Wireless Communications and Mobile Computing

Advanced Wireless Communications and Mobile Computing Technologies for the Internet of Things

Lead Guest Editor: Haiyu Huang
Guest Editors: Kejie Lu, Giovanni Pau, Yong Ren, and Pai-Yen Chen

Advanced Wireless Communications and Mobile Computing Technologies for the Internet of Things

Wireless Communications and Mobile Computing

Advanced Wireless Communications and Mobile Computing Technologies for the Internet of Things

Lead Guest Editor: Haiyu Huang
Guest Editors: Kejie Lu, Giovanni Pau, Yong Ren, and Pai-Yen Chen

Editorial Board

Contents

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 9693514, 2 pages
https://doi.org/10.1155/2018/9693514

Editorial

Advanced Wireless Communications and Mobile Computing Technologies for the Internet of Things

Haiyu Huang[1,2] **Kejie Lu**[3] **Giovanni Pau**[4] **Yong Ren**[5] **and Pai-Yen Chen**[6]

[1]*Department of Electrical and Computer Engineering, University of Texas, Austin, TX, USA*
[2]*Maxim Integrated Inc., San Jose, CA, USA*
[3]*Department of Electrical and Computer Engineering, University of Puerto Rico, Mayagüez, PR, USA*
[4]*Faculty of Engineering and Architecture, Kore University of Enna, Enna, Italy*
[5]*Department of Electronic Engineering, Tsinghua University, Beijing, China*
[6]*Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI, USA*

Correspondence should be addressed to Haiyu Huang; harryhuang@utexas.edu

Received 18 March 2018; Accepted 21 March 2018; Published 24 April 2018

Since its origination from P&G and MIT Auto-ID Center in 1999, the term "Internet of Things" (IoT) has been elevated from a specific application concept, based on RFID, to a vastly prominent phrase that represents the general direction to the future of many important aspects of human life. From the application perspective, all the major industries and sectors are now experiencing a certain level of paradigm shift thanks to IoT. For instance, the healthcare system is moving from hospital-centered care to distributed omnipresent care; the transportation is gradually replacing human involvement with driverless technologies and vehicle-to-vehicle (V2V) communications; in manufactory, Industry 4.0, which is empowered by IoT, is taking the leading role in changing the face of factories everywhere in the world. From technology perspective, trending research/development fields in electrical and computer engineering, such as Device-to-Device Network, Wireless Sensor Networks, 5G, LoRaWAN, Bluetooth LE, MIMO, deep learning, and distributed low power computing, have all been directly or indirectly contributing to the evolution of IoT.

Due to the timeliness and importance of the topics, we received a large number of submissions. In the review process, each paper was reviewed by multiple experts in relevant fields. After a rigorous two-round review process, we decided to accept 10 excellent articles addressing cutting-edge wireless communications and mobile computing technologies and applications around the latest trend of IoT. Since the topics of articles cover broad technology scopes, we will introduce them below according to the themes of IoT applications.

The first theme is device-to-device (D2D) communication. In this special issue, we have two papers investigating different D2D aspects. Specifically, in "Clustering Optimization for Out-of-Band D2D Communications," the authors focused on clustering optimization algorithms for Out-of-Band D2D communication. The results showed that well-known clustering algorithms can be employed to determine the cluster head which provides a near-optimal solution for throughput in channels between the cluster head and its members.

In "Performance Analysis of Three-Dimensional Clustered Device-to-Device Networks for Internet of Things," H. Jung and I.-H. Lee aimed at modeling and analyzing clustered D2D networks in three-dimensional space for scenarios where devices are stacked vertically and dispersed in the horizontal plane.

The second theme is low power/short distance wireless communications for IoT. In this issue, we accepted four papers that are related to low power and/or short distance wireless communication technologies. In "Maximum Power Plus RSSI Based Routing Protocol for Bluetooth Low Energy Ad Hoc Networks," the authors proposed an energy-conserving multihop routing protocol for Bluetooth Low Energy.

Next, in "A High Throughput Anticollision Protocol to Decrease the Energy Consumption in a Passive RFID System," the aim is an anticollision protocol that can solve tag collision problem in UHF RFID.

The paper "An Adaptive Scheduler for Real-Time Operating Systems to Extend WSN Nodes Lifetime" introduced an adaptive scheduler for RTOS used in WSN sensor node to reduce energy consumption and thus increase the lifetime of battery powered WSN sensor node.

In the paper "A Dual Key-Based Activation Scheme for Secure LoRaWAN," the authors focused on the security issue of LoRaWAN, which is one of the most promising Low Power Wide Area Network technologies for IoT, and they proposed a dual key-based activation scheme for LoRaWAN to improve its security.

The third theme in our special issue is MIMO and beamforming for 5G and IoT. There are three relevant papers in this category. The paper "5G MIMO Conformal Microstrip Antenna Design" demonstrated eight-element microstrip MIMO conformal antennas at 35 GHz well suited for the 5G MIMO communication.

In "Pipeline Implementation of Polyphase PSO for Adaptive Beamforming Algorithm," the authors investigated a low hardware cost realization of a partial Particle Swarm Optimization algorithm for an adaptive beamformer.

The paper "Optimized Power Allocation and Relay Location Selection in Cooperative Relay Networks" focused on power allocation optimization in cooperation communication used in MIMO.

Finally, yet importantly, the last theme is cyberphysical systems for IoT. Here we have one paper "SVM-Based Dynamic Reconfiguration CPS for Manufacturing System in Industry 4.0," in which a cyberphysical system framework was designed using learning algorithm SVM to support Industry 4.0.

Haiyu Huang
Kejie Lu
Giovanni Pau
Yong Ren
Pai-Yen Chen

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 4185650, 10 pages
https://doi.org/10.1155/2018/4185650

Research Article

An Adaptive Scheduler for Real-Time Operating Systems to Extend WSN Nodes Lifetime

Roberto Rodriguez-Zurrunero, Ramiro Utrilla, Elena Romero, and Alvaro Araujo

B105 Electronic Systems Lab, ETSI Telecomunicación, Universidad Politécnica de Madrid, Avenida Complutense 30, 28040 Madrid, Spain

Correspondence should be addressed to Roberto Rodriguez-Zurrunero; r.rodriguezz@b105.upm.es

Received 25 July 2017; Revised 29 December 2017; Accepted 9 January 2018; Published 6 February 2018

Academic Editor: Giovanni Pau

Wireless Sensor Networks (WSNs) are a growing research area as a large of number portable devices are being developed. This fact makes operating systems (OS) useful to homogenize the development of these devices, to reduce design times, and to provide tools for developing complex applications. This work presents an operating system scheduler for resource-constraint wireless devices, which adapts the tasks scheduling in changing environments. The proposed adaptive scheduler allows dynamically delaying the execution of low priority tasks while maintaining real-time capabilities on high priority ones. Therefore, the scheduler is useful in nodes with rechargeable batteries, as it reduces its energy consumption when battery level is low, by delaying the least critical tasks. The adaptive scheduler has been implemented and tested in real nodes, and the results show that the nodes lifetime could be increased up to 70% in some scenarios at the expense of increasing latency of low priority tasks.

1. Introduction

An operating system (OS) is a software layer that provides hardware abstraction and allows the developer to manage hardware resources. An OS also provides the developer standard mechanisms and services to ease and unify application development.

Therefore, the main advantages of using an OS are the software portability over heterogeneous hardware platforms and the ability to build application level developments regardless of the hardware used. OSes also provide other features such as multithreading capabilities or memory management. Wireless Sensor Networks (WSNs) are one of the most OSes demanding fields as these networks are composed by heterogeneous nodes where efficient hardware management is a main issue.

On the other hand, using an OS usually implies an overload in memory, CPU cycles, or energy consumption. This overload could be a critical issue in autonomous resource-constraint systems such as nodes present in WSNs. For this reason, OSes for WSNs must fulfil some specific requirements and features:

(i) *energy efficiency*, so the battery of autonomous wireless sensor nodes could last long periods;

(ii) *memory management tools,* in order to develop dynamic applications that use memory efficiently;

(iii) *real-time capabilities*, as most applications require bounded processing latencies of sensor data;

(iv) *wireless protocol stack,* which allows reliable and efficient communications in the nodes, while consuming low resources;

(v) *adaptability to the environment*, as WSN applications are heterogeneous and they usually operate in dynamic environments.

The scheduler is considered the core of an OS as it manages the tasks execution and could provide real-time management capabilities to the developer. Optimizing the scheduler is mandatory in OSes for WSNs in order to provide real-time multithread capabilities while using the lowest resources possible.

Our work proposes a scheduler that changes the task scheduling depending on environment conditions, which are

treated as inputs. Energy efficiency could be improved with this algorithm as the scheduler adapts dynamically to the environment when it changes. In this work, we use the device battery level and the tasks priorities as environment inputs in order to reduce energy consumption when the battery is running low, while maintaining minimum latencies for high priority tasks.

This paper is organized as follows. Section 2 presents the related works in the WSN OS field. In Section 3, the architecture of the scheduler is described. Section 4 shows the algorithm used for making scheduling decisions, while Section 5 describes the implementation of the algorithm in real nodes and the test scenario used. In Section 6, results are presented and discussed. Finally, the paper is finished in Section 7 with the work conclusion.

2. Related Work

Operating systems for WSNs are a highly studied area in the last decade since the first networks were deployed. Many of them have been developed during the last years, with TinyOS [1] and Contiki OS [2] being the most extended ones in WSN applications.

These OSes fit the requirements of WSN OSes, as they have a very small memory footprint while providing development abstraction. They also provide full network stack, simple memory management, and some multithreading capabilities. These OSes can run well in resource-constraint low-power microcontrollers, such as the Texas Instruments MSP430 used in TelosB, running at 8 MHz with 10 KB RAM.

However, new microcontrollers, such as low-power ARM Cortex-M ones, have increased available resources while maintaining very low energy consumption, reaching up to 120 MHz clock speed and 320 KB RAM. Therefore, these new devices allow the usage of more advanced OSes that employ fully preemptive threads and other features such as mutexes, semaphores, timers, or queues. Real-time operating systems (RTOS), such as open sourced FreeRTOS, may be used for WSN on these new microcontrollers. A priority scheduler or a round-robin scheduler may be used to implement a real-time OS. Several studies have been conducted to compare performance of both and to decide the best situation for using each method [3]. A round-robin scheduler shares the executing time with all active tasks, while a priority scheduler executes first higher priority tasks, reducing their latencies. Mixed strategies could be used as FreeRTOS does, where round-robin schedule is applied for same-priority tasks. However, for these OSes the main drawback is RAM usage, so several memory optimization techniques are presented by authors [4] to reduce it. Both thread optimization and memory allocation techniques could be useful for multithread RTOS.

Recent studies demonstrate that these real-time operating systems are being used in WSN monitoring systems [5], showing that a sensor network could be implemented even with real-time constraint over a wireless channel.

Other open issues regarding OSes are also being studied in the last years such as their steep learning curve and their power management features. RIOT OS [6] was developed in order to reduce the learning curve when programming IoT applications. This OS also provides real-time and built-in energy capabilities and energy-efficiency features. However, it uses a priority scheduler that does not share executing time between same-priority tasks and does not adapt their properties dynamically. On the other hand, improving power management of a multitasking WSN is also proposed by Brandolese et al. [7]. This management infrastructure and optimization model improves energy saving exploiting hibernation modes dynamically without memory retention. However, this method could cause losing real-time capabilities, as sensing tasks are grouped to improve energy efficiency, and they are not processed till a later time. Finally, CerberOS [8] presents a method to facilitate third party application design, by providing resource-secure capabilities in the nodes, allowing sharing them for different applications.

The distributed OSes for WSN field are also targeted by several research works in order to provide better management features and a highly transparent interface for the network developer [9, 10]. Load balancing of the nodes in distributed architectures has also been studied by Zoican et al. [11]. This work proposes a method for centralized task migration resulting in a final load near the average over all nodes of the network. Therefore, node cooperation and context aware methods will be critical issues for future WSN OSes developments.

Finally, other works target dynamic reconfiguration and operation of OSes. Lorien OS [12] was proposed as a fully component based operating system allowing efficient dynamic modules loading. On the other hand, an OS reconfiguration mechanism is proposed by Gasmi et al., [13] in order to provide efficient middleware that solves decision-making problems during reconfiguration stage. Besides, improving TinyOS tasks throughput while reducing energy consumption is achieved using a dynamic priority scheduler [14]. In this scheduler, the energy consumption is 1.14 times lower than the original TinyOS.

All these works show the interest in dynamic reconfiguration of OSes for WSNs. However, there are still open issues getting an adaptive scheduler that modifies its properties dynamically on changing environments. The main target of our work consists in improving WSN nodes lifetime in dynamic battery-operated environments by adapting dynamically a round-robin scheduler.

3. Scheduler Architecture

In this section, we explain an architecture for an adaptive scheduler that could modify its behaviour in changing environments. The global architecture is shown in Figure 1. The main idea consists in a module which accepts some environment inputs and makes scheduling decisions to modify scheduler properties. Some of the scheduler properties that could be dynamically modified are the duty cycle, the tasks priorities, and the scheduler system timer (Systick) period.

(i) The duty cycle is the portion the node CPU is running with respect to total time; the rest of the time the node is in low-power mode, also called sleep mode.

(ii) The tasks priorities allow managing tasks execution and real-time capabilities, as higher priority tasks are

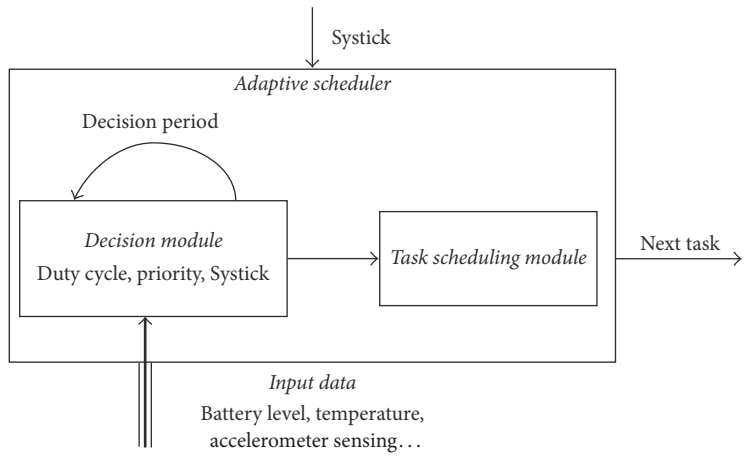


Figure 1: Adaptive scheduler architecture.

usually executed with lower latencies compared with low priority ones.

(iii) The Systick is the main timer of the scheduler, so the execution could be changed from one task to another in every system timer interrupt.

In order to make decisions, the scheduler architecture proposed uses a decision period. At the starting time of this period a decision is made in order to change scheduler properties. This decision period is a multiple of the Systick so we use only one timer for OS kernel management. The Systick period and the CPU clock frequency are constant in our scheduler, so the CPU does not change its executing speed dynamically.

This architecture is scalable to adapt dynamically any scheduler property although in this work we manage only the scheduler duty cycle. Our target consists in extending lifetime through scheduler duty cycling control while maintaining low latency for real-time tasks. Duty cycle represents the time when the microcontroller is active, so it is directly proportional with energy consumption.

In most schedulers, the active duty cycle is set by the tasks load, so all energy management is expected to be done by the programmer of each task. This way, a badly programmed task that never sleeps causes the CPU always to be active, so the duty cycle will be 100%. In the scheduler proposed, the node active time is fixed by the duty cycling decision module independently of the task load. A task which is in ready state could not be executed if the fixed active time has lapsed. This could cause increasing latencies for some tasks but save large amount of energy in some situations. To avoid this effect for real-time tasks the scheduler allows them to execute even if the active time has lapsed. Therefore, the adaptive scheduler saves energy by delaying low priority tasks. It should be noted that this is done without slowing them down as the CPU clock frequency does not change.

The flowchart of the adaptive scheduler process with duty cycling decision is shown in Figure 2. The schedule process is executed each Systick interrupt. First, it checks whether there is any active task or not. If not, it goes to sleep mode until the next Systick interrupt is triggered, so the process would start again.

On the other hand, if there are active tasks, it checks if the decision has lapsed. If it does, a decision must be taken in order to set a new duty cycle for the next decision period. This duty cycle sets the maximum available executing time for the tasks during this period. For example, if we set the Systick timer to 1 time unit and the decision period to 5 time units, we will have 2 time units as available active time if the decision-making process sets the duty cycle to 40%.

Whether or not a decision is made, the next step consists in checking if there is any available active time to execute tasks during this period. If not, the system checks if there is any task with highest priority, as we need them to be executed even if there is not available time for this period. If there are not highest priority tasks, the system goes to sleep mode until the current period finishes.

Finally, if there is any available time or there is any highest priority task, the next task to be executed will be scheduled in a priority-based round-robin way. The task will be executed during the Systick time, and when it lapses the process will start again.

In Figure 3 a time diagram example is presented comparing a priority round-robin scheduler with our adaptive scheduler. There are 3 tasks, with task 1 and task 2 being low priority tasks, having equal priority, and task 3 being the highest priority task. In this example, the decision period is 5 time units, while the decisions made set duty cycle to 40% for the three first ones and 20% for the last decisions. This way, the active time results in 2 and 1 time units, respectively. The Systick timer for both schedulers are set to 1 time unit.

The example shows the behaviour of our proposed scheduler. While round-robin scheduler executes all available tasks as soon as possible, the adaptive scheduler only executes the fixed duty cycle for each period. This causes low priority tasks to be delayed compared to round-robin scheduler. The result over large time scheduler operation will be a lower number of executions of these tasks which will lead to a large energy saving. On the other hand, high priority tasks, like task 3, execute the same way they do in a round-robin scheduler,

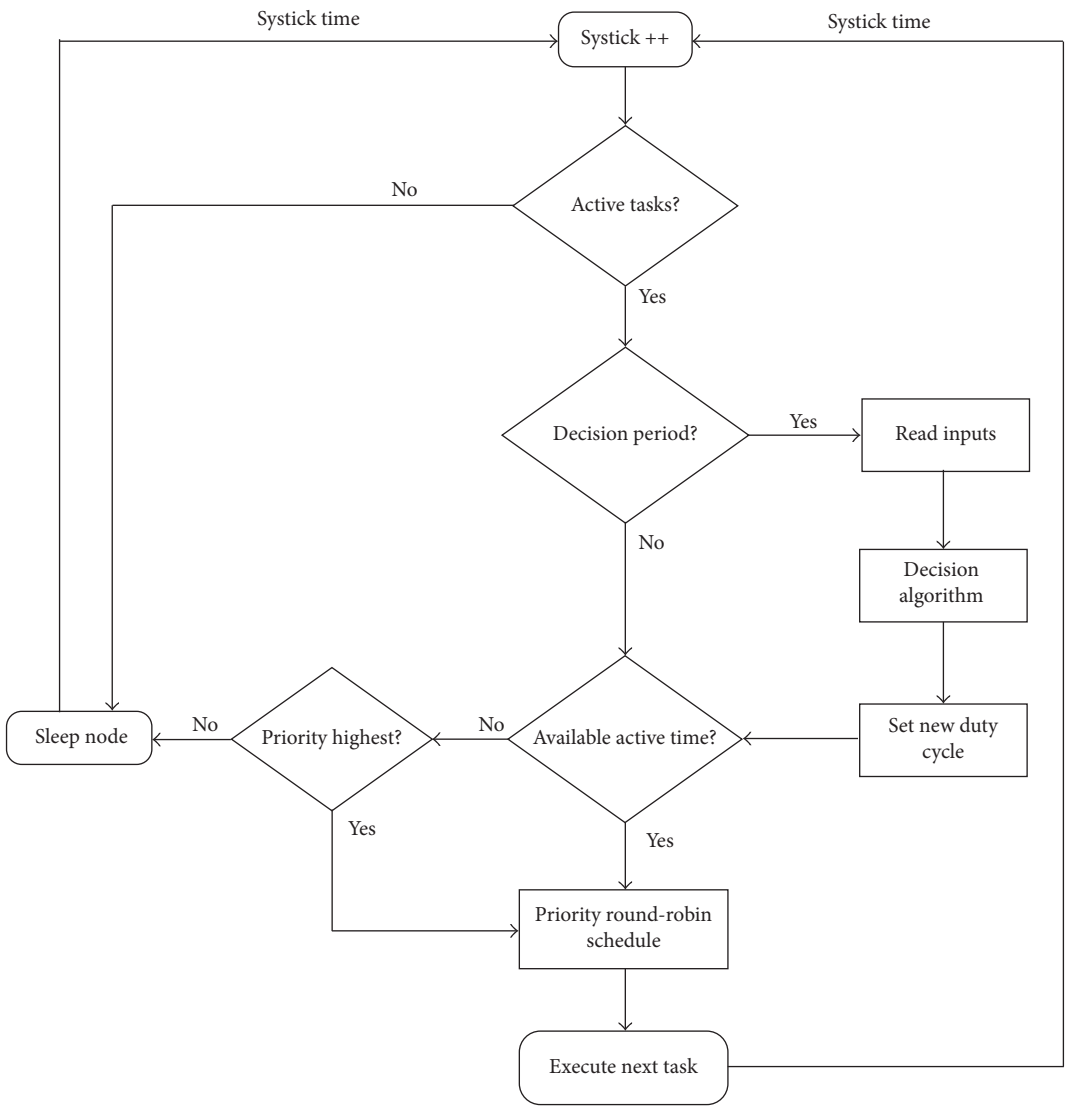
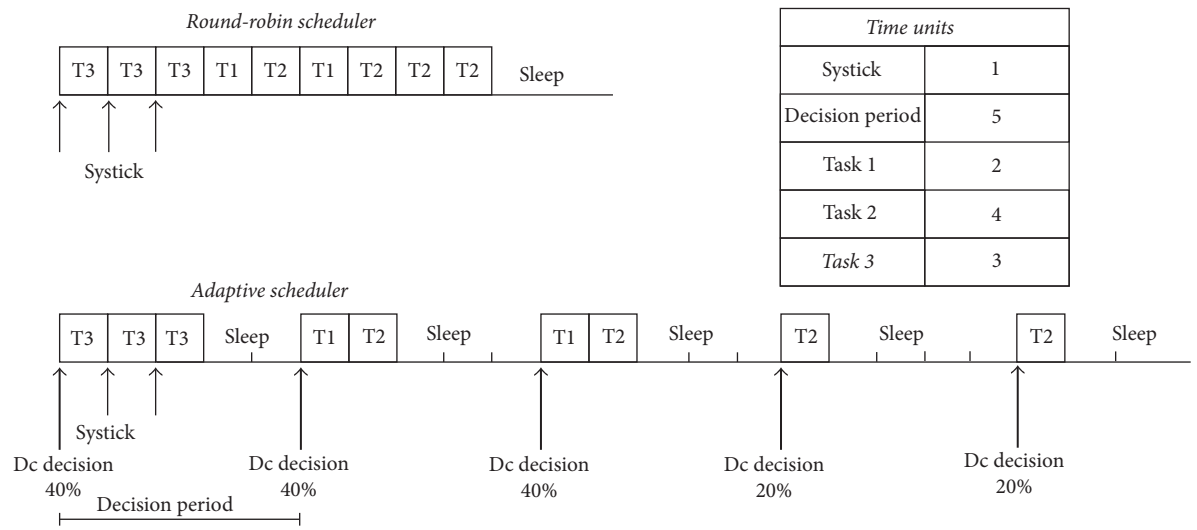


Figure 2: Adaptive scheduler flowchart.



Time units	
Systick	1
Decision period	5
Task 1	2
Task 2	4
Task 3	3

Figure 3: Time diagrams for round-robin and adaptive schedulers.

meaning no extra latency for them. In this example task 3 executes during 3 time units in both schedulers, so it is not delayed. On the other hand, task 1 lasts 6 time units in round-robin scheduler to complete execution, while it needs 11 time units in our scheduler. This delay in executing low priority tasks results in large energy saving as the system is in sleep state for a longer time.

In this work, we make duty cycle decisions, so node active time is changed dynamically. Input data used on this model could be either environmental parameters such as temperature, humidity, and RSSI or node parameters like battery level, energy consumption, tasks priorities, and execution state. This data could be collected each time a decision is made or could be stored in node memory and accessed by the decision module.

4. Duty Cycle Decision Algorithm

In this section, we present a duty cycle decision algorithm targeted at improving nodes lifetime. We use the approach proposed by Sirakoulis and Karafyllidis [15] which uses Public Goods Games (PGG) as a model to make decisions in power-aware embedded systems. This approach is based on Game Theory, which is a large field that studies mathematical models for making decisions in scenarios where rational players must use a shared resource. Players will take different decisions depending on the outcome of each one. On the PGG model, players compete for a shared resource and they cooperate optimizing their global outcome. This work [15] studies the effects of cooperation using a PGG applied to embedded systems on changing environments and presents a complete theoretical approach to these games. Besides, a global overview to Game Theory is also presented.

As described by authors of [15], the PGG is the most appropriate model for a scenario where there are power-aware jobs considered as players that should compete or cooperate for energy resources. Therefore, PGG provides a standardized formulation to solve the decisions proposed in our scheduler.

In our work, we propose a variation of the standard PGG problem in order to get a duty cycle value for each decision period. First, we define the global game parameters. The players of our game are each active task for the decision period and the shared resource is the execution time. The players could cooperate investing part of their available time in the decision cycle resulting in a global lower execution time for all tasks. This way, the lifetime could be extended when tasks decide to invest part of their time.

The investment done by each task in a decision cycle $I_i(t+1)$ is calculated in (1), where t and $t+1$ are indexes denoting the time step and i is the index denoting the current player (task). $I_i(t)$ is the investment done by a task in the previous decision cycle, adding a memory component to the algorithm, while F is a function of the reward $r_i(t)$ obtained each round for each task. The investment I_i is limited between 0 and 1 and represents the portion of time a task invests in order to save energy. The function F is then bounded in order to maintain the investment on its limits as shown in (2):

$$I_i(t+1) = I_i(t) + F\left(r_i(t)\right), \tag{1}$$

$$-I_i \leq F\left(r_i(t)\right) \leq 1 - I_i. \tag{2}$$

The parameter r_i represents the reward a task will obtain when investing part of its executing time. It is defined by the difference between the gain $g_i(t)$ obtained and the investment done in the previous decision period:

$$r_i(t) = \left(g_i(t) - I_i(t)\right). \tag{3}$$

The gain g_i depends on the state of the inputs defined for our scheduler decision module. Therefore, the gain changes in every cycle depending on the input values, so it defines the behaviour of the scheduler in changing environments. In our algorithm, we use the node battery level, the tasks priority, and a user-defined multiplication factor (MF) as input values.

In order to get the desired behaviour, the gain function is defined increasing with the multiplication factor and decreasing with task priority and battery level. This way, for high battery level or high task priority the gain value is low since the task is less likely to invest its executing time. We have defined the gain function in (4), where $M(t)$ is the multiplication factor, P_i is the normalized task priority, and $E(t)$ is the battery level. On the other hand, the parameters a, b, c, and k are fixed weights used to calibrate the behaviour of the scheduler:

$$g_i(t) = M(t)\left(\frac{a}{P_i} + \frac{b}{\left(E(t)+k\right)^2} + c\right). \tag{4}$$

It is important to note that the gain function defines the behaviour of the scheduler depending on the input data. This function could be modified in order to get a different behaviour or if we had other input data sources. Thereby both the function and its weights a, b, c, and k could be tuned depending on the desired behaviour. In our work, the gain function and its weights have been fixed to empirical values, which leads to a reasonable behaviour, in order to increase the gain when battery is low and the task priority is also low.

Finally, the reward function $F(r_i(t))$ is defined in (5). Its maximum and minimum values are F_{MAX} and $-F_{\mathrm{MAX}}$, respectively, that must meet the bounds set in (2). So, using the F_{MAX} value of (6) we can make sure the bounds are never exceeded. The reward function is linear between R_1 and R_2, which are fixed user-defined limits, and constant beyond these bounds. This function and the memory component of (1) make the scheduler response to changes slower, so the duty cycle will not change abruptly even if input values do:

$$F\left(r_i(t)\right) = \begin{cases} F_{\mathrm{MAX}} & r_i(t) \geq R_2 \\ \dfrac{2F_{\mathrm{MAX}}}{R_2 - R_1}\left(r_i(t) - R_2\right) + F_{\mathrm{MAX}} & R_1 \leq r_i(t) < R_2 \\ -F_{\mathrm{MAX}} & r_i(t) \leq R_1, \end{cases} \tag{5}$$

$$F_{\mathrm{MAX}} = I_i(t)\left(1 - I_i(t)\right). \tag{6}$$

Once the PGG has been formulated, the algorithm steps are presented in order to get the duty cycle of each period:

(1) Check the number of active tasks.

(2) For each active task,

(i) calculate reward value $r_i(t)$ from previous gain and investment values (3);

(ii) calculate reward function value $F(r_i(t))$ (5);

(iii) obtain the inversion of this cycle for this task $I_i(t+1)$ (1);

(iv) compute the gain value of this cycle $g_i(t)$ (4), which will be used in the next cycle.

(3) Calculate the arithmetic mean investment over all active tasks (7).

(4) Obtain the period duty cycle from the mean investment (8):

$$I_{\text{mean}}(t+1) = \frac{\sum_i^{N_{\text{tasks}}} I_i(t+1)}{N_{\text{tasks}}}, \tag{7}$$

$$D_T(t+1) = 1 - I_{\text{mean}}(t+1). \tag{8}$$

The duty cycle calculated has values between 0 and 1 as the investment has. The greater the investment made by all tasks is, the shorter the duty cycle of this period is. Therefore, by using this decision module in our adaptive scheduler architecture, the duty cycle is reduced when tasks decide to cooperate investing part of their time. This allows extending lifetime at the price of delaying low priority tasks that have decided to reduce their executing time.

5. Materials and Methods

The adaptive scheduler and the duty cycle decision algorithm proposed have been implemented in the YetiMote WSN node developed in the B105 Electronic Systems Lab which is shown in Figure 4. It is a custom-designed node composed by a high-performance low-power STM32L4 [16] microcontroller. The node runs up to 80 MHz with high memory capabilities (512 KB Flash, 128 KB RAM) and supports several low-power modes. In our test scenario, the microcontroller has 48 MHz system clock frequency. The node also has 2 accelerometers, a temperature sensor, an air quality sensor, a power management module, and 3 radio interfaces for 433 MHz, 868 MHz, and 2.45 GHz bands. A full version of FreeRTOS operating system is implemented on these nodes, which uses a priority-based round-robin scheduler. The tests performed are run on FreeRTOS scheduler in order to compare results with our adaptive scheduler.

The test scenario consists of 16 periodic tasks running a fixed time of 60 seconds for each test. The tasks periods are all different as well as their executing time in order to get the most realistic scenario possible when the tasks do not execute synchronously. In our tests, three scenarios have been defined depending on average task load. The task load is defined as the sum of the tasks active times divided by the total test time. Therefore, the tests have been performed using low task load (5%), medium task load (10%), and high task load (25%). Although 25% may not seem as a high executing load for most systems, in an energy constrained WSN scenario this task load is considered very high.

Figure 4: YetiMote WSN node used for testing the adaptive scheduler.

The input values used in our adaptive scheduler are the battery level $E(t)$, the normalized task priority P_i, and the user-defined multiplication factor $M(t)$. The battery level is limited to 0 when battery is discharged and 1 when it is fully charged. The scheduler implemented has 6 different task priorities, from 1 to 6, so the normalized task priority P_i is the quotient of the task priority and the total number of priorities. We have set the maximum task priority to 6 and the lowest task priority to 1. In all the scenarios 3 tasks are defined as high priority tasks, with priority levels 5 and 6, and the remaining 13 tasks have random priority values from 1 to 4. Finally, the multiplication factor allows the user to tune the scheduler behaviour dynamically, and it could have any positive value.

The adaptive scheduler parameters have been set to fixed values for all tests as well as the duty cycle decision algorithm parameters. The decision period value is 10 ms, while the Systick time value is 250 μs. On the other hand, gain function (4) parameters are $a = 0.5$, $b = 3.5$, $c = -2.5$, and $k = 0.7$, while reward function (5)bounds are $R_1 = -0.6$ and $R_2 = 1.2$. These values were empirically obtained after numerous tests in order to get a specific scheduler behaviour. Different values could be used to tune the scheduler if other behaviour is desired.

Each test measures the energy consumption and each one lasts 60 seconds. Therefore, different battery level or multiplication factor values are fixed for each test so we can evaluate energy saving on different input conditions. The energy consumption is obtained counting the time the microcontroller is in sleep mode during the test time. For that reason, we need to suppose 30 mW average power consumption when microcontroller is running and zero milliwatts when it is sleeping.

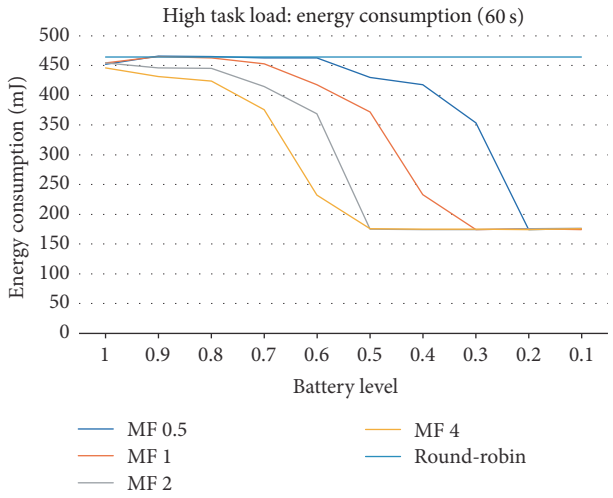


Figure 5: High task load test: energy consumption.

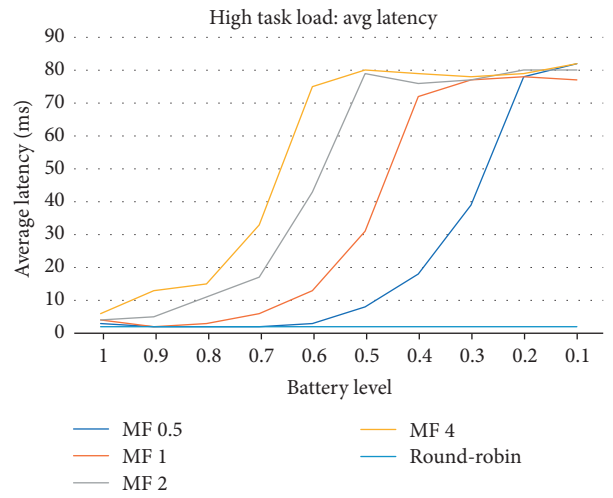


Figure 6: High task load test: average latency.

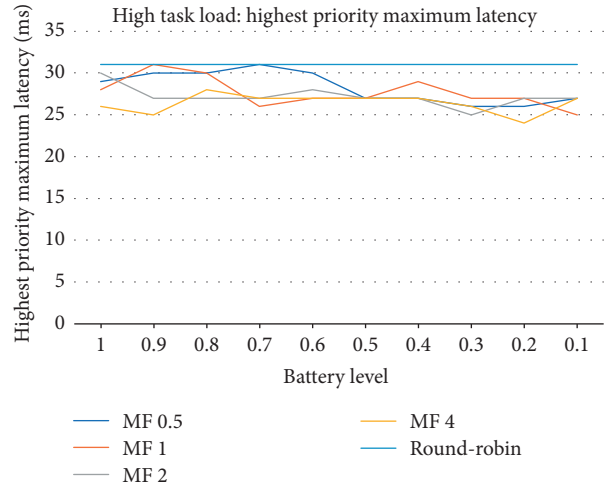


Figure 7: High task load test: maximum latency achieved by highest priority tasks.

The tasks latencies are also measured in order to evaluate how much the tasks execution is delayed in the proposed scheduler. We have measured the maximum and average latencies reached over all tasks during a test as well as the maximum and average latencies reached only by highest priority tasks, which should not be delayed in our adaptive scheduler.

For these tests, we use a modified version of FreeRTOS with most OS functionalities—in addition to the scheduler—such as memory management, tasks management, tasks communications, device drivers, and wireless stack. The tests are performed using the default FreeRTOS priority round-robin scheduler and using our adaptive scheduler in order to compare the performance of both.

6. Results and Discussion

For each of the three proposed scenarios, with different task load, tests have been performed varying the battery level from 1 to 0, with a step of 0.05. Therefore, up to 20 tests are executed for each scenario with different battery level values. Besides, the tests have been carried out with different multiplication factor values: 0.5, 1, 2, and 4. The priority round-robin scheduler has also been tested in order to compare the results with our scheduler.

First, we discuss the high task load scenario results. In Figure 5, the energy consumption is presented for this scenario over different battery levels and multiplication factor values. It can be seen that energy consumption is reduced when battery is discharging. That allows saving energy in low battery charge situations. The effect of the multiplication factor can also be noticed, represented in the figures as MF. Different MF values maintain the global behaviour. However, the scheduler starts saving energy at different battery level depending on MF. This way, the multiplication factor input may be used from user level to dynamically tune the scheduler behaviour.

Moreover, Figure 6 represents the average latency over all tasks and it can be seen that latencies are highly increased when battery level is low. However, Figure 7 shows that for highest priority tasks the maximum latencies are not increased as they have almost the same values as they do in the round-robin scheduler.

From now on, we will present the results only for multiplication factor 1, as this factor just tunes the scheduler behaviour maintaining the same functionality. For medium task load and low task load the results are quite similar, but moving the average energy consumption and task latencies to lower levels.

The results for medium task load are presented in Figures 8 and 9. The energy consumption is reduced when the battery level runs low and the task latencies are increased. The same behaviour can be seen in Figures 10 and 11 when the task load is low but displaced to lower values. Therefore, the results show that the scheduler behaviour is the same for different tasks sets, making the scheduler suitable for various applications with different tasks loads.

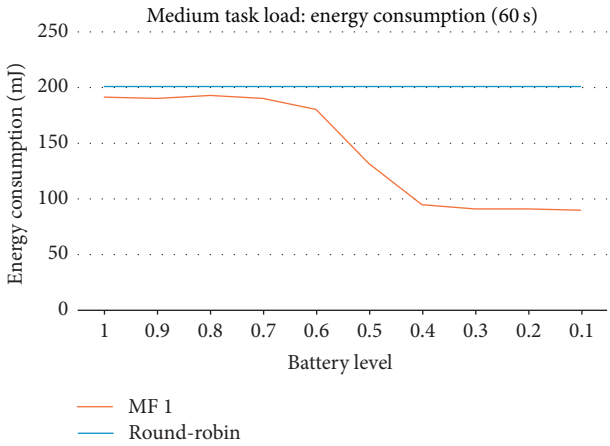


Figure 8: Medium task load test: energy consumption.

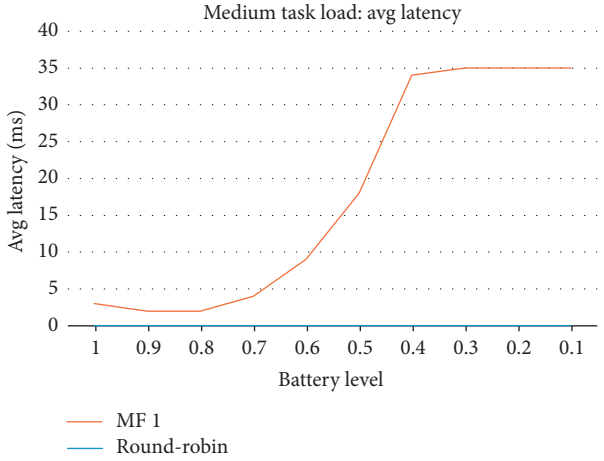


Figure 9: Medium task load test: average latency.

The latencies for highest priority tasks stand in the same level as in the round-robin scheduler, so real-time jobs could be performed with our scheduler even at low battery levels.

We also measure the overhead introduced by our scheduler in order to compare it to the overhead of a round-robin scheduler. In the tests performed the round-robin scheduler expends 58 milliseconds in the scheduling routines over 60-second tests. This time supposes 0.098% of the time which is despicable over the total time. On the other hand, our adaptive scheduler takes 83.4 ms during the scheduling routines and duty cycle decision algorithm. That means 0.14% of total time, which could be still considered despicable.

Finally, we obtain the expected lifetime of a node running our scheduler supposing it is powered by a 3000 mAh battery. Figure 12 shows the battery discharge rate of the round-robin scheduler compared with our proposed one for the three test scenarios and with a multiplication factor value of 1.

In this case the lifetime is extended from 48 days to 82 days, which means up to 71% increment. The lifetime is also obtained for medium and low task loads, which leads to 57% and 21% improvement, respectively. That means that our scheduler performs better with a higher task load.

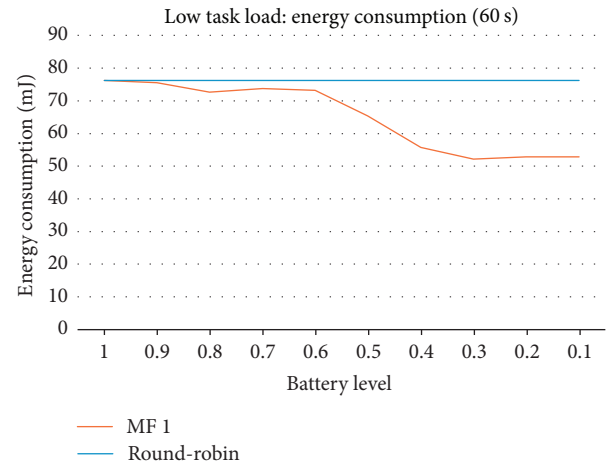


Figure 10: Low task load: energy consumption.

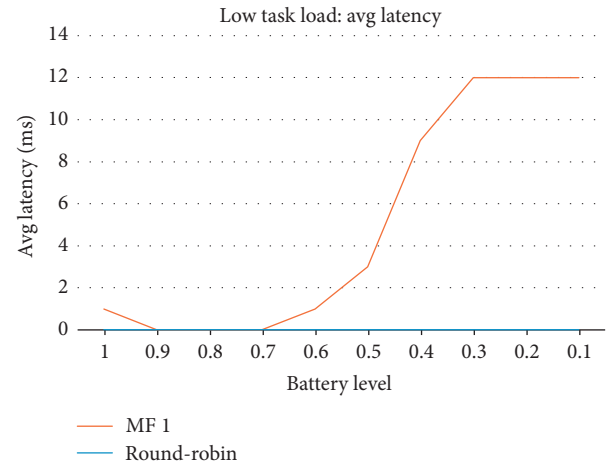


Figure 11: Low task load: average latency.

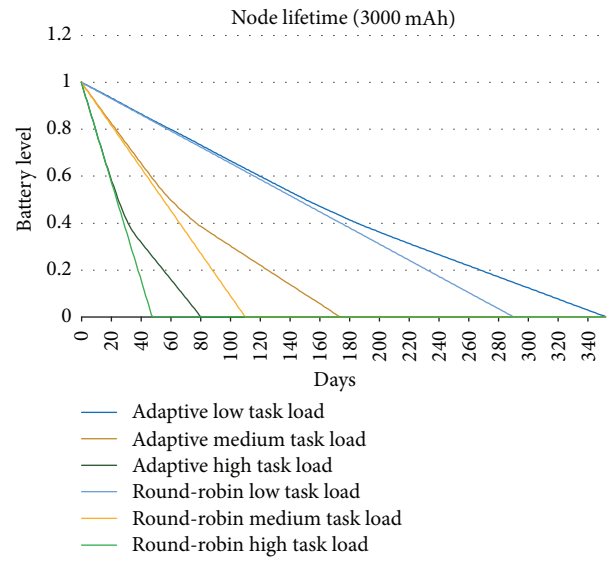


Figure 12: Node lifetime for test scenarios.

As the results obtained show, this scheduler could be highly suitable for battery-operated scenarios with energy harvesting sources. For example, if the nodes have solar panels as energy source, the battery is expected to be full during daylight hours, so the scheduler will run similar to a priority round-robin scheduler and low priority tasks will not be delayed. However, during night hours, the battery level will decay and the scheduler will start saving energy by delaying low priority tasks. This way, we could prevent the node running out of battery in cloudy days or in winter station when the night lasts longer than the day.

7. Conclusion

In this paper, we have proposed an adaptive scheduler architecture which makes possible change the task scheduling dynamically depending on the environment conditions. This could be very useful for WSN applications where changing environments are common. Specifically, we have targeted our scheduling algorithm at improving nodes lifetime, while it could be used for other optimization techniques in future works. The proposed scheduler changes dynamically its active duty cycle depending on battery level and tasks priorities. This leads to a large energy saving when battery charge is low and normal operation when battery is charged. For this duty cycle decisions, a PGG based algorithm is used and it is integrated in our scheduler architecture.

The scheduler proposed delays low priority tasks to achieve lower energy consumption, so they are executed with a higher period during low battery level states, which gives large energy saving. However, this latency does not affect high priority tasks as they are executed in all conditions, even when battery level is low.

Finally, the adaptive scheduler presented has been implemented and tested in real WSN nodes. The results show higher latencies when using our scheduler compared to a round-robin for low priority tasks. On the other hand, large energy saving is achieved and we can increase nodes lifetime up to 71% depending on the scenario.

The OS scheduler proposed is useful in many WSN scenarios to prevent nodes running out of battery by delaying noncritical tasks, while keeping high priority tasks running. This could lead to controlled degradation mechanisms for network nodes as they could maintain just critical functionality before the nodes run out of battery.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partially funded by the Spanish Ministry of Economy and Competitiveness, under RETOS COLABORACION program (Reference Grants SONRISAS: RTC-2015-3601-3, All-in-One: RTC-2016-5479-4 and EASYSAFE RTC-2015-3893-4), and the Spanish Ministry of Industry, Energy, and Tourism through the Strategic Action on Economy and Digital Society (AEESD) under DEPERITA: TSI-100503-2015-39 and SENSORIZA: TSI-100505-2016-10 projects.

References

[1] P. Levis, S. Madden, J. Polastre et al., "TinyOS: an operating system for sensor networks," in *Journal of Ambient Intelligence and Smart Environments*, pp. 115–148, Springer, Berlin, Germany, 2005.

[2] A. Dunkels, B. Grönvall, and T. Voigt, "Contiki—a lightweight and flexible operating system for tiny networked sensors," in *Proceedings of the 29th IEEE Annual International Conference on Local Computer Networks (LCN '04)*, pp. 455–462, November 2004.

[3] M. Chovanec and P. Šarafín, "Real-time schedule for mobile robotics and WSN aplications," in *Proceedings of the Federated Conference on Computer Science and Information Systems, FedCSIS 2015*, pp. 1199–1202, Poland, September 2015.

[4] X. Liu, K. M. Hou, C. De Vaulx, H. Zhu, and J. Liu, "Memory optimization techniques for multithreaded operating system on wireless sensor nodes," in *Proceedings of the 2014 2nd IEEE International Conference on Progress in Informatics and Computing, PIC 2014*, pp. 503–508, China, May 2014.

[5] S. P. Patil and S. C. Patil, "A real time sensor data monitoring system for wireless sensor network," in *Proceedings of the 2015 IEEE International Conference on Information Processing, ICIP 2015*, pp. 525–528, India, December 2015.

[6] O. Hahm, E. Baccelli, H. Petersen, M. Wählisch, and T. C. Schmidt, "Demonstration abstract: Simply RIOT - Teaching and experimental research in the Internet of Things," in *Proceedings of the 13th IEEE/ACM International Conference on Information Processing in Sensor Networks, IPSN 2014*, pp. 329-330, Germany, April 2014.

[7] C. Brandolese, W. Fornaciari, and L. Rucco, "Power management support to optimal duty-cycling in stateful multitasking wsn," in *Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2013*, pp. 1123–1132, Australia, July 2013.

[8] S. Akkermans, W. Daniels, G. S. Ramachandran, B. Crispo, and D. Hughes, "CerberOS: a resource-secure OS for sharing IoT devices," in *EWSN 2017*, 2017.

[9] A. Sleman and R. Moeller, "SOA distributed operating system for managing embedded devices in home and building automation," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 2, pp. 945–952, 2011.

[10] B. Pasztor and P. Hui, "OSone: a distributed operating system for energy efficient sensor network," in *Proceedings of the 2013 25th International Teletraffic Congress, ITC 2013*, China, September 2013.

[11] S. Zoican, R. Zoican, and D. Galatchi, "Improved load balancing and scheduling performance in embedded systems with task migration," in *Proceedings of the 12th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services, TELSIKS 2015*, pp. 354–357, Serbia, October 2015.

[12] B. Porter and G. Coulson, "Lorien: A pure dynamic component-based operating system for wireless sensor networks," in *Proceedings of the MidSens'09 - 4th International Workshop on Middleware Tools, Services and Run-Time Support for Sensor Networks, Co-located with the 10th ACM/IFIP/USENIX International Middleware Conference*, pp. 7–12, USA, December 2009.

[13] M. Gasmi, O. Mosbahi, M. Khalgui, L. Gomes, and Z. Li, "R-Node: New Pipelined Approach for an Effective Reconfigurable Wireless Sensor Node," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–14.

[14] Z. Jing, X. Leng, H. Fan, and C. Yi, "TQS-DP: A lightweight and active mechanism for fast scheduling based on WSN operating system TinyOS," in *Proceedings of the 27th Chinese Control and Decision Conference, CCDC 2015*, pp. 1470–1475, China, May 2015.

[15] G. C. Sirakoulis and I. G. Karafyllidis, "Cooperation in a power-aware embedded-system changing environment: public goods games with variable multiplication factors," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 42, no. 3, pp. 596–603, 2012.

[16] STMicroelectronics, http://www.st.com/en/microcontrollers/stm32l476re.html.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 5795037, 13 pages
https://doi.org/10.1155/2018/5795037

Research Article

SVM-Based Dynamic Reconfiguration CPS for Manufacturing System in Industry 4.0

Hyun-Jun Shin, Kyoung-Woo Cho, and Chang-Heon Oh

Department of Electrical, Electronics and Communication Engineering, Korea University of Technology and Education, 1600 Gajeon-ri, Byeongcheon-myeon, Dongnam-gu, Cheonan-si, Chungcheongnam-do 31253, Republic of Korea

Correspondence should be addressed to Chang-Heon Oh; choh@koreatech.ac.kr

Received 28 July 2017; Accepted 18 December 2017; Published 29 January 2018

Academic Editor: Yong Ren

CPS is potential application in various fields, such as medical, healthcare, energy, transportation, and defense, as well as Industry 4.0 in Germany. Although studies on the equipment aging and prediction of problem have been done by combining CPS with Industry 4.0, such studies were based on small numbers and majority of the papers focused primarily on CPS methodology. Therefore, it is necessary to study active self-protection to enable self-management functions, such as self-healing by applying CPS in shop-floor. In this paper, we have proposed modeling of shop-floor and a dynamic reconfigurable CPS scheme that can predict the occurrence of anomalies and self-protection in the model. For this purpose, SVM was used as a machine learning technology and it was possible to restrain overloading in manufacturing process. In addition, we design CPS framework based on machine learning for Industry 4.0, simulate it, and perform. Simulation results show the simulation model autonomously detects the abnormal situation and it is dynamically reconfigured through self-healing.

1. Introduction

The term Industry 4.0 refers to a strategy of German manufacturing industries in which strategy copes with a change such as social, technological, economic, ecological, and political using Information Communication Technology (ICT). The aim of Industry 4.0 is to primarily create a smart factory that will use ICT technologies actively, such as Internet of Things (IoT), enterprise software, location information, security, cloud, big data, and virtual reality. The Cyber-Physical System (CPS) plays a critical role in realizing Industry 4.0. CPS acts as a medium to link physical world, such as sensors, actuators, and mobile devices, with Internet service and also to mirror what happens in the real world to a cyber space to process preinspection, real-time management, and postmortem. Europe, Sweden, US, China, and South Korea use CPS in an attempt to realize Industry 4.0 [1, 2]. Recently, manufacturing countries in an industrially advanced nation are rapidly shrinking production populations, and the rate of elderly dependency is soaring. This decrease in production population is affecting the labor productivity, which is the foundation of a manufacturing industry. In this regard, Industry 4.0 emerged so that manufacturing evolution can complement future competitiveness.

The manufacturing facility is generally operated by a preset program under existing factory automation system. On the other hand, the manufacturing facility must decide how to operate autonomously in Industry 4.0. Smart manufacturing by a smart factory involves facilities and processing of an individual factory and shares and uses all production information by combining ICT with traditional manufacturing, thereby making it possible to achieve optimal production and operation. At the same time, it also connects related factories to establish a production system which will allow continued collaboration through extension of the smart manufacturing concept [3].

CPS refers to a computer-based component and system that closely connects various complicated processes and information of real space with the cyber space that provides data access and processing services through Internet. The smart factory CPS helps making optimal decision for the network connecting the manufacturing equipment as well as their design and operation through intelligent context awareness, decision making, and execution [4, 5]. In spite of

being old itself, CPS can be used to develop a new technology by interfacing it with existing technologies, such as multiagent systems (MASs), service-oriented architectures (SOAs), wireless sensor networks (WSNs) [6], Internet of Things (IoT) [7, 8], cloud computing [9–14], augmented reality, big data [15], machine-to-machine (M2M), and mobile Internet [16]. Still, there are important tasks such as safety, security, and interoperability that need to be considered [17].

In the past decade, research on CPS concept, modeling method, and application method was broadly divided into studies on the integration of CPS technology with other ICT technologies or existing systems for application in manufacturing. The most commonly used keywords are cyber model, digital twin, real-time modeling, and analysis [18]. Studies on the application in manufacturing primarily involved problems such as aging of equipment and prediction of problems, and they were solved by using machine learning and artificial intelligence. Prior reports showed that a few actual manufacturing cases were solved, but such papers are a few in number and most of them focused on CPS methodology. In the early stage, the conceptual approach of the whole system or presentation of design methodology and partial application of elemental technology are mainstream, and more specifically, integrated and empirical research is needed.

In this paper, CPS was applied to shop-floor as a part of CPS research. The overall goal was to use machine learning to enable self-management functions, such as self-healing, and to prevent the system from further degradation, thereby, providing active self-protection and self-healing. To achieve this, we executed shop-floor modeling and applied self-healing in the modeling. For this purpose, 5C's CPS architecture model of Lee et al. was used. The 5C's CPS architecture model consists of Connection, Conversion, Cyber, Cognition, and Configuration. We have reconstructed the manufacturing process based on this. The manufacturing site modeled the conveyor belt manufacturing system using the M/D/1 queue, and the parameters used were μ, λ, and ρ. SVM, a machine learning method, was used to predict the occurrence of abnormal conditions, and an abnormal situation was detected through the change of ρ. These concepts and researches can serve as reference models for building CPS and can be useful in the design step before starting the application.

Section 2 will describe a related architecture research and basic research for implementing CPS. A framework for dynamically reconfiguring CPS-based shop-floor will be introduced in Section 3. Section 4 will explain the proposed system and its results. Finally, Section 5 will complete this with conclusions.

2. Related Research Work

In Section 2, we will describe three related studies for CPS implementation. Section 2.1 describes the architecture underlying the dynamic reconfiguration framework, Section 2.2 describes the Queuing Theory on which the simulation model is based, and Section 2.3 deals with related machine learning that is the basis for self-healing.

2.1. Cyber-Physical System. Figure 1 shows the results of Lee et al., who proposed CPS architecture of an Industry 4.0 based manufacturing system [19]. The architecture comprises 5 levels, which is "connection," "conversion," "cyber," "cognition," "configuration." It consists of methodologies and guidelines for CPS deployment for manufacturing from step-by-step design and data collection for analysis and final value creation. The paragraphs below explain the function of each level in detail.

2.1.1. Connection Level. Acquiring accurate and reliable data from machines and their components is the first step in developing a Cyber-Physical System application. The data might be directly measured by sensors or obtained from controller or enterprise manufacturing systems, such as ERP, MES, SCM, and CMM.

2.1.2. Conversion Level. Meaningful information needs to be inferred from the data. Currently, there are several tools and methodologies available to draw inference from the data in the information conversion level.

2.1.3. Cyber Level. The cyber level acts as central information hub in this architecture. Information is being pushed to it from every connected machine to form a machines network. Having massive information gathered, specific analytics have to be used to extract additional information that provides better insight on the status of individual machines among the fleet.

2.1.4. Cognition Level. Implementing CPS in this level generates a thorough knowledge of the monitored system. Proper presentation of the acquired knowledge to expert users supports leads to correct decision of the users. Since comparative information as well as individual machine status is available, decisions based on priority of tasks can be made taken to sustain optimal maintaining process.

2.1.5. Configuration Level. The configuration level is the feedback from cyber space to physical space and acts as a supervisory control to make machines self-configuring and self-adaptive. This stage acts as resilience control system (RCS) to apply the corrective and preventive decisions, which have been made in cognition level, to the monitored systems.

Lee et al. proposed a 5C's CPS architecture to achieve the goal of resilient, intelligent, and self-adaptable system. CPS in a manufacture and automation environments can be applied to diverse processes including simulation, design, control, and verification. In manufacturing, CPS can improve quality and productivity through smart presymptom and diagnosis using big data from different machines, network sensors, and systems. In addition to this, various related studies have been carried out, but the focus was primarily on the role of the CPS in methods for applications connected with technologies, such as manufacturing, application scenarios, conceptual or architectural design, and big data, analysis, IoT, and human-machine interface (HMI) [20, 21]. Additionally, the degree of CPS implementation in the enterprises is still low. These

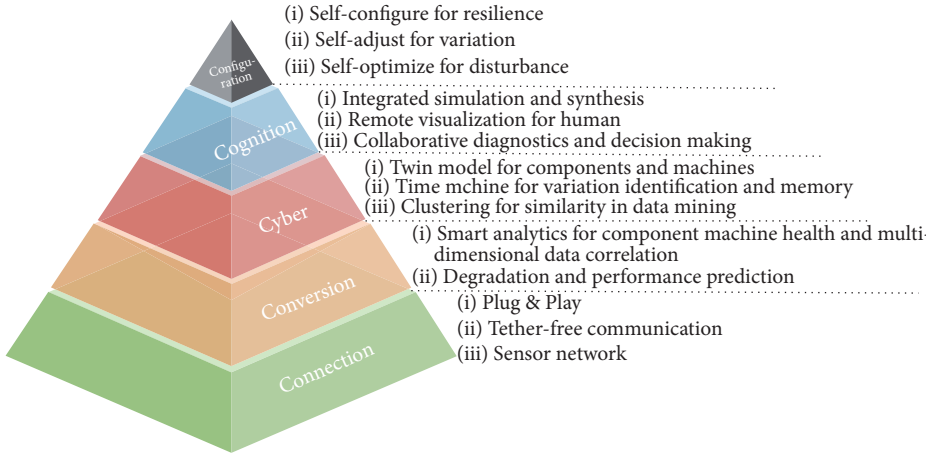


Figure 1: "5C" architecture of cyber-physical systems.

concepts and studies can act collectively as reference models for building CPS and can be useful in the design phase before starting an application. However, they dealt with issues such as cyber model which is essential for more practical implementation of CPS. In the present work, the production site of the manufacturing process was implemented through CPS. This can help to reduce the technology gap in the stage of technological innovation.

2.2. Queuing Theory. The Queuing Theory creates models (consequent insights) that are useful in predicting behavior of systems which provide services to randomly generate demand. It is also important to consider the statistical distribution of production operations (ex, process time, process cycle, and production mix) that allow for a description of the complex environment. When actually modeling a production system, the main benefits of Queuing Theory are the probability, average time of the system, average service time, average work time, work time, average number of customers in the system, and the probability of number of customers who will be in the system.

The use of Queuing Theory allows rapid modeling of a production system even when there are certain uncertainties in the environment. These uncertainties can be managed by statistical distribution of parameters, such as arrival and service rate of the queuing model.

Figure 2 shows a typical Queuing Theory, comprising input, output, queue, and service time of a production.

Table 1 shows the parameters associated with the adopted notation. The most commonly used parameters are λ, μ, and ρ. ρ is an important parameter that describes how busy a server is during a period of time.

This Queuing Theory is used in systems such as logistics service, AGV, Less Than Truckload (LTT), conveyor belt for assembling parts, airports with a queue for runway access, elevators of banks, and warehouses. The stochastic of transport routes, arrivals, and service times is mainly studied. However, there are a few studies on ρ of server. In this paper, we used a utilization of server to change the manufacturing process efficiently.

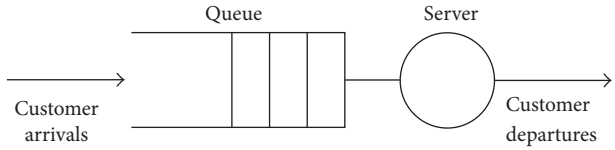


Figure 2: Queue of machine.

Table 1: Notation of the queuing models.

Symbol	Units	Description
λ	Job/h	Mean arrival rate of *jobs* at the system
μ	h	Mean service of *jobs* in the system
ρ	%	Utilization coefficient of the department

2.3. Machine Learning in CPS. Artificial intelligence techniques, such as artificial neural networks, inductive learning methods, case-based reasoning, and genetic algorithms, have been applied to the prediction field to recognize, predict, and reconstruct the present situation. Odom and Sharda were the first to apply artificial neural networks to predictions [22]. They compared prediction rates by applying discriminant analysis and artificial neural network model, between which the artificial neural network model showed better results than the discriminant analysis. Tam and Kiang applied artificial neural networks and compared the results with those of discriminant analysis, Logit, k-nearest neighbor, and inductive reasoning. As a result, the model based on artificial neural network showed better results in prediction and adaptability than other methods.

Despite the excellent predictive accuracy of the artificial neural networks, the main limitation is that it is difficult to explain the cause of the prediction results and the possibility

of generalization is also reduced. Furthermore, another disadvantage is that a lot of time and effort are required to design an artificial neural network structure and excessive suitability problem in constructing an artificial neural network model.

In this paper, we have proposed a solution to the above-mentioned problems by recognizing the present situation using support vector machine (SVM).

SVM proposed by Vapnik in 1995 is a learning algorithm that first divides input data into two groups and then analyzes them [23]. Figure 3 shows a typical SVM. To separate the data, the support vector which is the farthest away from the opposite group of data is found, the hyperplane, which is the criterion for dividing into two groups, is determined, and the margin is then calculated. There can be multiple hyperplanes, but there is one hyperplane that maximizes the margins and the distance between the support vector and the hyperplane. In our study, we found the hyperplanes and separated the data.

We give a brief mathematical summary of the classical SVM for binary-class classification. Assume there is a group of independent training samples, as shown in the following equation [24]:

$$\left\{x_i, y^i\right\}, \quad x_i \in R^n, \ y_i = \pm 1, \ i = 1, 2, \ldots, l. \tag{1}$$

Given that the adopted classification method of the samples is proposed as shown in the following equation:

$$f(x) = \operatorname{sng}(w \cdot x + b), \tag{2}$$

so, SVM line subclassification can convert a quadratic regression which can be recorded as

$$\begin{aligned} \min \quad & \left(\frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{l}\xi_i\right)^P\right), \\ \text{s.t.} \quad & y_i(w \cdot x + b) \geq 1 + \xi_i, \\ & \xi_i \geq 0, \ i = 1, 2, \ldots, l, \end{aligned} \tag{3}$$

where C stands for the penalty factor, the greater its experience error value is, the greater the penalty will be. By the application of Lagrange's multiplier method, (3) can be changed into a Wolfe Dual Planning shown as

$$\max \quad \left(\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{i}\alpha_i\alpha_j x_i \cdot x_j\right). \tag{4}$$

α_i and α_j stand for Lagrange multipliers. In this way, after the adoption of dual planning, the research separates the SVM and the input sample dimensions, thus to avoid the appearance of so-called "Dimension Disasters." The final liner function for SVM can be shown as in the following equation:

$$f(x) = \operatorname{sng}(w \cdot x + b) = \operatorname{sng}\left(\sum_{i=1}^{l}\alpha_i\alpha_j y_i x_i + b\right). \tag{5}$$

For nonlinear problems, substituting the kernel function $k(x_i, x_j)$ into (6), one can obtain a final nonlinear function for SVM as shown in the following equation:

$$f(x) = \operatorname{sng}\left(\sum_{i=1}^{l}\alpha_i y_i k\left(x_i, x_j\right) + b\right). \tag{6}$$

The most important part of CPS is self-healing. There are various ways to solve problems when they occur, and self-healing using machine learning is becoming more popular these days [25, 26]. However, they are limited to real-time monitoring, as they do not only detect and diagnose machine failures, defective products, or training and test predefined dataset. Therefore, it is necessary to study the dynamic reconfiguration of manufacturing process based on CPS when an abnormal situation occurs.

3. Machine Learning Based Self-Aware Machines

Smart Factory is a manufacturing CPS that integrates physical objects, such as machines, conveyors, and products with information systems to enable flexible and agile production. In this section, a framework and shop-floor modeling for smart factory will be proposed. Discrete event simulation will be used to evaluate the proposed model.

3.1. Framework. The concept of smart manufacturing is actually based on the integration of IoT and CPS concepts. IoT's vision is to interconnect millions of devices and interconnect them with enterprise systems. The combination of IoT and CPS is essential to provide users with the data provided by the millions of devices in the shop-floor. Applying the general concept of CPS to the manufacturing system is called cyber-physical production system (CPPS). CPPS is the factor that enables IoT in the manufacturing process. Thus, the CPPS concept allows for high level integration and interoperability of manufacturing applications and systems by improving autonomy and flexibility in industrial environments. As the network communication technology developed, the virtual world that emerged as IoT and the real world have a vision to harmonize with each other. This indicates that it ensures a smooth data flow between real-time data on the shop-floor and information of the management system. Figure 4 shows the work type of shop-floor. In the manufacturing system, conveyor belt, AGV, warehouse, and machine exist. In order to obtain data of each equipment, data should be provided to users through wired/wireless communication networks based on smart object.

Figure 5 shows that a smart factory framework consists of physical layer and cyber layer. The physical layer transmits the actual data generated at the shop-floor to the cyber layer through an industrial network [27]. Shop-floor based real data is collected in real time on all elements in the factory layout, from automation facilities to equipment operated by the operator, work performed by the operator, warehouse,

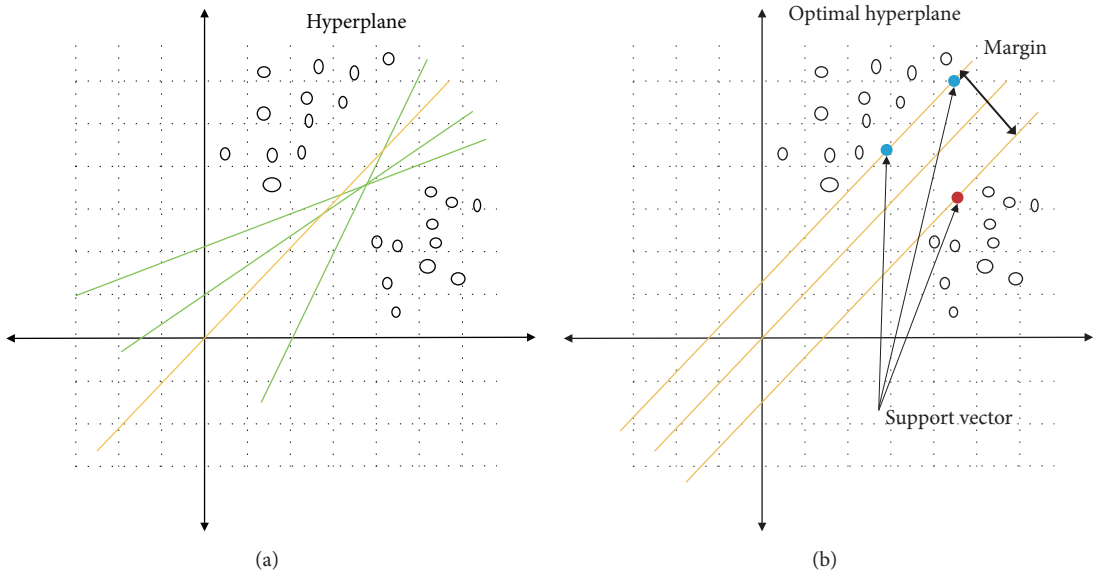


Figure 3: Compositions of support vector machine.

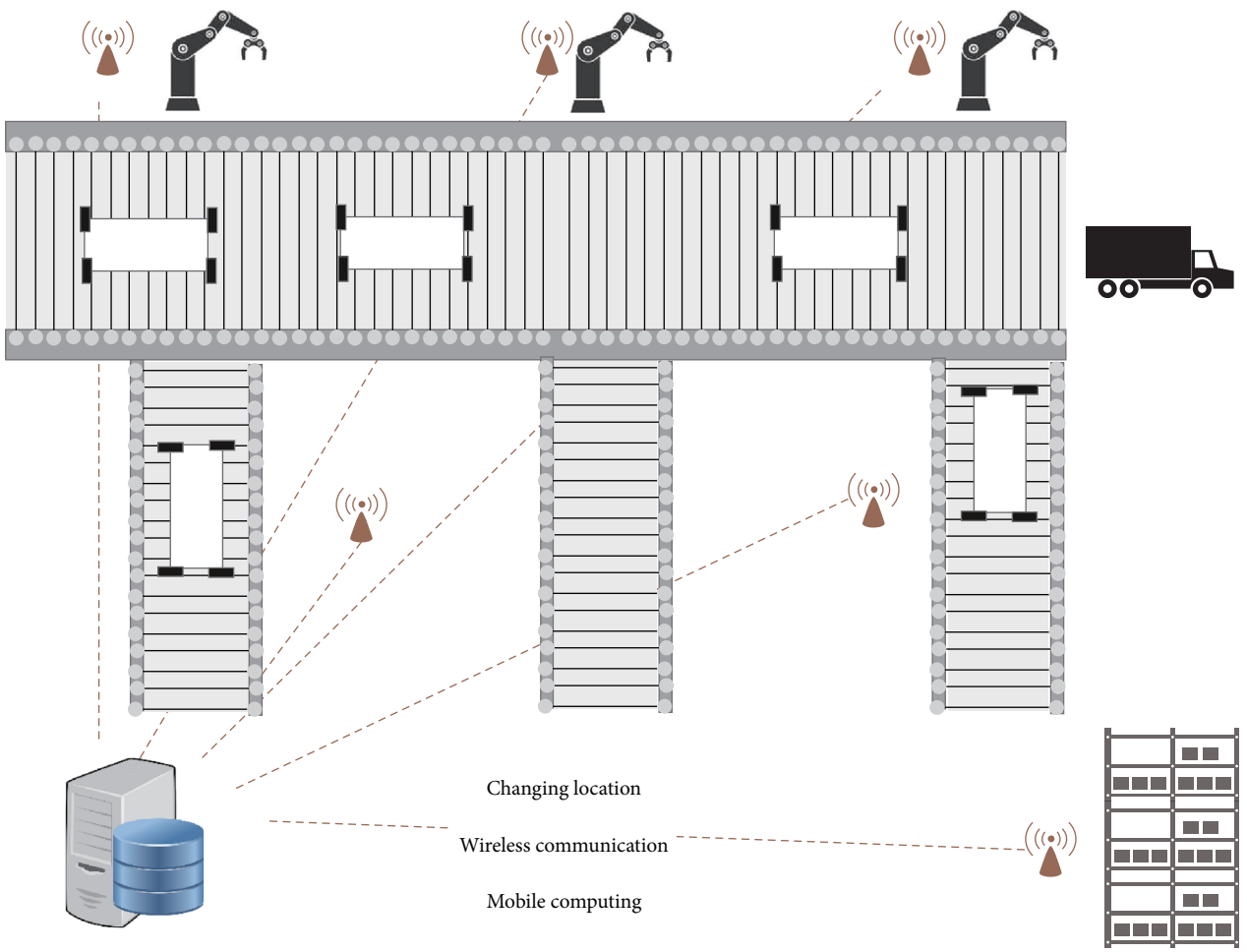


Figure 4: Manufacturing system case study.

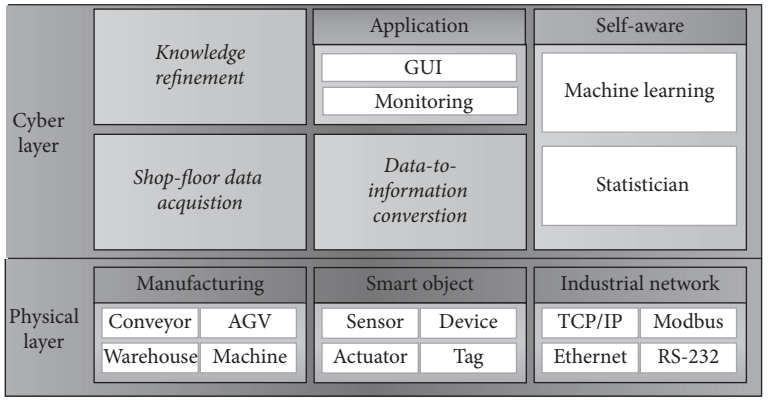


Figure 5: Framework of reconfiguration CPS.

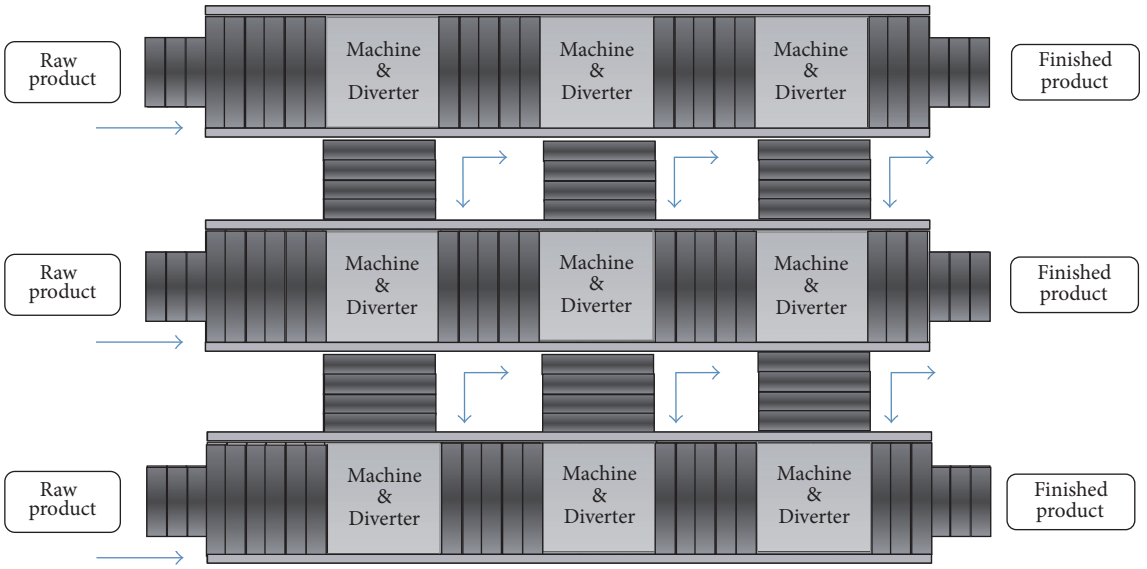


Figure 6: Shop-floor field modeling.

buffer, conveyor, and logistics facilities such as AGV. Data must be collected through smart objects such as sensors, devices, actuators, and tags. A smart object is an intelligent electronic device that has built-in Internet access control function which makes it easy to access online anytime and anywhere, thereby enabling data collection by equipment in the cyber layer. The cyber layer collects all the data from the industrial site and converts it into meaningful information [28]. Actually, the data generated in the physical layer is diverse and very large. Thus, it is necessary to reduce and convert the data to make it suitable for techniques such as machine learning and big data analysis [29]. The transformed information is trained through machine learning technology to generate a model and the generated model is then tested. The output data may be used for monitoring or GUI provided for user's service. The output data is also kept for future knowledge improvement.

3.2. Shop-Floor Modeling. Figure 6 shows a virtual model of the conveyor belt shop-floor. There are three products in the model, each product is manufactured and transported to the next line. If the process time of a particular device is long or short, there may be a change in the input quantity, which may indicate out of order of the machine. In such a case, it is necessary to stop the machine or change the order of operations with other equipment. The path of the model is changed through the diverter.

Figure 7 shows the open queuing network model. This shows the conveyor belt in Figure 6 as a queuing model. We will assume that a single server, and all nodes operate according to a FIFO queuing discipline. It is assumed that the data of all nodes are transmitted by wireless communication. The following assumptions are followed for modeling implementation and testing.

(1) Input product arrives at the system following a Poisson distribution.

(2) The machine's queue and server follow the M/D/1 standby queue.

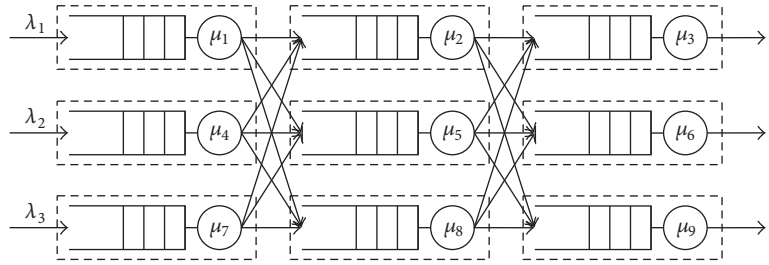


Figure 7: Queuing network model.

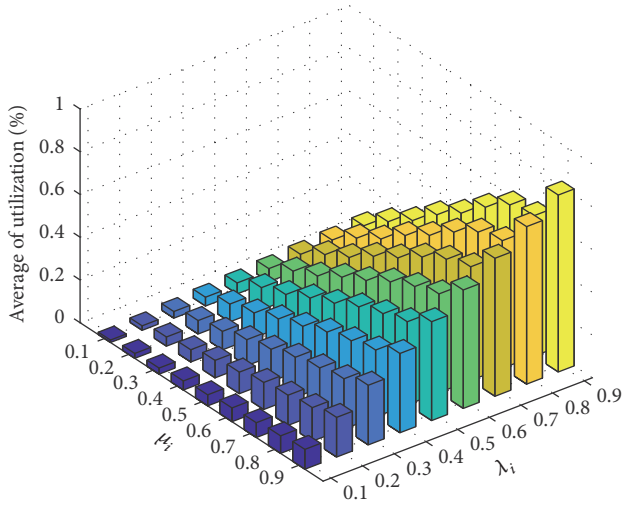


Figure 8: Parameter correlation.

The Poisson distribution is a discrete probability distribution that represents how many events occur within a unit of time. If the probability is sufficiently large or the probability is small enough, the Poisson distribution can approximate the problem. The M/D/1 queue is a model used when the service time is deterministic rather than random. It is a single server which sets the machine's working time constant in the production system and the number of machines as one.

Table 2 shows the average arrival rate (λ) and the average service rate (μ) as parameters of M/D/1 used in the model. The following relationship can be obtained in the M/D/1 queue [30].

$$\rho = \frac{\lambda}{\mu}. \tag{7}$$

If $\rho = 1$, it means that the server is operating 100 percentages, and if $\lambda > \mu$, the service of the equipment is blocked. In this paper, we do not consider the ratio of the server over 100 percentages and since μ is set to 1 at maximum, λ is specified as 0.1~0.9, according to (7).

Figure 8 shows the correlation of three parameters through (7). In order to verify the quality of the manufacturing process using SVM, a machine learning technology, the input data needs to be divided into two groups. The input parameter is required to divide into two groups, ρ, λ, where ρ is the percentage of time that the server works on all of the time. The results of ρ obtained according to the ratio of λ and the ρ obtained by changing μ during the manufacturing process are placed in two groups. Then, test is done through the newly modified μ.

Table 2: M/D/1 queue parameter.

Input parameter		
λ	μ	ρ
0.1~0.9	0.1~1	0~1

Figure 9 shows the SVM-based dynamic reconfiguration CPS flowchart. When the shop-floor shown in Figure 5 was initially constructed, the process proceeded to the M/D/1 queue and the data (λ, ρ, and μ) of the generated queues was input to the SVM training module. The SVM training module finds a support vector for the input data, divides the input vector into two groups, and calculates hyperplanes and margins. The data in the queue which will be processed in future is input to the SVM test module so that it belongs to one of the two groups "class 1" and "class 2" generated in the SVM training module. Then if the SVM test result belongs to "class 1," it decided that there is no abnormality in the equipment, whereas if it belongs to "class 2," it decided that the equipment is abnormal. If an abnormality is decided, it needs to be checked whether the average ρ of the equipment is out of the range of "$\lambda \pm$ threshold" and then change the path after confirming whether the state of the peripheral equipment is normal.

4. Simulation and Results

In this paper, we implemented the model through Matlab SimEvent of Mathworks, discrete event simulation software [31–34]. The remainder of this section describes the verification and validation of the simulation model and some preliminary results. In order to implement CPS-based environment, a network system capable of systematically managing collected data using smart objects, such as sensors and actuators and industrial networks, is needed. However, in this study, simulation software is used to collect data because there is no environment that can obtain data from factories through sensors [35, 36].

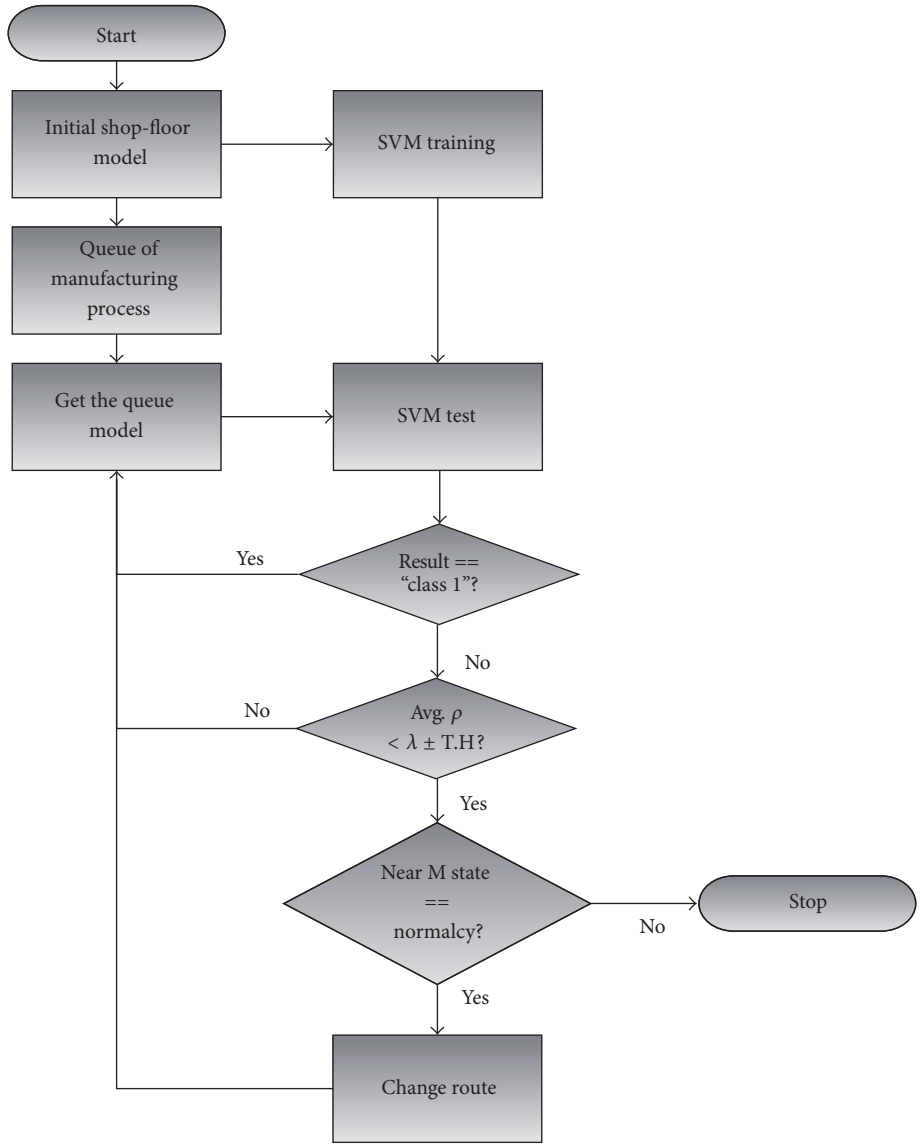


Figure 9: Flowchart of SVM-based dynamic reconfiguration CPS.

Figure 10 shows implementation of a conveyor belt at the shop-floor. The production time of the initial product and the process time of each equipment can be adjusted, and the number of production of the product can be confirmed. Exponential Arrival Time (EAT) can be generated with a Poisson distribution of 0.1 to 0.9, and Stamp Entity (SE) can cause an event to change μ during the manufacturing process.

Figure 11 shows the inside of each machine block. After fixing the machine service based on the M/D/1 system, μ is changed according to the event occurrence. If one needs to change the conveyor path by changing μ in the machine, the path can be changed through the entity output switch, which acts as a diverter. This signifies transportation of product to another line.

SVM is a machine learning algorithm that analyzes and classifies various input variables, as mentioned above. In this problem, λ and μ are used as input variables. As shown in Figure 7, training was performed through a predetermined λ and μ was changed in the manufacturing process.

Based on the input/output variables defined in Table 3, proceed according to the process represented in Figure 9. As mentioned above, if the machine shows no change in service time, it is assumed that it is under normal condition. On the other hand, if there is a change, it is assumed that it is an abnormal situation. Therefore, the output data is set as training and test output values before and after the change of μ.

Figure 12 shows the training results, where the x-axis represents λ and the y-axis represents the mean value of ρ from 0.1 to 0.9. The value of ρ was obtained by repeated experiment from 0.1 to 0.9 after fixing μ and λ. In this paper, "class 1" was used when there was no abnormal situation of the machine, and "class 2" referred to when the value of μ was changed. To note, the newly input data has been

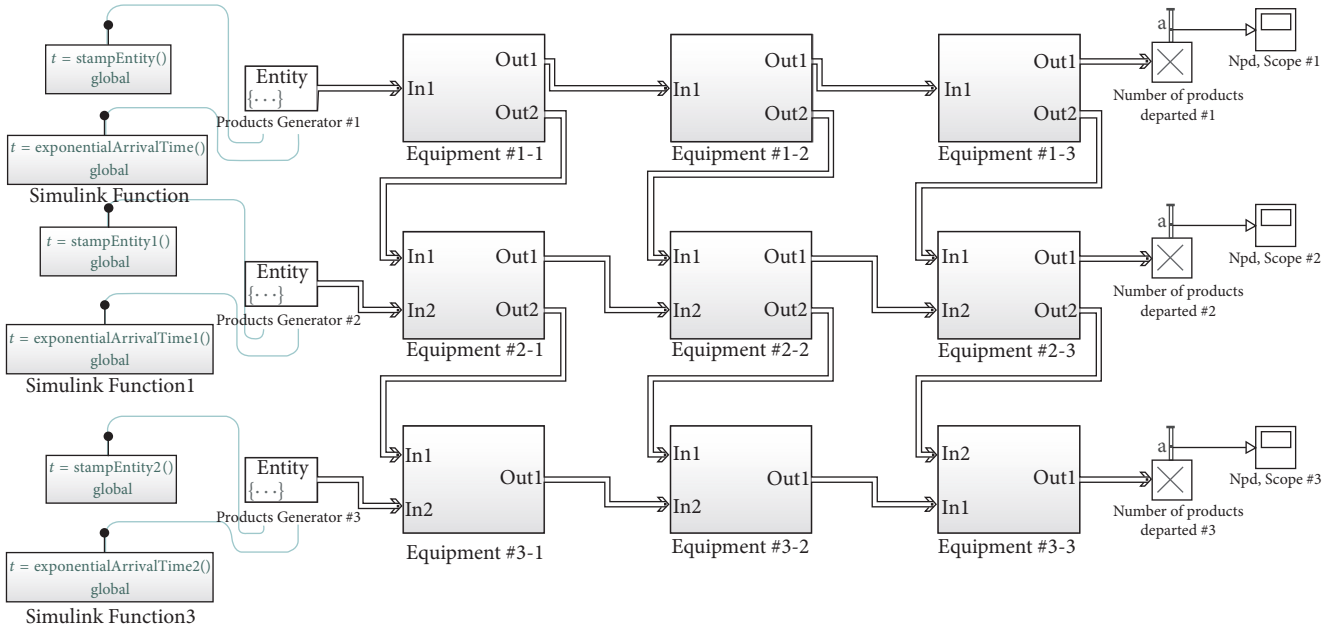


Figure 10: Matlab-based shop-floor field.

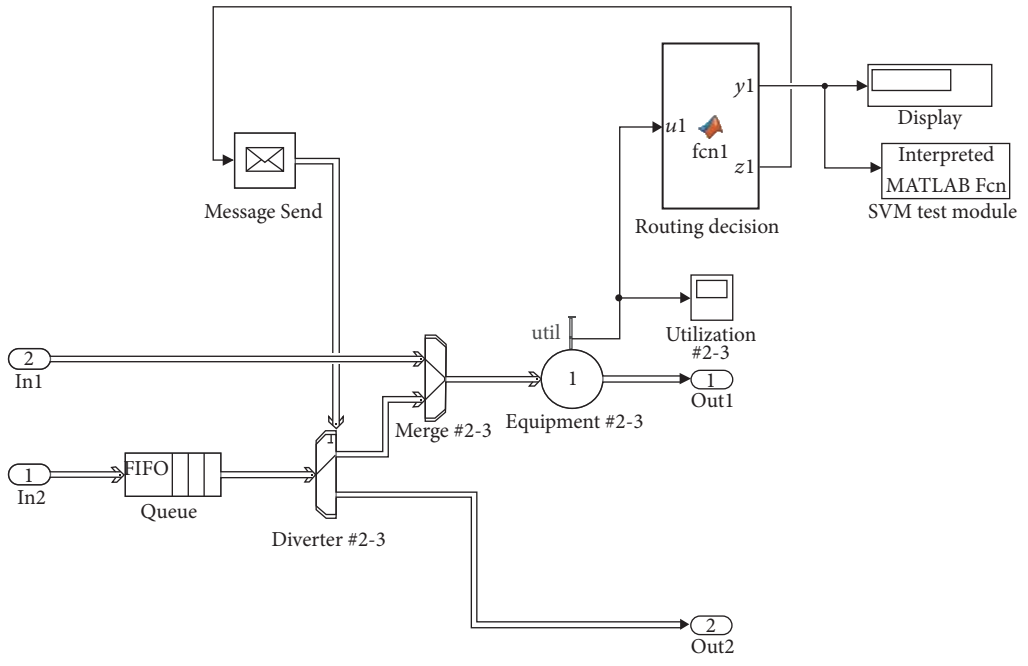


Figure 11: Machine process block.

Table 3: The variables of input and output data set.

	Input	Output
Before μ change (normal state)	λ_i	SVM training results
	ρ_i	
After μ change (abnormal detection)	λ_i	SVM test results
	ρ_i	

classified as "class 2" because the new data is located below the hyperplane.

In SVM modeling, property selection and parameter setting are important. These two have a decisive influence on the efficiency and accuracy of SVM classification. We used the Grid-search (GS) algorithm for parameter optimization. The Grid-search method is a method of finding optimal parameters by attempting a discrete value of a suitable interval within a predetermined range.

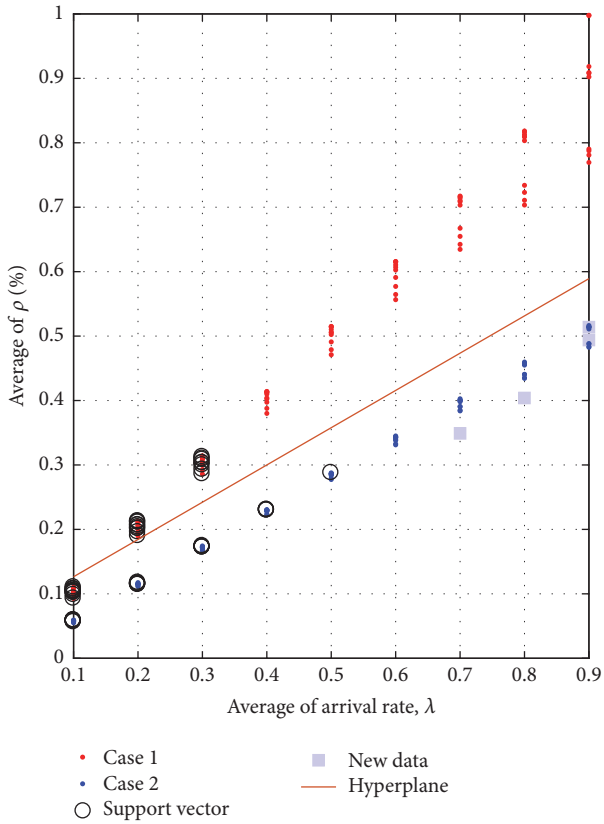


FIGURE 12: Detect abnormal situation using SVM.

```
*.*
optimization finished, #iter = 313
nu = 0.310837
obj = −312.668385, rho = 0.976663
nSV = 57, nBSV = 52
Total nSV = 57
```

Box 1: Result of classification using LibSVM.

Two parameters (C, r) are required to execute the SVM using the Radial Basis Function (RBF) kernel. C is the penalty parameter of the SVM, and r is the kernel parameter. In the GS, basically, (C, r) pair with the highest cross-validation accuracy is chosen. Thus exponentially increasing (C, r) values finds the optimal parameter. In this paper, C and R were obtained using GS during training.

Box 1 shows the model result obtained after training. From the output, obj is optimal objective value of the dual SVM problem. The value ρ is $-b$ in the decision function. nSV and nBSV are number of support vectors and bounded support vectors, respectively.

In order to verify the performance of the SVM-based dynamic reconfiguration production system proposed, we compared the server ρ before and after the abnormal situation occurred and then proceeded with the reconfiguration process of the production system. Abnormal situation means that the process rate of the machine is overloaded or the rate of service is changed due to decrease in speed. The processing time was 10,000 sec and the time and place of occurrence of the abnormal situation occurred randomly.

Figure 13 shows the server ρ of machines #1-3 and #2-3 when no abnormalities occur. The λ of each machine were 0.7 and 0.8. It was observed that ρ was similar to λ when no abnormal situation occurred. The average value of ρ were classified as "class 1" in Figure 12.

Figure 14 shows the variation of ρ after (a) and (b) occurred at t = 3,000 and 4,000. Abnormal situations indicate situations such as overloading or slowing down of the machine, which lowers ρ. The average value of this ρ has been classified as "class 2" in Figure 12.

Figure 15 shows the ρ after the change of the production route after the abnormal situation occurs. After the abnormal situation occurred at t = 3,000 on machine #1-3, the average ρ was out of the range of $\lambda \pm$ threshold, and the production route was changed at t = 8,000 to the surrounding machine. On machine #2-3, the production route has not changed after the abnormal situation occurred at t = 4,000, because ρ has not exceeded threshold. The product of machine #1-3 flow into machine #2-3 and increased at t = 8000. This indicates that the simulation model has been reconstructed by recognizing machine #1-3 as an error in the model test.

Figure 16 shows a number of products produced after an abnormal situation has occurred and the production route

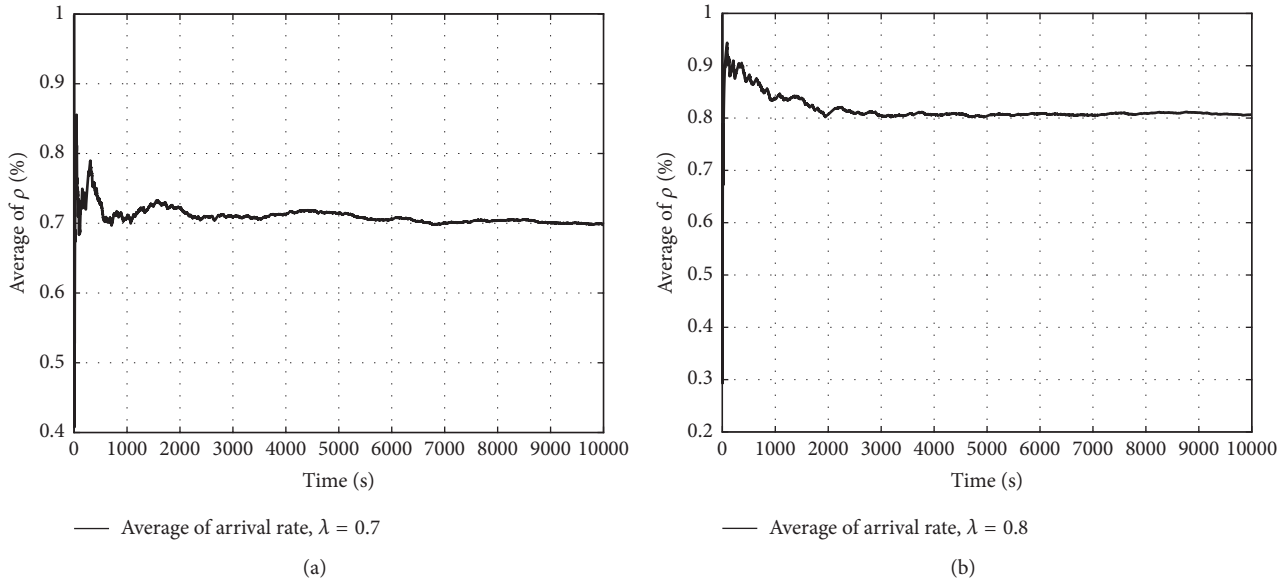


FIGURE 13: Simulation results for server utilization when no event occurred (%): (a) server utilization in machine #1-3 ($\lambda = 0.7$); (b) server utilization in machine #2-3 ($\lambda = 0.8$).

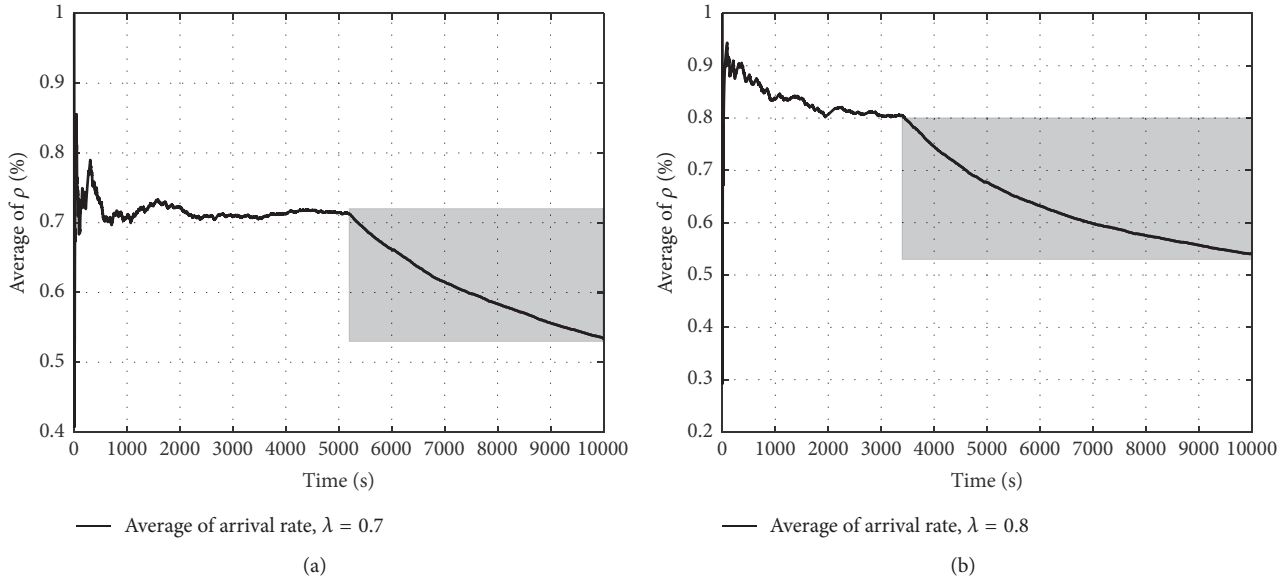


FIGURE 14: Simulation results for ρ during event occurrence: (a) event occurrence in machine #1-3 ($t = 5{,}200$); (b) event occurrence in machine #2-3 ($t = 3{,}400$).

has changed. (a) shows that machine #1-3 is stopped at $t = 8{,}000$, and (b) shows that the number of products increases because the products have flowed from machine #1-3 to machine #2-3. This indicates that the simulation model has been reconstructed.

5. Conclusions

In this paper, for development of CPS, we modeled and simulated conveyor belt manufacturing system based on M/D/1 queue and decided the occurrence of abnormal situation due to equipment overload at shop-floor using SVM. SVM is trained by using μ, λ, and ρ of M/D/1 queue as input parameters. As a result, it was possible to decide whether the condition was normal or abnormal. For any abnormality, the situation was solved by reconfiguring the manufacturing system. This enabled a flexible system even if an abnormal situation occurred in a CPS-based manufacturing system. Future research will explore ways to use multiple decisions by adding different types of decision making. It is expected that CPS will be useful for further research and development

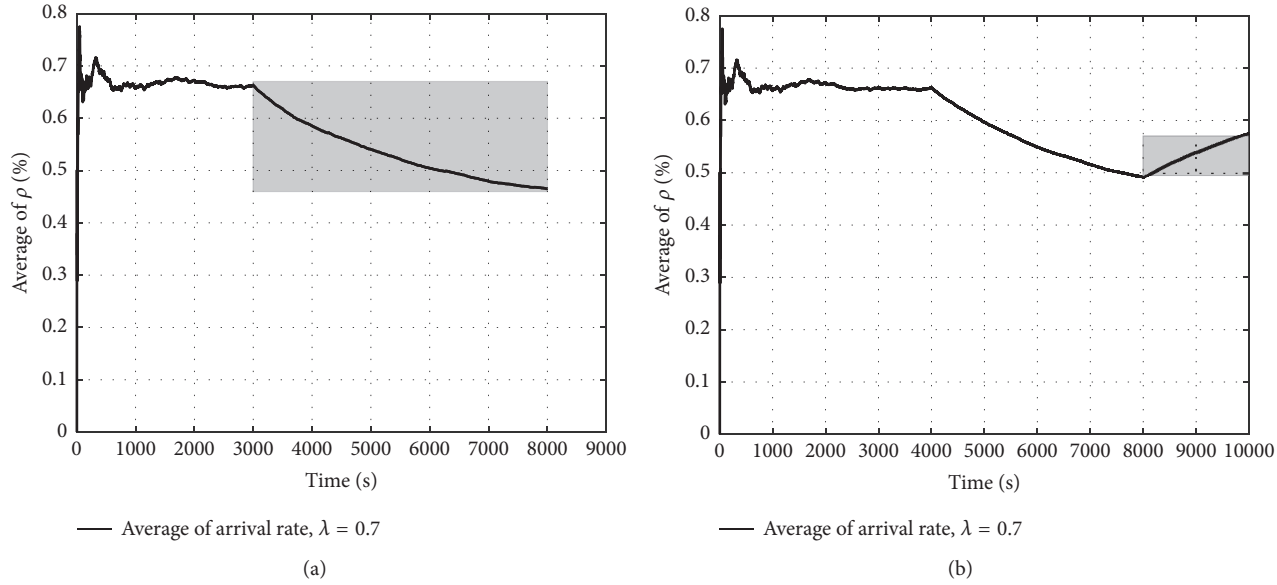


FIGURE 15: The utilization after the change of the production route in the abnormal situation occurs: (a) event occurrence in machine #1-3 (t = 3,000); (b) utilization changed due to abnormal situation recognition (t = 8,000).

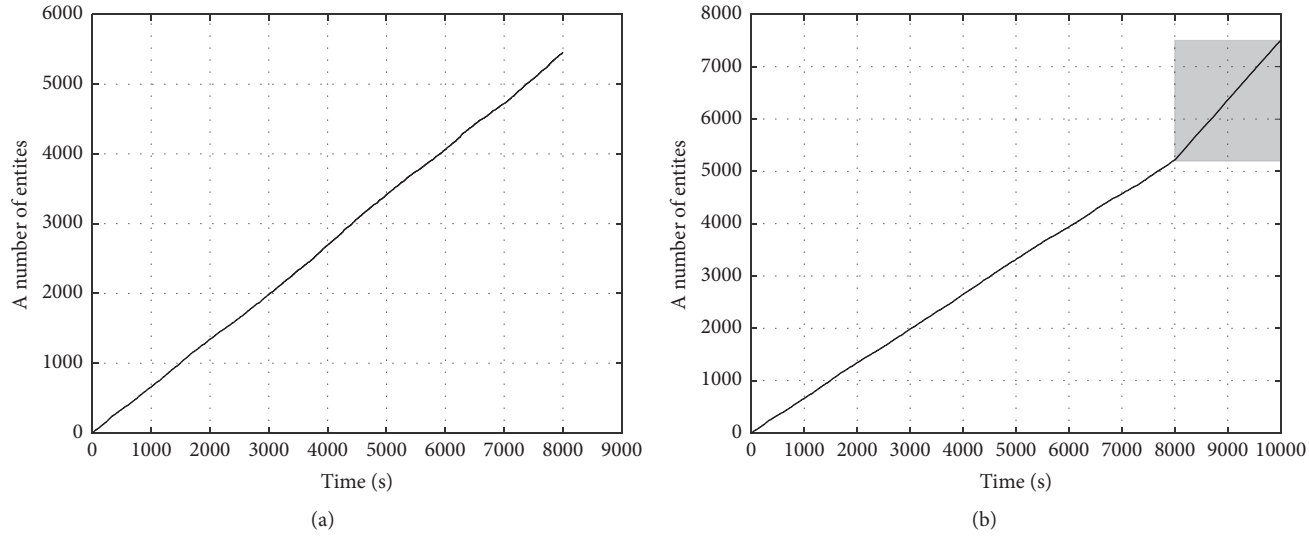


FIGURE 16: A number of entities after the change of the production route after the abnormal situation occurs: (a) event occurrence in machine #1-3; (b) a number of entities changed due to abnormal situation recognition (t = 8,000).

because it is a technology applicable to various fields as well as Industry 4.0 and is indispensable in fields requiring prediction and self-healing.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The present research has been conducted by the Graduate Research Project of KOREATECH in 2016.

References

[1] M. Taisch, B. Stahl, and G. Tavola, "ICT in manufacturing: trends and challenges for 2020—an European view," in *Industrial Informatics (INDIN), Proceedings of the IEEE International Conference on Industrial Informatics*, pp. 941–946, 2012.

[2] J. Park, "Technology and issue on embodiment of smart factory in small-medium manufacturing business," *The Journal of Korean Institute of Communications and Information Sciences*, vol. 40, no. 12, pp. 2491–2502, 2015.

[3] H. Syed, P. Athul, K. Andreas, P. Apostolds, and J. S. Song, "Recent trends in standards related to the internet of things and machine-to-machine communications," *Journal of Information*

and Communication Convergence Engineering, vol. 12, pp. 228–236, 2014.

[4] W. Guo, Y. Zhang, and L. Li, "The integration of CPS, CPSS, and ITS: a focus on data," *Tsinghua Science and Technology*, vol. 20, no. 4, pp. 327–335, 2015.

[5] X. Jin, B. A. Weiss, D. Siegel, J. Lee, and J. Ni, "Present status and future growth of advanced maintenance technology and strategy in US manufacturing," *International Journal of Prognostics and Health Management*, vol. 7, 2016.

[6] M. Qiu and E. H.-M. Sha, "Energy-aware online algorithm to satisfy sampling rates with guaranteed probability for sensor applications," in *Proceedings of the High Performance Computing and Communications*, pp. 156–167.

[7] F. Tao, Y. Zuo, L. D. Xu, and L. Zhang, "IoT-based intelligent perception and access of manufacturing resource toward cloud manufacturing," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1547–1557, 2014.

[8] Q. Jing, A. V. Vasilakos, J. Wan, J. Lu, and D. Qiu, "Security of the internet of things: perspectives and challenges," *Wireless Networks*, vol. 20, no. 8, pp. 2481–2501, 2014.

[9] X. Xu, "From cloud computing to cloud manufacturing," *Robotics and Computer-Integrated Manufacturing*, vol. 28, no. 1, pp. 75–86, 2012.

[10] Q. Liu, J. Wan, and K. Zhou, "Cloud manufacturing service system for industrial-cluster-oriented application," *Journal of Internet Technology*, vol. 15, no. 3, pp. 373–380, 2014.

[11] J. Wan, D. Zhang, Y. Sun, K. Lin, C. Zou, and H. Cai, "VCMIA: a novel architecture for integrating vehicular cyber-physical systems and mobile cloud computing," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 153–160, 2014.

[12] Z. Zhang, H. Lim, and H. J. Lee, "The design of an efficient proxy-based framework for mobile cloud computing," *Journal of Information and Communication Convergence Engineering*, vol. 13, no. 1, pp. 15–20, 2015.

[13] Z. Yang, M. Awasthi, M. Ghosh, and N. Mi, "A fresh perspective on total cost of ownership models for flash storage in datacenters," in *Proceedings of the 8th IEEE International Conference on Cloud Computing Technology and Science*, pp. 245–252, December 2016.

[14] J. Bhimani, Z. Yang, M. Leeser, and N. Mi, "Accelerating big data applications using lightweight virtualization framework on enterprise cloud," in *Proceedings of the IEEE High-Performance Extreme Computing Conference*, pp. 1–7, September 2017.

[15] H. Gao, Z. Yang, J. Bhimani et al., "AutoPath: harnessing parallel execution paths for efficient resource allocation in multi-stage big data frameworks," in *Proceedings of the 26th International Conference on Computer Communication and Networks*, pp. 1–9, July 2017.

[16] F. Soliman and M. A. Youssef, "Internet-based E-commerce and its impact on manufacturing and business operations," *Industrial Management & Data Systems*, vol. 103, no. 8-9, pp. 546–552, 2003.

[17] Gartner's 2015 Hype Cycle for Emerging Technologies Identifies the Computing Innovations that Organizations Should Monitor, http://www.gartner.com/newsroom/id/3114217.

[18] H. S. Kang, J. Y. Lee, S. Choi et al., "Smart manufacturing: past research, present findings, and future directions," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 13, pp. 111–128, 2016.

[19] J. Lee, C. Jin, and B. Bagheri, "Cyber physical systems for predictive production systems," *Production Engineering Research and Development*, vol. 11, no. 2, pp. 155–165, 2017.

[20] B. Dworschak and H. Zaiser, "Competences for cyber-physical systems in manufacturing-first findings and scenarios," *Procedia CIRP*, vol. 25, pp. 345–350, 2014.

[21] L. Monostori, "Cyber-physical production systems: roots, expectations and R&D challenges," *Procedia CIRP*, vol. 17, pp. 9–13, 2014.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] M. D. Odom and R. Sharda, "A neural network model for bankruptcy prediction," in *Proceedings of the International Joint Conference on Neural Prediction Networks (IJCNN '90)*, pp. 163–168, 1990.

[24] J. C. Beard, C. Epstein, and R. D. Chamberlain, "Online automated reliability classification of queueing models for streaming processing using support vector machines," in *Proceedings of the International Conference on Parallel and Distributed Computing*, pp. 325–328, 2015.

[25] C. J. Lee, S. O. Song, and E. S. Yoon, "The monitoring of chemical process using the support vector machine," *Korean Chemical Engineering Research*, vol. 42, pp. 538–544, 2004.

[26] Y. Oh, H. S. Park, A. Yoo et al., "A product quality prediction model using real-time process monitoring in manufacturing supply chain," *Journal of Korean Institute of Industrial Engineers*, vol. 39, no. 4, pp. 271–277, 2013.

[27] S. Neumeyer, K. Exner, S. Kind, H. Hayka, and R. Stark, "Virtual prototyping and validation of cpps within a new software framework," *Computation*, vol. 5, no. 1, article 10, 2017.

[28] B. Bagheri, S. Yang, H.-A. Kao, and J. Lee, "Cyber-physical systems architecture for self-aware machines in Industry 4.0 environment," *International Federation of Automatic Control*, vol. 48, no. 3, pp. 1622–1627, 2015.

[29] M. Marques, C. Agostinho, R. Poler, G. Zacharewicz, and R. Jardim-Goncalves, "An architecture to support responsive production in manufacturing companies," in *Proceedings of the 8th IEEE International Conference on Intelligent Systems*, pp. 40–46, September 2016.

[30] U. N. Bhat, *An Introduction to Queueing Theory: Modeling and Analysis in Applications*, Springer, New York, NY, USA, 2008.

[31] MathWorks, SimEvents User's Guide, https://kr.mathworks.com/help/pdf_doc/simevents/simevents_ug.pdf.

[32] M. A. Gray, "Discrete event simulation: a review of SimEvents," *Computing in Science & Engineering*, vol. 9, pp. 62–66, 2007.

[33] A. A. Alsebae, M. S. Leeson, and R. J. Green, "SimEvents-based modeling and simulation study of stop-and-wait protocol," in *Proceedings of the 5th International Conference on Modelling, Identification and Control*, pp. 239–244, September 2013.

[34] N. Galaske, D. Strang, and R. Anderl, "Response behavior model for process deviations in cyber-physical production systems," in *Proceedings of the World Congress on Engineering and Computer Science*, pp. 443–455, 2015.

[35] S. C. Lee, T. G. Jeon, H. S. Hwang, and C. S. Kim, "Design and implementation of wireless sensor based-monitoring system for smart factory," in *Proceedings of the International Conference on Computational Science and Its Applications*, pp. 584–592, 2007.

[36] J. Jang and E. J. Kim, "Survey on industrial wireless network technologies for smart factory," *Journal of Platform Technology*, vol. 4, pp. 3–10, 2016.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 3926821, 12 pages
https://doi.org/10.1155/2017/3926821

Research Article

Pipeline Implementation of Polyphase PSO for Adaptive Beamforming Algorithm

Shaobing Huang, Li Yu, Fangjian Han, and Yiwen Luo

School of Electronic Science and Engineering, National University of Defense Technology, Changsha, China

Correspondence should be addressed to Li Yu; yuli@nudt.edu.cn

Received 14 March 2017; Accepted 13 September 2017; Published 19 December 2017

Academic Editor: Haiyu Huang

Adaptive beamforming is a powerful technique for anti-interference, where searching and tracking optimal solutions are a great challenge. In this paper, a partial Particle Swarm Optimization (PSO) algorithm is proposed to track the optimal solution of an adaptive beamformer due to its great global searching character. Also, due to its naturally parallel searching capabilities, a novel Field Programmable Gate Arrays (FPGA) pipeline architecture using polyphase filter bank structure is designed. In order to perform computations with large dynamic range and high precision, the proposed implementation algorithm uses an efficient user-defined floating-point arithmetic. In addition, a polyphase architecture is proposed to achieve full pipeline implementation. In the case of PSO with large population, the polyphase architecture can significantly save hardware resources while achieving high performance. Finally, the simulation results are presented by cosimulation with ModelSim and SIMULINK.

1. Introduction

Potential interference has been the major concern for system designers in military and critical civilian wireless communication since it may obscure the original received signal. As we all know, the traditional filters process signals in frequency domain, which are usually incapable of interference cancellation in cases when the interference signals occupy the same frequency band as the desired signal. In this case, if we attempt to suppress high-power interferences, the low-power signals of interest will be eliminated. Adaptive beamforming [1], known as a spatial filtering method, has been a powerful technique to enhance signals of interest while suppressing the interference and the noise signal as a result of the linear combination of the array antenna. Most of the adaptive beamforming algorithms, according to whether the training sequence is used or not, could be divided into two classes [2]: blind adaptive algorithm and nonblind adaptive algorithm. And, in our research, the nonblind algorithms are employed.

LMS [3] approaches may have been the most widely used nonblind adaptive beamforming algorithm in engineering applications due to its robustness and simplicity. However, it exhibits a slow convergence and easily tracks into local optimal solution, which would be a fatal flaw when the digital wireless communication system has a high-performance requirement of real-time implementation.

Particle Swarm Optimization (PSO), which was proposed by Professors Eberhart and Kennedy in 1995 [4], is now one of the most important and widely used swarm intelligence algorithms. Using some simple principles, the PSO algorithms mimic the behavior of birds flocking to guide the swarm particles to search for global optimal solution. Compared to other evolutionary algorithms such as the genetic algorithms [5], simulated annealing algorithms [6], ant colony algorithms [7], and others, the PSO algorithms is much easier to implement and shows great performance in convergence speed and in searching global optimal solutions. Therefore, it has been successfully used in many engineering applications in recent years, including adaptive filters, which can be regarded as real-world optimization problems [8–13].

Similar to other iterative evolutionary computation approaches, the PSO algorithm is also a population-based optimization technique, the main drawback of which is long execution times, specifically when solving large scale complex engineering problems. Therefore, with the advantage of naturally parallel searching capabilities, parallel implementation of the PSO algorithms has been proposed to overcome the problems mentioned above, achieving high performance in

comparison with software solutions [14–17]. However, the PSO algorithm's hardware cost will increase rapidly when its population enlarges, since every increase in swarm size will result in a linear increase in the consumption of hardware resources. This weakness has restricted the use of the PSO algorithm in wide applications of digital signal processing methods.

Recently, advances in Very Large Scale Integration (VLSI) technology have seen significant interest in using Field Programmable Gate Array (FPGA) to speed up scientific and engineering computation with its parallel implementation and configurable hardware technology [18–20]. Taking advantage of powerful designed architecture, such as pipelining and parallel computing, FPGA could achieve much greater processing speed than common software solutions.

FPGA implementation of the PSO algorithm is a feasible and cheap solution because of its parallel high-performance computing and configurable character. Several different parallel architectures have been proposed to implement the PSO algorithm. Most of the previous work dealing with the implementation of the PSO algorithms based on FPGA uses fixed-point arithmetic since the conventional FPGA technology just provides integer and fixed-point arithmetic [21–25]. This approach could reduce the hardware cost in the logic area; however, the simplification is likely to result in resolution degradation because of its small dynamic range. A simple implementation of adaptive filters with the PSO algorithm based on FPGA has been presented in literature [23]. In the anti-interference communication field, especially in military wireless communication, the narrow interference signal's power is usually more than 30 dB higher than the signal of interest which requires a large dynamic range; namely, the algorithm operates over small and large numbers during the PSO execution. In addition, the iterative PSO algorithm needs high precision to offset the effect of update error. Obviously, fixed-point arithmetic could not satisfy these two requirements. Hence, we propose the adaptive beamforming algorithm with PSO using the user-defined floating-point arithmetic which would reduce the loss of precision while decreasing the consumption of hardware resources as much as possible. Although few previous works [18, 26] have implemented the PSO based on floating-point arithmetic, they are still presented using common parallel architectures in which each particle has to use independent hardware units to achieve signal processing. This results in a large consumption of hardware resources and power, which is an adverse issue for digital communication systems.

In this paper, we present a novel pipelined architecture based on FPGA to implement an adaptive beamforming algorithm using PSO based on the minimum mean square error (MSE) criterion. The proposed architecture is based on user-defined floating-point arithmetic [8]. This implementation architecture mainly applies to modern digital anti-interference communication systems in which the baseband chip cycle is much greater than the system clock period. As a consequence, a large time redundancy is generated, of which full use could be made. Essentially, this novel architecture reuses hardware resources meaning that all particles share the same hardware units to evaluate fitness and update position. This hardly makes any difference in achieving high performance of the system because of the large fixed time redundancy. Using digital polyphase filtering signal processing technology could save a large amount of hardware resources and power consumption since essentially only one hardware processing unit i is needed for one particle. In addition, the existing floating-point arithmetic on FPGA designed by XILINX executes a formatting operation after finishing every addition or multiplication operation which would no doubt increase the consumption of resources. Further, the existing floating-point arithmetic uses the IEEE-754 standard, which may not be enough to achieve large dynamic range and high precision. For the two reasons given above, the implementations of adaptive beamforming with the PSO algorithm are based on suitable user-defined floating-point arithmetic.

The remainder of this paper is organized as follows. The model of adaptive beamforming and the PSO algorithm is presented in Section 2. Section 3 describes the related operations covering FPGA implementation of adaptive beamforming with the PSO algorithm. Section 4 provides the entire proposed implementation architectures. The simulation methods and results are given in Section 5. Finally, we present our conclusions in Section 6.

2. Adaptive Beamforming

In a real digital anti-interference communication system, an adaptive beamformer only processes baseband signals rather than the RF (Radio Frequency) signals or IF (Intermediate Frequency) signals. Figure 1 shows the entire simplified adaptive beamforming system based on the Uniform Linear Array (ULA) with N isotropic antennas. The output of the ULA $x(t)$ is given by [10]

$$X(t) = S(t)\, a(\theta_d) + \sum_{i=1}^{L} S_i(t)\, a(\theta_i) + n(t), \tag{1}$$

where $S(t)$ denotes the signal of interest with the Direction of Arrival (DOA) θ_d and $S_i(t)$ denotes the interference signals with the DOA θ_i. $a(\theta_d)$ and $a(\theta_i)$ denote the steering vectors for the signal of interest and interfering signals, respectively. $n(t)$ is the additive white Gaussian noise (AWGN). The RF signals from the ULA will be mixed with the LOF (Local Oscillator Frequency) by the local oscillator and then output the specified IF signals. Signal $\mathbf{x}(n) = [x_0(n), x_1(n), x_2(n), \ldots, x_{N-1}(n)]^T$ is the output of the AD converter, as the input signal for the Digital Downconverter (DDC). The main role of DDC is to transform the IF discrete signals $\mathbf{x}(n)$ down to the complex baseband signal $\mathbf{X}(n)$ (where $\mathbf{X}(n) = [X_0, X_1, X_2, \ldots, X_{N-1}]^T$), which is the input signal for the adaptive beamformer.

Figure 2 shows the working principle of the adaptive beamformer. The aim of adaptive beamforming is to use an a priori desired signal $S_d(t)$ to estimate the signal of interest from the received signal outside of the interference and noise.

As shown in Figure 2, the output of the adaptive beamformer is the linear combination of the weight vectors and the output of DDC. The criterion is to maximize the output in the

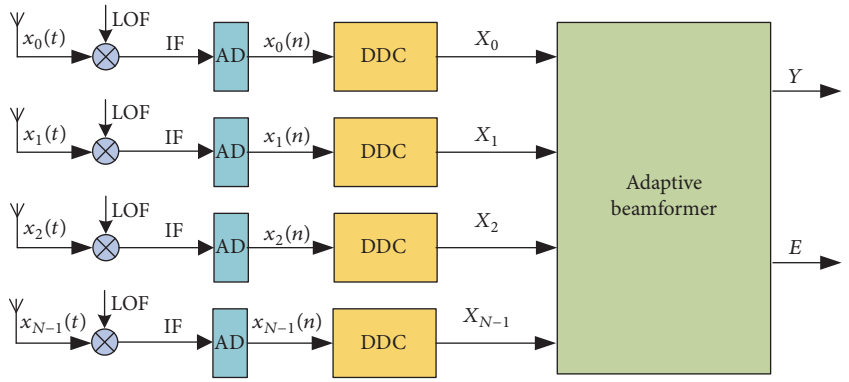


Figure 1: Baseband signal processing by adaptive beamformer.

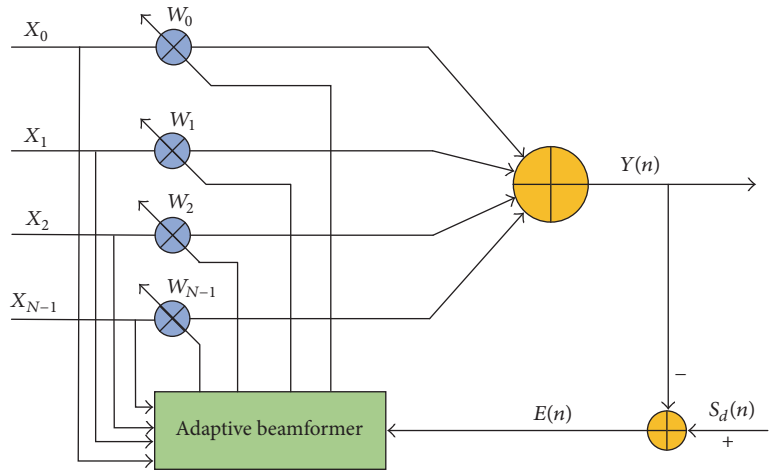


Figure 2: Adaptive beamforming network.

direction of signal of interest and to get null in the direction of the interferences. The weight vectors are updated in each iteration by using the adaptive beamforming algorithm based on the minimum mean square error (MSE) criterion. Therefore, the adaptive beamforming problem can be described as follows: the output of the adaptive beamforming Y is the linear combination of the input signal $\mathbf{X}(n)$ with complex weight vectors $\mathbf{W}(n)$ (where $\mathbf{W}(n) = [W_0, W_1, W_2, \ldots, W_{N-1}]^T$). Then the error signal $E(n)$ is minimized between the desired signal $D(n)$ and the output $Y(n)$. Finally, $E(n)$ is used to update the weight vectors $W(n)$.

As described above, a simple example using LMS based on MSE criterion as the adaptive beamforming algorithm can be expressed as [1]

$$Y(n) = W^H(n) X(n), \tag{2}$$

where

$$W(n) = [w_0(n), w_1(n), \ldots, w_{N-1}(n)]^T,$$
$$X(n) = [x_0(n), x_1(n), \ldots, x_{N-1}(n)]^T \tag{3}$$

where H denotes Hermitian transpose and T denotes transpose. The error signal $E(n)$ is given by

$$E(n) = D(n) - Y(n). \tag{4}$$

And the weight vectors updated equation is presented in the following:

$$W(n+1) = W(n) + \mu X(n) E^*(n), \tag{5}$$

where parameter μ is the correlation of the power spectrum of the input signal, representing the step size which controls the convergence speed.

3. Adaptive Beamforming Based on PSO Algorithm

In this section, an adaptive beamforming algorithm using PSO based on MSE criterion is proposed. Searching the optimal solution for adaptive beamforming can be regarded

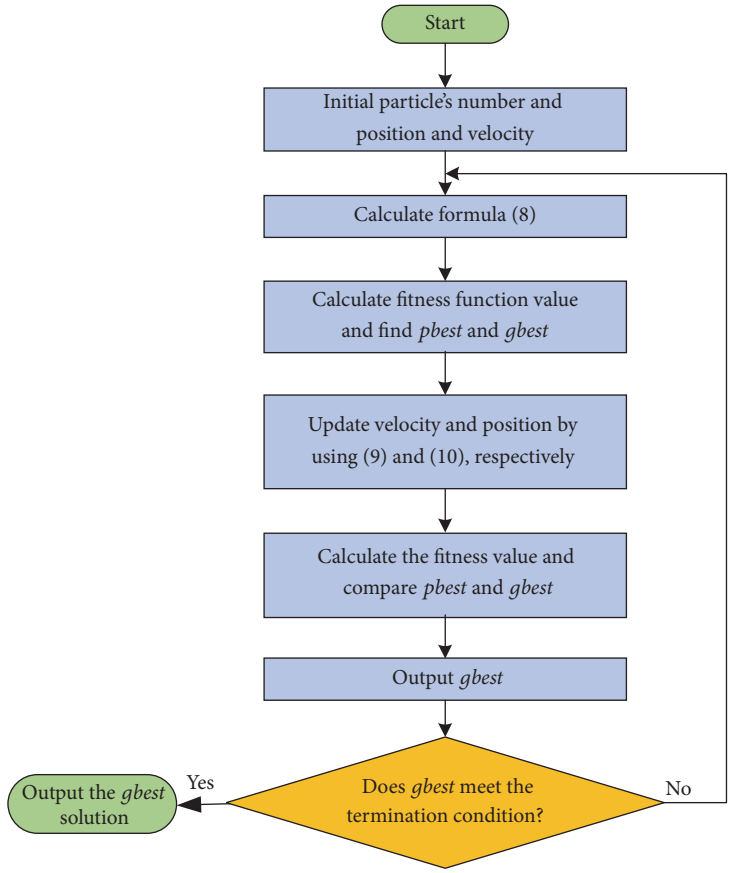


Figure 3: The flowchart of the algorithm.

as a Multiobject Optimization Problem (MOP). And the MOP can be described in the following formula [26, 27]:

$$\begin{aligned} V\text{-min} \quad & f(t) = \left[f_1(t), f_2(t), \ldots, f_{N-1}(t)\right]^T \\ \text{s.t.} \quad & t \in X,\ X \in R^m, \end{aligned} \tag{6}$$

where t is the feasible solution and V-min means the minimization of the functions group. As for the adaptive beamformer, the object is to search for the optimal weight vectors using the given input signals and the desired signals. The weight vectors $W(n)$ can be regarded as a set of the functions $f_i(t)$. Therefore, the criterion is described as follows:

$$\min J = \min \left|E_p(n)\right|^2, \tag{7}$$

where J is the fitness function of PSO and $p = 0, 1, \ldots, N-1$. Considering formulas (2) and (4), (7) is rewritten as

$$\min J = \min \left|D(n) - W_p^H(n) X(n)\right|^2, \tag{8}$$

where $W(n)$ means the position vector in the PSO algorithm.

To solve the MOP model of the adaptive beamforming by PSO, we consider a D-dimensional problem space. The position of the ith particle is expressed as $S_i = (s_{i1}, s_{i2}, \ldots, s_{iD})$, which is represented as a weight vector and the speed of the change of position of X_i is $V_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$. $i = 0, 1, \ldots, N-1$, where N is the population size. In each iteration k, the PSO update equation is expressed as

$$\mathbf{V}^{k+1} = \omega \mathbf{V}^k + c_1 r_1 \left(\mathbf{P}_{pbest}^k - \mathbf{S}^k\right) + c_2 r_2 \left(\mathbf{P}_{gbest}^k - \mathbf{S}^k\right) \tag{9}$$

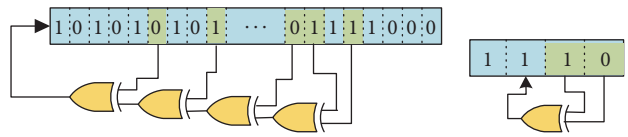
$$\mathbf{S}^{(k+1)} = \mathbf{S}^k + \mathbf{V}^{(k+1)}, \tag{10}$$

where ω is the inertia weight and it mainly plays the role of balancing the local search and global search [14]. c_1 and c_2 represent the acceleration constants, usually both set to 2, which is easy to implement by a shift operation on FPGA. $r_1 = U[0,1]$ and $r_2 = U[0,1]$ are two random numbers ranging from 0 to 1 [8]. $P_{pbesti} = (p_{best1i}, p_{best2i}, \ldots, p_{bestDi})$ represents the individual best position, and $P_{gbest} = (p_{gest1}, p_{gest2}, \ldots, p_{gestD})$ represents the best global position in the search space.

As for the specified optimization problem by the PSO algorithm, the fitness function could be described as (8), in which parameter p is the particles' population. The flowchart of the adaptive beamforming based on the PSO algorithm is given in Figure 3.

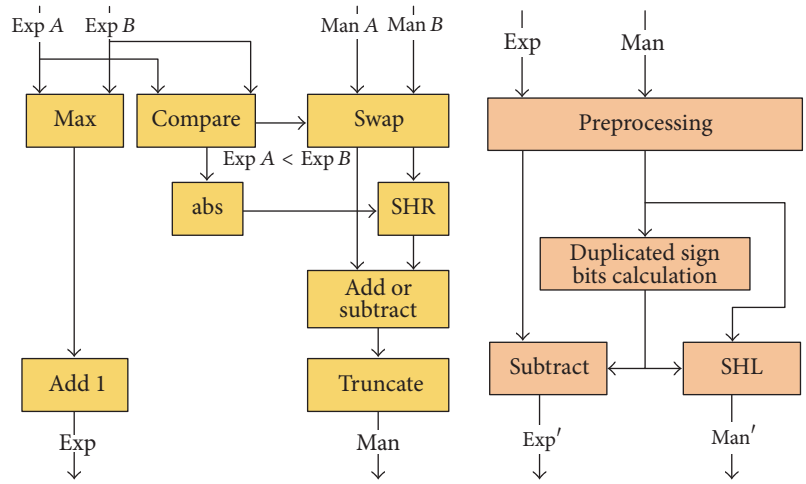
4. Related Operation Based on Floating-Point Arithmetic

The algorithms implemented on FPGA heavily depend on the algorithmic precision. The user-defined floating-point



(a) Mantissa generate (b) Exponent generate

Figure 4: User-defined floating-point pseudorandom number generator.



(a) Basic-adder architecture (b) Formatting operation

Figure 5: User-defined floating-point addition operation.

arithmetic allows the designer to make appropriate use of the bit-width of the floating-point representation according to the balance of logic area consumption and the precision requirement of the algorithm implementation. As stated in (9) and (10), the related operations of the algorithm include multiplication, addition, and random number generation. For the user-defined floating-point data, the multiplication operation is easy to realize by multiplying the IP (Intellectual Property) core provided by XILINX. Therefore, our main work focuses on the pipeline addition operation and random number generation.

4.1. Floating-Point Uniform Random Number Generator. PSO is a stochastic searching algorithm, which is based on several particles randomly moving in a feasible space. In order to compute (9) and (10), where r_1 and r_2 are supposed to be set randomly, we need to use the uniform Random Number Generators (RNGs). The position and velocity of the population in PSO also require the RNGs to generate uniform random initial values. In our proposed scheme, the RNG module is built by the configurable bit-width Linear Feedback Shift Registers (LFSRs), whose input is commonly driven by the feedback XOR (exclusive OR) function of several bits of the overall shift registers. The mantissa is a period of $2^{\text{bit-width}}$. LFSRs on FPGA are operated on fixed-point data. Hence, we could define two LFSRs to generate the mantissa and exponent in floating-point format, respectively. For the sake of simplicity of computation, all signals are power normalized; therefore, as presented in Figure 4, the signed bit of the exponent LFSR is set as 1. That is to say, the generating exponent is always a negative integer. To avoid an integer that is too small, the bit-width of LFSR's exponent is set as 4. And the bit-width of its mantissa is supposed to be configurable enough according to the requirement of precision. In this way, the algorithm avoids the fixed-point-to-float-point conversion.

4.2. User-Defined Floating-Point Pipeline Addition Operation. The floating-point addition operation consists of the sequence of mantissa and exponent operations: shift, swap, round, and format [28, 29].

The floating-point pipeline addition in our proposed implementation architectures consists mainly of two parts: the basic-adder and formatting operation shown in Figure 5, in which SHR and SHL mean shift to the right and left, respectively. The basic adder first compares the exponents of two input operands; then the bigger one is incremented as the exponent of the output sum. At the same time, according to the compared result, it swaps and shifts the mantissa of the smaller number to align the two-incoming numbers. Then the two mantissas are added and the sum is truncated by discarding the lowest bit. The formatting operation first preprocesses the exponent and mantissa of the sum of the basic adder. Then it calculates how many duplicated sign bits

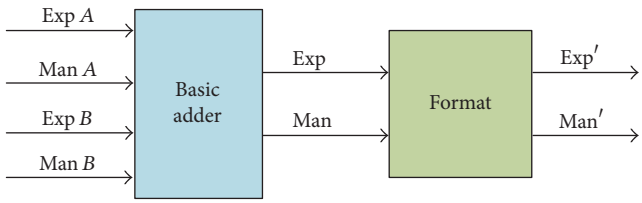


Figure 6: Architecture of two-incoming floating-point adder.

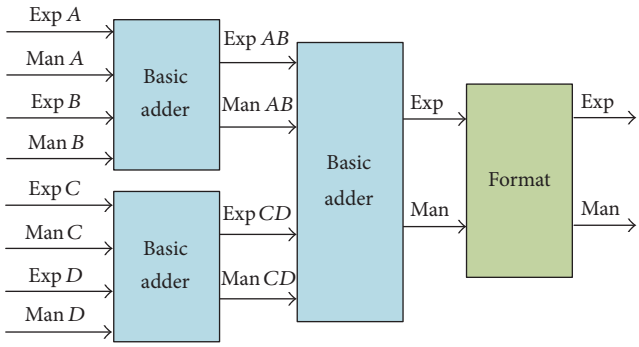


Figure 7: Architecture of four-incoming floating-point adder.

there are and finally outputs the exponent by a subtraction operation and the mantissa by SHL operation according to the number of duplicated sign bits.

A conventional floating-point addition IP core, provided by XILINX, does the formatting operation after every two-incoming addition operation. However, the formatting operation consumes much more hardware resources compared to the basic adder because of the operation of calculating the duplicated sign bits.

In our proposed architecture, we use the eight-incoming floating-point adder in formula (2) and the two-incoming and four-incoming floating-point adder for others. However, the use of the formatting operation should be minimized since it consumes greater resources compared to the floating-point adder based on the standard IEEE-754. Hence the architectures of two-incoming and four-incoming floating-point adders can be implemented in the way shown in Figures 6 and 7, in which it is unnecessary to conduct formatting after every basic-adder operation; instead it conducts formatting after summing all incoming numbers. In this way, the architecture of eight-incoming floating-point adder is presented in Figure 8.

5. Pipeline Polyphase Architecture of PSO

In this section, we first explain what the time redundancy is and then discuss how to use it to achieve a novel pipeline polyphase PSO (PPPSO) architecture for adaptive beamforming. The polyphase term is derived from polyphase filtering, a time-sharing multiplex technology which can make good use of hardware resources units while not affecting the high performance of the algorithm in our proposed

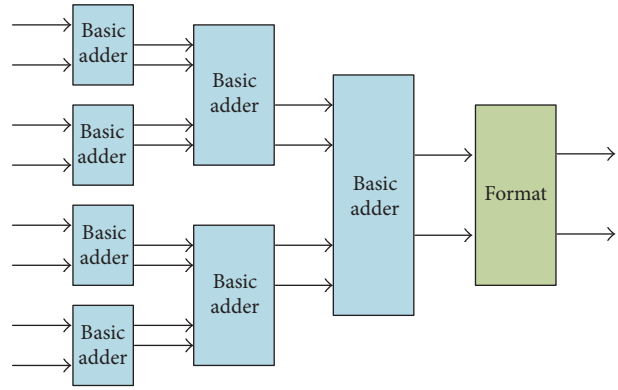


Figure 8: Architecture of eight-incoming floating-point adder.

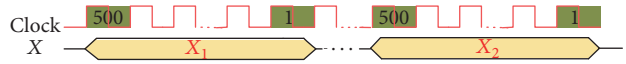


Figure 9: System clock cycle and data chip rate.

architecture. Finally, one particle's whole hardware unit and its main parts are presented.

5.1. Time Redundancy. In modern digital communication system, AD converters sample signals very fast as a consequence of extremely high requirement of data throughput and huge hardware resources consumption if it processes the signals directly after AD converters. In fact, it is not a feasible solution since there are not enough hardware resources, and it is unnecessary as well. In general, the sampling signals will be transformed by DDC and achieve the baseband signals with a low chip rate (e.g., 500 K chip/s). However, the system clock of a 7-series XILINX FPGA can easily achieve 250 M rate which is 500 times faster than the chip rate. An example is shown in Figure 9.

As we can see in Figure 4, every baseband chip continues for 500 system clock cycles, only one of which is needed in a conventional digital signal processing (DSP) scheme based on FPGA. And this leads to a large time redundancy; that is to say, the baseband signal chip is invalid in the other 499 system clock cycles, which is no doubt an enormous waste of hardware resources. Therefore, we propose a pipeline polyphase scheme to make full use of this part of resources. To make a better illustration, we define Time Redundancy Rate (TRR) as follows:

$$\mathrm{TRR} = \left\lfloor \frac{\text{baseband chip cycle}}{\text{system clock cycle}} \right\rfloor, \tag{11}$$

where $\lfloor \cdot \rfloor$ means rounding down to the nearest integer.

One of the most important characteristics of PSO is that all particles in the same population are independent of the optimal solution (exchanging information only by *Pgbest*). Therefore, only one hardware unit is shared by all particles to evaluate fitness value and update positions. This greatly reduces the use of hardware resources. Undoubtedly,

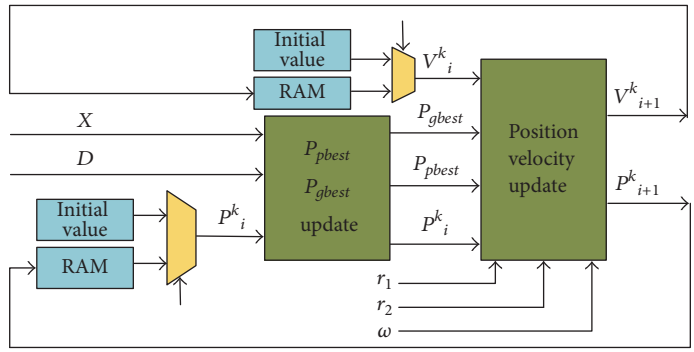


Figure 10: Whole architecture of PPPSO.

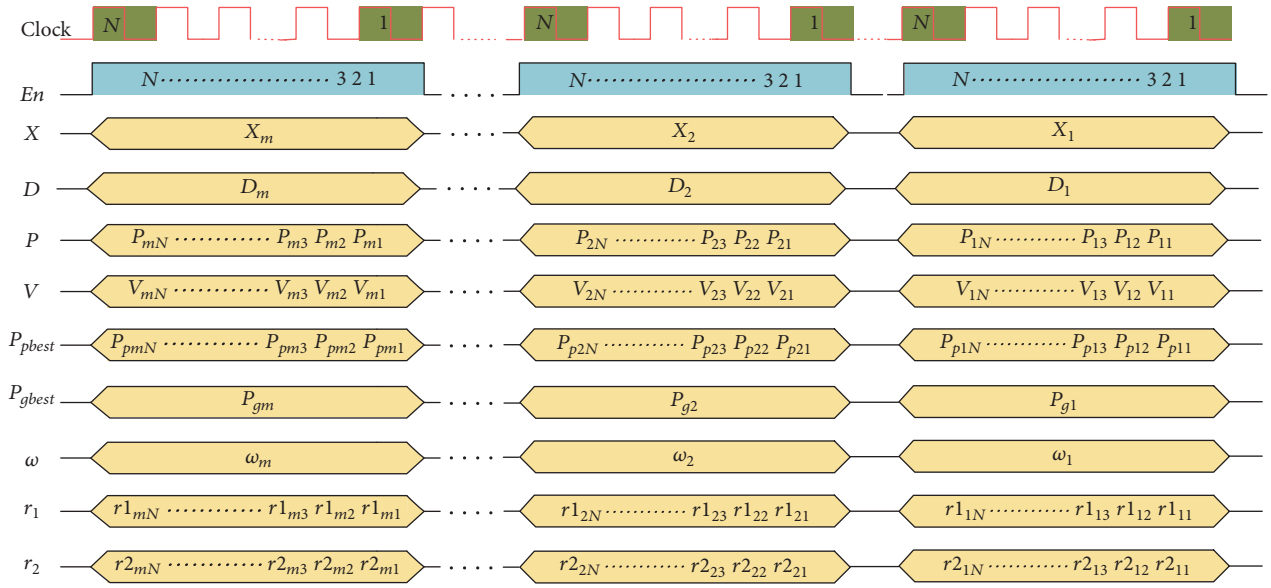


Figure 11: Timing diagram of critical signals.

the greater the TRR is, the larger the population of the PSO algorithm can be set, and higher performance can be achieved, theoretically.

5.2. Adaptive Beamformer with PPPSO Algorithm. The whole architecture of the adaptive beamformer based on the PPPSO algorithm is presented in Figure 10. It consists of three parts: (1) *Ppbest* and *Pgbest* updating module: the individual best and global best values update or not according to the evaluation of the fitness function value; (2) position and velocity update module: the swarm particle updates according to formulas (9) and (10); formula (3) signals storage module: it mainly makes use of Random Access Memory (RAM).

As depicted in Figure 10, the *Ppbest* and *Pgbest* updating module receives the input signals (shown as X) and the desired signals (shown as D), then calculates fitness values according to the fitness evaluation function, and finally updates the values of individual best and global best. The individual and global best, together with r_1, r_2, and ω, apply to the position and velocity updating module to accomplish the updating process. Finally, our proposed pipeline polyphase architecture requires storage of all critical coefficients, including position, velocity, *Ppbest*, and *Pgbest*.

r_1 and r_2 will be generated at each system clock cycle by RNG function mentioned above. As for the inertia coefficient ω, the proposed PPPSO architecture adopts the suggestion from [14], setting it as a dynamic function of iteration index, given by

$$\omega^k = \omega_{\max} - k \times \frac{\omega_{\max} - \omega_{\min}}{K}, \tag{12}$$

where $\omega_{\max}$ and $\omega_{\min}$ represent the maximum and minimum value of ω, respectively; k is the current iteration index of the PSO algorithm; and K is the maximum iteration index when the iterative process ends.

The timing diagram of the critical signals in our proposed PPPSO architecture is presented in Figure 11. As depicted in Figure 11, a polyphase period has an N (population size of the algorithm) system clock cycle, in which the PPPSO algorithm finishes one iteration.

That is to say, each particle will independently (they only exchange searching information at the end of a polyphase

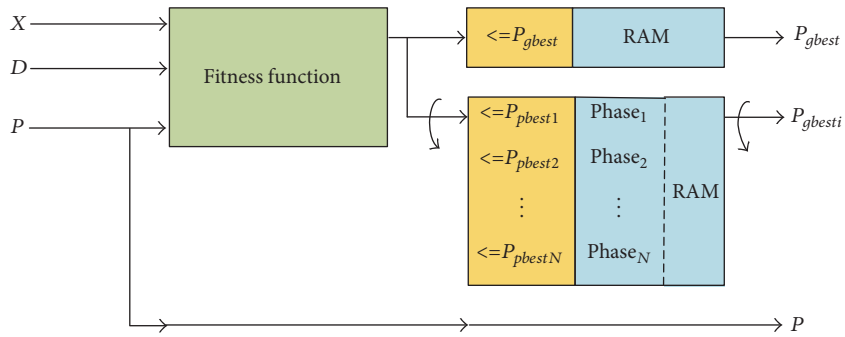


Figure 12: The *Ppbest* and *Pgbest* update module.

period by the global best) finish its individual best update in a system clock cycle, benefited from the pipeline polyphase signal processing technique. In a polyphase period, the input signals, desired signals, and the inertia ω remain unchanged for all particles. Take position (shown as P in Figure 11) as an example to show how the pipeline architecture works. Pij represents the ith data of the jth ($1 \leq j \leq N$) phase data channel which means jth particle's position value. Therefore, in a whole polyphase period, every particle would receive the same X and D as the input of the whole architecture to finish the update process. Since it is a pipeline process, every particle in a specified phase channel could share the same hardware units to achieve its own update using the previous position's value in the same phase channel and they do not affect each other. In this way, when one polyphase period finishes, each particle will have finished searching its own individual best value and finished searching the optimal solution.

5.3. The Individual Best and Global Best Update Module. The individual best and global best update module (depicted as Figure 12) is another critical step of the whole pipeline polyphase architecture. It contains the fitness function to evaluate the fitness value of every particle's position. The individual best and global best are considered to update or not according to computed value of fitness function.

As shown in Figure 12, the fitness function value is calculated using X, D and P as incoming data. Then the individual and global best will update or not according to the new evaluated value of the fitness function. The global best just updates one time at the end of the polyphase period according to the compared result of the global best position's fitness value of the current and previous iteration. However, it must compare the corresponding evaluated value at every different specified phase channel, which is the method to update the individual best. Hence each individual best of the particles will be stored in RAM in order, as shown in Figure 12, to achieve the comparison of each particle's individual best position of the current and last iteration.

In our proposed adaptive beamformer with the PPPSO algorithm, a four-antenna simple ULA is applied. Hence, each particle has four dimensions.

The fitness function uses the MSE criterion to minimize the error value as stated in formula (8), in which W denotes the position of the particle in the PPPSO algorithm. It is

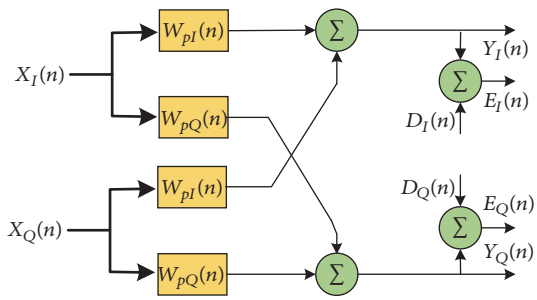


Figure 13: Method to calculate the complex error.

noted that the adaptive beamformer with the PPPSO is a complex-based algorithm so that all baseband signals are complex-based. And the complex error is calculated as shown in Figure 13.

5.4. Particle Update Equation. The position and velocity update module is shown in Figure 14. As stated in formulas (9) and (10), the updating process of each particle in each dimension requires five additions (or subtractions), three multiplications, and two uniform RNGs.

As depicted in Figure 14, all operations in the hardware units of position and velocity updating module work in a full-parallel pipeline way. These operation hardware units need to work together in every system clock cycle because of the pipeline requirement. However, our scheme makes all particles share just one particle updating module, which makes good use of the pipeline polyphase implementation.

6. Simulation Results and Analysis

The proposed architectures for an adaptive beamformer based on the PPPSO algorithm have been developed in hardware description language using Verilog HDL and VHDL (Very High Speed Integrated Circuits Hardware Description Language). All the architectures are synthesizable in the XILINX ISE 14.7 tool and are based on the parameterizable floating-point packages with user-defined bit-width. Our proposed architecture mainly aims to the PPPSO algorithm with a large scale population (more than 64 in size). As mentioned above, the TRR is easy to achieve 500 in XILINX

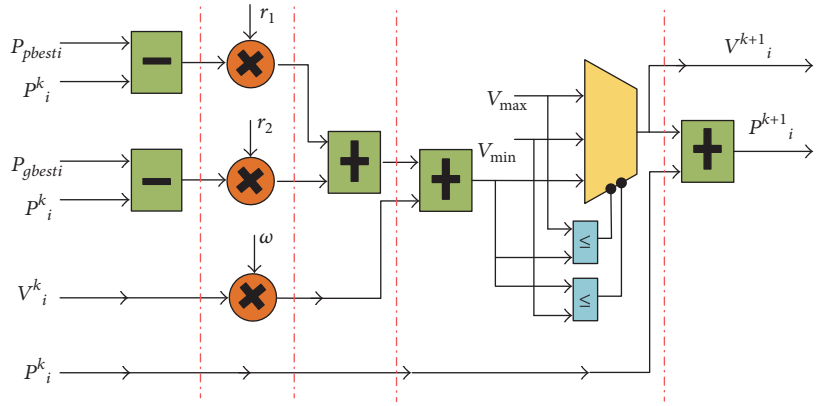


Figure 14: Position and velocity update module.

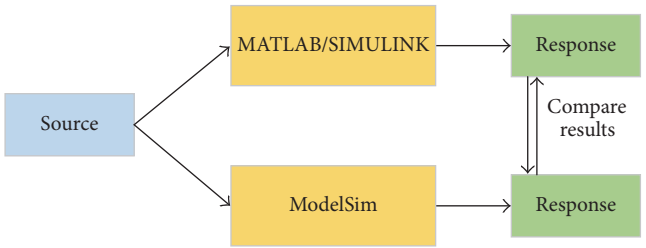


Figure 15: Cosimulation with ModelSim and MATLAB/SIMULINK.

7-serias devices. Hence, the population can reach a scale of 500 in size theoretically.

Mentor Graphics ModelSim is the most conventional HDL simulator to validate the timing of signals in the whole architecture. However, it is complicated if only ModelSim is used to validate the results of the whole adaptive beamforming system. Hence, for convenience and simplification, a cosimulation technique by ModelSim and MATLAB/SIMULINK with HDL Verifier is applied to verify the simulated results. HDL verifier automates Verilog and VHDL design verification and analyzes its response, providing interfaces to link MATLAB/SIMULINK with ModelSim. In this way, we are able to compare the complete calculated results from ModelSim and the theory results from MATLAB to verify the responses. The cosimulation schematic diagram is depicted in Figure 15.

We consider a ULA with four elements for simulating the real situation. The SNR (Signal-to-Noise Ratio) is set as 1 dB and the ISR (Interference Signal Ratio) is 30 dB. The azimuth (AZ) of the signal and interference are set as 0° and 60°, respectively. The desired signals and interference signals are composed of PN sequences and sine function, respectively. The system supposes that the ULA receives signals including AWGN and horizontal narrow-band interference so that the pitch angle is 90°. These parameters are shown in Table 1.

As for the initial parameters, Huang et al. [8] suggested a fixed value of 2.0 for both acceleration coefficients. The inertia weight is set to be ranging from 0.9 to 0.4, as a linear function

Table 1: Simulation parameter.

N	SNR/dB	ISR/dB	Signal AZ/°	Interference AZ/°
4	1	30	0	60

Table 2: Synthesis results for the PPPSO with different population size.

Population size	FF	LUTs	DSP48	RAM
128	45840 (5.29%)	51948 (11.99%)	138 (3.88%)	25 (1.7%)
192	45840 (5.29%)	51976 (12.00%)	138 (3.88%)	25 (1.7%)
256	45840 (5.29%)	51985 (12.00%)	138 (3.88%)	25 (1.7%)
320	45840 (5.29%)	52009 (12.01%)	138 (3.88%)	25 (1.7%)

as stated in formula (11). The maximum and minimum of the velocity are 0.125 and −0.125, respectively.

6.1. Synthesis Results. The synthesis results for the FPGA implementation, based on a double precision PPPSO architecture, are presented in Table 2. The hardware resource consumption is reported in terms of FF (Flip-Flops), LUTs (Look-Up Tables), DSP blocks, and RAM memory for a XILINX Virtex7 xc7vx690tffg1926. As depicted in Table 2, it can be observed that the implementation architecture requires around 5.29% FF, 12% LUTs, 3.88% DSP48, and 1.7% RAM of available resources for all different polyphase (population size) implementations. No matter how large the population size is, the cost in logic area does not increase very much (the small increment for LUTs for the signals register can be ignored). The proposed architecture is effectively implemented in the hardware when a large scale population is required. Table 3 depicts the synthesis results for architectures based on user-defined floating-point arithmetic with various bit-width when the polyphase number (population size) is 128.

Table 3: Synthesis results for architectures based on user-defined floating-point arithmetic with various bit-width.

Bit-width (man + exp)	LUTs	FF
32 + 8	32481 (7.50%)	31193 (3.60%)
36 + 8	36836 (8.50%)	33489 (3.87%)
40 + 8	40698 (9.39%)	35889 (4.14%)
48 + 8	46654 (10.77%)	40538 (4.68%)
52 + 12	45840 (11.99%)	51948 (5.29%)

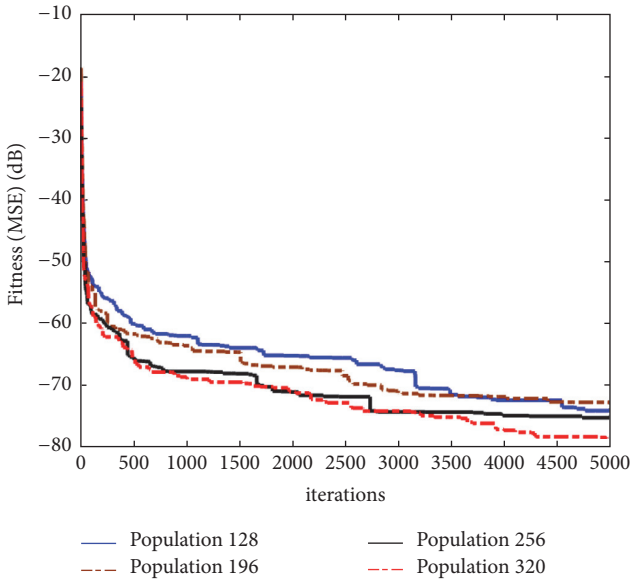


Figure 16: Fitness function (MSE) performance.

The hardware resources cost of DSP48 and RAM is unchangeable because of the fixed use of multipliers and RAM. As shown in Table 3, the cost of LUTs and FFs is gradually decreased with shorter bit-widths of mantissa and exponent. Taking into consideration Table 3, designers have the option to balance the hardware unit consumption and performance of precision. We suggest that the algorithm should use much shorter bit-widths of mantissa and exponent while not doing so affects convergence of the algorithms. From our simulation results, a value of 36 for bit-width of mantissa and a value of 8 for bit-width of exponent would already satisfy the requirement for the precision.

6.2. Simulation Results. As mentioned above, it is convenient to verify the simulation results using cosimulation technology with ModelSim and SIMULINK as shown in Figure 15. All simulation results are based on user-defined floating-point arithmetic with a value of 36 for bit-width of mantissa and a value of 8 for bit-width of exponent.

Figure 16 depicts the results of the MSE performance of the PPPSO algorithm with different sizes of population (128, 196, 256, and 320, resp.). A 10-ensemble Monte Carlo Method is applied to our simulation with different initial values for all swarm particles. It can be observed in Figure 16 that MSE learning curves of the PPPSO are very steep since the number of swarm particles is very large. Although the MSE learning curves shown in Figure 16 are closely convergent, the convergent speed of the algorithm with 256-size and 320-size populations is obviously greater than it is with 128-size and 192-size populations.

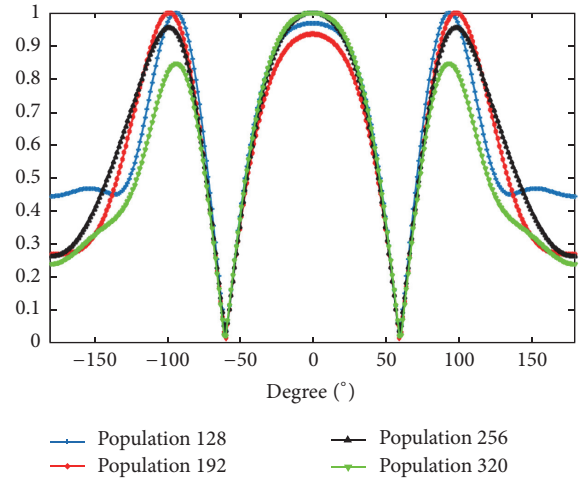


Figure 17: ULA amplitude pattern for PPPSO with different population size.

Figure 17 shows the amplitude pattern of the PPPSO algorithm with different population sizes by using global best position, in the situation that signals are amid interferer and AWGN with a SIR and SNR values mentioned in Table 1. The algorithm in all of these situations can achieve a great performance to null the signals from 60° direction, namely, the interferer's direction, and achieve a high gain for signals at 0° direction, namely, the interested signal's direction. In general, they all are able to achieve a wider main lobe and can null the signal at direction of the interferer while attempting to achieve maximum reception in the specified direction of desired signal.

7. Conclusions

This paper describes a pipeline polyphase PSO architecture implementation on FPGA for an adaptive beamformer, using the efficient user-defined floating-point arithmetic. The user-defined floating-point arithmetic can perform computations with a large dynamic range and suitable precision while saving hardware resources consumption for the digital anti-interference communication application. The major advantage of our proposed architecture is to allow the use of the PSO algorithm with a large scale population by polyphase signal processing technology. In order to use polyphase architectures to implement the proposed algorithm rather than a full-parallel architecture, a pipeline hardware architecture of one swarm particle's processing unit is required, in which the hardware processing unit could be shared by all other swarm particles, with the consequence of saving a large cost of logic area.

Synthesis results demonstrate that using FPGA to implement the adaptive beamformer based on the PSO algorithm is an entirely acceptable solution. Moreover, the proposed architecture allows the designers to explore the balance of precision and performance by using the user-defined floating-point arithmetic.

In order to simplify the simulation process, the cosimulation technique with ModelSim and SIMULINK is applied to validate the results of the whole adaptive beamforming system with a four-antenna ULA. The PPPSO architectures with various large scale populations are simulated. The MSE learning curve and amplitude pattern are applied to measure performance. The simulation results demonstrate that it is efficient to implement the PPPSO algorithm with large scale populations.

In the future, we intend to explore the balance for exactly suitable precision requirement and the hardware logic area. Furthermore, a complicated time-varying situation is also supposed to take more real scenario into account.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] S. Haykin, "Adaptive Filter Theory," in *Person Education*, pp. 83–87, Asia, 4th edition, 2002.

[2] S. Hossain, M. T. Islam, and S. Serikawal, "Adaptive beamforming algorithms for smart antenna systems," in *Proceedings of the 2008 International Conference on Control, Automation and Systems, ICCAS 2008*, pp. 412–416, Republic of Korea, October 2008.

[3] A. Senapati, K. Ghatak, and J. S. Roy, "A comparative study of adaptive beamforming techniques in smart antenna using LMS algorithm and its variants," in *Proceedings of the 1st International Conference on Computational Intelligence and Networks, CINE 2015*, pp. 58–62, India, January 2015.

[4] R. C. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Proceedings of the 6th International Symposium on Micro Machine and Human Science (MHS '95)*, pp. 39–43, Nagoya, Japan, October 1995.

[5] D. Beasley, R. R. Martin, and D. R. Bull, "An overview of genetic algorithms," in *Part1. Fundamentals , University computing*, pp. 58-58, An overview of genetic algorithms, Part1. Fundamentals, 1993.

[6] R. A. Rutenbar, "Simulated annealing algorithms: an overview," *IEEE Circuits and Devices Magazine*, vol. 5, no. 1, pp. 19–26, 1989.

[7] G. Bilchev and I. C. Parmee, "The ant colony metaphor for searching continuous design spaces," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 993, pp. 25–39, 1995.

[8] S. Huang, L. Yu, F.-J. Han, and W. Ding, "Adaptive beamforming algorithm for interference suppression based on partition PSO," in *Proceedings of the 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference, IEEE IEMCON 2016*, Canada, October 2016.

[9] D. J. Krusienski and W. K. Jenkins, "A particle swarm optimization - Least mean squares algorithm for adaptive filtering," in *Proceedings of the Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, pp. 241–245, November 2004.

[10] U. Mahbub, C. Shahnaz, and S. A. Fattah, "An adaptive noise cancellation scheme using particle swarm optimization algorithm," in *Proceedings of the 2010 IEEE International Conference on Communication Control and Computing Technologies, ICCCCT 2010*, pp. 683–686, India, October 2010.

[11] Z. Zhao, S. Xu, S. Zheng, and J. Shang, "Cognitive radio adaptation using particle swarm optimization," *Wireless Communications and Mobile Computing*, vol. 9, no. 7, pp. 875–881, 2009.

[12] J. J. Liang, A. K. Qin, P. N. Suganthan, and S. Baskar, "Comprehensive learning particle swarm optimizer for global optimization of multimodal functions," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 3, pp. 281–295, 2006.

[13] Q. Qi, J. Wang, Q. Li, T. Li, and Y. Cao, "Resource orchestration for multi-Task application in home-To-home cloud," *IEEE Transactions on Consumer Electronics*, vol. 62, no. 2, pp. 191–199, 2016.

[14] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *Proceedings of the IEEE International Conference on Evolutionary Computation and IEEE World Congress on Computational Intelligence*, (Cat. No.98TH8360), pp. 69–73, Anchorage, Alaska, USA, May 1998.

[15] C.-J. Lin and H.-M. Tsai, "FPGA implementation of a wavelet neural network with particle swarm optimization learning," *Mathematical and Computer Modelling*, vol. 47, no. 9-10, pp. 982–996, 2008.

[16] S. Mehmood, S. Cagnoni, M. Mordonini, and M. Farooq, "Particle swarm optimisation as a hardware-oriented metaheuristic for image analysis," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 5484, pp. 369–374, 2009.

[17] B.-I. Koh, A. D. George, R. T. Haftka, and B. J. Fregly, "Parallel asynchronous particle swarm optimization," *International Journal for Numerical Methods in Engineering*, vol. 67, no. 4, pp. 578–595, 2006.

[18] D. M. Muñoz, C. H. Llanos, L. Dos S. Coelho, and M. Ayala-Rincón, "Comparison between two FPGA implementations of the particle swarm optimization algorithm for high-performance embedded applications," in *Proceedings of the 2010 IEEE 5th International Conference on Bio-Inspired Computing: Theories and Applications, BIC-TA 2010*, pp. 1637–1645, China, September 2010.

[19] N. K. Quang, N. T. Hieu, and Q. P. Ha, "FPGA-based sensorless PMSM speed control using reduced-order extended Kalman filters," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 12, pp. 6574–6582, 2014.

[20] A. Cilardo, "New techniques and tools for application-dependent testing of FPGA-based components," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 1, pp. 94–103, 2015.

[21] H. Guo, H. Chen, F. Xu, F. Wang, and G. Lu, "Implementation of EKF for vehicle velocities estimation on FPGA," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 9, pp. 3823–3835, 2013.

[22] G. Kókai, T. Christ, and H. H. Frhauf, "Using hardware-based particle swarm method for dynamic optimization of adaptive array antennas," in *Proceedings of the 1st NASA/ESA Conference on Adaptive Hardware and Systems, AHS 2006*, pp. 51–58, Turkey, June 2006.

[23] Z. Gao, X. Zeng, J. Wang, and J. Liu, "FPGA implementation of adaptive IIR filters with particle swarm optimization algorithm," in *Proceedings of the 2008 11th IEEE Singapore International Conference on Communication Systems, ICCS 2008*, pp. 1364–1367, China, November 2008.

[24] P. Reynolds, R. Duren, M. Trumbo, and R. Marks, "FPGA implementation of particle swarm optimization for inversion of large neural networks," in *Proceedings of the 2005 IEEE Swarm Intelligence Symposium, 2005. SIS 2005.*, pp. 389–392, Pasadena, CA, USA.

[25] X. Cai, S. Ngah, H. Zhu, Y. Tanabe, and T. Baba, "Pipeline architecture of particle swarm optimization," in *Proceedings of the 9th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2010*, pp. 3–8, Japan, August 2010.

[26] H. Tamaki, H. Kita, and S. Kobayashi, "Multi-objective optimization by genetic algorithms: a review," in *Proceedings of the 1996 IEEE International Conference on Evolutionary Computation, ICEC'96*, pp. 517–522, May 1996.

[27] K. Tang, Z. Li, L. Luo, and B. Liu, "Multi-strategy adaptive particle swarm optimization for numerical optimization," *Engineering Applications of Artificial Intelligence*, vol. 37, pp. 9–19, 2015.

[28] S. R. Vangal, Y. V. Hoskote, N. Y. Borkar, and A. Alvandpour, "A 6.2-GFlops floating-point multiply-accumulator with conditional normalization," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 10, pp. 2314–2322, 2006.

[29] A. Beaumont-Smith, N. Burgess, S. Lefrere, and C. C. Lim, "Reduced latency IEEE floating-point standard adder architectures," in *Proceedings of the 14th IEEE Symposium on Computer Arithmetic, ARITH-14*, pp. 35–42, April 1999.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 7616825, 11 pages
https://doi.org/10.1155/2017/7616825

Research Article

5G MIMO Conformal Microstrip Antenna Design

Qian Wang,[1] Ning Mu,[1] LingLi Wang,[1] Safieddin Safavi-Naeini,[2] and JingPing Liu[1]

[1]*School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China*
[2]*Electrical and Computer Engineering Department, University of Waterloo, Waterloo, ON, Canada N2L 3G1*

Correspondence should be addressed to JingPing Liu; liujingpin2002@aliyun.com

Received 14 June 2017; Revised 29 October 2017; Accepted 23 November 2017; Published 17 December 2017

Academic Editor: Pai-Yen Chen

With the development of wireless communication technology, 5G will develop into a new generation of wireless mobile communication systems. MIMO (multiple-input multiple-output) technology is expected to be one of the key technologies in the field of 5G wireless communications. In this paper, 4 pairs of microstrip MIMO conformal antennas of 35 GHz have been designed. Eight-element microstrip Taylor antenna array with series-feeding not only achieves the deviation of the main lobe of the pattern but also increases the bandwidth of the antenna array and reduces sidelobe. MIMO antennas have been fabricated and measured. Measurement results match the simulation results well. The return loss of the antenna at 35 GHz is better than 20 dB, the first sidelobe level is −16 dB, and the angle between the main lobe and the plane of array is 60°.

1. Introduction

Multiple-input multiple-output (MIMO) technology is originated from wireless communication antenna diversity technology and intelligent antenna technology. It is a combination of multiple-input single-output (MISO) and single-input multiple-output (SIMO) and therefore has the advantages and characteristics of the two [1, 2]. The MIMO system is equipped with multiple antennas at the transmitter and the receiver. It can improve the quality of wireless communication and the rate of data exponentially without increasing the bandwidth and transmitted power [3, 4]. Multiantenna system is an important part of MIMO technology. MIMO wireless system is not only affected by the multipath characteristics of the wireless communication channel but also depends on the design and layout of the multiantenna system. The research of MIMO multiantenna design mainly includes the form of antenna element, the layout of multiple antennas, and the mutual coupling analysis. At present, the research of MIMO multiantenna is focused on the exploration of low cost and high performance designs of antenna and layout [5–7].

Antennas are usually placed on the surface of the carrier to achieve the desired electromagnetic performance. To this end, the conformal antenna was designed [8]. Conformal antenna can be designed on the surface of the carrier, which will not damage the mechanical structure of the carrier and can save space [9–11]. It can be placed anywhere on the surface of the carrier. Conformal antennas are usually microstrip antenna, stripline antenna, or crack antenna. The microstrip antenna has many advantages such as low profile, small size, light weight, and ease to integrate with other carriers. It is therefore more suitable for conformal antenna [12, 13]. In addition, the millimeter-wave band has attracted a lot of attention due to the advent of 5G technology and its inherent characteristics, such as short wavelength, wide frequency band, and propagation characteristics in the fog, snow, and dust environment [14–16]. Therefore, there has been extensive research on millimeter-wave microstrip antenna.

This paper presents the design of a MIMO conformal antenna for 5G. The frequency is 35 GHz, the carrier of conformal is a cylinder, and the angle between the main lobe of pattern and the carrier axis is 60°. The sidelobe characteristics of the antenna significantly affect the interference of the system and the suppression of the clutter. The antenna designed in this paper requires the first sidelobe level to be about −18 dB. In view of this characteristic, a series-fed standing wave antenna array with Taylor distribution is designed. Considering the influence of coupling, 4 pairs of antennas are designed. The results of the research are well suited for the 5G MIMO communication.

2. 5G MIMO Conformal Antenna Design

2.1. Radiation Elements Design. The first part of the MIMO conformal antenna design is the radiation elements. The design uses microstrip patch antenna as radiation elements. There are two main steps in the rectangular microstrip antennas design. The first step is theoretical analysis, and the second step is simulation and optimization.

Firstly, we choose dielectric substrate. For microstrip circuit, the loss of the microstrip is very large in the millimeter-wave band. The loss can be divided into dielectric loss, conductor loss, and radiation loss [17]. Substrates with low loss tangent dielectric are usually chosen to reduce the dielectric loss. When the dielectric constant is low, the total loss of the microstrip would not change with the characteristic impedance. On the contrary, when the substrate has high dielectric constant, the loss of the microstrip will change rapidly with the characteristic impedance. Thicker substrate will increase radiation losses and the surface wave is more serious. A smaller height is more effective in suppressing the higher mode and reducing the radiation loss. Additionally, the thinner substrate with good flexibility is good for conformal antenna [18]. Taking these factors into consideration, we use RT/duroid5880 ($\varepsilon_r = 2.2, \tan\delta = 0.0009$) as substrate, and the height of the dielectric substrate is 0.5 mm.

Next the width W of the patch elements is determined. Directivity factor of microstrip antenna, radiation resistance, and other characteristics will vary with the change of W. These characteristics directly affect the frequency bandwidth and radiation efficiency of the antenna. In order to get the desired frequency bandwidth and radiation efficiency, the choice of W is particularly important. The size of the width should meet the following requirement [18]:

$$W \leq \frac{c}{2f_r}\left(\frac{\varepsilon_r + 1}{2}\right)^{-1/2}, \tag{1}$$

where c is speed of light and f_r is the resonant frequency. In this design, f_r is 35 GHz.

The next step is to determine the length L of the patch elements. The size of unit length L is determined by the effective dielectric constant and the operating frequency. Effective dielectric constant of substrate is defined as ε_e, which is given by the following [18]:

$$\varepsilon_e = \frac{1}{2}\left[\varepsilon_r + 1 + (\varepsilon_r - 1)\left(1 + \frac{12h}{W}\right)^{-1/2}\right]. \tag{2}$$

The length L of the rectangular microstrip patch antenna is approximately $\lambda_g/2$ and is given by the following [18]:

$$\Delta l = 0.412\frac{(\varepsilon_e + 0.3)(W/h + 0.264)}{(\varepsilon_e - 0.258)(W/h + 0.8)}h$$
$$L = \frac{c}{2f_r\sqrt{\varepsilon_e}} - 2\Delta l. \tag{3}$$

From (1)–(3), W is chosen to be 3.38 mm and L is 2.8 mm.

There are three main methods to feed the rectangular microstrip patch. Microstrip line feeding is usually used to design array elements, coaxial feeding is usually used for single microstrip antenna, and the electromagnetic coupling feeding is usually used in the microstrip antenna of double structure [18]. Microstrip patch element designed here is a radiation element in the antenna array, so microstrip is chosen as feeding method.

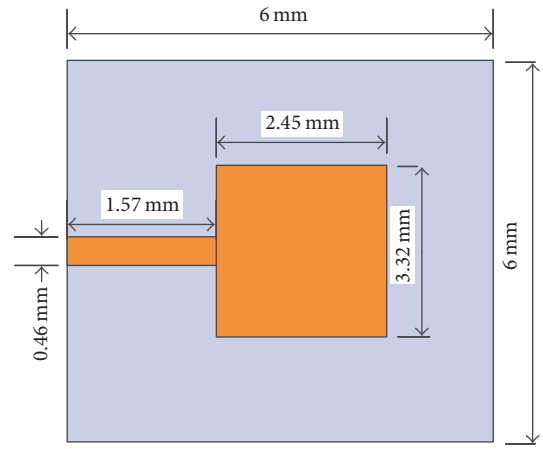


Figure 1: The model of rectangular microstrip patch antenna.

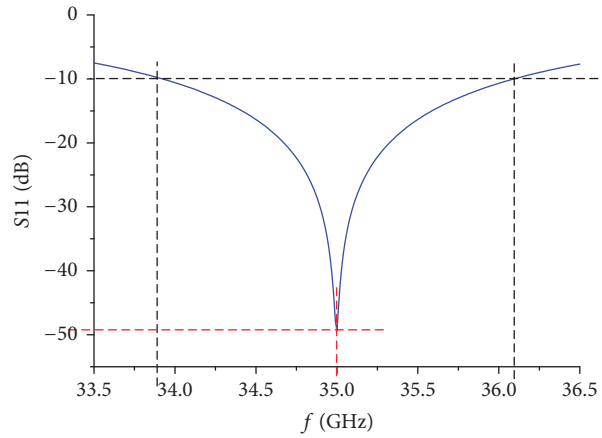


Figure 2: S-parameter of rectangular microstrip patch antenna.

The rectangular microstrip antenna element is shown in Figure 1. W and L of the patch element are adjusted during the simulation process. The size of L mainly affects the resonant frequency, and W mainly affects impedance matching. Through simulation and optimization, we get that $W = 3.32$ mm and $L = 2.45$ mm, the width of feed line $w = 0.46$ mm, and the center of the feed line is midpoint of W.

All simulations are performed using HFSS in the following. Figure 2 is S-parameter of rectangular microstrip patch antenna. It can be seen from the figure that return loss has reached −49 dB at 35 GHz. The relative bandwidth for $|S_{11}| < -10$ dB can be found from the figure to be 6.6%.

In this design, the line width of the microstrip line is 0.46 mm, and the characteristic impedance is 50 Ω. Therefore, the input impedance of the patch element needs to be close to 50 Ω. Figure 3 is the input impedance of rectangular microstrip patch antenna. From the figure, the input

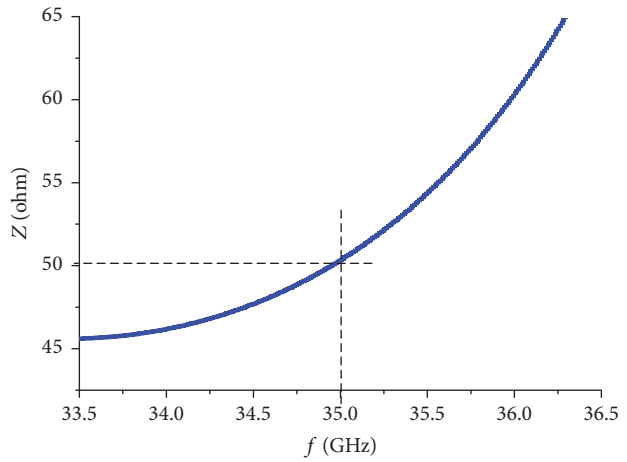


Figure 3: The input impedance of rectangular microstrip patch antenna.

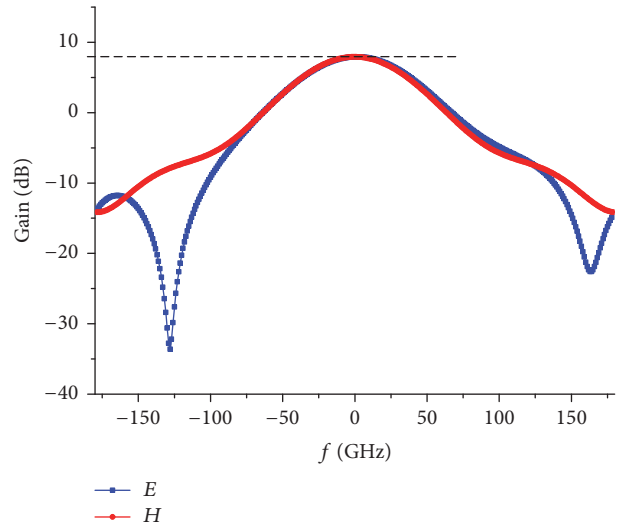


Figure 4: The gain of rectangular microstrip patch antenna.

impedance is about 50.34 Ω, which matches well to the characteristic impedance of the microstrip line.

Figure 4 is the gain of rectangular microstrip patch antenna, which shows that the maximum gain of the patch element is 7.93 dB, the 3 dB lobe-width of E-Plane is 76°, and the 3 dB lobe-width of H-plane is 73°. The pattern of the patch is consistent with the theory and the main lobe-width of E-plane is slightly larger than the H-plane.

The frequency band width of the microstrip antenna is not enough for the whole system. Therefore, in the following, we will analyze the array antenna to meet the requirements of the frequency bandwidth.

2.2. Theoretical Analysis of Series-Fed Array. The second part of MIMO conformal antenna design is the microstrip series-fed array. To get high gain, low sidelobe, beam scanning, and beam control, we need to use the discrete radiating element to form the array according to the appropriate excitation and distance. In this paper, the requirements of the microstrip

Figure 5: The radiating element connected with a fine line to realize the feeding.

array are as follows: the gain is 10 dB, the angle between the main lobe and plane of array is not less than 10% ($|S_{11}| < -10$ dB), and the first sidelobe level is about −18 dB. The design of microstrip is divided into three steps. The first step is to select the feed method of the linear array, the second step is to realize the offset of the main lobe, and third step is to reduce the first sidelobe level.

For the antenna array, the feeding method can be formed with parallel feed and series feed or the combination of the two [18, 19]. In this paper, we use series feed, as shown in Figure 5. The radiating elements are connected through a microstrip line, and the end is open circuit. The first radiating element is fed by a coaxial line. In order to avoid the influence of the microstrip on the antenna radiation, it is necessary to make it as thin as possible.

The second step is to realize the offset of the main lobe of the antenna pattern. There are typically series-fed traveling-wave array and the series-fed standing wave array. For series-fed traveling-wave array, the distance between radiating elements can be adjusted to achieve the offset of the main lobe. However, in the design of standing wave array with series-feeding, as long as the input impedance of the last radiating element of the array is designed to be consistent with the characteristic impedance of the microstrip line, it can also play a role in impedance matching to achieve the effect of traveling-wave array. This design uses a series of standing wave array, as shown in Figure 5. The advantage of this array is that it does not require the addition of terminal load. In the design of the radiating element, the input impedance of the radiating element is designed near the characteristic impedance of the microstrip. Therefore, the radiating element can be regarded as the matched load. We can change the phase relationship between elements by adjusting the distance between them in order to realize the arbitrary beam direction and achieve the effect of the main lobe [20].

For the design of a series-fed traveling-wave array, assuming that the main lobe angle from the end fire direction is θ, the relationship between the main lobe direction angle and the radiating element spacing is as follows:

$$\cos\theta = \frac{\lambda}{\lambda_g} - \frac{\lambda}{S}, \tag{4}$$

where S is the distance between the radiating elements and λ_g is the effective wavelength in the medium. When the distance $S < \lambda_g$, the main lobe biases feed; otherwise, it biases load. Element spacing S is an important parameter influencing the radiation characteristics of an antenna array. In order to avoid grating lobes, radiating element spacing S needed to meet the following formula:

$$S < \frac{\lambda_0}{1 + |\cos\theta_m|}. \tag{5}$$

Table 1: Normalized current value of eight-element Taylor array.

Unit number	1	2	3	4	5	6	7	8
Normalized current	0.6	0.63	0.83	1	1	0.83	0.63	0.6

Table 2: S_{12} of eight-element Taylor array.

Unit number	2	3	4
S_{12}	−3.1792	−2.8252	−2.2721

Formulas (4) and (5) are for traveling wave. In this paper, the design of the standing wave array can also use these two formulas. The distance between the radiating elements of the standing wave array obtained by the above two formulas is 4.41 mm.

The third step is to reduce the sidelobe level. The antenna design is based on 8-element linear array. Because of the need to achieve the main lobe deviation, the distance between the radiating elements is consistent. The sidelobe amplitude can be reduced by controlling the current. The current amplitude distribution design is based on the Taylor distribution [21, 22]. In the comprehensive design of the Taylor, it is necessary to determine the ratio R of the main lobe level to the sidelobe level. The value of A is calculated by R. Under the guarantee of $\overline{n} \geq 2A^2 + 1/2$, selecting appropriate $\overline{n}$ (the value of $\overline{n}$ increases, the value of σ decreases, and the lobe-width narrows down). The value of $\overline{n}$ should not be too large; otherwise, the amplitude distribution of the current will change dramatically. After selecting $\overline{n}$, beam broadening factor σ and current amplitude distribution of each radiating element can be calculated. Ratio of the main lobe level to the sidelobe level is $R = -18$ dB, $\overline{n} = 4$. The normalized current values of all levels are shown in Table 1.

There are two methods to change the current amplitude distribution. The first one is $\lambda/4$ impedance transformer, and the second is the patch width distribution method. Due to the relatively small spacing of the radiating elements, the $\lambda/4$ section cannot be added, so the patch width distribution method is used to change the current amplitude distribution. In fact, the change of current amplitude distribution can be realized by changing the radiation admittance of the element. Firstly, S_{12} of eight-element Taylor array at all levels should be calculated according to the current distribution. Feeding in this paper is from the center to both ends of the array. Therefore, according to the symmetry, only half of the array needs to be considered where calculating S_{12}. The calculated results are shown in Table 2. According to these values, the width of each radiating element can be adjusted, and the appropriate size can be found to satisfy the current amplitude distribution through simulation and optimization.

2.3. Simulation and Analysis of Microstrip Series-Fed Linear Array. The model of rectangular array with uniform distribution of one-end feeding is shown in Figure 6. *XOY* plane is the plane of the array. The rectangular microstrip patches with the same shape are used to design the array element. The spacing between patches is the same. First, we adjust the unit spacing to meet the main direction deflection of the beams of the microstrip series-fed linear array. Then, the array is connected to the external 50 ohm coaxial line. Because the impedance of the whole array is not matched to the 50 ohm coaxial line, an impedance transforming section should be added at the front of the array to match the impedance. The length of the section is $\lambda_g/4$. After optimization, the distance between the radiating element and the radiating element is 4.19 mm.

The S-parameter of a rectangular array with uniform distribution of one-end feeding is shown in Figure 7. It can be seen from the figure that, at the center frequency 35 GHz, the return loss is −27.7 dB. The relative bandwidth $|S_{11}| < -10$ dB can be found from the figure to be 26.14%.

The impedance of a rectangular array with uniform distribution of one-end feeding is shown in Figure 8. It can be seen that the antenna is well matched at 48.8 Ω.

The E-plane gain is shown in Figure 9. The maximum gain is 13.79 dB. The first sidelobe level is −13.2 dB. The main lobe deflection offset is achieved on E-plane, the angle is about 60°, and the 3 dB lobe-width in the E-plane is 16.8°.

The first sidelobe level is higher, which cannot meet the design requirements. We need to find the appropriate spacing between elements to meet the main beam deflection. Then the current distribution is designed to reduce the sidelobe level.

In this paper, the Taylor current distribution is used to reduce the sidelobe level. Taylor distribution is usually used in the form of intermediate feed. The spacing between radiation units usually takes one wavelength. As the design needs to achieve main beam offset, the spacing is no longer a wavelength. The feed position is required to be transferred to the center of the array, and the form needs to be adjusted. For an array which makes main beam deviation through changing the spacing between the radiation units, we in fact change the current phase difference between the radiating elements and then the main beam is offset. Considering the current phase difference for the whole array, add serpentine at one side of the array to adjust the phase difference and, at another side of the serpentine, the phase difference of 180 degrees should be added at the beginning of the unit. Then, adjust the current phase difference between the left and right arrays.

From the simulation results, we know that the unit spacing which can satisfy the main beam offset is 4.19 mm. The next step is to transfer the feed position to the center of the array, and add serpentine at one side of the array, determine the length of serpentine through the simulation. Before the design of Taylor matrix, we need to design a uniformly distributed rectangular array with intermediate feed to decide the length of serpentine. The design is shown in Figure 10; *XOY* plane is the model plane.

The rectangular microstrip patches with the same shape are used to design the array element. The spacing between antennas is the same. The feed structure of this array is in the middle of the array and is connected to a 50 Ω coaxial line. The design process is similar to that of a rectangular array with uniform distribution at one end of the feed. The design is divided into two parts. As can be seen from Figure 10, the left end of the feed is the same as the uniform distribution of one-end feeding. In the right end, serpentine lines are added to realize phase array progression, so as to realize the beam

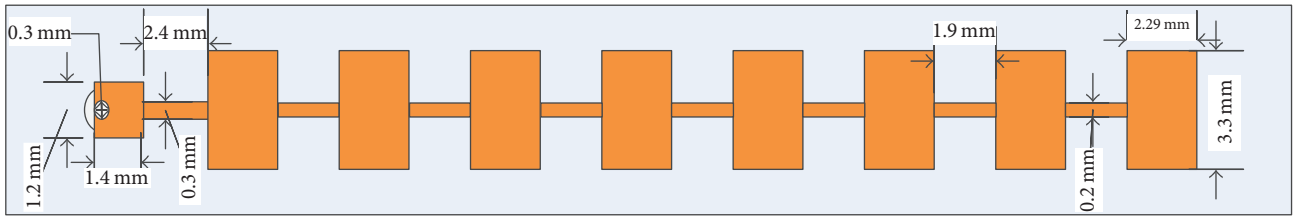


Figure 6: Model of rectangular array with uniform distribution of one-end feeding.

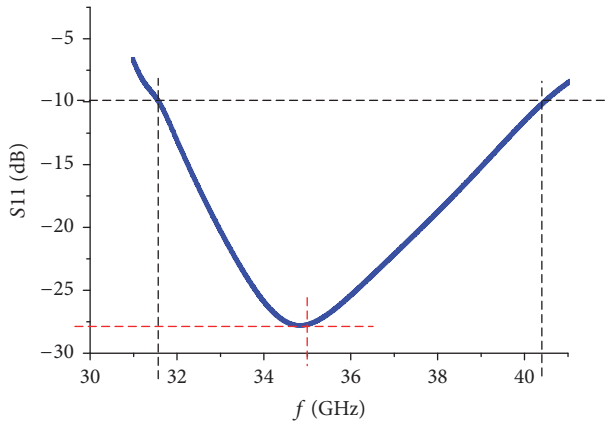


Figure 7: S-parameter diagram of a rectangular array with uniform distribution of one-end feeding.

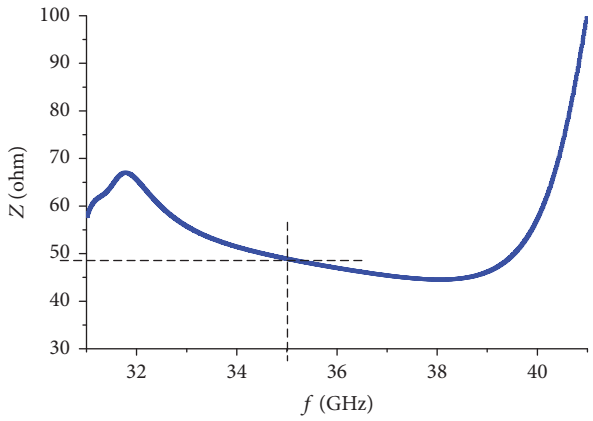


Figure 8: Impedance of a rectangular array with uniform distribution of one-end feeding.

deviation. After the simulation, the length of the serpentine is found to be 6.51 mm.

The S-parameter of a rectangular array with uniform distribution and intermediate feeding is shown in Figure 11. It can be seen from the plot that, at the center frequency of 35 GHz, the return loss is up to −31.87 dB. The relative bandwidth $|S_{11}| < -10$ dB can be found to be 21.7%.

The impedance is shown in Figure 12. It can be seen from the figure that the antenna is well matched at 51.5 Ω. The gain graph of E-plane is shown in Figure 13. It can be seen from the figure that the maximum gain is 13.36 dB, the first sidelobe level is −13.7 dB, the main beam deflection offset is achieved

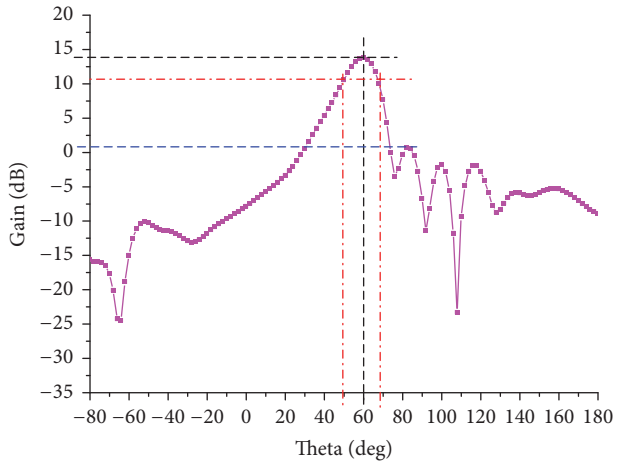


Figure 9: The gain of E-plane of a rectangular array with uniform distribution of one-end feeding.

on E-plane, the angle is about 62°, and the 3 dB lobe-width in the E-plane is 18°.

The designed array has satisfied the requirement of the main beam offset, but the sidelobe level is still too high. The feeding position is in the middle of the array, which satisfies the design of Taylor distribution. The design of Taylor distribution is carried out based on this array. The model diagram is shown in Figure 14; XOY plane is the plane of the array.

The form of the array is the same as that of the middle feed, and there is a difference in the size of the array. The size of the array element is designed according to Taylor current distribution regulation S_{12} of the elements at all levels which can be obtained from Table 2. By adjusting the radiation edge size of each element, the radiation admittance of each element can be changed, and the corresponding value of the radiation edge size can be obtained.

After simulation and optimization, from the feed point to the right, the sizes of radiation side are $W_1 = 3.7$ mm, $W_2 = 3.4$ mm, $W_3 = 4.1$ mm, $W_4 = 3.2$ mm.

The S-parameter diagram of a rectangular array with Taylor distribution and intermediate feed is shown in Figure 15. It can be seen from the figure that at the center frequency the return loss is very high (21.2 dB at 35 GHz). The relative bandwidth $|S_{11}| < -10$ dB can be found from the picture to be 11.6%.

The impedance is shown in Figure 16. It can be seen that the antenna is well matched at 50.1 Ω. The gain graph of E-plane is shown in Figure 17. It can be seen from the figure

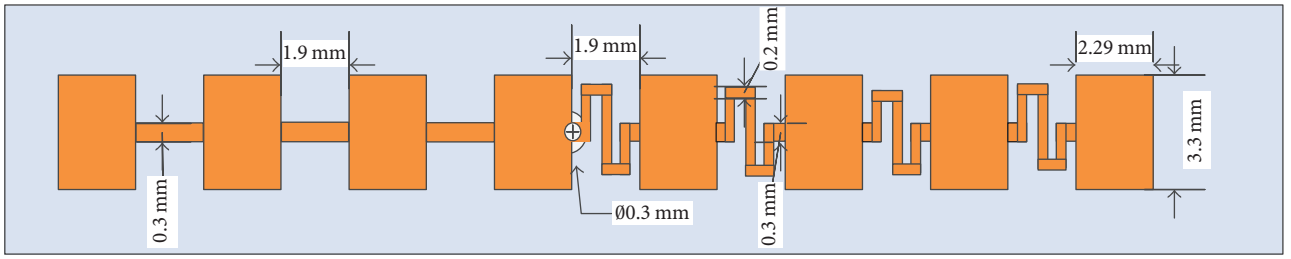


Figure 10: A uniformly distributed rectangular array with intermediate feeding.

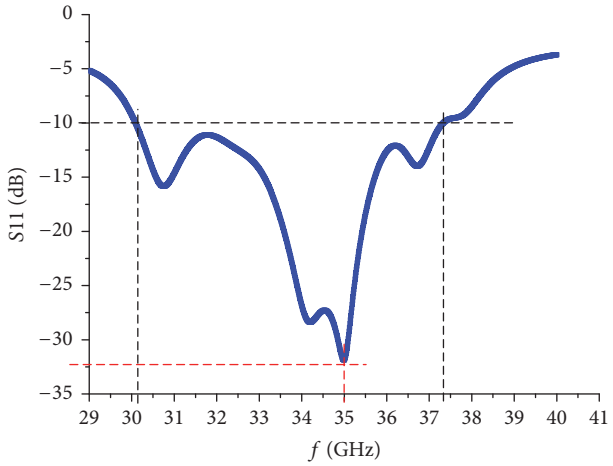


Figure 11: *S*-parameter of a rectangular array with uniform distribution with intermediate feeding.

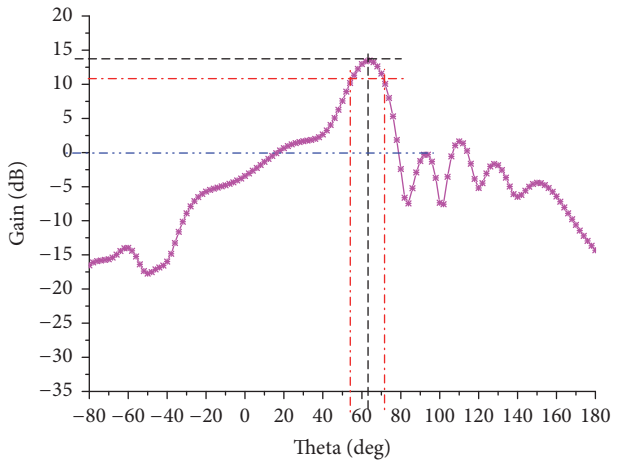


Figure 13: The gain of *E*-plane of a rectangular array with uniform distribution with intermediate feeding.

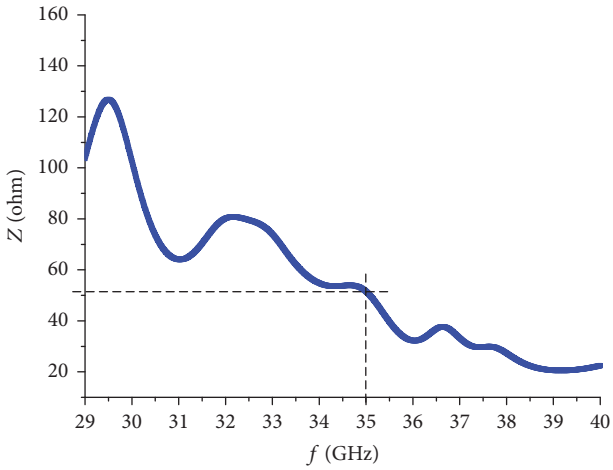


Figure 12: Impedance of a rectangular array with uniform distribution with intermediate feeding.

that the maximum gain is 13.9 dB, the first sidelobe level is −15.6 dB, the main beam deflection offset is achieved on *E*-plane, the angle is about 60°, and the 3 dB lobe-width in the *E*-plane is 20°.

Three kinds of arrays are given in this design. The first two arrays actually provide reference for the Taylor distribution matrix. The first array provides an appropriate spacing of the radiating elements. The second one determines the length of the serpentine. The final array form is based on the two arrays to adjust the radiation side of each radiating element to realize the current redistribution. In the array of rectangular patch, the gain is higher in the form of uniform distribution with intermediate feed. The lowest sidelobe level is the Taylor distribution with intermediate feed to reduce the sidelobe. The narrowest beam and the best matched impedance are the uniform distribution with one end of the feed. It can be seen that the reduction of the first sidelobe level is at the expense of width of the main lobe.

3. Design of Conformal Arrays

The conformal array is designed in the third part of the MIMO conformal antenna, and the conformal carrier is the cylinder with a diameter of 60 mm [23].

The center frequency of the design is 35 GHz, and the dielectric substrate with relative dielectric constant $\varepsilon_r = 2.2$ is selected. Thickness of the substrate is 0.5 mm. According to the design of the radiation unit, the size of the microstrip patch antenna is only about 3 mm. The curvature of 60 mm cylindrical diameter is smaller than the microstrip patch antenna. So the antenna can be regarded as a planar antenna and analyzed by the theory of planar antenna. The design needs to achieve a specific beam direction, which is 60° to the conformal vector axis, and can be realized by conformal

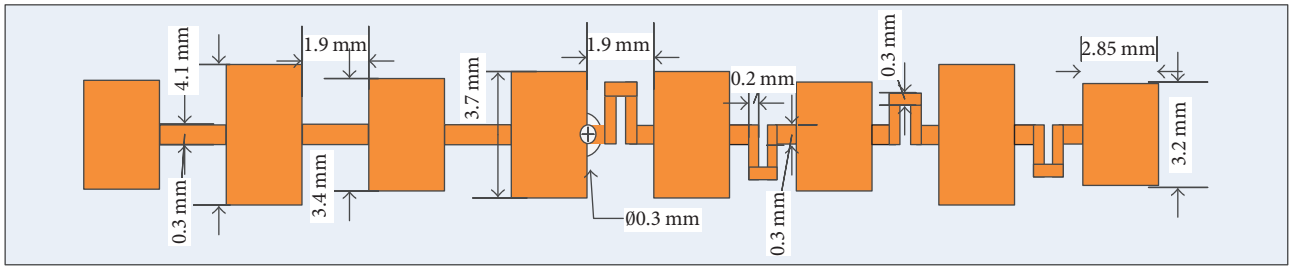


Figure 14: A rectangular array with Taylor distribution with intermediate feed.

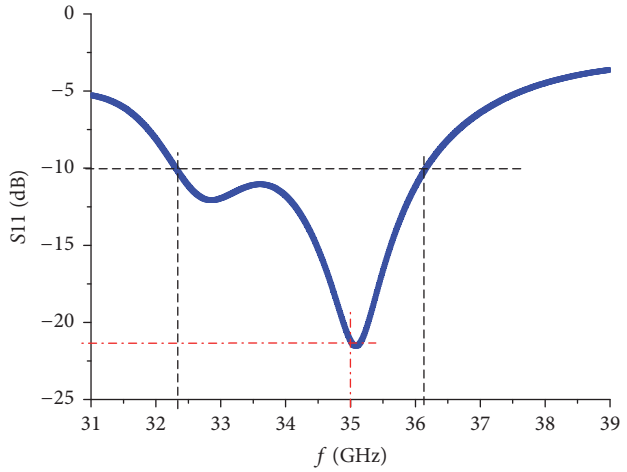


Figure 15: S-parameter of a rectangular array with Taylor distribution with intermediate feeding.

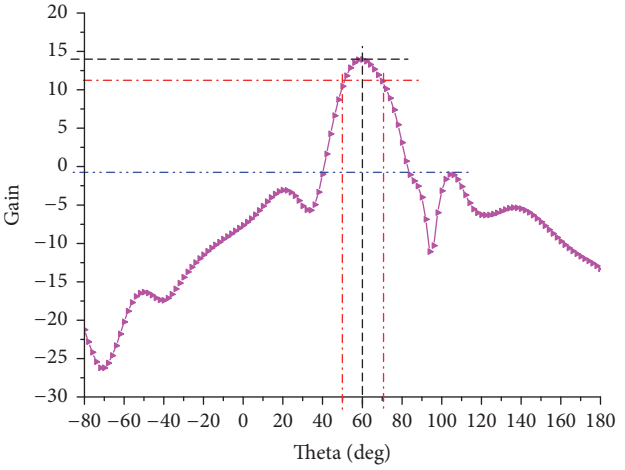


Figure 17: The gain of E-plane of a rectangular array with Taylor distribution with intermediate feeding.

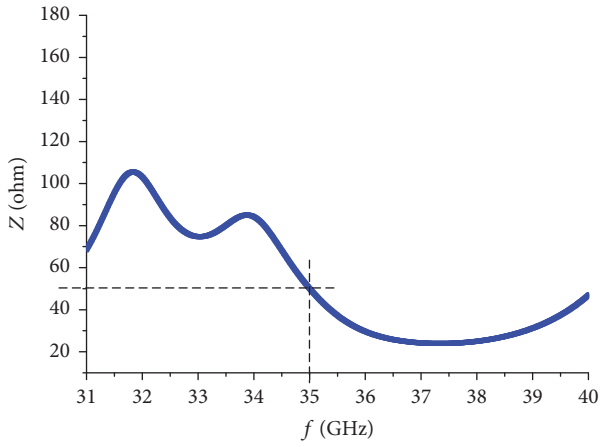


Figure 16: Impedance of a rectangular array with Taylor distribution with intermediate feeding.

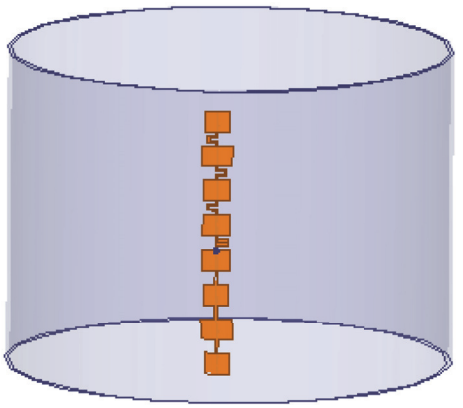


Figure 18: The microstrip antenna conformal array.

array. According to the analysis of the series-feed array, the microstrip patch antenna can be composed of a series-feed array to achieve such a beam direction. It can be realized by adjusting the spacing between the elements. The low sidelobe can be realized by the Taylor synthesis method. The distribution current of the antenna array is tapered to reduce the sidelobe level.

The microstrip antenna conformal array is shown in Figure 18.

Through simulation and optimization, the radiation characteristic of microstrip antenna array is obtained. The S-parameter is shown in Figure 19. The return loss of the conformal array is down to 14 dB at 35 GHz. The relative bandwidth of $|S_{11}| < -10$ dB can be calculated to be 11.8%.

The input impedance of the array is shown in Figure 20. It can be seen that the whole antenna is matched to 45 Ω. The

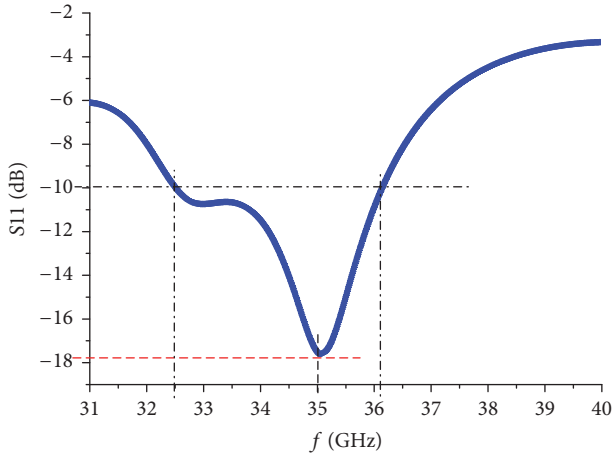


Figure 19: *S*-parameter of conformal microstrip antenna array.

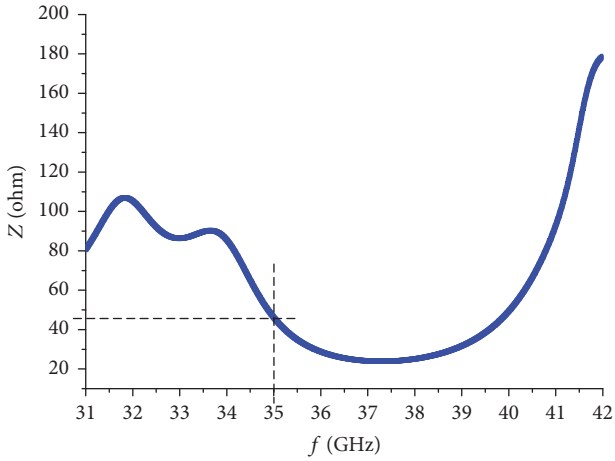


Figure 20: Input impedance of conformal microstrip antenna array.

gain of *E*-plane microstrip antenna conformal array is shown in Figure 21. It can be seen that the maximum gain of the array is 13.3 dB, and the first sidelobe level is −16 dB. The *E*-plane realized the main beam offset. The offset angle is about 62°. The 3 dB beam width of *E* plane is 18.2°.

We fabricated a pair of 8-element series-fed conformal antenna array, as shown in Figure 22. The analysis and test results are compared.

S-parameters are shown in Figure 23. It can be seen from the results of the *S*-parameters of the antenna that the resonance point of the measurement and simulation is consistent. The measurement results of the antenna relative bandwidth is about 11%, which is slightly less than the simulation results.

The comparison of measurement results and simulation results of normalized pattern of *E* plane is shown in Figure 24. It can be seen that the radiation plot of *E*-plane achieves the main beam offset. The angle is about 62° and this meets the requirements. The first sidelobe level rose to −14 dB. The 3 dB lobe-width of *E*-plane is about 17°. The measurement results are in good agreement with the simulation results.

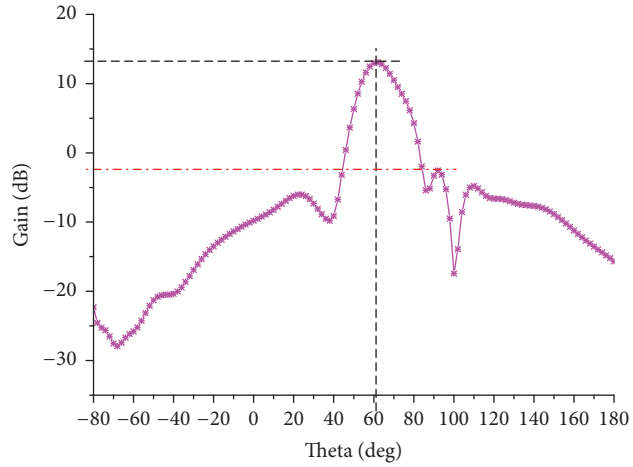


Figure 21: Gain of *E*-plane of conformal microstrip antenna array.

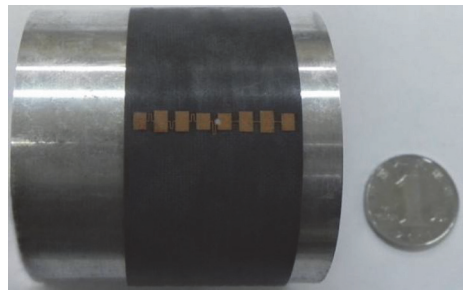


Figure 22: Eight-element series-fed conformal antenna array.

The gain measurement is performed by comparing to a standard horn antenna. The gain of the antenna is 12.2 dB.

4. Coupling Analysis of 5G MIMO Conformal Antenna

It is also very important to decide the number of the antennas in the design of MIMO conformal antenna. Antenna coupling has significant impact on radiation pattern. When using multiple antennas, the cross coupling between the antennas should be discussed and the coupling needs to be minimized. The main factor that affects the coupling is the distance between antennas. The closer the distance between antennas, the stronger the coupling. Therefore, we should find the appropriate antenna spacing to meet the performance requirement of the antenna coupling. At the same time, this determines the maximum number of antennas [24]. For conformal system, the diameter of the conformal carrier is 60 mm. The carrier space is limited, so the number of RF circuits is limited. At the same time, too many RF circuits will increase the cost of the system. Therefore, the number of antennas needs to be determined by considering these factors.

As we know, the energy of array antenna can be coupled by space wave or surface wave, when the coupling level is greater than −20 dB, the performance of the antenna will be greatly affected [25]. In this design, each antenna is used

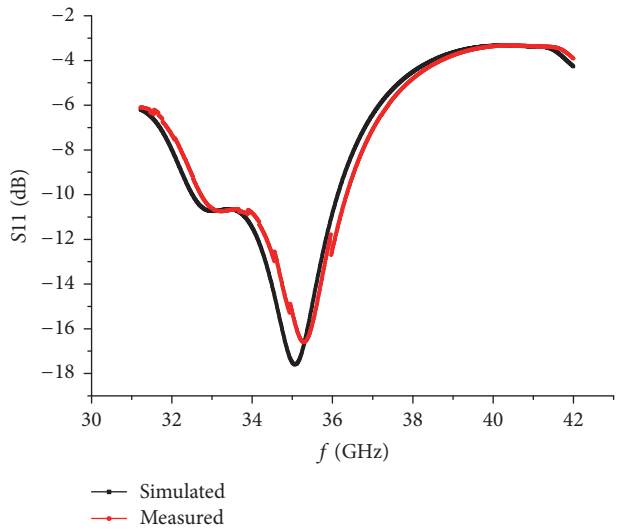


Figure 23: Comparison of S-parameters and simulation results of the 8-element series-feed conformal antenna array.

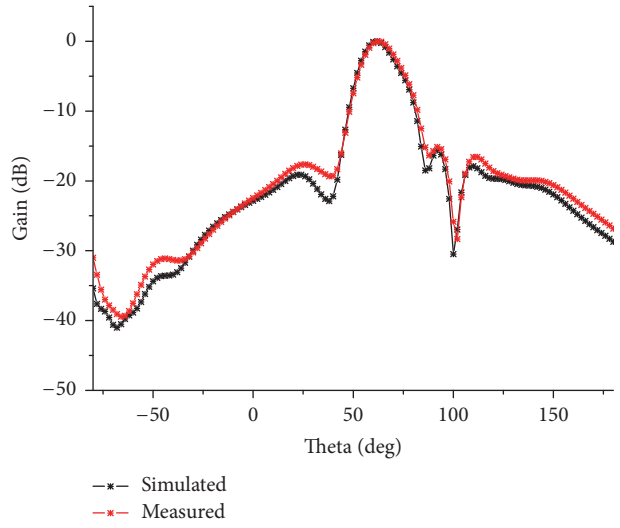


Figure 24: Comparison of the measurement and simulation results of the E surface orientation of the 8-element series-fed conformal antenna array.

as an independent antenna, so the requirement of coupling between the antennas is low, and the coupling between antennas is −40 dB. Through simulation and optimization, we find that when the number of the antennas is eight, it can meet the requirement of the coupling of antenna to be less than −40 dB. Eight antenna arrays can be placed on the conformal carrier. Actually we cannot put so many antennas. First of all, to consider the cost of the RF circuit, eight arrays of antennas require eight RF links. Secondly, to consider the volume of the conformal carrier, not more than four radio frequency links can be placed in this limited space. Therefore this design uses four arrays. Combined with the spatial symmetry of the antenna, the four pairs of antennas are distributed with equal distance in the conformal cylindrical carrier. The simulation model is shown in Figure 25.

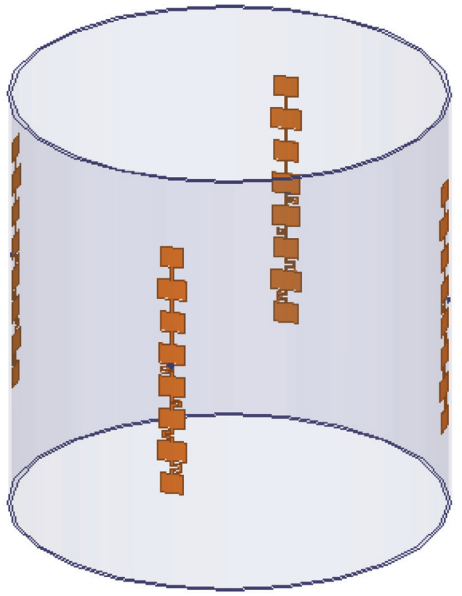


Figure 25: MIMO radar conformal antenna.

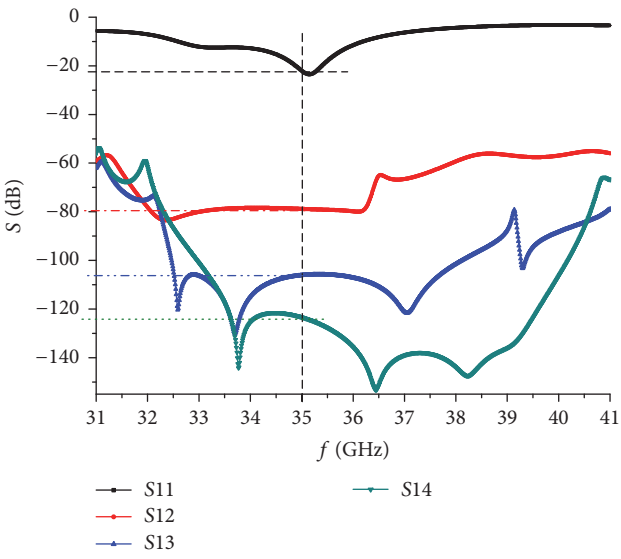


Figure 26: S-parameters of MIMO conformal antenna 1.

The S-parameters of the first antenna in the MIMO radar conformal antenna are shown in Figure 26. It can be seen from the figure that the antenna's reflection parameters are −21.5 dB. Considering the coupling of antennas, it is clear that the cross coupling satisfies the requirements of −40 dB.

The radiation plot of the first antenna in the MIMO radar conformal antenna is shown in Figure 27. It can be seen from the figure that, by considering the coupling effects, the first sidelobe level has been significantly improved to −11.5 dB. This is consistent with the theoretical analysis.

The fabricated conformal antenna with 4 arrays of antennas is shown in Figure 28.

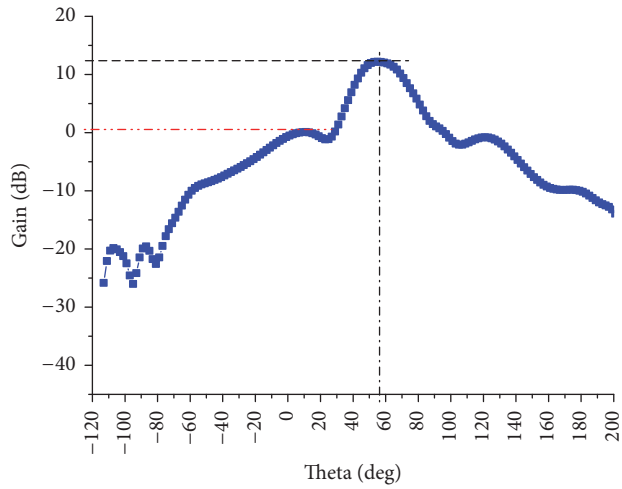


Figure 27: Gain of *E*-plane MIMO conformal antenna 1.

Figure 28: Planar expansion of four pairs conformal microstrip antenna.

5. Conclusion

In this paper, in order to meet the requirements of the antenna system, an 8-cell series-fed microstrip standing wave antenna array by traveling-wave theory has been designed in Ka band (35 GHz). The deflection of the main lobe and the plane of array is realized by adjusting the spacing between the elements. At the same time, the Taylor distribution is used for the antenna synthesis, and the first sidelobe level is reduced by controlling the current amplitude of the unit. Next, MIMO conformal antenna at 35 GHz is designed. The bandwidth of antenna is greater than 10%, the gain is greater than 10 dB, and the first sidelobe level is reduced to −16 dB. The angle between the main lobe and the carrier axis is 60°. The measurement results agree well with the simulation results, which meet the requirements of the system to the antenna performance. Considering the cost of system, space limitation, and antenna coupling, we design four 8-cell series-fed microstrip standing wave antenna arrays. The four antenna arrays are evenly distributed on the conformal of carrier, and the cross coupling of the antenna is lower than −40 dB.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] YaoHuan Gong, "Multiple input multiple out of smart antenna technology," *ZTE Technology*, vol. 6, pp. 19–21, 2002.

[2] Li Gang Ren and Mei Song, "MIMO technology in mobile communication [J]," *Modern telecommunication technology*, vol. 1, pp. 42–45, 2004.

[3] G. J. Foschini, "Layered space-time architecutre of wireless communication in a fding environment when using multi-element antennas," *Bell Labs Technical Journal*, vol. l, no. 2, pp. 41–59, 1996.

[4] G. J. Foschini, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, 1998.

[5] WeiHong Xiao, *Multi Antenna Design for MIMO Mobile Communication System*, Xidian University, Xi'an, China, 2006.

[6] YanJie Zhang, *Research on Coupling Characteristics of MIMO Antena*, Xidian University, Xi'an, China, 2012.

[7] Z. Yang, H. Yang, and H. Cui, "A compact MIMO antenna with inverted C-shaped ground branches for mobile terminals," *International Journal of Antennas and Propagation*, vol. 2016, Article ID 3080563, 2016.

[8] Kyungjung Kim., Sarkar, M. C. Wicks et al., "DOA Estimation Utilizing Directive Elements on a Conformal Surface," in *Radar Conference, 2003. Proceeding of the 2003 IEEE*, pp. 91–96, Huntsville, Alabama, 2003.

[9] R. K. Hersey, W. L. Melvin, and J. H. McClellan, "Clutter-limited detection performance of multi-channel conformal arrays," *Signal Processing*, vol. 84, no. 9, pp. 1481–1500, 2004.

[10] Z. Li, X. Kang, J. Su, Q. Guo, Y. L. Yang, and J. Wang, "Clutter Limited Detection Performance of Multi-channel Conformal Arrays," *International Journal of Antennas and Propagation*, vol. 2016, Article ID 9812642, 2016.

[11] T. E. Morton and K. M. Pasala, "Pattern synthesis of conformal arrays for airborne vehicles," in *Proceedings of the 2004 IEEE Aerospace Conference Proceedings*, pp. 1030–1038, Big Sky, Montana, USA, March 2004.

[12] R. K. Mishra and A. Patnaik, "Design of circular microstrip antanna using neural networks," *IETE Journal of Research*, vol. 44, no. 1-2, pp. 35–39, 2015.

[13] A. Sayed, R. Ghonam, and A. Zekry, "Design of a Compact Dual Band Microstrip Antenna for Ku-Band Applications," *International Journal of Computer Applications*, vol. 115, no. 13, pp. 699–702, 2015.

[14] DongLiang Zhao, "Thoughts on the development of 5G mobile communication," *Information Communication*, vol. 9, 2015.

[15] HongJie Yao, "The key technology and process of 5G mobile communication," *Communication World*, vol. 6, 2015.

[16] Warren L.Stutzman and Gary A. thiele, *Encyclopedia of RF and Microwave Engineering*, People's Posts and Telecommunications Publishing House, China, 2nd edition, 2006.

[17] YuFang Tang, *Theoretical study and engineering application of microstrip line loss*, Nanjing University of Science and Technology, 2009.

[18] Jun zhang and KeCheng Liu, *Microstrip antenna theory and Engineering*, National Defense Industry Press, China, 1988.

[19] ShuJie Li, *Research on Microstrip Planar Array Antenna in Ku Band*, Xidian University, Xi'an, China, 2006.

[20] DaJun Wu, *Design And Research of Millimeter Wave Cylindrical Conformal Microstrip Antenna*, Nanjing University of Science and Technology, Nanjing, China, 2007.

[21] XingJian Kang, *Principle And Design of Antenna*, National Defense Industry Press, China, 1995.

[22] ChuFang Xie, *Modern Antenna Theory*, Chengdu Telecommunication Engineering Institute Press, Sichuan, China, 1987.

[23] Song Zhu, "Development of conformal antenna and its application in electronic warfare," *Journal of the Chinese Academy of Electronic Science*, vol. 12, 2007.

[24] ChangSheng Shi, "Isolation of antenna," *Electronic Science and Technology Revie*, vol. 11, pp. 16–18, 1997.

[25] QiaoLong Lan, *Research on Millimeter Wave Microstrip Antenna*, University of Electronic Science and technology, Chengdu, China, 2006.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 9843825, 13 pages
https://doi.org/10.1155/2017/9843825

Research Article

Maximum Power Plus RSSI Based Routing Protocol for Bluetooth Low Energy Ad Hoc Networks

Changsu Jung and Kijun Han

School of Computer Science and Engineering, Kyungpook National University, Daegu, Republic of Korea

Correspondence should be addressed to Kijun Han; kjhan@knu.ac.kr

Received 20 July 2017; Revised 26 October 2017; Accepted 27 November 2017; Published 13 December 2017

Academic Editor: Haiyu Huang

This paper proposes a novel energy-conserving multihop routing protocol to maximize network lifetime and consume the battery in a distributed manner during route discovery in energy-constrained Bluetooth Low Energy ad hoc networks. Furthermore, a flooding avoidance approach is adopted by the proposed scheme to minimize the number of route request packets flooded. In addition, the proposed scheme maximizes network lifetime by using residual battery and RSSI as a route metric. The simulation results confirm that our proposed scheme has a surpassing performance with regard to network lifetime, evenly distributed battery consumption, and route discovery latency compared with the traditional on-demand routing protocol.

1. Introduction

Bluetooth Low Energy (BLE) [1] is a notable wireless communication technology with low-power, low-cost, and low complexity and it is regarded currently by many researchers as the ideal technology for realizing IoT, M2M, and energy-constrained applications [2, 3]. In particular, mobile ad hoc networks (MANET) and wireless sensor networks (WSN) are commonly operated by battery-powered devices in energy-constrained networks. Therefore, low-power and low-cost communication technologies are necessary to connect sensors and mobile nodes. Additionally, multihop communication should be considered to communicate with out-of-the-radio-coverage devices in the massive sensor networks and MANET. Due to its low-power consumption, BLE technology is adequate to implement MANET and WSN.

For instance, sensors and mobile nodes are distributed in ad hoc environment, which has no fixed infrastructure, no information of the entire network topology, and constant transformation of the network topology. Moreover, a device depends on its battery as the power source; thus, the whole network may collapse due to a node's depleted energy. Therefore, network lifetime depends on each node's residual battery level. Due to the limitations of ad hoc networks mentioned beforehand, energy-conserving scheme is a very important issue [4].

A few scenarios using BLE technology can be considered. Firstly, BLE technology can be applicable in the environment monitoring systems transmitting changes of the physical world periodically. Secondly, a disaster area without network facilities can use BLE technology to deliver short messages using BLE embedded gadgets such as wide-spread smartphones and tablets. In such ad hoc environment, multihop communication is necessary to deliver short messages to a long distance.

However, Bluetooth 4.0 only supports single role and one hop communication, which is called a piconet. The single-hop communication nature of BLE restricts multihop communication and was not solved until the release of Bluetooth version 4.0. The latest attributes of Bluetooth 4.1 specification permit BLE devices to have binary roles and participate in multiple piconets. Thus, corresponding and conveying data to a BLE device beyond single-hop distance has become possible for BLE devices [5]. Although the multihop routing is supported from BLE version 4.1, the BLE specification does not include a specific algorithm for the aforementioned purpose.

A routing protocol is a requisite to transfer data from a source to a multihop distance destination to support multihop communication among devices. Proactive (table-driven) and reactive (on-demand) approaches are the two categories of routing protocols of MANET [5]. In the former method,

each node retains a whole routing information of the network so the routing discovery latency is very short. However, each node consumes more energy to renew its routing table at any time in which the topology of the network is modified.

Contrarily, MANET extensively adopts reactive routing protocols to minimize power usage by establishing connections in on-demand manner. On account of on-demand route discovery features, a route path is constructed once the node generates a route request. Thus, on-demand approach is an appropriate multihop communication protocol under the power-restricted environment of BLE networks. Ad hoc On-demand Distance Vector (AODV) protocol [6] is another ad hoc routing protocol, which is well known for its simplicity and efficiency in MANET [7]. However, the traditional AODV has a flooding issue. A source floods route request messages (RREQ) to the entire networks to discover a path to a destination. Consequent to the flooding mechanism, AODV degrades the performance in the energy-constrained networks [8]. In addition, Temporally Ordered Routing Algorithm (TORA) [9] is a source initiated on-demand routing protocol in ad hoc networks which uses a broadcasting scheme to find a destination using query (QRY) and update (UPD) packets. If a source node wants to find a destination to transmit data, the source broadcasts a QRY packet to its neighboring nodes and the destination or adjacent nodes broadcast an UPD packet to the upstream nodes. When an intermediate node receives an UPD packet from a downstream node, the intermediate node calculates the height value which is the hop count number to the destination and broadcasts UPD including its height. After the route discovery procedure, TORA organizes a directed acyclic graph (DAG) which is rooted at the destination. Due to the broadcasting approach of the route discovery procedure, TORA also consumes more energy in the energy limited networks.

In this paper, we propose a new energy-conserving routing protocol including multihop communication scheme in the power-restricted BLE networks. We focus on evenly distributed energy consumption among BLE devices and maximize the network lifetime in the energy-constrained networks such as MANET. The proposed protocol utilizes residual battery power and Received Signal Strength Indication (RSSI) as route metrics. In addition, to decrease the volume of route request packets which cause significant power usage, the flooding avoidance approach is adopted.

We organized the remaining parts of this study in this manner. Related works are illustrated in Section 2. Section 3 demonstrates the state transition of BLE node and node discovery procedure. Section 4 illustrates the energy-conserving multihop routing protocol and a route metric proposed. Finally, Section 5 evaluates the proposed routing protocol and depicts the outcomes of simulation and the conclusion is shown in Section 6.

2. Related Works

We introduce a short illustration of the existing energy-conserving mechanisms in ad hoc networks. Further, we reviewed multihop routing protocol in Bluetooth networks [7, 10–15], since fewer studies have been performed on BLE networks [5].

Toh [16] proposed a routing protocol utilizing battery capacity as a route metric to satisfy two critical challenges in wireless ad hoc networks, namely, maximizing lifetime and evenly distributed power consumption. The study proposed Conditional Max-Min Battery Capacity Routing (CMMBCR) to select the energy efficient route path with the aid of a threshold value. The scheme selects the shortest path excluding any nodes that operate under the defined battery capacity to extend the network lifetime and operation time. The study mentioned that power-aware routing protocols are inclined to have longer route paths that reduce the lifetime of nodes.

A route metric that occupies a combination of Minimum Drain Rate (MDR) and the remaining battery level was proposed to estimate the lifespan of individual node regarding recent traffic loads [17]. The proposed metric decides which nodes can participate in a route. Moreover, the study introduced Conditional Minimum Drain Rate (CMDR) to reduce the whole energy consumption during transmission. CMDR selects a path which is longer than a defined threshold among all the possible paths. However, if nodes which belong to all the possible paths do not satisfy the threshold, the proposed scheme follows MDR mechanism.

Zhang and Riley [18] proposed an energy-aware on-demand routing protocol in Bluetooth sensor networks. When data is required to be sent to a sink node, a source discovers its neighboring nodes. When a source sends a route request message to a relay node, it selects a node with the highest remaining current level. A route request message is delivered to an intermediate node; it stores the information of the source and the previous node in its neighbor list for route reply messages and flooding loop prevention. Forwarding or discarding decision is dependent on the surplus power of a node along the route path to extend the network lifetime.

A power-efficient reactive routing protocol was introduced in ad hoc networks [19]. The research used average energy consumption and link error rate as route metrics to find an energy efficient route path. In addition, total energy for data transmission and packet error rate were calculated for a reliable route path. The performance was evaluated in terms of throughput and average energy costs using three different mobility models. In the performance evaluation, a few routing schemes such as shortest-delay, power-aware, and retransmission energy were adopted.

A routing protocol based on transmission power was proposed for mobile ad hoc networks [20]. The proposed scheme uses a RSSI and residual battery level as a route metric to find an energy efficient route path. To discover the nearest neighboring node, a node broadcasts route request messages and it calculates RSSI and distance using replied messages from the neighboring nodes. A node selects the next hop which has a higher RSSI and adjusts its transmission power to save its energy consumption.

A reactive routing protocol supporting multihop communication with scatternet formation was proposed in BLE wireless sensor networks [5]. For neighbor discovery procedure, each node alternates its role as an advertiser and scanner. To find a route path of a destination, a source delivers a route

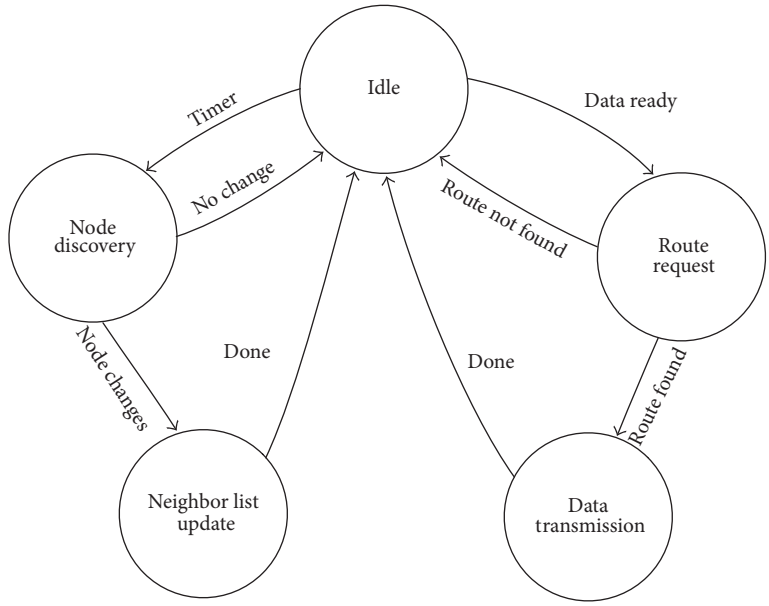


Figure 1: The state transition diagram of a BLE node.

request message to its master in its piconet. If a destination is not found in a piconet, the master sends the message to a relay node in another piconet. After a destination receives multiple route request messages, it sends route reply messages to all possible paths. Finally, a source selects the shortest path from the route response messages.

An enhanced routing protocol based on the traditional AODV was introduced to improve the flooding issue during route request procedure in Bluetooth ad hoc networks. Three metrics, namely, hop count, battery capacity, and average traffic loads, are adopted to handle multiple route request packets. A route request packet contains three metrics to find the lowest cost path. The route metrics are updated with the node's values as a route request packet, which includes three metrics, arrived at an intermediate node. When multiple route request messages are arrived at a destination, the proposed route selection algorithm chooses the lowest cost path. Moreover, route request packets are only delivered to bridge nodes to minimize the number of route request packets. However, the proposed routing protocol does not present the configuration of bridge nodes and network topology.

An on-demand multihop routing protocol using scatternet structure was introduced in Bluetooth networks [21]. A route path is established whenever data transmission is required and only contains the nodes along the route path. When communication starts, a route path is set up dynamically and released after the transmission. A node can have a single role and double roles for the scatternet route structure. To connect piconets in critical areas, double role nodes are allowed in this research.

3. Node Discovery Procedure

In this section, we describe the state changes and neighbor discovery of BLE nodes in the discovery procedure. BLE nodes are presumed to be distributed randomly in a substantial indoor structure.

3.1. State Transition of BLE Node. When BLE devices turn on, they enter "Idle" state to save their battery consumption. Each BLE device discovers neighbor nodes within a defined time using the discovery timer. If the timer expires or the neighbor list is unchanged, BLE devices return to "Idle" state. If a BLE device detects some changes of incoming or outgoing devices within the radio range, it updates the neighbor list to include the detected information. After updating the neighbor list, a BLE device goes back to "Idle" state. When a device wants to transmit data to a destination, it moves to "route request" state and sends data in "data transmission" state. After transmitting data, a BLE device returns to "Idle" state. Figure 1 illustrates the state transition of a BLE device in the proposed scheme.

3.2. Node Discovery Procedure Using Random Role Switching. Node discovery procedure is collecting adjoining nodes' ID and their residual battery levels by reciprocating advertisement messages. A BLE node updates a neighboring node's residual battery level in its neighbor list after receiving an advertisement message from a nearby node. The residual battery level is used as a route metric with RSSI to establish the minimum battery consumption route.

Random operation function in Figure 2 determines each BLE node's role. Then, a node changes its role at random after the operation time within the expected discovery time. To fulfill the maximum advertisement interval of BLE specification, the operation time is settled at 30 ms. The nodes which operate as advertisers broadcast advertising messages through the advertisement channels (37, 38, and 39) during the operation time including its own id and battery level. Similarly, scanners wait to receive advertising messages from

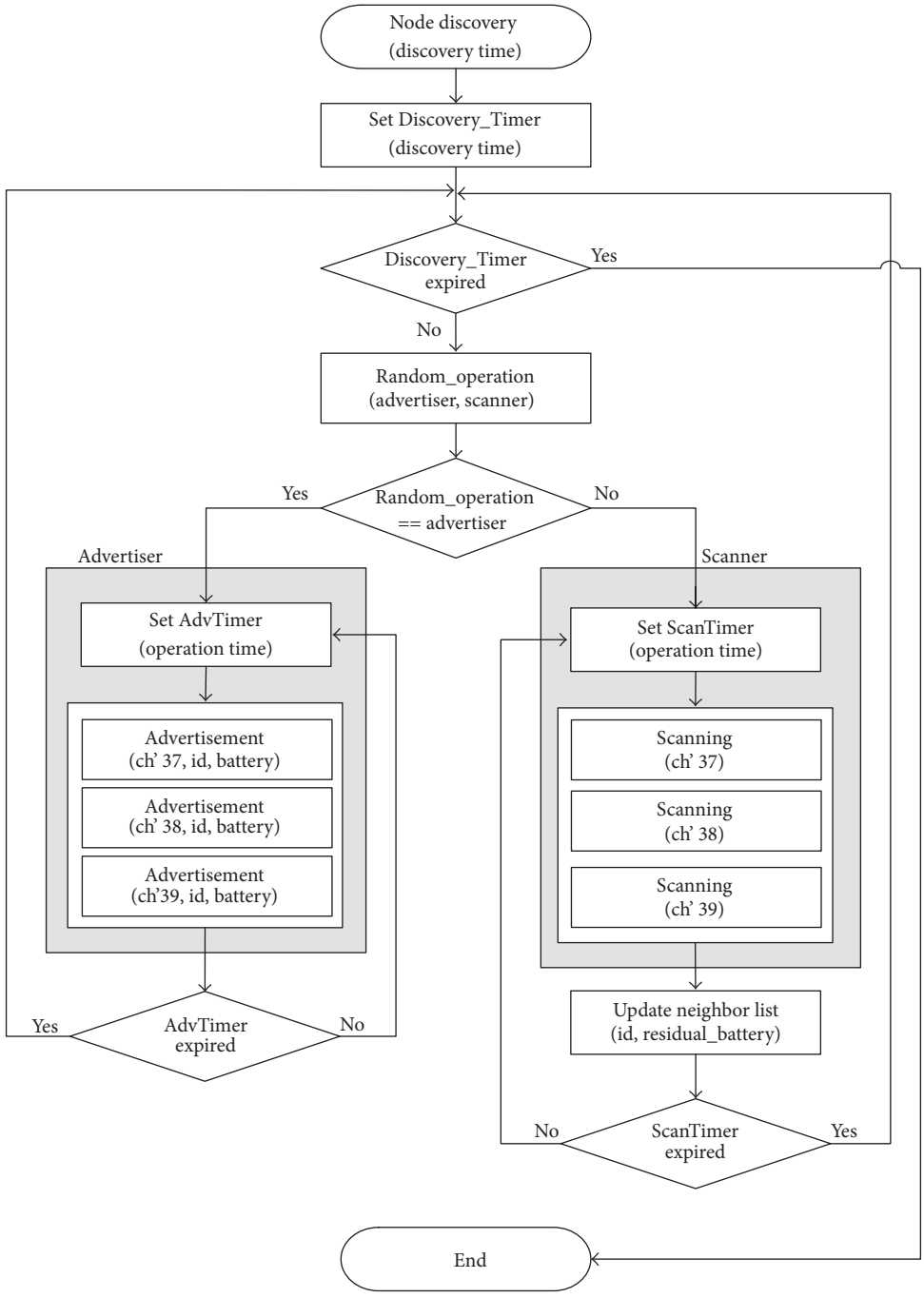


Figure 2: Flowchart of random role switching procedure.

neighboring nodes that are sent via the advertising channels during the operation time. This process continues through the discovery time as shown in Figure 2.

When a BLE node obtains advertisement packets from nearby nodes, it calculates weight of the advertising nodes with the residual battery level and RSSI. Sequentially, it updates the neighbor list entry for the corresponding advertising node. The proposed routing protocol utilizes two parameters as a route metric and chooses the maximum residual battery route using the weight value. The neighbor list has four fields, that is, device id, weight, residual battery, and RSSI as described in Table 1. After the node discovery procedure, each node completes configuring its neighbor list.

4. Energy-Conserving Multihop Routing Protocol

We introduce the energy-conserving on-demand multihop routing protocol in BLE ad hoc networks and how to calculate the route metric using residual battery and RSSI. The

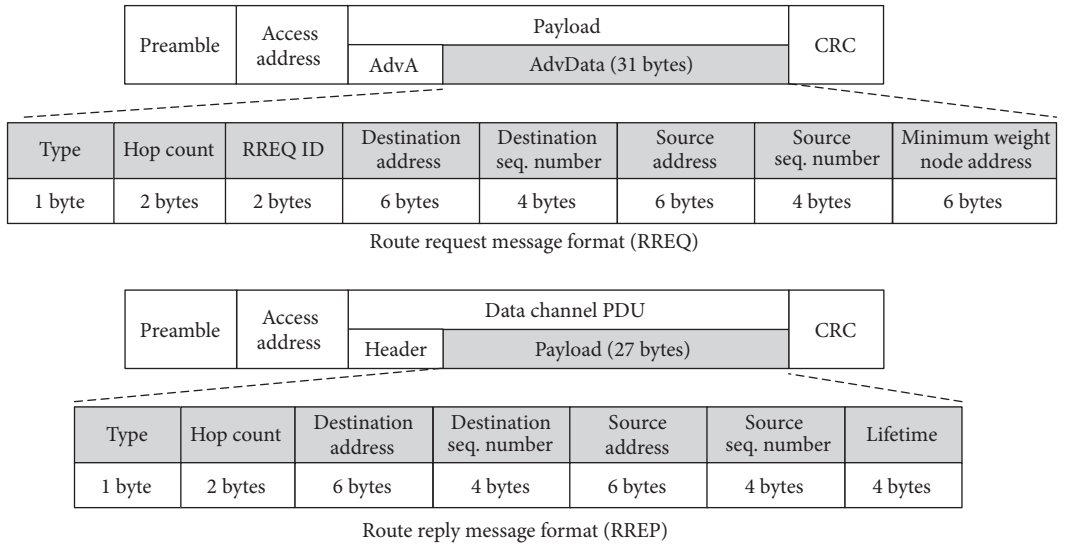


Figure 3: Route request and reply message formats.

Table 1: Neighbor list table.

Device ID	Weight	Residual battery	RSSI
6 bytes	2 bytes	4 bytes	2 bytes

proposed scheme chooses the minimum weight node to discover a route from a source to a destination. A route request packet is broadcasted using ADV_IND message and data channel PDU is used for transmitting a route reply packet. The two packet formats for the route discovery procedure are illustrated in Figure 3.

4.1. Route Metrics: Maximum Power Plus RSSI (MPPR). The route metric of this paper includes two metrics, that is, Maximum Power Plus RSSI (MPPR). If a source transmits data to a destination, a source finds a neighbor node with the minimum weight value in its neighbor list. Then, the source sends a route request message to the node having the minimum weight. The route metric is computed by using

$$\begin{aligned} \text{Weight}_i &= \alpha \times \left(1 - \frac{P_i}{P_{\max}}\right) + (1-\alpha) \times \frac{\text{RSSI}_i}{\text{RSSI}_{\min}}, \\ \text{Route}_j &= \min_{j \in \text{neighbors of node}_i} \left(\text{Weight}_j\right). \end{aligned} \tag{1}$$

P_i is the current residual battery of node i and $P_{\max}$ is the maximum battery level of a coin cell battery. RSSI_i is a RSSI value of node i and $\text{RSSI}_{\min}$ is the minimum RSSI value. We used 230 mA and −100 dBm, respectively, for $P_{\max}$ and $\text{RSSI}_{\min}$. The route metric is already calculated by exchanging advertisement messages during the node discovery procedure. If a node's residual battery (P_i) is close to $P_{\max}$ and RSSI value (RSSI_i) is large (close to 0), the weight of (1) reaches zero. Hence, if a node with the maximum residual battery (230 mA) and the larger RSSI value is subjected to be chosen during the route discovery procedure, a node searches its neighbor list and transmits the route request message to a nearby node having the minimum weight (Route_j). The MPPR routing scheme uses Prim's algorithm for forwarding route request packets. In addition, the proposed scheme utilizes a priority parameter (α) to compare the performance when α changes from 0 to 1.

4.2. Flooding Avoidance Route Discovery Scheme. The flooding RREQ packets consume substantial power of BLE nodes during route discovery procedure so this paper proposes the flooding prevention mechanism to minimize energy consumption. The proposed mechanism discards RREQ packets in the intermediate nodes using the minimum weight node address as illustrated in Figure 4. Whenever a source tries to discover a routing path to a specific destination, the source identifies the minimum weighted node in its neighbor list. Then, the source broadcasts a route request message including the address of the minimum weight node as shown in Figure 3. When a route request message is delivered to an intermediate node, the intermediate node checks the minimum weight node address. If the minimum weight node address of a RREQ packet is identical to the node receiving a RREQ message, then the node forwards the route request message.

For example, when node 6 forwards a RREQ message, node 6 looks up its neighbor list table and finds that node 7 has the minimum weight value. Then, node 6 sets the minimum weight node address as node 7's address in a RREQ message (Min. node = 7) and broadcasts a RREQ message. When neighboring nodes (nodes 3, 4, 5, and 8) of node 6 receive the RREQ message, they discard the route request message because their addresses are not equal to the minimum weight node address in the route request message. In addition, the proposed scheme uses the sequence number to avoid a route path loop like AODV protocol. When a node receives a RREQ message with the same sequence number which it already forwarded, the node discards the RREQ message to prevent the route loop.

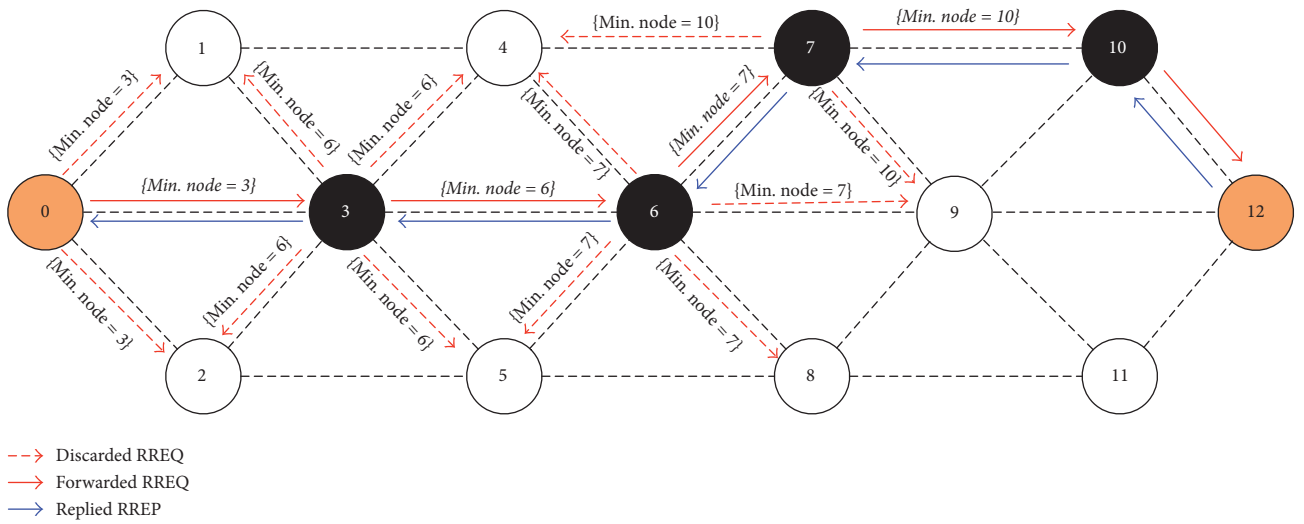


Figure 4: Flooding avoidance mechanism of route discovery procedure.

The procedure of flooding avoidance mechanism in an intermediate node is described in Figure 5. An intermediate node discards the route request message if the minimum weight node address of the packet is not identical to its address.

In this route discovery scheme, a source chooses a neighbor with the maximum residual power and the strongest RSSI value for the route request procedure. Whenever an intermediate node's battery level is lower than the neighboring nodes, the intermediate node is eliminated from the route path. This scheme can distribute energy consumption among BLE nodes.

4.3. Route Discovery Procedure. Aligning with the Bluetooth 4.1 specification, we considered route request and reply transmission time and delays for the proposed system. A RREQ message is simply broadcasted through advertisement channels. A route reply message (RREP) is delivered after establishing unicast connections along the route path. Thus, the route reply procedure is more complicated than the route request procedure as illustrated in Figure 6.

In this paper, we assumed that a RREQ message can be received within the advertisement period (T_{ADV}). Upon receiving a RREQ message, the node changes its role from a scanner to an advertiser and broadcasts it to the neighbors. Hence, we assumed that the role switch time (T_{ROLE_SW}) is one clock period. Moreover, we presumed that a unicast connection is established by exchanging ADV_DIRECT_IND and CONNECT_REQ messages during three advertisement periods (T_{CONN_SETUP}). After the connection setup, the mandatory delay (T_{POST_CONN}) was considered in accordance with the Bluetooth specification. The RREP transmission time (T_{REP}) is the RREQ packet size over BLE data rate (1 Mbps). Table 2 shows the description and time of route discovery latency.

4.4. Energy Consumption. The measurements of a CC2541 BLE chip of Texas Instruments [22] are referred to as the power usage model of this paper. In reference to these measurements, we determined the present current consumption and time of each phase in the course of route discovery process.

In the proposed routing protocol, RREQ messages are broadcasted during BLE advertisement period. Contrastingly, RREP messages are delivered via unicast connections during the route reply procedure. Accordingly, the route reply procedure exchanges more control message and consumes additional energy. Figure 7 illustrates the current level and time of each phase, respectively, and shows extra energy consumption of route reply procedure.

Table 3 describes the time and current consumption for each phase in an intermediate node during route discovery procedure as depicted in Figure 7.

5. Performance Evaluation

In this section, we describe the simulation environments and the performance evaluation of the energy-conserving routing protocol by modifying the priority parameter (α) in (1). We evaluated network lifetime, average residual battery, energy consumption, average number of route requests, and the latency of route discovery procedure. The proposed scheme adopted on-demand routing protocol to reduce network connectivity and maximize network lifetime. In this section, we compare the performance with other on-demand ad hoc routing protocols such as AODV and TORA. However, we excluded Dynamic Source Routing (DSR) protocol because DSR includes the addresses of intermediate nodes to transmit data. If a routing path has more than four hops, DSR over BLE cannot contain all nodes' address in its data packet because BLE payload is only 27 bytes long.

5.1. Simulation Environments. In the proposed scheme, all BLE devices are presumed to be spread randomly within 100 × 100 meters and the radio coverage is 20 meters. The BLE nodes increased in the increments of 10 from 50 to 100 and

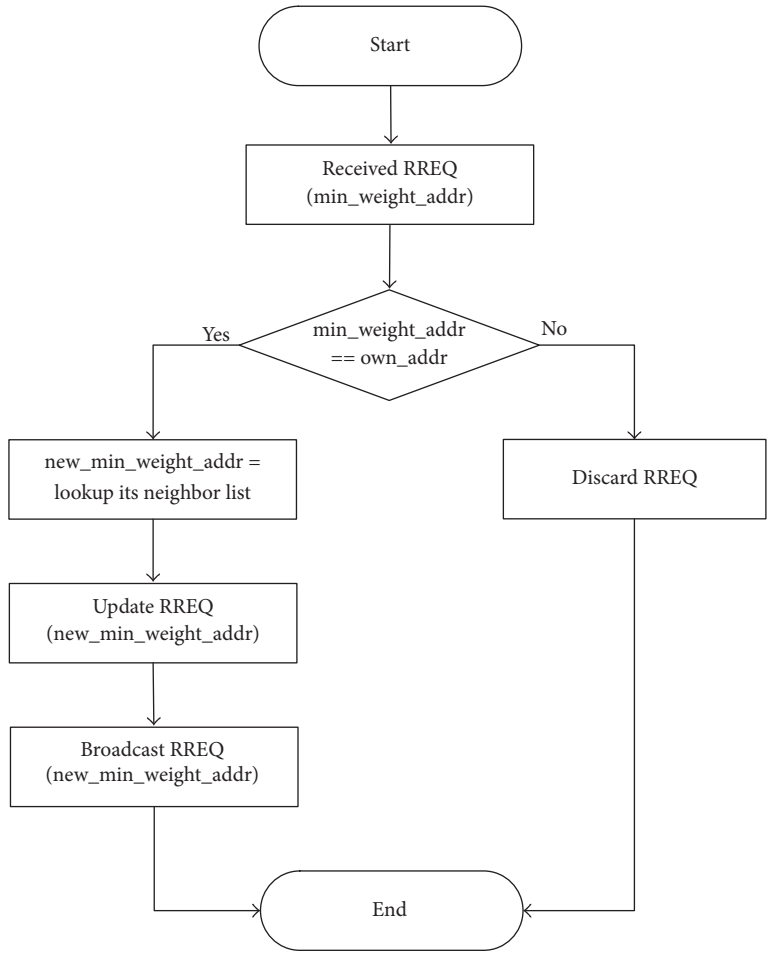


Figure 5: Handling RREQ message in an intermediate node.

Table 2: List of route discovery latency.

Notation	Description	Time (ms)
T_{ADV}	Advertisement interval	30
T_{IFS}	Inter-frame space	0.15
T_{ROLE_SW}	Role switch time	1.25
T_{CONN_SETUP}	Connection setup time	30
T_{POST_CONN}	Mandatory delay after connection request	1.25
T_{REP}	Transmission time of RREP message	0.296

the weight values for the route discovery were calculated by changing alpha values (0.3, 0.5, 0.7, and 1.0) in (1). The route discovery was performed every second after choosing a source and a destination randomly using Poisson process. Thus, the route discovery is assumed to occur frequently.

We adopted WINNER-II path loss model [23] for RSSI calculation between two nodes and the RSSI value was used to compute the node's weight mentioned in Section 4. For each BLE node, we collected 300 records of data. For the performance evaluation, a simulator using Python and NetworkX [24] was implemented to gauge the operation of the proposed scheme, AODV, and TORA protocol. The route paths and network topology are illustrated in Figure 8.

In this simulation, we adopted B3 scenario of WINNER-II path loss model to calculate the RSSI value between two nodes within the radio coverage. The scenario is suitable for large indoor structures such as airport, factories, and stations sized from 20×20 m to 100×100 m. Equation (2) shows the calculation of path loss in WINNER-II model and parameters are fixed in accordance with the scenario selected. The path loss exponent is A, while B acts as the intercept, C is the frequency dependent parameter, d is a distance, fc is the

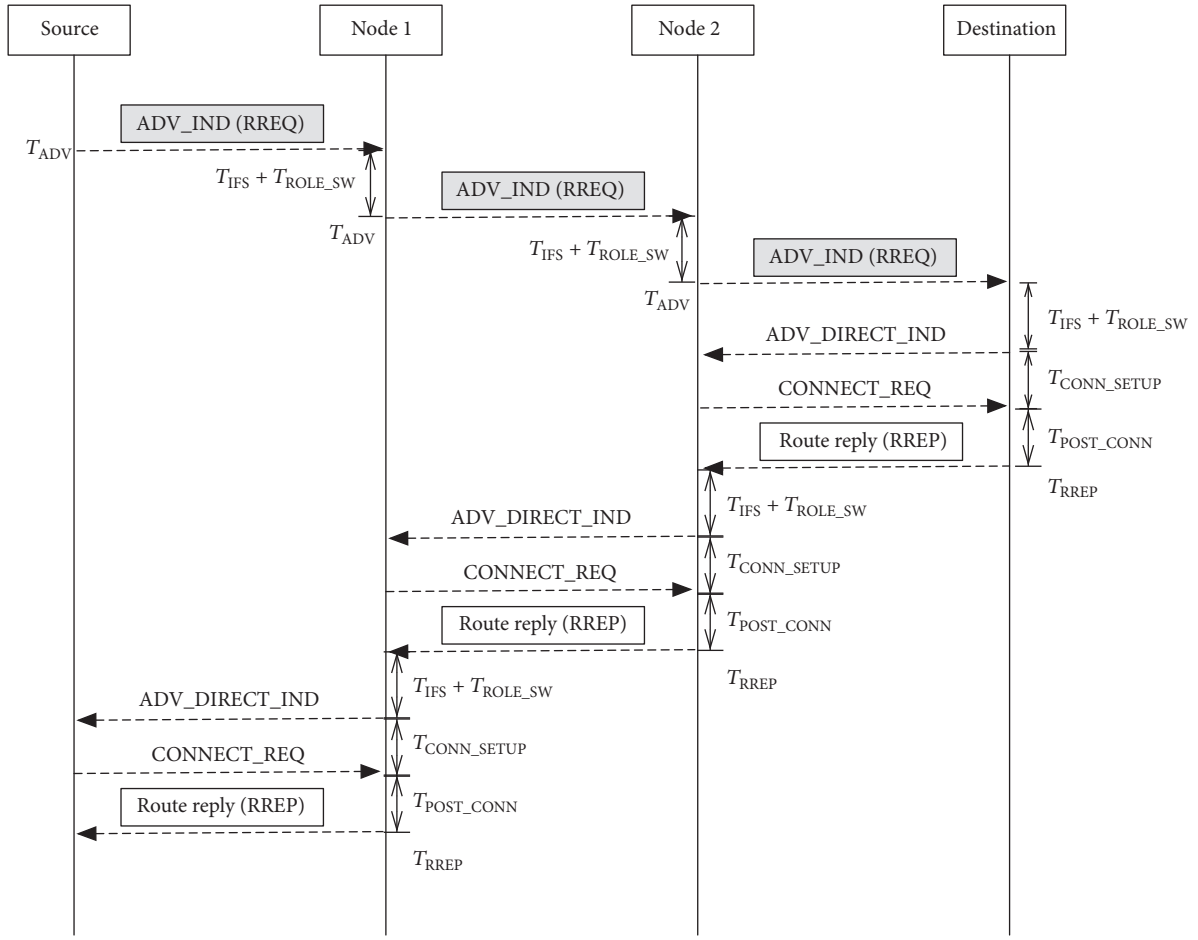
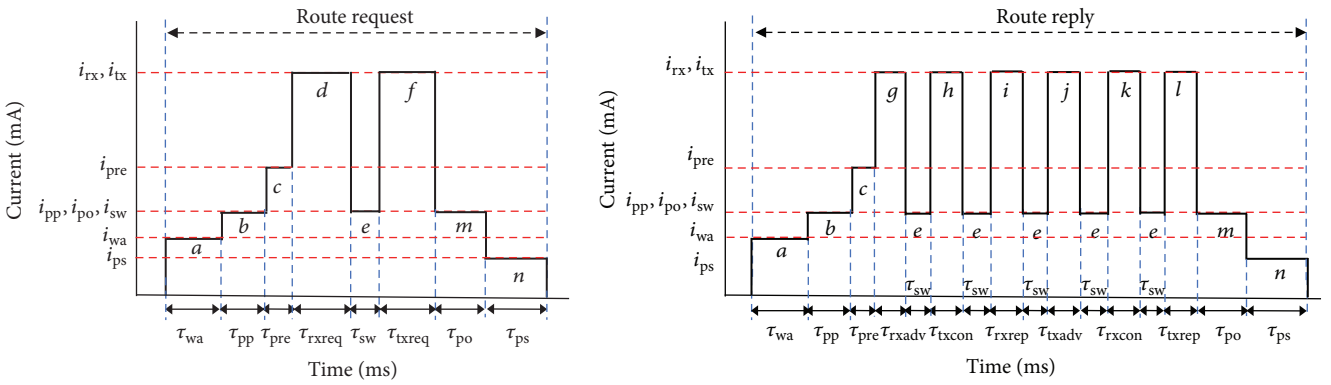


Figure 6: Message flows of route discovery procedure.



(a) Current and time of route request procedure

(b) Current and time of route reply procedure

Figure 7: Current waveform of an intermediate node.

system frequency of BLE, and X is an optional parameter. These parameters are set to $A = 13.9$, $B = 64.4$, $C = 20$, $fc = 2.4$, and $X = 0$ in the B3 Line of Sight (LOS) scenario. Further, the transmission power of each BLE device was fixed at 4 dBm in the simulation. As the path loss model has a correlation with the distance between two nodes, RSSI values can be applied to the proposed scheme. Thus, the proposed routing scheme considered the distance as well as the residual power.

$$\mathrm{PL} = A\log_{10}(d\,[\mathrm{m}]) + B + C\log_{10}\left(\frac{f_c\,[\mathrm{GHz}]}{5.0}\right) + X, \tag{2}$$

$$\mathrm{RSSI} = Tx\mathrm{Power} - \mathrm{PL}.$$

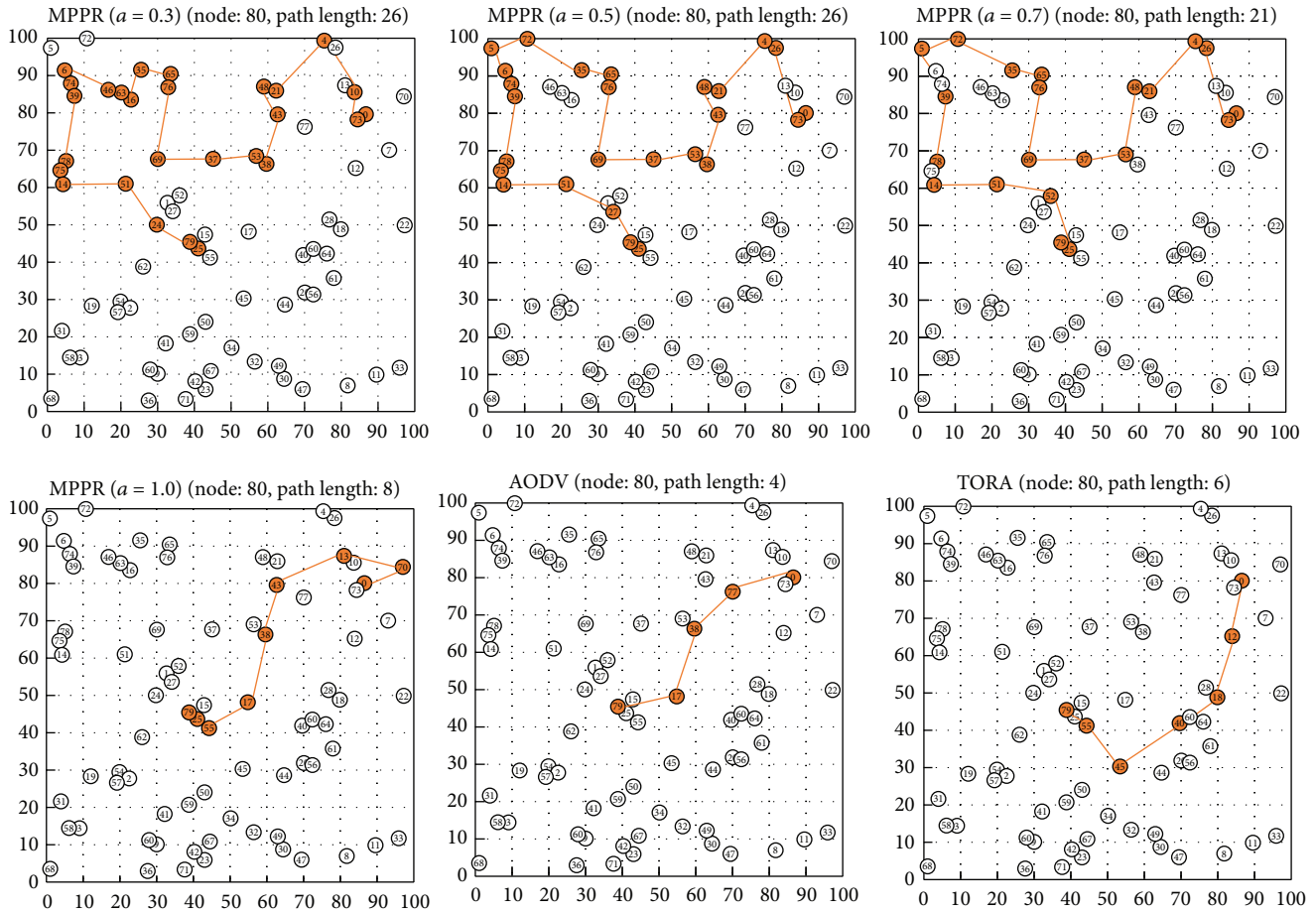


Figure 8: Network topology and the comparison of route paths.

Table 3: Time and current consumption of an intermediate node.

Category	State	Phase	Time notation	Current notation
Common	a	Wake-up	τ_{wa}	i_{wa}
	b	Preprocessing	τ_{pp}	i_{pp}
	c	Pre-Tx/Rx	τ_{pre}	i_{pre}
	e	Tx-to-Rx	τ_{sw}	i_{sw}
	m	Postprocessing	τ_{po}	i_{po}
	n	Pre-sleep	τ_{ps}	i_{ps}
Route request	d	Rx (RREQ)	τ_{rxreq}	i_{rx}
	f	Tx (RREQ)	τ_{txreq}	i_{tx}
Route reply	g	Rx (ADV_DIRECT_IND)	τ_{rxadv}	i_{rx}
	h	Tx (CONNECT_REQ)	τ_{txcon}	i_{tx}
	i	Rx (RREP)	τ_{rxrep}	i_{rx}
	j	Tx (ADV_DIRECT_IND)	τ_{txadv}	i_{tx}
	k	Rx (CONNECT_REQ)	τ_{rxcon}	i_{rx}
	l	Tx (RREP)	τ_{txrep}	i_{tx}

Table 4: Average current consumption of each phase in a route path.

Phase	Source (mA)	Intermediate (mA)	Destination (mA)	Neighbor (mA)
Route request	0.047	0.069	0.047	0.047
Route reply	0.032	0.048	0.032	None
Data transmission	0.02	0.025	0.02	None

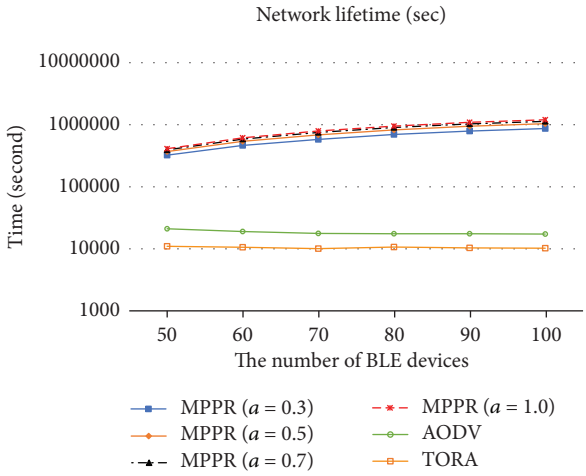


Figure 9: Network lifetime.

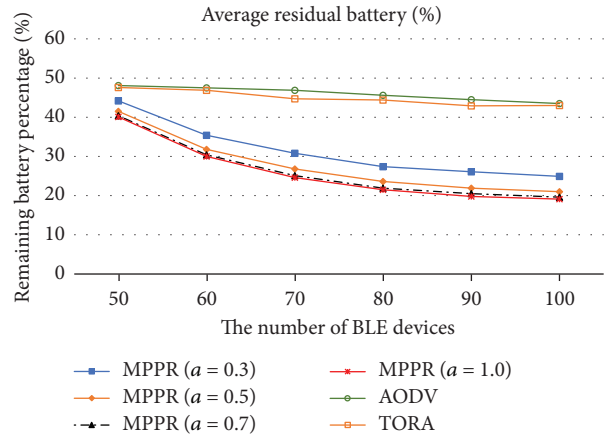


Figure 10: The percentage of residual battery.

Table 4 shows the energy consumption of each phase along the route path. At the time of the route request process, a neighboring node of the route path exhausts energy. This is because the neighbor nodes do not forward route request messages when the minimum weight node address in a route request message is not equivalent to its address.

5.2. Network Lifetime. The period when the first node is depleted of its energy is defined as the network lifetime. We have simulated the network lifetime by applying different priorities between residual battery and RSSI by changing α (α = 0.3, 0.5, 0.7, 1.0) in (1). A source and a destination are chosen randomly and route request is generated by Poisson process. After successful route discovery, a 20-byte data packet was delivered. When it comes to the energy consumption in each node, whenever a RREQ or RREP message is acquired by a node during route discovery procedure, node's battery level is decreased by the calculated level as shown in Table 4. If a node's battery capacity becomes zero in the network, this simulation stops and calculates the lifetime.

The proposed scheme has the longest network lifetime when α is 1 (α = 1.0) as illustrated in Figure 9. When route decision is made only by the residual battery during route discovery procedure, the performance is at its best. In addition, the result of the proposed scheme shows that the lifetime is improved as the number of BLE devices increases regardless of the value of α. The increment of lifetime explains that the proposed scheme consumes the energy of BLE devices evenly.

However, AODV and TORA protocols show the gradual decline of the network lifetime when the number of BLE devices is increasing. The network lifetime is reduced in AODV protocol when the number of BLE devices grows, due to its RREQ broadcasting mechanism in AODV. Moreover, route request messages are concentrated on specific BLE devices and their energies are depleted rapidly in AODV protocol. In TORA's case, the network life time is the lowest among other protocols because it adopts the broadcasting scheme during the route request and route reply procedure using QTY and UPD packets, respectively. Therefore, TORA consumes more energy during the route discovery procedure than the proposed scheme and AODV.

5.3. Average Residual Battery. Average residual battery is measured when the network lifetime simulation is done. We have calculated the average residual battery level of all nodes. As described in Figure 10, the residual battery of the proposed scheme (α = 1.0) has the minimum battery level at approximately 20%. The residual battery of AODV and TORA protocols is bigger even though their network lifetime is much shorter than the proposed scheme as shown in Figure 9.

On the contrary, the proposed scheme has a longer lifetime with a remaining battery around 18%. The simulation result demonstrates that the proposed scheme has consumed energy in a distributed way compared with AODV and TORA protocols.

5.4. Power Consumption for Route Discovery. During the route discovery process, power consumption (P_{route}) is computed using (3). Energy consumption for each phase is mentioned in Section 4.4. A source node consumes energy (P_{source}) to broadcast a RREQ packet using BLE advertisement channels. The energy consumption of neighboring nodes (P_{neighbor}) occurs when they receive RREQ messages

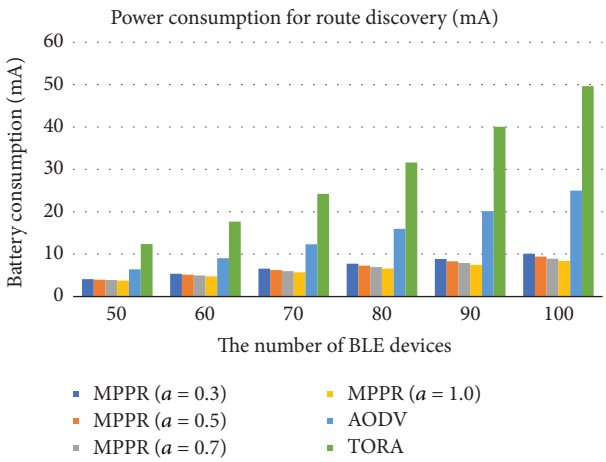


Figure 11: Energy consumption for route discovery.

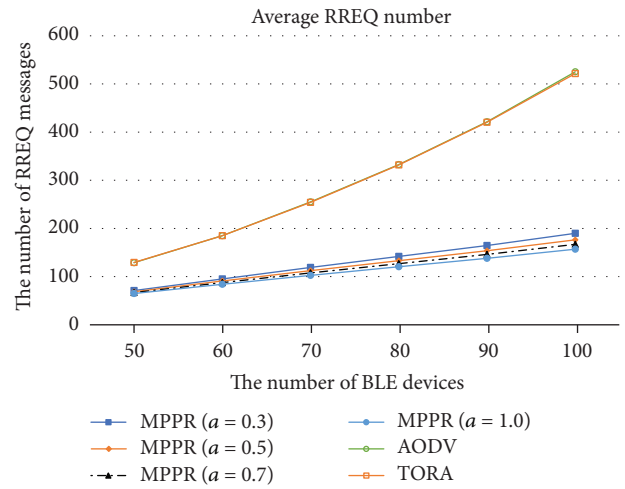


Figure 12: Average RREQ number.

from their neighbors. The neighboring nodes do not broadcast RREQ messages unless they have the minimum weight value among the neighbors. Intermediate nodes along the route path consume more energy (P_{interm}) to broadcast RREQ messages and receive RREP messages. The energy consumption of a destination node (P_{dest}) takes place when it receives a RREQ message and responds with a RREP message via a unicast connection.

$$\begin{aligned} P_{\text{source}} &= P_{\text{RREQ\_TX}} + P_{\text{RREP\_RX}} + P_{\text{Data\_TX}}, \\ P_{\text{interm}} &= \left(P_{\text{RREQ\_RX}} + P_{\text{RREQ\_TX}}\right) \\ &\quad + \left(P_{\text{RREP\_RX}} + P_{\text{RREP\_TX}}\right) \\ &\quad + \left(P_{\text{DATA\_RX}} + P_{\text{DATA\_TX}}\right), \\ P_{\text{dest}} &= P_{\text{RREQ\_RX}} + P_{\text{RREP\_TX}} + P_{\text{Data\_RX}}, \\ P_{\text{route}} &= P_{\text{source}} + \left[\left(N_{\text{hop}} - 1\right) \times P_{\text{interm}}\right] \\ &\quad + \left[\left(N_{\text{RREQ}} - N_{\text{hop}}\right) \times P_{\text{neighbor}}\right] + P_{\text{dest}}. \end{aligned} \tag{3}$$

Figure 11 clearly presents that TORA protocols consume the highest energy compared to the proposed scheme and AODV due to the QRY and UPD flooding mechanism. AODV only broadcasts RREQ messages during the route request procedure but TORA adopts the broadcast mechanism for the route request and reply procedure. Therefore, when the number of nodes increases, TORA consumes more energy during the route discovery procedure. Contrastingly, the proposed scheme shows that its energy consumption decreases when the value α is changed from 0.3 to 1.

5.5. Average Number of Route Request Messages. The average number of RREQ messages (N_{RREQ}) is determined during the network lifetime simulation. A flooding approach of AODV and TORA is used to disseminate route request messages over the advertisement interval. For instance, the average route requests number is $(N - 1) \times 2$ when N numbers of BLE nodes are distributed and connected in AODV and TORA. Contrarily, route request messages are transmitted to nearby nodes and only a node with the minimum weight among neighboring nodes forwards a RREQ message to find a route path in the proposed scheme. Equation (4) describes the way of calculating RREQ numbers of a route path. Hence, AODV and TORA protocols have greater values than the proposed scheme in the average number of RREQ messages and show similar result because of the broadcasting approach for their route request procedure. The steady progress of route request messages with regard to the surge of BLE devices of the proposed routing protocol is described in Figure 12.

$$N_{\text{RREQ}} = \sum_{i=\text{source}}^{\text{dest}-1} \text{Number of Route}_i\text{'s neighbor.} \tag{4}$$

5.6. Route Discovery Latency. The elapsed time between delivering RREQ and receiving RREP at a source is defined as route discovery latency. The latency was calculated using (5) for the proposed scheme, AODV, and TORA protocol. We assumed that a RREQ message is delivered during the advertisement period ($T_{\text{ADV}} = 30$ ms) and the role switch time ($T_{\text{ROLE\_SW}}$) is 1.25 ms. The number of RREQ messages and the route path length are represented as N_{RREQ} and L_{PATH}, respectively. TORA broadcasts UPD packets for the route reply procedure so we consider the delivery time of UPD packets ($T_{\text{TORA\_UPD}}$) as described in (5). We also assumed that an UPD packet of TORA is delivered during the advertisement period because a route is not established unlike the proposed scheme. Other parameters were already mentioned in Table 2.

$$\begin{aligned} T_{\text{RREQ}} &= \left(T_{\text{ADV}} + T_{\text{IFS}} + T_{\text{ROLE\_SW}}\right) \times N_{\text{RREQ}}, \\ T_{\text{RREP}} &= \left(T_{\text{CONN\_SETUP}} + T_{\text{IFS}} + T_{\text{ROLE\_SW}} + T_{\text{POST\_CON}}\right. \\ &\quad \left. + T_{\text{REP}}\right) \times L_{\text{PATH}}, \\ T_{\text{TORA\_UPD}} &= \left(T_{\text{ADV}} + T_{\text{IFS}} + T_{\text{ROLE\_SW}}\right) \times N_{\text{UPD}}. \end{aligned} \tag{5}$$

As shown in Figure 13, the results clearly show the steady increase of the proposed scheme, while AODV and TORA

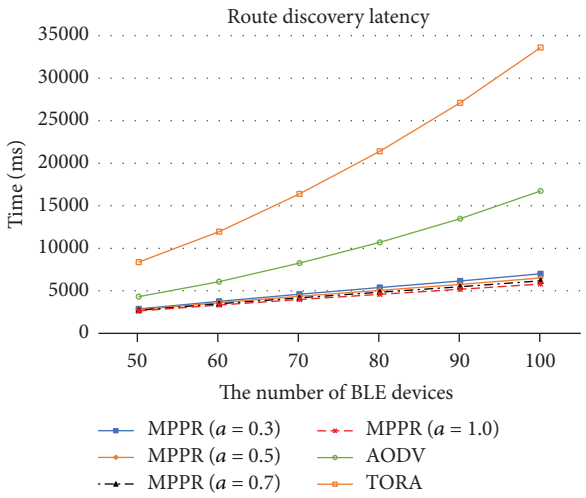


Figure 13: Route discovery latency.

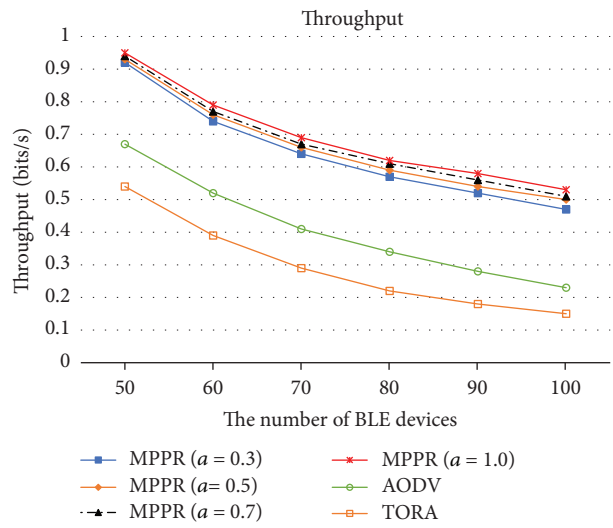


Figure 14: Throughput.

protocols follow an exponential growth as the number of BLE nodes increases. In particular, TORA shows a higher exponential growth because of the broadcast scheme during the route request and reply procedure. This exponential growth in the latency resulted from the increasing amount of RREQ and UPD messages.

5.7. Average Throughput. The average throughput is calculated by the ratio of the data packet size delivered over the route discovery latency plus packet delivery time ($T_{\text{DATA}i}$) during the network lifetime as depicted in (6). The total number of transmitted data is represented as N during the network lifetime of each protocol.

$$\text{Average throughput} = \frac{1}{N}\sum_{i=1}^{n}\left(\frac{\text{Data size}}{T_{\text{RREQ}i} + T_{\text{RREP}i} + T_{\text{DATA}i}}\right). \quad (6)$$

The average throughput decreases as the number of BLE device increases due to the augmented route discovery latency in the proposed scheme, AODV, and TORA. Specifically, the lowest throughput of TORA is caused by the increased route discovery time. The average throughput of AODV is also affected by the broadcast approach of route request message. The proposed scheme shows that the average throughputs have no big differences related to α value as described in Figure 14.

6. Conclusions

We proposed an energy-conserving multihop routing protocol based on maximum power and RSSI in BLE ad hoc networks. This study is designed to avoid the flooding mechanism that causes substantial energy consumption. Similarly, the proposed scheme focused on magnifying network lifespan under power-constrained ad hoc networks by using the residual battery level and RSSI as a route metric. In addition, the energy-conserving routing protocol minimizes the number of route request messages which are solely forwarded by the minimum weighted node. Therefore, the outcome of the minimized number of route request messages causes less power exhaustion and shorter route discovery latency compared with the typical ad hoc routing protocol.

The performance evaluation revealed that our suggested energy-conserving protocol improves performances with regard to network lifetime, average residual battery, power usage, route discovery latency, and the volume of route request messages under power-restricted BLE ad hoc networks. Furthermore, the battery consumption of the proposed scheme is evenly distributed during route discovery procedure and it results in less residual battery level after a longer network lifetime.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study 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, Korea (21A20131600005), by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1B03933566), and by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (no. 2017-0-00770).

References

[1] *Bluetooth SIG, Bluetooth Core Specification, Version 4.1*, December 2013.

[2] K.-H. Chang, "Bluetooth: a viable solution for IoT? [Industry Perspectives]," *IEEE Wireless Communications Magazine*, vol. 21, no. 6, pp. 6-7, 2014.

[3] J. Nieminen, C. Gomez, M. Isomaki et al., "Networking solutions for connecting bluetooth low energy enabled machines to the internet of things," *IEEE Network*, vol. 28, no. 6, pp. 83–90, 2014.

[4] D.-W. Kum, A.-N. Le, Y.-Z. Cho, C. K. Toh, and I.-S. Lee, "An efficient on-demand routing approach with: Directional flooding for wireless mesh networks," *Journal of Communications and Networks*, vol. 12, no. 1, pp. 67–73, 2010.

[5] Z. Guo, I. G. Harris, L.-F. Tsaur, and X. Chen, "An on-demand scatternet formation and multi-hop routing protocol for BLE-based wireless sensor networks," in *Proceedings of the 2015 IEEE Wireless Communications and Networking Conference, WCNC 2015*, pp. 1590–1595, March 2015.

[6] C. Perkins, E. Belding-Royer, and S. Das, *Ad hoc On-Demand Distance Vector (AODV) Routing*, RFC 3561, July 2003, http://www.rfc-editor.org/info/rfc3561.

[7] S. Sharafeddine, I. Al-Kassem, and Z. Dawy, "A scatternet formation algorithm for Bluetooth networks with a non-uniform distribution of devices," *Journal of Network and Computer Applications*, vol. 35, no. 2, pp. 644–656, 2012.

[8] O. Al-Jarrah and O. Megdadi, "Enhanced AODV routing protocol for Bluetooth scatternet," *Computers and Electrical Engineering*, vol. 35, no. 1, pp. 197–208, 2009.

[9] V. Park and S. Corson, *Temporally-ordered routing algorithm (TORA) Version1 Functional Specification, draft-ietf-manet-tora-spec-04.txt*, July 2001, https://tools.ietf.org/html/draft-ietf-manet-tora-spec-04.

[10] C. Yu and Y. Yu, "Joint Layer-Based Formation and Self-Routing Algorithm for Bluetooth Multihop Networks," *IEEE Systems Journal*, pp. 1–11.

[11] C.-M. Yu and J.-H. Lin, "Enhanced Bluetree: A mesh topology approach forming Bluetooth scatternet," *IET Wireless Sensor Systems*, vol. 2, no. 4, pp. 409–415, 2012.

[12] A. M. E. Ejmaa, S. Subramaniam, Z. A. Zukarnain, and Z. M. Hanapi, "Neighbor-Based Dynamic Connectivity Factor Routing Protocol for Mobile Ad Hoc Network," *IEEE Access*, vol. 4, pp. 8053–8064, 2016.

[13] C.-M. Yu and Y.-B. Yu, "Reconfigurable algorithm for bluetooth sensor networks," *IEEE Sensors Journal*, vol. 14, no. 10, pp. 3506-3507, 2014.

[14] A. Jedda and H. T. Mouftah, "Forming MS-Free and Outdegree-Limited Bluetooth Scatternets in Pessimistic Environments," *IEEE Internet of Things Journal*, vol. 2, no. 6, pp. 538–550, 2015.

[15] C. Yu and Y. Lee, "A Reconfigurable Formation and Disjoint Hierarchical Routing for Rechargeable Bluetooth Networks," *Energies*, vol. 9, no. 5, p. 338, 2016.

[16] C.-K. Toh, "Maximum battery life routing to support ubiquitous mobile computing in wireless ad hoc networks," *IEEE Communications Magazine*, vol. 39, no. 6, pp. 138–147, 2001.

[17] D. Kim, J. J. Garcia-Luna-Aceves, K. Obraczka, J.-C. Cano, and P. Manzoni, "Routing mechanisms for mobile ad hoc networks based on the energy drain rate," *IEEE Transactions on Mobile Computing*, vol. 2, no. 2, pp. 161–173, 2003.

[18] X. Zhang and G. F. Riley, "Energy-aware on-demand scatternet formation and routing for bluetooth-based wireless sensor networks," *IEEE Communications Magazine*, vol. 43, no. 7, pp. 126–133, 2005.

[19] T. Nadeem, S. Banerjee, A. Misra, and A. Agrawala, "Energy-efficient reliable paths for on-demand routing protocols," *IFIP Advances in Information and Communication Technology*, vol. 162, pp. 485–496, 2005.

[20] P. Ramachandran and M. Dinakaran, "Signal Strength and Residual Power Based Optimum Transmission Power Routing for Mobile Ad hoc Networks," in *Proceedings of the 2nd International Conference on Intelligent Computing, Communication and Convergence, ICCC 2016*, pp. 168–174, India, January 2016.

[21] Y. Liu, M. J. Lee, and T. N. Saadawi, "A Bluetooth scatternet-route structure for multihop ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 2, pp. 229–239, 2003.

[22] S. Kamath and J. Lindh, *Measuring Bluetooth Low Energy Power Consumption*, Application Note AN092, http://www.ti.com.cn/cn/lit/an/swra347a/swra347a.pdf.

[23] P. Kyösti et al., *WINNER II channel models*, IST-4-027756 WINNER II Deliverable D1.1.2, V1.2.4.2, 2008, http://www.ist-winner.org/deliverables.html.

[24] "NetworkX website," http://networkx.github.io/index.html.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 9628565, 10 pages
https://doi.org/10.1155/2017/9628565

Research Article

Performance Analysis of Three-Dimensional Clustered Device-to-Device Networks for Internet of Things

Haejoon Jung[1] and In-Ho Lee[2]

[1]*Department of Information and Telecommunication Engineering, Incheon National University, Incheon 22012, Republic of Korea*
[2]*Department of Electrical, Electronic and Control Engineering, Hankyong National University, Anseong 17579, Republic of Korea*

Correspondence should be addressed to In-Ho Lee; ihlee@hknu.ac.kr

Received 27 April 2017; Accepted 10 September 2017; Published 13 November 2017

Academic Editor: Haiyu Huang

Internet of things (IoT) is a smart technology that connects anything anywhere at any time. Intelligent device-to-device (D2D) communication, in which devices will communicate with each other autonomously without any centralized control, is an integral part of the Internet of Things (IoT) ecosystem. Thus, for D2D applications such as local file sharing or swarm sensing, we study communications between devices in proximity in ultra-dense urban environments, where devices are stacked vertically and dispersed in the horizontal plane. To reflect the spatiotemporal correlation inherently embedded in the D2D communications, we model and analyze clustered D2D networks in three-dimensional (3D) space based on Thomas cluster process (TCP), where the locations of clusters follow Poisson point process, and cluster members (devices) are normally distributed around their cluster centers. We assume that multiple device pairs in the network can share the same frequency band simultaneously. Thus, in the presence of cochannel interference from both the same cluster and the other clusters, we investigate the coverage probability and the area spectral efficiency of the clustered D2D networks in 3D space.

1. Introduction

Fifth-generation (5G) networks are being developed to support dramatically increasing data traffic with various multimedia applications [1]. As more devices are embraced to connect everything, everywhere, and everyone, networks become dense with unprecedented rise of mobile traffic. In this context, device-to-device (D2D) communication, which relieves the burden of base stations (BSs), is an important feature for various types of mobile networks in the future cellular systems [2–4]. Through the D2D communication, wireless devices can constantly interact to each other as well as with their environments, which is the key 5G enabler for the Internet of Things (IoT) [5–7]. The D2D communications to create, gather, and share information involve various types of devices such as sensors, smartphones, cars, health care gadgets, and home appliances [8].

Motivated by such emerging applications of the D2D communications, in this paper, we model and analyze D2D networks in three-dimensional (3D) space based on stochastic geometry [9]. To be specific, we consider 3D multicluster D2D networks, where devices in close proximity form a clustered network architecture. Poisson point process (PPP) is a widely used to analyze various types of networks (e.g., [10, 11]) including D2D networks, for its mathematical tractability. However, it cannot capture the fact that a device typically has multiple proximate devices, any of which is a potential serving device, with correlation in space and time.

To overcome this limitation, the authors in [12] develop a more realistic model for two-dimensional (2D) D2D networks, where the devices locations are modeled as a Poisson cluster process, in particular a variant of a Thomas cluster process [9], where the D2D network consists of multiple clusters, and cluster members (devices) are normally distributed around the center of clusters. Different from the widely used uniform spatial distribution assumption with PPP as in [13], the model proposed in [12] reflects the spatiotemporal correlation in the content demand in D2D networks in the IoT environments as indicated in [14, 15]. Using this model, they investigate 2D clustered D2D networks for local information sharing with each cluster [16–18].

However, as highlighted in [19–22], a 2D space model assumed in [12] may not be suitable for dense urban environments with high-rise buildings, where both devices and small-cell BSs are distributed over the 3D space. In [12], the coverage probability of wireless networks has been studied for various 2D deployment scenarios without much consideration for the vertical component of node distributions. However, to better model the future wireless environments (especially for the IoT applications) with ultra-dense deployments of devices and BSs, we need to consider the spatial distribution in the vertical space as well as the horizontal plane, as noted in [19–22]. For this reason, we extend the analytic framework of [12] in 2D space (on the horizontal plane) into 3D space. To our knowledge, this is the first study to model 3D D2D networks using TCP.

The contributions of this paper are fourfold. First, we derive the probability distributions of distance between two devices that belong to (i) the same cluster and (ii) two different clusters in the 3D space. Second, we provide the exact mathematical expressions of the coverage probability and the area spectral efficiency of the 3D clustered D2D networks. Third, the approximate upper and lower bounds of the coverage probability are obtained, which are useful in the coverage analysis to gain insights into system design guidelines. Moreover, we present numerical and simulation results to validate our analysis and compare the 2D and 3D TCP models with various system parameters.

2. System Model

We consider a D2D network in 3D space, where the devices participating communications exist in clusters by the nature of D2D communications [12]. We assume that each device communicates with other devices in the same cluster, while the devices across clusters do not communicate directly (or, the intercluster communications may use orthogonal channels). As shown in Figure 1, the locations of the devices in 3D space are modeled by a TCP, where the cluster centers follow a homogeneous PPP Φ_c with density λ_c. Also, the cluster members (devices) are independent and identically distributed (i.i.d.) according to a symmetric normal distribution with variance σ^2 around each cluster center $x \in \Phi_c$ with the density function of the device locations $y \in \mathbb{R}^3$ relative to a cluster center as

$$f_Y(y) = \frac{1}{(2\pi)^{3/2}\sigma^3} \exp\left(-\frac{\|y\|^2}{2\sigma^2}\right), \tag{1}$$

where σ is the scattering parameter.

The devices in the cluster of $x \in \Phi_c$ are denoted by $\mathcal{N}^x$, which has two subsets: (i) transmitting devices $\mathcal{N}_t^x$ and (ii) receiving devices $\mathcal{N}_r^x$. Suppose the set of simultaneously transmitting devices in the cluster is $\mathcal{B}^x \subseteq \mathcal{N}_t^x$, and its cardinality $|\mathcal{B}^x|$ follows a Poisson distribution with mean λ_t. In other words, the number of simultaneously active transmitting devices (Dev-Txs) inside each cluster is a Poisson random variable (RV) with mean λ_t. Therefore, excluding the serving (or desired) Dev-Tx, we assume that the number of interfering devices follows a Poisson distribution with mean

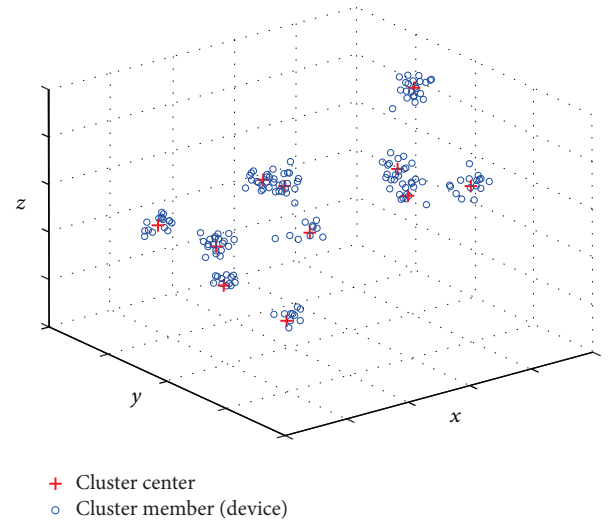


Figure 1: An example illustration of a three-dimensional clustered D2D network based on TCP.

$(\lambda_t - 1)$. As in [12], without loss of generality, we perform analysis based on a typical device in a representative cluster $x_0 \in \Phi_c$, where the typical device is regarded as the device receiver of interest. We assume that the typical device is located at the origin.

We assume that the serving Dev-Tx is located at y_0 inside the cluster $x_0 \in \Phi_c$. Thus, the distance between the serving Dev-Tx and the typical device is denoted by $r = \|y_0 + x_0\|$. Hence, with the transmit power of each device denoted by P_0, the received power at the typical device is

$$S = P_0 h_0 r^{-\alpha} = \frac{P_0 h_0}{\|x_0 + y_0\|^\alpha}, \tag{2}$$

where α is the path-loss exponent and h_0 is the power gain of small scale fading channel, which follows exponential distribution with unit mean, as in [12, 19–21]. The typical device suffers from two types of cochannel interference: (i) intracluster interference caused by the simultaneously active Dev-Txs in the same cluster and (ii) intercluster interference caused by the Dev-Txs in the other clusters, which are represented as

$$I_{\text{intra}} = \sum_{y \in \mathcal{B}^{x_0} \setminus y_0} \frac{P_0 h_{y_{x_0}}}{\|x_0 + y\|^\alpha}, \tag{3}$$

$$I_{\text{inter}} = \sum_{x \in \Phi_c \setminus x_0} \sum_{y \in \mathcal{B}^x} \frac{P_0 h_{y_x}}{\|x + y\|^\alpha}, \tag{4}$$

respectively. Consequently, assuming interference-limited networks, the signal-to-interference-ratio (SIR) at the typical device is

$$\begin{aligned} \mathtt{SIR}(r) &= \frac{S}{I_{\text{intra}} + I_{\text{inter}}} \\ &= \frac{h_0 / \|x_0 + y_0\|^\alpha}{\sum_{y \in \mathcal{B}^{x_0} \setminus y_0} \left(h_{y_{x_0}} / \|x_0 + y\|^\alpha\right) + \sum_{x \in \Phi_c \setminus x_0} \sum_{y \in \mathcal{B}^x} \left(h_{y_x} / \|x + y\|^\alpha\right)}, \end{aligned} \tag{5}$$

where P_0 is canceled, since we assume the fixed transmit power for all Dev-Txs.

3. Distance Distributions

In this section, we derive the probability distributions of the distances from the typical device to intra- and intercluster devices for system performance analysis associated with SIR. We assume that the content of interest for a typical device in a given cluster is available at a device chosen uniformly at random in the cluster, as in [12]. Based on this assumption, we derive the distance distributions from the typical device to the serving Dev-Tx, intra- and intercluster interferers.

3.1. Distances between Typical Device and Intracluster Dev-Txs. For the intracluster devices, let $\mathscr{D}_t^{x_0}$ be the set $\{D_i\}_{i=1:|\mathscr{N}_t^{x_0}|}$ of distances from the typical device to the set of possible Dev-Txs $\mathscr{N}_t^{x_0}$ in the cluster $x_0 \in \Phi_c$, where $d_i = \|x_0 + y\|$ is the realization of D_i. We note that the index i will be omitted when it is clear from the context. To delve into the distance statistics of D2D links, we first derive the probability distribution function (PDF) of the distance $v_0 = \|x_0\|$ between the cluster center x_0 and the typical device at the origin. Then, using this result, the PDF of the separation between the intracluster Dev-Tx and the typical device will be derived.

Lemma 1 (probability distribution of $v_0 = \|x_0\|$). *The PDF of v_0 is given by*

$$f_{v_0}(x) = \sqrt{\frac{2}{\pi}}\frac{x^2}{\sigma^3}\exp\left(-\frac{x^2}{2\sigma^2}\right), \tag{6}$$

where $x \geq 0$.

Proof. Based on the 3D Gaussian distribution defined in (1), $Z = v_0^2/\sigma^2$ is the squared sum of three i.i.d. standard (zero mean and unit variance) Gaussian random variables, which corresponds to the PDF as

$$f_Z(z) = \sqrt{\frac{z}{2\pi}}e^{-z/2}, \tag{7}$$

where $z \geq 0$. Therefore, by the change of variables, we can obtain the PDF in (6). □

Lemma 2 (probability distribution of $D = \|x_0 + y\|$). *The PDF of the separation between the typical device and the Dev-Tx in the same cluster is given by*

$$f_D(d) = \frac{d^2}{2\sqrt{\pi}\sigma^3}\exp\left(-\frac{d^2}{4\sigma^2}\right), \tag{8}$$

where $d \geq 0$.

Proof. The locations of the cluster center x_0 and the Dev-Txs y are i.i.d. random vectors in $\mathbb{R}^3$, where the three components follow i.i.d. Gaussian distributions with zero mean and variance of σ^2. Suppose $Z = \|x_0 + y\|^2/2\sigma^2$, which is the squared sum of three i.i.d standard Gaussian random variables. Thus, Z follows a chi-squared distribution with 3 degrees of freedom with the PDF:

$$f_Z(z) = \sqrt{\frac{z}{2\pi}}e^{-z/2}. \tag{9}$$

Therefore, the PDF of $D = \sigma\sqrt{2Z}$ in (8) can be derived by the change of variables (it is noted that the PDF and conditional PDF of D are, resp., obtained by extending the probability distribution analysis in 2D to 3D space). □

3.2. Conditional Distribution of D Given $\|x_0\|$. The distances of the typical device to the Dev-Txs in the same clusters, which are required to calculate S and I_{intra} in SIR, are correlated because of the common factor x_0. Therefore, conditioning the relative location of the cluster center, x_0, to typical device, we can treat the locations of the intracluster devices as i.i.d. RVs, which means that the distances between the typical device and the intracluster devices are i.i.d. To exploit this property, the following lemma gives the conditional distribution of D given $\|x_0\|$.

Lemma 3 (conditional probability distribution of $D = \|x_0 + y\|$ given $\|x_0\|$). *The conditional PDF of D for a given $\|x_0\|$ is derived as*

$$f_D(d \mid v_0) = \frac{d^{3/2}}{\sigma^2\sqrt{v_0}}e^{-(v_0^2+d^2)/2\sigma^2}I_{1/2}\left(\frac{v_0 d}{\sigma^2}\right), \tag{10}$$

where $d \geq 0$ and $I_{1/2}(t) = \sum_{k=0}^{\infty}(1/k!\Gamma(t+k+1))(t/2)^{1/2+2k}$ is the modified Bessel function with order $1/2$.

Proof. Let $Z = \|x_0 + y\|^2/\sigma^2$. Because $\|y\|^2/\sigma^2$ is the squared sum of three i.i.d standard Gaussian RVs, conditioned on $v_0 = \|x_0\|$, Z follows a noncentral chi-square distribution with the PDF:

$$f_Z(z \mid v_0) = \frac{1}{2}\left(\frac{\sigma^2 z}{v_0^2}\right)^{1/4}e^{-(v_0^2+\sigma^2 z)/2\sigma^2}I_{1/2}\left(\frac{v_0\sqrt{z}}{\sigma}\right). \tag{11}$$

Since $D = \|x_0 + y\| = \sigma\sqrt{Z}$, its PDF in (10) can be obtained by the change of variables (it is noted that the PDF and conditional PDF of D are, resp., obtained by extending the probability distribution analysis in 2D to 3D space). □

3.3. Distances to Serving Dev-Tx and Interferers: r, w, and u. Let the distances from the typical device to the serving Dev-Tx and intracluster interferer be $r = \|x_0 + y_0\|$ and $w = \|x_0 + y\|$, respectively. Their conditional PDFs given that $v_0 = \|x_0\|$ are same as (10). In other words, $f_R(r \mid v_0) = f_D(r \mid v_0)$ and $f_W(w \mid v_0) = f_D(w \mid v_0)$. In addition, conditioned on the distance $v = \|x\|$ between one of the other clusters $x \in \Phi_c$ and the typical device, the distances $\{u = \|x + y\|,\ \forall y \in \mathscr{B}^x\}$ between the typical device and the intercluster interfering Dev-Txs in $x \in \Phi_c$ are i.i.d., following the conditional PDF $f_U(u \mid v) = f_D(u \mid v_0 = v)$ given in (10). Also, the PDF of $v = \|x\|$ is identical to the PDF of $v_0 = \|x_0\|$ defined in (6).

4. Performance Analysis: $\mathtt{P_c}$ and ASE

In this section, we investigate the coverage probability, $\mathtt{P_c}$, and the area spectral efficiency, ASE, of the clustered D2D network. We first find the Laplace transforms of the two interference terms to characterize SIR. Then, we derive the exact expressions of $\mathtt{P_c}$ and ASE.

4.1. Laplace Transform of Intracluster Interference. Conditioned on $v_0 = \|x_0\|$, we first derive the Laplace transform of I_{intra} as

$$\begin{aligned}
\mathscr{L}_{I_{\text{intra}}}(s \mid v_0) &= \mathbb{E}\left[e^{-sI_{\text{intra}}}\right] \\
&= \mathbb{E}\left[\prod_{y\in\mathscr{B}^{x_0}\setminus y_0} \mathbb{E}_{h_{y_{x_0}}}\left[\exp\left(\frac{-sh_{y_{x_0}}}{\|x+y\|^{\alpha}}\right)\right]\right] \\
&\overset{(a)}{=} \mathbb{E}_{\mathscr{B}^{x_0}}\left[\prod_{y\in\mathscr{B}^{x_0}\setminus y_0} \frac{1}{1+s\|y+x_0\|^{-\alpha}}\right] \\
&\overset{(b)}{=} \exp\left((1-\lambda_t)\int_{\mathbb{R}^3} \frac{s\|y+x_0\|^{-\alpha}}{1+s\|y+x_0\|^{-\alpha}} f_Y(y)\,dy\right) \\
&\overset{(c)}{=} \exp\left((1-\lambda_t)\int_0^{\infty} \frac{sw^{-\alpha}}{1+sw^{-\alpha}} f_W(w \mid v_0)\,dw\right),
\end{aligned} \tag{12}$$

where (a) follows from the exponentially distributed h_{x_0} with unit mean and (b) follows from the probability generating functional (PGF) of Poisson process of the intracluster interferers with mean $(\lambda_t - 1)$. Also, (c) follows from $w = \|x_0 + y\|$.

4.2. Laplace Transform of Intercluster Interference. The Laplace transform of I_{inter} is given by

$$\begin{aligned}
\mathscr{L}_{I_{\text{inter}}}(s) &= \mathbb{E}\left[e^{-sI_{\text{inter}}}\right] \\
&= \mathbb{E}_{\Phi_c}\left[\prod_{x\in\Phi_c\setminus x_0} \mathbb{E}_{\mathscr{B}^x}\left[\prod_{y\in\mathscr{B}^x} \mathbb{E}_{h_{y_x}}\left[\exp\left(\frac{-sh_{y_x}}{\|x+y\|^{\alpha}}\right)\right]\right]\right] \\
&\overset{(a)}{=} \mathbb{E}_{\Phi_c}\left[\prod_{x\in\Phi_c\setminus x_0} \mathbb{E}_{\mathscr{B}^x}\left[\prod_{y\in\mathscr{B}^x} \frac{1}{1+s\|y+x\|^{-\alpha}}\right]\right] \\
&\overset{(b)}{=} \mathbb{E}_{\Phi_c}\left[\prod_{x\in\Phi_c\setminus x_0} \exp\left(\int_0^{\infty} \frac{-\lambda_t su^{-\alpha}}{1+su^{-\alpha}} f_U(u \mid v)\,du\right)\right] \\
&\overset{(c)}{=} \exp\left(4\pi\lambda_c \int_0^{\infty} (\kappa(v)-1)\,v^2 dv\right),
\end{aligned} \tag{13}$$

where $\kappa(v) = \exp(\int_0^{\infty}(-\lambda_t su^{-\alpha}/(1+su^{-\alpha}))f_U(u \mid v)du)$ and (a) follows from the exponentially distributed h_{x_0} with unit mean. Also, (b) and (c) follow from the PGF of Poisson process (with the mean of λ_c and $(\lambda_t - 1)$ resp.).

4.3. Coverage Probability and Area Spectral Efficiency. Letting β denote the SIR threshold for successful decoding at the receiver, which is a function of modulation and coding, the coverage probability is

$$\begin{aligned}
\mathtt{P_c} &= \mathbb{P}\left[\mathtt{SIR} > \beta\right] = \mathbb{E}_R\left\{\mathbb{P}\left[\mathtt{SIR}(R) > \beta \mid R\right]\right\} \\
&= \mathbb{E}_R\left\{\mathbb{P}\left[h_0 > \beta r^{\alpha}\left(I_{\text{intra}} + I_{\text{inter}}\right) \mid R = r\right]\right\} \\
&= \mathbb{E}_R\left\{\mathbb{E}\left\{e^{-\beta r^{\alpha}(I_{\text{intra}}+I_{\text{inter}})} \mid R = r\right\}\right\} \\
&= \int_0^{\infty}\int_0^{\infty} \mathscr{L}_{I_{\text{inter}}}(\beta r^{\alpha})\,\mathscr{L}_{I_{\text{intra}}}(\beta r^{\alpha} \mid v_0) \\
&\quad \times f_R(r \mid v_0)\, f_{v_0}(v_0)\,dr\,dv_0.
\end{aligned} \tag{14}$$

Therefore, letting the area spectral efficiency be defined as the average achievable rate per unit bandwidth per unit area as in [12], the area spectral efficiency is given by

$$\mathtt{ASE} = \lambda_t\lambda_c \log_2(1+\beta)\,\mathtt{P_c}, \tag{15}$$

where $\lambda_t\lambda_c$ is the average density of simultaneously active Dev-Txs of the whole D2D network.

5. Approximate Upper and Lower Bounds of $\mathtt{P_c}$

Because the exact expressions of $\mathtt{P_c}$ and ASE are unwieldy, we provide easy-to-compute upper and lower bounds of $\mathtt{P_c}$. In particular, the lower bound is in a closed form, which can be readily evaluated. As stated in Section 2, r and w are correlated because of the common factor x_0. For analytical tractability to derive the two approximate bounds, we allow separate deconditioning on r and w as in [12], which implies that r and w are i.i.d. following the PDF in (8).

5.1. Upper Bound of $\mathtt{P_c}$. Since the intracluster interferers are significantly closer to the typical device compared to the intercluster Dev-Txs, I_{intra} is dominant in the denominator of SIR. Thus, we can derive the approximate upper bound of SIR by ignoring I_{inter}, which corresponds to the upper bound of $\mathtt{P_c}$. By the i.i.d. assumption of r and w, the Laplace transform of I_{intra} can be approximated as $\widetilde{\mathscr{L}}_{I_{\text{intra}}}(s) = e^{(1-\lambda_t)\int_0^{\infty}(sw^{-\alpha}/(1+sw^{-\alpha}))f_W(w)dw}$, where $f_W(w)$ follows the PDF in (8). Thus, the upper bound of $\mathtt{P_c}$ is given by

$$\begin{aligned}
\widetilde{\mathtt{P_c}} &= \mathbb{E}_R\left\{\mathbb{P}\left[h_0 > \beta r^{\alpha} I_{\text{intra}} \mid R = r\right]\right\} \\
&= \int_0^{\infty} \widetilde{\mathscr{L}}_{I_{\text{intra}}}(\beta r^{\alpha})\, f_R(r)\,dr,
\end{aligned} \tag{16}$$

where $f_R(r)$ follows the PDF in (8).

5.2. Lower Bound of $\mathtt{P_c}$. We first derive lower bounds of $\mathscr{L}_{I_{\text{intra}}}(s)$ and $\mathscr{L}_{I_{\text{inter}}}(s)$ in closed forms. Then, using the two, the lower bound of $\mathtt{P_c}$ will be obtained.

Lemma 4 (lower bound of $\mathscr{L}_{I_{\text{intra}}}(s)$). *The lower bound on the Laplace transform of I_{intra} is*

$$\mathscr{L}_{I_{intra}}(s) \geq \mathscr{L}^{*}_{I_{intra}}(s) = \exp\left[\frac{1-\lambda_t}{6\sqrt{\pi}\sigma^3}\frac{s^{3/\alpha}(3\pi/\alpha)}{\sin(3\pi/\alpha)}\right]. \quad (17)$$

Proof. See Appendix A. □

Lemma 5 (lower bound of $\mathscr{L}_{I_{\text{inter}}}(s)$). *The lower bound on the Laplace transform of I_{inter} is given by*

$$\mathscr{L}_{I_{inter}}(s) \geq \mathscr{L}^{*}_{I_{inter}}(s) = \exp\left[-\frac{4}{3}s^{3/\alpha}\frac{\lambda_c\lambda_t\left(3\pi^2/\alpha\right)}{\sin(3\pi/\alpha)}\right]. \quad (18)$$

Proof. See Appendix B. □

With (17) and (18) along with the independent deconditioning assumption, we can obtain the approximate lower bound of $\mathtt{P}_c$ in a closed form as

$$\begin{aligned}\mathtt{P}_c &\geq \int_0^{\infty} \mathscr{L}^{*}_{I_{\text{inter}}}(\beta r^{\alpha})\,\mathscr{L}^{*}_{I_{\text{intra}}}(\beta r^{\alpha})\,f_R(r)\,dr \\ &\overset{(a)}{=} \int_0^{\infty} \exp\left(-\rho r^3\right)\frac{r^2}{2\sqrt{\pi}\sigma^3}\exp\left(-\frac{r^2}{4\sigma^2}\right)dr,\end{aligned} \quad (19)$$

where $\rho = (3\pi\beta^{3/\alpha}/\alpha\sin(3\pi/\alpha))((\lambda_t - 1)/6\sqrt{\pi}\sigma^3 + 4\pi\lambda_c\lambda_t/3)$ and (a) follows from $f_R(r)$ following (8). Because $\rho \geq 0$ ($\because \alpha \geq 2$ and $\lambda_t \geq 1$), we can obtain the lower bound in (20), where $\mathrm{Bi}(a) = (1/\pi)\int_0^{\infty}\cos(t^3/3 + at)dt$ is the Airy function, the derivative of which is $\mathrm{Bi}'(a)$. Moreover, ${}_2F_2$ is the generalized hypergeometric function [23].

$$\begin{aligned}&\mathtt{P}_c^{*} \\ &= \frac{108\rho^{4/3}\sigma^4\,{}_2F_2\left(1/2, 1; 1/3, 2/3; -1/432\rho^2\sigma^6\right) + e^{-1/864\rho^2\sigma^6}\left(3^{2/3}\pi\,\mathrm{Bi}\left(1/48\sqrt[3]{3}\rho^{4/3}\sigma^4\right) - 12\sqrt[3]{3}\pi\rho^{2/3}\sigma^2\mathrm{Bi}'\left(1/48\sqrt[3]{3}\rho^{4/3}\sigma^4\right)\right)}{648\sqrt{\pi}\rho^{7/3}\sigma^7}.\end{aligned} \quad (20)$$

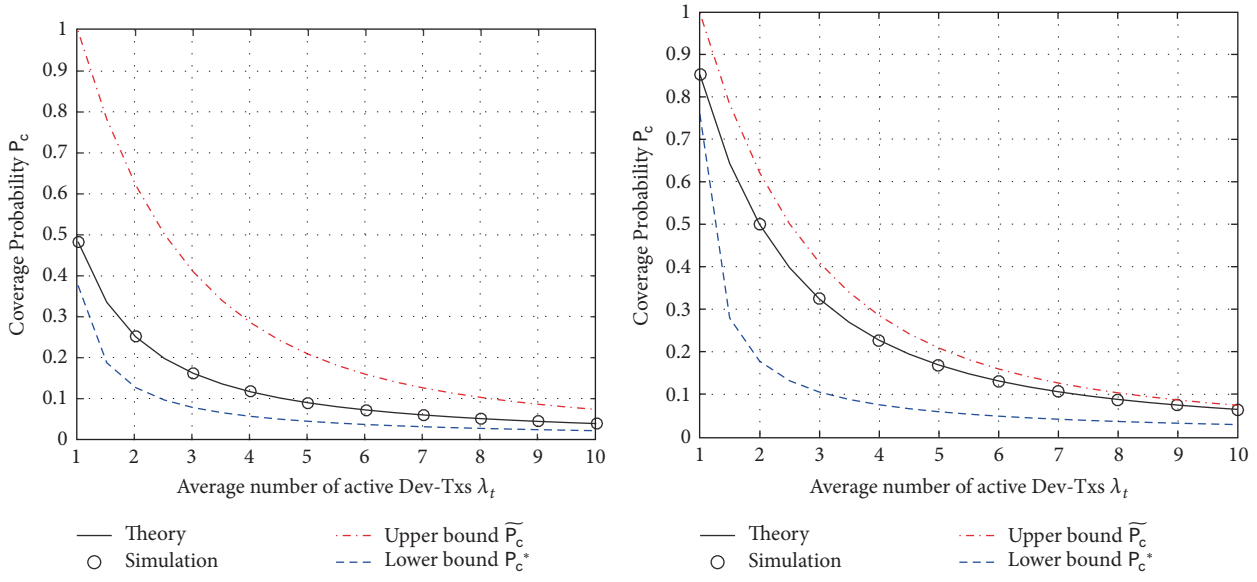
6. Numerical Results

In this section, we present numerical results to validate our analysis and discuss the impacts of system parameters. For simulations, the device locations are randomly drawn from a TCP over $100 \times 100 \times 100\ \mathrm{m}^3$ cube. The cluster centers follow PPP with intensity λ_c, and devices are normally distributed around their cluster centers. Moreover, the number of the Dev-Txs in each cluster follows a Poisson distribution with mean λ_t. Also, we assume the path-loss exponent α of 4, as in [12, 19, 20]. The simulation results are obtained from 10^6 random realizations of device distribution (network topology) and Rayleigh fading channel.

6.1. Impacts of System Parameters. Figures 2(a) and 2(b) show how the coverage probability $\mathtt{P}_c$ varies, as the average number of simultaneously active Dev-Txs λ_t increases, with $\lambda_c = 0.3$ and 0.05, respectively. In the figures, the circles indicate the simulation results, while the solid line represents the theoretical results obtained numerically using (14). Moreover, the dash-dotted and dashed curves correspond to the upper and lower bounds $\widetilde{\mathtt{P}_c}$ and $\mathtt{P}_c{}^{*}$ in (16) and (20), respectively. In both figures, the simulation results show the excellent agreements with the theoretical results, which verifies our analysis. Moreover, the approximate upper and lower bounds of $\mathtt{P}_c$ derived in the previous section are validated. Specifically, comparing the two figures, when λ_c is large, the actual $\mathtt{P}_c$ is closer to the lower bound $\mathtt{P}_c{}^{*}$ compared to the upper bound $\widetilde{\mathtt{P}_c}$, as in Figure 2(a), because the large λ_c results in the higher intercluster interference I_{inter}, which is ignored in the $\widetilde{\mathtt{P}_c}$. On the other hand, for small λ_c, the gap between the exact $\mathtt{P}_c$ and its upper bound $\widetilde{\mathtt{P}_c}$ is significantly smaller compared to the difference from its lower bound $\mathtt{P}_c{}^{*}$, as in Figure 2(b), since the intracluster interference I_{intra} is dominant relative to the intercluster interference I_{inter}. In either case, the exact $\mathtt{P}_c$ curve is always bounded by $\widetilde{\mathtt{P}_c}$ and $\mathtt{P}_c{}^{*}$.

In Figures 3 and 4, we observe the impacts of σ and λ_c on exact $\mathtt{P}_c$, numerically obtained by (14), respectively. In the figures, we consider three scenarios in the presence of (i) only intracluster interference, (ii) only intercluster interference, and (iii) both intra- and intercluster interferences, which correspond to the dashed, dash-dotted, and solid lines, respectively. Moreover, the triangles and circles indicate the corresponding simulation results. In both figures, when λ_t grows, the intracluster interference, indicated by the dashed line, is dominant compared to the intercluster interference, which is indicated by the dash-dotted line. Also, in Figure 3, the larger σ, which means the larger spatial scattering of the devices from the cluster center, results in the lower $\mathtt{P}_c$. This can be attributed to the increased impact of I_{inter}, while the $\mathtt{P}_c$ curves only with I_{intra} do not change as indicated by the dashed curves in the figure. The $\mathtt{P}_c$ curves only with I_{intra} stay the same regardless of λ_t, because the variations of the serving and interfering Dev-Txs cancel each other. We can observe the same trend in Figure 4: as λ_c increases, $\mathtt{P}_c$ decreases because of the increased intercluster interference I_{inter}. On the other hand, the coverage probability $\mathtt{P}_c$ only with the intracluster interference I_{intra} does not vary under the variation in the cluster density λ_c.

Figures 5 and 6 show the exact area spectral efficiency ASE, numerically obtained by (15), versus the average number of simultaneously active Dev-Txs λ_t. In the figure, the horizontal axis indicates λ_t, while the vertical axis is ASE. Also, the solid, dashed, and dash-dotted lines represent the theoretical results with different system parameters (λ_c, σ, and β), while the circle and triangle markers represent the corresponding simulation results. Lastly, the optimal λ_t in each graph is indicated by the "x"-marker.



(a) P_c with $\lambda_c = 0.3$ clusters/m^3, $\sigma = 0.3$, and $\beta = 0$ dB

(b) P_c with $\lambda_c = 0.5$ clusters/m^3, $\sigma = 0.3$, and $\beta = 0$ dB

FIGURE 2: P_c versus λ_t: comparison with the upper and lower bounds.

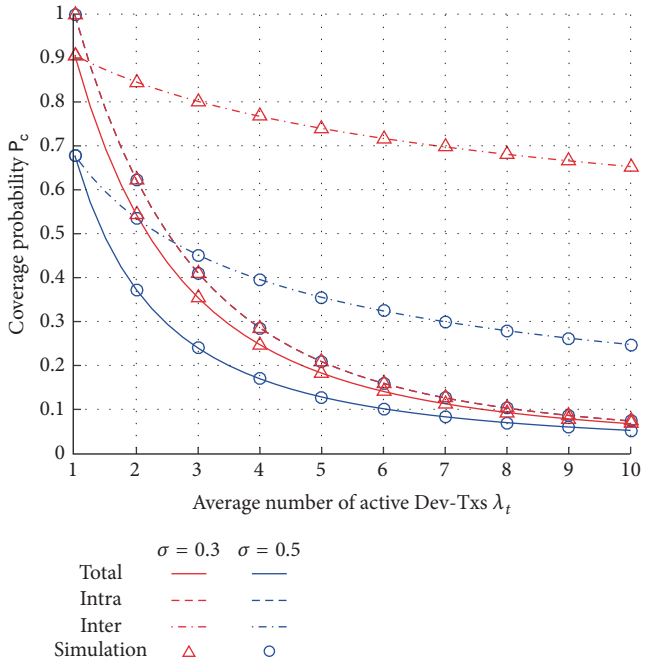


FIGURE 3: P_c versus λ_t with $\beta = 0$ dB, $\lambda_c = 0.03$ clusters/m^3, and $\sigma = \{0.3, 0.5\}$.

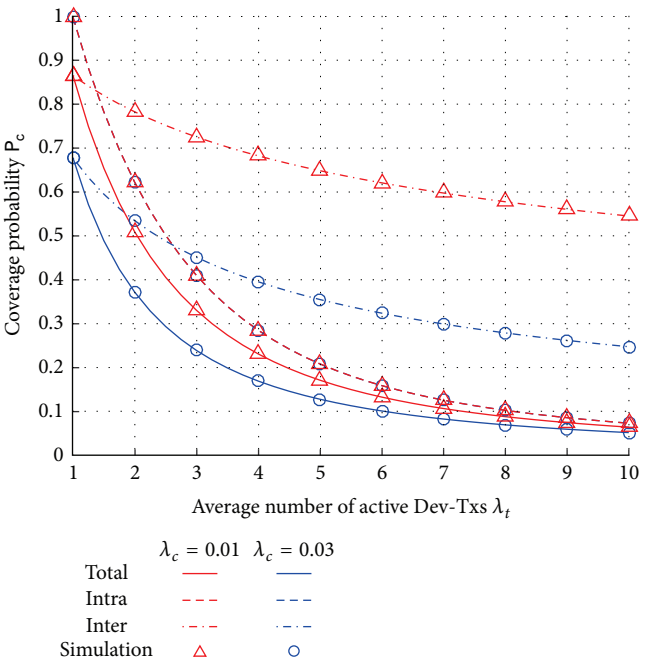


FIGURE 4: P_c versus λ_t with $\beta = 0$ dB, $\sigma = 0.5$, and $\lambda_c = \{0.01, 0.03\}$ clusters/m^3.

As shown in Figures 2, 3, and 4, we can observe the great correlation between the simulation and theoretical results. If comparing the curves with $\sigma = 0.3$ and 0.5 with the same λ_c in Figure 5, ASE increases, as σ decreases, which is expected from the result of P_c in Figure 3. Moreover, for the fixed $\sigma = 0.5$, the curves with $\lambda_c = 0.03$ show significantly higher ASE compared to the curve with $\lambda_c = 0.01$, because of the greater multiplication factor in (15). In Figure 6, the higher β makes ASE increase for small λ_t, while the curves with higher β decrease more rapidly compared to the curves with smaller β, as λ_t increases. One of the most important design aspects is the optimal λ_t to maximize the ASE, which determines network operation and wireless resource allocation. In Figure 5, $\lambda_t \approx 2$ gives the best ASE for all the three curves, which indicates its low sensitivity to σ and λ_c. On the other hand, if we increase the SIR threshold

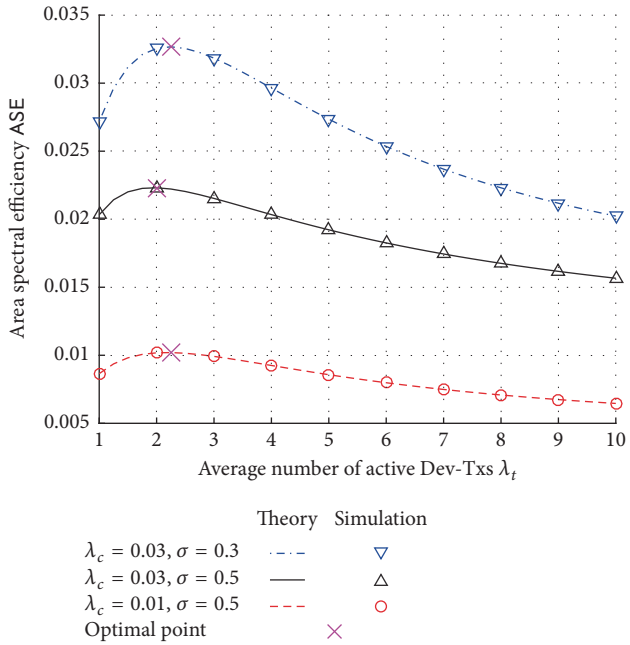


FIGURE 5: ASE versus λ_t; variations in σ and λ_c with $\beta = 0$ dB.

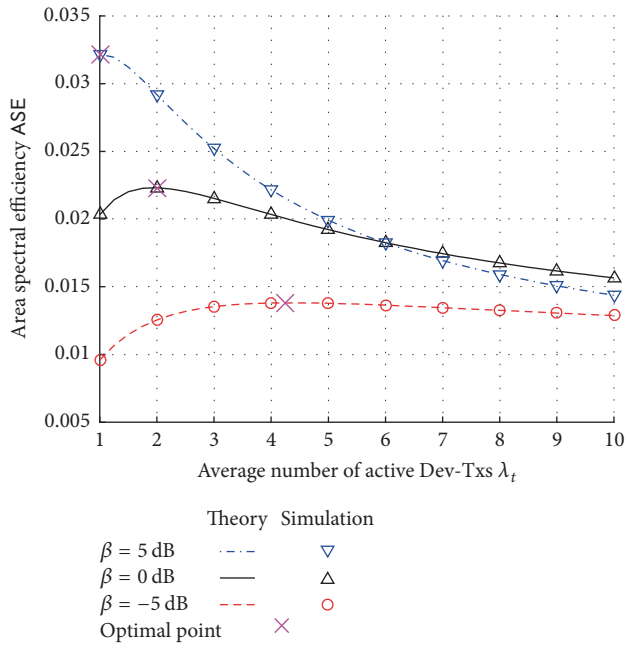


FIGURE 6: ASE versus λ_t with $\sigma = 0.5$, $\lambda_c = 0.03$ clusters/m^3, and $\beta = \{-5, 0, 5\}$ dB.

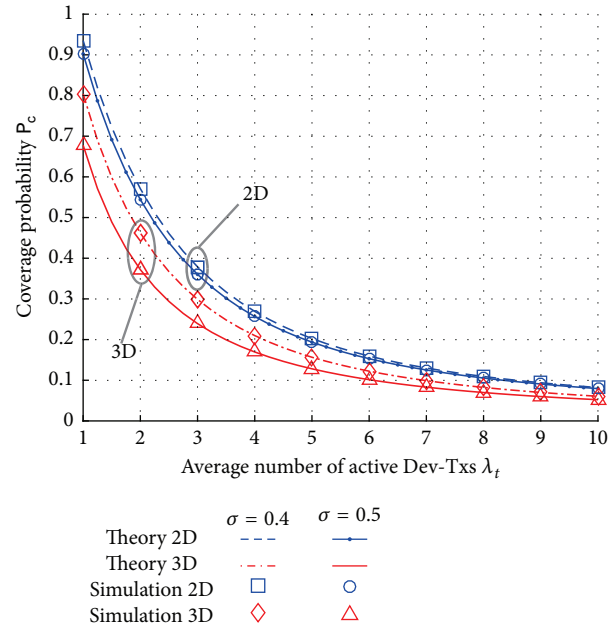


FIGURE 7: Comparison of 2D and 3D with $\beta = 0$ dB, $\lambda_c = 0.03$, and $\sigma = \{0.4, 0.5\}$.

β as in Figure 6, the optimal value of λ_t decreases (4.25, 2, 1 for $\beta = -5, 0, 5$ dB, resp.), because lower β can accommodate more simultaneous users (devices).

6.2. Comparison between 2D and 3D TCP Models. In this section, we compare the performance of the 3D clustered D2D networks with the 2D clustered networks studied in [12]. For the comparison, we set the same cluster density per unit space λ_c (clusters/m^2 and clusters/m^3 in 2D and 3D spaces, resp.). Figures 7 and 8 show the coverage probability P_c versus the average number of simultaneously active Dev-Txs λ_t in the 2D and 3D spaces. In the figures, the lined curves represent the theoretical results, while the markers indicate the simulation results. As shown in the figure, we observe that the analytical and simulation results are consistent with each other both for the 2D and 3D cases. For the same parameter set, P_c of the 2D TCP is higher compared to P_c of the 3D TCP, which is consistent with the results assuming uniform node distribution following PPP in [20]. This can be explained by more number of interferers inside volumes with the same radius from the typical device in 3D space compared to 2D space even with the same cluster density λ_c per unit space. From the figure, the gap between the 2D and 3D curves grows for the larger σ and λ_c. Furthermore, compared to the 3D space results, the P_c performances in the 2D space are less sensitive to the change in σ and λ_c as observed in Figures 7 and 8, respectively.

Furthermore, Figure 9 displays the area spectral efficiency ASE versus λ_t graphs of the 2D TCP under the change in β using the same parameters as the 3D TCP case shown in Figure 6: $\sigma = 0.5$, $\lambda_c = 0.03$ clusters/m^2, and $\beta = \{-5, 0, 5\}$ dB. Overall, the ASE in the 2D TCP is greater compared to the 3D case, because of the higher coverage probability P_c as indicated in Figures 7 and 8. That is, if we use the 2D TCP model for ultra-dense urban environments where the devices exist in 3D space, which will be common in the future wireless networks, both the coverage probability and the area spectral efficiency are overestimated. When comparing the impact of β, we can observe the similar trend in the 2D and 3D models that the higher β gives the higher ASE with small λ_t. However, while the curves with $\beta = 0$ dB and 5 dB cross over at around

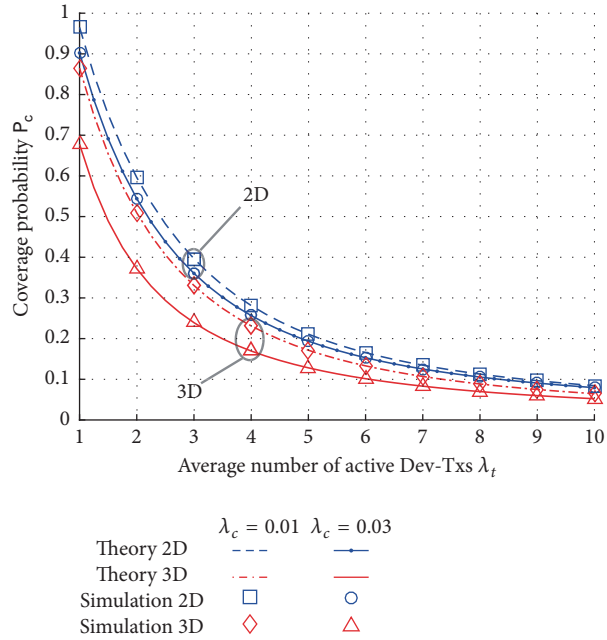


Figure 8: Comparison of 2D and 3D with $\beta = 0$ dB, $\sigma = 0.5$, and $\lambda_c = \{0.01, 0.03\}$.

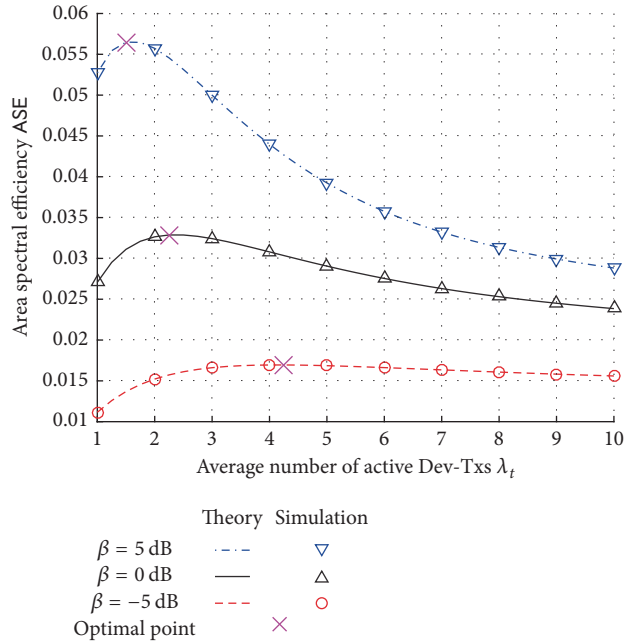


Figure 9: ASE versus λ_t in two-dimensional (2D) space with $\sigma = 0.5$, $\lambda_c = 0.03$ clusters/m^2, and $\beta = \{-5, 0, 5\}$ dB.

$\lambda_t = 6$ in Figure 6, the two curves in Figure 9 keep the measurable gap. Interestingly, the optimal λ_t to maximize the ASE shows the similar (but not exactly the same) trend to that seen in Figure 6. Thus, in a nutshell, the 2D TCP model can be used to estimate the optimal number of simultaneously active Dev-Txs in the 3D clustered networks; however both P_c and ASE estimated based on the 2D model are significantly overestimated compared to the actual performances of the 3D clustered D2D networks.

7. Conclusion

In this paper, we have studied clustered D2D networks in 3D space modeled by TCP for dense urban environments, where devices are distributed over the 3D space. Using stochastic geometry, we have analyzed P_c and ASE of the D2D network in the presence of cochannel interference from both the same cluster and the other clusters. We have derived the exact mathematical expressions of P_c and ASE, which were verified with the simulation results. Moreover, the approximate upper

and lower bounds on $\mathtt{P_c}$ have been derived, which provide design insights. Both the numerical and simulation results indicate that $\mathtt{P_c}$ in 3D space is significantly lower compared to 2D space for the same cluster density λ_c per unit space because of the more interferers within a certain distance. In addition, compared to the 3D space, the 2D TCP model is less sensitive to the system parameters such as the spatial scattering of the devices σ and the cluster density λ_c. Comparing the two models, we can conclude that the optimal numbers of simultaneously active devices λ_t to maximize ASE can be similar in the 2D and 3D models. However, it is not appropriate to use the 2D TCP to estimate $\mathtt{P_c}$ and ASE of the D2D networks following the 3D TCP especially for the large σ and λ_c.

The study in this paper provides guidelines on how to operate D2D networks in the presence of cochannel interference among devices, which are distributed in clusters in 3D space. The most significant aspect is how much simultaneous traffic to accommodate using the same channel. Through analysis and simulation, we have shown that there exist an optimal number of the simultaneously active D2D links to maximize ASE, and the optimum is smaller in the 3D D2D networks compared to the 2D D2D networks. Based on this result, one can determine the number of the cochannel D2D pairs to allow communicating in each cluster at the same time, which impacts the higher layer design such as wireless resource allocation for given cluster density λ_c, spatial scattering of devices σ, and quality of service (QoS) requirement characterized by β.

Appendix

A. Proof of Lemma 4

$$
\begin{aligned}
&\mathscr{L}_{I_{\text{intra}}}(s)\\
&= \int_{\mathbb{R}^3} \exp\left(\int_{\mathbb{R}^3} \frac{(1-\lambda_t) f_Y(y)\, dy}{1+\|y+x_0\|^{\alpha}/s}\right) \cdot f_Y(x_0)\, dx_0\\
&= \int_{\mathbb{R}^3} \exp\left(\int_{\mathbb{R}^3} \frac{(1-\lambda_t)}{1+\|z\|^{\alpha}/s} f_Y(z-x_0)\, dz\right)\\
&\quad \cdot f_Y(x_0)\, dx_0\\
&\overset{(a)}{\geq} \exp\left(\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{(1-\lambda_t)}{1+\|z\|^{\alpha}/s} f_Y(z-x_0) f_Y(x_0)\, dx_0 dz\right)\\
&\overset{(b)}{\geq} \exp\left(\int_{\mathbb{R}^3} \frac{(1-\lambda_t)}{1+\|z\|^{\alpha}/s} \|f_Y * f_Y\|_{\infty}\, dz\right)\\
&= \exp\left(\frac{(1-\lambda_t)}{8\pi\sqrt{\pi}\sigma^3}\int_{\mathbb{R}^3} \frac{1}{1+\|z\|^{\alpha}/s} dz\right)\\
&= \exp\left[\frac{1-\lambda_t}{6\sqrt{\pi}\sigma^3}\frac{s^{3/\alpha}(3\pi/\alpha)}{\sin(3\pi/\alpha)}\right],
\end{aligned}
\tag{A.1}
$$

which corresponds to $\mathscr{L}^{*}_{I_{\text{intra}}}(s)$ in (17). (a) follows from Jensen's inequality, and (b) follows from Holder's inequality.

B. Proof of Lemma 5

$$
\begin{aligned}
&\mathscr{L}_{I_{\text{inter}}}(s)\\
&\overset{(a)}{\geq} \exp\left(4\pi\lambda_c \int_0^{\infty}\int_0^{\infty} \frac{-\lambda_t s u^{-\alpha}}{1+s u^{-\alpha}} f_U(u \mid v)\, du v^2 dv\right)\\
&\overset{(b)}{=} \exp\left(-4\pi\lambda_c\lambda_t \int_0^{\infty} \frac{s u^{-\alpha}}{1+s u^{-\alpha}} u^2 du\right)\\
&= \exp\left[-\frac{4}{3} s^{3/\alpha} \frac{\lambda_c\lambda_t\left(3\pi^2/\alpha\right)}{\sin(3\pi/\alpha)}\right],
\end{aligned}
\tag{B.1}
$$

which is $\mathscr{L}^{*}_{I_{\text{inter}}}(s)$ in (18). (a) follows from the Taylor expansion of an exponential function, and (b) is based on the property of the PDF in (10) that $\int_0^{\infty} f_U(u \mid v) v^2 dv = u^2$.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

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 (Grant no. 2016R1D1A1B03930060).

References

[1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, 2014.

[2] J. G. Andrews, S. Buzzi, and W. Choi, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[3] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 4, pp. 1801–1819, 2014.

[4] Y. Li and W. Wang, "Message dissemination in intermittently connected D2D communication networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3978–3990, 2014.

[5] S. Mumtaz, K. M. S. Huq, M. I. Ashraf, J. Rodriguez, V. Monteiro, and C. Politis, "Cognitive vehicular communication for 5G," *IEEE Communications Magazine*, vol. 53, no. 7, pp. 109–117, 2015.

[6] M. R. Palattella, M. Dohler, A. Grieco et al., "Internet of things in the 5G era: enablers, architecture, and business models," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 510–527, 2016.

[7] B. Bangerter, S. Talwar, R. Arefi, and K. Stewart, "Networks and devices for the 5G era," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 90–96, 2014.

[8] O. Bello and S. Zeadally, "Intelligent device-to-device communication in the internet of things," *IEEE Systems Journal*, vol. 10, no. 3, pp. 1172–1182, 2014.

[9] M. Haenggi, *Stochastic Geometry for Wireless Networks*, Cambridge University Press, Cambridge, UK, 2012.

[10] H. Jung and I.-H. Lee, "Outage analysis of millimeter-wave wireless backhaul in the presence of blockage," *IEEE Communications Letters*, vol. 20, no. 11, pp. 2268–2271, 2016.

[11] H. Jung and I. Lee, "Outage Analysis of Multihop Wireless Backhaul Using Millimeter Wave under Blockage Effects," *International Journal of Antennas and Propagation*, vol. 2017, Article ID 4519365, 9 pages, 2017.

[12] M. Afshang, H. S. Dhillon, and P. H. Joo Chong, "Modeling and performance analysis of clustered device-to-device networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4957–4972, 2016.

[13] H. Jung and I.-H. Lee, "Connectivity Analysis of millimeter-wave device-to-device networks with blockage," *International Journal of Antennas and Propagation*, vol. 2016, Article ID 7939671, 9 pages, 2016.

[14] M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahnt, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, pp. 1–14, October 2007.

[15] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of YouTube videos," in *Proceedings of the 16th International Workshop on Quality of Service (IWQoS '08)*, pp. 229–238, IEEE, Enschede, The Netherlands, June 2008.

[16] T. Koskela, S. Hakola, T. Chen, and J. Lehtomäki, "Clustering concept using device-to-device communication in cellular system," in *Proceedings of the IEEE Wireless Communications and Networking Conference 2010, WCNC 2010*, pp. 1–6, April 2010.

[17] Y. Zhang, E. Pan, L. Song, W. Saad, Z. Dawy, and Z. Han, "Social network aware device-to-device communication in wireless networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 177–190, 2015.

[18] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: basic principles and system performance," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 176–189, 2016.

[19] A. K. Gupta, X. Zhang, and J. G. Andrews, "Potential throughput in 3D ultradense cellular networks," in *Proceedings of the 49th Asilomar Conference on Signals, Systems and Computers, ACSSC 2015*, pp. 1026–1030, usa, November 2015.

[20] Z. Pan and Q. Zhu, "Modeling and analysis of coverage in 3-d cellular networks," *IEEE Communications Letters*, vol. 19, no. 5, pp. 831–834, 2015.

[21] A. Omri and M. O. Hasna, "Modeling and performance analysis of D2D communications with interference management in 3-D HetNets," in *Proceedings of the 59th IEEE Global Communications Conference, GLOBECOM 2016*, pp. 1–7, December 2016.

[22] Z. Pan and Q. Zhu, "Energy efficiency optimization in 3-D small cell networks-based sleep strategy," *IEEE Communications Letters*, vol. 21, no. 5, pp. 1131–1134, 2017.

[23] G. E. Andrews, R. Askey, and R. Roy, *Special Functions*, Cambridge University Press, Cambridge, UK, 1999.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 9727360, 10 pages
https://doi.org/10.1155/2017/9727360

Research Article

Optimized Power Allocation and Relay Location Selection in Cooperative Relay Networks

Jianrong Bao,[1,2] Jiawen Wu,[1] Chao Liu,[1] Bin Jiang,[1] and Xianghong Tang[1]

[1]*School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China*
[2]*National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China*

Correspondence should be addressed to Bin Jiang; jiangbin@hdu.edu.cn

Received 4 April 2017; Revised 2 July 2017; Accepted 11 October 2017; Published 9 November 2017

Academic Editor: Haiyu Huang

An incremental selection hybrid decode-amplify forward (ISHDAF) scheme for the two-hop single relay systems and a relay selection strategy based on the hybrid decode-amplify-and-forward (HDAF) scheme for the multirelay systems are proposed along with an optimized power allocation for the Internet of Thing (IoT). Given total power as the constraint and outage probability as an objective function, the proposed scheme possesses good power efficiency better than the equal power allocation. By the ISHDAF scheme and HDAF relay selection strategy, an optimized power allocation for both the source and relay nodes is obtained, as well as an effective reduction of outage probability. In addition, the optimal relay location for maximizing the gain of the proposed algorithm is also investigated and designed. Simulation results show that, in both single relay and multirelay selection systems, some outage probability gains by the proposed scheme can be obtained. In the comparison of the optimized power allocation scheme with the equal power allocation one, nearly 0.1695 gains are obtained in the ISHDAF single relay network at a total power of 2 dB, and about 0.083 gains are obtained in the HDAF relay selection system with 2 relays at a total power of 2 dB.

1. Introduction

Recently, multiple-input multiple-output (MIMO), as a milestone in the development of wireless communications, brought an efficient transmission rate and reliability. To put it into practice, a cooperative communication scheme was then proposed in time [1], and it had been widely used and rapidly developed. In cooperative communications, the diversity gain was obtained, when the relay node forwarded messages and the destination node combined the received signals from both the source and relay nodes. According to different strategies for processing signals at the relay nodes, there are mainly three cooperation schemes, such as the amplify-and-forward (AF) [1], the decode-and-forward (DF) [2], and the coded cooperation (CC) [3]. To solve the deficiency of AF relay amplifying both the noises and signals, causing the incorrect DF relay decoding and also the error propagation phenomenon, an incremental relay protocol [4], accompanied by a hybrid decode-amplify-and-forward (HDAF), was proposed [5]. For the shortage of the incremental relay protocol, the incremental selection amplify-and-forward (ISAF) [6] was investigated, which selected the proper occasion to retransmit the messages in the source according to the channel estimation, when the direct transmission between the source and destination was failed. But the noise amplification still remained. Then, an incremental selection hybrid decode-amplify forward (ISHDAF) scheme was proposed in [7], where the HDAF scheme was combined with incremental selection strategy. Compared with the aforementioned ISAF scheme, the ISHDAF scheme had a significant improvement in bit error rate (BER) and outage probability, since both the BER and outage probability of a cooperative transmission system in the DF strategy were lower than those in the AF strategy. According to the principle of incremental relay, the average spectral efficiency of the ISHDAF scheme was also higher than that of the HDAF scheme. However, all gains were obtained under equal power allocation of the source and relay nodes for the low systematic complexity, which caused the deficiency of only few performance improvements. To improve the

spectral efficiency of the system, there was also a signal-to-noise ratio (SNR) based incremental hybrid decode-amplify forward (IHDAF) protocol proposed in [8]. Furthermore, the SNR thresholds, the power allocation schemes, and the relay locations were studied to optimize the outage probability and BER performance.

Meanwhile, power allocation in cooperation communications had always been one of the research hot-spots. In wireless uplink transmissions, successive interference cancellation (SIC) was combined with the power allocation method to obtain the optimal power allocation ratio for efficient resource allocation [9]. For relay forwarding systems, two power allocation methods, by the Lagrange multiplier method and the differential algorithm, respectively, were also proposed for the lower bound of symbol error rate (SER) in the HDAF relay cooperative networks [10]. Xiao and Ouyang in [11] developed a two-source-destination-pair cooperative network with the HDAF protocol. And a closed-form expression of the outage probability was derived and thus a minimal total power was obtained under the constraints of the supposed outage probability. Similarly, for the two-source-destination-pair system, there was also a parallel shift water filling algorithm proposed for the power allocation [12]. The advantage of it over the conventional ones was the reduced complexity by just eliminating the iterative searching process. However, the cost is a little performance decrease. In [13], a multirelay selection scheme with joint power allocation was proposed, featured with the significantly decreased complexity of computation. To achieve this effect, it used a simple power reallocation during the multirelay selection process. Also a joint relay selection and power allocation scheme for cooperative wireless sensor networks was proposed in [14]. It adaptively chose the proper relays and their transmission power to maximize the signal-to-noise ratio (SNR) at the destination by the channel state information (CSI). Then a SNR-based relay selection for IHDAF cooperative diversity protocol was also proposed in [15], and the closed-form expressions of average channel capacity and outage probability were derived simultaneously. Also a swarm intelligence-based power allocation and relay selection algorithm could be used for wireless cooperative networks [16]. It could not only reduce the computational complexity effectively, but also select the optimal relay nodes to solve the nonlinear optimization problems by a fast global search with low cost. In [17], with a AF protocol based two-hop multiple energy-harvesting relays network, an improved power allocation was investigated to improve the whole outage performance. The innovation in the proposed scheme lied in the fact that it jointly maximized the transmit power under the constraints of limited individual relay energy. Also the power allocation and relay selection strategies in both dual-hop and multihop scenarios in cognitive relay networks were researched [18]. They achieved the good features of both minimal total transmit power and maximal entire network capacity. For the relay selection optimization, there had been a dynamic strategy to choose the best relay node and path under the constraints of total power and power allocation for each relay [18]. In addition, an improved relay selection strategy of HDAF scheme was proposed to improve the BER and outage probability [19]. It can adaptively select the AF or DF forward strategy for all relays according to the channel quality. Then the best relay was chosen to forward signals. However, it still used the equal power allocation to reduce the complexity. In [20], a power allocation algorithm by the lowest average bit error was proposed for these infrastructure-less networks using unbalanced communication links. To investigate the influence of the links on the system performance, the location of the relay node with respect to the source and destination nodes was also studied. Unfortunately, only the effect of certain node locations, rather than the optimal relay location, was determined. Subsequently, a much more detailed study about the optimal relay location was presented in [21], but under simply fixed ratio nodes power.

In this paper, by analyzing a two-hop single relay cooperative network with an ISHDAF scheme and the multirelay selection strategy with a HDAF scheme, an optimized power allocation is proposed. The main contributions are summarized as follows:

(1) The power allocation is optimized in both the ISHDAF single relay and the HDAF multirelay systems. In the case of link status change, the preferred links are allocated with much more power for transmission according to the well-known water filling principle in information theory, which reduces the entire power consumption under the same system performance. The proposed scheme also provides a new hybrid automatic repeat request (ARQ) retransmission and relay forward mechanism, where the source node can retransmit messages to the destination node when the first direct transmission failed. It differs from the source node sending new messages to the destination node directly in the incremental relaying protocol. So it can be more suited for all kinds of the multiple relay channel status and obviously improves the systematic outage probability without any complexity increase.

(2) The approximate closed-form expression of the systematic outage probability with relation to the node power and channel coefficients is derived by the equivalent infinitesimal replacement of the probability distribution function at high SNR. And it can be taken as the objective function of the optimization. Then, the minimization is achieved under fixed total power by Lagrange multiplier method, and the objective function is related to the power of the source node and the relay nodes. The power allocation coefficients between the source and the relay nodes are then obtained to achieve optimized power allocation. Moreover, the power allocation changes the location selection of the relay nodes, which can be calculated indirectly from the above closed-form expression. It can adaptively satisfy the link conditions to optimize the entire system performance.

(3) By introducing the path loss factor, the powers of the source and the relay nodes are modeled as the objective function related to the distance among all nodes. According to both the property of the objective

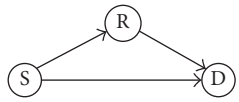


Figure 1: System model of a single relay communication.

function and the related numerical analyses, the relationships of the varied power to the distance of the nodes are obtained. Then, the optimized node positions are obtained to improve the power efficiency with minimized systematic outage probability. Also the power allocation of all relay nodes with respect to their relative location to the source and destination nodes can be clearly and quantitatively analyzed by this model. Therefore, the link status associated with the proposed relay position obviously affects the selection of the cooperative schemes, which can be adopted in practice.

This paper is organized as follows. In Section 2, a single relay cooperation system with the ISHDAF scheme is introduced. Section 3 presents a multirelay selection strategy based on the HDAF scheme. Subsequently, the analytical expressions of the outage probability of both the ISHDAF and the HDAF relay system are derived in Section 4. In this section, an optimized power allocation using Lagrange Method is also proposed to minimize the outage probability. Simultaneously, the close-form analytical expression of the optimal relay location of the proposed algorithm is given to manifest the relationship between the relay location and the outage probability. After that, in Section 5, the simulation results and analyses are presented to verify the good outage probability and optimal power allocation brought by the proposed algorithm. The optimal relay location by the proposed method is also given and tested to be effective. Finally, Section 6 concludes the whole paper.

2. Single Relay Model and ISHDAF Mutual Information Evaluation

For a classic three-node relay model shown in Figure 1, it consists of a source node S, a relay node R, and a destination node D. Equipped with a single omnidirectional antenna, all nodes communicates with each other. The ideal channel state information (CSI) can be obtained through channel training. For the independent links S-D, S-R, and R-D, their channel gains, that is, $|h_{sd}|^2$, $|h_{sr}|^2$, and $|h_{rd}|^2$, are subject to the exponential distribution with channel parameters as $1/{\sigma_{sd}}^2$, $1/{\sigma_{sr}}^2$, and $1/{\sigma_{rd}}^2$, respectively. At a flat Rayleigh fading channel, the noise is an additive white Gaussian noise (AWGN), with zero mean and variance N_0.

In an ISHDAF cooperative network, the whole transmission is divided into two time slots. In the first slot, node S sends a signal to node R and node D, while in the second slot, either node S or node R sends the signal to node D, which depends on the link status. Suppose that the transmitted power of the source node S is P_{S1}, and the information transmission rate is R bit/s. There are two main situations according to the decoding results of the destination node.

For the first situation, if the destination node successfully receives the signal sent by the source in the first slot, the transmission from node S to node D is not interrupted. In this case, the mutual information is defined in [22] as

$$I_{\mathrm{DT}} = \frac{1}{2} \cdot \log_2 \left(1 + \frac{|h_{sd}|^2 P_{S1}}{N_0} \right), \tag{1}$$

which needs to be larger than R according to the information theory. By simplifying (1) and the condition of $I_{\mathrm{DT}} > R$, the relationship of $|h_{sd}|^2 > (2^{2R} - 1)N_0/P_{S1}$ is obtained. Given the threshold as $T_1 = (2^{2R} - 1)N_0/P_{S1}$, and $|h_{sd}|^2 > T_1$, the source node S keeps transmitting directly to the destination node D in the second slot, and the relay node R remains inactive.

For another situation, if the direct transmission fails in the first slot, or the destination does not receive the correct information from the source, the source would retransmit the message to the destination in the second slot. In this case, the mutual information is deduced and presented in [6] as

$$I_{\mathrm{DRT}} = \frac{1}{2} \cdot \log_2 \left(1 + \frac{2\,|h_{sd}|^2 P_{S1}}{N_0} \right). \tag{2}$$

To ensure the success retransmission, (2) should also be larger than R and it can obtain $|h_{sd}|^2 > T_1/2$. When $T_1/2 < |h_{sd}|^2 \leq T_1$, the source retransmits message and the relay remains inactive in the second slot.

The relay node starts the cooperative transmission when there is $|h_{sd}|^2 \leq T_1/2$. Then, if the relay can correctly decode the information from the source, the DF scheme is adopted in the second slot. It needs to satisfy the relationship of $1/2 \cdot \log_2(1 + |h_{sr}|^2 P_{S1}/N_0) > R$, or $|h_{sr}|^2 > T_1$, which is the required condition for the relay to forward the correct signal in the DF protocol. Otherwise, when $|h_{sr}|^2 \leq T_1$, the AF protocol is used instead.

If the DF scheme is adopted for the cooperative transmission, the mutual information is obtained by the maximum ratio combining (MRC) for the destination as

$$I_{\mathrm{DF}} = \frac{1}{2} \cdot \log_2 \left(1 + \frac{|h_{sd}|^2 P_{S1}}{N_0} + \frac{|h_{rd}|^2 P_{R1}}{N_0} \right). \tag{3}$$

If the relay node transmits in the AF protocol, the mutual information is expressed in [6] as

$$I_{\mathrm{AF}} = \frac{1}{2} \cdot \log_2 \left[1 + \frac{|h_{sd}|^2 P_{S1}}{N_0} + f\left(\frac{|h_{sr}|^2 P_{S1}}{N_0}, \frac{|h_{rd}|^2 P_{R1}}{N_0} \right) \right], \tag{4}$$

where $f(x, y) = xy/(1 + x + y)$.

In summary, the mutual information in the ISHDAF cooperative network is concluded as follows. When $|h_{sd}|^2 > T_1$, the source directly transmits to the destination

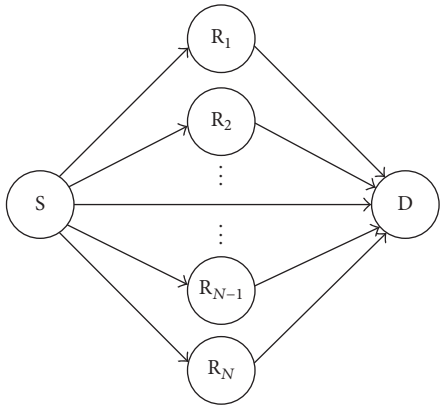


Figure 2: System model of the multiple relay communication.

successfully, and the mutual information is I_{DT}. When $T_1/2 < |h_{sd}|^2 \le T_1$, the direct transmission in link S-D failed. But the retransmission is successful in the second slot, and the mutual information is I_{DRT}. When $|h_{sd}|^2 \le T_1/2$ and $|h_{sr}|^2 > T_1$, the DF protocol is used to forward the messages and the mutual information is given as I_{DF}. When $|h_{sd}|^2 \le T_1/2$ and $|h_{sr}|^2 \le T_1$, the AF protocol is employed to forward the messages, and the mutual information is expressed as I_{AF}.

3. Multiple Relay System Model and Relay Selection Strategy

There is a typical two-hop multirelay cooperative network shown in Figure 2, consisting of the source node S, the destination node D, and N relay nodes R_i ($i = 1, 2, \ldots, N$). Equipped with single omnidirectional antenna, all the above nodes operate in a half-duplex mode. Hence the entire transmission procedure is also divided into two slots, under the independent and identically distributed (i.i.d.) AWGN channel noise. In addition, all channels are also supposed to be the flat Rayleigh fading channels, with fixed channel gains and independent channel status in each transmission. The destination node can select the appropriate relay nodes and notify them to forward the source information, where the channel status information (CSI) is available between all nodes through training sequence feedback.

Based on the system model, there is a relay selection strategy as the HDAF scheme, which chooses the AF or DF scheme to forward signals adaptively according to the channel status. If the channel status of link S-R_i is good enough for the relay to decode the source information, the DF protocol is selected to forward signals in the relay. Otherwise, the AF protocol is just used to prevent from the error propagation. According to the above strategy, the N relays can be divided into two sets for comparison. The optimal relay is then selected among the N relays in the premise of the maximum SNR at the destination. Finally, the specific process of the relay selection is listed as follows.

3.1. Determination of the Cooperative Relay Schemes. At first, there are some symbol definitions about the transmission power of the source and relay, respectively, that is, P_{S2} and P_{R2}, as well as the information transmission rate R bit/s. The channel noise is the AWGN with zero mean and variance N_0. Three channel parameters, such as $|h_{sd}|^2$, $|h_{sri}|^2$, and $|h_{rid}|^2$, are the channel gains of the links S-D, S-R_i, and R_i-D, respectively. They are subjected to the exponential distribution with parameters of $1/{\sigma_{sd}}^2$, $1/{\sigma_{sri}}^2$, and $1/{\sigma_{rid}}^2$. If the relays decode the signals from the source successfully, there will not be any interruption for transmission between source node S and relay node R_i. For this case, the mutual information in the transmission is deduced as

$$I_{sri} = \frac{1}{2} \cdot \log_2 \left(1 + \frac{P_{S2} \left| h_{sri} \right|^2}{N_0} \right). \tag{5}$$

Equation (5) should be larger than R to ensure that the transmission is not interrupted. And it can be transformed as $|h_{sri}|^2 > (2^{2R} - 1)N_0/P_{S2}$. Thus the threshold value can be set as $T_2 = (2^{2R} - 1)N_0/P_{S2}$. When $|h_{sri}|^2 \ge T_2$, the relay can decode the signal successfully. So the DF protocol is selected to forward the signal to avoid noise amplification. When $|h_{sri}|^2 < T_2$, the decoding in the relay failed. And the AF protocol is adopted to prevent error propagation.

Therefore, the candidate relays are divided into two sets according to whether successful decoding occurs or not in the relays, where the relays in set Ω_{DF} select the DF scheme to forward the signals in the second slot and others in set Ω_{AF} use the AF scheme. They are expressed as

$$\begin{aligned} \Omega_{\mathrm{DF}} &= \left\{ R_i : \left| h_{sri} \right|^2 > T_2 \right\}, \\ \Omega_{\mathrm{AF}} &= \left\{ R_i : \left| h_{sri} \right|^2 \le T_2 \right\}. \end{aligned} \tag{6}$$

3.2. Optimal Relay Selection. Since the optimal relay means to the maximized SNR in the destination, there are two steps to obtain it. Firstly, the best relays of ${R_b}^{\mathrm{DF}}$ and ${R_b}^{\mathrm{AF}}$ are chosen from the sets Ω_{DF} and Ω_{AF}, respectively. Then, the optimal relay R_b can be chosen between relay ${R_b}^{\mathrm{DF}}$ and relay ${R_b}^{\mathrm{AF}}$.

For the cooperation system with the AF scheme, the destination combines the signals from the source and the relay together by the maximum ratio combination (MRC) mechanism, and the instantaneous SNR at the destination is expressed as

$$\gamma^{\mathrm{AF}} = {\gamma_1}^{\mathrm{AF}} + {\gamma_2}^{\mathrm{AF}}, \tag{7}$$

where the instantaneous SNR is expressed as ${\gamma_1}^{\mathrm{AF}} = P_{S2}|h_{sd}|^2/N_0$ in the first slot and ${\gamma_2}^{\mathrm{AF}} = P_{S2}P_{R2}|h_{sri}|^2|h_{rid}|^2/[N_0(P_{S2}|h_{sri}|^2 + P_{R2}|h_{rid}|^2 + N_0)]$ in the second slot. Thus for the relay selection in set Ω_{AF}, the instantaneous SNR γ^{AF} is maximized to get the best relay, so the candidate relay ${R_b}^{\mathrm{AF}}$ with largest SNR ${\gamma_2}^{\mathrm{AF}}$ is obtained and expressed as

$$\begin{aligned} &{R_b}^{\mathrm{AF}} \\ &= \arg \max_{R_i \in \Omega_{\mathrm{AF}}} \left\{ \frac{P_{S2}P_{R2} \left| h_{sri} \right|^2 \left| h_{rid} \right|^2}{N_0 \left(P_{S2} \left| h_{sri} \right|^2 + P_{R2} \left| h_{rid} \right|^2 + N_0 \right)} \right\}. \end{aligned} \tag{8}$$

For the cooperation system with the DF protocol, the signals from the source and the relay are combined by the MRC scheme, and the instantaneous SNR at the destination is obtained as

$$\gamma^{\mathrm{DF}} = \min\left(\gamma_1^{\mathrm{DF}}, \gamma_2^{\mathrm{DF}}\right), \tag{9}$$

where the instantaneous SNR from the first slot is $\gamma_1^{\mathrm{DF}} = P_{S2}|h_{sri}|^2/N_0$ and that of the second slot is $\gamma_2^{\mathrm{DF}} = P_{S2}|h_{sd}|^2/N_0 + P_{R2}|h_{rid}|^2/N_0$. If all relays in set Ω_{DF} can succeed in decoding the signals, the instantaneous SNR at the destination is γ_2^{DF}. Then the optimal relay R_b^{DF} is represented as

$$R_b^{\mathrm{DF}} = \arg\max_{R_i \in \Omega_{\mathrm{DF}}} \left\{ \frac{P_{S2}\left|h_{sd}\right|^2}{N_0} + \frac{P_{R2}\left|h_{rid}\right|^2}{N_0} \right\}. \tag{10}$$

Finally, the optimal relay R_b can be chosen as the larger instantaneous SNR between R_b^{DF} and R_b^{AF}. Then it forwards the signal from the source in the corresponding mode. And it is expressed as

$$R_b = \max\left\{R_b^{\mathrm{AF}}, R_b^{\mathrm{DF}}\right\}. \tag{11}$$

Meanwhile, the mutual information of the cooperative transmission of the HDAF scheme by the proposed relay selection strategy is

$$I_{\mathrm{HDAF}} = \begin{cases} I_{\mathrm{DF}'}, & \left|h_{srb}\right|^2 \geq T_2 \\ I_{\mathrm{AF}'}, & \left|h_{srb}\right|^2 < T_2, \end{cases} \tag{12}$$

where $|h_{srb}|^2$ is the channel gain of link S-R_b.

4. Optimization of Power Allocation in the Relay Selection

The power allocation is optimized to obtain high power efficiency, where the entire power is taken as the constraint condition and the outage probability as the objective function. Then, the outage probability of the whole cooperation system is deduced analytically. And the Lagrange multiplier method is used to solve the optimal power allocation equation.

4.1. Deduction of Outage Probability. Outage probability is defined as the probability of failure in a transmission, which is one of the most used measures to evaluate the entire wireless communications. The transmission interruption occurs when the link capacity can not attain the required user rate. In other words, the mutual information of the transmission channel is smaller than the actual transmission rate. For a single relay network in the ISHDAF scheme, the direct source-destination transmission or retransmission is premised on the successful decoding of the received signals in the destination. Hence, the interruption only exists in the cooperative transmission. Based on the analysis in Section 3, the outage probability of the ISHDAF scheme can be deduced as

$$\begin{aligned} P_{\mathrm{out}}^{\mathrm{ISHDAF}} &= \Pr\left(\left|h_{sd}\right|^2 \leq \frac{T_1}{2},\ \left|h_{sr}\right|^2 > T_1,\ I_{\mathrm{DF}} < R\right) \\ &\quad + \Pr\left(\left|h_{sd}\right|^2 \leq \frac{T_1}{2},\ \left|h_{sr}\right|^2 \leq T_1,\ I_{\mathrm{AF}} < R\right). \end{aligned} \tag{13}$$

By replacing (3) and (4) into (13), and letting $\gamma = 2^{2R} - 1$, it gets

$$\begin{aligned} P_{\mathrm{out}}^{\mathrm{ISHDAF}} &= \Pr\left(\left|h_{sd}\right|^2 \leq \frac{T_1}{2},\ \left|h_{sr}\right|^2 > T_1,\ \frac{\left|h_{sd}\right|^2 P_{S1}}{N_0} \right. \\ &\quad \left. + \frac{\left|h_{rd}\right|^2 P_{R1}}{N_0} < \gamma\right) + \Pr\left(\left|h_{sd}\right|^2 \leq \frac{T_1}{2},\ \left|h_{sr}\right|^2 \right. \\ &\quad \left. \leq T_1,\ \frac{\left|h_{sd}\right|^2 P_{S1}}{N_0} + f\left(\frac{\left|h_{sr}\right|^2 P_{S1}}{N_0}, \frac{\left|h_{rd}\right|^2 P_{R1}}{N_0}\right) \right. \\ &\quad \left. < \gamma\right). \end{aligned} \tag{14}$$

The *probability density function* (PDF) of $|h_{sd}|^2$, $|h_{sr}|^2$, and $|h_{rd}|^2$ is expressed as

$$\begin{aligned} f_{|h_{sd}|^2}(x) &= \lambda_1 e^{-\lambda_1 x}, \\ f_{|h_{sr}|^2}(x) &= \lambda_2 e^{-\lambda_2 x}, \\ f_{|h_{rd}|^2}(x) &= \lambda_3 e^{-\lambda_3 x}, \\ & \qquad x > 0, \end{aligned} \tag{15}$$

where $\lambda_1 = 1/\sigma_{sd}^2$, $\lambda_2 = 1/\sigma_{sr}^2$, $\lambda_3 = 1/\sigma_{rd}^2$. Given the exponential distribution X and Y with parameters θ_1 and θ_2, respectively, the PDF of Z ($Z = X + Y$) is deduced by the integral equation $f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) dy$, and it is expressed in [23] as

$$f_Z(z) = \frac{\theta_1\theta_2}{(\theta_2 - \theta_1)} \cdot \left(e^{-\theta_1 z} - e^{-\theta_2 z}\right), \quad z > 0. \tag{16}$$

In addition, the probability distribution function W is approximately represented as $F_W(x) \approx (\theta_1 + \theta_2)x$ on condition of high SNR, when $W = XY/(1 + X + Y)$ [23].

Based on the above discussion, at high SNRs, the outage probability of the ISHDAF scheme can be calculated as follows:

$$\begin{aligned} \Pr\left(\left|h_{sd}\right|^2 \leq \frac{T_1}{2}\right) &= 1 - \exp\left(-\frac{T_1}{2\sigma_{sd}^2}\right) \\ &= 1 - \exp\left(-\frac{\gamma N_0}{2P_{S1}\sigma_{sd}^2}\right) \approx \frac{\gamma N_0}{2P_{S1}\sigma_{sd}^2}, \end{aligned} \tag{17}$$

and similar result is obtained as $\Pr(|h_{sr}|^2 \le T_1) \approx \gamma N_0/(P_{S1}\sigma_{sr}^2)$.

According to (16), take $N_0/(P_{S1}\sigma_{sd}^2)$ as θ_1 and $N_0/(P_{R1}\sigma_{rd}^2)$ as θ_2; there is the following deduction:

$$\begin{aligned}
&\Pr\left(\frac{|h_{sd}|^2 P_{S1}}{N_0} + \frac{|h_{rd}|^2 P_{R1}}{N_0} < \gamma\right) \\
&= \int_0^{\gamma} \frac{\theta_1\theta_2 \cdot \left(e^{-\theta_1 z} - e^{-\theta_2 z}\right)}{\theta_2 - \theta_1} dz \\
&\approx \frac{\theta_1\theta_2}{\theta_2 - \theta_1} \cdot \int_0^{\gamma} \left[(1-\theta_1 z) - (1-\theta_2 z)\right] dz \\
&= \frac{\theta_1\theta_2}{\theta_2 - \theta_1} \cdot \int_0^{\gamma} (\theta_2 - \theta_1)\, z\, dz = \frac{\theta_1\theta_2\gamma^2}{2} \\
&= \frac{\gamma^2 {N_0}^2}{2P_{S1}P_{R1}\sigma_{sd}^2\sigma_{rd}^2}.
\end{aligned} \tag{18}$$

With (18) and the descriptions mentioned above, there is

$$\begin{aligned}
&\Pr\left(\frac{|h_{sd}|^2 P_{S1}}{N_0} + f\left(\frac{|h_{sr}|^2 P_{S1}}{N_0}, \frac{|h_{rd}|^2 P_{R1}}{N_0}\right) < \gamma\right) \\
&= \frac{\gamma^2}{2} \cdot \frac{N_0}{P_{S1}\sigma_{sd}^2}\left(\frac{N_0}{P_{S1}\sigma_{sr}^2} + \frac{N_0}{P_{R1}\sigma_{rd}^2}\right).
\end{aligned} \tag{19}$$

Therefore, the outage probability of a single relay ISHDAF cooperative network is expressed as

$$\begin{aligned}
P_{\text{out}}^{\text{ISHDAF}} &= \frac{\gamma N_0}{2P_{S1}\sigma_{sd}^2}\left[\left(1 - \frac{\gamma N_0}{P_{S1}\sigma_{sr}^2}\right)\frac{\gamma^2 {N_0}^2}{2P_{S1}P_{R1}\sigma_{sd}^2\sigma_{rd}^2}\right. \\
&\left. + \frac{\gamma N_0}{P_{S1}\sigma_{sr}^2}\frac{\gamma^2}{2}\frac{N_0}{P_{S1}\sigma_{sd}^2}\left(\frac{N_0}{P_{S1}\sigma_{sr}^2} + \frac{N_0}{P_{R1}\sigma_{rd}^2}\right)\right] \\
&\approx \frac{(\gamma N_0)^4}{4P_{S1}^2\sigma_{sd}^4\sigma_{sr}^2}\left(\frac{1}{P_{S1}} + \frac{\sigma_{sr}^2}{P_{R1}\sigma_{rd}^2}\right).
\end{aligned} \tag{20}$$

Similarly, the outage probability in the HDAF relay selection strategy is denoted as

$$\begin{aligned}
P_{\text{out}}^{\text{HDAF}} &= \Pr\left(|h_{sr}|^2 > T_2,\ \frac{|h_{sd}|^2 P_{S2}}{N_0} + \frac{|h_{rd}|^2 P_{R2}}{N_0}\right. \\
&\left. < \gamma\right) + \Pr\left(|h_{sr}|^2 \le T_2,\ \frac{|h_{sd}|^2 P_{S2}}{N_0}\right. \\
&\left. + f\left(\frac{|h_{sr}|^2 P_{S2}}{N_0}, \frac{|h_{rd}|^2 P_{R2}}{N_0}\right) < \gamma\right).
\end{aligned} \tag{21}$$

It is obvious that $P_{\text{out}}^{\text{HDAF}}$ just lacks the part of "$|h_{sd}|^2 \le T_1/2$" when compared with $P_{\text{out}}^{\text{ISHDAF}}$. Finally, according to the above analyses, the outage probability is deduced as

$$P_{\text{out}}^{\text{HDAF}} \approx \frac{(\gamma N_0)^3}{2P_{S2}\sigma_{sd}^2\sigma_{sr}^2}\left(\frac{1}{P_{S2}} + \frac{\sigma_{sr}^2}{P_{R2}\sigma_{rd}^2}\right). \tag{22}$$

4.2. Optimization of Power Allocation. Using the Lagrange multiplier method, the optimized power allocation among the source and relay nodes to minimize the outage probability is produced as follows. For the ISHDAF scheme, with entire power as the constraint, as long as the fixed power P with $P_{S1} + P_{R1} = P$, the optimization problem can be denoted as

$$\begin{aligned}
&\min \quad \frac{(\gamma N_0)^4}{4P_{S1}^2\sigma_{sd}^4\sigma_{sr}^2}\left(\frac{1}{P_{S1}} + \frac{\sigma_{sr}^2}{P_{R1}\sigma_{rd}^2}\right) \\
&\text{s.t.} \quad P_{S1} + P_{R1} = P.
\end{aligned} \tag{23}$$

Let $P_{S1} = a_s P$, $P_{R1} = a_r P$, where $a_s + a_r = 1$. The Lagrange function is established as

$$\begin{aligned}
L(P_{S1}, P_{R1}, \lambda) &= \frac{(\gamma N_0)^4}{4P_{S1}^2\sigma_{sd}^4\sigma_{sr}^2}\left(\frac{1}{P_{S1}} + \frac{\sigma_{sr}^2}{P_{R1}\sigma_{rd}^2}\right) \\
&- \lambda P\left(a_s + a_r - 1\right).
\end{aligned} \tag{24}$$

Take partial derivation of (24) with respect to a_s and a_r, respectively, and then make them equal to zero; we can get

$$\frac{3}{a_s^4\sigma_{sr}^2} + \frac{2}{a_s^3 a_r\sigma_{rd}^2} - \lambda P = 0, \tag{25}$$

$$\frac{1}{a_s^2 a_r^2\sigma_{rd}^2} - \lambda P = 0. \tag{26}$$

By combining (25) and (26) together, and setting $e = a_s/a_r$, a quadratic equation with respect to variable e is obtained as

$$e^2 - 2e - \frac{3\sigma_{rd}^2}{\sigma_{sr}^2} = 0. \tag{27}$$

According to the root of (27) and $a_s + a_r = 1$, the optimized solutions of P_{S1} and P_{R1} satisfying (23) are obtained, respectively, as

$$P_{S1} = P \cdot \frac{\sigma_{sr}^4 + \sqrt{\sigma_{sr}^2 + 3\sigma_{rd}^2}}{2\sigma_{sr}^4 + \sqrt{\sigma_{sr}^2 + 3\sigma_{rd}^2}}, \tag{28}$$

$$P_{R1} = P \cdot \frac{\sigma_{sr}^4}{2\sigma_{sr}^4 + \sqrt{\sigma_{sr}^2 + 3\sigma_{rd}^2}}. \tag{29}$$

From (28) and (29), the powers P_{S1} and P_{R1} of the ISHDAF relay network in the optimized power allocation mainly depend on the channel coefficients of link S-R and link R-D, but not on that of link S-D.

Similarly, the power allocation optimization for the HDAF scheme can be defined as

$$\begin{aligned}
&\min \quad \frac{(\gamma N_0)^3}{2P_{S2}\sigma_{sd}^2\sigma_{sr}^2}\left(\frac{1}{P_{S2}} + \frac{\sigma_{sr}^2}{P_{R2}\sigma_{rd}^2}\right) \\
&\text{s.t.} \quad P_{S2} + P_{R2} = P.
\end{aligned} \tag{30}$$

Finally, the optimized powers P_{S2} and P_{R2} in the above power allocation are resolved as

$$P_{S2} = P \cdot \frac{\sigma_{sr}^4 + \sqrt{\sigma_{sr}^2 + 8\sigma_{rd}^2}}{2\sigma_{sr}^4 + \sqrt{\sigma_{sr}^2 + 8\sigma_{rd}^2}}, \tag{31}$$

$$P_{R2} = P \cdot \frac{\sigma_{sr}^4}{\left(2\sigma_{sr}^4 + \sqrt{\sigma_{sr}^2 + 8\sigma_{rd}^2}\right)}. \tag{32}$$

4.3. Optimal Relay Location. The power allocation depends on the channel coefficients, which are related to the distance between the relay and the source or the destination. To obtain the maximum outage probability gain by the proposed power allocation, an optimal relay location is deduced as follows.

To simplify the analysis of power allocation, we just constrain the situations where the distance between the link S-D and the link S-R-D is approximately equal, especially when the distance is quite large. Otherwise, under the same channel noise variance N_0 in the assumed condition, the transmission of the link S-R-D is much worse than that of link S-D, which loses the sense of relay selection. And it has been adopted similarly in [21]. Then, given normalization distance of link S-D (or approximate link S-R-D) as $d_{\mathrm{SD}} = 1$, and the distance x of link S-R, there is $d_{\mathrm{SR}} = x$ and $d_{\mathrm{RD}} = 1 - x$, where $0 < x < 1$. When the path loss factor is considered as $\alpha = 4$, the channel coefficients are obtained as $\sigma_{sr} = x^{-4}$, $\sigma_{rd} = (1-x)^{-4}$. Then, for the ISHDAF scheme, the transmit power of the source and relay is presented as

$$P_{S1} = P \cdot \frac{x^{-16} + \sqrt{x^{-8} + 3(1-x)^{-8}}}{2x^{-16} + \sqrt{x^{-8} + 3(1-x)^{-8}}}, \tag{33}$$

$$P_{R1} = P \cdot \frac{x^{-16}}{2x^{-16} + \sqrt{x^{-8} + 3(1-x)^{-8}}}. \tag{34}$$

Taking the derivation of variable x in (33), it obtains

$$\frac{dP_{S1}}{dx} = P \cdot \left(\frac{16x^{-17}}{2x^{-16} + \sqrt{x^{-8} + 3(x-1)^{-8}}} - \frac{\left(\left(4x^{-9} + 12(x-1)^{-9}\right)/\sqrt{x^{-8} + 3(x-1)^{-8}}\right) + 32x^{-17}}{x^{16}\left(2x^{-16} + \sqrt{x^{-8} + 3(x-1)^{-8}}\right)^2} \right). \tag{35}$$

Equation (35) is always greater than zero in the interval of $0 < x < 1$. So (33) is easily recognized as the monotone increasing function since the derivation of it, (35), is greater than zero. Then, P_{S1} is kept at about $0.5P$, when x is less than 0.5, and it tends to be P, when x is greater than 0.9. So the transmit power of the source is always larger than that of the relay in the proposed algorithm of the ISHDAF scheme.

Substituting (33), (34), and $\sigma_{sr} = x^{-4}$, $\sigma_{rd} = (1-x)^{-4}$ into (20), it gets

$$P_{\mathrm{out}}^{\mathrm{ISHDAF}} = \frac{(\gamma N_0)^4 x^8}{4P^3} \left(1 + \frac{x^{-16}}{x^{-16} + \sqrt{x^{-8} + 3(1-x)^{-8}}} + x^8 (1-x)^8 \left(2x^{-16} + \sqrt{x^{-8} + 3(1-x)^{-8}}\right) \right). \tag{36}$$

From (36), the systematic outage probability is relatively small in the case that the distance of link R-D is larger than that of link S-R. In other words, the relay node R is relatively close to the destination node D for better outage probability. However, when the relay node is approximately located in the middle between the source and destination node, the entire outage probability is minimal. Simultaneously, a theoretical analysis about the outage probability of the HDAF system just resembles that of the ISHDAF system. But when the relay node is close to the destination node, the outage performance decreases, and the optimal relay location closely approaches to the source node than that of the ISHDAF system. Since the order of (36) is too high to obtain the analytic solution, only numerical results are available and they will be given in the successive simulation related in Section 5.

4.4. Diversity Gain. Given the diversity gain in the proposed ISHDAF scheme, it should be divided into three cases as follows.

First, when $|h_{sd}|^2 > T_1$, the source transmits directly to the destination successfully in the first time slot, and I_{DT} in (1) shows that the signal is transmitted only in one path. So it extracts one diversity gain in the direct transmission.

Second, the direct transmission is failed, but the retransmission is successful in the second time slot, when there is $T_1/2 < |h_{sd}|^2 \leq T_1$. The SNR received by the destination node is twice as straightforward, while the transmission still experiences only one path. Therefore, the diversity gain is still one.

Third, the relay node starts to forward signals in the AF protocol or the DF protocol. In these two cases, the destination node receives signals from two links. So the system achieves two diversity gains in the cooperative transmission by the relay nodes. In addition, for the multirelay selection strategy under the HDAF scheme, it employs DF or AF mode to forward signals adaptively according to the channel status. Because both forward modes are required for R-D transmission, the full diversity gain of 2 is then obtained.

In summary, the proposed ISHDAF scheme obtains much more outage probability performance gain by the direct link retransmission rather than the relay forwarding, when compared with the IHDAF one in [8]. In other words, the diversity gain in the ISHDAF scheme is better than that of the IHDAF one, because the overall channel transmission effect

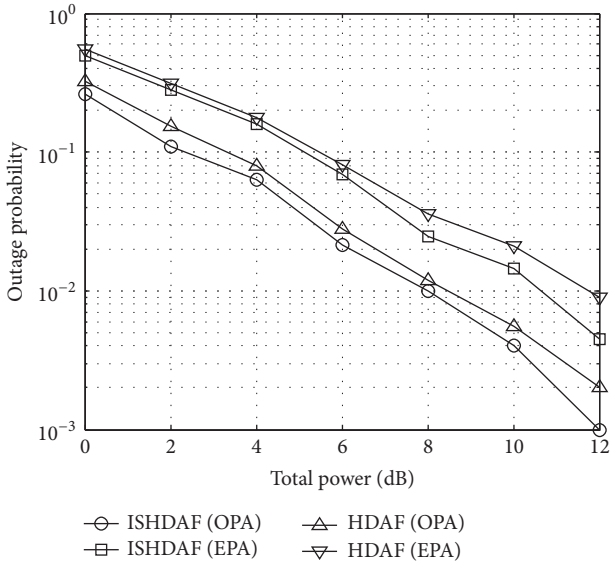


Figure 3: Outage probabilities between the OPA and the EPA in different forwarding strategies.

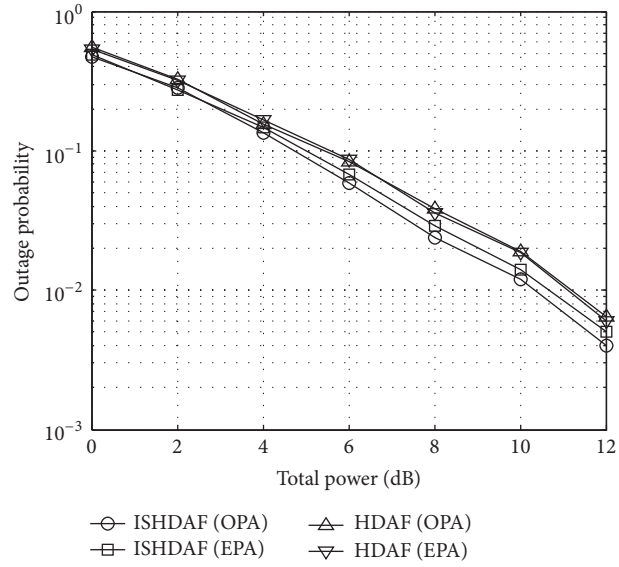


Figure 4: Outage probabilities between the OPA and the EPA under the specific relay location.

(retransmission and then cooperative relay communication) in the former is superior to that (just cooperative relay communication) in the latter.

5. Simulation Results and Analysis

To validate the proposed power allocation optimization algorithm for the ISHDAF and the HDAF relay selection strategy, two typical kinds of cooperative network are simulated and analyzed. For a single relay network, the outage performances by the proposed HDAF and ISHDAF strategy are compared. Besides, the optimized power allocation (OPA) and the equal power allocation (EPA) algorithms are employed in the two strategies, respectively, for comparison. In addition, for a HDAF multirelay selection network, the outage performances of the whole system with different relay numbers are simulated and compared. And the simulations for the validation of the optimal relay location are also performed in both the HDAF and the ISHDAF single relay network.

The simulation parameters are set as follows. The transmission rate is set as $R = 1$ bit/s. The node distance is fixed as $d_{SD} = 1$, where d_{ij} is the normalized distance between node i and node j. All channels are Rayleigh flat fading channels, and $\alpha = 4$ stands for the path loss factor. The channel noise is an AWGN with zero mean and variance $N_0 = 1$. By the Binary Phase Shift Keying (BPSK) modulation, the results are simulated and shown as follows.

Figure 3 shows the comparison of the outage probability between the OPA and the EPA scheme. They are at both the ISHDAF and the HDAF single relay network, respectively, with distance parameters of $d_{SR} = 0.8$ and $d_{RD} = 0.2$. From Figure 3, in the ISHDAF strategy, the OPA scheme achieves a much better outage performance gain over the EPA scheme. And it is also true in the HDAF strategy. Moreover, the ISHDAF strategy has larger gain over the HDAF strategy, when they are under the same power allocation. There are nearly 0.1695 gains for the OPA scheme compared with the EPA scheme in the ISHDAF strategy and about 0.045 gains for the ISHDAF strategy compared with the HDAF strategy employed in the OPA scheme, at a total power of 2 dB. With the increasing of the total power, the outage probability decreases, and the gains become gradually small. The reason is that when the relay node is far from the source, the outage performance of link S-R is poor. So the threshold is decreased too in the OPA scheme, which guarantees the direct transmission or retransmission in link S-D with the ISHDAF strategy. Also, from (14) to (18), there is a condition as $|h_{sd}|^2 \leq T_1/2$ for the ISHDAF to calculate the outage probability. So the outage performance of the ISHDAF scheme outperforms that of the HDAF one in the EPA scheme.

The outage performances of different schemes with the distance parameters of $d_{SR} = 0.2$ and $d_{RD} = 0.8$ are compared in Figure 4. At both the ISHDAF and the HDAF cooperation relay network, the OPA and EPA scheme have almost the same performance. Since the relay transmission opportunity increased when the relay node approaches the source node, the whole transmission in the ISHDAF scheme is similar to that of the HDAF one. And from (25) to (29), the node power based on the OPA scheme is also similar to that of the EPA scheme. At the same time, the OPA algorithm in the ISHDAF strategy achieves a better gain, when the distance of link S-R is larger than that of link R-D.

There is also a comparison of outage performance between the OPA and the EPA scheme, in the HDAF multirelay selection network. They are simulated with different number of relays, under the distance parameters of $d_{SR} = 0.8$ and $d_{RD} = 0.2$ and the results are shown in Figure 5. From Figure 5, the outage probability is reduced with the increased number of the relays. This is because many more numbers

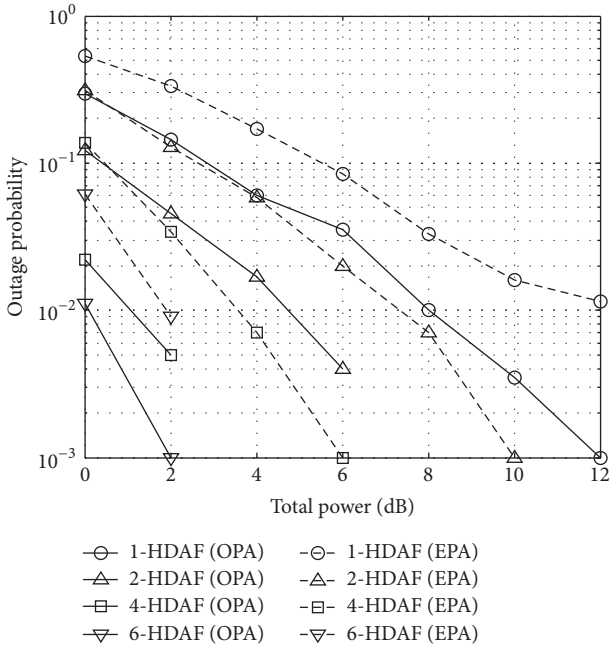


Figure 5: Outage probabilities among different number of the relays in the HDAF scheme.

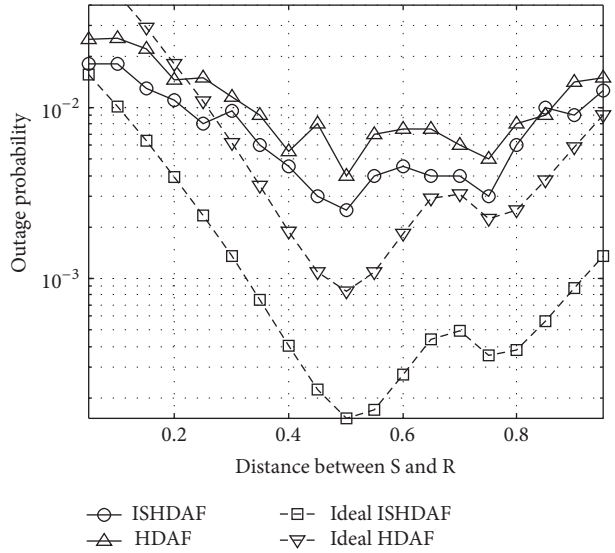


Figure 6: Outage probabilities among different relay locations in the OPA scheme.

of relays result in better channel quality of the best selected relay. It also leads to the increased mutual information in transmission; thus the outage probability of the system reduces correspondingly. In addition, Figure 5 shows that, with the same relay number in the HDAF cooperative system, OPA scheme has a significant outage performance gain than that of the EPA scheme. For instance, at a total power of 2 dB, there are about 0.083 gains for OPA scheme when compared with those of the EPA scheme under 2 relays. Since the declined threshold results in many more opportunities for the DF protocol employed at the relay node, at such relay location, in this case, the DF protocol outperforms the AF protocol similar to that in [24].

To verify the theoretical analysis of the optimal relay location for the proposed algorithm, some simulations are performed in the cooperative single relay network. For the different relay locations, the outage probabilities of the ISHDAF and the HDAF strategies are illustrated in Figure 6, with the fixed total power of 10 dB. From the results, the outage probability is really low when the relay node R is relatively close to the destination node D. It turns out to be the lowest (i.e., best) one for the relay node R at the middle position between the source node S and the destination node D, which is consistent with the theoretical analysis indicated by (32). The most possible reasons mainly rely on the following reasons. When the relay is a little far from the source, the cooperative scheme at the relay performs better under the node power allocation, which is related to the link performance. Moreover, when the relay node is just in the middle between them, the performance of both link S-R and link R-D is good. The whole system can thus obtain maximum benefit in the proper power allocation ratio of the source and the relay under the proposed algorithm. In addition, the simulation shows that the outage probability performance of the ISHDAF scheme is always better than that of the HDAF scheme in the proposed algorithm.

6. Conclusion

In this paper, an optimized power allocation algorithm is proposed, which mainly employs the two-hop single relay network with ISHDAF scheme and the multirelay selection strategy with HDAF scheme. The optimization of the proposed algorithm is just to minimize the outage probability of system under the constraints of total power of the source and relay nodes. In addition, the proposed scheme can only occupy a small amount of time complexity to obtain the power allocation optimization in a cooperative communication system. In the simulations, the proposed algorithm is applied in the ISHDAF and the HDAF scheme with the well-known three-node models, respectively. Simulation results show that the proposed algorithm can achieve much larger gain by the ISHDAF scheme than that by other ones. Also, for different number of the relay nodes in a cooperative network, the simulation comparisons show that the proposed algorithm by the HDAF relay selection strategy has a significant validity in power allocation. Simultaneously, the optimal relay location by the suggested algorithm is also established for an even better gain over current schemes. Therefore, the proposed optimized power allocation and relay location selection algorithm can be effectively adopted in cooperative IoT relay systems in practice for high power efficiency and good outage probability performance.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Zhejiang Provincial Natural Science Foundation of China (no. LZ14F010003, no. LY17F010019), the National Natural Science Foundation of China (no. 61471152), the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University (no. 2014D02), and the Zhejiang Provincial Science and Technology Plan Project (no. 2015C31103, no. LGG18F010011).

References

[1] J.-S. Han, J.-S. Baek, S. Jeon, and J.-S. Seo, "Cooperative networks with amplify-and-forward multiple-full-duplex relays," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 2137–2149, 2014.

[2] P. K. Sharma and P. Garg, "Performance analysis of full duplex decode-and-forward cooperative relaying over Nakagami-m fading channels," *Transactions on Emerging Telecommunications Technologies*, vol. 25, no. 9, pp. 905–913, 2014.

[3] T. X. Vu, P. Duhamel, and M. D. Renzo, "Performance analysis of network coded cooperation with channel coding and adaptive DF-based relaying in Rayleigh fading channels," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1354–1358, 2015.

[4] S. S. Ikki, M. Uysal, and M. H. Ahmed, "Performance analysis of incremental-relay-selection decode-and-forward technique," in *Proceedings of the 2009 IEEE Global Telecommunications Conference, GLOBECOM 2009*, December 2009.

[5] T. T. Duy and H. Y. Kong, "On performance evaluation of hybrid decode-amplify-forward relaying protocol with partial relay selection in underlay cognitive networks," *Journal of Communications and Networks*, vol. 16, no. 5, pp. 502–511, 2014.

[6] Q. F. Zhou and F. C. M. Lau, "Two incremental relaying protocols for cooperative networks," *IET Communications*, vol. 2, no. 10, pp. 1272–1278, 2008.

[7] J. Wu, D. He, J. Bao, X. Xu, and B. Jiang, "Optimization of power allocation and outage probability of cooperative relay networks," in *Proceedings of the 8th International Conference on Wireless Communications and Signal Processing, WCSP 2016*, October 2016.

[8] Z. Bai, J. Jia, and C. X. Wang, "Performance analysis of SNR-based incremental hybrid decode-amplify-forward cooperative relaying protocol," *IEEE Transactions on Communications*, vol. 63, no. 6, pp. 2094–2106, 2015.

[9] L. Wu, Y. Wang, J. Han et al., "Optimal power allocation for wireless uplink transmissions using successive interference cancellation," *KSII Transactions on Internet and Information Systems*, vol. 10, no. 5, 2016.

[10] K. K. Gurrala and S. Das, "Minimized ser based power allocation for multi HDAF relay cooperative network using differential evolution algorithm," in *Proceedings of the 11th IEEE India Conference, INDICON 2014*, December 2014.

[11] H. Xiao and S. Ouyang, "Power allocation for a hybrid decode-amplify-forward cooperative communication system with two source-destination pairs under outage probability constraint," *IEEE Systems Journal*, vol. 9, no. 3, pp. 797–804, 2015.

[12] A. Kwolek-Folland, "Power allocation for two source destination pair cooperative communication system under the usage probability constraint," *Journal of the Japanese Association for Petroleum Technology*, vol. 4, no. 2, pp. 74–83, 2015.

[13] Z.-K. Zhou and Q. Zhu, "Joint power allocation and multi-relay selection scheme based on system outage probability," *Journal of China Universities of Posts and Telecommunications*, vol. 21, no. 5, pp. 9–16, 2014.

[14] M. Qian, C. Liu, Y. Fu, and W. Zhu, "A relay selection and power allocation scheme for cooperative wireless sensor networks," *KSII Transactions on Internet and Information Systems*, vol. 8, no. 4, pp. 1390–1405, 2014.

[15] T. T. Duy and H.-Y. Kong, "Performance analysis of hybrid decode-amplify-forward incremental relaying cooperative diversity protocol using SNR-based relay selection," *Journal of Communications and Networks*, vol. 14, no. 6, pp. 703–719, 2012.

[16] Y. Xing, Y. Chen, C. Lv, Z. Gong, and L. Xu, "Swarm intelligence-based power allocation and relay selection algorithm for wireless cooperative network," *KSII Transactions on Internet and Information Systems*, vol. 10, no. 3, pp. 1111–1130, 2016.

[17] H. Wang and Q. Zhu, "Power Allocation Scheme based on System Outage Probability for Energy-Harvesting Cooperative Networks," in *Proceedings of the 2015 4th National Conference on Electrical, Electronics and Computer Engineering*, pp. 617–622, Xi'an, China, December 2015.

[18] Q. Zhang, Z. Feng, T. Yang, and W. Li, "Optimal power allocation and relay selection in multi-hop cognitive relay networks," *Wireless Personal Communications*, vol. 86, no. 3, pp. 1673–1692, 2016.

[19] J. Zhang, J. Jiang, J. Bao, B. Jiang, and C. Liu, "Improved relay selection strategy for hybrid decode-amplify forward protocol," *Journal of Communications*, vol. 11, no. 3, pp. 297–304, 2016.

[20] M. H. D. Khan and M. S. Elmusrati, "Performance analysis of power allocation and relay location in a cooperative relay network," in *Proceedings of the 17th IEEE International Conference on Advanced Communications Technology, ICACT 2015*, pp. 444–449, July 2015.

[21] L. Han, J. Mu, S. Liu et al., "Optimization of power allocation and relay location for decode-and-forward relaying in the presence of co-channel interference," in *Proceedings of the Third International Conference on Communications, Signal Processing, and Systems*, pp. 319–327, Springer International Publishing, 2015.

[22] Y. W. P. Hong, W. J. Huang, and C. C. J. Kuo, *Cooperative Communications and Networking*, Cambridge University Press, Cambridge, UK, 2009.

[23] W. Qingtao, "Research on outage probability in relay systems," *Computer Engineering and Applications*, vol. 49, no. 11, pp. 24–26, 2013.

[24] O. J. Pandey, A. Trivedi, and M. K. Shukla, "Outage performance of decode-forward and amplify-forward protocols in cooperative wireless communication," in *Proceedings of the 10th IEEE and IFIP International Conference on Wireless and Optical Communications Networks, WOCN 2013*, pp. 1–5, July 2013.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 2135182, 10 pages
https://doi.org/10.1155/2017/2135182

Research Article

A High Throughput Anticollision Protocol to Decrease the Energy Consumption in a Passive RFID System

Hugo Landaluce,[1,2] Laura Arjona,[1,2] Asier Perallos,[1,2] Lars Bengtsson,[3] and Nikola Cmiljanic[1,2]

[1]*DeustoTech-Deusto Foundation, Avda. Universidades, 48007 Bilbao, Spain*
[2]*University of Deusto, Avda. Universidades, 48007 Bilbao, Spain*
[3]*Department of Computer Science and Engineering, Chalmers University of Technology, Göteborg, Sweden*

Correspondence should be addressed to Hugo Landaluce; hlandaluce@deusto.es

Received 28 April 2017; Accepted 12 October 2017; Published 8 November 2017

Academic Editor: Haiyu Huang

One of the main existing problems in Radio Frequency Identification (RFID) technology is the tag collision problem. When several tags try to respond to the reader under the coverage of the same reader antenna their messages collide, degrading bandwidth and increasing the number of transmitted bits. An anticollision protocol, based on the classical Binary Tree (BT) protocol, with the ability to decrease the number of bits transmitted by the reader and the tags, is proposed here. Simulations results show that the proposed protocol increases the throughput with respect to other recent state-of-the-art protocols while keeping a low energy consumption of a passive RFID system.

1. Introduction

The Radio Frequency Identification (RFID) technology use is being increased in applications where autoidentification methods are needed [1–5]. RFID is used to identify codes stored in devices, called tags, using radio frequency waves. These tags are attached to different objects which can be identified without an existing line of sight. RFID is a low intrusive technology which can be easily adapted to the Internet of Things [2].

An RFID system is basically composed of two elements: a reader and one or more tags. The reader is an electronic device with a RF module, a control unit, and one or more antennas which establish a bidirectional communication with the second device, the tags; tags include an IC-chip and an antenna and are attached to the object to be identified. Tags can be active (battery-powered) or passive (obtain power from reader's signal). Passive tags are widely used due to their low price and the absence of batteries; however, they have a lower coverage than the active ones. This paper is focused on passive RFID systems consisting of a reader and different numbers of passive tags.

Typically, an RFID system may contain several tags coexisting and transmitting to the reader at the same time using the same channel (the air). This fact can cause the cancellation of tags' responses. The reader may not be able to decode their waveforms and tags will be forced to retransmit their messages causing a decrease in the time needed to be identified and an increase in the energy consumed by the reader. This problem is called the tag collision problem [1].

This problem is faced using anticollision protocols. Several protocols are presented in the literature and can be mainly classified into Aloha based and tree based protocols [3]. Aloha based protocols, classified as probabilistic since tags' responses, are randomly organized and distribute responses among slots. The current standard EPC global Class 1 Gen 2 [6] belongs to this category. Research in these types of protocols is focused on the optimal distribution of tags' responses in a timeline. Tree based protocols, on the other hand, are classified as deterministic since they are supposed to read all the tags in the interrogation zone [3]. These protocols usually split the set of tags upon collisions until achieving a successful response from all the tags; however, some recent solutions provide an estimation phase to define

initial subsets which are easier to be identified by the reader [7, 8].

The energy consumption of an anticollision protocol has been directly related to active RFID systems due to the use of batteries in active tags [9–12]. Passive RFID systems are increasingly being used with portable readers which means that the energy consumed by the reader is becoming important, and the anticollision protocol used affects it [13]. A deep analysis of the energy consumption of anticollision protocols in passive RFID is given in [14]. The window procedure is presented and applied to the query tree (QT) protocol [15], in the query window tree protocol (QwT), to decrease the energy consumed by a passive RFID system. QwT manages to decrease the number of bits transmitted by the tags which produces a significant decrease in the energy consumed by the reader. The use of the window forces the anticollision protocol to add a new type of slot, called go-on slot, which increases the total number of slots and, therefore, the total number of bits transmitted by the reader. This is specially problematic when the reader is asking the tags for the last part of their identification code (ID), since it needs to transmit longer commands [14].

This paper presents the contribution of a new protocol called binary window tree (BwT). This protocol proposes adding an additional internal counter to the tags which indicates the last bit of their transmitted ID. This modification allows the adoption of the window in BT. Additionally, the reader is adapted to working with the window which is tuned to dynamically adapt its size during the procedure of the identification using tags with memory. This protocol is later compared to other state-of-the-art protocols in the literature. The results show that the novel proposed protocol increases the throughput of the RFID system while it saves energy using passive tags.

Subsequently, the rest of the paper is organized as follows. Section 2 provides background information and related work on anticollision protocols. Section 3 presents the proposed BwT protocol. In Section 4 the simulation results of the comparison with state-of-the-art protocols are presented. And Section 5 concludes this paper.

2. Background

In order to properly understand the proposed work, some concept definitions are introduced here:

(i) A *slot* determines the period of time that includes a reader command and a tag response. This time is usually fixed, but some state-of-the-art works consider it dynamic, as is the case of this work. Three types of slots are usually considered upon the number of tags' responses: a collision occurs when more than one tag responds in the same slot period; a success slot is given when only one tag responds to a reader command; and an idle slot occurs when no tag responds to a reader command.

(ii) An *interrogation cycle* is the period of time the reader needs to identify the whole set of available tags. An interrogation cycle is composed of several slots.

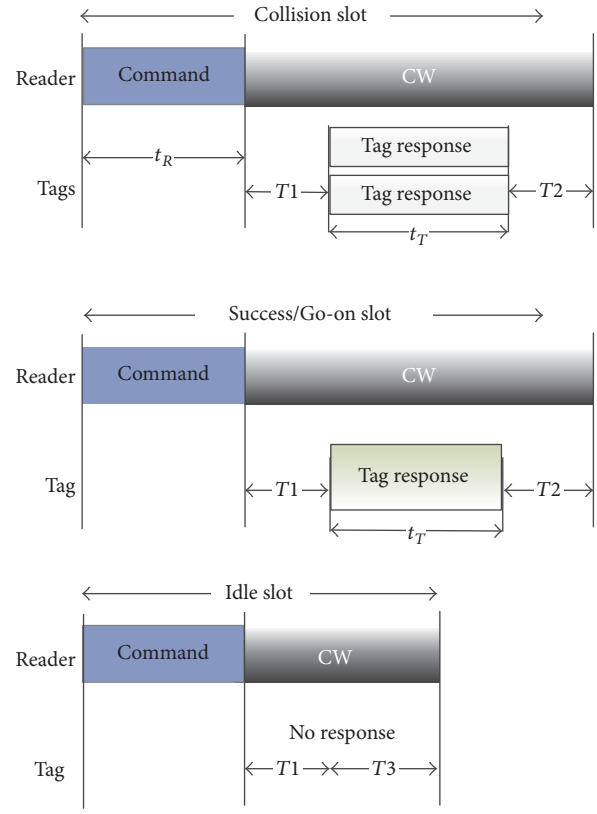


Figure 1: Example of a collision/idle/success/go-on slot in the transmission model used.

(iii) The metric *throughput* is conceived as the number of read tags in the unit of time.

Using these main concepts the transmission model between the reader and the tags can be explained. This model assumes an ideal channel for transmissions. No physical-layer and no capture effect (when the reader decodes only the tag response with the highest power in a collision slot) are considered here. All tags in the antenna range remain correctly energized during the interrogation cycle; all tags' responses are synchronized; and lastly, a collision occurs only when two different messages or bits are simultaneously transmitted since error transmissions are not considered. These assumptions are extensively made in other similar anticollision protocol works proposed [7, 8, 14, 16]. The transmission model is explained below.

2.1. Transmission Model. The transmission model used is defined in [6], which corresponds to the EPC global C1G2 standard.

Figure 1 shows the link timing of the three types of slots mentioned (collision, idle, and success) and the go-on slot explained below. Also, in Definition of Symbols and Variables a list of all the variables used in the paper is included. The reader starts transmissions using commands during time t_R. The reader holds the RF downlink carrier, also called continuous wave CW, so that tags can harvest energy and respond with their ID. After every reader command there is a

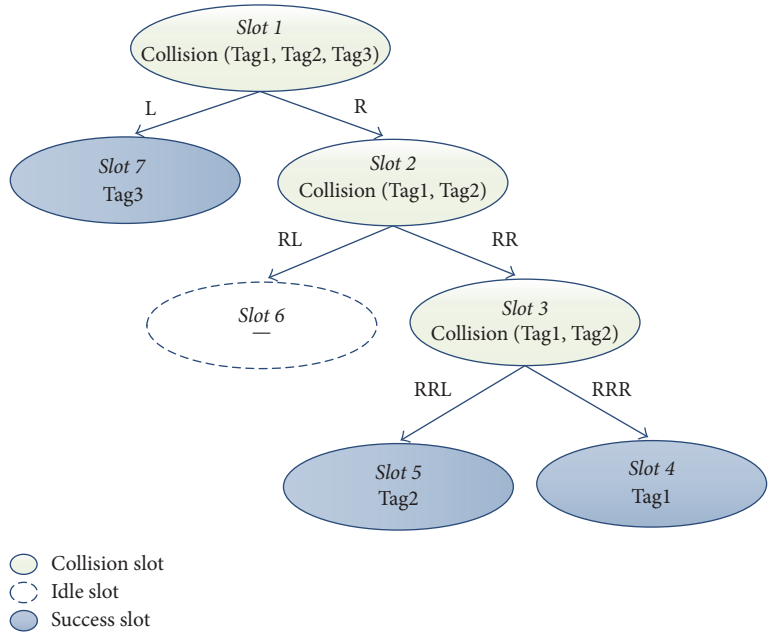


Figure 2: Example of an identification cycle with BT.

time T_1 needed for the tags to generate their responses and a time T_2 needed for the reader to receive all the transmissions; a slot will be considered idle when the reader waits for the tags' responses for a time T_3. Additionally, the tag's response is produced during time t_T.

With respect to the energy consumption model, E represents the energy consumed by the reader. It is a function of the time it spends transmitting and receiving information. An energy model is proposed where the reader transmits the command and the CW to power up passive tags with power P_{tx}. In addition, the reader will require extra power P_{rx} when receiving data from the tags. Therefore, the expression used to calculate the total energy consumed during the interrogation cycle is shown in

$$\begin{aligned} E &= E_c + E_i + E_s \\ &= \sum_{j=0}^{c+s} \left[P_{tx} \times \left(t_{Rj} + T_1 + t_{Tj} + T_2 \right) + P_{rx} \times t_{Tj} \right] \\ &\quad + \sum_{j=0}^{i} \left[P_{tx} \times \left(t_{Rj} + T_1 + T_3 \right) \right]. \end{aligned} \tag{1}$$

Here, E_c, E_s, and E_i represent the energy consumed during collision, success, and idle slots and c, s, and i represent the number of collision, success, and idle slots, respectively.

2.2. Related Work. Here, some of the most relevant tree based protocols in the literature are presented. This will later be simulated and compared in Section 4.

2.2.1. Binary Tree Protocol. The Binary Tree (BT) protocol was firstly applied to RFID by Hush and Wood in [17]. It uses a tree to organize and identify all the tags into the reader's antenna range. Every time a collision occurs between tags' responses the responding tags are split into two different subgroups. These subgroups become increasingly smaller until they are split into two tags, which can therefore be identified (see Figure 2).

The reader consecutively interrogates all these subgroups and tags outside these groups which wait until their subgroup is chosen for the interaction with the reader. Every time the protocol reaches a leave of the tree, the reader identifies the tag and goes back to the last subgroup produced which starts to be split in a similar manner.

Bertsekas and Gallager then included an internal counter on every tag which they modify upon the reader commands [18]. The reader can indicate the three possible slot states. Upon an idle slot, tags decrease their counters by 1; upon a collision slot, transmitting tags choose between 0 and 1 randomly, and waiting tags increase their counter by 1; upon a success, the transmitting tag goes to sleep mode and the rest of the tags decrease their counters by 1.

2.2.2. Fast Tree Traversal Protocol. Choi et al. proposed the Fast Tree Traversal Protocol (FTTP) [7] that uses the Maximum Likelihood Estimation (MLE) to calculate the number of available tags on the internal nodes of the left branch of the tree.

The protocol uses BT until the first tag is identified. This process covers the full left branch of the tree. Then, the protocol goes step by step back on the different internal nodes and calculates the number of available tags on the right branch of those nodes using MLE. FTTP estimates that the number of tags available on the right branch should be equal to the number of tags on the left one which is already known. This is used to modify the internal counters of the tags

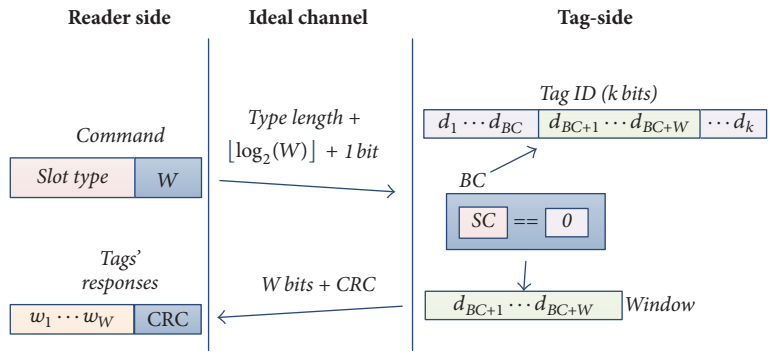


Figure 3: Example of a communication slot between the reader and a tag.

and separate their responses. FTTP then uses BT whenever a collision is produced during the identification.

2.2.3. Optimal Binary Tracking Tree Protocol. A BT based protocol with a bit estimator and bit tracking technology, referred to as the Optimal Binary Tracking Tree (OBTT) protocol, is given in [8]. It first employs a bit estimation algorithm to estimate the existing number of tags. Tags choose to swap one of the bits in a 1-bit string. This string is split into $x = 1/k - 1$ segments of the tag ID length k and each portion is transmitted separately to the reader. The reader uses Chebyshev's inequality [16] to calculate the estimated number of bits $\widehat{n}$ with the number of selected and nonselected bits received in the tags' responses. Once $\widehat{n}$ is obtained, the optimal number of queries m to initialize a query stack is calculated using $m = 0.595824\ \widehat{n}$. Afterwards, m queries are generated and pushed onto the stack. The rest of the process is solved using BT. Although the slot efficiency obtained by OBTT is very high, the preprocessing increases the energy consumption of the protocol, especially when $x > 1$.

2.2.4. Query Window Tree (QwT). A protocol named Query window Tree (QwT) protocol which uses a methodology to manage the number of bits transmitted by the tags is presented in [14]. The "window" is a bit string of size W, where $0 < W < k$, responded to by the tags instead of their full ID. The QwT is a modification of the query tree (QT) protocol [15] which includes the window methodology in tags' responses. As in QT, the reader transmits a command filled with a bit string, called query of length L, and the value of W calculated at the reader side. Tags compare the query with their initial part of their ID and those matching the reader's query respond exclusively to the following number of bits specified by W from the ID.

Three possible slot statuses can happen at the reader side, idle, collision, and success, as well as in QT. However, QwT introduces a new type of slot called go-on, when the reader receives similar windows of one or more tags simultaneously and obeys the following condition $L + W < k$. In this case, the reader has not received the full ID of the tag; therefore it cannot consider the tag identified.

The window alleviates tags from transmitting large number of bits upon a collision. However, low W values can cause the number of go-on slots to increase. Heuristic function (2) is proposed in [14] which provide an updated W value in case of a go-on slot to decrease its number. β parameter is used to tune the heuristic function.

$$W = \frac{k}{(k-\beta)^2} \times L^2. \tag{2}$$

Using the window, tags transmit fewer bits than that of the QT to be identified. However, the number of reader bits required by QwT is larger than that of QT since it uses a lot of longer queries than the latter.

3. The Proposed BwT Protocol

Here the BwT protocol is presented. This protocol implements the window methodology into BT. BwT tags use two counters: a slot counter SC and a bit counter BC; and the reader, on the other hand, uses another counter rBC. All tags update their SC on every slot and transmit when $SC = 0$. The BC indicates the first bit of its ID to transmit and is updated adding the W value only when the tag is in the transmitting state ($SC = 0$) and a notification of a go-on slot is received. Therefore, whenever a tag has its counter $SC = 0$, it transmits W bits starting from BC bit (see Figure 3) and a cyclic redundancy check code (CRC), similar to that demanded on the EPC C1G2 standard [6], to differentiate their responses at the reader.

The reader of BwT follows the same procedure on every slot. First, it receives the window transmitted by the tags and checks the type of slot it has received. According to the type of slot, the reader updates its counter rBC to monitor the length of the already acquired ID of the transmitting tag or tags and calculates the new W value to be transmitted. Then, it transmits a new command containing the status of the last slot received and the newly calculated W.

As mentioned before, slots can be idle, success, collision, and go-on. Pseudocode for the BwT reader and tags is shown in Algorithms 1 and 2 to perform an identification cycle. This pseudocode is executed during a specified time until all the tags in the interrogation zone are identified. In the beginning, the reader is initialized with $rBC = 0$ and $W = 1$. The new command is assembled with the type of the last slot and the value of W and is broadcast to the tags.

```
(1) Command = ε
(2) W = 1
(3) rBC = 0
(4) tagID = []
(5) k = ID.length
(6) while(unidentified tags) do
(7)     broadcast(Command,W)
(8)     [winID,crcOK] = receiveResponses
(9)     if isempty(winMatch) then Command = Idle
(10)    else crcOK==0 then
(11)      Command = Collision; W = 1
(12)      LIFOpush(tagID)
(13)    else
(14)      tagID = tagID+winID
(15)      rBC = tagID.length
(16)      if rBC + W < k then
(17)        Command = Go-On
(18)        W = f(rBC)
(19)      else
(20)        Command = Success
(21)        W = 1; LIFOpop(tagID)
```

Algorithm 1: Pseudocode of BwT. Reader procedure.

```
(22) sleep = false
(23) SC = 0; BC = 0
(24) nW = 1; oW = 1
(25) while (not sleep) do
(26)   receive(Command,nW)
(27)   switchCommand
(28)   case Idle:
(29)     SC=SC−1
(30)   case Collision:
(31)     ifSC==0 then SC=rand()%2
(32)     else SC=SC+1
(33)   case Go-On:
(34)     ifSC==0 then BC = BC + oW
(35)   case Success:
(36)     ifSC==0 then sleep=true
(37)     else SC=SC−1
(38)   ifSC==0 then
(39)     backscatter(ID[BC:BC+nW]); oW = nW
```

Algorithm 2: Pseudocode of BwT. Tag procedure.

After a certain period of time or after receiving a response, the reader identifies the type of slot according to the CRC consistency and takes the following actions depending on the slot identified:

(i) Idle slot (Algorithm 1 line (9)): when no response is received, the reader broadcasts a new command "Idle" with an invariant W.

(ii) Collision slot (Algorithm 1 lines (10)–(12)) occurs when the reader decodes the received windows and these are not CRC consistent. The reader needs to remember the already acquired ID since it may belong to different tags with a common partial ID. Therefore, the accumulated tag ID received at that point is stored into a Last Input First Output (LIFO) stack. Then the reader indicates the new command "Collision" with W set to 1 and broadcasts it to the tags.

(iii) Go-on slot (Algorithm 1 lines (16)–(18)) is when the CRC validates the received window and the condition $rBC + W < k$ is met. The reader updates the partial ID received and its length with the received window (Algorithms 1 lines (14)-(15)). Then, W is updated using the exponential heuristic function shown in (3). How W is adjusted is given in Section 3.1.

$$f(rBC) = k\left(1 - e^{-\beta \times rBC}\right). \tag{3}$$

This expression is a practiced deduction to balance the number of tag transmitting bits and go-on slots. It allows the reader to choose small W when the probability of collision is prone to increase, providing a small colliding tag bit rate; while it offers larger W when rBC increases (and, thus, the probability of collision decreases), contributing to decrease of the number of go-on slots. In addition, this W value is always delimited by the expression

$$W = \begin{cases} f(rBC), & W \leq k - rBC \\ k - rBC, & W > k - rBC. \end{cases} \tag{4}$$

Lastly, a new command "go-on" with the calculated W is broadcast to the tags.

(iv) Success slot (Algorithm 1 lines (20)-(21)) is when the CRC validates the received window and the tag ID is uniquely defined: $rBC+W = k$. Then, the ID is stored in a database and a new partially received ID is popped from the LIFO stack to continue the identification of the rest of the tags.

The tags' operation in Algorithm 2 line (25) starts receiving the reader's command and acts differently if they have transmitted in the previous slot ($SC = 0$).

(v) If tags remained silent in the previous slot, they update their SC counter adding in case of collision (Algorithm 2 line (32)) or subtracting for idle or success (Algorithm 2 lines (29), (37)).

(vi) If a tag transmitted in the previous slot ($SC = 0$), it performs differently. Upon a collision, the tag chooses a new SC randomly between 0 and 1; in case of success it goes to "sleep" state until the next interrogation cycle; and in case of a go-on command, the tag updates its BC counter with the previous W transmitted (oW), transmits by backscattering the immediately received W bits (nW) from the bit indicated by BC, and attaches the calculated CRC in the response.

An example of an identification of three tags using BwT is shown in Table 1. This example shows how the counters are updated upon the different types of slots.

3.1. Tuning β in BwT. As previously explained, the reception of a CRC consistent tag response does not necessarily mean the identification of a tag. The use of the window can cause the tag response to be only part of the ID; that is, $rBC + W < k$.

There is a need of dynamically recalculating W in order to decrease the number of go-on slots while keeping the number of tag transmitting bits low. The higher the value of W, the

Table 1: Example of an identification cycle of BwT.

Slot	Reader				Tag 1-0001011		Tag 2-1001010		Tag 3-1110001		Reader interpretation
	rBC [tagID]	LIFO	Command	*W*	*SC*	*BC*	*SC*	*BC*	*SC*	*BC*	
(1)	0 [ ]	{ }	ε	1	0	0	0	0	0	0	X
(2)	0 [ ]	{ }	Collision	1	1	0	0	0	0	0	1
(3)	1 [1]	{ }	Go-on	6	1	0	0	1	0	1	X
(4)	1 [1]	{1}	Collision	1	2	0	1	1	0	1	1
(5)	2 [11]	{1}	Go-on	5	2	0	1	1	0	2	10001
(6)	1 [1]	{ }	Success	1	1	0	0	1	Sleep	—	0
(7)	2 [10]	{ }	Go-on	5	1	0	0	2			1010
(8)	0 [ ]	{ }	Success	1	0	0	Sleep	—			0
(9)	1 [0]	{ }	Go-on	6	0	1					1011
(10)	0 [ ]	{ }	Success	1	Sleep	—					—

lower the number of go-on slots and the higher the number of tag transmitting bits. A desirable behavior can be found using exponential equation (3) to search for a balance between these parameters. The parameter β is tuned in order to seek for that balance.

The simulation results of the energy consumed by the RFID reader to identify 1 tag, the throughput of the system, and the number of slots and reader bits needed per tag for different values of β are shown in Figure 4 under a set of 1000 tags (more details about the simulation parameters are also given in Section 4). The results show that for β = [0,13–0,27] the energy consumed by the reader, and the number of slots, and reader bits per tag are at their lowest values and the throughput shows the highest values achieving a compensated balance between go-on slots and tag transmitting bits. Therefore, a value of β = 0,2 is chosen for the comparison with the state-of-the-art protocols.

4. Simulation Results

This section presents the simulation results of the proposed protocol using Matlab R2016b with an evaluation of the outcomes. A comparison between the proposed BwT protocol and the presented protocols in Section 2.2, BT [17], FTTP [7], OBTT [8], and QwT [14], is presented here.

A scenario with one reader and a varying set of tags from 100 to 1000 tags is proposed. These tags are uniformly distributed and k is assumed as 128 bits since it is the most common ID length that is currently used in the standard EPC C1 G2 (96 bits of Electronic Product Code + 16 bits of Protocol Control + 16 bits of CRC) [6]. The tag IDs are uniformly distributed and dynamically generated with varying random seed values for every simulation iteration. The simulated responses have been averaged over 100 iterations for accuracy in the results. Table 2 shows the parameters used in the simulations. Tari, the time interval for a data 0 transmission, is set to the standard's minimum of 6.25 μs for the highest data rate (same for reader and tag), conditioning, RTCal, TRCal, T1, T2, and T3 in accordance with the EPC standard [6]. P_{tx} and P_{rx} were obtained from [15].

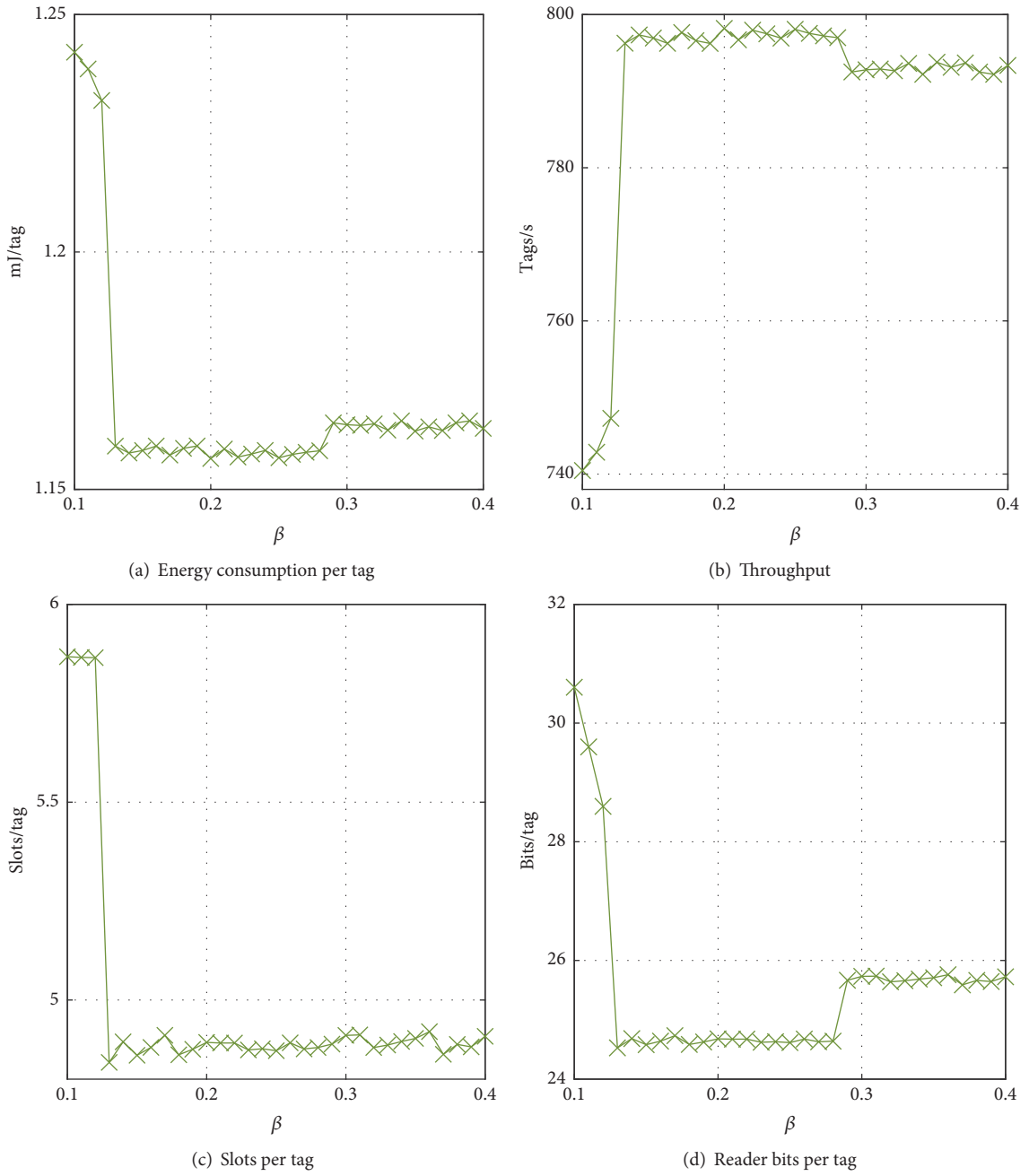
Presented in Figure 1 is the link timing of the four typical types of slots to perform identification time calculations and (1) for energy calculations. The duration of each slot can be different, and bits 0 and 1 have been considered as 1 Tari for easiness in calculations. This, in fact, has been applied to all the protocols, ensuring fairness in the comparison.

Table 2: Parameters used in simulations.

Parameter	Value
Tari	6.25 μs
data rate	160 kbps
RTCal	18.75 μs
TRCal	24.38 μs
$T1$	18.86 μs
$T2$	8.13 μs
$T3$	37.5 μs
P_{tx}	825 mW
P_{rx}	125 mW

The length of the reader commands is set to 3 bits for all the compared protocols, enough to encode all the needed commands. BwT, in addition, attaches W represented with $\log_2 W + 1$ bits; and QwT uses the length of the query on every slot plus the corresponding W bits. Tag responses are k bits long, except for BwT and QwT which use W bits and a CRC of 5 bits and OBTT which transmits the following ID bits to the received query after the estimation phase.

Figure 5 shows the throughput and the total number of bits used by the RFID system in the comparison of the simulated protocols. The calculation of the throughput is based on the total number of bits transmitted by the reader and the tags. BwT shows the highest throughput, slightly over OBTT. The use of the window increases the number of slots needed to identify the set of tags due to the generation of go-on slots. OBTT splits the initial set of tags in smaller subsets and therefore decreases the number of collisions and total slots. BwT, however, reduces the length of tag responses causing a BwT collision to spend less time than that of an OBTT. This is directly reflected on the throughput, meaning that the time BwT spends with go-on slots plus the time saved in collisions with low W values is less than the time OBTT spends on collisions.



(a) Energy consumption per tag

(b) Throughput

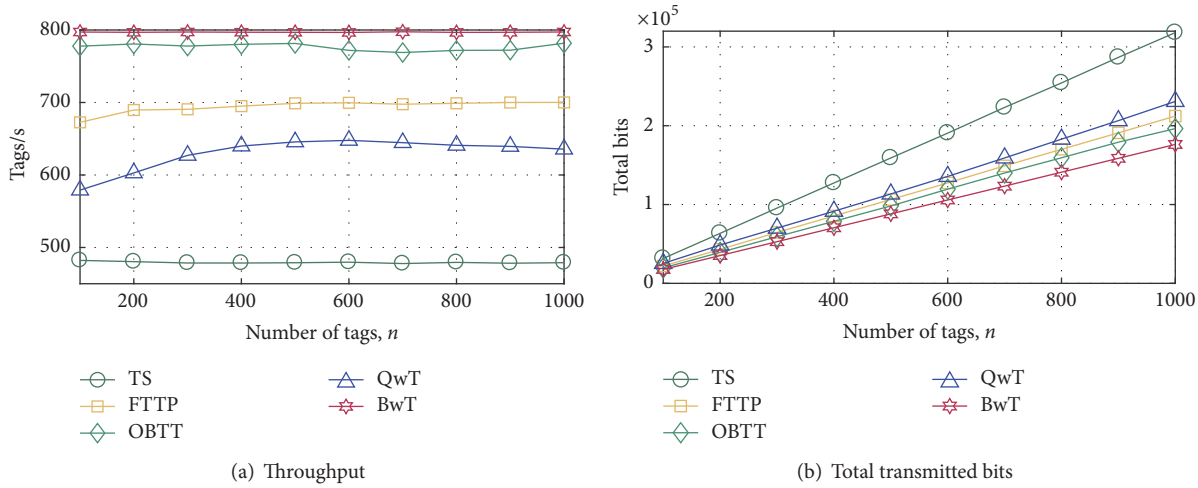
(c) Slots per tag

(d) Reader bits per tag

FIGURE 4: Selected $\beta = 0{,}2$ to obtain a high throughput, and a reduced energy consumption, low number of slots, and reader bits per tag in an interrogation cycle.

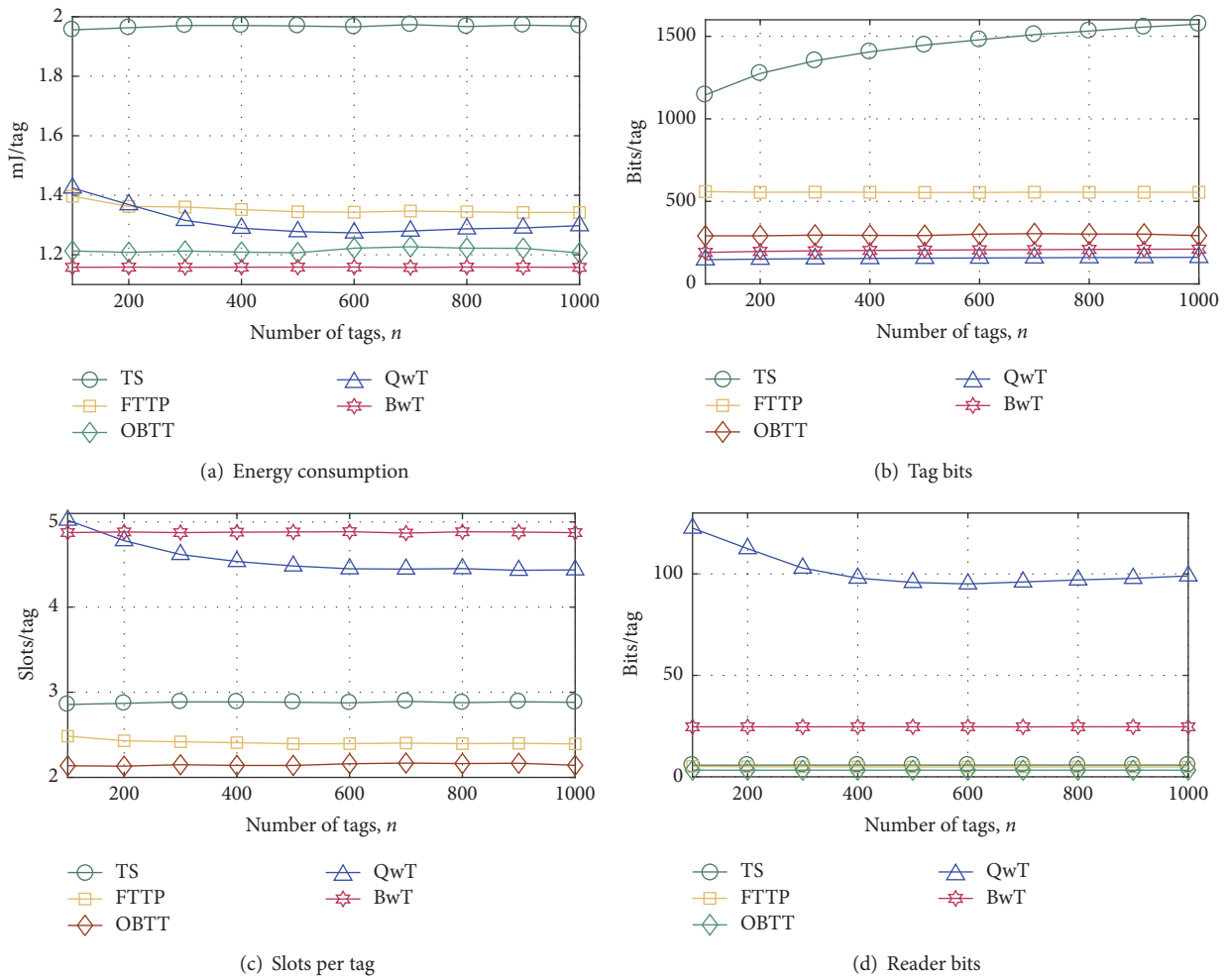
The window, therefore, contributes to reducing the number of tag transmitting bits, which is shown on Figure 5(b) showing BwT as the least bit consuming protocol. Although QwT also uses the window, the excessive number of bits demanded by the reader transmitting queries causes a higher increase in the total number of bits than that of the BwT, decreasing also the throughput of the system. OBTT presents good results also in both metrics thanks to its estimation phase at the beginning of the identification and the use of Manchester coding in the interrogation of the tags. This codification helps the reader to track collisions bit by bit, which in the end affects the total number of bits transmitted by reducing them. FTTP also presents an estimation phase in the beginning; however, it does not use Manchester coding and cannot reach the throughput of OBTT.

Simulations results per tag are shown in Figure 6. The energy consumed by the reader to identify 1 tag is shown in Figure 6(a). The energy consumed has been calculated using (1). In this comparison, BwT outperforms the other compared protocols for all the different sets of tags. The



(a) Throughput (b) Total transmitted bits

FIGURE 5: Simulation results of the throughput (a) and the total number of bits transmitted in the identification of several sets of tags (b).



(a) Energy consumption (b) Tag bits

(c) Slots per tag (d) Reader bits

FIGURE 6: Simulation results of the energy consumption (a), the tag transmitting bits (b), the number of slots (c), and the reader transmitting bits (d) needed to identify 1 tag.

energy consumed to identify 1 tag is not affected by the total number of tags existing in the range of the antenna in BwT. OBTT and FTTP, however, show slight increases in dense and low populated tag environments, respectively. QwT shows a high increase when there are a few tags on the range of the reader's antenna and BT shows the worst results of the comparison.

As the bit window technique modifies the number of tag transmitting bits per slot, Figure 6(b) shows the improvement caused by the window in QwT and BwT as the least tag bit transmitting protocols, quite the opposite of BT. Although BwT and QwT tags need to use CRCs, they transmit the lowest number of bits.

The decrease in tag transmitting bits shown by the windowed protocols BwT and QwT is achieved at the cost of an increase in go-on slots. Figure 6(c) presents both protocols as the most slot consuming protocols to identify a tag. FTTP and OBTT, using the estimation phase, provide the best performance in terms of slots, where the bit tracking protocol OBTT stands out with only 2 slots to identify a tag. Notice also that despite the need of both windowed protocols of the highest number of slots, BwT saves more than 50% of the reader bits transmitted compared with that of QwT. QwT tags demand that the reader transmit long queries, which increases the number of reader bits.

Summing up, BwT reduces the number of tag transmitting bits thanks to the use of the window and avoids transmitting queries. This fact results in a deep reduction of the total transmitted bits reducing the energy consumed by the reader and increasing the throughput of the RFID system. These results show evidence of BwT being a good candidate which seeks for a high throughput under low energy consumption.

5. Conclusions

An anticollision protocol, called BwT, has been presented in this paper. BwT applies the window procedure to the BT protocol including an additional counter in the tags in order to manage the number of ID bits they transmit and the bits they have already transmitted. BwT has been compared to several state-of-the-art anticollision protocols outperforming them in terms of throughput and decreasing the energy consumed by the reader to identify 1 tag. Therefore, simulations showed that BwT can be considered as a good RFID anticollision candidate in passive RFID systems.

Definition of Symbols and Variables

t_R: Time needed to transmit a reader command
t_T: Time needed to transmit a tag response
T_1: Time the tags need to generate a response
T_2: Time the reader needs to receive a response
T_3: Max. time a reader waits before considering the slot idle
P_{tx}: Reader transmission power
P_{rx}: Reader reception power
E: Energy consumed by reader
k: Length of a tag ID
L: Length of a query
W: Size of the window.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] K. Finkenzeller, "RFID Handbook: Fundamentals and Applications in Contactless Smart Cards, Radio Frequency Identification and near-Field Communication," *RFID Handbook: Fundamentals and Applications in Contactless Smart Cards, Radio Frequency Identification and near-Field Communication*, 2010.

[2] R. Want, B. N. Schilit, and S. Jenson, "Enabling the internet of things," *The Computer Journal*, vol. 48, no. 1, pp. 28–35, 2015.

[3] D. K. Klair, K.-W. Chin, and R. Raad, "A survey and tutorial of RFID anti-collision protocols," *IEEE Communications Surveys & Tutorials*, vol. 12, no. 3, pp. 400–421, 2010.

[4] H. Gao, Z. Yang, J. Bhimani et al., "AutoPath: Harnessing Parallel Execution Paths for Efficient Resource Allocation in Multi-Stage Big Data Frameworks," in *Proceedings of the 2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–9, July 2017.

[5] Z. Yang, J. Wang, D. Evans, and N. Mi, "AutoReplica: Automatic data replica manager in distributed caching and data processing systems," in *Proceedings of the 35th IEEE International Performance Computing and Communications Conference, IPCCC 2016*, December 2016.

[6] GS1, 2015, EPC™ Radio-Frequency Identity Protocols Generation-2 UHF RFID - Specification for RFID Air Interface - Protocol for Communications at 860 MHz–960 MHz.

[7] J. Choi, I. Lee, D.-Z. Du, and W. Lee, "FTTP: A fast tree traversal protocol for efficient tag identification in RFID networks," *IEEE Communications Letters*, vol. 14, no. 8, pp. 713–715, 2010.

[8] Y. C. Lai, L. Y. Hsiao, and B. S. Lin, "Optimal slot assignment for binary tracking tree protocol in RFID tag identification," *IEEE/ACM Transactions on Networking*, vol. 23, no. 1, pp. 255–268, 2015.

[9] V. Namboodiri and L. Gao, "Energy-aware tag anticollision protocols for RFID systems," *IEEE Transactions on Mobile Computing*, vol. 9, no. 1, pp. 44–59, 2010.

[10] X. Yan and X. Liu, "Evaluating the energy consumption of the RFID tag collision resolution protocols," *Telecommunication Systems*, vol. 52, no. 4, pp. 2561–2568, 2013.

[11] D. K. Klair, K.-W. Chin, and R. Raad, "An investigation into the energy efficiency of pure and Slotted Aloha based RFID anticollision protocols," in *Proceedings of the 2007 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, WOWMOM*, 4, 1 pages, June 2007.

[12] Z. Yang, M. Awasthi, M. Ghosh, and N. Mi, "A fresh perspective on total cost of ownership models for flash storage in datacenters," in *Proceedings of the 8th IEEE International Conference on Cloud Computing Technology and Science, CloudCom 2016*, pp. 245–252, December 2016.

[13] F. Hessar and S. Roy, "Energy based performance evaluation of passive EPC Gen 2 class 1 RFID systems," *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1337–1348, 2013.

[14] H. Landaluce, A. Perallos, E. Onieva, L. Arjona, and L. Bengtsson, "An Energy and Identification Time Decreasing Procedure for Memoryless RFID Tag Anticollision Protocols," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 4234–4247, 2016.

[15] C. Law, K. Lee, and K.-Y. Siu, "Efficient memoryless protocol for tag identification," in *Proceedings of the 4th International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications*, pp. 75–84, Boston, Mass, USA, August 2000.

[16] H. Vogt, "Efficient object identification with passive RFID tags," in *Pervasive Computing*, vol. 2414 of *Lecture Notes in Computer Science*, pp. 98–113, Springer, Berlin, Germany, 2002.

[17] D. R. Hush and C. Wood, "Analysis of tree algorithms for RFID arbitration," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT '98)*, pp. 107–114, IEEE, Cambridge, Mass, USA, August 1998.

[18] D. Bertsekas and R. Gallager, *Data Networks*, Pearson, 2nd edition.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 6590713, 12 pages
https://doi.org/10.1155/2017/6590713

Research Article

A Dual Key-Based Activation Scheme for Secure LoRaWAN

Jaehyu Kim and JooSeok Song

Yonsei University, 3rd Engineering Building C505, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea

Correspondence should be addressed to Jaehyu Kim; jaehyu_kim@yonsei.ac.kr

Received 28 April 2017; Accepted 12 October 2017; Published 6 November 2017

Academic Editor: Haiyu Huang

With the advent of the Internet of Things (IoT) era, we are experiencing rapid technological progress. Billions of devices are connected to each other, and our homes, cities, hospitals, and schools are getting smarter and smarter. However, to realize the IoT, several challenging issues such as connecting resource-constrained devices to the Internet must be resolved. Recently introduced Low Power Wide Area Network (LPWAN) technologies have been devised to resolve this issue. Among many LPWAN candidates, the Long Range (LoRa) is one of the most promising technologies. The Long Range Wide Area Network (LoRaWAN) is a communication protocol for LoRa that provides basic security mechanisms. However, some security loopholes exist in LoRaWAN's key update and session key generation. In this paper, we propose a dual key-based activation scheme for LoRaWAN. It resolves the problem of key updates not being fully supported. In addition, our scheme facilitates each layer in generating its own session key directly, which ensures the independence of all layers. Real-world experimental results compared with the original scheme show that the proposed scheme is totally feasible in terms of delay and battery consumption.

1. Introduction

Today, we are living in the Internet of Things (IoT) era where billions of IoT devices are deployed all over the world. According to a report from Ericsson [1], the number of connected IoT devices will reach 28 billion by 2021. These devices produce a massive amount of data and transfer the information to cloud servers, which can be accessed anytime and anywhere. Revolutionary changes are brought into our lives.

Many approaches have been taken to realize various types of communication used in the IoT environment. In the last few years, short-range communication technologies, such as Bluetooth, ZigBee, and Z-Wave, have been popular for utilizing resource-constrained IoT devices because of their low energy consumption [2]. However, their short communication range makes them difficult to use for important IoT applications that require a wide communication range, such as smart city [3]. Although cellular networks provide a wide coverage area, they are also not fully suitable for the IoT environment because of its complexity and cost [4]. To complement the shortcomings of these conventional approaches, Low Power Wide Area Networks (LPWAN) technologies have recently been developed. They are devised to enable long range communication with low battery consumption. With these technologies, even resource-constrained small sensors or actuators can send messages up to tens of kilometers and survive for several years even without a power source [5].

Among the recently proposed LPWAN technologies, such as SigFox, LoRa, Weightless, Ingenu, and Telensa, LoRa is one of the most competitive technologies because of its low power consumption and low cost design [6]. LoRa is a physical layer protocol that enables low power and long-distance communication up to 15 km using chirp spreading spectrum modulation [4]. LoRaWAN is an upper layer protocol based on LoRa that defines the structure and operation of the entire system [7]. LoRaWAN's asynchronous communication scheme enables much longer battery lifetime by reducing the overhead caused by synchronization [5].

While many LPWAN technologies are primarily focused on issues such as battery consumption and communication range, security is also an important issue. In the IoT environment, the importance of security becomes much greater than ever before. The IoT can be a big threat to privacy because it is

closely related to a user's real life. Moreover, the damage from security incidents can be unprecedentedly enormous due to the large scale and connectivity of the IoT environment. To prepare for this situation, previous studies on IoT security [8–10] have addressed some important factors, one of which is key management. According to the research, cryptographic keys can be leaked through various attacks, considering that IoT sensing devices are usually deployed where the attacker can access them. This can be applied to LoRaWAN as well. LoRaWAN specifications [7] emphasize that the key must be uniquely managed to minimize the damage caused by key leakage. This means that when the key is extracted from an end node, it should not affect the other nodes. However, it is not enough and problems still occur with key updates. Although LoRaWAN uses cryptographic keys for several security mechanisms, such as authentication, encryption, and integrity checking, the current LoRaWAN specifications provide update of these keys partially. In some cases, an end node may have to keep using certain keys without changing them during its lifetime. Thus, at some point in the future, if the key is leaked, all the data that the end node has transferred may be passed on to the attacker. To prepare for the attack, keys must be updated periodically, as pointed out in many previous studies [11–14]. How the session key is created in the current LoRaWAN is also problematic. As depicted in Figure 1, each session key is used in a different layer. Thus, the current way in which both session keys are only created by a network server can violate the independence between layers. According to [15], this system could lead to a conflict of interest between the network server and the application server.

In this paper, we propose a dual key-based activation scheme with a new key called a network key (NwkKey). Our scheme resolves the problem that key updates are not available in some cases. We also redefine the operation of each server in the key generation process. In our scheme, a network server and an application server generate a network session key (Nwk_SKey) and an application session key (App_SKey), respectively, so that each layer works completely independently. Moreover, our scheme does not require any additional entities such as a trusted third party. Finally, we demonstrate the feasibility of our scheme through a real-world test. To the best of our knowledge, this is the first attempt to improve the security of LoRaWAN activation.

The rest of this paper is organized as follows. Section 2 provides related works. Section 3 is about LoRaWAN architecture. In Section 4, we provide basic information about LoRaWAN end node activation. In Section 5, a detailed explanation of our proposed scheme is provided. Section 6 is a security analysis of our scheme. In Section 7, we evaluate the performance of our scheme. Section 8 provides our conclusion about this research.

2. Related Works

A security report [16] written by Miller of MWR Infosecurity provides LoRaWAN's possible vulnerabilities and countermeasures as well as basic description of LoRaWAN security. According to the report, all LoRaWAN entities should be prepared for vulnerabilities that can occur during key management, communications, and Internet connection. Especially in case of an end node, the report emphasizes that even if cryptographic keys are leaked through side channel attacks, this should not affect other parts of the system.

In [15], Girard of Gemalto pointed out a problem with LoRaWAN's key provisioning method. In the current LoRaWAN, the network server generates both session keys. This means that the network server generates even the application session key to be used by the application server. According to [15], this could lead to a conflict of interest between the network server and the application server. As a solution to this problem, the author proposes a new LoRaWAN network structure with the trusted third party.

In [17], Zulian analyzed the DevNonce of LoRaWAN. The DevNonce is a random number generated by the end node. It is used for replay attack prevention as well as session key generation. Replay attack prevention works in such a way that the network server determines an invalid message by checking whether previously used DevNonce is contained or not. The author mathematically analyzed the method and determined that the end node can be unavailable with a certain probability under the current DevNonce system. To alleviate this problem, the author proposed increasing the size of the DevNonce field to 24 or 32 bits.

Naoui et al. proposed a new security architecture for LoRaWAN [18]. Their scheme uses the concept of a proxy node, which performs several other functions, including the basic function of the conventional LoRaWAN gateway. In particular, proxy nodes evaluate each other's trustworthiness to create a table and forward it to the end node. The end node can then communicate through the proxy node that has the highest trust value.

The current LoRaWAN has problems with key update and session key generation. In the case of the key update, it has not been addressed in any LoRaWAN security study, despite its seriousness. The solution proposed by [15] to solve the problem of session key generation also has some disadvantages. Because of the newly added trusted third party, the whole join procedure becomes more complex and communication overhead is increased. It is also difficult to be applied to the existing LoRaWAN network already deployed without the trusted third party. In this paper, we propose a dual key-based activation scheme that fully supports the key update and resolves the problem of session key generation without any additional entities.

3. LoRaWAN Architecture

In this section, we briefly describe the architecture of the LoRaWAN network and its entities. We also provide a description of the protocol architecture and message format that are used in the LoRaWAN network environment.

3.1. LoRaWAN Network Architecture. As shown in Figure 1, the LoRaWAN network uses a star topology in which an end node can send messages to multiple gateways that communicate with the network server. Since an end node does not belong to a specific gateway, more than one gateway

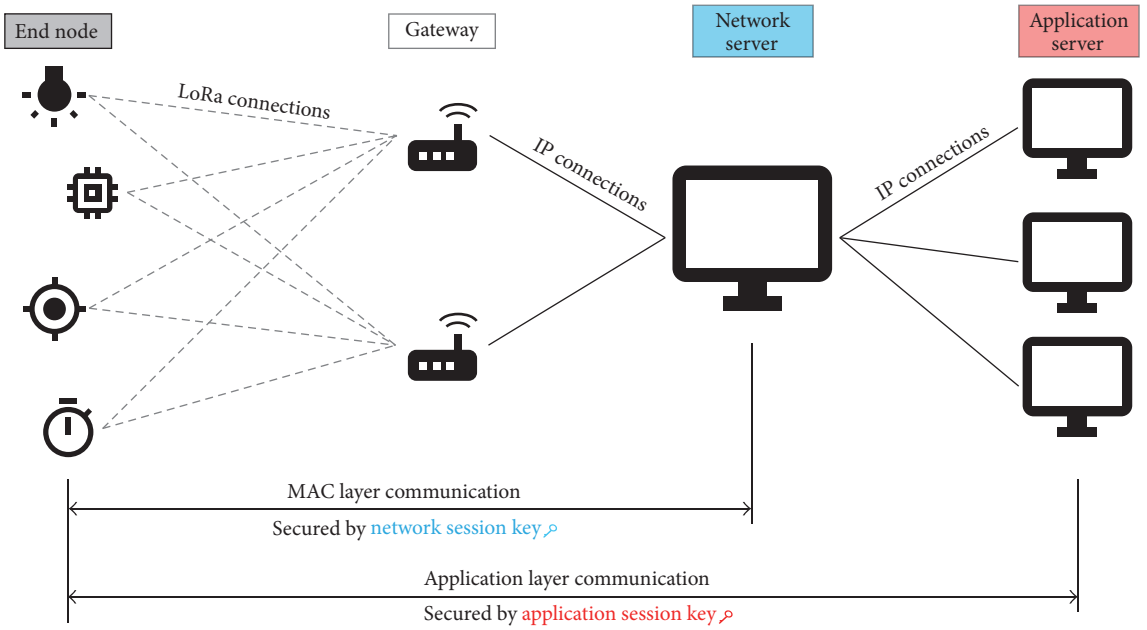


Figure 1: LoRaWAN network structure.

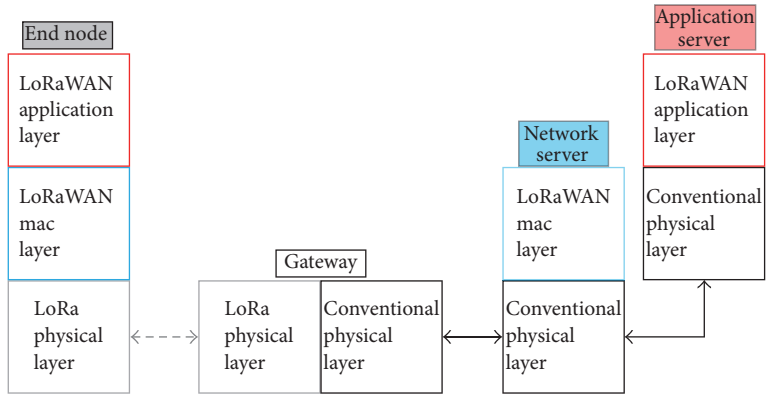


Figure 2: LoRaWAN protocol architecture.

can receive a message sent by an end node [19]. LoRa radio technology is used in communications between an end node and the gateways. The gateways and network server are connected via standard IP connections. The following is a brief description of the entities defined in the LoRaWAN specifications [7].

(i) End node: a LoRaWAN end node is typically used to send small amounts of data at low frequencies over long distances. It can be utilized in various fields such as smart city, smart building, factory automation, farm automation, and logistics.

(ii) Gateway: a LoRaWAN gateway receives packets from the end node via a LoRa radio link. It then forwards them to the network server through the IP connection.

(iii) Network server: the LoRaWAN network server manages the entire network. When it receives packets, it removes the redundancy of packets and performs a security check and then determines the most suitable gateway to send back an acknowledgement message.

3.2. LoRaWAN Protocol Architecture. Figure 2 shows the protocol architecture of LoRaWAN. As shown in this figure, LoRaWAN's protocol consists of a MAC layer and an application layer, and it operates based on the LoRa physical layer. The packet format is displayed in Figure 3. The maximum payload lengths M and N vary with the data rate. It is specified in [20].

(i) MAC layer: the packet processed in the MAC layer consists of a MAC Header (MHDR), a MAC Payload, and a Message Integrity Code (MIC). In a join procedure for end node activation, the MAC Payload can be replaced by join request or join accept messages. The entire MAC Header and MAC Payload portion is used to compute the MIC value with a network session key

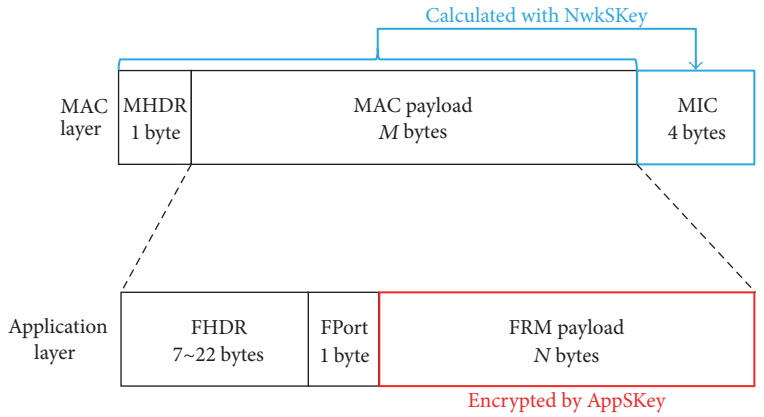


Figure 3: LoRaWAN packet format.

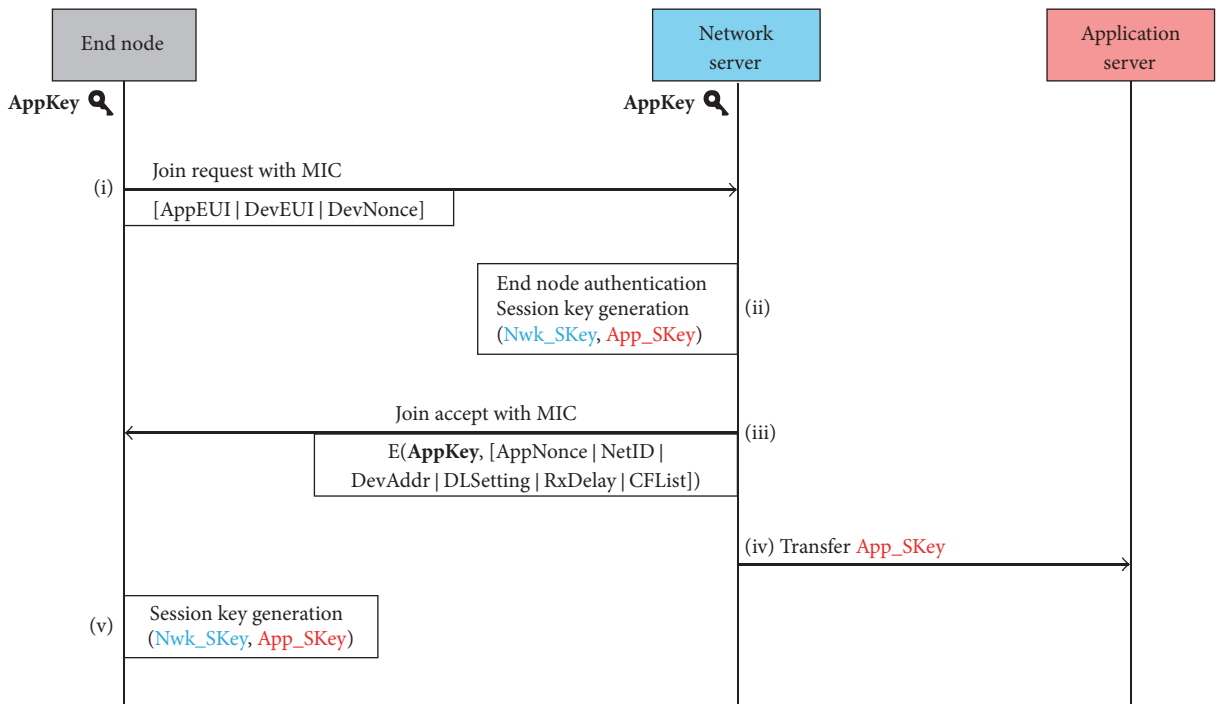


Figure 4: LoRaWAN join procedure.

(Nwk_SKey). The MIC value is used to prevent the forgery of messages and authenticate the end node.

(ii) Application layer: the MAC Payload handled by the application layer consists of a FRM Header (FHDR), an FPort, and a FRM Payload. The FPort value is determined depending on the application type. The FRM Payload value is encrypted with an application session key (App_SKey). This encryption is based on the AES_128 algorithm.

4. LoRaWAN End Node Activation

When a new end node is added to a LoRa network, it should go through an activation process. Through the activation process, both session keys are shared between the end node and the network server. Currently, LoRaWAN provides two types of activation methods. One is over-the-air activation (OTAA) and the other is activation by personalization (ABP).

4.1. Over-the-Air Activation. In the OTAA mode, an end node communicates with the network server to perform the activation process, which is called join procedure. According to the LoRaWAN specifications [7], the OTAA mode is used when an end node is deployed or reset. Figure 4 shows the LoRaWAN join procedure. A detailed explanation of each step is as follows.

(i) Join request message: by sending a join request message, the end node starts the join procedure.

DevEUI, AppEUI, and DevNonce are included in the join request. DevEUI and AppEUI refer to the global end node and application identifier, respectively. They follow the IEEE EUI-64 address space format. The DevNonce is a random number generated by the end node. The MIC value of join request is calculated by the following formula:

$$\begin{aligned} cmac &= aes128\_cmac\left(AppKey, MHDR \mid AppEUI \mid DevEUI \mid DevNonce\right) \\ \mathbf{MIC} &= cmac\left[0 \cdots 3\right]. \end{aligned} \tag{1}$$

An application key (AppKey) is preshared between the end node and the network server.

(ii) After the network server receives the join request, it performs the replay attack prevention process, which is based on the DevNonce. If the DevNonce in the join request is previously used, the network server determines that the message is invalid and that the join process will fail. If the message is valid, the network server authenticates the end node with the MIC value. If the end node passes the authentication, the network server generates an Nwk_SKey and an App_SKey by the following formula:

$$\begin{aligned} Nwk\_SKey &= aes128\_encrypt\left(AppKey, \texttt{0x01} \mid AppNonce \mid NetID \mid DevNonce \mid pad_{16}\right) \\ App\_SKey &= aes128\_encrypt\left(AppKey, \texttt{0x02} \mid AppNonce \mid NetID \mid DevNonce \mid pad_{16}\right). \end{aligned} \tag{2}$$

AppNonce is a random number generated by the network server. NetID is a 24-bit field. Its 5 LSBs are called NwkID which is used to separate addresses of geographically duplicated LoRa networks. The other bits of NetID can be freely determined by the network server.

(iii) Join accept message: a join accept message contains AppNonce, NetID, DevAddr, DLSettings, RxDelay, and CFList. The DevAddr is a 32-bit identifier of the end node within the current network. The 7 MSBs of DevAddr are referred to as the NwkID, which is also contained in NetID. The other bits can be arbitrarily chosen by the network server. DLSettings contains several values related to the downlink configuration. RxDelay is a delay between the transmission and reception process. CFList is an optional field that is about channel frequencies. Finally, the whole join accept message is encrypted with the AppKey.

(iv) Transfer App_SKey: since the App_SKey is devised to secure end-to-end communications between the end node and the application server, it should be transferred from the network server to the application server. The LoRaWAN specification does not specify when and how to exchange App_SKey with the application server. We thought it is an essential part and so included it in the join procedure.

(v) After receiving the join accept message, the end node decrypts it and generates session keys using extracted parameters.

4.2. Activation by Personalization. ABP is the way in which an end node can belong to a particular LoRa network without performing a join procedure under certain circumstances. In the ABP mode, the end node does not have DevEUI, AppEUI, and AppKey, which are essential for join procedure. Instead, both session keys required for LoRaWAN communications and DevAddr are preloaded on the end node.

4.3. Problem Statement

(1) OTAA key update: in the OTAA mode, authentication and session key agreement is performed using the AppKey preshared between the end node and the network server. In this process, one of the most critical problems is that updating the AppKey is not supported by the LoRaWAN specifications. Under the current standard, session keys can be updated several times, but the AppKey that is used to generate them cannot be updated. In other words, the end node has to use only one AppKey for a lifetime. As pointed out in several previous IoT security studies [8–10], we have to prepare for key leakage, which can have various causes, such as node capture attacks and side channel attacks.

In this respect, LoRaWAN AppKey, which cannot be updated, can cause serious security problems. If the AppKey is leaked by an attacker, the attacker can get the contents of all join accept messages that have been sent up to that point. As shown in (2), AppNonce, NetID, and DevNonce are used to generate session keys. Among them, the AppNonce and NetID are contained in the join accept message. The DevNonce can easily be obtained in the join request transmitted without encryption. By using these parameters, the attacker can restore all the session keys used in the past. Thus, the attacker can steal all the data that the target node had previously transmitted. From this perspective, many previous studies on key management [11–13] and NIST [14] have emphasized that the key must be updated periodically.

(2) ABP key update: in the ABP mode, the AppKey is not preloaded on the end node. Since the AppKey is essential to the join procedure, the end node cannot perform it, which means that there is no way of updating session keys. Therefore, in the ABP mode, the end node must use the same session key throughout its lifetime. This can also pose a similar security threat to the end node, as discussed in (1). If the attacker successfully steals these keys, he or she can get all the data sent from the target node that were

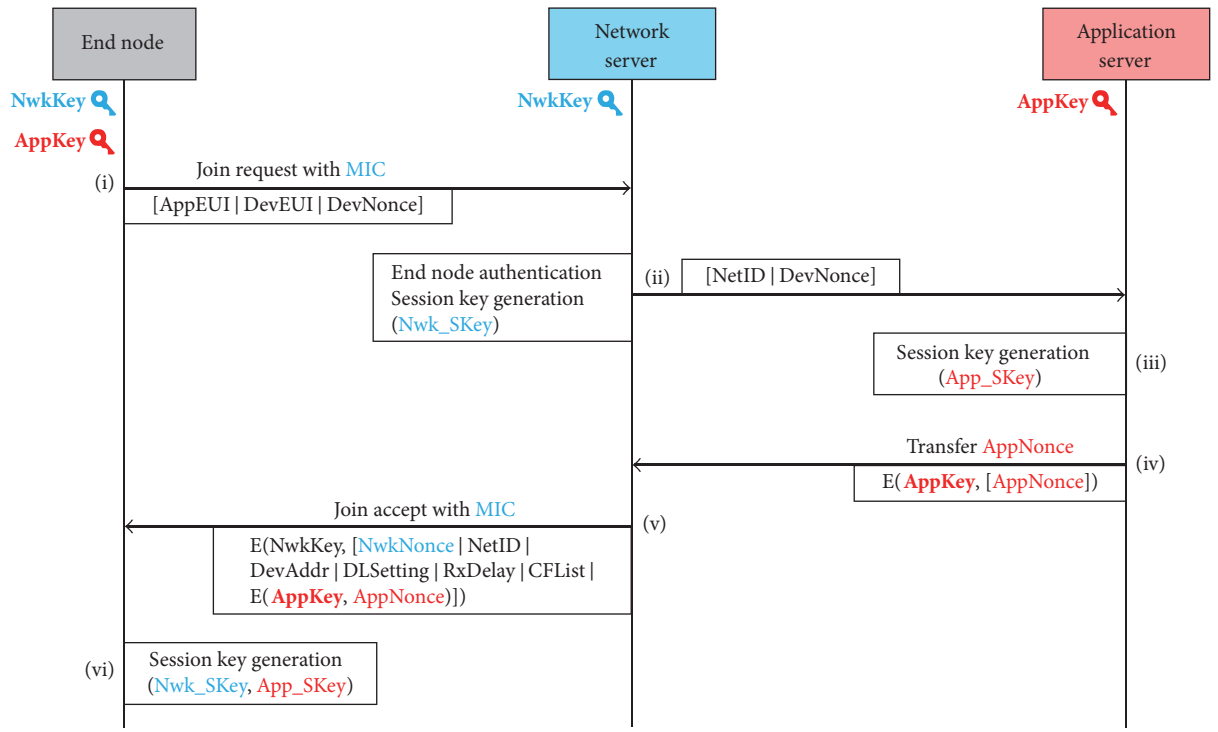


Figure 5: Dual key join procedure.

protected by the session keys. This is why supporting the key update in the ABP mode is an important issue.

(3) Session key generation: under the current LoRaWAN session key generation system, the network server generates all the session keys alone. Since the Nwk_SKey and the App_SKey are used in different layers, this system does not guarantee independence between layers. According to [15], there is even the risk that the network server with the App_SKey can intercept the application layer data. Thus, we need to construct a new system in which each layer independently generates its own session key.

5. Dual Key-Based End Node Activation

5.1. Dual Key-Based Over-the-Air Activation. Compared to the original join procedure, the notable feature of our scheme is the existence of the NwkKey. The NwkKey has the same properties as the AppKey. They are of the same length and should not be deduced from the public information of the end node. They should not also be shared with other end nodes. In our scheme, the NwkKey and the AppKey are preloaded together on the end node. The NwkKey is shared with the network server, and the AppKey is with the application server. During the proposed join procedure, Nwk_SKey and App_SKey are generated from the NwkKey and the AppKey, respectively.

Our scheme works in two modes, initial and noninitial.

(1) Initial Join Procedure. The initial mode is applied when an end node performs the join procedure for the first time. Figure 5 represents the proposed initial join procedure, and the details are as follows.

(i) Join request: the join request message is created in the same manner as the original scheme. The message includes AppEUI, DevEUI, and DevNonce. The MIC is also calculated in the same way, except that the NwkKey is used instead of the AppKey. In our scheme, preshared key between the end node and the network server is not the AppKey but the NwkKey.

$$cmac = aes128\_cmac\left(\mathbf{NwkKey}, MHDR \mid AppEUI \mid DevEUI \mid DevNonce\right) \quad (3)$$

$$\mathbf{MIC} = cmac\left[0 \cdots 3\right].$$

(ii) On receipt of the join request, the network server authenticates the end node by recalculating the MIC value. Since the end node and the network server share the NwkKey, the end node can be authenticated. If the message is valid, the network server generates the Nwk_SKey with the NwkKey and transfers NetID and DevNonce to the application server.

$$Nwk\_SKey = aes128\_encrypt\left(\mathbf{NwkKey}, \texttt{0x01} \mid \mathbf{NwkNonce} \mid NetID \mid DevNonce \mid pad_{16}\right). \quad (4)$$

Compared to the original join procedure, the NwkNonce and the NwkKey are used instead of the AppKey and the AppNonce. The NwkNonce

is a random number that has essentially the same properties as the AppNonce.

(iii) Application server generates an App_SKey after receiving the NetID and DevNonce from the network server. The generation method is as follows:

$$AppSKey = aes128\_encrypt\left(\mathbf{AppKey}, \texttt{0x01} \mid \mathbf{AppNonce} \mid NetID \mid DevNonce \mid pad_{16}\right). \quad (5)$$

(iv) When the App_SKey generation is completed, the application server sends the AppNonce to the network server. At this time, the AppNonce is encrypted with the AppKey. Since the network server does not have the AppKey, it cannot decrypt the ciphertext.

(v) Join accept message: after receiving the encrypted AppNonce from the application server, the network server sends a join accept message. In the proposed scheme, the NwkNonce and the encrypted AppNonce are included. The rest is the same as the original join accept message. The entire message is encrypted with the NwkKey before transmission.

(vi) The end node decrypts the join accept message. After that, it generates the Nwk_SKey and the App_SKey with the extracted parameters.

When communication is initiated with the newly created session keys, the end node and both servers immediately discard the NwkKey and the AppKey. This is to prevent the key from being leaked by the attacker in the future.

(2) Noninitial Join Procedure. If the end node already joined to the network through the initial join procedure needs to perform join procedure again, the noninitial join procedure is performed. The noninitial mode is almost the same as the initial mode, except that the session keys created in the previous join procedure are used instead of the NwkKey and the AppKey. In other words, the noninitial join procedure is the process of creating new session keys from the old session keys. The joined end node no longer has the NwkKey and the AppKey. Therefore, subsequent join procedures are performed in the noninitial mode. Through this process, the end node can update the keys used for LoRaWAN's security mechanisms.

5.2. Dual Key-Based Activation by Personalization. In the current LoRaWAN's ABP mode, both session keys and DevAddr are directly mounted on the end node. Since there are no DevEUI, AppEUI, and AppKey, the join procedure cannot be performed. This means that the session keys mounted on the end node cannot be automatically updated. For the absence of an AppKey, our proposed noninitial join procedure that utilizes both session keys can be a solution. However, a problem still remains, in that the join request cannot be made due to the absence of DevEUI and AppEUI. Therefore, we propose a new join request for ABP mode as follows:

$$JoinReqestforABP = [DevAddr \mid DevNonce]. \quad (6)$$

It uses DevAddr as an identifier. All the other steps, such as session key generation and join accept processing, can be done in the same manner as the noninitial join procedure. As a result, the end node activated via ABP mode can also update its session keys.

6. Security Analysis

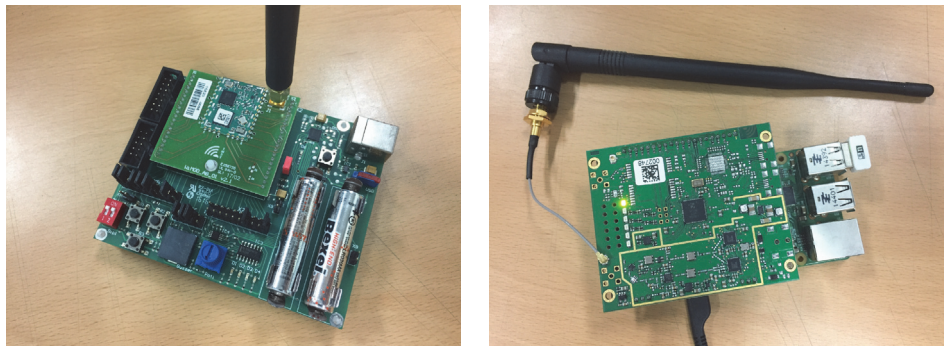
6.1. Basic Security Mechanisms. Our scheme satisfies the same security requirements as LoRaWAN, such as authentication, message integrity, data confidentiality, and replay attack prevention. End node authentication is achieved by using the NwkKey and the MIC value. When a join request arrives, the network server authenticates the end node by recalculating the MIC value with the NwkKey. The MIC is also used for message integrity checking. If the recalculated MIC value is different from the transmitted one, this means that the message is manipulated by unauthorized entities. Application data is encrypted by the App_SKey. In the current LoRaWAN, since the network server generates the App_SKey, application data confidentiality may not be perfectly guaranteed. However, in the proposed scheme, the App_SKey is only shared between the end node and the application server. Thus, application data confidentiality is guaranteed. Replay attack prevention is provided by the DevNonce. When a previously used DevNonce is contained in a join request, the network server considers it invalid.

6.2. Key Update. In the current LoRaWAN activation process, the key update is not fully supported. In the OTAA mode, the end node cannot update the preloaded AppKey and in the ABP mode, preloaded session keys cannot be updated. The end node must use these keys for a lifetime without updating. Thus, if the key is stolen by the attacker, he or she can steal all the data that the target node had previously transmitted.

On the contrary, in our scheme, the end node can update keys in any cases. Regardless of the activation mode in which the end node is activated, the end node can update the key using the proposed initial or noninitial join procedure. Another important aspect of our scheme is that once the key is updated, the previously used key is discarded. In case of the initial join, preloaded NwkKey and AppKey are discarded after the procedure is done. Through the noninitial join, both previously used session keys are discarded. All keys are valid only for the session in our scheme. Thus, the data transmitted in the previous sessions can be protected even if the current session keys are leaked.

6.3. Session Key Generation. LoRaWAN network communication consists of two layers. Each layer has a different session key. Under the current LoRaWAN session key generation system, the network server creates both session keys. Therefore, the security mechanism of each layer is not completely isolated. The network server with the App_SKey can access the application layer data, which should not be permitted.

However, in our scheme, each layer generates its own key. The network server creates the Nwk_SKey, and the application server creates the App_SKey. This means that the network server is no longer involved in creating App_Skey



(a) End node: SK-IM880B demo board (b) Gateway: Raspberry Pi2 with IMST IC880A board

Figure 6: LoRaWAN devices.

and cannot access the application layer data. Thus, compared with the previous method, each layer independently operates a security mechanism.

7. Performance Evaluation

We performed a real-world experiment to demonstrate the feasibility of the proposed scheme. In this section, we provide detailed information on the experimental environment and an analysis of the experimental results.

7.1. Experimentation Environment

(1) Hardware Environment. We have installed a private LoRaWAN network consisting of four entities: the end node, gateway, network server, and application server. Figure 6 shows LoRaWAN devices that we used in this experiment. Hardware specification is shown in Table 1.

(i) End node: we installed the end node by uploading the source code [21] provided by Semtech to the demo board included in SK-IM880B [22].

(ii) Gateway: we made a gateway device by connecting the Raspberry Pi 2 and IMST IC880A [23] board according to the tutorial [24, 25] provided by Semtech and The Things Network. We uploaded the LoRaWAN gateway source code [26, 27] provided by Semtech to complete the gateway installation.

(iii) Network server: there are open source projects for implementing LoRa network server [28, 29]. We established the network server by installing these source codes on Ubuntu OS.

(iv) Application server: there is also an open source project for LoRa application server [30]. It is installed on our application server computer, which runs on Ubuntu OS.

We implemented the proposed scheme by modifying the source code on each entity.

(2) Network Environment. We installed the network environment according to [20]. In this document, parameters related to LoRa transmission, such as default channels, frequency, date rate, and delays, are specified.

Table 1: Hardware specification.

End node	STM32L151CB MCU, 128K Flash, 10K RAM, IM880B-L Module
Gateway	Raspberry Pi 2 with IC880A, Wi-Fi connection
Network server	Intel Core i5-470UM 1.33 GHz CPU, 4 G RAM, Ethernet connection
Application server	Intel Core 2 6600 2.4 GHz CPU, 4 G RAM, Ethernet connection

(i) Band: in South Korea where we have conducted experiments, the LoRa dedicated band is 920–923 MHz [20]. Currently, however, student researchers face difficulty in obtaining experimental equipment for the band. Therefore, we conducted experiments with EU 863–870 MHz band equipment, which can be easily purchased online. Although the band is being used for other purposes in South Korea, we determined that simply verifying the feasibility of our scheme is possible.

(ii) Reception windows: Figure 7 shows when reception windows open in an end node. After completion of the join request transmission, the end node opens two reception windows. A join accept packet can be received only when the reception window is open. The first reception window (RX1) is opened Join Accept Delay1 seconds after the completion of the join request transmission. The Join Accept Delay1 is defined as 5 seconds for EU band. By default, the join accept packet received by RX1 uses the same frequency and same date rate as the join request. The second reception window (RX2) opens after the Join Accept Delay2 and uses the predefined channel. The Join Accept Delay2 is defined as 6 seconds for EU band. The network server decides which reception window to use when creating a join accept message. Since the predefined channel for RX2 is not contained

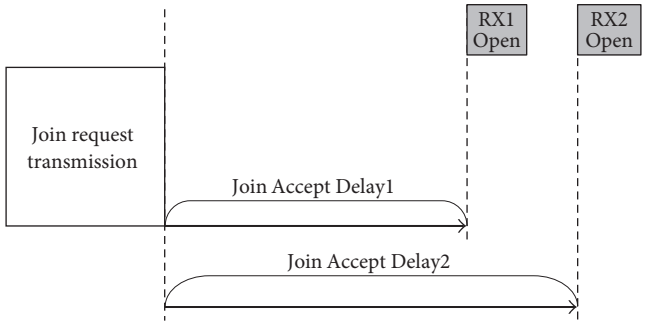


Figure 7: LoRaWAN reception windows.

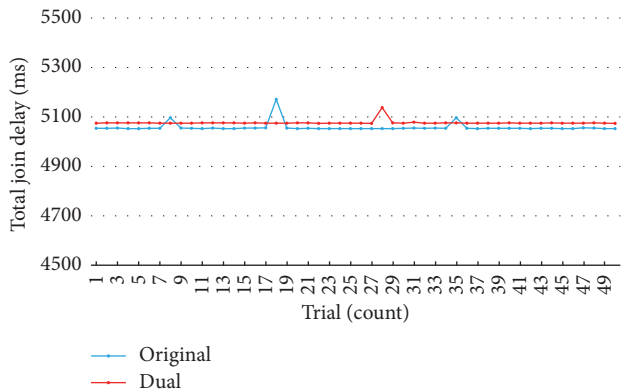


Figure 8: Total join delay.

in the default channel list provided in [20], we use RX1 as a default setting in our experiment.

7.2. Experimentation Results. In this section, we provide the experimentation results. Our proposed scheme has additional work on the end node and server-side compared to the original scheme. This may increase delay and battery consumption. Thus, we compared with the proposed initial join procedure and the original join procedure in terms of delay and battery consumption, which are key elements of feasibility.

(1) Delay. At first, we measured the total join delay from the end of the join request transmission until the end of the join procedure on the end node. Figure 8 shows the experimental result after performing the original join procedure and the proposed initial join procedure 50 times, respectively. Some of the abnormal values seemed to be caused by the temporarily impaired LoRa wireless link. Table 2 shows the average time spent on each scheme. According to the results, the total join delay in the proposed initial join procedure increased by about 18 ms on average. To find out the cause of the increase in delay, we analyzed how the delay occurs at each step of the join procedure. Figure 9 shows the structure of the delay that occurs in the join procedure. Both original

Table 2: Total join delay.

	Total join delay
Original join	5057.92 ms
Proposed initial join	5076.64 ms

and proposed join procedure follow this structure. According to the structure, the total delay can be expressed as follows.

$$\begin{aligned} TotalJoinDelay = {} & ComDelay_{joinreq} \\ & + JoinDelay_{gateway} \\ & + ProcTime_{joinacpt} \\ & + ComDelay_{joinacpt}. \end{aligned} \tag{7}$$

$ComDelay_{joinreq}$ means join request communication delay between the end node and gateway. $JoinDelay_{gateway}$ means gateway join delay. $ProcTime_{joinacpt}$ means the join accept processing time of the end node. $ComDelay_{joinacpt}$ means join accept communication delay between the gateway and end node.

(1) $ComDelay_{joinreq}$: the join request had no difference between the proposed scheme and the original scheme. Therefore, it can be assumed that the same delay occurs.

(2) $JoinDelay_{gateway}$: as depicted in Figure 9, the gateway join delay is the time when the join request packet is received on the gateway until the join accept packet starts sending. This value mainly consists of server-side processing time and gateway waiting time. At the end node, RX1 is opened 5 seconds after the join request is sent, and the join accept message must arrive at this time. As shown in Figure 10 and Table 3, the server-side processing time is less than 1 second. So the gateway must wait until RX1 is open on the end node. If the gateway immediately sends a join accept message without the waiting time, a message cannot be received on the end node where RX1 has not yet been opened. Thus, proper waiting time on the gateway is essential for the successful join procedure. In our experiment, the network server deliberately set the gateway join delay to 5 seconds, which is the same as Join Accept Delay1, to generate a proper waiting time on the gateway. According to this mechanism, the gateway waits until the gateway join delay becomes 5 seconds and then sends the join accept message. As a result, the gateway join delay of both schemes is equally 5 seconds.

(3) $ProcTime_{joinacpt}$: as can be seen in Table 4, the join accept processing time increased by about 0.4 ms on average in the proposed scheme. This is because it performed AES encryption once more. However, since the increased value was very small, the effect on the overall delay was negligible.

(4) $ComDelay_{joinacpt}$: according to our analysis, the other factors constituting the total join delay had little effect

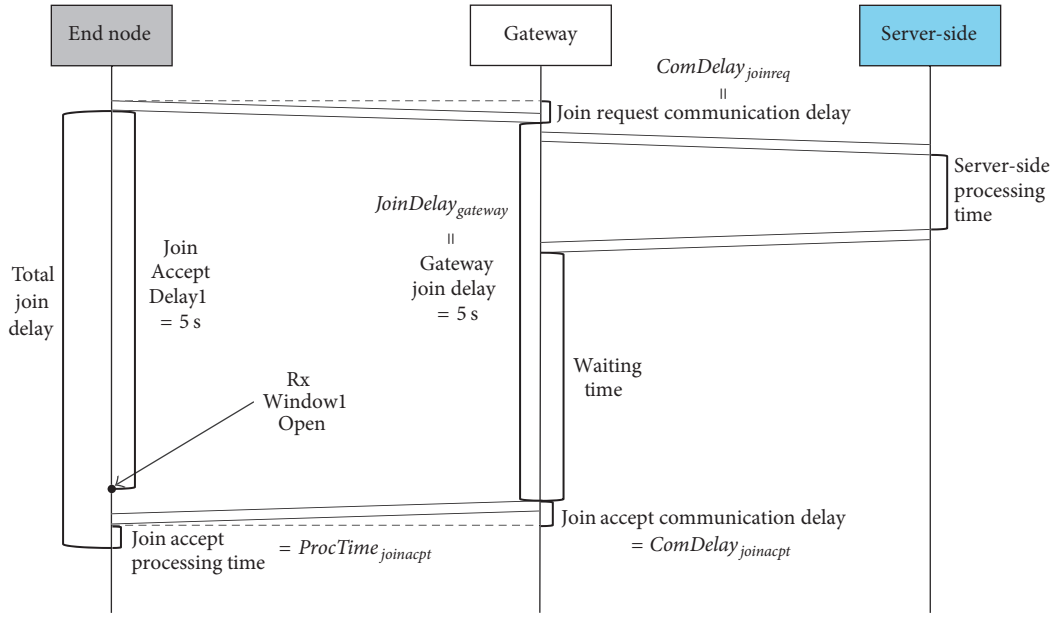


Figure 9: Delay structure.

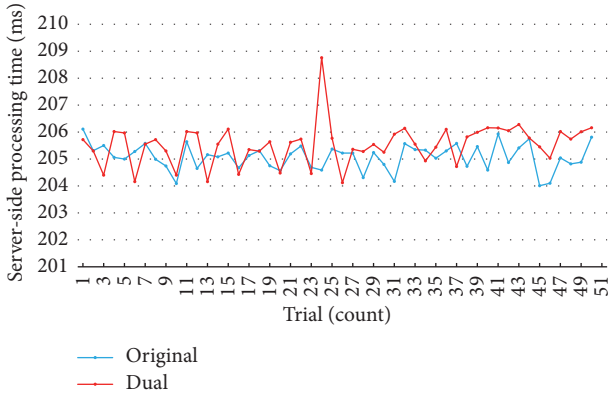


Figure 10: Server-side processing time.

Table 3: Server-side processing time.

	Server-side processing time
Original join	205.07 ms
Proposed initial join	205.53 ms

Table 4: Join accept processing time.

	Join accept processing time
Original join	1.3 ms
Proposed initial join	1.78 ms

on the increased delay of the proposed scheme. Therefore, the join accept communication delay can be inferred as the most decisive factor on the increased delay. The only change in the join accept message of the proposed scheme was the payload length. As shown in Table 5, the payload length of the join accept message increased by 16 bytes. We concluded that this is the main cause of the delay increase.

Table 5: Join accept packet size (without CFList field).

	Join accept payload length
Original join	12 bytes
Proposed initial join	28 bytes

For further analysis, we carried out a simulation with the LoRa Modem Calculator [31] provided by Semtech. The result shown in Figure 11 cannot be directly applied to our experimental results because the calculator does not support the SX1257 transmitter that is equipped in our IC880A gateway board. However, at least we can determine how the transmission time was changed according to the payload length.

We also considered the maximum payload length. According to [20], the maximum payload length varies from 59 bytes to 230 bytes depending on the data rate. Thus, we can ensure that the increased payload length did not affect feasibility because the 28-byte payload length satisfied the maximum payload length in any cases.

As a result, the total join delay of our scheme increased by about 18 ms because the payload length of the join accept packet increased. This payload length completely satisfied the maximum payload length criterion specified by LoRaWAN. Therefore, in terms of delay, our scheme is feasible.

(2) Battery Consumption. Our end node uses two AAA-sized 1.5-V batteries. The source code [21] provided by Semtech has

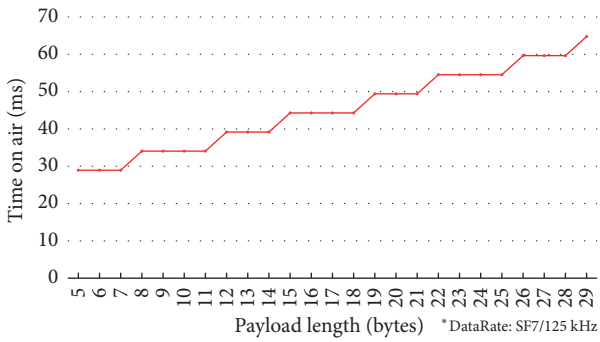


Figure 11: LoRa Modem Calculator result.

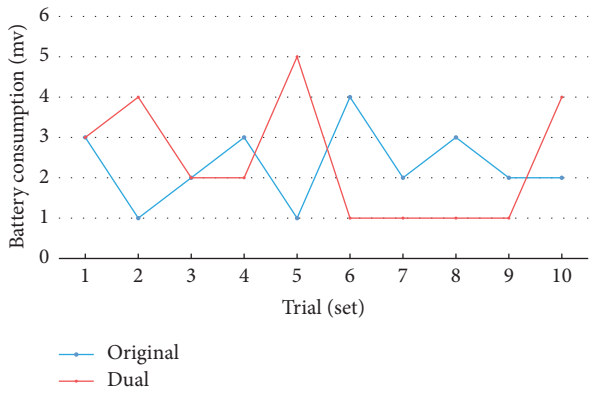


Figure 12: Battery consumption.

Table 6: Average battery consumption.

	Average battery consumption
Original join	0.23 mv
Proposed initial join	0.24 mv

a function that measures the remaining battery capacity of the end node in millivolts. We use this function to measure the battery consumed during the join procedure. However, we cannot obtain meaningful values in one or two join procedures because the battery consumption is less than one millivolt. To overcome the limitation of measurement method, we decide to measure after performing 10 consecutive join procedures. Figure 12 is the result of 10 sets of experiments, 10 times per set.

Due to transmission errors, function errors, and so on, the battery consumption value per set seemed not to be constant. We performed 10 sets, a total of 100 experiments, to ensure that these external factors are equally applied to both schemes. Therefore, the relative battery consumption of the proposed scheme for the original scheme is trustworthy. Table 6 shows that the battery consumption of the proposed scheme is not significantly different from the original scheme. Therefore, in terms of battery consumption, the proposed scheme is feasible.

8. Conclusion

In this paper, we proposed a dual key-based activation scheme. Our scheme uses NwkKey and AppKey to perform the initial join procedure. From the second join procedure, session keys are used, which are created in the previous join procedure. This resolves the key update problem, which was not fully supported in the original scheme. In addition, our scheme makes each layer generate its own session key so that the layers can work independently. We compared the performance of the original scheme and the proposed scheme through a real-world experiment. According to the experimental results, our scheme is feasible in terms of delay and battery consumption.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1D1A1A01058928).

References

[1] "Ericsson Mobility Report: On the Pulse of the Networked Society," 2015. http://www.ericsson.com/res/docs/2015/mobility-report/ericsson-mobility-report-nov-2015.pdf.

[2] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: a survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[3] X. Xiong, K. Zheng, R. Xu, W. Xiang, and P. Chatzimisios, "Low power wide area machine-to-machine networks: Key techniques and prototype," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 64–71, 2015.

[4] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi, "Long-range communications in unlicensed bands: The rising stars in the IoT and smart city scenarios," *IEEE Wireless Communications Magazine*, vol. 230, no. 5, pp. 60–67, 2016.

[5] U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low power wide area networks: an overview," *IEEE Communications Surveys & Tutorials*, 2017.

[6] O. Georgiou and U. Raza, "Low power wide area network analysis: can LoRa scale?" *IEEE Wireless Communications Letters*, vol. 6, no. 2, pp. 162–165, 2017.

[7] N. Sornin, M. Luis, T. Eirich, T. Kramp, and O. Hersent, "LoRaWAN Specification V1.0.2," *LoRa Alliance*, 2016.

[8] Z.-K. Zhang, M. C. Y. Cho, and S. Shieh, "Emerging security threats and countermeasures in IoT," in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security (ASIA CCS '15)*, pp. 1–6, ACM, April 2015.

[9] M. M. Hossain, M. Fotouhi, and R. Hasan, "Towards an Analysis of Security Issues, Challenges, and Open Problems in the Internet of Things," in *Proceedings of the IEEE World Congress on Services, SERVICES 2015*, pp. 21–28, July 2015.

[10] K. Zhao and L. Ge, "A survey on the internet of things security," in *Proceedings of the 9th International Conference on Computational Intelligence and Security, CIS 2013*, pp. 663–667, December 2013.

[11] S.-H. Seo, J. Won, S. Sultana, and E. Bertino, "Effective key management in dynamic wireless sensor networks," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 2, pp. 371–383, 2015.

[12] S. Agrawal, R. Roman, ML. Das, A. Mathuria, and J. Lopez, "A novel key update protocol in mobile sensor networks," in *Proceedings of the International Conference on Information Systems Security*, pp. 194–207, Springer, 2012.

[13] J. He, X. Zhang, and Q. Wei, "EDDK: Energy-efficient distributed deterministic key management for wireless sensor networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2011, 2011.

[14] E. Barker, W. Barker, W. Burr, W. Polk, and M. Smid, *Recommendation for Key Management Part 1: General (Revision 4)*, NIST Special Publication, 2016.

[15] P. Girard, "Low Power Widw Area Networks security," 2015. https://docbox.etsi.org/Workshop/2015/201512_M2MWORKSHOP/S04_WirelessTechnoforIoTandSecurityChallenges/GEMALTO_GIRARD.pdf.

[16] R. Miller, "LoRa Security: Building a Secure LoRa Solution," 2016. https://labs.mwrinfosecurity.com/publications/lo/.

[17] S. Zulian, *Security Threat Analysis and Countermeasures for LoRaWAN Join Procedure*, 2016, http://tesi.cab.unipd.it/53210/.

[18] S. Naoui, M. E. Elhdhili, and L. A. Saidane, "Enhancing the security of the IoT LoraWAN architecture," in *Proceedings of the 5th IFIP International Conference on Performance Evaluation and Modeling in Wired and Wireless Networks, PEMWN 2016*, November 2016.

[19] "LoRaWAN - What is it: A technical overview of LoRa and LoRaWAN," 2015. https://www.lora-alliance.org/portals/0/documents/whitepapers/LoRaWAN101.pdf.

[20] LoRa Alliance Technical committee, LoRaWAN Regional Parameters, 2016.

[21] Semtech, *LoRaWAN endpoint stack implementation and example projects*, 2013. https://github.com/Lora-net/LoRaMac-node.

[22] *SK-iM880B –Long Range Radio Starter Kit* https://wireless-solutions.de/products/starterkits/sk-im880b.html.

[23] *iC880A - LoRaWAN Concentrator 868MHz* https://wireless-solutions.de/products/radiomodules/ic880a.html.

[24] Semtech, *Use with Raspberry Pi*, 2016. https://github.com/Lora-net/packet_forwarder/wiki/Use-with-Raspberry-Pi.

[25] The Things Network, *From zero to LoRaWAN in a weekend*, 2016. https://github.com/ttn-zh/ic880a-gateway/wiki.

[26] Semtech, *LoRa Gateway project* https://github.com/Lora-net/lora_gateway.

[27] Semtech, *Lora network packet forwarder project* https://github.com/Lora-net/packet_forwarder.

[28] Brocaar, *LoRa Gateway Bridge* https://github.com/brocaar/lora-gateway-bridge.

[29] Brocaar, *LoRa Server* https://github.com/brocaar/loraserver.

[30] Brocaar, *LoRa App Server* https://github.com/brocaar/lora-app-server.

[31] Semtech, *LoRa Calculator* http://www.semtech.com/wireless-rf/rf-transceivers/sx1272/.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 6747052, 11 pages
https://doi.org/10.1155/2017/6747052

Research Article

Clustering Optimization for Out-of-Band D2D Communications

A. Paramonov,[1] O. Hussain,[1] K. Samouylov,[2] A. Koucheryavy,[1] R. Kirichek,[1] and Y. Koucheryavy[3]

[1]*The Bonch-Bruevich State University of Telecommunication, 22 Prospekt Bolshevikov, St. Petersburg, Russia*
[2]*Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St., Moscow, Russia*
[3]*Tampere University of Technology (TUT), Korkeakoulunkatu 10, Tampere, Finland*

Correspondence should be addressed to R. Kirichek; kirichek@sut.ru

Received 24 April 2017; Revised 12 August 2017; Accepted 12 September 2017; Published 22 October 2017

Academic Editor: Pai-Yen Chen

Significant increase in multimedia traffic challenges 5G networks in terms of capacity and correspondent QoS parameters. Device-to-device communication paradigm has already become an integral part of 3GPP standards; nevertheless it has not yet been widely deployed due to many different reasons. D2D is expected to leverage implementation of many qualitatively new services and to efficiently accomplish it D2D devices are supposed to form clusters. Due to practical limitations, current D2D implementations are mostly out-of-band and use Wi-Fi Direct. In this paper, we propose a novel model for throughput optimization in out-of-band D2D clusters. We delivered numerical results for different typical cluster member distributions and revealed key functional dependencies. Further, for the first time we compare clustering algorithms for out-of-band D2D and identify effective clustering algorithm that increases network resource utilization rate.

1. Introduction and Rationale

Integration of device-to-device (D2D) communication technology became a mainstream direction for fifth-generation (5G) communication networks. Driven by a huge increase in demand of multimedia traffic transfer, D2D communication allows saving scarce network resources by transferring data directly between devices either in-band or out-of-band, and D2D communications allow significantly reducing traffic between base station (BS) and end-user device [1].

According to the 3rd-Generation Partnership Project (3GPP) [2], D2D is a flexible paradigm of direct communication between devices which is open for use and based on cellular communication technologies (in-band D2D communication) and also WLAN technologies which are IEEE 802.11 standardization (out-of-band D2D communication) [3].

The last approach has recently become attractive due to the ease of implementation compared to in-band D2D, where end-user device shall be equipped with appropriate uplink/downlink functionality. In case of in-band D2D communications, the transmission power should be properly regulated so that the D2D transmitter does not interfere with the cellular UE communication while maintaining the minimum SINR requirement of the D2D receiver [4]. This significantly complicates feasibility of in-band D2D wide-scale implementation at least for the time being. Out-of-band D2D can be easily implemented with network assistance option; hence cellular operators are able to control out-of-band sessions. For the obvious reasons, IEEE 802.11-based Wi-Fi is taken as the transmission technology for implementation of out-of-band D2D functionality [5]. Operators of communication networks can encourage regular users to use D2D technology in order to improve the overall performance of the communication system in return for rewards proportional to their contributions.

Typically, geographically beside D2D nodes can form a cluster (Figure 1), where traffic circulates between cluster nodes directly, and outside-of-cluster traffic is forwarded to BS via relay node, so-called cluster head. A number of algorithms for cluster head selection are available today, e.g., [6, 7]. The decision on selection of a particular cluster member as a cluster head affects at least network efficiency, energy expenditures, and quality of service (QoS) offered to all members within the cluster. Generally, if all data transfer

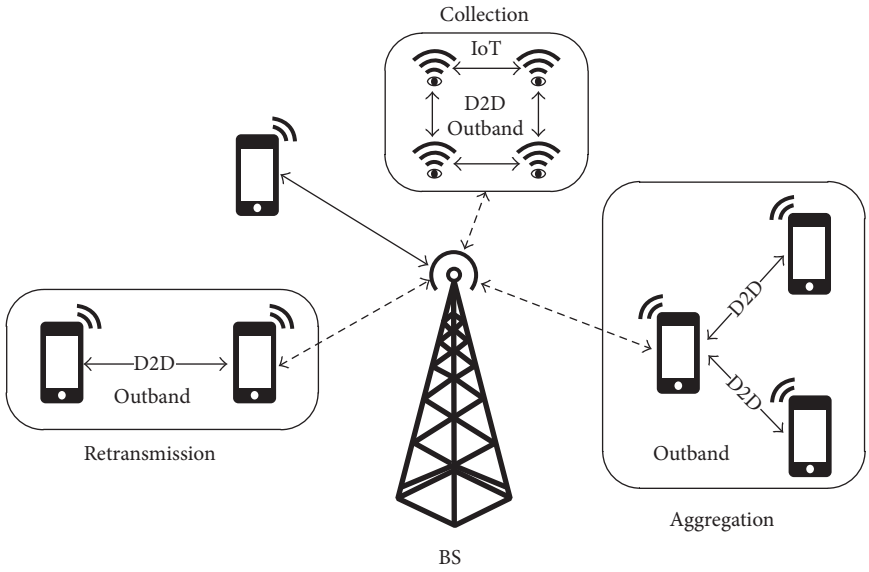


Figure 1: Out-of-band clustering of D2D components.

members are the members of the same cluster, the cluster can operate off-line, meaning without connection to network BS.

A larger number of cluster members are expected to lead to larger savings of the network resources. However, the maximum number of members in a cluster is restricted by coverage of selected D2D technology, channel throughput of cluster head and cluster traffic intensity, and cluster members physical location towards cluster head. Existing studies show that D2D clustering in 5G leads to reduction of signaling traffic and provides higher spectral efficiency and better energy performance than conventional cellular systems [7, 8]. Thus, efficient D2D clustering in 5G networks especially with high density of devices is of a paramount importance.

Multiple past works have concentrated on quantitative and qualitative analysis of cluster algorithms for D2D communications. In [1, 3] the authors provide comprehensive analysis of D2D communications. The use of out-of-band D2D communications and D2D clustering is discussed in detail given criteria of cluster head selection based on channel quality between cluster head and BS. In [9], the authors designed clustering algorithm for in-band D2D case, which increases system-level spectral efficiency. Numerical analysis and simulation modeling have shown that this proposal gives 66% gain in terms of throughput compared to traditional solutions, in the case where 20% of users use D2D communication. The authors derived the probability density formula (pdf) for the optimal number of repeater units in the cluster and have come up with the cross-cluster interaction scheme. Also via simulation the authors show that the proposed algorithm provides gains up to 40% in terms of network efficiency of resource use.

Different aspects of the out-of-band D2D communication are presented in [4, 10, 11]. In [4, 10], the authors developed analytical model of the network unloading for different D2D scenarios using stochastic geometry. The authors estimated

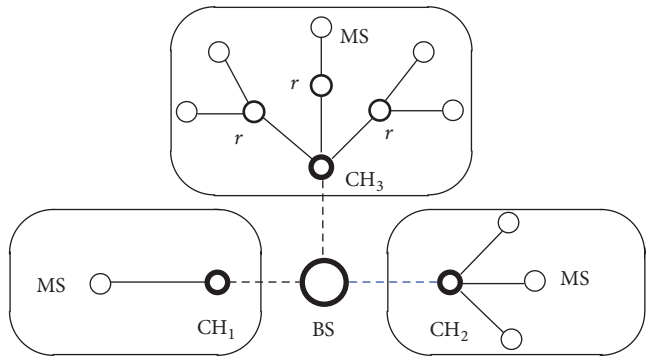


Figure 2: Possible cluster structures.

potential opportunities of the out-of-band D2D communication, using both the system level and the mathematical analysis. They show that at 30% of clustering productivity and energy network performance increase up to four and two times, respectively. In [11] the authors studied problems of implementation of network-assisted D2D communication while interacting in social networks. Besides, they use the existing experimental LTE testbed [12] for implementation of D2D system and show its performance evaluation in terms of latency and users satisfaction.

D2D transmission technology selection is still rather limited to Wi-Fi and Bluetooth due to wide implementation of those in consumer devices. Most recent of works, e.g., [13, 14], consider D2D devices forming clusters by using Wi-Fi Direct (see Figure 2). Due to features of radio channel, resources of channel between cluster member and cluster head may drastically vary for different nodes within one cluster. Therefore, while forming cluster, we suggest the way cluster head is selected shall be based primarily on anticipated

QoS parameters, not distance as many studies suggested up to date [15, 16]. The same applies to selection of cluster members. Clustering algorithm can be implemented for different target parameters such as cumulative throughput of cluster as a whole, maximum number of cluster nodes, and quality of service.

Summarizing the results of the analysis of publications, we can conclude that most of them are devoted to analysis and evidence of the effectiveness of using both in-band and out-of-band clustering. The efficiency of spectrum utilization is not the only criterion that should be considered when solving the clustering problem. Therefore, further in this paper we design clustering algorithm, which is characterized by bandwidth of the channels between cluster members and the cluster head.

The remainder of this paper is organized as follows. In Section 2, the reference scenario for D2D is defined and analytical model for cluster is introduced. The proposed model is delivered in Section 3. The required analytical and simulation performance evaluation campaign are reported in Section 4. Section 5 discusses clustering algorithms, whereas the concluding remarks and future research directions appear in Section 6.

2. Reference Scenario Definition

The use of D2D communication allows the increase in system effectiveness of cellular communication; moreover D2D directly influences at system level both efficiency and energy. The users are distributed on the BS coverage area randomly. Generally, network planning takes into account distribution of nodes in the geographical area letting operator provide at least wanted coverage and required throughput and QoS. The possible cluster structures are presented in Figure 2.

We assume that mobile *stations (MS)* can interact directly with base station (BS), through transit node (CH_1) or head node of a cluster (CH_2, CH_3). Generally, one cluster can have a star-like structure with single transit node, head node of cluster, or tree-like structure, including both head node and other transit nodes (r). The choice of the cluster structure can be done without the involvement of network functionality (for mobile stations) and involving the network functionality (for the BS). Shaping of the cluster consists of *the MS group* choice and distribution of their functionality within the cluster (terminal node, transit node, or head node of the cluster). To shape a cluster one needs to define the indicators that characterize the decision for cluster shaping (status indicators) and the criterion validity (quality) of decision and control settings that affect performance status and also method of finding the valid (optimal) solution.

The authors in [3, 4] suggested approach for the analysis and shaping of tree-like structure of clusters. Following the proposal, in this paper we focus attention on clusters with star-like structure with one head node; such structure is very useful for high density wireless environments such as apartment block houses, offices, and stadiums.

Quality of traffic service within a cluster and between cluster head node and BS depends on channels throughput between cluster participants (b_{ij}) and between head node and BS (b_k) and also on traffic intensity (a_i) produced by users. We suppose that in BS service area there are n of MS that support D2D mode. We indicate the set of MS as $M = \{m_1, m_2, \ldots, m_n\}$. Then the task of clustering consists of estimating a quantity of clusters (k) and choosing their structure when the best possible QoS is provided. We assume the efficiency of the solution is higher, if in a service area of BS there are a smaller number of channels (BS-CH) (i.e., quantity of clusters (k)) and a greater number of MS which are using D2D mode. At the same time QoS for the participants of clusters should not be below target value (the rule) b_{ij}^0. Generally, target values can be different for different users. They depend on characteristics of the produced traffic, that is, the type of services required.

We indicate the set of clusters in BS service area as $C = \{c_1, c_2, \ldots, c_k\}$. We assume that all clusters owned by multitude C are formed by elements of multitude M, and not all elements of M should be included in clusters, and the clusters have no general elements (they form disjoint subsets) $C \subseteq M$, $\forall c \subset C$ and $\cup C = \{\}$.

As it was said above, cluster shaping can be performed by various methods, the choice of which depends on desired outcome. Further we consider the possibility of use of certain centroid methods for the task of clustering objects [17, 18]. The solution of the clustering problem represents the solution to the optimization problem where certain metric $d(m, p_m)$ is minimized or maximized. This metric characterizes the "distance" between cluster participant and cluster center $p_m = (1/|c|) \sum_{m \in c} m$.

Distance, throughput, time delay, and so forth can be used as optimization metric. Generally, it is the task of nonconvex optimization which may not have a unique solution. As a rule, to solve this problem dynamic programming is required, which can minimize the parameter $d^2(m, p_m)$ for all clusters.

$$C = \min \sum_{c \in C} \sum_{m \in c} d^2 (m, p_m). \quad (1)$$

Well-known clustering algorithms allow us to find the particular (near-optimal) solution. As it was said above, clustering can be chosen in different scenarios, depending on goals and restrictions. Therefore, to solve the considered clustering task the analysis of possible solutions and approach to selecting of criteria and clustering method is required.

3. The Proposed Model Description

Quality of traffic service is instantiated by probability and time indexes as the probability of availability, discard probability, and data delivery time. These indexes depend on the traffic parameters and bandwidth of the network connection. Thus, for given traffic characteristics, the amount of bandwidth best describes the results of decision that is made in terms of quality of communication services. Under the throughput we assume the achievable data rate. The throughput is not a complete metric for the quality of service description. The quality of service depends on the traffic characteristics. To analyze such quality characteristics as

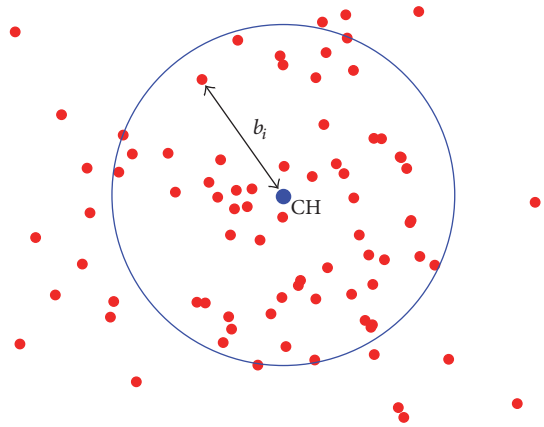


Figure 3: Cluster model.

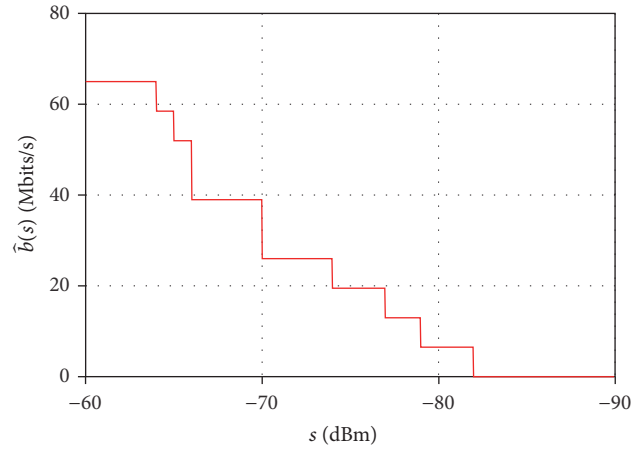


Figure 4: The data transfer rate dependence from the magnitude s.

packet loss ratio and delay it is necessary to build a complete model based on queuing theory. At this stage we concentrate on the initial model where known parameters such as number of users and distribution of users in the service area are available. We suppose that the initial clustering solution shall be done taking into account only throughput parameter in assumption that all users generate equal traffic flows. In fact, analytical results given below can be used for different types of traffic flows. As a target metric we consider throughput between network elements b_{ij}. In our analysis we consider the case when head node is already defined. We assume the communication area of head node represents as a circle with R radius, centered at the location of the head CH node, as it is shown on Figure 3.

Considering the IEEE 802.11 family standards as communication technology between D2D nodes, one shall define the type of dependence b_{ij}. According to [12], data transmission rate between two cluster members is defined by making the choice of modulation and coding scheme (MCS) according to receiving conditions (i.e., radio channel quality). These conditions are evaluated by signal power on receiver input $s = p_{rx}$ or the signal-to-noise ratio (SNR) or signal/noise + noise (SINR), where, in addition to the useful signal, the noise power is total power of all received signals, including noise. The noise power is the power created by nodes from neighboring clusters. According to [12] the model dependence of data transmission rate on signal power at the receiver input s is a jump function which increases with power growth. Figure 4 shows an example of dependence between data transfer rate and magnitude s for the IEEE 802.11n standard using the 20 MHz channel width.

Similar dependence can also be constructed for values of SNR and SINR. Signal power, SNR, and SINR are estimated by the equation given below.

$$
\begin{aligned}
s = p_{rx} &= 10 \lg \left(\frac{P_{rx}}{1\,\text{mW}} \right) \text{dBm}, \\
\text{SNR} &= 10 \lg \frac{P_{rx}}{P_N} = p_{rx} - 10 \lg \left(\frac{P_N}{1\,\text{mW}} \right) \text{dB}, \\
\text{SINR} &= 10 \lg \frac{P_{rx}}{P_N + P_I} = p_{rx} - 10 \lg \left(\frac{P_N + P_I}{1\,\text{mW}} \right) \text{dB},
\end{aligned}
\tag{2}
$$

where p_{rx} is the receiver input power (dBm); P_{rx} is the receiver input power (W); P_N is the input noise intensity (W); P_I is the receiver input interference power (W). The receiver input signal power can be described by the RCPI power indicator of receiving channel. According to [19] the value of this indicator is measured with an accuracy of ±5 dB (95% of confidential interval) taking into account the band noise, corresponding to the channel strip.

Along with the specified metrics, signal power on an input of the receiver can be described by the indicator of power of the accepted RSSI signal. For this value the exact compliance with a power of the accepted signal is not defined. According to the IEEE 802.11 standard it can vary from the minimum to the maximum value.

Each of the parameters specified above affects throughput of the channel and can be selected as a metric in the task of a clustering. The choice of specific parameter depends on possibility of its assessment and statements of the problem. In this article we generally restrict ourselves by the analysis of throughput in a cluster without considering a specific method of clusters shaping (this is a topic of further research). In the presented analysis as metrics we selected the input signal power on the receiver s which can be estimated by RCPI parameter.

The value of throughput which can be defined by this model (see Figure 4) depends on receiving conditions and, in practice, has the considerable dispersion. Taking this fact into account for analyzing the bandwidth we can assume that the jump model approximation by the continuous function does not propagate a significant error in the results but significantly simplifies the task.

Considering that most of D2D users are concentrated indoors, we describe the signal attenuation using the model recommended in [20] for indoor application (ITU-R 1238), thus considering out-of-band D2D:

$$
L(d) = 20 \lg (f) + N \lg (d) + P_f(n) - 28\,\text{dB}, \tag{3}
$$

where d is the distance in meters; f is the frequency in MHz; N is the back-off power; n is the number of obstacles; $P_f(n)$ is the loss of power parameter while passing over the

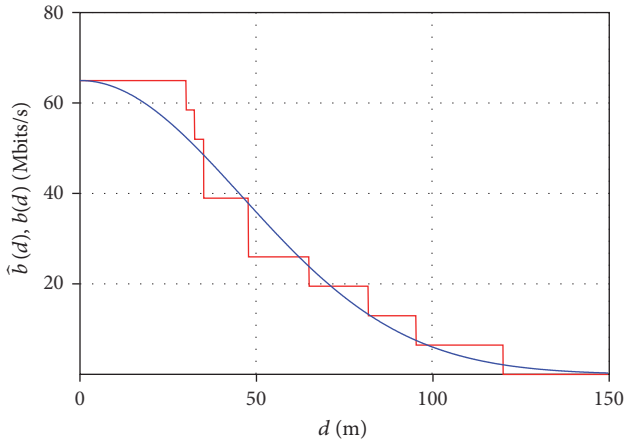


FIGURE 5: Dependence of data transfer rate on the distance to the device.

obstacle (dB). Considering the attenuation model we describe dependence of throughput on the distance of jump function (see Figure 5):

$$\widehat{b}(d) = \widehat{b}(L(d)). \tag{4}$$

The jump function was approximated by normal distribution [21], which reflects the trend of the throughput at varying distance.

$$b(d) = \begin{cases} 0 & d < 0 \\ b_{\max} e^{-d^2/2c^2} & 0 \le d \le R \text{ Mbps} \\ 0 & d > R, \end{cases} \tag{5}$$

where d is the distance in meters; D is the constant; $b_{\max}$ is the maximum possible data transfer rate (Mbps); c is the half-width of the curve in meters.

$$R = \arg\left\{\widehat{b}(d) = 0\right\} \text{ (m)}. \tag{6}$$

Generally, mobile stations are distributed across the service area in a random way; therefore, value of distance d and value of signal attenuation $L(d)$ in between are also a random value. In this case we do not consider signal depression, which also affects the character of a random value of attenuation and consequently throughput. In this analysis we consider only a factor of a relative positioning of users $b(d)$.

Since throughput is a function of a random value, the distribution function b can be determined according to [22] as

$$F(b) = \iint_{D_b} f(x, y)\, dx\, dy, \tag{7}$$

where D_b is the range of b values and $f(x, y)$ is the function of the user distribution within the circle with radius R.

The probability density of b can be determined as

$$f(b) = \frac{dF(b)}{db}. \tag{8}$$

The mathematical expectation of b is

$$M(b) = \int_0^{b_0} b \cdot f(b)\, db. \tag{9}$$

Further we consider two kinds of functions of the users distribution on the service area: (i) uniform distribution and (ii) normal probability distribution.

3.1. Uniform Distribution. The uniform distribution is described as a set inside of a circle, the square is S and the radius is R ($S = \pi R^2$), and the interval is $0 \le r \le R$. The radius R of the circle is defined as $R = \arg\{\widehat{b}(d) = 0\}$ (m).

Probability density function $f(r)$ inside the circle is constant:

$$f(r) = \frac{1}{S} = \frac{1}{\pi R^2}. \tag{10}$$

If the function expressing dependence of throughput b from distances to the base station has the form (5), that is,

$$b(d) = b_{\max} e^{-d^2/2c^2}, \tag{11}$$

where R is the radius of the service area, which is defined from the model of attenuation, we can express from (5) $d = c\sqrt{-2\ln(b/b_{\max})}$ (m).

Throughput distribution function b inside the circle R, according to (8), can be expressed as

$$\begin{aligned} F(b) &= \int_0^{2\pi} \int_{c\sqrt{-2\ln(b/b_{\max})}}^{R} \frac{1}{S} r\, dr\, d\theta \\ &= \frac{1}{2\pi R^2} r^2 \Bigg|_{c\sqrt{-2\ln(b/b_{\max})}}^{R} \cdot 2\pi \\ &= \frac{1}{R^2}\left(R^2 + 2c^2 \ln\left(\frac{b}{b_{\max}}\right)\right) \\ &= 1 + \frac{2c^2}{R^2} \ln\left(\frac{b}{b_{\max}}\right). \end{aligned} \tag{12}$$

The probability density function according to (9) is

$$f(b) = \frac{dF_b(r)}{dr} = \frac{d}{dr}\left(1 + \frac{2c^2}{R^2} \ln\left(\frac{b}{b_{\max}}\right)\right) = \frac{2c^2}{R^2 b}. \tag{13}$$

Throughput distribution and probability density functions are shown on Figure 6.

For example, for the IEEE 802.11g standard the mathematical expectation of the throughput $M(b)$ in the service area (circle R) according to (9) can be determined as

$$M(b) = \int_{b_{\min}}^{b_{\max}} b \frac{2c^2}{R^2 b} db = 2\frac{c^2}{R^2}(b_{\max} - b_{\min}) \text{ Mbps}. \tag{14}$$

By choosing approximation throughput function in the distance and using uniform users distribution in the service area, mathematical expectation of throughput, for IEEE 802.11n we obtain 19.51 Mbps.

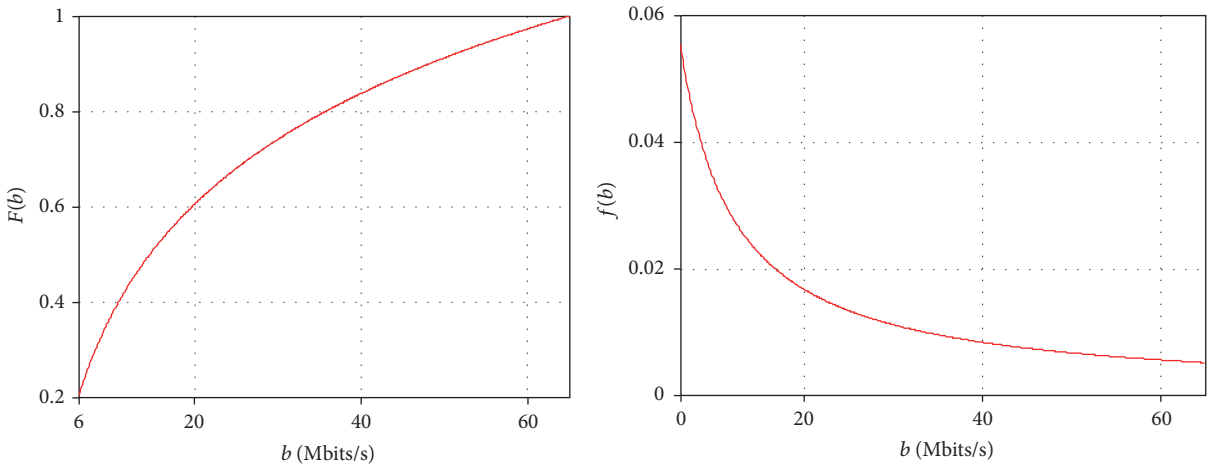


Figure 6: Throughput probability distribution and the probability density functions.

3.2. Normal Distribution. It is supposed that the traffic intensity in each point on the surface is a random value and is given by random, independent x and y coordinates. Then, the probability density function will be determined by the joint distribution function of random x and y. For a normal distribution with a center of dispersion in the center of the circle and circular dispersion (equality of the variance of x and y), the density distribution is equal to

$$f(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}, \qquad r = \sqrt{x^2+y^2}, \tag{15}$$

where σ is the root-mean-square deviation.

For throughput analysis, we assume that the probability of penetration inside the circle R (which is the service area) is equal to 1. The normal law of probability distribution is infinite according to the x and y values. In those conditions, the probability distribution law cannot be normal. However, with sufficient accuracy it can be described by normal truncated distribution [21] as

$$f(x,y) = K(\sigma,R) \cdot \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}, \tag{16}$$

where

$$K(\sigma,R) = \frac{1}{\iint_{S_R} (1/2\pi\sigma^2)\, e^{-(x^2+y^2)/2\sigma^2} dx\, dy}, \tag{17}$$

where S_R is the area that is restricted by the circle with radius R.

The b is throughput distribution function inside the circle R.

From (5) it can be seen that $d = c\sqrt{-2\ln(b/b_{\max})}$, and from (5) according to (16) and (17)

$$\begin{aligned} F(b) &= K(\sigma,R) \int_0^{2\pi} \int_{c\sqrt{-2\ln(b/b_{\max})}}^{R} \frac{1}{2\pi\sigma^2} e^{-r^2/2\sigma^2} r\, dr\, d\theta \\ &= K(\sigma,R) \frac{1}{2\pi} e^{-r^2/2\sigma^2} 2\pi \Bigg|_{c\sqrt{-2\ln(b/b_{\max})}}^{R} \\ &= K(\sigma,R) \left(\left(\frac{b}{b_{\max}} \right)^{c^2/\sigma^2} - e^{-R^2/2\sigma^2} \right). \end{aligned} \tag{18}$$

The throughput probability density according to (8) is

$$f(b) = K(\sigma,R) \frac{c^2}{\sigma^2 b_{\max}} \left(\frac{b}{b_{\max}} \right)^{c^2/\sigma^2 - 1}. \tag{19}$$

Throughput distribution function and probability density results are presented on Figure 7.

The mathematical expectation of the throughput value according to (19) is

$$\begin{aligned} M(b) &= \int_0^{b_0} b \cdot f(b)\, db \\ &= K(\sigma,R) \frac{c^2}{\sigma^2 + c^2} \left(b_{\max} - b_{\min} \left(\frac{b_{\min}}{b_{\max}} \right)^{c^2/\sigma^2} \right). \end{aligned} \tag{20}$$

The values of $M(b)$ are 55.72; 47.28; and 25.15 Mbps when the root-mean-square deviation values are 20, 30, and 80 m, respectively. Thus, the average throughput using the normal law of user traffic distribution inside the service area depends on the variance (scattering), when some bigger throughput values happen while there are lower values of variance.

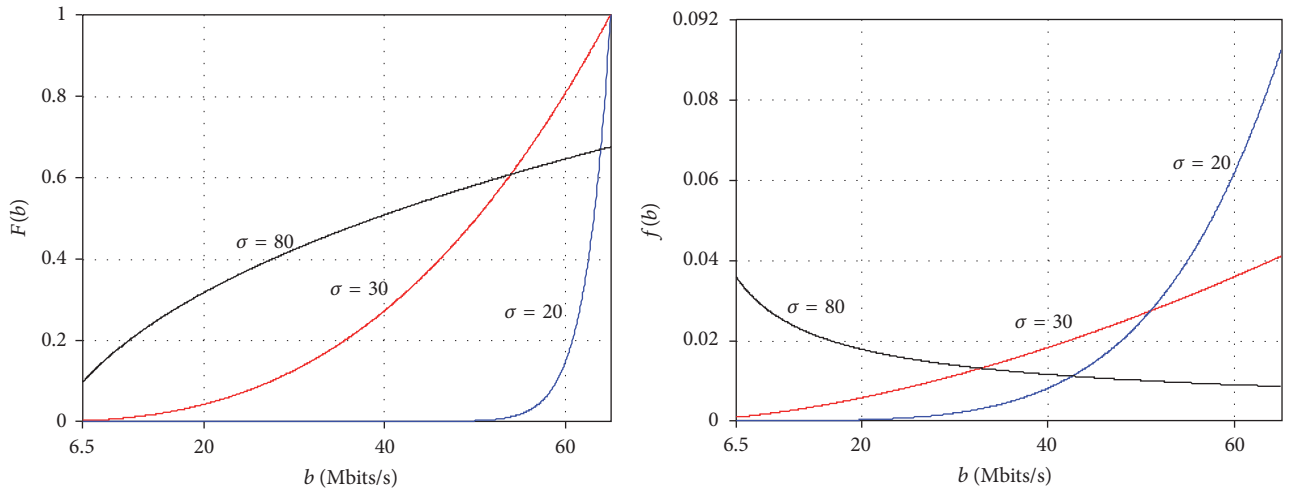


FIGURE 7: $F(b)$: throughput distribution function and throughput probability density $f(b)$.

Average throughput in the case of normal distribution always exceeds analogical value of uniform distribution.

From the analysis of the examples we can assume that the uniform distribution is the worst one according to the average amount of throughput criteria in comparison with other types of distributions, where the coordinates of the access point are coincident with the mathematical expectation of traffic distribution. The mathematical expectation of traffic distribution coordinate is the most appropriate base station installation point from the position of maximum throughput support (potential capacity), at least for unimodal distribution laws.

Using (12) or (18) depending on type of distribution of users we can estimate probability that the throughput not less than the given value:

$$P\left(b \geq B_{\min}\right) = 1 - F\left(b\right), \tag{21}$$

where $B_{\min}$ is the minimum throughput value permitted for a cluster member and $F(b)$ is the probability distribution given by (12) or (18).

In the condition of $B_{\min}$ restriction this probability (21) presents part of users included in clusters. It allows estimating a number of clustering users, choosing parameters of the network from (12) or (18). If in the network service area a number of randomly distributed clusters formed, using (21) we can estimate the number of users served with throughput no lower than $B_{\min}$.

4. Performance Evaluation

To verify the models above, a simulation campaign was performed. The following data represent the results of throughput simulation between user terminals and terminals within the communication area with uniform distribution. During simulation modeling there are permissible variations about the level of received signal that are verified by short fading effect and represent random value with a Nakagami distribution [22].

$$f\left(x, \alpha, \beta\right) = \frac{2}{\Gamma\left(\alpha\right)} \left(\frac{\alpha}{\beta}\right)^{\alpha} x^{2\alpha-1} e^{-(\alpha/\beta)x^2}, \tag{22}$$

where β is the shape parameter; α is the scale parameter; α and β are associated with the throughput by the formula given below:

$$\overline{b} = \frac{\Gamma\left(\alpha + 1/2\right)}{\Gamma\left(\alpha\right)} \sqrt{\frac{\beta}{\alpha}}. \tag{23}$$

Figure 8 represents analytical and simulation results for the case of uniformly distributed users: dependence between throughput and distance between cluster head and cluster member (a); probability density function (b), obtained via simulation (red) and analysis according to the formula (13) (blue). The average throughput for this case is 17.2 Mbps. Figure 9, in turn, represents simulation results for the case of normally distributed users: dependence between throughput and distance between cluster head and cluster member (a) and probability density function (b), obtained via simulation (red) and analysis according to formula (19) (blue). The average throughput for this case is 47.4 Mbps. Therefore, one can conclude that in case of normal distribution the average throughput between cluster head and cluster member is larger than 2.5 times that for the case with uniform distribution. Also, as it can be seen from the simulation and analytical distributions, the probability density function of throughput is sufficiently close to the simulation modeling results. This allows us to make a conclusion on credibility of obtained models for uniform and normal distributions of user devices. Hence, both analytical and simulation models allow defining the throughput in clusters with different user distribution in the service area of the head node.

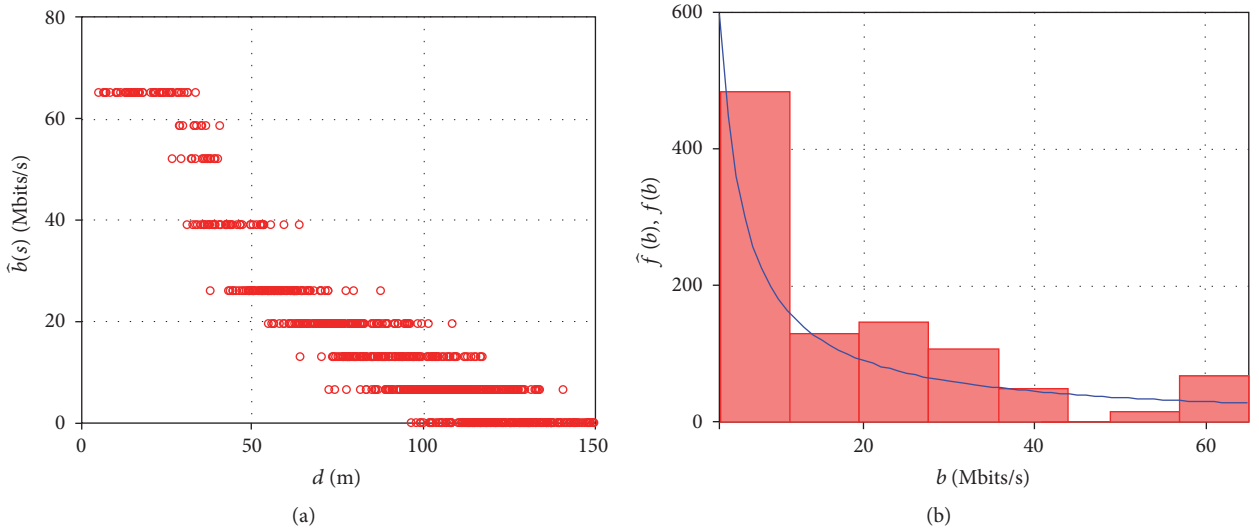


FIGURE 8: Dependence of throughput on the distance and probability density in case of uniform distribution of user devices.

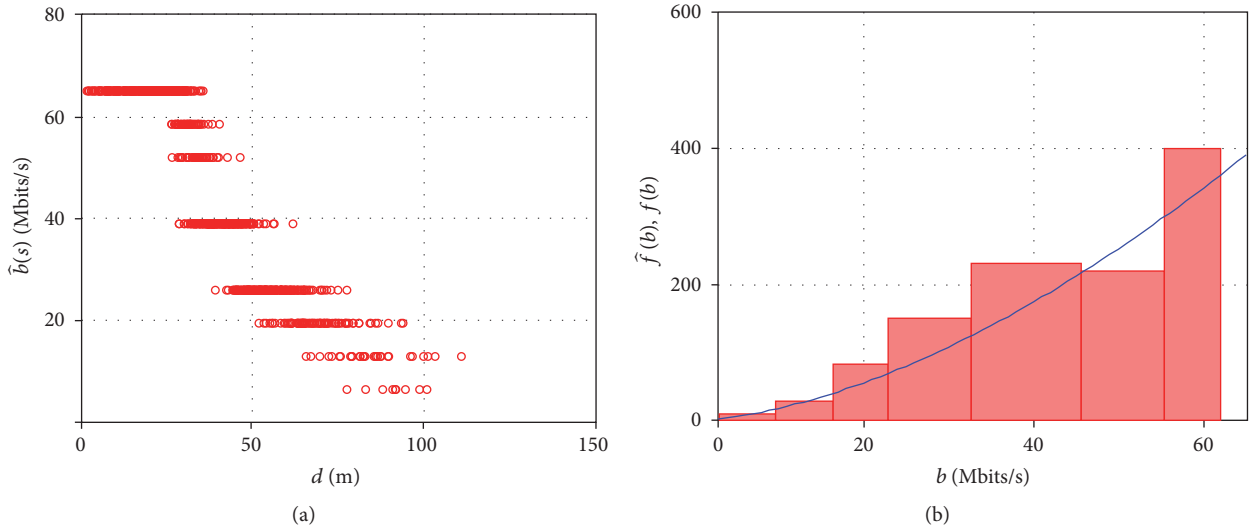


FIGURE 9: Dependence of throughput on the distance and probability density in case of normal distribution of user devices.

5. Clustering Method Selection

Figure 10 shows an example of clustering of 5,000 objects using two different algorithms [23, 24] (k-means algorithm on the left and FOREL algorithm on the right), obtained as a result of our simulation modeling. This is a theoretical example that does not reflect the real sizes of clusters in the network. However, this method quite clearly demonstrates the difference of the result, by choosing different methods of clustering.

As it can be seen from Figure 10, with an equal number of clusters (25 clusters in total) the shape of clusters is marked visually differently in the first and second cases.

To compare these methods we performed analysis of cluster members distribution processed relative to the cluster centers. Distribution was obtained by simulation modeling. For center-of-mass we used the following expression:

$$s = gm_j - Cm_i, \quad i = 1 \cdots |C|, \quad j = 1 \cdots n_i, \quad m_j \in c_i, \quad (24)$$

where Cm_i the is center-of-mass coordinate of cluster i; gm_j is the coordinate of m_j element of c_i cluster; c_i is cluster i; m_j is element j of cluster i; n_i is the number of elements within the cluster i; $|C|$ is the number of clusters.

Figure 11 shows the empirical probability density of cluster members coordinates relative to cluster center (histogram) and its approximation by the normal distribution (smooth curve) for k-means and FOREL cases, respectively. The diagrams show that in both the first and second cases the distribution of elements within clusters (relative to the centers

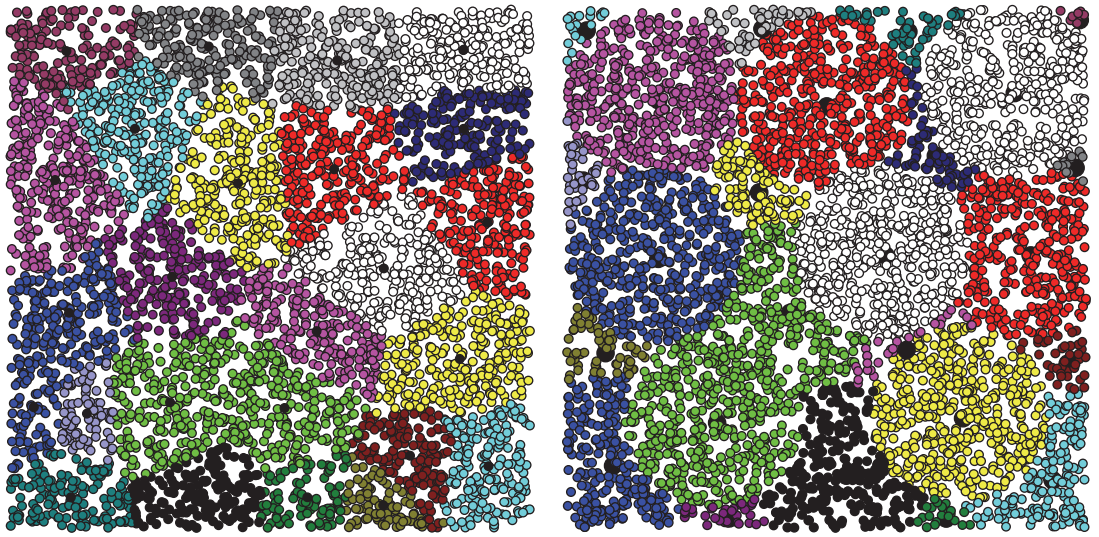


Figure 10: Clustering of 5,000 of elements using two different algorithms.

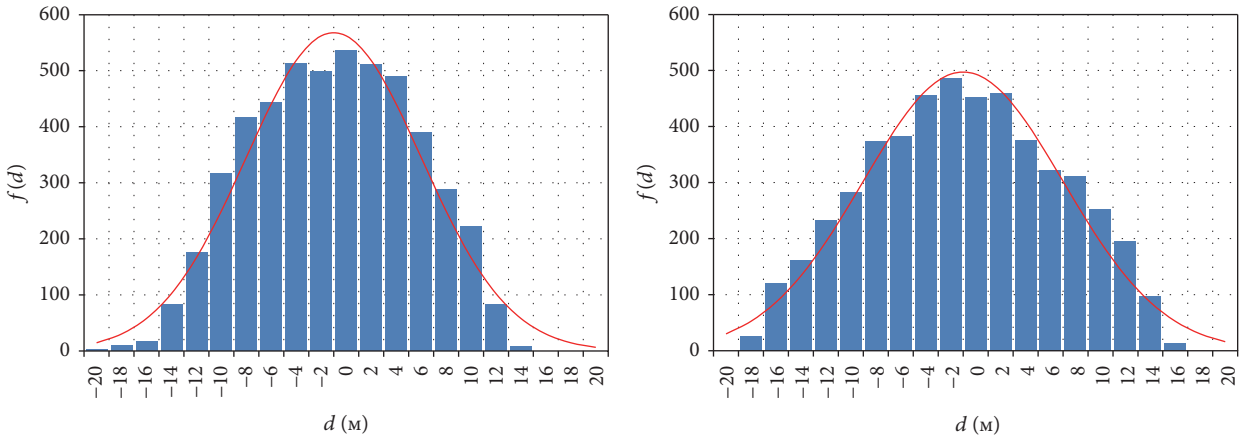


Figure 11: The node distribution within the clusters by using different methods of clustering (25 clusters in total).

of clusters) is close enough to the normal. It is worth noting that for the considered clustering algorithms dispersion of elements in clusters is remarkably different, which can be definitely seen from Figure 11.

Figure 12 shows an example for small number of clusters, 2 and 5 clusters, respectively. The figure shows that different algorithms form clusters in different ways. k-means split two clusters of a similar size, FOREL in turn has combined most of the elements in one cluster of a given size, and for the remaining elements 4 clusters were formed.

As it is shown on Figure 13, with a small number of clusters, the distribution of the elements inside them is closer to uniform than to normal, and more specifically it is closer to the original distribution of elements in a clustering area. It should be noted that the shape of the boundaries of the clusters when using k-means is close to the Voronoi diagram [17], which is constructed relative to the cluster centers. Therefore, in case of large number of clusters the distribution of cluster members can be approximated by normal law, in case of small number of clusters by uniform distribution.

6. Conclusions and Future Work

In this paper, we propose and evaluate a novel model for throughput estimation in out-of-band D2D clustering. Compared to existing studies where a distance between cluster head and cluster members was used, we suggest using a throughput. We obtained the numerical results for different types of typical cluster members distributions, uniform and normal. We delivered closed-form analytical expressions for probability distribution of throughput in the cluster for a given distribution and density of users. Through analytical and simulation studies, we show that the average throughput between cluster head and cluster member for the normal distribution is 2.5 times larger than for the case with uniform distribution. The obtained results show that known clustering

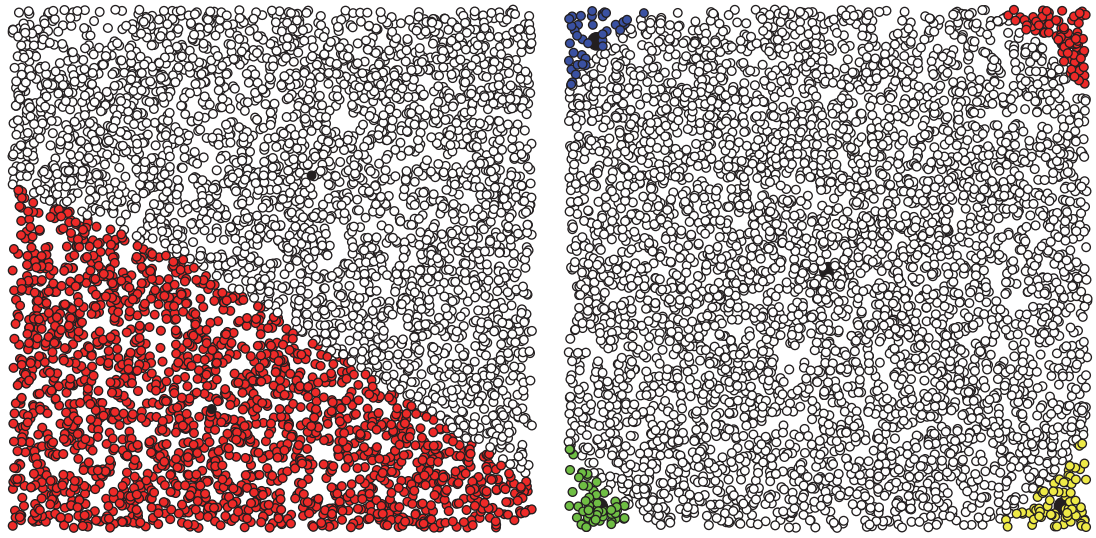


Figure 12: The node distribution within clusters by using different clustering methods, 2 and 5 clusters.

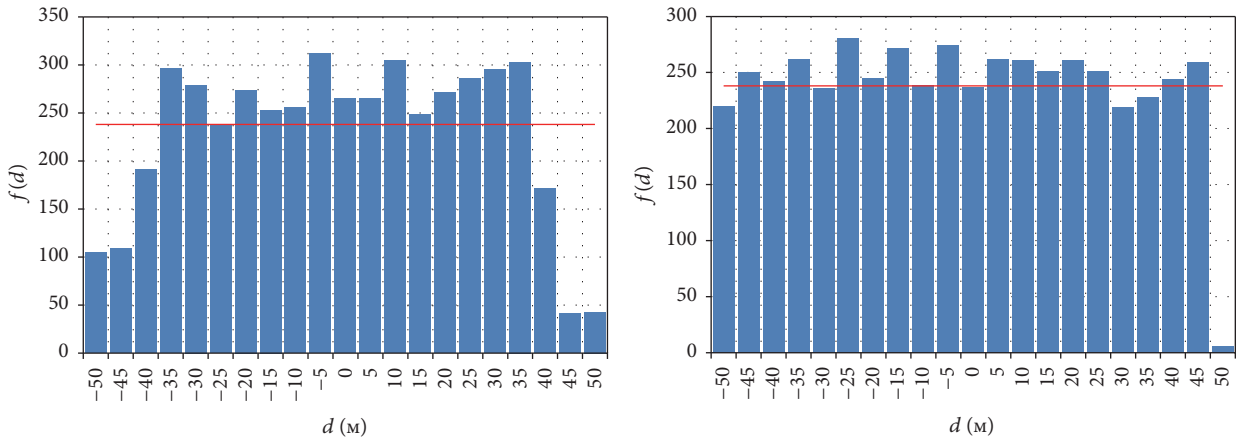


Figure 13: The node distribution within clusters by using different clustering methods, 2 and 5 clusters.

algorithms can be used to choose cluster head which provides near-optimal solution for throughput of channels between cluster head and cluster members.

Further, by using well-known clustering algorithms k-means and FOREL we obtained distribution of cluster members for large and small number of clusters. We show that, for the case with large number of cluster members, compared to FOREL k-means algorithm gives remarkably smaller dispersion of elements in clusters. Thus k-means constructs cluster of similar sizes, compared to FOREL that constructs clusters of very different sizes. Therefore, for out-of-band D2D case k-means provide better clustering considering target even resource distribution and increase of network resource utilization rate.

Our future work will concentrate on further enhancements of the delivered model by introducing more QoS metrics such as packet loss ratio, delay, and energy as optimization parameters. Also, a system-level performance evaluation would be needed to understand ultimate implications of the suggested model. In addition, we plan to consider 3D cases for out-of-band D2D scenarios, as well as cluster member location dynamics, which will allow addressing advanced drone-based scenarios of D2D communications.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] A. Asadi and V. Mancuso, "Network-assisted outband D2D-clustering in 5G cellular networks: theory and practice," *IEEE Transactions on Mobile Computing*, vol. 16, no. 8, pp. 2246–2259, 2017.

[2] "3rd generation partnership project; technical specification group services and system aspects; policy and charging control architecture (Release 13)," Tech. Rep. 23., 2015.

[3] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1801–1819, 2014.

[4] S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina, and Y. Koucheryavy, "Cellular traffic offloading onto network-assisted device-to-device connections," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 20–31, 2014.

[5] A. Gupta and R. K. Jha, "A survey of 5G network: architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[6] B. Zhou, H. Hu, S.-Q. Huang, and H.-H. Chen, "Intracluster device-to-device relay algorithm with optimal resource utilization," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 2315–2326, 2013.

[7] M. Ashraf, W.-Y. Yeo, M. Woo, and K.-G. Lee, "Smart energy efficient device-to-multidevice cooperative clustering for multicasting content," *International Journal of Distributed Sensor Networks*, vol. 2016, Article ID 3727918, 9 pages, 2016.

[8] L. Wang, G. Araniti, C. Cao, W. Wang, and Y. Liu, "Device-to-device users clustering based on physical and social characteristics," *International Journal of Distributed Sensor Networks*, vol. 2015, Article ID 165608, 14 pages, 2015.

[9] J. Seppälä, T. Koskela, T. Chen, and S. Hakola, "Network controlled Device-to-Device (D2D) and cluster multicast concept for LTE and LTE-A networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '11)*, pp. 986–991, Cancún, Mexico, March 2011.

[10] S. Andreev, O. Galinina, A. Pyattaev, K. Johnsson, and Y. Koucheryavy, "Analyzing assisted offloading of cellular user sessions onto D2D links in unlicensed bands," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 1, pp. 67–80, 2015.

[11] S. Andreev, J. Hosek, T. Olsson et al., "A unifying perspective on proximity-based cellular-assisted mobile social networking," *IEEE Communications Magazine*, vol. 54, no. 4, pp. 108–116, 2016.

[12] A. Pyattaev, J. Hosek, K. Johnsson et al., "3GPP LTE-assisted Wi-Fi-direct: Trial implementation of live D2D technology," *ETRI Journal*, vol. 37, no. 5, pp. 877–887, 2015.

[13] H. Shimodaira, J. Kim, and A. S. Sadri, "Enhanced next generation millimeter-wave multicarrier system with generalized frequency division multiplexing," *International Journal of Antennas and Propagation*, vol. 2016, Article ID 9269567, 11 pages, 2016.

[14] J. Kim, J.-J. Lee, and W. Lee, "Strategic control of 60 GHz millimeter-wave high-speed wireless links for distributed virtual reality platforms," *Mobile Information Systems*, vol. 2017, Article ID 5040347, 10 pages, 2017.

[15] K. S. Hassan and E. M. Maher, "Device-to-device communication distance analysis in interference limited cellular networks," in *Proceedings of the Tenth International Symposium on Wireless Communication Systems (ISWCS '13)*, 2013.

[16] M. Afshang, H. S. Dhillon, P. Han, and J. Chong, "Coverage and area spectral efficiency of clustered device-to-device networks," in *Proceedings of the 2015 IEEE Global Communications Conference (GLOBECOM '15)*, pp. 1–6, San Diego, Calif, USA, August 2015.

[17] F. Aurenhammer, R. Klein, and D.-T. Lee, *Voronoi Diagrams and Delaunay Triangulations*, World Scientific Publishing Co., Hackensack, NJ, USA, 2013.

[18] A. Muthanna, P. Masek, J. Hosek et al., "Analytical evaluation of D2D connectivity potential in 5G wireless systems," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 9870, pp. 395–403, 2016.

[19] IEEE Std 802.11™, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," 2012.

[20] ITU-R P.1238-8, "Propagation data and prediction methods for the planning of indoor radio communication systems and radio local area networks in the frequency range 300 MHz to 100 GHz".

[21] Wolfram MathWorld, "Gaussian Function," http://mathworld.wolfram.com/GaussianFunction.html.

[22] D. Laurenson, "Nakagami Distribution," *Indoor Radio Channel Propagation Modelling by Ray Tracing Techniques*, 1994.

[23] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York, NY, USA, 2005.

[24] J.-O. Kim and C. W. Mueller, *Factor Analysis: Statistical Methods and Practical Issues*, Sage Publications, Newbury Park, Calif, USA, 1978.