschmitt, "Convergence properties of an adaptive digital lattice filter," *IEEE Transactions on Circuits and Systems*, vol. 28, no. 6, pp. 482–493, 1981.

[18] WiMAX Forum, *Mobile WiMAX-Part I: A. Technical Overview and Performance Evaluation*, 2006.

[19] T. B. Sorensen, P. E. Mogensen, and F. Frederiksen, "Extension of the ITU channel models for wideband OFDM systems," in *Proceedings of the 62nd IEEE Vehicular Technology Conference (VTC-Fall '05)*, vol. 1, pp. 392–396, Dallas, Tex, USA, September 2005.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2013, Article ID 473609, 6 pages
http://dx.doi.org/10.1155/2013/473609

Research Article

Selective Cooperative Transmission in Ad Hoc Networks with Directional Antennas

Eui-Jik Kim[1] and Sungkwan Youm[2]

[1] *Department of Ubiquitous Computing, Hallym University, 39 Hallymdaehak-gil, Chuncheon-si, Gangwon-do, 200-702, Republic of Korea*
[2] *Telecommunication Systems Division, Samsung Electronics Co, Ltd., Dong Suwon P.O. Box 105, 416 Maetan-3dong, Yeongtong-gu, Suwon-si, Gyeonggi-do, 443-742, Republic of Korea*

Correspondence should be addressed to Sungkwan Youm; skyoum@gmail.com

Received 24 September 2013; Accepted 10 November 2013

Academic Editor: Ken Choi

This paper presents a selective cooperative transmission scheme (abbreviated SCT) for ad hoc network with directional antennas that leverages the benefits of directional-only antenna approach and cooperative communication. The main feature of SCT is its adaptability to the channel condition in the network. In other words, when the node sends data, SCT determines its transmission strategy on either direct or cooperative transmission via a relay node called a forwarder, depending on the transmission time. Simulation results are provided to validate the effectiveness of the proposed scheme.

1. Introduction

The use of directional antennas in wireless ad hoc networks has advantages such as high antenna gain, high spatial reuse, and extended transmission range. Especially, it is indispensable in the emerging millimeter-wave (mmWave) systems, which operate in the license-free frequency band from 57 to 64 GHz, because the communication signals at 60 GHz suffer from high path loss [1]. However, the asynchronous feature for the antenna beam directions and transmission ranges between the nodes causes new challenge such as deafness problem, since the nodes do not have prior knowledge about their neighbors.

In order to resolve the deafness problem, Choudhury et al. [2] and Nasipuri et al. [3] present the antenna mode selection scheme, in which the node chooses either the omnidirectional or directional antenna modes, depending on the received control message. This approach definitely alleviates the deafness problem, but maintaining both types of antennas not only induces a high cost but also results in the asymmetric-in-gain problem, which is another form of the deafness problem caused by the different antenna transmission range sizes of the two nodes. To overcome the asymmetric-in-gain problem, Shihab et al. [4] and Jakllari et al. [5] present a directional-only antenna approach where the nodes operate with a single directional antenna. However, their performance can be significantly degraded, due to the high control message overhead accrued for searching neighbors.

Liu et al. [6] and Zhu and Cao [7] propose a two-hop relay transmission approach for ad hoc networks with the omnidirectional-only antenna. They improve the network performance by reducing the transmission time of data packets via a concept of cooperative transmission. However, under high path loss conditions such as mmWave systems, they still suffer from the disadvantages of omnidirectional antennas.

In this paper, we present a selective cooperative transmission scheme (abbreviated SCT) for ad hoc network with directional antennas that leverages the benefits of directional-only antenna approach and cooperative communication. The main feature of SCT is its adaptability to the channel condition in the network. In other words, when the node sends data, SCT determines its transmission strategy on either direct or cooperative transmission via a relay node called a forwarder, depending on the transmission time.

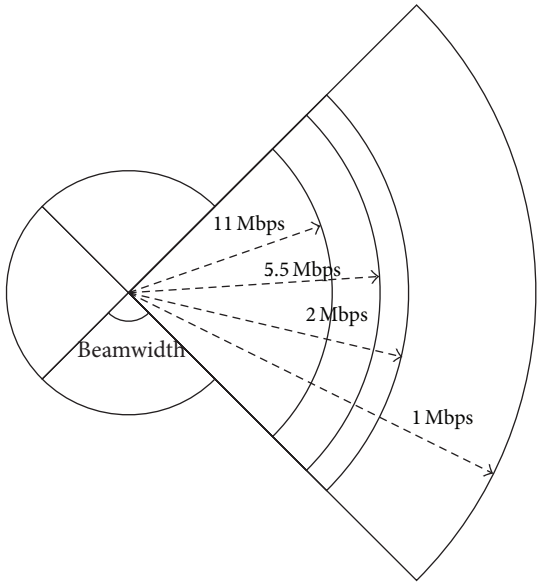


Figure 1: Directional antenna model supporting multiple data rates.

Simulation results are provided to validate the effectiveness of the proposed scheme.

This paper is organized as follows: Section 2 gives an overview of the directional antenna model and the neighbor discovery procedure. SCT scheme is described in Section 3. The performance evaluations of SCT are presented in Section 4. The conclusions are given in Section 5.

2. System Model

2.1. Directional Antenna. In our work, we consider the distributed wireless network where every node is equipped with a single directional antenna. Figure 1 shows a directional antenna model supporting multiple data rates. For simplicity of analysis, we employ the flat-top antenna model, where the antenna gain is a constant within the beamwidth and zero outside the beamwidth. Therefore, for a beam with beamwidth θ, the antenna gains of the mainlobe and sidelobe are given by $G_m = 2\pi/\theta$ and $G_s = 0$, respectively. Moreover, the radius of each sector, namely, transmission range, can be calculated as follows:

$$r = 10^{K/10n}, \quad K = P_t + G_t + G_r - I_L - PL_0 - \gamma, \qquad (1)$$

where P_t is the transmission power and G_t and G_r are the antenna gains of the transmitter and receiver, respectively. I_L is the implementation loss. PL_0 is the reference path loss at 1 meter and n is the path loss exponent. γ is the receiver sensitivity.

2.2. Neighbor Discovery. Due to the asynchronous feature of beam direction in the different nodes, neighbor discovery procedure has to be considered. In SCT, the directional network allocation vector (DNAV) [2] and the one-way neighbor discovery (one-way ND) [8] are used for neighbor discovery. With DNAV, the nodes keep a record of the ongoing transmissions by their neighbors in each direction. As a node starts up, it first runs the one-way ND procedure to create the neighbor table, whose entries include the sector number, node ID, distance, sequence number, DNAV status, and sector-switching timing information. Note that the sector-switching timing information indicates the time difference between its sector-switching schedule and neighbor's one [9]. In one-way ND, each node in the network periodically transmits an advertisement message to announce its presence and discovers its neighbors by receiving advertisement messages from the other nodes. Once a node receives a neighbor's advertisement message, it caches the information in its neighbor table. If a node receives a previously cached advertisement message, it updates the corresponding fields of the table. Consequently, all the nodes in the network can sustain the latest information for their neighbors via this preliminary procedure.

3. Design of SCT

SCT is inherently a cooperative communication solution under a single directional antennas condition. In SCT, high data rate nodes, called forwarders, assist the transmission of low data rate nodes by forwarding their traffic. Major components of SCT design are the mechanism for each node to learn about candidate forwarder nodes, and the corresponding data structures, called a SctTable, used to store the information related to those identified candidates. Using SCT, the node in the network can choose a forwarder from this list of potential forwarders to use at the time of its transmissions, depending on the possibility of reducing the transmission time for the packet. In the following, we present the operation of the SCT, in detail.

3.1. Forwarder Selection. Each node in the network should maintain a table, referred to as the SctTable, whose entries include node ID, sequence number, R_{FR} (the data rate between the forwarder candidate and the receiver), and R_{SR} (the data rate between the sender and the receiver). Note that the specific values of SctTable entries are updated to reflect the current channel conditions and retrieved from the neighbor table, which is earlier described in Section 2.2.

Every node within the antenna sector coverage of the sender toward the receiver can be a forwarder node, except for those that set DNAV. To select the forwarder, the sender checks its neighbor table and creates the entries of SctTable. Note that if one sets DNAV, it is deleted from the SctTable entries. The sender finds that the nodes belong to the sector including the receiver by referring to the sector number field of the neighbor table and then brings them to the SctTable entries. R_{FR} and R_{SR} can be obtained from the distance field of neighbor table. The SctTable entry that minimizes the total transmission time is selected as a forwarder. If two or more nodes can be selected as forwarders, the sender chooses the latest one, by referring to the sequence number field. The transmission time can be calculated as follows:

$$E[T_{TX}] = \frac{8L}{R_{\mathrm{SF}}} + \frac{8L}{R_{\mathrm{FR}}} + T_{\mathrm{ACK}} + 3T_{\mathrm{SIFS}} + E[T_{\mathrm{OH\text{-}C}}], \qquad (2)$$

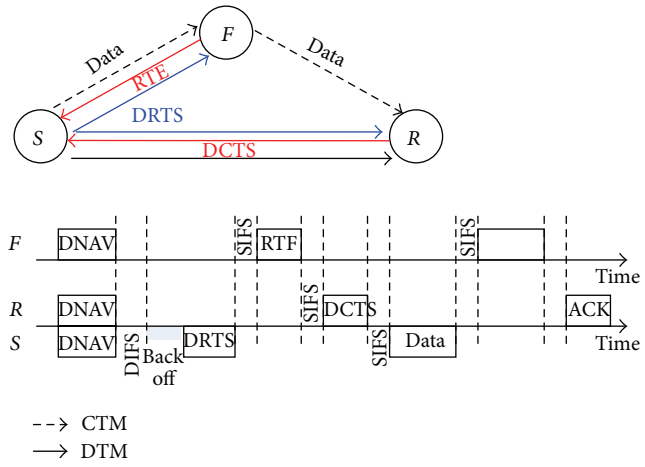


Figure 2: Exchange of control messages.

where L is the data size in octets, R_{SF} is the data rate between sender and forwarder, R_{FR} is the data rate between forwarder and receiver, and $E[T_{OH\text{-}C}]$ denotes the average overhead time for the cooperative transmission.

3.2. Transmission Mode Determination. As mentioned above, we use the concept of cooperative communication to improve the network performance by reducing the transmission time of data packets, under a single directional antenna condition. The idle nodes scan a sector for a specific time period and continuously switch their beams in a clockwise or anticlockwise direction in order to circumvent the deafness problem. The basic exchange procedure of control messages in SCT is shown in Figure 2. Assuming that the number of antenna sectors of every node is M, DRTS messages are repeatedly sent to the receiver up to M times. Note that, the scanning time duration for all sectors is equal to the time from the sending moment of the first DRTS to that of the Mth DRTS. Cooperative transmission mode (CTM) via a forwarder is selected if it satisfies

$$\frac{8L}{R_{SR}} + E\left[T_{OH\text{-}D}\right] > \frac{8L}{R_{SF}} + \frac{8L}{R_{FR}} + T_{SIFS} + E\left[T_{OH\text{-}C}\right], \quad (3)$$

where $E[T_{OH\text{-}D}]$ is the average overhead time for direct transmission mode (DTM). Otherwise, the sender uses DTM. Once the transmission mode is determined, the sender fixes its antenna beam direction toward the receiver and starts sending the DRTS. Under directional antennas condition, due to the asynchronous beam feature, more control packets are required to establish the directional link. In the worst case, the sender might transmit M DRTSes until the receiver receives a DRTS successfully. Due to the transmission of these multiple DRTSes, an additional backoff procedure has to be considered. Thus, we employ the backoff scheme similar to that used in [4], in which the average backoff times for DTM and CTM are, respectively, given by

$$E\left[T_{BO\text{-}D}\right] = \frac{1}{M^2}\sum_{i=1}^{M}\frac{iCW_{\max}T_{slot}}{2},$$
$$E\left[T_{BO\text{-}C}\right] = \frac{1}{M^3}\sum_{i=1}^{M}\frac{(2i-1)\,CW_{\max}T_{slot}}{2}. \quad (4)$$

The DRTS includes the information on the forwarder, receiver, and payload size. The forwarder overhears the DRTS while the sender sends multiple DRTSes. Upon overhearing the DRTS, the forwarder sends a Ready-To-Forward (RTF) message to the sender. When the receiver receives the DRTS, it responses with the DCTS and then fixes its beam direction toward the forwarder to receive the data packet. If the sender receives both the RTF and DCTS messages, it first sends the data packet to the forwarder, which subsequently transfers it to the receiver. If nonparticipating idle nodes overhear the control packet, DNAV for the transmission duration is set up.

In (3), $E[T_{OH\text{-}C}]$ can be expressed by using $E[T_{CS}]$ and $E[T_{CF}]$, which are the average times for successful transmission and failed transmission in CTM, respectively, given by

$$E\left[T_{CS}\right] = E\left[T_{BO\text{-}C}\right] + T_{DRTS} + 2T_{SIFS} + T_{RTF} + T_{DCTS} + T_{DIFS},$$
$$E\left[T_{CF}\right] = E\left[T_{BO\text{-}C}\right] + T_{DRTS} + T_{DIFS}, \quad (5)$$

where $E[T_{CS}]$ is the average time for initiating communication with only one DRTS. Therefore, the average time for initiating communication with M DRTSes is given by $E[T_{CS}] + (M-1)E[T_{CF}]$. Then, we can easily infer $E[T_{OH\text{-}D}]$ as follows:

$$E\left[T_{OH\text{-}D}\right] = E\left[T_{DS}\right] + \frac{(M-1)\,E\left[T_{DF}\right]}{2}, \quad (6)$$

where

$$E\left[T_{DS}\right] = E\left[T_{BO\text{-}D}\right] + T_{DRTS} + T_{SIFS} + \mathrm{T}_{DCTS} + \mathrm{T}_{DIFS},$$
$$E\left[T_{DF}\right] = E\left[T_{BO\text{-}D}\right] + T_{DRTS} + T_{DIFS}. \quad (7)$$

The average overhead time in CTM is longer than that in DTM, because the sender has to receive the responses for the DRTSes from both forwarder and receiver. During the unit scanning time for one sector, one DRTS can be transmitted on average, since the scanning time for $M sectors is equal to the transmission time for M DRTSes. Thus, in order to receive responses from both nodes (i.e., the forwarder and receiver) for a DRTS transmission, the antenna beams of both nodes have to point toward the sender before the DRTS transmission is completed. Note that the probability that an idle node selects a sector is $1/M$, and the probability that two nodes select the same sector simultaneously is $1/M^2$. As the number of DRTSes needed to initiate the communication increases, the number of cases that two nodes select the sector

Table 1: Simulation parameters.

Parameter	Value	Parameter	Value
Simulation area	$300 \times 300\ m^2$	Slot time	20 μsec
MAC header	28 Octets	DIFS	50 μsec
DRTS	30 Octets	SIFS	10 μsec
DCTS, RTF	18 Octets	aCWMin, aCWMax	31 slots, 64 slots

Table 2: Antenna transmission range.

Data rate	11 Mbps	5.5 Mbps	2 Mbps	1 Mbps
4 sectors	96.8 m	133.6 m	149.9 m	199.9 m
6 sectors	153.4 m	211.7 m	231.8 m	267.7 m
8 sectors	234.5 m	275.5 m	291.8 m	337 m

corresponding to the direction of the sender at the same time also increases. So, we can obtain $E[T_{\text{OH-C}}]$ as follows:

$$E\left[T_{\text{OH-C}}\right] = E\left[T_{\text{CS}}\right] + \frac{E\left[T_{\text{CF}}\right]}{M^3}\sum_{i=1}^{M}(i-1)(2i-1). \quad (8)$$

4. Simulation Results

We evaluated the performance of the SCT through experimental simulations using the QualNet 5.0 simulator [10]. We created the azimuth file shown in Figure 3 to implement the flat-top directional antenna model. Common simulation parameters are listed in Table 1. Note that, due to the limitations of the simulator that does not support mmWave transmissions, we implement our work on top of the IEEE 802.11b physical layer model.

4.1. Performance Study in DTM. In ad hoc networks that support multiple data rates, as the data rate or the number of antenna sectors increases, the antenna transmission range decreases. The relation between the data rates/the number of sectors and the transmission ranges is shown in Table 2. So, it can be inferred that the narrow beamwidth of the sector makes its gain relatively higher. To analyze the throughput performance for various numbers of sectors, we perform a simple experiment, in which the packet payload size and the arrival rate are 1250 bytes and 200 packets/sec, respectively. Figure 4 shows the throughput for varying distances between sender and receiver. In the case of small number of sectors (e.g., $M = 2$), the nodes can transmit the data and maintain high throughput in only a short distance. On the other hand, in the case of large number of sectors (e.g., $M = 8$), they shows the opposite behavior. Note that at the same distance value, the latter case exhibits lower throughput than the former case. From this, we can infer that as the number of sectors increases, the messaging overhead (e.g., DRTS, DCTS) also increases.

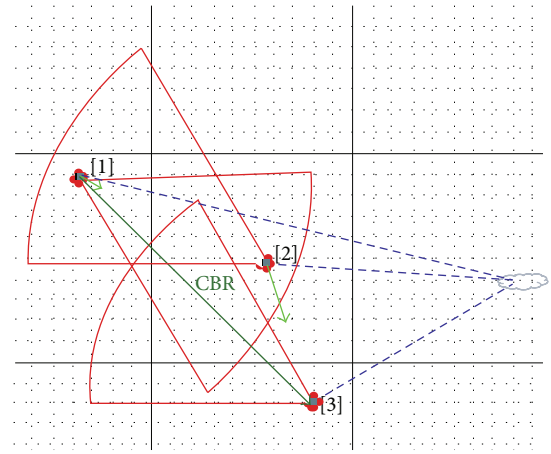


Figure 3: Flat-top directional antenna model in Qualnet 5.0.

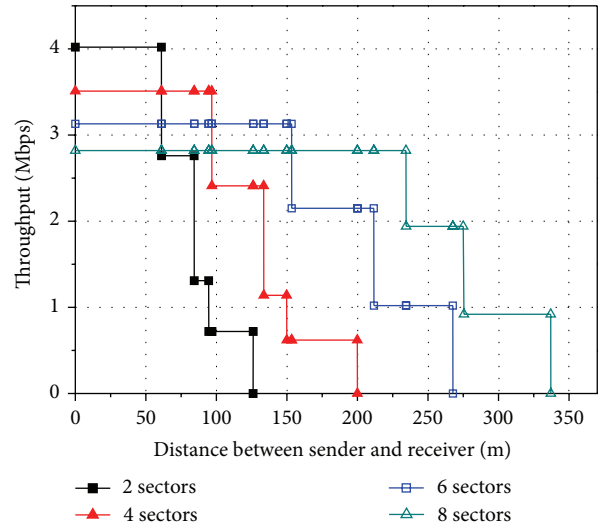


Figure 4: Throughput as a function of distance between sender and receiver.

4.2. Performance Comparison of DTM and CTM. The throughput is expressed as the average payload size during the average transmission time of a data packet and is given by

$$TH = \frac{P_s P_{tr} E\left[L\right]}{\left(1-P_{tr}\right)T_{\text{slot}} + P_s P_{tr} E\left[T_s\right] + P_{tr}\left(1-P_s\right)E\left[T_f\right]}, \quad (9)$$

where T_{slot}, P_s, P_{tr}, $E[L]$, $E[T_s]$, and $E[T_f]$ denote the slot time, the probability of successful transmission, the probability that there exists at least one transmission in the slot time, the average payload size, the average time for successful transmission, and the average time for failed transmission. In order to obtain the throughput in CTM, the nodes' sector number and the multiple DRTSes should be considered.

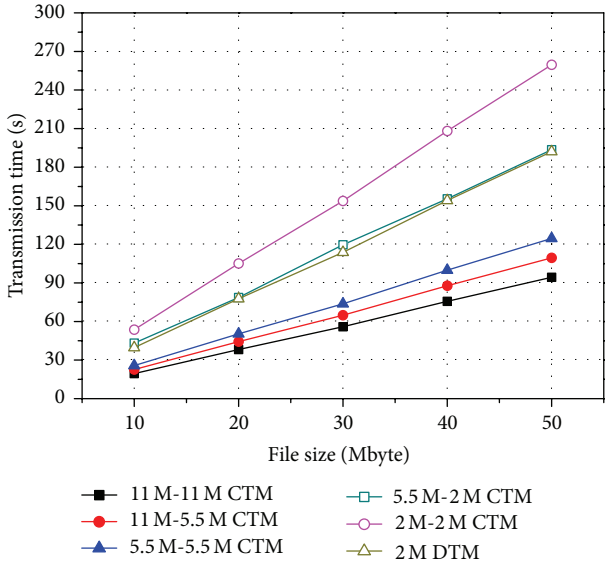


Figure 5: Transmission time with various data rates.

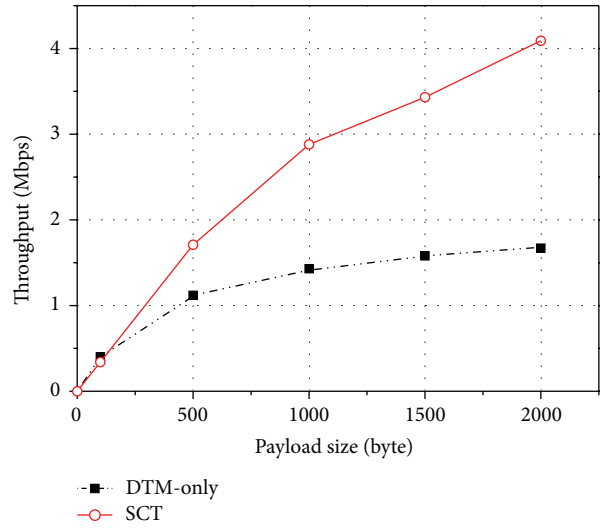


Figure 6: Throughput as a function of payload size.

Since the nodes in the network are assumed to be randomly deployed, we can obtain $E[T_s]$ as

$$E[T_s] = f_{11}T_{11} + f_{5.5}T_{5.5} + f_2T_2 + f_1T_1 + T_{\text{ACK}} + 3T_{\text{SIFS}} + E[T_{\text{OH-C}}], \quad (10)$$

where f_x is the probability that the forwarder is located in the x Mbps area from the sender and T_x is the transmission time when the nodes use the data rate of x Mbps. $E[T_f]$ is given by $M \times E[T_{\text{CF}}]$. To evaluate the transmission time, we set the following parameters; the packet arrival rate is 1000 packets/s, the number of sectors is 4, the distance between the sender and receiver is 145 meters, and the packet payload size is 1250 bytes. The sender can transmit its data to the receiver directly with a 2 Mbps data rate.

Figure 5 shows the transmission time for various transmission modes. Since the node can use various data rates according to the position of the forwarder in CTM, we change the forwarder's position from the 2 Mbps area to the 11 Mbps area in the experiment. In the figure, the case of CTM where both sender and receiver use the 11 Mbps link exhibits the shortest transmission time. In DTM, the node takes less time for the transmission of data packets than CTM with the 2 Mbps-2 Mbps, because the overhead time for exchanging the control messages is relatively small in DTM. Figure 6 shows the throughput performance for various payload sizes. Overall, SCT exhibits higher network throughput than the DTM stand-alone solution, because SCT can be supported by the high data rate link (e.g., 11 Mbps link). However, in the cases that the payload size is below 100 bytes, the throughput of the DTM-only solution is slightly higher than SCT, due to the control message overhead.

Figure 7 shows the performance comparison of average overhead time for initiating the cooperative transmissions. In directional-only antenna approach, each node may have different sector-switching timing; thus the overhead time

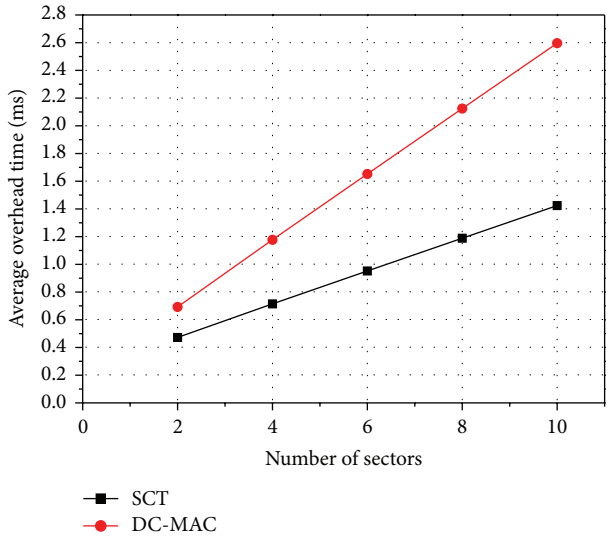


Figure 7: Comparison of average overhead time.

for initiating cooperative transmission is mainly affected by sector-scanning time of the node, which is the time that the sender needs to transmit one or more DRTS messages at each sector for trying cooperative transmission with the receiver. We measure the overhead time of SCT and DC-MAC [11] for varying numbers of sectors. On the whole, SCT exhibits better performance compared to DC-MAC. In SCT, it is assumed that the advertisement message of neighbor discovery procedure contains the information for sector-switching timing; thus the sender can initiate cooperative transmission with one DRTS message. On the other hand, in DC-MAC, the node sends two DRTS messages at each sector to keep its high probability of successful transmission, which leads the increase of average sector-scanning time.

5. Conclusion

This paper presents SCT, which is a selective cooperative transmission scheme for ad hoc network with directional antennas. Under directional antenna only environments, SCT improves the network performance by reducing the transmission time of data packets through adopting selectively either CTM or DTM. The simulation results verify that the proposed scheme exhibits high network performance in terms of both the throughput and transmission time.

Acknowledgments

This research was supported by Hallym University Research Fund, 2013 (HRF-201309-002). A preliminary version of this paper [11] appeared in IEEE ICOIN, Feb. 1–3, 2012, Bali, Indonesia. This version includes an updated protocol design and its performance evaluation results.

References

[1] K. Shin, Y. Kim, and C.-H. Kang, "Adaptive directional multicast scheme in mm wave WPANs with directional antennas," *IEICE Transactions on Communications*, vol. E95-B, no. 5, pp. 1834–1838, 2012.

[2] R. Choudhury, X. Yang, R. Ramanathan, and N. H. Vaidya, "Using directional antennas for medium access control in ad hoc networks," in *Proceedings of The 8th Annual International Conference on Mobile Computing and Networking (MOBICOM '02)*, pp. 59–70, September 2002.

[3] A. Nasipuri, S. Ye, J. You, and R. E. Hiromoto, "A MAC protocol for mobile ad hoc networks using directional antennas," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '00)*, pp. 1214–1219, September 2000.

[4] E. Shihab, L. Cai, and J. Pan, "A distributed asynchronous directional-to-directional MAC protocol for wireless ad hoc networks," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 9, pp. 5124–5134, 2009.

[5] G. Jakllari, L. Wenjie, and S. V. Krishnamurthy, "An integrated neighbor discovery and MAC protocol for ad hoc networks using directional antennas," *IEEE Transactions on Wireless Communications*, vol. 6, no. 3, pp. 1114–1124, 2007.

[6] P. Liu, Z. Tao, S. Narayanan, T. Korakis, and S. S. Panwar, "CoopMAC: a cooperative MAC for wireless LANs," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 2, pp. 340–353, 2007.

[7] H. Zhu and G. Cao, "rDCF: a relay-enabled medium access control protocol for wireless ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 5, no. 9, pp. 1201–1214, 2006.

[8] X. An, R. Venkatesha Prasad, and I. Niemegeers, "Neighbor discovery in 60 GHz wireless personal area networks," in *Proceedings of the IEEE International Symposium on "a World of Wireless, Mobile and Multimedia Networks" (WoWMoM '10)*, pp. 1–8, June 2010.

[9] E. Felemban, R. Murawski, E. Ekici et al., "SAND: sectored-antenna neighbor discovery protocol for wireless networks," in *Proceedings of the 7th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON '10)*, pp. 1–9, June 2010.

[10] http://www.scalable-networks.com/.

[11] J.-H. Kwon, E.-J. Kim, H. Park, S.-H. Lee, and C.-H. Kang, "DC-MAC: directional cooperative MAC for ad-hoc networks," in *Proceedings of the International Conference on Information Networking (ICOIN '12)*, pp. 19–24, February 2012.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2013, Article ID 639325, 6 pages
http://dx.doi.org/10.1155/2013/639325

Research Article

Hybrid MAC Scheme for Vehicular Communications

Woong Cho

Department of Computer System Engineering, Jungwon University, 85 Munmu-ro, Goesan-eup, Goesan-gun, Chungbuk 367-805, Republic of Korea

Correspondence should be addressed to Woong Cho; wcho@jwu.ac.kr

Received 19 August 2013; Accepted 15 October 2013

Academic Editor: Jongsung Kim

Vehicular communications have been creating new applications in intelligent transportation systems (ITS) areas by converging information and communication technology (ICT) with automobile and road industries. In this paper, we introduce hybrid MAC scheme for vehicular communications which enhances the performance of IEEE 802.11p based communication system. The proposed MAC scheme supports high throughput by combining carrier sensing multiple access/collision avoidance (CSMA/CA) with TDMA. The benefits of proposed schemes are verified by simulations. In addition, we discuss some implementation issues including several application scenarios.

1. Introduction

Vehicular communications have been creating various communication services by combining communication technologies with existing automobile industry. There are two categories of vehicular communications, which are vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications. V2V application can build vehicle multihop networking for transmitting safety messages, in which each vehicle detects and transmits emergency messages to neighboring vehicles by multihop communications. V2I application can adopt bidirectional communications for telematics and ITS, where both vehicle and infrastructure can request and receive traffic information or other data such as probe data, toll information, and multimedia data.

By focusing on vehicular safety, standards for V2V and V2I communications have been developed in the 5.9 GHz frequency band, and the representative standard is referred to as wireless access in vehicular environments (WAVE) which consists of IEEE 802.11p for PHY/MAC layers and IEEE 1609 family for higher layers. The core structure of PHY/MAC layers in WAVE standard originates from the orthogonal frequency division multiplexing (OFDM) PHY layer in IEEE 802.11 standard [1]. To support reliable packet transmission in vehicular environments, some modifications have been adopted from wireless local area network (WLAN) technique. The performance of vehicular communication has been analyzed extensively in [2–5]. Recently, V2V-based vehicle ad hoc network (VANET) has attracted much attention for vehicle safety applications by automobile industry because of its standalone vehicle networking benefit without connecting infrastructure network. WAVE technology may be a good candidate for VANET. However, its drawback in terms of packet transmission delay renders the performance of WAVE insufficient in vehicular communications. WAVE adopts CSMA/CA MAC protocol, which has good throughput in the case that the number of nodes is small. However, as the number of nodes increases, the throughput decreases dramatically, which results in long packet transmission delay. To overcome this drawback, we propose hybrid MAC scheme which combines CSMA/CA and TDMA.

The rest of this paper is organized as follows. We first describe the overall system architecture and hybrid MAC scheme in Section 2. In Section 3, the advantages of the proposed scheme are verified with simulations. In Section 4, some application scenarios are discussed with practical issues for implementation. Finally, the conclusion remarks are given in Section 5.

2. System Configuration

We consider the communication systems with vehicles and road side equipment (RSE)s. Each vehicle is equipped with a vehicle terminal and on-board unit (OBU) which consists

Table 1: General features and basic requirements of the communication systems.

General features	Basic requirements
RF frequency: 5.9 GHz	Communications: broadcasting, unicasting
Channel bandwidth: 10 MHz	Latency: Less than 100 msec
Modulation: OFDM (BPSK, QPSK, 16QAM, 64QAM)	Networking: V2V/V2I (I2V)
MAC protocol: CSMA/CA, EDCA	Up to the speed of 200 km/h

of modem, MAC, and radio frequency (RF) module. Vehicle terminal displays communication status, vehicle status, and other information such as GPS. Vehicle information can be collected through information gathering device where engine control unit/transmission control unit/microcontrol unit (ECU/TCU/MCU) provides detailed vehicle status. In the road side, an infrastructure has RSE which also has communication module, and RSE is connected to the server and ITS center via backbone network. With above communication system architecture, our system has the following general features and satisfies basic requirements for various safety related applications [6] as we indicated in Table 1.

It is worth stressing that the above features satisfy IEEE 802.11p protocol in [7]. In addition to the above features, we propose hybrid MAC with priority control in this paper.

VANET has short lifetime of communication link and dynamic network topology due to the rapid changes of vehicle traffic. The main application areas of vehicular communications are safety services. Most of safety services require point-to-multipoint communication scheme rather than point-to-point communication since safety messages are aimed at general drivers not a specific driver [8]. Therefore, reliable broadcasting plays an important role in safety services. Currently, CSMA/CA-based distributed coordination function (DCF) has been considered for a standard MAC scheme [6, 7]. However, the throughput of this protocol decreases as the number of user increases, as shown in [9, 10]. To diminish this drawback, we propose hybrid MAC which combines TDMA and CSMA/CA.

Figure 1(a) shows the channel structure of the proposed scheme. Unlike CSMA/CA scheme, channel is composed of time frames, and that time frame is composed of time slots where the number of time slots in a time frame can be configured. To manage time frames and time slots, synchronization is needed. In general, GPS can be used for synchronization, where GPS receivers typically provide a precise 1 pulse per second (PPS) UTC signal (with an error less than 100 ns), and these precise 1 PPS signals can be used for timing and synchronization. In WAVE, the guard interval (3 msec) and channel switching time (maximum: 1 msec) are specified for multichannel operation in IEEE Std 1609.4. In addition to this time, DCF of the proposed scheme enable to use GPS in the proposed MAC scheme.

Based on the aforementioned channel structure, every node selects the time slot on which node has ownership and has transmission priority on that time slot at the beginning of each time frame. Then, given the time frame, each node becomes a slot owner and a nonslot owner for the time slot which has been chosen by the selection algorithm and the time slot for the other slots, respectively. The slot owner has transmission priority over nonslot owner. Notice that, in TDMA scheme, each node monopolizes the assigned time slot. However, multiple nodes can be slot owners of the specific time slot in the proposed scheme. In this case, each slot owner competes with the other slot owner using DCF scheme, and the node with the smallest backoff counter transmits the frame. Hence, there is no guarantee that the slot owner must have opportunity to transmit the frame in the time slot selected by node.

Basically, the proposed MAC scheme adopts random back-off process of DCF. The difference is that the proposed scheme adaptively assigns the contention window (CW) based on its slot ownership. Figure 1(b) shows the relationship of CW between the slot owner and the non-slot owner. The proposed MAC scheme uses different CW range for the slot owner and the non-slot owner. The slot owner has smaller CW range than the non-slot owner. Then, the slot owner has priority for frame transmission. As mentioned earlier, multiple slot owners can exist for the specific time slot. In this case, the slot owner that selects the smallest CW value transmits a frame. If the slot owner has no frame to transmit, the non-slot owner obtains the chance of transmit. Therefore, the proposed MAC scheme gives an opportunity to the non-slot owner which compensates a drawback of usability in TDMA. In the proposed MAC scheme, every node should select time slots in each time frame. The time slot selection algorithm must be needed and various manners can be possible. For example, random selection or selection based on network density can be used. In this paper, we use the random selection manner.

3. Simulations

In this section, we present the performance of proposed system with simulations. To evaluate the performance of hybrid MAC, we simulate and compare the throughput of IEEE 802.11 protocol and hybrid MAC using QualNet 4.0. For hybrid MAC, each time frame consists of four time slots, and each time slot has 10 ms of duration. A node is designed to select a time slot randomly, where the time slot has an ownership for each time frame. With this setup, we apply IEEE 802.11a PHY layer with 6 Mbps data rate for simulation. In MAC layer, we use DCF/hybrid MAC, and constant bit rate (CBR) is applied in application layer. For unicast, 20 pairs of unicast flow are considered using 40 nodes, where each pair consists of one source-destination link using two nodes. For broadcast, each node generates a broadcast frame by considering a maximum of 80 nodes. Figure 2 represents network setup for simulations.

Figure 3 represents throughput and delay of unicast depending on the number of flows. In Figure 3, hybrid MAC shows a little bit better performance than IEEE 802.11 protocol. However, the difference of performance is almost the same as the number of flows increases. Since the retransmission is possible in unicast, there is no big difference in the overall throughput. Figure 4 shows that throughput and delay

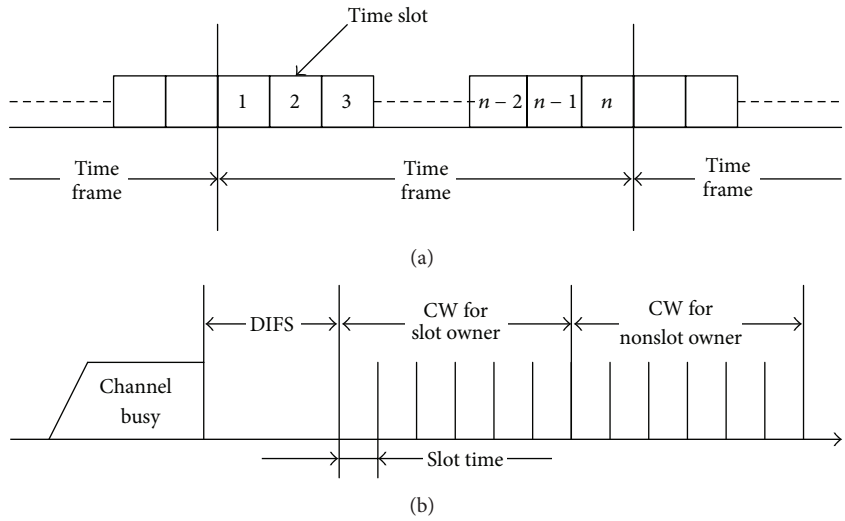


Figure 1: Hybrid MAC. (a) Structure of channel; (b) contention window of the hybrid MAC.

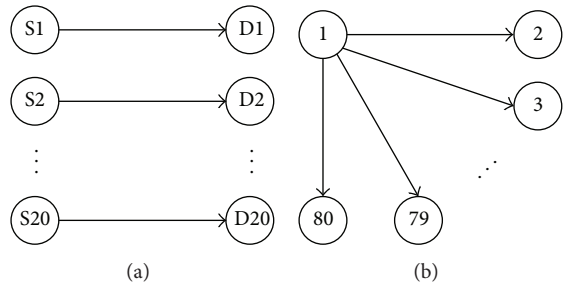


Figure 2: Network setup for simulations. (a) Unicast; (b) broadcast.

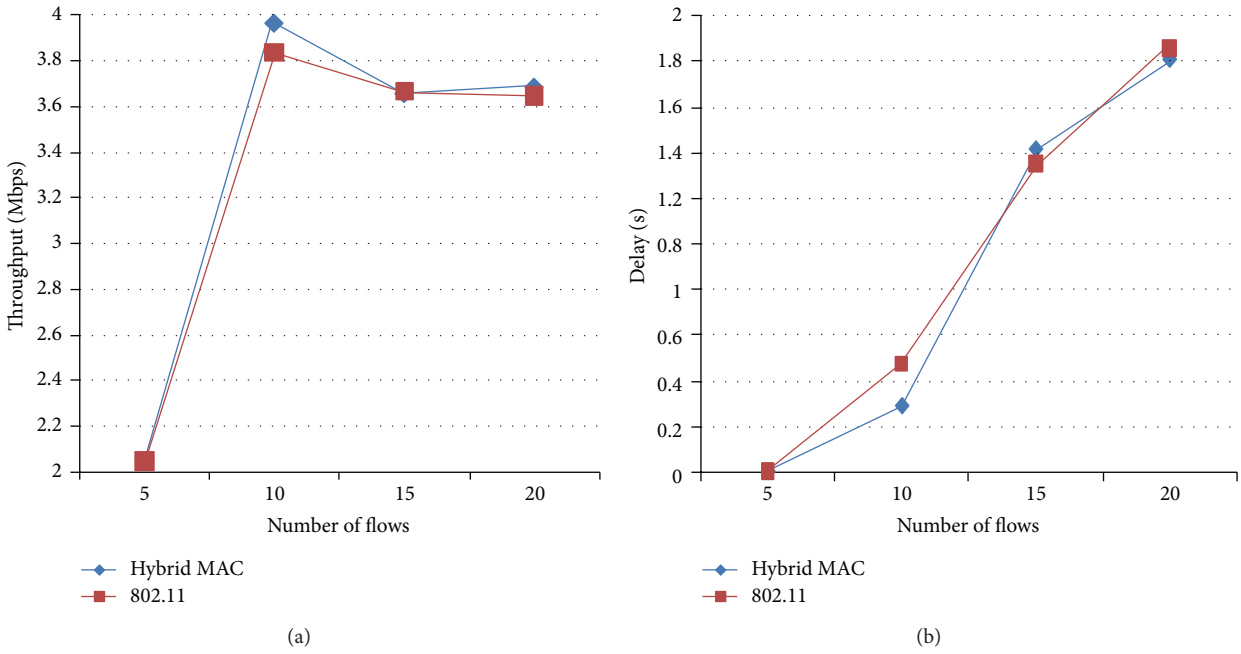


Figure 3: Throughput and delay depending on the number of nodes: unicast (a) throughput; (b) delay.

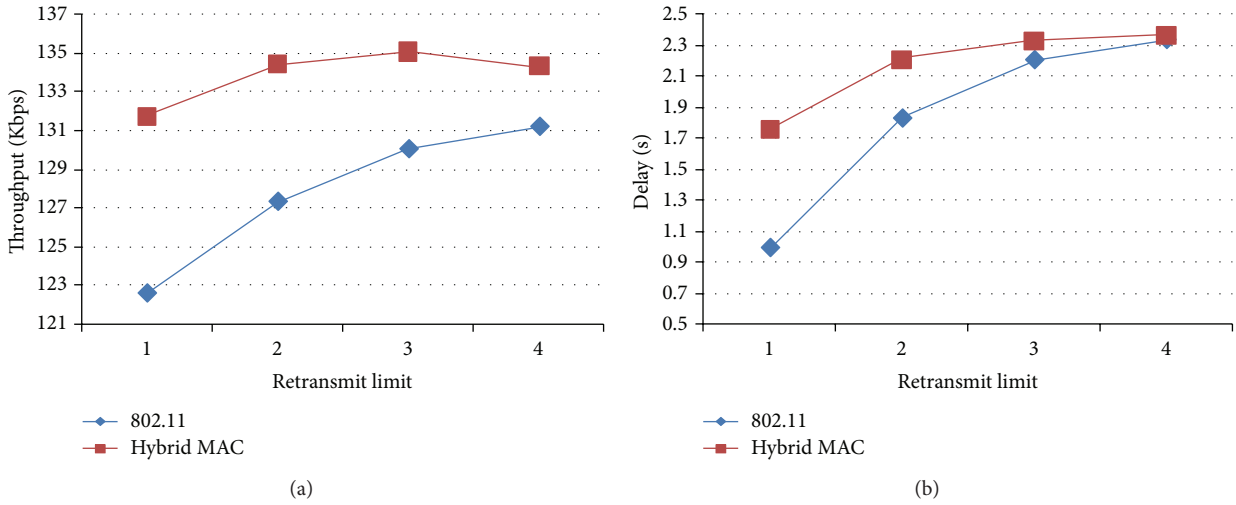


FIGURE 4: Throughput and delay depending on the number of retransmission: unicast (a) throughput, (b) delay.

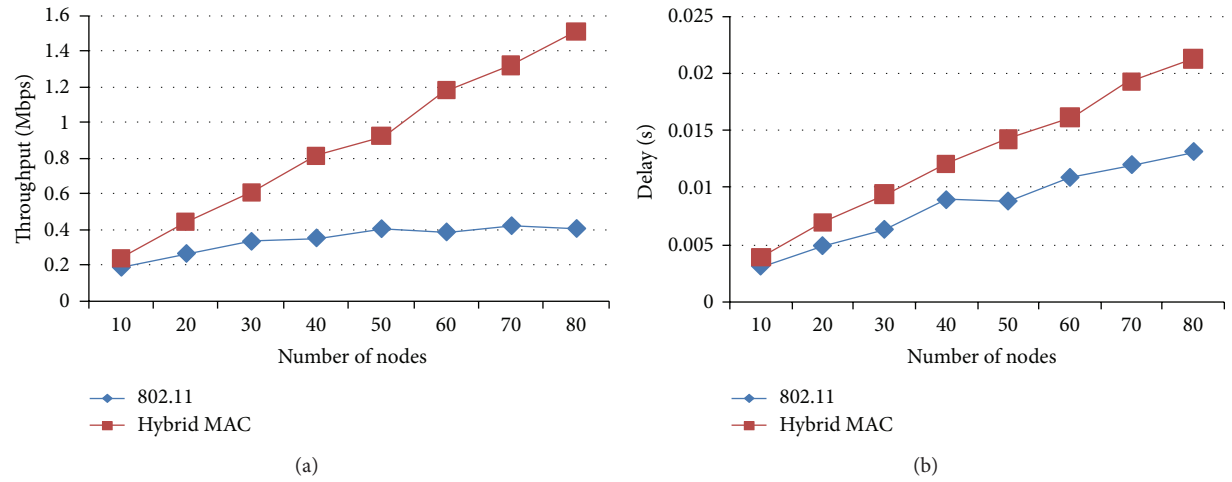


FIGURE 5: Throughput and delay depending on the number of nodes: broadcast, 100 ms periodic transmission (a), throughput (b) delay.

of unicast depend on the retransmission limit where throughput of hybrid MAC provides better than IEEE 802.11 protocol in general. In IEEE 802.11 protocol, when the transmission of frames failes, the retransmission of the corresponding frame occurs where the number of retransmission can be adjusted. If the number of nodes increases, the probability of collision also increases due to the nature of CSMA/CA, which results in decrement of successful transmission. Thus, throughput of unicast is handled by the retransmission limit, that is, small number of retransmission induces throughput decrement, whereas hybrid MAC distributes the trial number of access to the nodes at the given time by dividing the frame into several time slots; which increases throughput. Hybrid MAC has longer time delay than IEEE 802.11 since delay is dependent on time slot, that is, 4 time slots are equally used regardless of retransmit limit in these simulations, which results in long delay compared with IEEE 802.11 protocol.

Figure 5 represents throughput and delay of broadcast depending on the number of nodes when each node broadcasts every 100 ms. This scenario is corresponding to situation that every vehicle broadcasts frame periodically for safety services. In this scenario, the number of broadcast frames increases as the number of nodes increases; this results in increment of network density. The figure shows that throughput has a saturation point as the number of nodes increases in IEEE 802.11. This implies failure of transmission increases as the number of collision increases since there in no retransmission in broadcasting. However, in hybrid MAC, the overall network throughput increases linearly as the number of nodes increases. This is due to the decrement of frame collisions since each node transmits frames with different access priority based on slot ownership. Figure 6 shows throughput and delay of broadcast depending on the number of nodes when each node broadcasts every 10 ms. The figure reveals that the network will be saturated as the number of nodes increases and frame transmission rate decreases. From Figures 5 and 6, it is observed that hybrid MAC has superior performance than IEEE 802.11 protocol in broadcast.

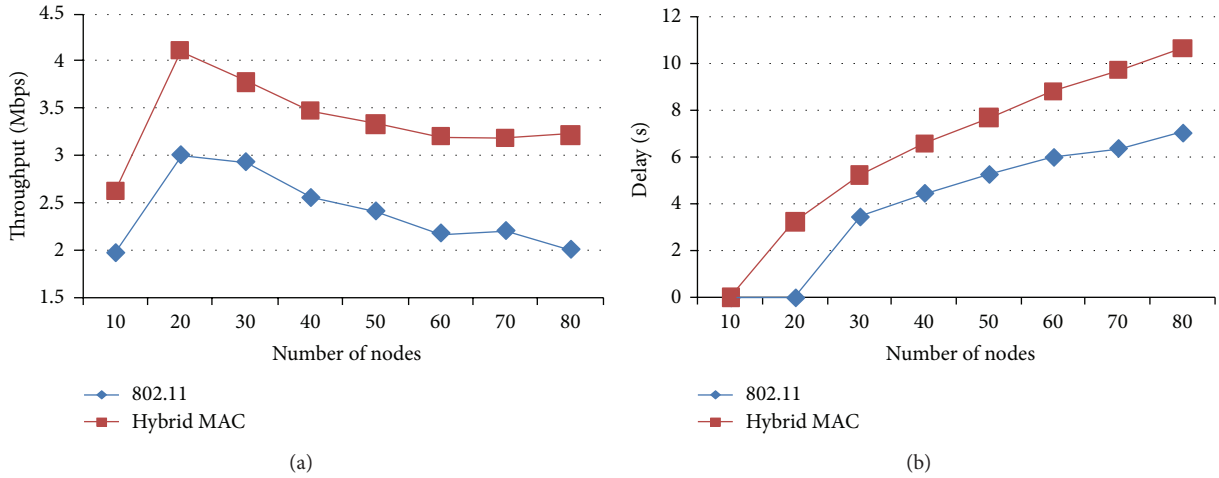


Figure 6: Throughput and delay depending on the number of nodes: broadcast, 10 ms periodic transmission (a) throughput; (b) delay.

In summary, hybrid MAC has better performance than IEEE 802.11 protocol in broadcast, although the performance in unicast is almost the same. However, it is worth stressing that most safety-related applications and information announcement-related applications are required in broadcast. In the final report of VSC project Task 3 [8], it is reported that most of applications uses broadcast.

4. Application Scenarios and Implementation Issues

Until now, we have discussed about advanced technologies for vehicular communication and their advantages. In this section, we will consider some application scenarios including implementation issues.

4.1. Application Scenarios. With V2V communication, we implemented anticollision application by multi-hop communication with very low latency. A safety message such as an accident alarm and vehicle approaching is generated by vehicle, and this message is transmitted to the following/ahead vehicles by multi-hop communication. In the vehicle terminal, the information of my vehicle, front vehicle, and rear vehicle is displayed. When front/rear vehicle approaches to a certain range of my vehicle, the warning message is generated in both vehicles, that is, my vehicle and front/rear vehicle. If the front/rear vehicle moves very close to my vehicle, the very emergency warning is generated. Otherwise, the warning message disappears. This application can prevent or decrease consecutive collisions by informing emergency situation in advance, where this application is developed and under testing in some projects [11–13]. It is worth stressing that the location information is critical in this application since the distance between the vehicles is based on the location. Although reliable communications are guaranteed, this application cannot be realized without accurate location information. We use GPS for deciding location. With GPS, we have seen that error occurs sometimes. To solve this problem, the technique for finding location with very high resolution has to be developed. With V2I communication, we are developing vehicle information-based service (VIS) and intersection safety service (ISS) to seek market-valued key applications. In VIS, vehicle information generated by ECU/TCU/MCU in the vehicle can be delivered to the server/ITS center, and this vehicle information is used for diagnosing and analyzing the vehicle status remotely as well as gathering of traffic information. By providing bidirectional communication, the diagnosed information will be reported to the corresponding vehicle. Then, the driver can monitor the status of vehicle. In ISS, safety messages, that is, vehicle approaching, traffic signal information, pedestrians alerting, and so forth, are broadcasted to the corresponding vehicles by infrastructure. This application will guarantee safety enhancement at the intersection.

4.2. Implementation Issues. In this subsection, we will discuss other technologies which can enhance the system performance. Vehicular communication experiences rapid changes in network topology. Hence, dynamic ad hoc routing has to be considered in VANET. By keeping in mind this fact, the current WAVE will have to be enhanced to meet the three key factors; latency, high mobility, and dynamic VANET.

Besides the proposed MAC scheme, we also have to consider the following technologies: adaptive beam forming in vehicular environments, precise location information in vehicular environments, and non-line-of-sight (NLOS) problem. For adaptive beam forming, we are considering multiple antennas. By applying beam forming algorithm and selecting the specific antenna, the receiver sensitivity can be increased. For example, at the receiver, the initial mode may use multiple antennas. When the receiver tunes to the specific direction, the receiver increases the receiver's gain for the corresponding direction by adjusting antenna pattern. At the transmitter, we may increase the transmitter antenna gain by deciding transmission direction in advance. Thus, the error performance will be improved by supporting high capacity in the network. For various use cases, the location of vehicle will be a critical factor especially in safety-related applications

such as V2V anticollision warning, sudden stop warning, and road working warning. Nowadays, GPS is commonly used for finding the location. However, GPS does not provide accurate information and has several meters of error. To address this issue, the technique for finding location with very high resolution has to be adopted by maintaining the cost of location management with appropriate level. The other issue is NLOS problem in vehicular environments. Since 5.9 GHz frequency band is used for vehicular communications, it is hard to support reliable communication links without LOS. In the vehicular environments, there exist many blocking objects such as big trucks and curve areas. To overcome this problem, we may use a relay node at the road or vehicle. A relay node may regenerate the received signal, or simply amplify and forward that signal to another vehicle. Then, the shadowed area and NLOS may be removed which guarantees the reliable communication links.

5. Conclusions and Future Work

In this paper, we investigated a novel MAC scheme for vehicular communications. To improve the performance vehicular communication systems, hybrid MAC protocol is suggested in MAC layer, which supports access priority using time slot and CSMA/CA. Our simulations revealed that hybrid MAC provides better throughput over CSMA/CA by time slot scheduling and priority channel access especially in broadcasting. In addition, we also addressed some practical implementation issues including application scenarios.

In future work, it may be very useful if the proposed scheme is implemented in practical systems since the proposed scheme in this paper only shows simulation results. It is also worth to consider various algorithms in slot owner selection to increase the performance of overall system with the aforementioned implementation issues.

References

[1] IEEE 802 Standard for information technology—telecommunications and information exchange between systems—local and metropolitan area networks-specific requirements, Part 11, Wireless LAN Medium Access Control and Physical layer specifications, 2008.

[2] F. Bai and H. Krishnan, "Reliability analysis of DSRC wireless communication for vehicle safety applications," in *Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC '06)*, pp. 355–362, September 2006.

[3] S. Biswas, R. Tatchikou, and F. Dion, "Vehicle-to-vehicle wireless communication protocols for enhancing highway traffic safety," *IEEE Communications Magazine*, vol. 44, no. 1, pp. 74–82, 2006.

[4] T. L. Willke, P. Tientrakool, and N. F. Maxemchuk, "A survey of inter-vehicle communication protocols and their applications," *IEEE Communications Surveys and Tutorials*, vol. 11, no. 2, pp. 3–20, 2009.

[5] B. Jarupan and E. Ekici, "Location- and delay-aware cross-layer communication in V2I multihop vehicular networks," *IEEE Communications Magazine*, vol. 47, no. 11, pp. 112–118, 2009.

[6] ASTM Standard E 2213-03, Standard specification for telecommunications and information exchange between roadside and vehicle systems. 5 GHz band Dedicated Short Range Communications (DSRC) Medium Access Control (MAC) and Physical Layer (PHY) specifications, 2003.

[7] IEEE Std P802.11p, IEEE standard for information technology-telecommunications and information exchange between systems-local and metropolitan area networks-specific requirements, Part 11, Amendment 6: Wireless Access in Vehicular Environments, 2010.

[8] The CAMP Vehicle Safety Communications Consortium, Vehicle Safety Communications Project Task 3 Final Report, U.S. Department of Transportation, 2005.

[9] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.

[10] I. Rhee, A. Warrier, M. Ais, and J. Min, "Z-MAC: a hybrid MAC for wireless sensor networks," in *Proceedings of the 3th International Conference on Embedded Networked Sensor Systems*, pp. 90–101, 2005.

[11] M. Bauer, "Redefine business models as V2V and V2I converge," V2X for auto safety and mobility Europe 2013, 2013.

[12] F. Ibrahim, "US trials, testbeds, impact in Europe," V2X for auto safety and mobility Europe 2013, 2013.

[13] F. Forsterling, "Challenges for C2X based ITS solutions," V2X for auto safety and mobility Europe 2013, 2013.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2013, Article ID 634531, 8 pages
http://dx.doi.org/10.1155/2013/634531

Research Article

Location Estimation Using Space-Time Signal Processing in RFID Wireless Sensor Networks

Chang-Heon Oh

School of Electrical, Electronics and Communication Engineering, Korea University of Technology and Education, 1600 Chungjelolno, Byungchun-myun, Chonan-Si, Choongnam 330-708, Republic of Korea

Correspondence should be addressed to Chang-Heon Oh; choh@koreatech.ac.kr

Received 3 September 2013; Accepted 7 October 2013

Academic Editor: Jongsung Kim

We consider real-time locating systems (RTLS) consisting of radio frequency identification (RFID) tags and multiantenna readers to propose a novel location estimation algorithm based on the concept of space-time signature matching for multipath environments. In contrast to previous fingerprint-based approaches that rely on received signal strength (RSS) information only, the proposed algorithm uses angle, delay, and RSS information from the received signal to form a signature, which in turn is utilized for location estimation. We evaluate the performance of the proposed algorithm in terms of the average error probability. Simulations and analytic results confirm the effectiveness of the proposed algorithm for location estimation even in non-line-of-sight (NLOS) environments.

1. Introduction

RFID technology has greatly extended our ability to monitor and control the physical world and has been proposed for various applications including search, rescue, disaster relief, and target tracking. The inherent characteristics of these applications make an object's location an important element [1, 2].

Wireless localization, the location estimation of an object in wireless networks, has been extensively investigated. The Federal Communications Commission (FCC) has recently amended its rules to permit the use of improved RFID systems to facilitate seaport security efforts [3]. The rules will permit the use of more powerful RFID devices with commercial shipping containers, which, in turn, will offer improved security at the nation's ports, rail yards, and warehouses. One of the representative standards of these applications is the RTLS. It is a real-time location estimation system based on RFID technology at a 2.45 GHz band that is intended to provide an approximate location (within 3 m in line-of-sight (LOS) environment) with frequent updates. It operates in conformance with the International Committee for Information Technology Standards (INCITS) 371 RTLS standard [4].

Many wireless location techniques have been investigated and can be divided into two approaches: geometric-based techniques, such as angle of arrival (AOA) [5], time of arrival (TOA) [6], time difference of arrival (TDOA) [7], and location fingerprinting approaches [8, 9]. In the case of AOA, TOA, and TDOA, location estimation is based on triangulation techniques that require LOS between the transmitter and the receiver. However, in practical wireless environments, it is difficult to guarantee a LOS path between the transmitter and the receiver. Therefore, the time and angle of arrival signal is affected by multipaths, and the estimation accuracy is generally considerably reduced in NLOS environments. On the other hand, location fingerprinting techniques solve the problems related to NLOS and multipath propagation by using a "radio map" of the RSS for a target environment. Although frequent and extensive site measurements are often required, location fingerprinting performs well in NLOS conditions [8]. Thus, In this paper, we focus on location fingerprinting approaches since they are more suitable for NLOS applications. Many fingerprinting-based techniques have been proposed for location estimation in wireless networks, with a special emphasis on wireless local area network (WLAN) applications [8, 10, 11]. Unfortunately, most existing research related to location fingerprinting only uses the RSS information. The location is estimated by comparing the current measured data with the "radio map" of the premeasured

database. This approach is simple to implement and gives reasonable accuracy in small area applications. However, it has limited value for some applications, such as outdoor scattering environments where radio signal propagation is very complicated because of severe multipath effects. In this case, unfortunately, the RSS information is not enough for accurate location estimation. To estimate a location more precisely, we need more information to characterize the received signal's features.

In this paper, we investigate a location estimation algorithm for RTLS consisting of RFID tags and multiantenna readers to propose novel location estimation algorithms based on the concept of space-time signature matching in NLOS multipath environments. The proposed algorithm uses angle, delay, and RSS information from the received signal; this is in contrast to previous fingerprint-based approaches [8–10], which rely on RSS information only to form a signature that is in turn utilized for location estimation. We evaluate the performance of the proposed algorithm in terms of the average error probability. The organization of this paper is as follows. Section 2 provides a description of the system model, an overview of RTLS and the basic space-time communication model in the proposed approach. Section 3 describes the proposed location estimation algorithm based on the space-time signature matching technique and presents the detection method. Section 4 describes multipath channel generation based on the ray-tracing technique, and simulation results are presented to verify the effectiveness of the proposed algorithm. Conclusions are given in Section 5.

2. System Model

2.1. Overview of RTLS. A real-time locating system is an automatic system that continually monitors the locations of objects. The system continually updates the database with current tag locations as frequently as every several seconds or as infrequently as every few hours depending on the mobility of the target tags. In typical applications, systems are required to track thousands of tags simultaneously, and the average tag battery must last for five or more years [12]. The RTLS infrastructure, as shown in Figure 1, typically consists of RTLS transmitters (radio tags), RTLS receivers (readers), and the RTLS server. RTLS transmitters blink (or transmit) a direct sequence spread spectrum (DSSS) signal, and RTLS readers, whose locations are fixed, receive signals from the tags. The RTLS server aggregates data from the RTLS readers and determines the tag's current location. Each DSSS transmission from a transmitter contains a "blink" packet containing subblinks. All subblinks within a blink are identical to provide time diversity.

Each subblink includes the RTLS transmitter's 32-bit identification (ID), 4 bits of status data, cyclical redundancy check (CRC) data, and optional telemetry data. The number of subblinks per blink, N_s, and the blink interval are configurable. The RTLS transmitters transmit at a power level that can ensure successful reception at readers located at least 300 meters away [4]. In this paper, we assume that each reader is equipped with an array of M antennas that can adapt space-time signal processing. We consider not only LOS propagation but also the more complicated NLOS multipath propagation scenario.

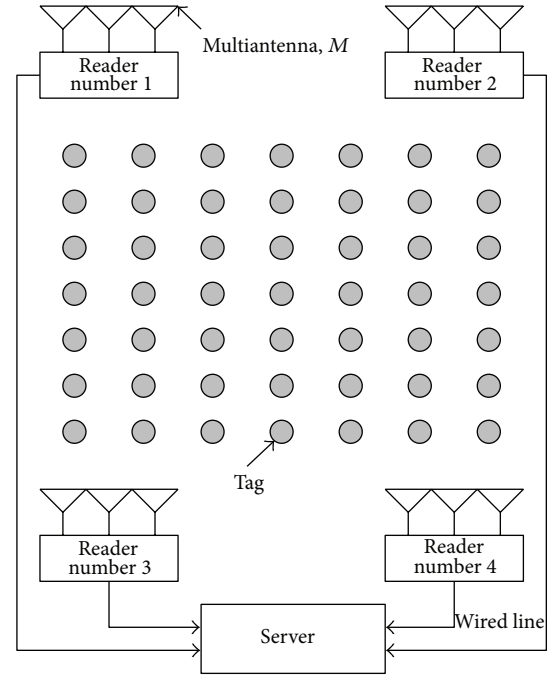


Figure 1: Elements of RTLS infrastructure: tags are regularly spaced for illustration.

2.2. Space-Time Communication. We assume that N_r tags are randomly distributed in a certain area and K readers are placed near the circumference of the area, for example, on the corners of the service area (see Figure 1). Each reader, equipped with an M element antenna array, is assumed sufficiently far from most of the tags in the same plane so that far-field assumptions apply. K readers receive the signal from a tag to estimate its location. Each tag transmits a spread spectrum signaling waveform $s(t)$ of duration T [s] and (two-sided) bandwidth W [Hz]. Let $N = TW \gg 1$ denote the time-bandwidth product of the signaling waveforms representing the approximate dimension of the spatiotemporal signal space. Thus, the signal space of the space-time waveforms has the dimension $MN = MTW$. We make the practical assumption that the readers and tags are frequency (f_c) but not phase synchronized. Furthermore, we assume that the phase offset between each tag and the reader stays constant for at least the packet duration T. A tag transmits the spread spectrum signal to update its status. Denoting the ith tag's data by c_i, the transmitted signal $s_i(t)$, $i = 1, 2, \ldots, N_r$ is expressed as

$$s_i(t) = c_i q(t), \tag{1}$$

where $q(t)$ is the spreading signal, that is, a pseudorandom noise (PN) sequence of N rectangular pulses or chips of duration T_c each. In most applications, $T = NT_c$ for some integer N. The spreading signal can be expressed as [13]

$$q(t) = \sum_{j=0}^{N-1} x_j p_{T_c}(t - jT_c), \quad 0 < t \le T, \tag{2}$$

where $\{x_j\}$ is a binary (± 1) spreading sequence and $p_{T_c}(t)$ is a rectangular pulse of duration T_c; that is, $p_{T_c}(t) = 1/\sqrt{T_c}$ when $0 \le t \le T_c$ and $p_{T_c}(t) = 0$ otherwise.

The transmitted signal $s_i(t)$ undergoes multipath propagation due to scattering off of objects, for example, densely stacked containers that are randomly distributed. Thus, each reader receives a superposition of multiple waveforms from the tag. In this paper, we assume for the sake of simplicity that only one tag transmits a signal at a certain instance; this is reasonable because each tag is active for only short periods of time, and in many applications they do not have to update the central server very often. Thus, the received signal at each reader can be expressed as

$$\mathbf{r}(t) = \sum_{n=1}^{N_p} \beta_n e^{-j\phi_n} a(\theta_n) c_n q(t - \tau_n) + w(t), \tag{3}$$

where β_n and ϕ_n denote the magnitude and phase, respectively, of the complex path gain corresponding to nth path with a normalized AOA θ_n and delay τ_n. N_p denotes the number of paths, and $w(t)$ denotes an additive white Gaussian noise (AWGN) vector process representing the noise at different antenna elements of variance σ^2. In (3), $a(\theta_n)$ is the steering/response vector of the antenna array toward angle θ_n; this is explained in further detail later. For simplicity, we consider a one-dimensional uniform linear array (ULA) with spacing d and assumed M to be odd without loss of generality; we then define $\widetilde{M} = (M-1)/2$. The array steering/response vector for a ULA is given by

$$a(\theta) = \left[e^{j2\pi\widetilde{M}\theta}, \ldots, 1, \ldots, e^{-j2\pi\widetilde{M}\theta}\right]^T, \tag{4}$$

where the normalized angle θ is related to the physical angle of arrival φ measured w.r.t. broadside as $\theta = d\sin(\varphi)/\lambda$ with $\lambda = 1/f_c$. The steering/response vector represents the relative phases across antenna for the receiver from angle θ. We assume that a $d = \lambda/\sqrt{2}$ spacing, which corresponds to the maximum angular spread (90°) at the antenna array; larger spacing can be used for smaller angular spreads ($d = \lambda/(2\sin(\varphi_{\max}))$ spacing results is a one-to-one mapping between $\theta \in [-0.5, 0.5]$ and $\varphi \in [-\varphi_{\max}, \varphi_{\max}] \subset [-\pi/4, \pi/4]$).

From (3), the physical channel can be expressed as

$$h(t) = \sum_{n=1}^{N_p} \alpha_n \delta(t - \tau_n) a(\theta_n), \tag{5}$$

where $\alpha_n = \beta_n e^{-j\phi_n}$ and the $M \times 1$ vector $\mathbf{h}(t)$ represent the impulse response for the space-time multipath channel. Let $\min_n \tau_n = 0$ and $\max_n \tau_n = \tau_{\max}$; the delay spread of the channel is then $\tau_{\max}$, and we assume that the packet signaling duration $T \gg \tau_{\max}$.

3. Proposed Location Estimation Algorithm

This section begins with an overview of conventional fingerprint-based algorithms and our algorithm, followed by the spatiotemporal partitioning technique. This technique serves as a key role in our algorithm because the angle and delay associated with each scatterer can be obtained through this. The details of our algorithm are then presented.

3.1. Fingerprint-Based Location Estimation. Conventional fingerprint-based location estimation algorithms rely solely on the received RSS measured at multiple single-antenna receivers. First, an RSS database is formed by transmitting signals from a large number of locations called reference points and storing the RSS measured at all receivers. Once the database is ready, the system is able to estimate the location of an object at an unknown location by finding the location in the database whose RSS values most closely match the current RSS values.

However, the estimation accuracy of the fingerprint technique is limited by the fact that it relies only on RSS. Although RSS is an important characteristic for a location, RSS-based algorithms cannot distinguish two locations if their RSSs' are similar, even if they induce different angles of arrivals and delays. This is especially true in the presence of multipaths. Thus, in this paper, we propose a novel method based on space-time signature matching to accurately estimate the location by utilizing not only the RSS but also the angle and delay information of the received signal. Two key differences are that our algorithm can (1) utilize angular information captured through multiple antennas at readers and (2) resolve powers contributed by each ray or path in contrast to RSS, which is the aggregate of the path contributions.

The proposed space-time signature matching algorithm is carried out in two steps. In the first step, reference signatures are generated for every reference point. A tag at an unknown location can then be estimated by comparing the signature generated from the received signal with reference signatures. This is explained in more details later. We assume that there are K readers and N_r reference locations.

Step 1. Reference signature database generation

(i) Place a transmitter at one of the reference positions and let it transmit a known spread spectrum waveform.

(ii) Readers receive the waveform and generate a signature vector of this location. Denoted by $\psi_i^{(k)}$, the signature vector of the ith reference point is measured at the kth reader. Details on generating it are explained in Section 3.3.

(iii) These signatures $\psi_i^{(k)}$, $k = 1, 2, \ldots, K$ along with the corresponding reference coordinates (x_i, y_i) are saved in the database located in the server.

(iv) Move to the next reference point until all reference points, $i = 1, 2, \ldots, N_r$, are visited.

(v) Thus, for N_r reference points, the database consists of $N_r K$ signature vectors, $\psi_i^{(k)}$, $i = 1, 2, \ldots, N_r$, $k = 1, 2, \ldots, K$.

(vi) Once the reference signatures are obtained, we can locate a tag at unknown position as follows.

Step 2. Location estimation process.

(i) A tag at unknown location transmits a known waveform, and readers compute a signature $\widetilde{\psi}^{(k)}$, $k = 1, 2, \ldots, K$ based on the received signal.

(ii) Compare the currently computed signature with the reference signatures in the database.

(iii) Estimate the unknown tag's location by finding the location in the database whose signatures are closest to the currently obtained signature. All information (i.e., angle, delay, and RSS information at each reader) is used or combined to estimate the tag location.

3.2. Path Partitioning in Angle and Delay. The main concept behind our algorithm is that the estimation accuracy can be greatly improved by working with a detailed map or signature of a scattering environment. The detailed signature consists of the angle, delay, and path gain associated with each scatterer, which provides more information to the estimator than the conventional RSS-based methods. Such information buried in the regular channel vector or matrix clearly shows up in the virtual channel vector, which is explained next.

A key property of the virtual channel representation is that its coefficients represent a resolution for the multipath in angle and delay commensurate with the signal space parameters M and W, respectively [14, 15]. The virtual representation in the angle corresponds to beam forming in M fixed virtual directions: $\widetilde{\theta}_m = m/M$, $m = -\widetilde{M}, \ldots, \widetilde{M}$. The $M \times M$ unitary (DFT) matrix is defined as

$$A = \frac{1}{\sqrt{M}} \left[a\left(-\frac{\widetilde{M}}{M}\right), \ldots, 1, \ldots, a\left(\frac{\widetilde{M}}{M}\right) \right]. \quad (6)$$

Its columns are the normalized steering vectors for the virtual angles, and they form an orthonormal basis for the spatial signal space. Using this, a virtual spatial vector $h_V(t)$ can be defined as

$$h_V(t) = A^H h(t) \longleftrightarrow h(t) = A h_V(t), \quad (7)$$

and its coefficients are related to the physical paths in

$$h_V(m;t) \approx M \sum_{n \in S_{\theta,m}} \alpha_n g\left(\theta_n - \frac{m}{M}\right) \delta(t - \tau_n), \quad (8)$$

where $g(\theta) = (1/M)(\sin(\pi M\theta)/\sin(\pi\theta))$ is the Dirichlet sinc function that captures the interaction between the fixed virtual beams and true signal directions, $\delta(\cdot)$ denotes the Dirac delta function, and $S_{\theta,m} = \{n \in \{1, \ldots, N_p\} : -1/2M < \theta_n - m/M \leq 1/2M\}$ represents the set of all paths whose angles lie in the mth spatial resolution bin of width $\Delta\theta = 1/M$, centered around the mth beam. Thus, the virtual spatial representation partitions the multipath signals in the angle.

The multipath responses within each spatial beam can be further partitioned by resolving their delays with resolution $\Delta\tau = 1/W$. Let $L = \lceil \tau_{\max} W \rceil$ be the normalized delay spread of the channel. The impulse response $h_V(m;t)$ of the channel corresponding to the mth direction can be further processed

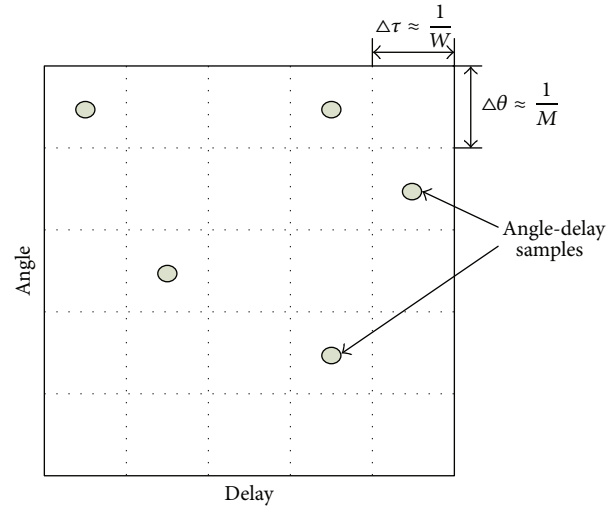


Figure 2: Angle-delay resolution bins.

to yield entries $h_V(m,l)$ corresponding to the mth direction and lth uniformly spaced delays as [16]

$$\begin{aligned} h_V(m;t) &\approx \sum_{l=0}^{L-1} h_V(m,l)\, \delta\left(t - \frac{l}{W}\right), \\ h_V(m,l) &= M \sum_{n=1}^{N_p} \alpha_n g\left(\theta_n - \frac{m}{M}\right) \operatorname{sinc}(W\tau_n - l) \\ &\approx M \sum_{n \in S_{\theta,m} \cap S_{\tau,l}} \alpha_n g\left(\theta_n - \frac{m}{M}\right) \operatorname{sinc}(W\tau_n - l), \end{aligned} \quad (9)$$

where $\operatorname{sinc}(x) = \sin(\pi x)/\pi x$ captures the interaction between the fixed virtual and true signal delays and $S_{\tau,l} = \{n : -1/2W < \tau_n - l/W \leq 1/2W\}$ is the set of paths whose relative delays lie within the lth delay resolution bin of width $\Delta\tau = 1/W$ centered around the lth fixed virtual delay $\widetilde{\tau}_l = l/W$.

Thus, the angle-delay virtual representation partitions the multipath responses into distinct angle-delay resolution bins: the virtual coefficient $h_V(m,l)$ is a superposition of all multipath responses whose angles and delays lie in the intersection of the mth angle and lth delay bin (see Figure 2). The generation of a signature vector based on virtual representation and the location estimation algorithm using signature matrices are presented next.

3.3. Generation of Signature Vector. The signature vector is what we eventually use to estimate the location. It is computed from the received signal at each reader, as illustrated in Figure 3.

Let $r(t)$ be the $M \times 1$ vector of the received signal from the antenna array of reader as in (3). For a given reference tag location, we ignore the dependency on reader (k) and reference location (i). After a unitary spatial transform, we obtain $r_V(t) = A^H r(t)$, which represents the signal in the

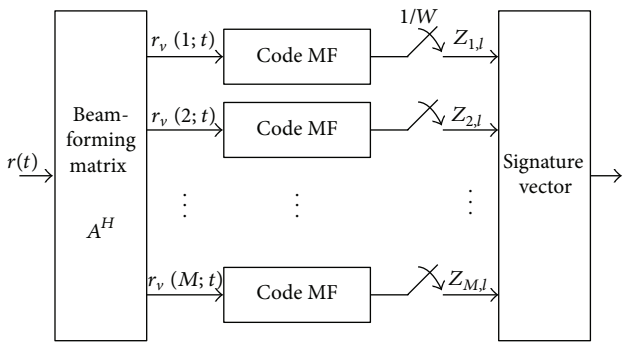


Figure 3: Computation of the signature vector: angle-delay matched filtering.

virtual angle domain, that is, each row corresponds to a virtual angle. Using (5) and (7), $r_V(t)$ can be expressed as

$$r_V(t) = \sum_{l=0}^{L-1} h_V(l)\, s\left(t - \frac{l}{W}\right) + w_V(t), \tag{10}$$

where $w_V(t) = A^H w(t)$ is the noise in the virtual angle domain.

Note that $s(t-l/W) = c_i q(t-l/W)$ is nothing but a delayed and BPSK-modulated version of a pseudorandom waveform $q(t)$. Hence, we have the following property

$$\left\langle s\left(t - \frac{l}{W}\right), s\left(t - \frac{l'}{W}\right)\right\rangle \approx \delta_{l-l'}. \tag{11}$$

Thus, correlating each $r_V(m;t)$ with delayed versions of $s(t)$ yields a signature $\{z_{m,l} : m = 1,\ldots,M; l = 0,\ldots,L-1\}$. Consider

$$\begin{aligned} z_{m,l} &= \left\langle \mathbf{r}_V(m;t), s\left(t - \frac{l}{W}\right)\right\rangle \\ &= \int_0^{T+\tau_{\max}} \mathbf{r}_V(m,t)\, s^*\left(t - \frac{l}{W}\right) dt. \end{aligned} \tag{12}$$

$\{z_{m,l}\}$ are stacked to generate the $ML \times 1$ dimensional signature vector $\psi_i^{(k)}$. The signature vector $\psi_i^{(k)}$ for the ith reference point at the kth reader is formed by placing those matched-filter outputs as $\psi_i^{(k)}(Ml + m) = z_{m,l}$. This yields a database of $N_r K$ signatures for N_r reference locations and K readers. Figure 4 shows an example of the tag signature at four readers.

After obtaining signature vectors from all reference points, readers constantly monitor the field. When a strong signal is detected, each reader forms a signature vector $\widetilde{\psi}^{(k)}$ from the received signal. The maximum likelihood decision for the location is then made. We consider coherent detection as the detection method.

3.4. Coherent Detection. When a tag of unknown location is placed at one of N_r reference points, the detection of its location is equivalent to an N_r-ary hypothesis testing as

$$H_i : \widetilde{\psi} = \psi_i + w, \quad i = 1, 2, \ldots, N_r, \tag{13}$$

where $\widetilde{\psi} = [\widetilde{\psi}^{(1)^T}, \ldots, \widetilde{\psi}^{(K)^T}]^T$ and $\psi_i = [\psi_i^{(1)^T}, \ldots, \psi_i^{(K)^T}]^T$ represent an observed signature and reference signature at the ith reference point, respectively, and noise vector w consists of zero mean and σ^2 variance i.i.d. complex Gaussian random variables.

The probability density function (pdf) of $\widetilde{\psi}^{(k)}$ given H_i is

$$f_i(\widetilde{\psi}) = \frac{1}{(\pi\sigma^2)^{MLK}} \exp\left(-\frac{\|\widetilde{\psi} - \psi_i\|^2}{\sigma^2}\right), \tag{14}$$

where $\|\cdot\|$ represents the Euclidean norm of a vector. Hence, the maximum likelihood estimation of the tag location is made as follows:

$$\widehat{i}_{ML}(\widetilde{\psi}) = \arg\max_{i\in\{1,\ldots,N_r\}} f_i(\widetilde{\psi}) = \arg\min_i \|\widetilde{\psi} - \psi_i\|^2, \tag{15}$$

which implies that the maximum likelihood decision is equivalent to simply finding a reference signature that is closest to the observed signature.

We now analyze the error probabilities of this algorithm. When the true tag location is i, the error probability can be upper bounded using the union bound technique as

$$P(e \mid i) \le \sum_{j=1, j\neq i}^{N_r} P\left(\widehat{i} = j \mid i\right), \tag{16}$$

where i and $\widehat{i}$ denote the true and estimated tag positions, respectively. The pairwise error probability can be easily derived to be

$$P\left(\widehat{i} = j \mid i\right) = Q\left(\frac{\sqrt{\sum_{k=1}^{K} \left\|\psi_i^{(k)} - \psi_j^{(k)}\right\|^2}}{\sqrt{2}\sigma}\right), \tag{17}$$

where $Q(x) = (1/\sqrt{2\pi})\int_x^\infty e^{-t^2/2} dt$ is the Gaussian tail function. Since $Q(x)$ is monotonically decreasing, the error probability improves with increasing k, decreasing σ, and the norm of the difference vector. Assuming each location has an equal chance of being chosen, the average error probability can be upper bounded as

$$\begin{aligned} P_e &= \frac{1}{N_r}\sum_{i=1}^{N_r} P(e \mid i) \\ &\le \frac{1}{N_r}\sum_{i=1}^{N_r}\sum_{j=1, j\neq i}^{N_r} Q\left(\frac{\sqrt{\sum_{k=1}^{K} \left\|\psi_i^{(k)} - \psi_j^{(k)}\right\|^2}}{\sqrt{2}\sigma}\right). \end{aligned} \tag{18}$$

We show the above upper bound with the Monte-Carlo simulation in Section 4 to confirm the validity of the simulation.

4. Simulation and Discussions

The performance of the proposed algorithm is evaluated through simulations and theoretical analysis. This section

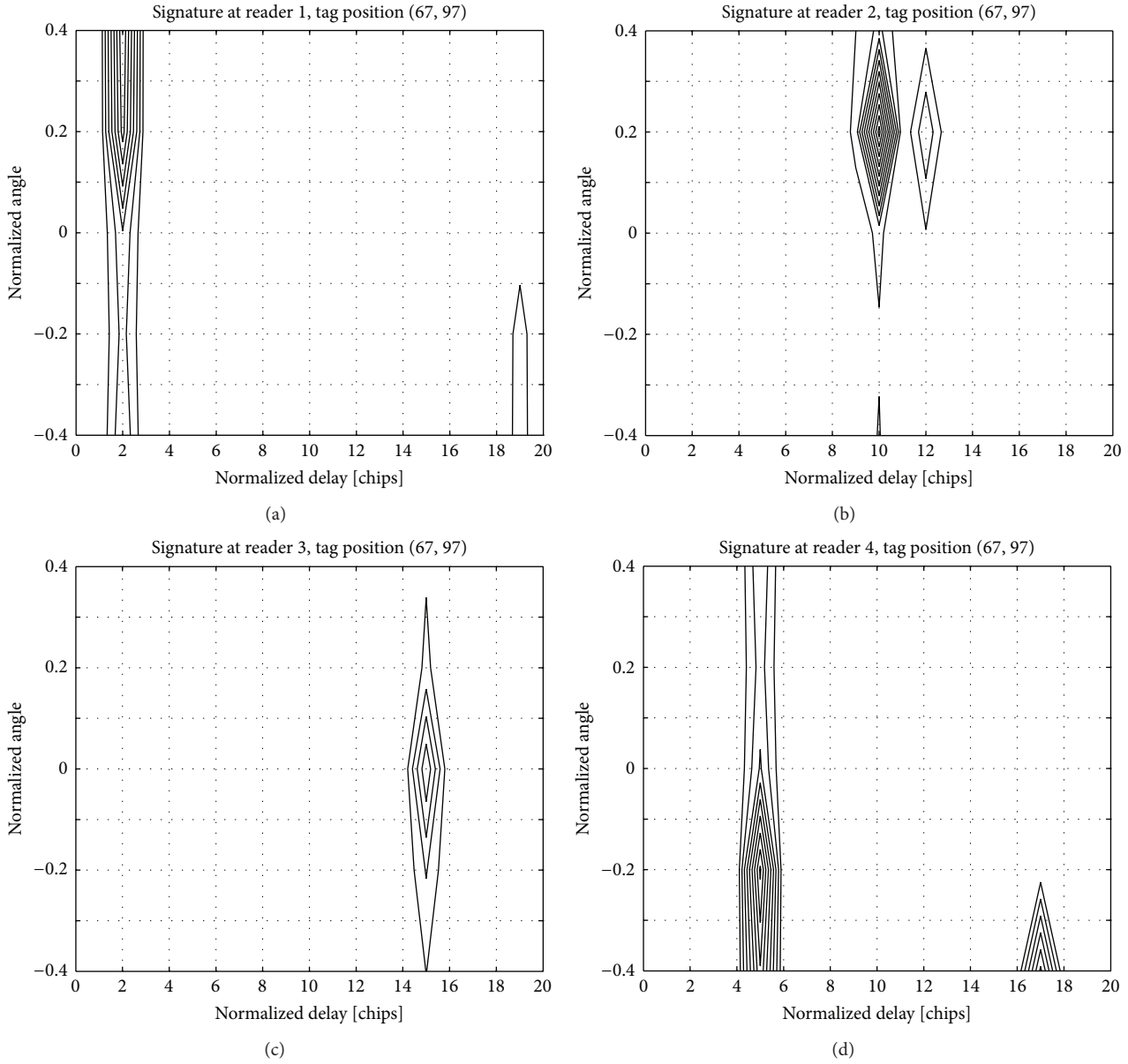


Figure 4: Example of the signature at each reader for a tag position.

first describes how we generate multipath channel realizations and then presents the simulation and analytic results with discussions.

4.1. Generation of Multipath Channel Realizations. We consider a field of a 300 m × 300 m square region. We assume that reference tags are randomly distributed over the region, and four readers equipped with five antennas each are placed in the corners of the region. Twenty rectangular-shaped scatterers are placed inside the region in a random manner. We measure signature vectors at reference tag positions separated by 3 m (see Figure 5).

In order to generate realistic multipath channel realizations that are suitable for seaport wireless environments, we develop a site-specific channel simulator based on the deterministic two-dimensional ray tracing technique. The ray-tracing technique has been used for accurately predicting site-specific radio propagation characteristics [17]. For each transmitter and receiver pair, ray-tracing simulates all paths arriving at the receiver to produce necessary information such as the RSS, multipath delay, AOA, and number of paths. In the ray-tracing algorithm, most rays arrive at the reader via multiple reflections due to the objects in the environment. We only consider reflections of the surface of scatterers, not diffraction. The angle of arrival, distance of travel (or delay), and number of reflections are recorded for every path and for every pair of a reference position and reader. To find a path, a reference point draws a line from itself along a certain angle. There can be three cases. (1) If the line does not intersect a reader

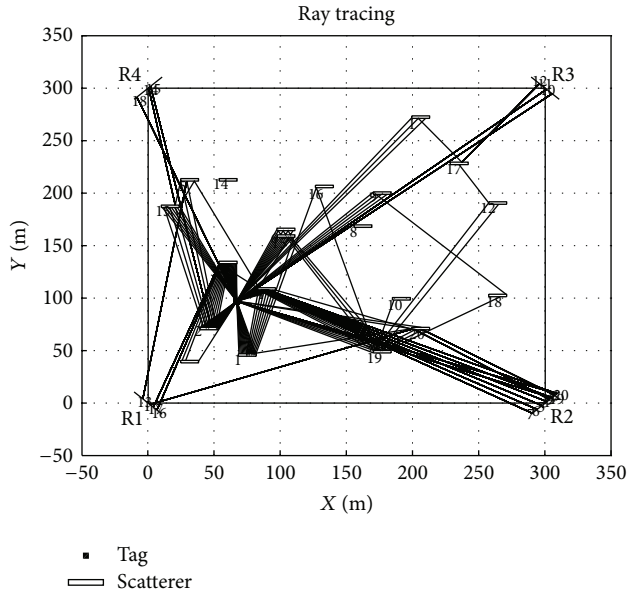


Figure 5: Sample snapshot of ray tracing at each reader.

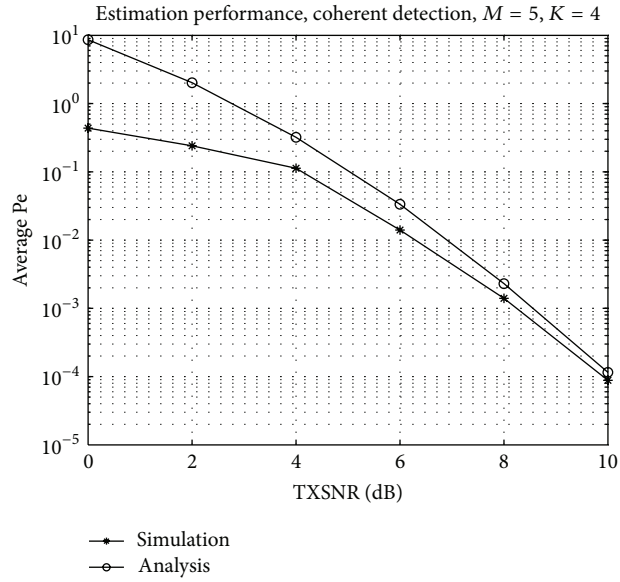


Figure 6: Average error probability of location estimation; comparison between simulation results and analytical results.

or scatterer, we examine the next angle, which is the previous angle incremented by a small amount (e.g., 1°) until all 360° have been examined. (2) If it intersects a reader, we are done with this angle, so we store all the information and examine the next angle. (3) If it intersects a scatterer, a new line is drawn according to the law of reflection, and the search continues until (1) or (2) is encountered. Figure 5 shows a sample snapshot of the ray-tracing results. In Figure 5, dark lines represent paths that eventually arrive at a reader, and light lines are those that initially intersect a scatterer but do not finally arrive at a reader. In this paper, we assume there is no signal fluctuation for at least the packet duration T because there is seldom movement of a tag in an RTLS environment.

4.2. Performance Evaluation. The effectiveness of the proposed algorithm for enhancing the accuracy of location estimation in NLOS environments is validated through computer simulation and theoretical analysis. In the simulation, 1,135 tags are considered as reference positions, and the transmit signal-to-noise ratio (TXSNR) was used as an SNR criterion. The estimation performance of the proposed algorithm is evaluated in terms of the average error probability. Figures 6 and 7 show the results of the proposed algorithm when there is no movement, that is, tag locations are fixed. In this case, we consider coherent detection as a detection method.

Figure 6 illustrates the estimation performance of the proposed algorithm in terms of average error probability for $M = 5$ and $K = 4$. For verification purposes, the analytical results obtained in Section 3 are also compared with the simulation results in Figure 6. As shown in Figure 6, the simulation results closely match the analytic upper bound, especially at high TXSNR. As expected, the estimation performance improves as the TXSNR increases. The estimation error probability is about 10^{-3} at TXSNR = 8 dB. This indicates that the proposed algorithm can reliably estimate the tag location even in NLOS environments.

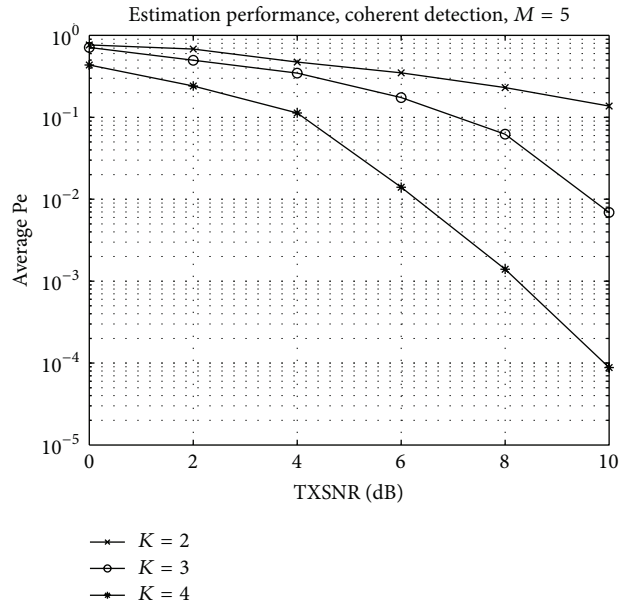


Figure 7: Average error probability of location estimation for different number of collaborating reader.

Figure 7 shows the effect of collaborative processing (combing) of the multiple readers on the average error probability for $M = 5$. As predicted, the estimation performance improves as the number of collaborating readers increases for a given TXSNR. This indicates that the more readers that collaborate, the more accurate estimation of the location that can be obtained; furthermore, the improvement is huge, on the order of a magnitude. These improvements are as expected from (17) since increasing k also increases the argument of the Q function.

5. Conclusions

In this paper, we investigated location estimation algorithms for real-time location systems and proposed novel location estimation algorithms based on the space-time signature matching in NLOS environments. Unlike existing works, we incorporated all available information (angle, delay and RSS) into a signature vector via spatiotemporal processing, and we did not require line-of-sight environment for successful operation. We derived the upper bound of the average error probability for coherent detection and carried out simulations. The two-dimensional ray-tracing technique was used to generate multipath channel realizations. From the results, we confirmed that in a fixed tag location the proposed algorithm could reliably estimate the target tag location even in the NLOS environments and the estimation accuracy greatly improved as the number of collaborating readers increased.

References

[1] T. G. Kanter, "Attaching context-aware services to moving locations," *IEEE Internet Computing*, vol. 7, no. 2, pp. 43–51, 2003.

[2] J. Schiller and A. Voisard, *Location-Based Services*, Morgan Kaufmann, 2004.

[3] L. M. Campos, J. L. Harris, and C. L. Rissetto, "FCC provides access to better RFID technology to port security: federal funding for this new tool may be available," ReedSmith, London, UK, 2004, http://m.reedsmith.com/files/Publication/fb0465f7-d047-4695-8165-d8bad31667a3/Presentation/PublicationAttachment/4d2c47b4-8dfd-4b38-89c3-5fd45b9b666b/bull0420.pdf.

[4] "Information technology automatic identification and data capture techniques-real time locating systems (RTLS). Part 2: 2. 4GHz air interface," ISO/IEC JTC 1/SC 31/WG 5, 2005.

[5] K. J. Krizman, T. E. Biedka, and T. S. Rappaport, "Wireless position location: fundamentals, implementation strategies, and sources of error," in *Proceedings of the 47th IEEE Vehicular Technology Conference*, vol. 2, pp. 919–923, May 1997.

[6] P.-C. Chen, "A non-line-of-sight error mitigation algorithm in location estimation," in *Proceedings of the IEEE Wireless Communications and Networking Conference*, vol. 1, pp. 316–320, 1999.

[7] L. Cong and W. Zhuang, "Non-line-of-sight error mitigation in TDOA mobile location," in *Proceedings of the IEEE Global Telecommunicatins Conference (GLOBECOM '01)*, pp. 680–684, November 2001.

[8] T.-N. Lin and P.-C. Lin, "Performance comparison of indoor positioning techniques based on location fingerprinting in wireless networks," in *Proceedings of the International Conference on Wireless Networks, Communications and Mobile Computing*, vol. 2, pp. 1569–1574, June 2005.

[9] K. Kaemarungsi and P. Krishnamurthy, "Modeling of indoor positioning systems based on location fingerprinting," in *Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, pp. 1012–1022, March 2004.

[10] A. Taheri, A. Singh, and E. Agu, "Location fingerprinting on infrastructure 802.11 Wireless Local Area Networks (WLANs) using locus," in *Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks (LCN '04)*, pp. 676–683, November 2004.

[11] K. Kaemarungsi, "Distribution of WLAN received signal strength indication for indoor location determination," in *Proceedings of the 1st International Symposium on Wireless Pervasive Computing*, pp. 1–6, January 2006.

[12] http://www.aimglobal.org/?page=RTLS.

[13] M. B. Pursley, "Direct-sequence spread-spectrum communications for multipath channels," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, no. 3, pp. 653–661, 2002.

[14] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Transactions on Signal Processing*, vol. 50, no. 10, pp. 2563–2579, 2002.

[15] A. M. Sayeed, "A virtual representation for time- and frequency-selective correlated MIMO channels," in *Proceedings of the IEEE International Conference on Accoustics, Speech, and Signal Processing*, vol. 4, pp. 648–651, April 2003.

[16] T. Sivanadyan and A. M. Sayeed, "Active wireless sensing: space-time information retrieval from sensor ensembles," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, pp. IV965–IV968, May 2006.

[17] B. Alavi, *Distance measurement error modeling for time-of-arrival based indoor geolocation [Ph.D. thesis]*, Worcester Polytechnic Institute, 2006.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2013, Article ID 145863, 7 pages
http://dx.doi.org/10.1155/2013/145863

Research Article

An Efficient WSN Simulator for GPU-Based Node Performance

An Na Kang,[1] Hyun-Woo Kim,[1] Leonard Barolli,[2] and Young-Sik Jeong[1]

[1] *Department of Multimedia Engineering, Dongguk University, 30 Pildongro 1 Gil, Jung-Gu, Seoul 100-715, Republic of Korea*
[2] *Department of Information and Communication Engineering, Fukuoka Institute of Technology (FIT), 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka, Japan*

Correspondence should be addressed to Young-Sik Jeong; ysjeong@dongguk.edu

Received 14 August 2013; Accepted 13 September 2013

Academic Editor: Ken Choi

In wireless sensor network, when these sensors are wrongly placed in an observation region, they can quickly run out of batteries or be disconnected. These incidents may result in huge losses in terms of sensing data from numerous sensors and their costs. For this reason, a number of simulators have been developed as tools for effective design and verification before the actual arrangement of sensors. While a number of simulators have been developed, simulation results can be fairly limited and the execution speed can be markedly slow depending on the function of each simulator. In this regard, to improve the performance of existing simulators, this research aimed to develop a parallel calculation simulator for independent sensor (PCSIS) that enables users to selectively use the GPU mode and, based on this mode, enables parallel and independent operations by matching GPU with many cores in order to resolve the slowdown of the execution speed when numerous sensor nodes are used for simulations. The PCSIS supports the GPU mode in an environment that allows the operation of compute unified device architecture (CUDA) and performs the parallel simulation calculation of multiple sensors using the mode within a short period of time.

1. Introduction

Today, wireless sensor networks (WSNs) are utilized by their integration into various fields in the real world. The data collected by sensing via WSNs are used in various service areas such as individual research, national projects, energy saving, luxury automobile systems, important social infrastructure (e.g., electricity, water), manufacturing, communication systems, weapon systems, robots with distributed processing on multiple computers, transportation control, and elderly people [1–6].

A number of factors should be considered to establish effective WSNs. Basic questions include which type of sensors to select for a target sensing region, which protocol to select for communication, how to arrange sensors, how many sensors to be used, how to decide the density of sensors, and how much budget to arrange for network establishment. Therefore, topology configuration is not an easy task. For this reason, various tools have been developed to arrange sensors and design and verify communication protocols. These include various types of simulators such as GloMoSim [7], ATEMU [8], NS2 [9], TOSSIM [10], AVRORA [11], SWANS [12], and SENSE [13].

Despite the development of a number of tools, their processing speed for large or small networks is not fast enough. In addition, as simulations are performed based on the information of simulator-dependent sensor nodes, substantially limited results are obtained. In addition, while the functional aspect of simulators is important, their execution speed cannot be neglected. Sensors in a simulator are based on independent operations, and thus, one sensor uses one thread. For this reason, an increase in the number of nodes significantly slows down the simulation speed [1, 6].

To overcome a marked slowdown of the execution speed according to an increase in the number of nodes set to run simulations in a WSN simulator, this paper intended to develop a PCSIS that offers the function of using many GPU cores. The PCSIS enables parallel operations by applying a node to each thread of the GPU at the ratio of 1 : 1. Using this sensor, sequential calculations of a large scale of sensor nodes can be processed simultaneously. Therefore, users are allowed to perform more accurate and speedy simulations.

This paper is organized as follows. Section 2 explains the operating mode of existing WSN simulators and briefly introduces CUDA to use GPUs. Section 3 explains the use of a GPU to improve the performance of a PCSIS proposed by this study. Section 4 introduces the design of the PCSIS. Section 5 presents a comparative explanation on the construction of the PCSIS and resulting improvements in the performance of existing simulators. Section 6 finally presents a summary and future research tasks.

2. Related Works

This section reviews CUDA to use the operating mode of existing WSN simulators that have been developed by relevant studies and GPUs.

2.1. Existing WSN Simulators. A number of tools have been developed as WSN simulators including GloMoSim [7], ATEMU [8], NS2 [9], TOSSIM [10], AVRORA [11], SWANS [12], and SENSE [13]. Table 1 examines the operating modes of the above simulators.

In addition, a research [14], which aimed for the speedier performance of AVRORA simulators, used a super computer. Java-based AVRORA sensor nodes and Java threads are configured at the ratio of 1 : 1. The results of an experiment that matched them with CPUs showed that an increase in the number of CPUs does not improve simulations. This is because context switches frequently occur as the number of threads exceeds the number of CPUs. In other words, this phenomenon occurs as the processing capacity for synchronization surpasses the improved processing speed of CPUs.

We propose a PCSIS that enables the fast simulation of threads with a maximum number of GPU cords by using a GPU that basically has more cores than a CPU. The PCSIS is a Java-based simulator that uses JCUDA (Java bindings for CUDA) to support the GPU.

2.2. CUDA for GPU. CUDA is the technology that implements parallel computing using GPUs, which was launched by NVDIA in November 2006. In an early stage, the company's CUDA was developed based on the C language but has evolved into CUDA 5.5 after continuous updates. This technology supports various standard programming languages such as C, C++, Java, Fortran, and Python. Its advantages include processing a large volume of parallel calculations and utilizing its full functions simply through the operation of NVIDIA's built-in devices. In addition, software, utilities, and example documents are provided free of charge [15].

CUDA is a powerful technology in terms of enabling parallel programming using GPUs. Programs supported by CUDA may differ slightly depending on the graphic device launched by NVIDIA. However, currently launched devices are upgraded from the previous versions and therefore can operate even formerly written programs. While programmers expect performance improvement through parallel teaks using CUDA, they might encounter lower levels of performance. Therefore, programmers should have the overall knowledge and understanding of CUDA. In addition, nvcc (NVIDIA C Compiler) should be installed to execute CUDA.

A general data flow for using CUDA is as follows.

(i) Allocate a memory necessary for tasks using CUDA in graphic cards.

(ii) Duplicate host data into device memory in the allocated memory (host refers to the RAM run in the CPU, and device refers to the RAM run in the GPU).

(iii) Perform calculations and tasks by summoning GPU kernels.

(iv) Duplicate the processing results from device to host to use them in the host (in case of graphic tasks, direct outputs can also be produced according to the interworking for types of use).

(v) The memory that was initially allocated to the graphic cards is reclaimed.

It should be noted that in performing parallel tasks for only calculations, improvement in the performance can be expected when minimizing the portion of duplications from host to device or from device to host and instead increasing the portion of parallel tasks.

As this study is focused on calculations through parallel tasks, it provides a solution to perform the area that requires a number of calculations for each sensor in the GPU.

3. GPU Functions on PCSIS

The PCSIS proposed, in this paper, was developed based on Java and provides the GPU mode to improve performance in existing WSN simulators.

Certain considerations are required for using the GPU mode as follows.

(i) First, as the PCSIS uses GPUs, an examination should be carried out to check whether GPUs can be used in the simulator.

(ii) An essential examination on operability is to check whether the graphic cards support CUDA.

(iii) Next, once an examination is finished on whether parallel thread execution (PTX) files can be run, the GPU mode is finally activated.

(iv) If the GPU is operable, selective GPU mode is activated. Otherwise, it is inactivated.

The number of nodes necessary for using the GPU mode in the PCSIS depends on graphic cards. Basically, to yield maximum efficiency in the parallel run of each node, each thread in the GPU should not be matched with more than one sensor node. In other words, each thread should perform calculations for a maximum of one sensor node. In addition, back collisions, registers, and local, shared, and global memories should be taken into account.

4. Design of PCSIS

In this paper, the PCSIS proposed is largely divided into the user interface, target area manager, interaction broker,

Table 1: Comparison with exiting WSN simulators.

Simulator	Function and operating mode
GloMoSim	(i) This is a simulator developed for large-scale simulation environments, and it uses the Parsec language to perform parallel simulations. (ii) The Parsec language was selected for the parallel tasks of each sensor. However, many sensor nodes that should be established in an actual environment with complete parallel tasks cannot be realized due to the limitation of parallel tasks in the CPU.
ATEMU	(i) This the first command-based simulator developed based on the C language. (ii) This can ensure cycle accuracy and set parameters for mutual different systems. (iii) This does not offer the GUI mode and produce outputs based on texts. As its sensor nodes perform sequential simulations, the execution speed is significantly low.
NS2	(i) This is a discrete event simulator that has a modular mode. (ii) Many occasions require the interaction between a network and application programs. In this respect, this simulator lacks application program models. (iii) Network animator (Nam) exists to support the GUI of NS2. However, this simulator stores event command files and reads them whenever the files are required. Therefore, it is ineffective by excluding simultaneity.
TOSSIM	(i) This runs the simulation of TinyOS, which is the OS of motes used in a sensor network. (ii) This simulation enables the inference of actual movements and the analysis of hardware-based influences. (iii) This cannot perform simulations in areas other than TinyOS and does not have cycle accuracy, an important factor for code debugging and functional verification. (iv) This has TinyViz, which was developed based on Java by providing a GUI for movements. However, it does not fully show the fluid movement of dynamic sensors.
AVRORA	(i) This is a Java-based simulator that enables the simultaneous simulations of multiple sensor nodes. (ii) In running simulations, each sensor node activates each thread. (iii) As this does not provide a GUI environment, users cannot quickly identify its operational status. Moreover, its CPU-dependent thread operating mode makes it difficult to obtain prompt simulation results.
SWANS	(i) Users can define models using a Java-based simulator. (ii) This graphically shows interactions and areas for the communication of networks. (iii) This can define various conditions of each sensor in terms of temperature, humidity, and movement. (iv) While each thread is logically judged as if run independently, the actual execution mode is CPU-dependent. Thus, this excludes parallel runs for a number of nodes.
SENSE	(i) This simulator was developed based on C++ in 2004. (ii) This lacks configurations for various WSNs and does not support visualization when using only SENSE. Therefore, it is incapable of the swift identification of simulations. (iii) While this has G-Sense as a visualization tool for SENSE, it does not properly process the dynamic movement of sensor nodes and is rather focused on results. (iv) Basically, this has a slow execution speed and thus requires improvement.

map manager, map controller, node manager, coordinate converter, and viewer based on functional terms. The user interface receives inputs from the user regarding the basic setting of sensor nodes and whether the GPU mode will be used or not. The target area manager manages sensing target regions established by the user. The interaction broker plays the role of connecting node and mapping values, which are set and input by the user, to the system. The map manager applies and manages the data of topographic information and performs calculations that apply mapping values input by the user. The node manager performs tasks by either using the CPU only or using the GPU installed to improve performance according to the information of model selection received from the user. The coordinate converter plays the role of processing data to send the operating conditions of simulations to the view in order to show them to the user. Finally, the viewer shows the conditions of simulations in the PCSIS to the user in a visual form. Figure 1 presents the architecture about the overall functions of the PCSIS.

The *user interface component* is further segmented into the map interface, node interface, and OP mode. The map interface is the place that receives the geography markup language (GML) that can be mapped on actual topography, which is input by the user. It can also set information on the location of preferred target regions. The node interface consists of range control, which can set basic sensor information, such as the sensing range, communication range, and supersonic wave range, and configuration on whether the status of range control should be shown to the viewer or not, which includes the sensing range view (SR-V), communication range view (CR-V), supersonic wave range (SWR-V), node trace line view (NTL-V), and node connection view (NC-V). The OP mode makes self-judgment on whether the JCUDA is operable to use the GPU. If the interface is operable, the OP mode provides two types of modes, which are base and GPU modes, to enable the user to selectively operate it.

The *target area manager component* reads actual topographic data via the GML importer of the map manager in order to provide more expanded tests using the actual data and set the area of target regions that require observations in relation to the analyzed GML documents.

The *interaction broker component* plays the role of a broker that analyzes the user-input basic setting of operating modes and sensors and map control messages and then

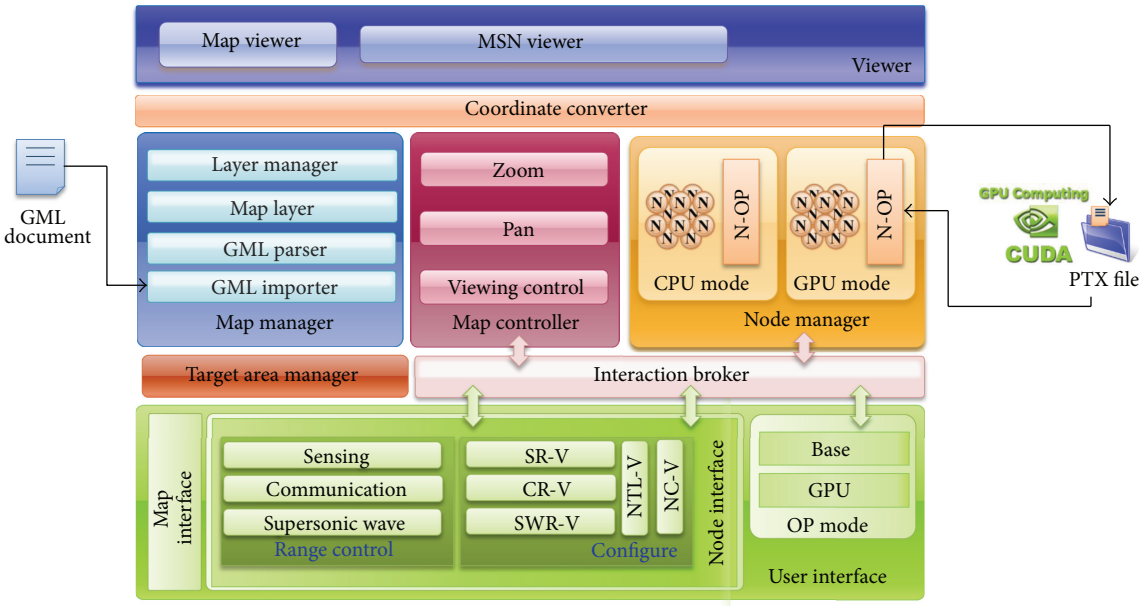


Figure 1: Architecture of PCSIS for high performance of WSN Simulator.

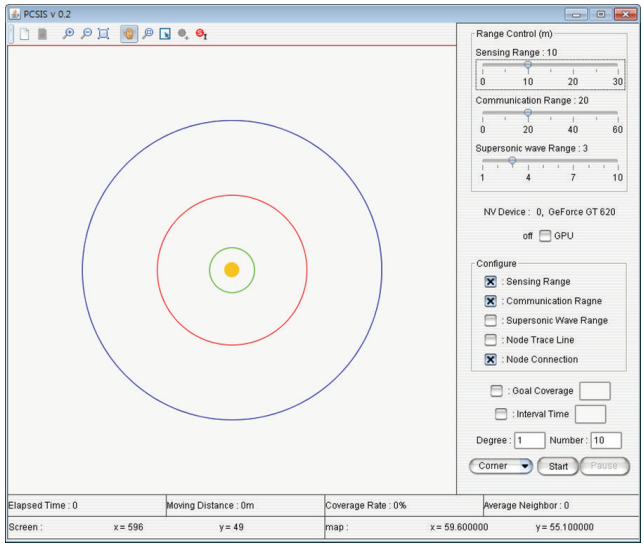


Figure 2: The initial execution status of PCSIS.

sends the analysis results to the map controller and the node manager.

The *map manager component* plays the role of applying managing GML documents that can be mapped on actual topography. In detail, this component consists of the following subcomponents: the GLM importer, which is necessary for adding GLM documents, which are selected by the user using the map interface of the user interface, to the PCSIS, the GML parser for analyzing the added GML documents to apply them to the PCSIS, the map layer for producing map objects by judging the presence of obstacles according to the objects of GML topographic data and then sending them to the layer manager, and the layer manager that provides and manages the polygons (e.g., building), polylines (e.g., road, street, and track), and texts (e.g., building name, road name) of topographic data received from the map layer according to the user's selection.

The *map controller component* performs expansion, contraction, area expansion, and selective movement for maps managed by the layer manager based on values defined and sent from the user. Such functions can be activated if the relevant data are received by the user's input via the map interface of the user interface. The corresponding results are displayed to the user by outputting them into the viewer using the coordinate converter.

The node manager component applies and manages node values input by the user. This is further divided into CPU and GPU modes and operates in one of the two modes using the OP mode of the user interface. In the CPU mode, the PCSIS operates in the same manner as general WSN simulators. In the GPU mode, the node manager component assesses whether the number of nodes designated for activation will be able to perform simulations. If no abnormality is detected, a common arithmetic unit across the nodes is produced into a PTX file. This PTX file is produced if the PCSIS is first run or no PTD files exist internally. If a PTX file already exists, the existing file is reused. The node-operator (N-OP) is the place to calculate the next location of mobile sensor nodes. In the GPU mode, this component converts data to enable parallel computation on the locations of sensor nodes. The converted data have the information necessary for the computation of each sensor node in the Java language.

The *coordinate converter component* plays the role of processing and sending data on the basic information of topography and nodes and the operating conditions of the nodes in a form that can be displayed at the viewer.

The *viewer component* visually presents the processed data delivered via the coordinate converter to the user.

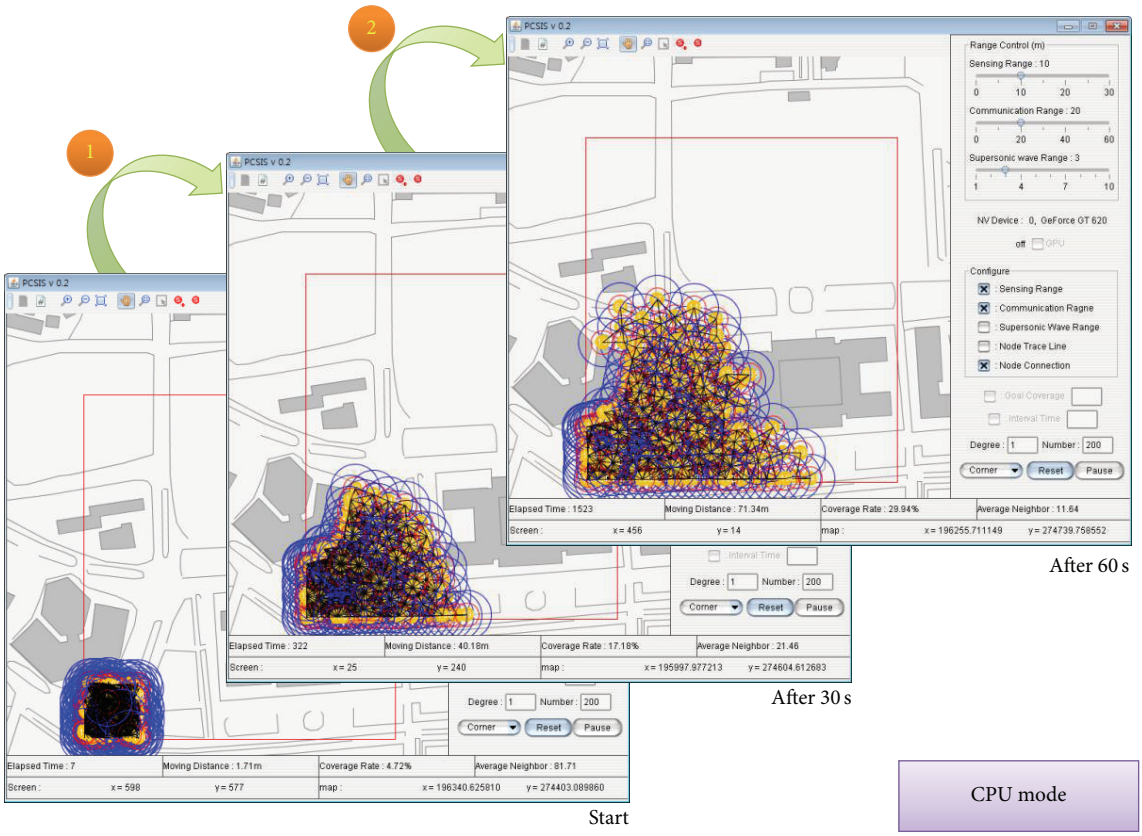


Figure 3: Simulation for 200 sensor nodes in the CPU mode.

The user can view the operating conditions of simulations through the viewer and draw expected problems by assessing and analyzing the conditions.

5. Implementation of PCSIS and Performance Evaluation

5.1. Implementation of PCSIS. The initial execution status of the PCSIS is shown in Figure 2. The tool bar in the upper part of Figure 2 consists of a button to read GML documents that can be mapped on actual topography, a button to expand, reduce, and move the views and select target observation regions, a button to add sensor nodes to an operating simulator, and a button for the matching of coordinates between moving sensor nodes and GML documents. The control view on the right side of Figure 2 shows the menu "range control" that enables the control of sensing ranges, communication, and supersonic wave ranges for sensor nodes. It also provides a checkbox to set whether the GPU mode will be used or not and selective visualization regarding the traces of range and movement for each sensor and the connection between sensors in order to configure the display of simulations during the run of simulations. The setup part for GPU mode in Figure 2 outputs the NV device that displays the identification number of each graphic device and the graphic model name in an operating environment. In addition, it can set the coverage of target regions as the range for the completion of simulations and define the basic number of operating sensor nodes when starting simulations. Figure 2 provides visualization on the variables of elapsed time, moving distance, coverage rate, average neighbor, and coordinates on the screen and actual maps.

Figure 3 shows the screens in which target regions and sensor nodes are set after reading GML documents that can be mapped on actual topography and run a simulation. ① and ② illustrate a sequential flow of views. Each view exhibits the conditions at the start of simulation, after 30 seconds of simulation, and after 60 seconds of simulation.

Figure 4 shows a simulation run by selecting the GPU mode in the same sensor condition as that in Figure 3. ① and ② exhibit a sequential flow of views. Compared with Figure 3, performance differences in the movement of each node can be confirmed with the naked eye.

5.2. Performance Evaluation. This section compares the execution speeds of using only the CPU and using the GPU in the PCSIS proposed by this study. The comparison was performed by examining performance differences in terms of temporal outputs through the computation of the coverage rate with time, while using the same number of sensor nodes within the same target regions. In addition, the status of performance according to the number of nodes was compared by increasing the number by 100 nodes each time.

Figure 5 presents the coverage rates with time when using 100 sensor nodes. The coverage rate is the percentage of the

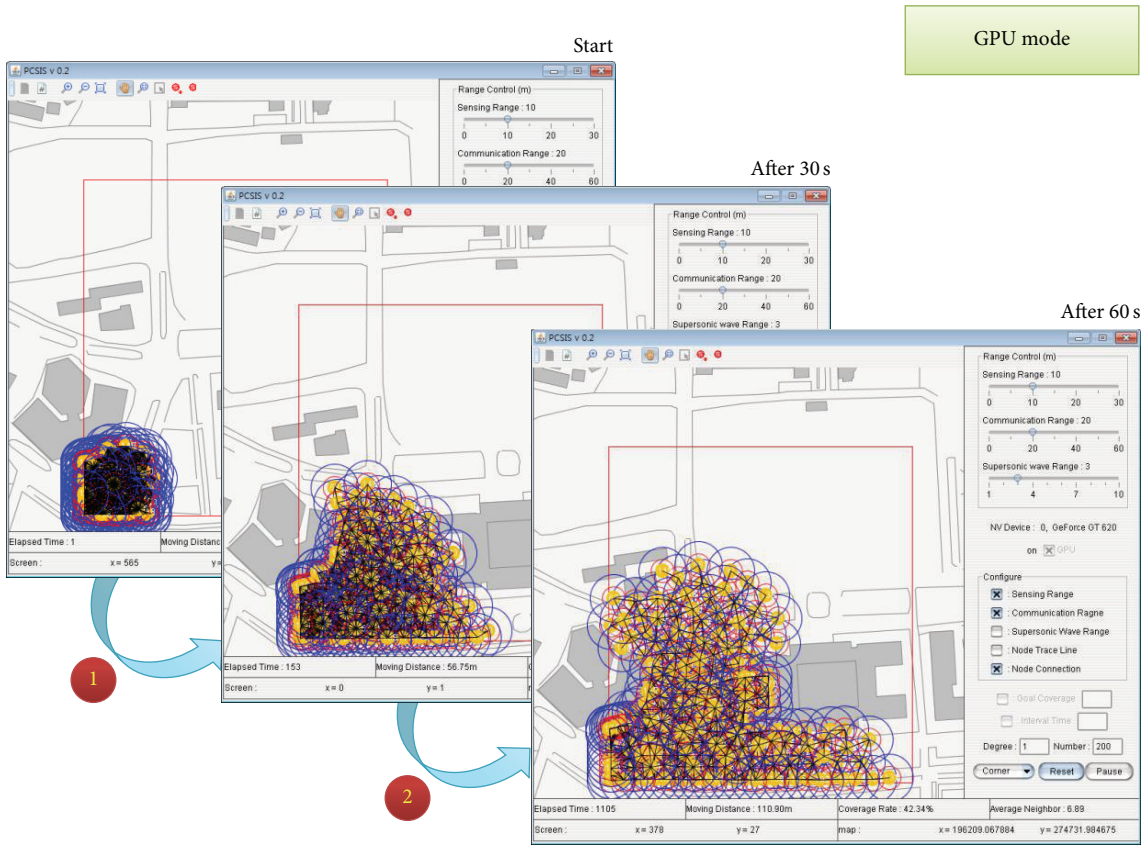


FIGURE 4: Simulation for 200 sensor nodes in the GPU mode.

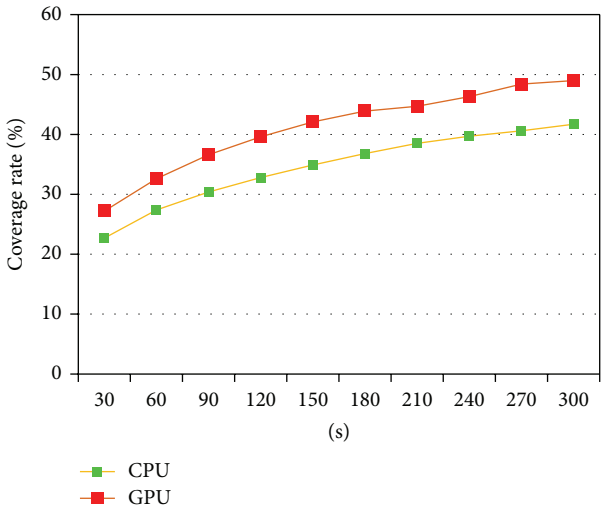


FIGURE 5: Coverage rates with time in the case of using 100 sensor nodes.

sensing area of sensor nodes to the target sensing region. The graph shows two situations of using the CPU and the GPU according to the calculation method for the next location of each sensor. Figure 5 exhibits that the coverage rates are similar between 60 seconds of using the GPU and 120 seconds of using the CPU.

This indicates that a task performed by using the CPU for 120 seconds can be achieved by using the GPU only for 60 seconds.

Figure 6 presents the results of using 100, 200, and 300 sensor nodes, respectively, by increasing the number of sensors by 100 nodes each time. The figure proves that each increase in the number of nodes resulted in a much greater level of performance improvement when using the GPU than using only the CPU.

However, if the number of nodes exceeds the number of threads that can be run at a time in a graphic device, their operation becomes rather ineffective. A key reason is that an increase in the number of nodes requires a corresponding increase in the data that should be duplicated from host to device, which, in turn, increases the delay time. In addition, the time required for running a number of threads in the CPU and showing the progress on the screen becomes longer than the time for the duplication of data transmitted from host to device. Moreover, an excess number of threads may cause a lot of processing, which subsequently increases the execution time.

6. Conclusion and Future Works

Existing WSN simulators have the disadvantage that an increase in the number of sensor nodes necessary for running simulations markedly slows down the execution speed. In this

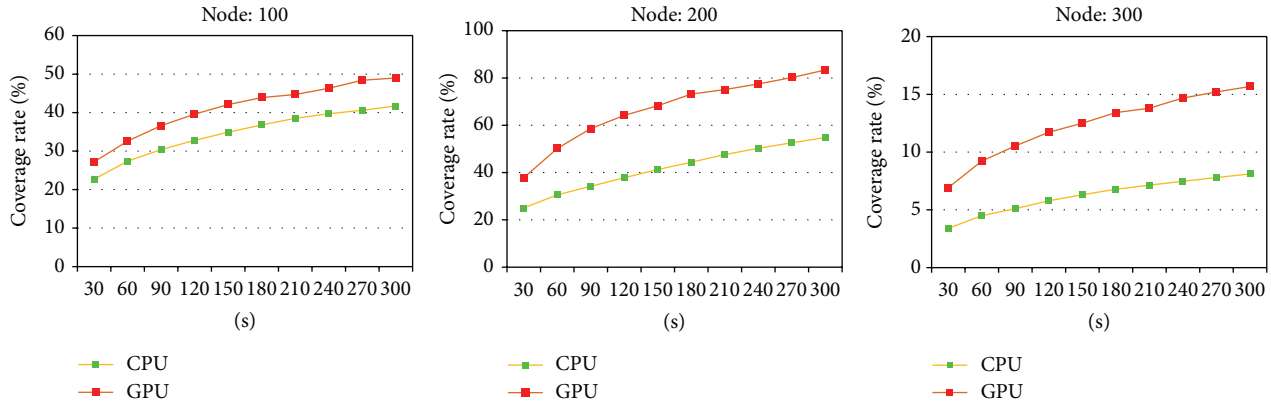


Figure 6: Performances by increasing the number of sensor nodes by 100 nodes.

respect, this study proposed a PCSIS that provides the GPU mode aimed at improving the performance of existing WSN simulators. The function of the GPU mode is the parallel processing of arithmetic units that should be separately calculated for each sensor node by using GPU cores. In addition, the results of increasing the number of nodes and the operating time of the simulator revealed that using the GPU yielded much faster execution speeds than using only the CPU. This demonstrates that the use of GPUs is effective for improving the performance of simulators. A follow-up study is planned to enable the incorporation of modularity into most WSN simulators by applying the parallel run of GPUs. In addition, given that the types of graphic devices increase with time, research will be performed to provide an interface that allows the user to define the number of grids, blocks, and threads per block.

Acknowledgments

This research is supported by the MSIP (Ministry of Science, ICT and Future Planning), Republic of Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2013-H0301-13-4007) supervised by the NIPA (National IT Industry Promotion Agency).

References

[1] B. Musznicki and P. Zwierzykowski, "Survey of simulators for wireless sensor networks," *International Journal of Grid and Distributed Computing*, vol. 5, no. 3, pp. 23–50, 2012.

[2] J. Chen, M. B. Salim, and M. Matsumoto, "A single mobile target tracking in Voronoi-based clustered wireless sensor network," *Journal of Information Processing Systems*, vol. 6, no. 4, pp. 17–28, 2010.

[3] A. U. Bandaranayake, V. Pandit, and D. P. Agrawal, "Indoor link quality comparison of IEEE 802. 11a channels in a multi-radio Mesh network testbed," *Journal of Information Processing Systems*, vol. 8, no. 1, pp. 1–20, 2012.

[4] S. Silas, K. Ezra, and E. B. Rajsingh, "A novel fault tolerant service selection framework for pervasive computing," *Human-Centric Computing and Information Sciences*, vol. 2, no. 5, pp. 1–14, 2012.

[5] X. Zhou, Y. Ge, X. Chen, Y. Jing, and W. Sun, "A distributed cache based reliable service execution and recovery approach in MANETs," *Journal of Convergence*, vol. 3, no. 1, pp. 5–12, 2012.

[6] Y. S. Jeong, Y. H. Han, J. J. Park, and S. Y. Lee, "MSNS: mobile sensor network simulator for area coverage and obstacle avoidance based on GML," *EURASIP Journal on Wireless Communications and Networking*, vol. 95, no. 1, pp. 1–15, 2012.

[7] L. Bajaj, M. Takai, R. Ahuja, K. Tang, R. Bagrodia, and M. Gerla, "GlomoSim: a scalable network simulation environment," UCLA CSD Technical Report #990027, UCLA, 1999.

[8] J. Polley, D. Blazakis, J. McGee, D. Rusk, J. S. Baras, and M. Karir, "ATEMU: a fine-grained sensor network simulator," in *Proceedings of the 1st Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks (IEEE SECON '04)*, pp. 145–152, October 2004.

[9] "The Network Simulator—ns—2," http://www.isi.edu/nsnam/ns/.

[10] P. Levis, N. Lee, M. Welsh, and D. Culler, "TOSSIM: accurate and scalable simulation of entire TinyOS applications," in *Proceedings of the 1st International Conference on Embedded Networked Sensor Systems (SenSys '03)*, pp. 126–137, November 2003.

[11] B. L. Titzer, D. K. Lee, and J. Palsberg, "Avrora: scalable sensor network simulation with precise timing," in *Proceedings of the 4th International Symposium on Information Processing in Sensor Networks (IPSN '05)*, pp. 477–482, April 2005.

[12] "Java in simulation time/scalable wireless Ad hoc network simulator," http://jist.ece.cornell.edu/.

[13] G. Chen, J. Branch, M. J. Pflug, L. Zhu, and B. K. Szymanski, "SENSE: a wireless sensor network simulator," http://www.ita.cs.rpi.edu/publications/sense-book-chapter.pdf.

[14] H. Joe, S. Y. Yoon, J. Hong, and H. Kim, "The sensor network simulation on the supercomputer," in *Proceedings of the Korea Computer Congress (KCC '11)*, vol. 38, pp. 442–445, 2011.

[15] "NVIDIA CUDA," http://www.nvidia.com/content/global/global.php.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2013, Article ID 169638, 9 pages
http://dx.doi.org/10.1155/2013/169638

Research Article

Security Analysis of Scalable Block Cipher PP-1 Applicable to Distributed Sensor Networks

Yuseop Lee,[1] Kitae Jeong,[1] Jaechul Sung,[2] Changhoon Lee,[3] Seokhie Hong,[1] and Ku-Young Chang[4]

[1] *Center for Information Security Technologies (CIST), Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Republic of Korea*
[2] *Department of Mathematics, University of Seoul, Jeonnong-dong, Dongdaemun-gu, Seoul 130-743, Republic of Korea*
[3] *Department of Computer Science and Engineering, Seoul National University of Science and Technology, 232 Gongneung-ro, Nowon-gu, Seoul 139-743, Republic of Korea*
[4] *Electronics and Telecommunication Research Institute, 218 Gajeong-ro, Yuseong-gu, Daejeon 305-700, Republic of Korea*

Correspondence should be addressed to Changhoon Lee; chlee@seoultech.ac.kr

Received 12 August 2013; Accepted 22 August 2013

Academic Editor: Jongsung Kim

PP-1 is a scalable block cipher which can be implemented on a platform with limited resource. In this paper, we analyze the security of PP-1 by using truncated differential cryptanalysis. As concrete examples, we consider four versions of PP-1, PP-1/64, PP-1/128, PP-1/192, and PP-1/256. Our attack is applicable to full-round versions of them, respectively. The proposed attacks can recover a secret key of PP-1 with the computational complexity which is faster than the exhaustive search. These are the first known cryptanalytic results on PP-1.

1. Introduction

Recently, the research on lightweight block ciphers has received considerable attention. Since these can be efficiently implemented under restricted resources such as low-cost, low-power, and lightweight platforms, they are applicable to low-end devices such as RFID tags, sensor nodes, and smart devices [1–6]. So far, many lightweight block ciphers (e.g., HIGHT [7], CLEFIA [8], KATAN/KTANTAN [9], PRINTCIPHER [5], and PP-1 [10]) have been proposed.

PP-1 is an involutional SPN block cipher which can be implemented on a platform with limited resources. It supports the scalability, which allows using different data block sizes and secret key sizes. In detail, PP-1 is an n-bit scalable block cipher and supports $n/2n$-bit secret keys. ($n = 64, 128, 192, \ldots$). It uses an 8×8 S-box which is an involution and a bit-oriented permutation which is also an involution. As a result, it is a totally involutional cipher. To our knowledge, there is no cryptanalytic result on PP-1.

In this paper, we analyze the security of PP-1 on truncated differential cryptanalysis. As concrete examples, we consider four versions of PP-1, PP-1/64, PP-1/128, PP-1/192, and PP-1/256. Here, 64, 128, 192, and 256 indicate the length of data blocks. Our attack is applicable to full-round versions of them, respectively. Our attack results are summarized in Table 1. Here, PP-1/n_k means an n-bit PP-1 which supports a k-bit secret key. Note that since our attacks do not use the property of the key schedule of PP-1, the data complexity and the memory complexity of the attacks on PP-1/n_k and PP-1/n_2k have the same value. From this table, our attacks can recover a secret key of PP-1 with the computational complexity which is faster than the exhaustive search. These results are the first known cryptanalytic results on PP-1.

The rest of this paper is organized as follows. In Section 2, we briefly present PP-1. In Section 3, differentials on PP-1 are derived, and their probabilities are computed. Truncated differential cryptanalysis on each version of PP-1 is proposed in Sections 4, 5, and 6, respectively. Finally, we give our conclusion in Section 7.

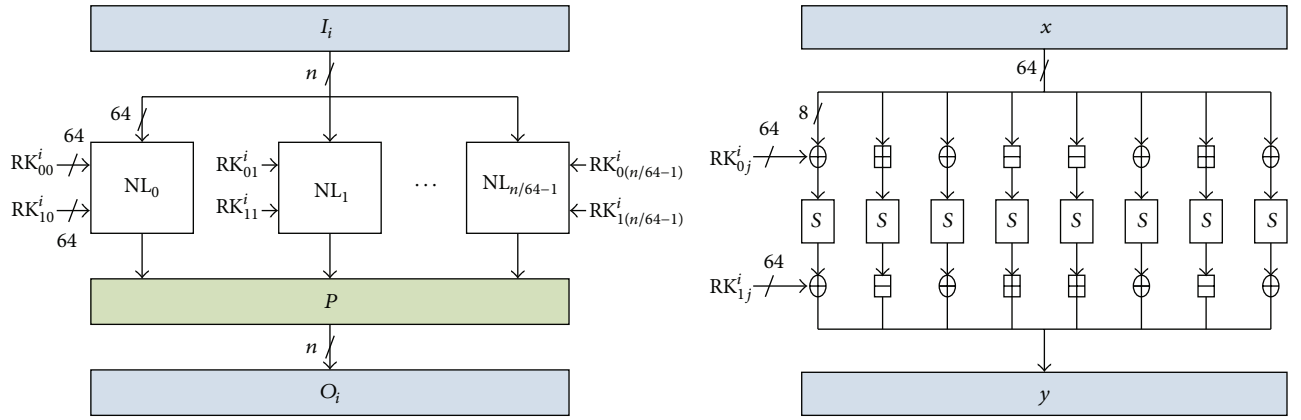


Figure 1: (a) The round function of PP-1 and (b) the nonlinear function NL_j.

Table 1: Our attack results on PP-1.

Target algorithm	Data complexity	Memory complexity	Computational complexity
PP-1/64_64	$2^{45.29}$ CP	$2^{41.29}$ bytes	$2^{45.29}$ encryptions
PP-1/64_128			$2^{48.21}$ encryptions
PP-1/128_128	$2^{103.45}$ CP	$2^{44.45}$ bytes	$2^{103.45}$ encryptions
PP-1/128_256			2^{168} encryptions
PP-1/192_192	$2^{157.85}$ CP	$2^{35.44}$ bytes	$2^{157.85}$ encryptions
PP-1/192_384			2^{296} encryptions
PP-1/256_256	$2^{210.84}$ CP	$2^{24.84}$ bytes	$2^{210.84}$ encryptions
PP-1/256_512			2^{432} encryptions

PP-1/n_k: an n-bit PP-1 with a k-bit secret key.
CP: chosen plaintexts.

2. Description of PP-1

PP-1 is an n-bit scalable block cipher and has r-round SPN structure ($n = 64, 128, 192, \ldots$). The length of a secret key is n or $2n$ bits. In [10], the designers of PP-1 proposed the following four versions of PP-1 as concrete examples.

(i) PP-1/64: $n = 64$, $r = 11$, that is, a 64-round block cipher with 64/128 bit secret keys and 11 rounds.

(ii) PP-1/128: $n = 128$, $r = 22$, PP-1/192: $n = 192$, $r = 32$, PP-1/256: $n = 256$, $r = 43$.

For the simplicity of notations, we denote an n-bit PP-1 with a k-bit secret key PP-1/n_k. And input/output values and a round key in round i are denoted by I^i, O^i, and RK^i, respectively ($i = 1, \ldots, r$).

We omit the key schedule of PP-1, as it is not effectively used in our attack.

2.1. The Round Function. As shown in Figure 1(a), the round function of PP-1 consists of (n/64) nonlinear NL functions ($NL_0, \ldots, NL_{(n/64-1)}$) and an n-bit involutional permutation P. For example, when $n = 64$, the round function uses only NL_0. Note that P is not conducted in the last round. ($NL_0, \ldots, NL_{(n/64-1)}$) have the same structure but different round keys are used.

In round i, two n-bit round keys $RK^i = (RK^i_0, RK^i_1)$ are used as follows:

$$RK^i_0 = RK^i_{00} \| \cdots \| RK^i_{0(n/64-1)},$$
$$RK^i_1 = RK^i_{10} \| \cdots \| RK^i_{1(n/64-1)}. \quad (1)$$

Here, NL_j takes 128 bit sub-round key (RK^i_{0j}, RK^i_{1j}) ($j = 0, \ldots, (n/64 - 1)$).

2.2. The Nonlinear Function NL. In round i, NL_j outputs a 64 bit value from a 64 bit input value and a 128 bit subround key (RK^i_{0j}, RK^i_{1j}) ($j = 0, \ldots, (n/64-1)$). It consists of one 8×8 S-box S, XOR($\oplus$), addition($\boxplus$), and substraction($\boxminus$) modulo 2^8 (see Figure 1(b)).

A 128 bit sub-round key (RK^i_{0j}, RK^i_{1j}) is divided into eight 8 bit elementary keys as follows, respectively:

$$RK^i_{0j} = RK^i_{0j0} \| \cdots \| RK^i_{0j7},$$
$$RK^i_{1j} = RK^i_{1j0} \| \cdots \| RK^i_{1j7}. \quad (2)$$

Thus, each elementary key is XORed or added or subtracted with an 8 bit intermediate value. For example, RK^i_{1j0} is XORed with the 8 bit output value of the first S-box.

2.3. The Permutation Function P. P is an n-bit involutional bit-oriented permutation. It is constructed by using two algorithms, the auxiliary algorithm (Algorithm 1) to compute auxiliary permutation Prm and the main algorithm (Algorithm 2) to compute permutation P.

For example, the 128 bit permutation P is obtained as a result of 64 calls of Algorithm 2 for a pair numbered as pno from 1 to 64, the number of block bits $nBb = 128$ and the number of S-box bits $nSb = 8$. When $pno = 2$, the value y of Prm is equal to 9 and the resultant pair $(px, py) = (3, 18)$. It means that the third bit of the input value is mapping to the eighteenth bit of the output value.

(1) $nS \leftarrow nBb \text{ div } nSb$.
(2) $Sno \leftarrow (x \bmod nS) + 1$.
(3) $Sb \leftarrow ((x-1) \text{ div } nS) + 1$.
(4) $y \leftarrow (Sno - 1) \cdot nSb + Sb$.
(5) Return y.

Algorithm 1: Prm(x, nBb, nSb).

(1) $y \leftarrow \text{Prm}(pno, nBb \text{ div } 2, nSb \text{ div } 2)$.
(2) $px \leftarrow 2 \cdot pno - 1$.
(3) $py \leftarrow 2 \cdot y$.
(4) Return (px, py).

Algorithm 2: $P(pno, nBb, nSb)$.

3. Construction of Differentials on PP-1

In this section, we introduce the methodology of constructing differentials on PP-1 used in our attacks. For the simplicity of notations, we define the following notations.

(i) $(\alpha \rightarrow \beta)_t$: a t-round differential characteristic where input/output differences are α and β, respectively.

(ii) $[\alpha \rightarrow \beta]_t$: a t-round differential where input/output differences are α and β, respectively.

(iii) (a, x): a byte string where the ath byte value is x and the other bytes are zero (the index of the left most byte is 0).

3.1. Differential Characteristic on PP-1. In general, a differential characteristic with the higher probability passes less nonlinear operations, such as S-box, than it with the lower probability. Recall that a *NF*-function consists of S-box, addition, substraction, and XOR. Among these operations, nonlinear operations are S-box, addition, and subtraction. Thus, in order to construct a differential characteristic with a high probability, we should avoid them.

We examined such differential characteristics on PP-1. As a result, we found several t-round differential characteristics with a probability of $2^{-7 \cdot t}$. For example, in the case of PP-1/64, we can construct $((7, \texttt{0x01}) \rightarrow (7, \texttt{0x01}))_t$ with a probability of $2^{-7 \cdot t}$. This characteristic passes only one S-box in each round and the probability that S-box outputs an output difference 0x01 from an input difference 0x01 is 2^{-7}.

We expect that this type of difference characteristics have the highest probability. That is, they pass least S-boxes, addition operations, and subtraction operations. We extend it to differentials in the next subsection.

3.2. Finding of Differentials with a High Probability. The probability of a differential is computed by adding the probabilities of all differential characteristics which are included in it. The more differential characteristics with a high probability a differential includes, the higher its probability is. Thus, in order to find a differential on PP-1 with a high probability, we consider the following criteria.

(i) In each round, a differential characteristic has only one active S-box.

(ii) In each round, a differential characteristic does not pass addition/subtraction operations.

The probabilities that all t-round differential characteristics satisfying the above criteria are at least $2^{-7 \cdot t}$, since the minimum probability from the difference distribution table on S-box is 2^{-7}. Thus, we measure the probability of a differential by counting only the number of differential characteristics which satisfy the above criteria and are included in it. That is, if there are w such differential characteristics, the probability of a differential including them is at least $w \cdot 2^{-7 \cdot t}$. On the other hand, in the case of differential characteristics which do not satisfy the above criteria, they pass the additional nonlinear operations in each round. In this case, the probabilities of them are much smaller than $2^{-7 \cdot t}$. Thus, we expect that differential characteristics which do not satisfy the above criteria depend on the probability of a differential less.

In order to count efficiently differential characteristics satisfying the above criteria, we consider differences δ's satisfying the following conditions.

(i) $\delta = (a, x)$ where $(a \bmod 8) \in \{0, 2, 5, 7\}$ and x is a nonzero byte value.

(ii) $P(\delta) = (b, y)$ where $(b \bmod 8) \in \{0, 2, 5, 7\}$ and y is a nonzero byte value.

Let $\mathscr{D}$ be a set containing such δ's. We can easily prove that all t-round differential characteristics satisfying the above criteria include $(\delta_i \rightarrow \delta_j)_1$ in each round $(\delta_i, \delta_j \in \mathscr{D})$. It means that we only need to consider $\mathscr{D}$ in order to find all differential characteristics holding the above criteria.

Let $N(\delta_i, \delta_j, t)$ be the number of $(\delta_i \rightarrow \delta_j)_t$. For each δ_i and δ_j included in $\mathscr{D}$, we compute w using the following recurrence relation:

$$N\left(\delta_i, \delta_j, t+1\right) = \sum_{\delta_k \in \mathscr{D}} N\left(\delta_i, \delta_k, 1\right) \cdot N\left(\delta_k, \delta_j, t\right). \quad (3)$$

Since $N(\delta_i, \delta_k, 1)$ is compute by using the difference distribution table for an S-box of PP-1, we can easily compute $N(\delta_i, \delta_j, t)$ for given t.

For example, we found twenty one δ's for PP-1/64 (see Table 2). As a simulation result, $N((7, \texttt{0x01}), (7, \texttt{0x09}), 8)$ is 8429. It means that the probability of $[(7, \texttt{0x01}) \rightarrow (7, \texttt{0x09})]_8$ is $2^{-42.96} (= 8429 \cdot 2^{-7 \cdot 8})$.

4. Truncated Differential Analysis on PP-1/64

In this section, we propose truncated differential analysis on full-round PP-1/64_64 and full-round PP-1/64_128. Since PP-1/64_64 and PP-1/64_128 have the same structure except the key schedule, the attack procedures on them are similar. Thus, we mainly introduce the attack procedure on PP-1/64_64.

Table 2: A difference set $\mathscr{D}$ for PP-1/64.

(0, 0x20)	(0, 0x04)	(0, 0x01)	(2, 0x40)	(2, 0x10)	(2, 0x20)	(2, 0x30)
(5, 0x02)	(5, 0x04)	(5, 0x80)	(5, 0x84)	(5, 0x01)	(5, 0x08)	(5, 0x09)
(7, 0x02)	(7, 0x04)	(7, 0x80)	(7, 0x84)	(7, 0x01)	(7, 0x08)	(7, 0x09)

By using the method in the previous section, we construct forty nine 8-round differentials $[P((7,\alpha)) \rightarrow (7,\beta)]_8$ ($\alpha, \beta \in$ {0x01, 0x02, 0x04, 0x08, 0x09, 0x80, 0x84}). And we extend these 8-round differentials to total 1785(= $7 \cdot 255$) 9-round differentials $[P((7,\alpha)) \rightarrow P((7, x_j^i))]_9$ ($x_j^i \in X^i$). Here, X^i's are defined as follows ($i = 1, \ldots, 6$):

$$
\begin{aligned}
X^0 &= \{00000000_2\},\\
X^1 &= \{0000?00?_2 \mid ? \in \{0,1\}\} - X^0,\\
X^2 &= \{0?00?00?_2 \mid ? \in \{0,1\}\} - \bigcup_{i=0}^{1} X^i,\\
X^3 &= \{??00??0?_2 \mid ? \in \{0,1\}\} - \bigcup_{i=0}^{2} X^i,\\
X^4 &= \{??00????_2 \mid ? \in \{0,1\}\} - \bigcup_{i=0}^{3} X^i,\\
X^5 &= \{???0????_2 \mid ? \in \{0,1\}\} - \bigcup_{i=0}^{4} X^i,\\
X^6 &= \{????????_2 \mid ? \in \{0,1\}\} - \bigcup_{i=0}^{5} X^i.
\end{aligned}
\tag{4}
$$

4.1. Construction of Structures. We consider a structure S_i consisting of 256 plaintext; that is, $S_i = \{(i \parallel j) \mid j = 0, 1, 2, \ldots, 255\}$ where i is a 56 bit fixed value. Then we can compose 2^{15} plaintext pairs for each structure. Among these plaintext pairs, there are 2^7 plaintext pairs where an input difference of round 2 is one of $P((7,\alpha))$ for each α. Table 3 presents the expected number of right plaintext pairs where an output difference of round 10 is included in $P((7, X^i))$. These values are computed as follows:

$$
\sum_{\alpha} \sum_{x_j^i \in X^i} 2^7 \cdot \Pr\left[\left[\alpha \longrightarrow x_j^i\right]_9\right]. \tag{5}
$$

4.2. Truncated Differential Analysis on PP-1/64. The main idea of our attack is to exploit the fact that the expected number of plaintext pairs where an output difference of round 10 is included in $P(X^i)$ is $2^{-30.66} \sim 2^{-35.29}$ for each structure (see Table 3).

In our attack on PP-1/64_64, we first obtain a 72 bit partial information on $\mathrm{RK}^{11} = (\mathrm{RK}_0^{11}, \mathrm{RK}_1^{11})$ and an 8 bit RK_{007}^1. The attack procedure is as follows (see Figure 2).

(1) Choose $2^{37.29}$ structures which are composed of 256 plaintexts and obtain the corresponding ciphertexts. From these ciphertexts, compute $2^{52.29} (= 2^{15} \cdot 2^{37.29})$ ciphertext pairs (C^i, C^{i*}) (note that we can compute total 2^{15} ciphertext pairs for each structure).

(2) Check that the difference ΔC^i between ciphertext pair (C^i, C^{i*}) is 0x????00??00??????; that is, $\Delta C_{02}^i = \Delta C_{04}^i =$ 0x00 for each i($? \in \{0,1\}^4$). We keep all ciphertext pairs passing Step (2) and the corresponding plaintext pairs in a table and call a set containing them $\mathscr{A}$.

(3) Filter out the ciphertext pairs where ΔC_{00}^i, ΔC_{01}^i, ΔC_{03}^i, ΔC_{05}^i and ΔC_{06}^i are not zero in $\mathscr{A}$. Do the following for the remaining ciphertext pairs:

  (a) Guess an 8 bit RK_{107}^{11} (note that this substep indicates "Guess 1" in Figure 2).

  (b) Partially encrypt all remaining ciphertext pairs with the guessed round key RK_{107}^{11} to get ΔI_{07}^{11}. Check that ΔI_{07}^{11} is included in $P(X^1)$ from (4). If it is included in $P(X^1)$, add the counter, corresponding to the guessed key, to one.

  (c) Output a guessed key which has the maximal counter as a right RK_{107}^{11}.

(4) From $\mathscr{A}$, filter out the ciphertext pairs where $\Delta C_{00}^i, \Delta C_{03}^i$, ΔC_{05}^i, and ΔC_{06}^i are not zero and the ciphertext pairs considered in Step (3). Do the following for the remaining ciphertext pairs:

  (a) Check that ΔC_{07} is a nonzero value for each remaining ciphertext pair. Partially encrypt the ciphertext pairs passing this test with the recovered round key RK_{107}^{11} to obtain ΔI_{07}^{11}. If this value is not included in $P(X^1)$, filter out the corresponding ciphertext pairs.

  (b) Guess 16 bit round keys $(\mathrm{RK}_{001}^{11}, \mathrm{RK}_{101}^{11})$. ("Guess 2" in Figure 2).

  (c) Similarly to Step (3)(b), partially encrypt the remaining ciphertext pairs with the guessed round keys to get ΔI_{01}^{11}. Check that ΔI_{01}^{11} is

Table 3: The expected number of right pairs for PP-1/64.

Set of output differences of round 10	The expected number of right pairs
$P((7, X^1))$	$2^{-35.29}$
$P((7, X^2))$	$2^{-35.27}$
$P((7, X^3))$	$2^{-33.19}$
$P((7, X^4))$	$2^{-32.65}$
$P((7, X^5))$	$2^{-31.53}$
$P((7, X^6))$	$2^{-30.66}$

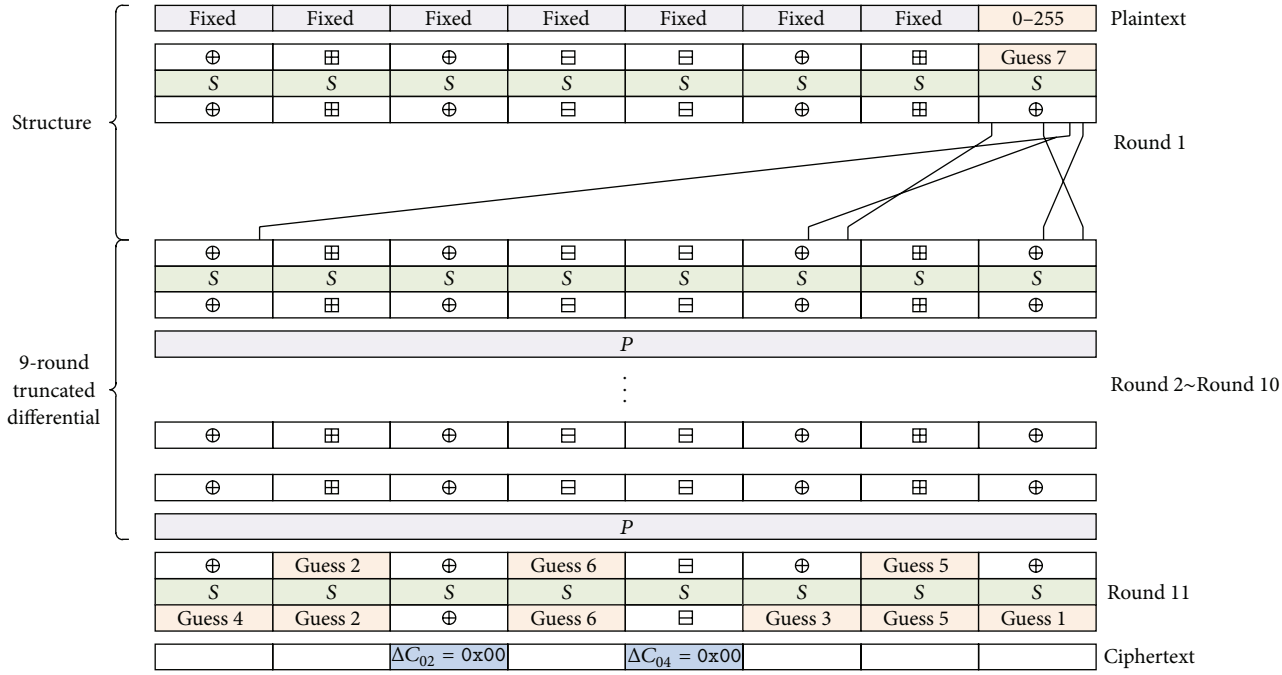


Figure 2: The attack procedure on full-round PP-1/64_64.

included in $P(X^2)$ from (4). If it is included in $P(X^2)$, add the counter, corresponding to the guessed key, to one.

(d) Output a guessed key which has the maximal counter as right RK_{001}^{11} and RK_{101}^{11}.

(5) From $\mathscr{A}$, discard the ciphertext pairs where ΔC_{00}^{i}, ΔC_{03}^{i}, and ΔC_{06}^{i} are not zero and the ciphertext pairs considered in Step (3) and (4). Do the following for the remaining ciphertext pairs:

(a) Similarly to Step (4)(a), filter out the ciphertext pairs where ΔI_{07}^{11} and ΔI_{01}^{11} are not included in $P(X^1)$ and $P(X^2)$, respectively.

(b) Guess an 8 bit round keys RK_{105}^{11}. ("Guess 3" in Figure 2).

(c) Similarly to Step (4)(c), check that ΔI_{05}^{11} is included in $P(X^3)$. Output a guessed key which has the maximal counter as a right RK_{105}^{11}.

(6) From $\mathscr{A}$, discard the ciphertext pairs where ΔC_{03}^{i} and ΔC_{06}^{i} are not zero and the ciphertext pairs considered in Step (3), (4) and (5). Do the following for the remaining ciphertext pairs:

(a) Similarly to Step (5)(a), filter out the ciphertext pairs where ΔI_{07}^{11}, ΔI_{01}^{11} and ΔI_{05}^{11} are not included in $P(X^1)$, $P(X^2)$ and $P(X^3)$, respectively.

(b) Guess an 8 bit round keys RK_{100}^{11}. ("Guess 4" in Figure 2).

(c) Similarly to Step (5)(c), check that ΔI_{00}^{11} is included in $P(X^4)$. Output a guessed key which has the maximal counter as a right RK_{100}^{11}.

(7) From $\mathscr{A}$, filter out the ciphertext pairs where ΔC_{03}^{i} is not zero and the ciphertext pairs considered in Step (3), (4), (5) and (6). Do the following for the remaining ciphertext pairs:

(a) Similarly to Step (6)(a), filter out the ciphertext pairs where ΔI_{07}^{11}, ΔI_{01}^{11}, ΔI_{05}^{11} and ΔI_{00}^{11} are not included in $P(X^1)$, $P(X^2)$, $P(X^3)$ and $P(X^4)$, respectively.

(b) Guess 16 bit round keys (RK_{006}^{11}, RK_{106}^{11}). ("Guess 5" in Figure 2).

(c) Similarly to Step (6)(c), check that ΔI_{06}^{11} is included in $P(X^5)$. Output a guessed key which has the maximal counter as right RK_{006}^{11} and RK_{106}^{11}.

(8) From $\mathscr{A}$, filter out the ciphertext pairs considered in Step (3), (4), (5), (6) and (7). Do the following for the remaining ciphertext pairs:

(a) Similarly to Step (7)(a), filter out the ciphertext pairs where ΔI_{07}^{11}, ΔI_{01}^{11}, ΔI_{05}^{11}, ΔI_{00}^{11} and ΔI_{06}^{11} are not included in $P(X^1)$, $P(X^2)$, $P(X^3)$, $P(X^4)$ and $P(X^5)$, respectively.

(b) Guess 16 bit round keys (RK_{003}^{11}, RK_{103}^{11}). ("Guess 6" in Figure 2).

(c) Similarly to Step (7)(c), check that ΔI_{03}^{11} is included in $P(X^6)$. Output a guessed key which has the maximal counter as right RK_{003}^{11} and RK_{103}^{11}.

(9) Get the the corresponding plaintext pairs to all ciphertext pairs considered in Step (3)(b), (4)(c), (5)(c), (6)(c), (7)(c) and (8)(c), respectively. With them, do the following:

(a) Guess an 8 bit RK_{007}^{1} ("Guess 7" in Figure 2).
(b) Partially encrypt the plaintext pairs in Step (9) with the guessed round key to obtain the output difference of the 8th S-box in round 1, that is, $(P^{-1}(I^2))_{07}$.
(c) If the computed difference is included in $\{0\mathtt{x}01, 0\mathtt{x}02, 0\mathtt{x}04, 0\mathtt{x}08, 0\mathtt{x}09, 0\mathtt{x}80, 0\mathtt{x}84\}$, add the counter, corresponding to the guessed key, to one.
(d) Output a guessed key which has the maximal counter as a right RK_{007}^{1}.

(10) With an 80 bit suggested round key, compute a secret key by operating the key schedule of PP-1_64. Output the computed secret key as a right secret key of PP-1_64.

In our attack on PP-1/64_64, we construct $2^{37.29}$ structures which are composed of 256 plaintexts. Thus, the data complexity of our attack is about $2^{45.29}(\approx 2^{37.29} \cdot 2^8)$ chosen plaintexts. We store all ciphertext pairs passing Step (2) and the corresponding plaintext pairs in a table. The probability that a ciphertext pair passes Step (2) is 2^{-16}. Thus, $2^{36.29}(\approx 2^{52.29} \cdot 2^{-16})$ ciphertext pairs pass this step. Hence, the memory complexity of this attack is about $2^{41.29}(\approx 2^{52.29} \cdot 2^{-16} \cdot 4 \cdot 8)$ memory bytes.

The computational complexity of our attack is dominated by Step (1). The computational complexity of Step (1) is about $2^{45.29}(\approx 2^{37.29} \cdot 2^8)$ encryptions. The probability that a wrong ciphertext pair passes Step (3) is 2^{-40}. Since the expected number of the remaining wrong ciphertext pairs is $2^{-3.71}(\approx 2^{36.29} \cdot 2^{-40})$, we expect that only right ciphertext pairs are survived. From (4), we can check easily that all ciphertext pairs where the corresponding ΔO^{10}'s are included in $P(X^1)$ pass Step (3). Thus, the expected number of right ciphertext pairs is $4(\approx 2^{37.29} \cdot 2^{-35.29})$ from Table 3. The computational complexity of Step (3)(b) is about $2^{3.54}(\approx 2^8 \cdot 4 \cdot 1/11 \cdot 1/8)$ encryptions. Similarly to Step (3)(b), the computational complexities of other steps are also small. Hence, the computational complexity of our attack on PP-1/64_64 is about $2^{45.29}$ encryptions.

In the case of the attack on PP-1/64_128, the data and memory complexities are the same as them of the attack on PP-1/64_64. However, the computational complexity of this attack is dominated by Step (1) and (10), since we should do an exhaustive search for the remaining 48 bit key information. The computational complexity of Step (1) is about $2^{45.29}(\approx 2^{37.29} \cdot 2^8)$ encryptions. In Step (10), the probability that a wrong key passes this step is 2^{-64}. Thus, it is sufficient to use just one plaintext/ciphertext pair. The computational complexity of Step (10) is about 2^{48}. Hence, the computational complexity of PP-1/64_128 is about $2^{48.21}(\approx 2^{45.29} + 2^{48})$ encryptions.

5. Truncated Differential Analysis on Full-Round PP-1/128

Our attacks on full-round PP-1/128_128 and full-round PP-1/128_256 use 20-round differentials which are constructed by using the method introduced in Section 3. In detail, in order to construct them, we consider twenty five 19-round differentials $[P((15,\alpha)) \rightarrow (15,\beta)]_{19}$ $(\alpha,\beta \in \{0\mathtt{x}01, 0\mathtt{x}08, 0\mathtt{x}09, 0\mathtt{x}10, 0\mathtt{x}80\})$. Then we extend these 19-round differentials to total $1275(= 5 \cdot 255)$ 20-round differentials $[P((15,\alpha)) \rightarrow P((15, y_j^i))]_{20}$ $(y_j^i \in Y^i)$. Here, Y^i's are defined as follows $(i = 1, \ldots, 7)$:

$$
\begin{aligned}
Y^0 &= \{00000000_2\},\\
Y^1 &= \{0000?00?_2 \mid ? \in \{0,1\}\} - Y^0,\\
Y^2 &= \{0?00?00?_2 \mid ? \in \{0,1\}\} - \bigcup_{i=0}^{1} Y^i,\\
Y^3 &= \{??00?00?_2 \mid ? \in \{0,1\}\} - \bigcup_{i=0}^{2} Y^i,\\
Y^4 &= \{??00??0?_2 \mid ? \in \{0,1\}\} - \bigcup_{i=0}^{3} Y^i,\\
Y^5 &= \{??00????_2 \mid ? \in \{0,1\}\} - \bigcup_{i=0}^{4} Y^i,\\
Y^6 &= \{???0????_2 \mid ? \in \{0,1\}\} - \bigcup_{i=0}^{5} Y^i,\\
Y^7 &= \{????????_2 \mid ? \in \{0,1\}\} - \bigcup_{i=0}^{6} Y^i.
\end{aligned}
\tag{6}
$$

In the similar manner to the previous section, we choose a structure S_i which consist of 256 plaintext; that is, $S_i = \{(i \parallel j) \mid j = 0, 1, 2, \ldots, 255\}$, where i is a 120 bit fixed value. Then, as shown in Table 4, we can calculate the expected number of right plaintext pairs where ΔO^{21} is included in $P(15, (Y^i))$ for each structure.

5.1. Truncated Differential Analysis on PP-1/128. Our attack on full-round PP-1/128_128 is similar to that on full-round PP-1/128_256. Thus, we mainly present the attack procedure on PP-1/128_128. Since it is similar to the attack procedure on full-round PP-1/64_64, we briefly introduce it. The attack procedure on full-round PP-1/128_128 is as follows (see Figure 3).

(1) Select $2^{95.45}$ structures which are composed of 256 plaintexts and get the corresponding ciphertexts.

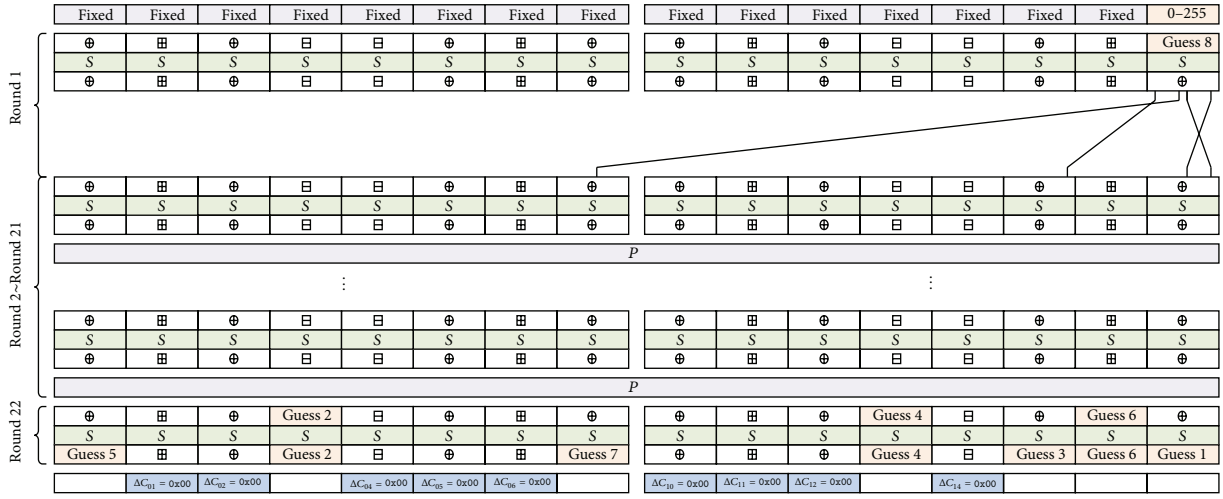


Figure 3: The attack procedure on full-round PP-1/128_128.

Table 4: The expected number of right pairs for PP-1/128.

Set of output differences of round 21	The expected number of right pairs
$P((15, Y^1))$	$2^{-93.33}$
$P((15, Y^2))$	$2^{-93.45}$
$P((15, Y^3))$	$2^{-92.64}$
$P((15, Y^4))$	$2^{-92.01}$
$P((15, Y^5))$	$2^{-96.86}$
$P((15, Y^6))$	$2^{-89.78}$
$P((15, Y^7))$	$2^{-88.92}$

From these ciphertexts, compute $2^{110.45}$ ciphertext pairs (C^i, C^{i*}).

(2) Check that $\Delta C^i_{01} = \Delta C^i_{02} = \Delta C^i_{04} = \Delta C^i_{05} = \Delta C^i_{06} = \Delta C^i_{10} = \Delta C^i_{11} = \Delta C^i_{12} = \Delta C^i_{14} = 0$ for each ciphertext pair. We keep all ciphertext pairs passing Step (2) and the corresponding plaintext pairs in a table and call a set containing them $\mathscr{A}$.

(3) From $\mathscr{A}$, filter out the ciphertext pairs where ΔC^i_{00}, ΔC^i_{03}, ΔC^i_{07}, ΔC^i_{13}, ΔC^i_{15}, and ΔC^i_{16} are not zero. Do the following for the remaining ciphertext pairs:

(a) Guess an 8 bit round key RK^{22}_{117} ("Guess 1" in Figure 3).

(b) Check that ΔI^{22}_{17} is included in $P(Y^1)$. Output a guessed key which has the maximal counter as right RK^{22}_{117}.

(4) Similarly to Step (3), determine sequentially the following right round keys.

(a) $(\mathrm{RK}^{22}_{003}, \mathrm{RK}^{22}_{103})$ ("Guess 2"),

(b) RK^{22}_{115} ("Guess 3").

(c) $(\mathrm{RK}^{22}_{013}, \mathrm{RK}^{22}_{115})$ ("Guess 4").

(d) RK^{22}_{101} ("Guess 5").

(e) $(\mathrm{RK}^{22}_{016}, \mathrm{RK}^{22}_{116})$ ("Guess 6").

(f) RK^{22}_{117} ("Guess 7").

(5) Get the the corresponding plaintext pairs to all ciphertext pairs considered in Steps (3) and (4). With them, do the following:

(a) Guess an 8 bit RK^1_{017} ("Guess 8").

(b) Partially encrypt the plaintext pairs in Step (5) with the guessed round key to obtain the output difference of the 8th S-box in second NF-function of round 1, that is, $(P^{-1}(I^2))_{17}$.

(c) If the computed difference is included in {0x01, 0x08, 0x09, 0x10, 0x80}, add the counter, corresponding to the guessed key, to one.

(d) Output a guessed key which has the maximal counter as a right RK^1_{017}.

(6) With an 88 bit suggested round key, do an exhaustive search for the remaining 40 bit key information by using one trial encryption. During this procedure, if a 128 bit secret key satisfies one known plaintext/ciphertext pair, output this 128 bit secret key as a right 128 bit secret key of full-round PP-1/128_128.

In this attack, we construct $2^{95.45}$ structures. Thus, the data complexity of our attack on full-round PP-1/128_128 is about $2^{103.45}(\approx 2^{95.45} \cdot 2^8)$ chosen plaintexts. In Step (2), since the probability that a ciphertext pair passes Step (2) is 2^{-72}, we store $2^{38.45}(\approx 2^{110.45} \cdot 2^{-72})$ ciphertext pairs pass this step and the corresponding plaintext pairs in a table. Thus, the memory complexity of this attack is about $2^{44.45}(\approx 2^{38.45} \cdot 4 \cdot 16)$ memory bytes. The computational complexity of this attack is dominated by Step (1), that is, about $2^{103.45}(\approx 2^{95.45} \cdot 2^8)$ encryptions.

In the case of the attack on full-round PP-1/128_256, the data and memory complexities are the same as them of the

attack on full-round PP-1/128_128. However, the computational complexity of this attack is dominated by Step (6), since we should do an exhaustive search for the remaining 168 bit key information. In Step (6), the probability that a wrong key that passes this step is 2^{-128}. Thus, this step needs two plaintext/ciphertext pairs. The computational complexity of Step (6) is about $2^{168}(\approx 2^{168} + 2^{168} \cdot 2^{-128})$ encryptions. Hence, the computational complexity of our attack on full-round PP-1/128_256 is about 2^{168} encryptions.

6. Truncated Differential Analysis on Full-Round PP-1/192 and PP-1/256

This section introduces our attack results on full-round PP-1/192 and full-round PP-1/256. Overall, the attack procedures on them are similar to the attack procedures on PP-1/64.

Our attacks on full-round PP-1/192 uses $2^{149.85}$ structures $S_i = \{(i \parallel j) \mid j = 0, 1, 2, \ldots, 255\}$, where i is a 184 bit fixed value. First, we construct thirty six 29-round differentials $[P((23, \alpha)) \rightarrow (23, \beta)]_{29}$ $(\alpha, \beta \in \{\texttt{0x01}, \texttt{0x02}, \texttt{0x08}, \texttt{0x09}, \texttt{0x40}, \texttt{0x80}\})$. Then we extend these 29-round differentials to total $1530(= 6 \cdot 255)$ 30-round truncated differentials $[P((23, \alpha)) \rightarrow P((23, y_j^i))]_{30}$ $(y_j^i \in Y^i)$. Note that Y^i's used in this attack are the same as them in the attack on full-round PP-1/128. Table 5 presents the expected number of right plaintext pairs where ΔO^{31} is included in $P(23, (Y^i))$ in each structure. The complexities of our attacks are as follows.

PP-1/192_192(PP-1/192_384)

(a) The data complexity: about $2^{157.85}$ chosen plaintexts.

(b) The memory complexity: about $2^{35.44}$ memory bytes.

(c) The computational complexity: about $2^{157.85}(2^{296})$ encryptions.

In the case of PP-1/256, we consider $2^{202.84}$ structure $S_i = \{(i \parallel j) \mid j = 0, 1, 2, \ldots, 255\}$, where i is a 248 bit fixed value. And we construct 204041-round differentials $[P((31, \alpha)) \rightarrow P((31, y_j^i))]_{41}$ $(\alpha \in \{\texttt{0x01}, \texttt{0x02}, \texttt{0x04}, \texttt{0x08}, \texttt{0x09}, \texttt{0x10}, \texttt{0x40}, \texttt{0x80}\})$ $(y_j^i \in Y^i)$. Note that Y^i's used in this attack are also the same as them in the attack on full-round PP-1/128. Table 6 presents the expected number of right plaintext pairs where ΔO^{42} is included in $P(31, (Y^i))$ in each structure. The complexities of our attacks are as follows.

PP-1/256_256(PP-1/256_512)

(a) The data complexity: about $2^{210.84}$ chosen plaintexts.

(b) The memory complexity: about $2^{24.84}$ memory bytes.

(c) The computational complexity: about $2^{210.84}(2^{432})$ encryptions.

Table 5: The expected number of right pairs for PP-1/192.

Set of output differences of round 31	The expected number of right pairs
$P((23, Y^1))$	$2^{-147.85}$
$P((23, Y^2))$	$2^{-147.01}$
$P((23, Y^3))$	$2^{-145.93}$
$P((23, Y^4))$	$2^{-145.48}$
$P((23, Y^5))$	$2^{-144.38}$
$P((23, Y^6))$	$2^{-143.26}$
$P((23, Y^7))$	$2^{-142.38}$

Table 6: The expected number of right pairs for PP-1/256.

Set of output differences of round 42	The expected number of right pairs
$P((31, Y^1))$	$2^{-200.73}$
$P((31, Y^2))$	$2^{-200.84}$
$P((31, Y^3))$	$2^{-199.89}$
$P((31, Y^4))$	$2^{-199.23}$
$P((31, Y^5))$	$2^{-198.10}$
$P((31, Y^6))$	$2^{-197.09}$
$P((31, Y^7))$	$2^{-196.18}$

7. Conclusion

In this paper, we have presented the first known cryptanalytic results of four concrete versions of a scalable block cipher PP-1, full-round PP-1/64, full-round PP-1/128, full-round PP-1/192, and full-round PP-1/256, by using truncated differential cryptanalysis. As summarized in Table 1, our attacks on these algorithms require computational complexities smaller than the exhaustive search. These results indicate that PP-1 is vulnerable to truncated differential cryptanalysis and that it is insecure.

Acknowledgments

This research was supported by the Ministry of Science, ICT and Future Planning (MSIP), Korea, under the Convergence Information Technology Research Center (C-ITRC) support Program (NIPA-2013-H0301-13-3007) supervised by the National IT Industry Promotion Agency (NIPA).

References

[1] J. Chen, B. Mariam, and M. Matsumoto, "A single mobile target tracking in voronoi-based clustered wireless sensor network," *Journal of Information Processing Systems*, vol. 1, pp. 17–28, 2011.

[2] C. Huang, R. H. Cheng, S. R. Chen, and C. Li, "Enhancing network availability by tolerance control in multi-sink wireless sensor network," *Journal of Convergence*, vol. 1, no. 1, pp. 15–22, 2010.

[3] P. Sarkar and A. Saha, "Security enhanced communication in wireless sensor networks using reed-muller codes and partially balanced incomplete block designs," vol. 2, pp. 23–30.

[4] D. Kumar, T. Aseri, and R. Patel, "Multi-hop communication routing (MCR) protocol for heterogeneous wireless sensor networks," *International Journal of Information Technology, Communications and Convergence*, vol. 1, pp. 130–145, 2010.

[5] H. Lim, K. Jang, and B. Kim, "A study on design and implementation of the ubiquitous computing environment-based dynamic smart on/off-line learner tracking system," *Journal of Information Processing Systems*, vol. 6, no. 4, pp. 609–620, 2010.

[6] B. Xie, A. Kumar, D. Zhao, R. Reddy, and B. He, "On secure communication in integrated heterogeneous wireless networks," *International Journal of Information Technology, Communications and Convergence*, vol. 1, no. 1, pp. 4–23, 2010.

[7] D. Hong, J. Sung, S. Hong et al., "HIGHT: a new block cipher suitable for low-resource device," *Cryptographic Hardware and Embedded Systems*, vol. 4249, pp. 46–59, 2006.

[8] T. Shirai, K. Shibutani, T. Akishita, S. Moriai, and T. Iwata, "The 128-bit blockcipher CLEFIA (extended abstract)," *Fast Software Encryption*, vol. 4593, pp. 181–195, 2007.

[9] C. de Cannière, O. Dunkelman, and M. Knežević, "KATAN and KTANTAN—a family of small and efficient hardware-oriented block ciphers," *Cryptographic Hardware and Embedded Systems*, vol. 5747, pp. 272–288, 2009.

[10] K. Bucholc, K. Chmiel, A. Grocholewska-Czuryło, E. Idzikowska, I. Janicka-Lipska, and J. Stokłosa, "Scalable PP-1 block cipher," *International Journal of Applied Mathematics and Computer Science*, vol. 20, no. 2, pp. 401–411, 2010.