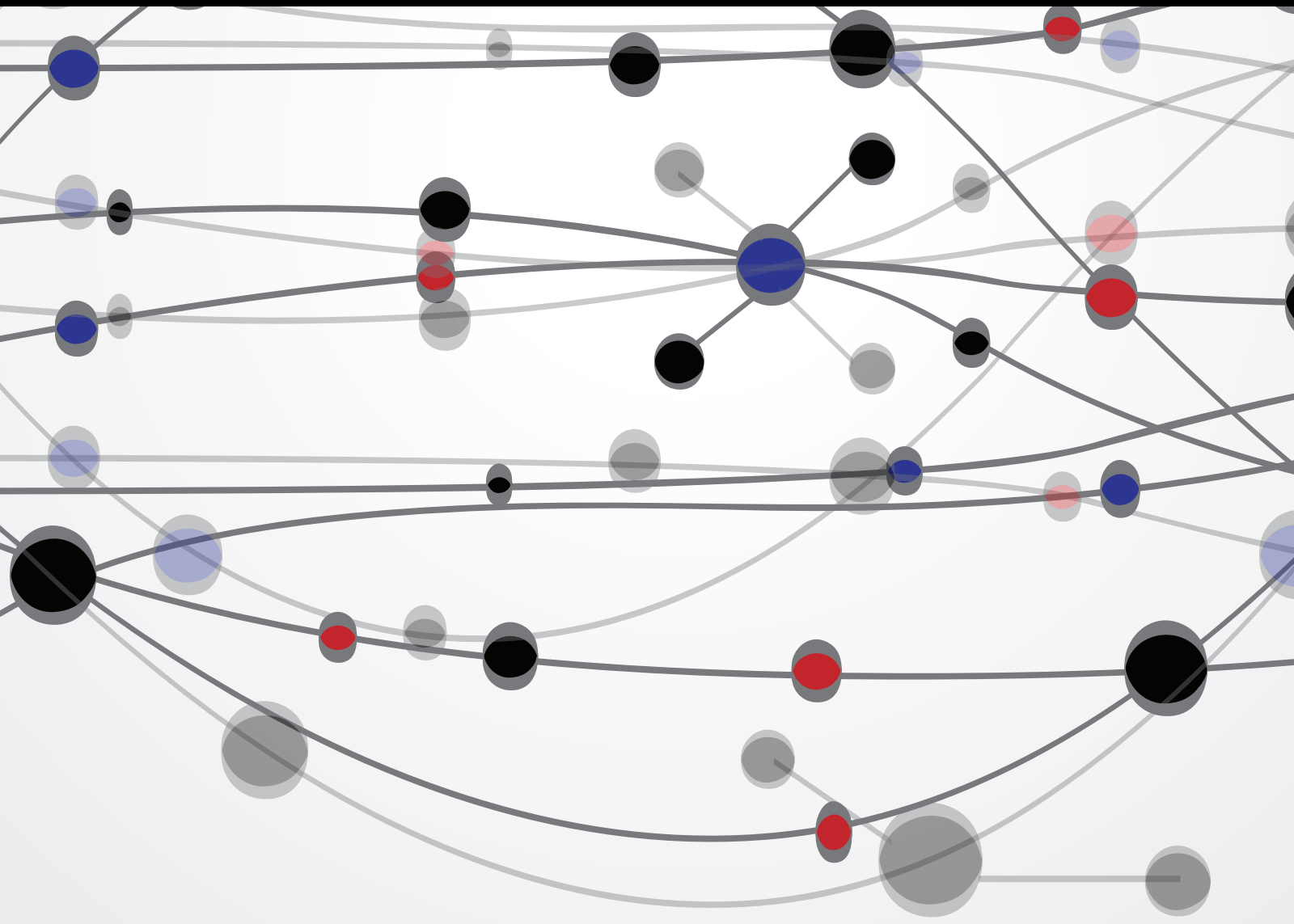
The Scientific World Journal

Next-Generation Internet and Communication

Guest Editors: Weifeng Sun, Guoqiang Zhang, Jingjing Zhou, and Vijay Bhuse



Next-Generation Internet and Communication

The Scientific World Journal

Next-Generation Internet and Communication

Guest Editors: Weifeng Sun, Guoqiang Zhang, Jingjing Zhou, and Vijay Bhuse

Contents

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 342471, 2 pages
http://dx.doi.org/10.1155/2014/342471

Editorial
Next-Generation Internet and Communication

Weifeng Sun,[1] Guoqiang Zhang,[2] Jingjing Zhou,[3] and Vijay Bhuse[4]

[1] *School of Software, Dalian University of Technology, Dalian 116620, China*
[2] *Nanjing Normal University, Jiangsu 210023, China*
[3] *School of Information and Electronic Engineering, Zhejiang Gongshang University, Zhejiang 310018, China*
[4] *Department of Computing, East Tennessee State University, Johnson City, TN 37614-1266, USA*

Correspondence should be addressed to Jingjing Zhou; zhoujingjing_cn@126.com

Received 12 March 2014; Accepted 12 March 2014; Published 3 June 2014

The Internet is one of the most key technical infrastructures in existence today. It will be a catalyst for much of our innovation and prosperity in the future. However, the present Internet may not be capable of meeting so many requests much longer. Fortunately, scientists and engineers believe that new technologies, protocols, and standards can be developed to meet tomorrow's demands. The research on this new Internet is considered as a global research challenge. Therefore, we have organized a special issue for researchers, engineers, and practitioners to share their work on next-generation Internet technologies.

Nowadays, the researches about next-generation Internet mostly focus on the design, engineering, protocols, and operation the of the new signal processing techniques in 3G/4G/B4G/LTE [1, 2], Ad Hoc Network and Wireless Mesh Network [3, 4], cloud computing [5], Internet of things (IoT) [6], channel allocation [7], and some other fields.

The papers in this special issue of next-generation Internet and communication cover a wide range of topics from underlying infrastructure and protocols to high-level systems design and analysis.

The new signal processing techniques in 3G/4G/B4G/LTE are a hot issue, and for this topic, we accepted three papers. They talk about the channel allocation and transmission problems in LTE.

Ad Hoc Network/Wireless Mesh Network has been a hot topic for several years and it is still a hot topic. The Wireless Mesh Network is Ad Hoc Network essentially, so we regard them as one topic. We accepted seven papers for this topic. They focus on the routing protocols and data delivery method in Ad Hoc Network or Wireless Mesh Network and propose some new routing protocols, such as a source-initiated on-demand routing algorithm.

Another important topic is channel allocation. We accepted six papers for this topic, including a paper about routing scheme. They talk about how to gain the network transfer capacity.

Security is another basic issue. For this topic, we accepted two papers. One is about defense model based on fuzzy evaluation. In the paper, first, the authors propose a distributed network security architecture, comprising a hybrid firewall, intrusion detection, virtual honey net projects, and connectivity and interactivity between these three components. The other one is about multiple-feature extracting modules. In the paper, they implement a crawler mining system that is equipped with SQL injection vulnerability detection, by means of an algorithm developed for the web crawler.

Cloud computing and IoT are very hot in recent years. We accepted two papers. One is about cloud computing. In the paper, the authors propose a novel unfair rating filtering method for a reputation revision system using prior knowledge as the basis of similarity when calculating the average rating, which facilitates the recognition and filtering of unfair ratings. The other one is about IOT. IOT accumulates large amounts of data by IOT users, which is a great challenge to mining useful knowledge from IOT. Classification is an effective strategy which can predict the need of users in IOT. In the paper, a new rule-based classification is proposed.

All the researches about Internet are for the purpose of getting better services, so the services of Internet are also

important topics. For this topic, we accepted three papers. They are about social network, content centric Internet, and power WebGIS platform.

We hope that readers will find in this special issue not only the new ideas, cutting-edge information, new technologies, and applications of next-generation Internet and communication, but also a special emphasis on how to solve various problems.

Weifeng Sun
Guoqiang Zhang
Jingjing Zhou
Vijay Bhuse

References

[1] A. Jayanthiladevi, H. M. Premlatha, and G. M. K. Nawaz, "Analysis study of seamless integration and intelligent solution in any situation by the future advanced mobile universal systems 4G—(FAMOUS 4G)," in *Proceedings of the International Conference on Electronics and Telecommunication System (ICEVENT '13)*, pp. 1–5, 2013.

[2] L. Duan, J. Huang, and J. Walrand, "Economic analysis of 4G network upgrade," in *Proceedings of the IEEE INFOCOM*, pp. 1070–1078, 2013.

[3] S. Fujiwara, T. Ohta, and Y. Kakuda, "An inter-domain routing for heterogeneous mobile ad hoc networks using packet conversion and address sharing," in *Proceedings of the 32nd International Conference on Distributed Computing Systems Workshops (ICDCSW '12)*, pp. 349–355, 2012.

[4] L. Song, C. Zhao, and C. Zheng, "Analysis and optimization model of cognitive wireless mesh networks," in *Proceedings of the International Conference on Industrial Control and Electronics Engineering (ICICEE '12)*, pp. 1426–1429, August 2012.

[5] M. Bourguiba, K. Haddadou, I. E. Korbi, and G. Pujolle, "Improving network I/O virtualization for cloud computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 3, pp. 673–681, 2014.

[6] S. Nastic, S. Sehic, M. Vogler, H. L. Truong, and S. Dustdar, "PatRICIA—a novel programming model for IoT applications on cloud platforms," in *Proceedings of the 2013 IEEE 6th International Conference on Service-Oriented Computing and Applications (SOCA '13)*, pp. 53–60, December 2013.

[7] S. N. Ohatkar and D. S. Bormane, "Channel allocation technique with genetic algorithm for interference reduction in cellular network," in *Proceedings of the 2013 Annual IEEE on India Conference (INDICON '13)*, pp. 1–6, December 2013.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 507517, 11 pages
http://dx.doi.org/10.1155/2014/507517

Research Article

An Optimization Algorithm for Multipath Parallel Allocation for Service Resource in the Simulation Task Workflow

Zhiteng Wang,[1] Hongjun Zhang,[1] Rui Zhang,[1] Yong Li,[2] and Xuliang Zhang[1]

[1] *PLA University of Science & Technology, Nanjing 210007, China*
[2] *Nanjing Artillery Academy, Nanjing 210110, China*

Correspondence should be addressed to Zhiteng Wang; zhitengwang@sina.com

Received 9 October 2013; Accepted 25 November 2013; Published 12 May 2014

Academic Editors: W. Sun, G. Zhang, and J. Zhou

Service oriented modeling and simulation are hot issues in the field of modeling and simulation, and there is need to call service resources when simulation task workflow is running. How to optimize the service resource allocation to ensure that the task is complete effectively is an important issue in this area. In military modeling and simulation field, it is important to improve the probability of success and timeliness in simulation task workflow. Therefore, this paper proposes an optimization algorithm for multipath service resource parallel allocation, in which multipath service resource parallel allocation model is built and multiple chains coding scheme quantum optimization algorithm is used for optimization and solution. The multiple chains coding scheme quantum optimization algorithm is to extend parallel search space to improve search efficiency. Through the simulation experiment, this paper investigates the effect for the probability of success in simulation task workflow from different optimization algorithm, service allocation strategy, and path number, and the simulation result shows that the optimization algorithm for multipath service resource parallel allocation is an effective method to improve the probability of success and timeliness in simulation task workflow.

1. Introduction

With the development of web technology, web service technology, SOA, grid, and cloud computation, there is a great revolution in modelling and simulation field. A framework of modelling and simulation based on simulation service, that is, service oriented modelling and simulation, is being formed. Service oriented simulation task framework from service oriented idea depends on information grid infrastructure with unified simulation description, access, and share standard to complete simulation task with the way of dynamic creation and running in the simulation resource interlink and share environment. Compared with traditional simulation framework, one of the service oriented modeling and simulation framework main characteristics is supporting dynamic creation and running simulation application and dynamic integration and running simulation in network by calling simulation service resource according to the simulation task workflow need.

In service oriented military modeling and simulation field, it is important to improve the probability of success and timeliness in simulation task workflow. Many researchers have done lots of research on how to improve the probability of success and timeliness in service composition in many stages. Some researchers try to improve the probability of simulation task workflow work success instead of failed service [1–3], but this method does not consider the timelessness of workflow. Therefore, it will produce extra time charge when the called service cannot respond or makes a mistake. Artificial intelligence (AI) planner is also applied to solve service composition combinatorial optimization problem [4–7]. But for the reason that it could immensely increase coupling between AI planner and other procedures when the special AI planner is used to solve automatic service composition problem, it will lead to limiting its web composition application filed. In addition, there are many differences in AI planner for applicability and designs performance, and this reason will result in biased service composition scheme and larger composition charge. Yajuan Song proposes a blend web service composition scheme based on buffer pool [8], in which dynamic service composition scheme is saved in buffer pool and user can use the composition scheme as well as use predefined workflow. But the composition algorithm should

be improved. ZHANG Bo investigates a service composition algorithm based on sub-Web service, in which the multi-input and multioutput parameter of web service is divided into sub-Web service with multi-input and single output parameters. In this way, it reduces dependencies of a Web service for multiple outputs with higher composition efficiency. In this way, we reduce dependencies of a Web service on multiple-output with higher composition efficiency; however, there exists redundancy in the composition route. Genetic algorithm (GA) and particle swarm optimization (PSO) [9] are typical evolution algorithm in service composition optimization filed, and PSO is more effective than GA with advantages such as too few arguments specified and the faster convergence speed. However, the local particle in the swarm is easy to be regarded as global best position leading to quickly converging to local best value. Zheng et al. [10] and Zeng et al. [11] use objective programming method to find the best path satisfied with condition by building the objective function with the condition and the probability of success and time requisition. The shortcomings of which may not complete the workflow requests.

Xiaohao et al. propose parallel allocation algorithm for service [12], in which one node starts many services to improve the probability of success and time requests in the simulation task workflow. The method gets good result but there exists exception probability in large-scale simulation task. Therefore, based on service parallel allocation in [12], this paper extends the scheme and proposes an optimization algorithm for multipath service resource parallel allocation scheme to further improve the probability of simulation task workflow work success and time and solve service scheduling problem.

The following section will investigate the systematic structure of simulation task body from two perspectives, that is, the reference model and the framework structure of simulation task body.

In service oriented military model and simulation field, it is important to ensure the timeliness and the probability of success in the simulation task workflow or it will greatly affect the workflow work efficiency and even lead to workflow which cannot complete the task. In fact, it could improve the timeliness and the probability of success by the optimization of the service composition in simulation task workflow. However, many optimization algorithms have deficiencies in solving this problem. For this reason, this paper proposes an optimization algorithm for multipath service resource parallel allocation for service resource to improve the timeliness and the probability of success in the simulation task workflow.

2. The Description of Service Composition Problem

2.1. The Concept of Service Scheduling and Simulation Task Workflow Framework. The concept of workflow originated from the organization of production and office automation field, in which work was separated into well-defined task, role, and the work is executed and monitored by specific rule and process to improve work probability and reduce production costs and raise enterprise competitive power. In military service oriented model and simulation field, workflow idea is introduced to arrange the role produce specific event under the simulation plot rule in simulation task. The event of role is related to simulation service agent, which is a service set and could call the simulation service by simulation service bus. The specific simulation service in service sets is eventual execution unit, and the purpose of service scheduling is to find the best path to allocate the simulation service in service sets to specific event of role in simulation plot. It can be shown in Figure 1.

The scheduling is to build a mapping relation from simulation event sets to simulation service sets by the scheduling algorithm. Its purpose is to choose the best service composition path under the simulation task limited condition to effectively complete the simulation. In the following part, a simulation task example of artillery firepower attacking is described. In this simulation task plot, artillery firepower attacking step is as follows.

(1) The artillery needs to get enemy's position by radar search service.

(2) The artillery needs to get enemy's global situation by situation analysis service.

(3) The artillery needs to prepare for firing by firing preparation service.

(4) The artillery needs to fire allocation scheme by fire allocation service.

In order to improve timeliness and the probability of success in simulation task, the probability of success and response time should be constrained. In the first phase, the longest echo time (LET) of radar service should be less than 10 seconds and the success call probability (SCP) should be more than 98%. In the second phase, for situation analysis service and firing preparation service, LET is required less than 30 seconds and SCP is required more than 98%. In the third phase, for situation analysis service and firing preparation service, LET is required less than 30 seconds and SCP is required more than 98%. For the whole of simulation task, LET is less than 40 seconds and SCP is required more than 98%. It can be shown in Figure 2.

2.2. The Description of Service Scheduling Model. Simulation task plot is used to arrange role completion action in specific time and position and it is important measure for workflow service scheduling. In fact, service scheduling is to optimize the simulation service composition, and service scheduling model can be simply described as the relation of simulation service composition optimization according to the simulation plot. Therefore, service scheduling optimization model can be described as: $P = \{S, P, C\}$ wherein S is representative of the service sets according to the event in plot, $S = \{S_1, S_2, S_3, \ldots, S_n\}$; P is representative of the mapping from the event relation to simulation service relation; C is representative of the constrained condition in the whole process.

2.3. The Optimization of Service Composition Path. The service composition path is selected one by one from simulation

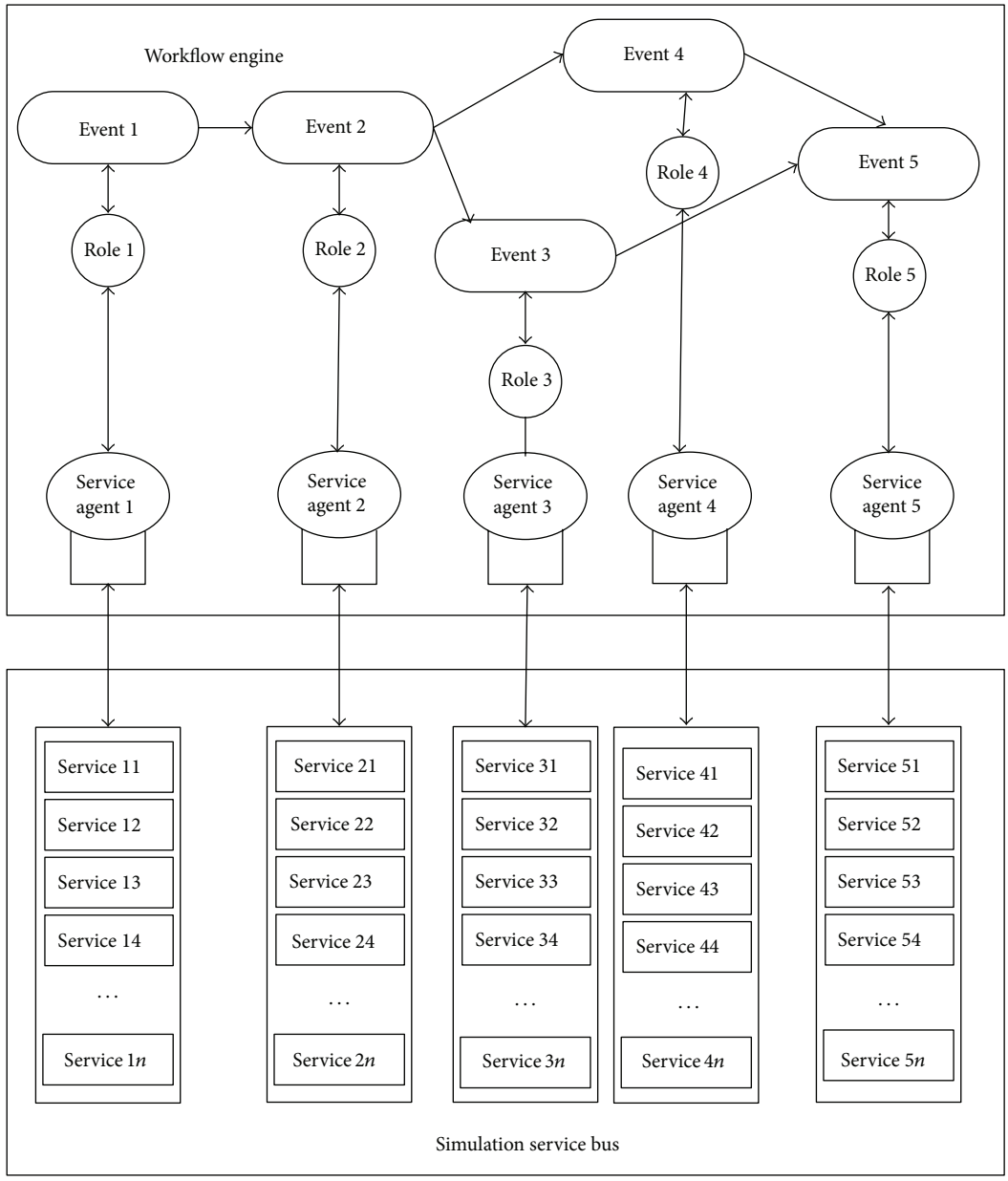


Figure 1: Simulation task workflow running framework.

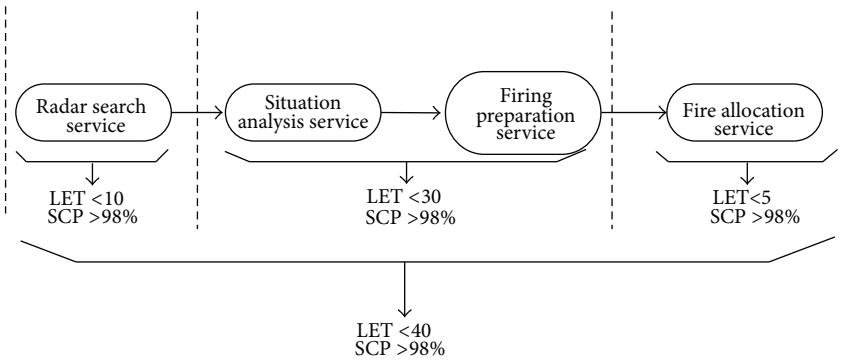


Figure 2: The service composition model under condition.

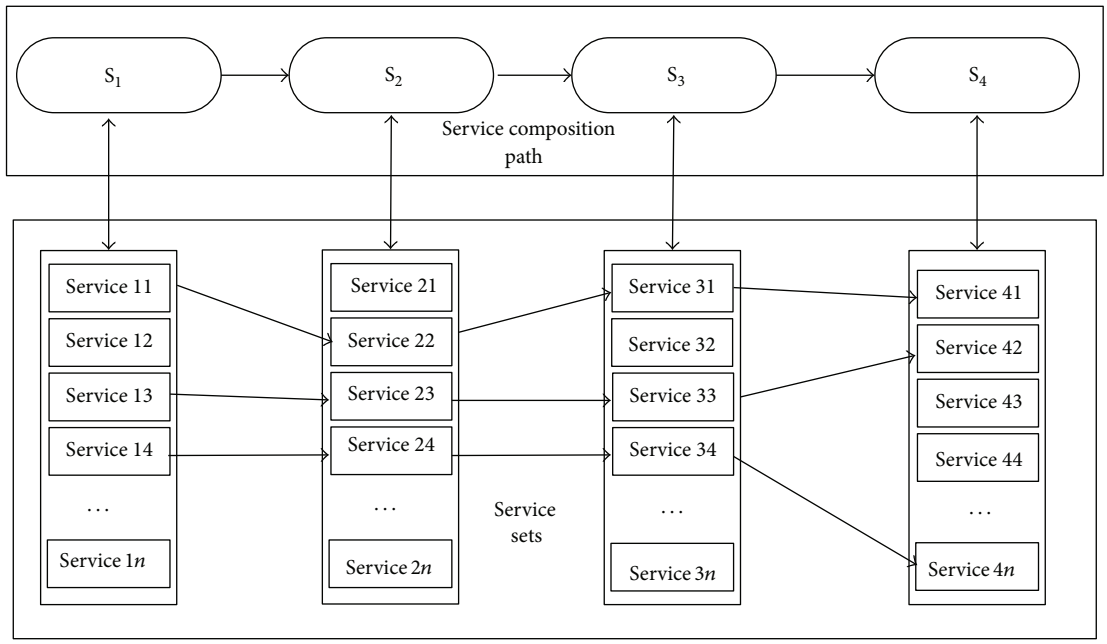


Figure 3: Service composition path choosing.

service sets to form a service composition path to complete simulation task. It can be shown in Figure 3.

From Figure 3, we can see that many paths can be selected by the methods of exhaustion, intelligence algorithm, and so forth to compile statistics quality of service and get the service composition path. However, these methods cannot ensure that the path is the best choice, because these methods are only considering the ideal condition without considering unexpected exception in practice.

3. The Strategy of Multipath Parallel Allocation for Service

When the simulation service is called, various problems may occur in the simulation task workflow during run time and the task may not be completed. In order to solve this problem, many researchers have taken a variety of approaches such as failed service instead, recall service again, and reselection of the path of service composition. The method of failed service instead is to replace the simulation service which is unable to continue to provide the service in process of the simulation task, and then another appropriate simulation service which has the same function in the service sets is called to complete the service function which the exception simulation service should complete. It can be shown in Figure 4.

The method of recall service is to recall service when the simulation service cannot complete the appropriate service function and surpass maximum response time. This method can be applied when the simulation service in service sets is not enough or it requires a higher limitation of simulation to use. Because the time expense for this method is small, if we recall the service again, it will restore the function promptly, so the simulation task workflow may restore the work in the shortest time.

The method of reselection of the path of service composition is to choose a service composition path again under the constrained condition and then run the simulation task again. This method's time expenses are large. Because it means the simulation service in original path should complete the task again in the new service composition path, so this method is used in the situation in which both of the expiration substitution method and recall service method are unable to complete the simulation task.

In fact, although these methods may reduce the influence to some extent on the simulation task workflow when the simulation service makes the mistake, it also takes time expenses as price. This is not permitted for real time simulation task, so it needs to use a more nimble method to solve this problem. In order to improve the probability of success and timeliness in simulation task workflow, this paper proposes an optimization algorithm for multipath service resource parallel allocation to handle service failure to be called in the service composition process, and each path is a parallel distribution services allocation plot. In the practice, various traditional methods such as failed service instead and recall service can be used synthetically. For example, two paths are selected and each node has two parallel services in each path. It can be shown in Figures 5(a) and 5(b).

This method is simultaneously calling many paths, and in each path simulation service simultaneously is called with many parallel services in service sets. The shortest response time of the first path in a number of paths can be selected as the main implementation of the path, in which any service can be selected as a main service when it has the shortest response time. If other main service is failed, other parallel service will be selected as main service. The difference from traditional instead methods is the parallel services running, and the workflow work cannot be effected

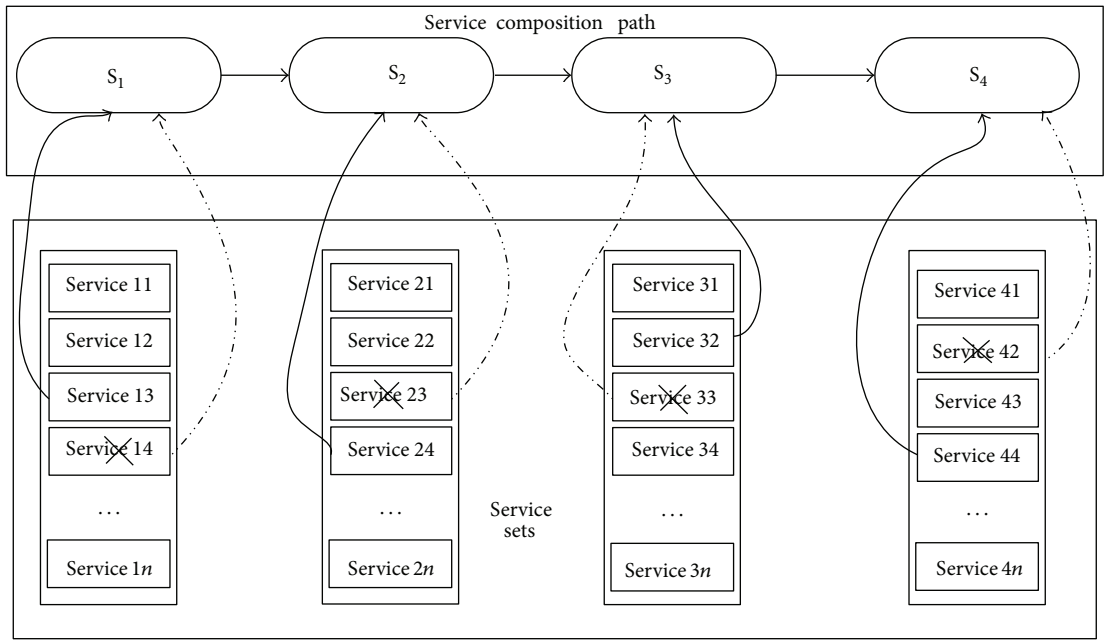


Figure 4: Failed service instead sketch.

even if one service is exception. Therefore, it nearly does not spend other expenses. In particular, because many paths run simultaneously, if the service in the main path produces an unexpected result, the service in another path can also run instead of it in time to effectively improve probability of success and timeliness for simulation task.

4. The Model of Multipath Service Parallel Allocation

Suppose there exist n agents in accordance with n service sets in workflow. S_i $(i = 1, 2, 3, \ldots, n)$ represents the service in accordance with the ith service agent. S_{ij} $(j = 1, 2, 3, \ldots, m)$ represents the ith service agent in accordance with m services in simulation service bus and the j represents the jth service in m services called. Suppose the probability of success service S_{ij} is $f(i, j)$ and its completion time follows a normal distribution under mean as u_{ij} and variance as σ_{ij}^2. When the workflow simultaneously call m service in S_i, its probability can be shown in the following formula:

$$\mathrm{Suc}(i) = 1 - \prod_{j=1}^{m} \left(1 - f(i, j)\right). \tag{1}$$

$M_i = \mathrm{Min}(t_{i,1}, t_{i,2}, \ldots, t_{i,m})$ is used to represent time distribution function in accordance with the ith service agent, where $t_{i,j}$ is the ith service in m services called by service agent. It is normal distribution and can be shown in the following formula:

$$\begin{aligned} T_i(t, j) &= \mathrm{Suc}\left(M_i \le t\right) \\ &= 1 - \mathrm{Suc}\left(M_i \ge t\right) \\ &= 1 - \mathrm{Suc}\left[\bigcap_{j=1}^{m} \left(t_{i.j} \ge t\right)\right] \\ &= 1 - \prod_{j=1}^{m} \mathrm{Suc}\left(t_{i,j} \ge t\right) \\ &= 1 - \prod_{j=1}^{m} \left[1 - \mathrm{Suc}\left(t_{i,j} < t\right)\right] \\ &= 1 - \prod_{j=1}^{m} \left[1 - T_{i,j}(t)\right] \\ &= 1 - \prod_{j=1}^{m} \left[1 - \frac{1}{\sqrt{2\pi}\sigma_{i,j}} \int_0^t e^{-(t-u_{i,j})^2/2\sigma_{i,j}^2} dt\right]. \end{aligned} \tag{2}$$

The success probability of simulation service S_i in the ith simulation in the waiting time of d_i is as follows:

$$\begin{aligned} \mathrm{Suc}T_i(d_i, i) &= \mathrm{Suc}(i) * T_i(t, j) \\ &= \left[1 - \prod_{j=1}^{m} \left(1 - f(i, j)\right)\right] \\ &\quad \times \left[1 - \prod_{j=1}^{m} \left(1 - \frac{1}{\sqrt{2\pi}\sigma_{i,j}} \times \int_0^{d_i} e^{-(t-u_{i,j})^2/2\sigma_{i,j}^2} dt\right)\right]. \end{aligned} \tag{3}$$

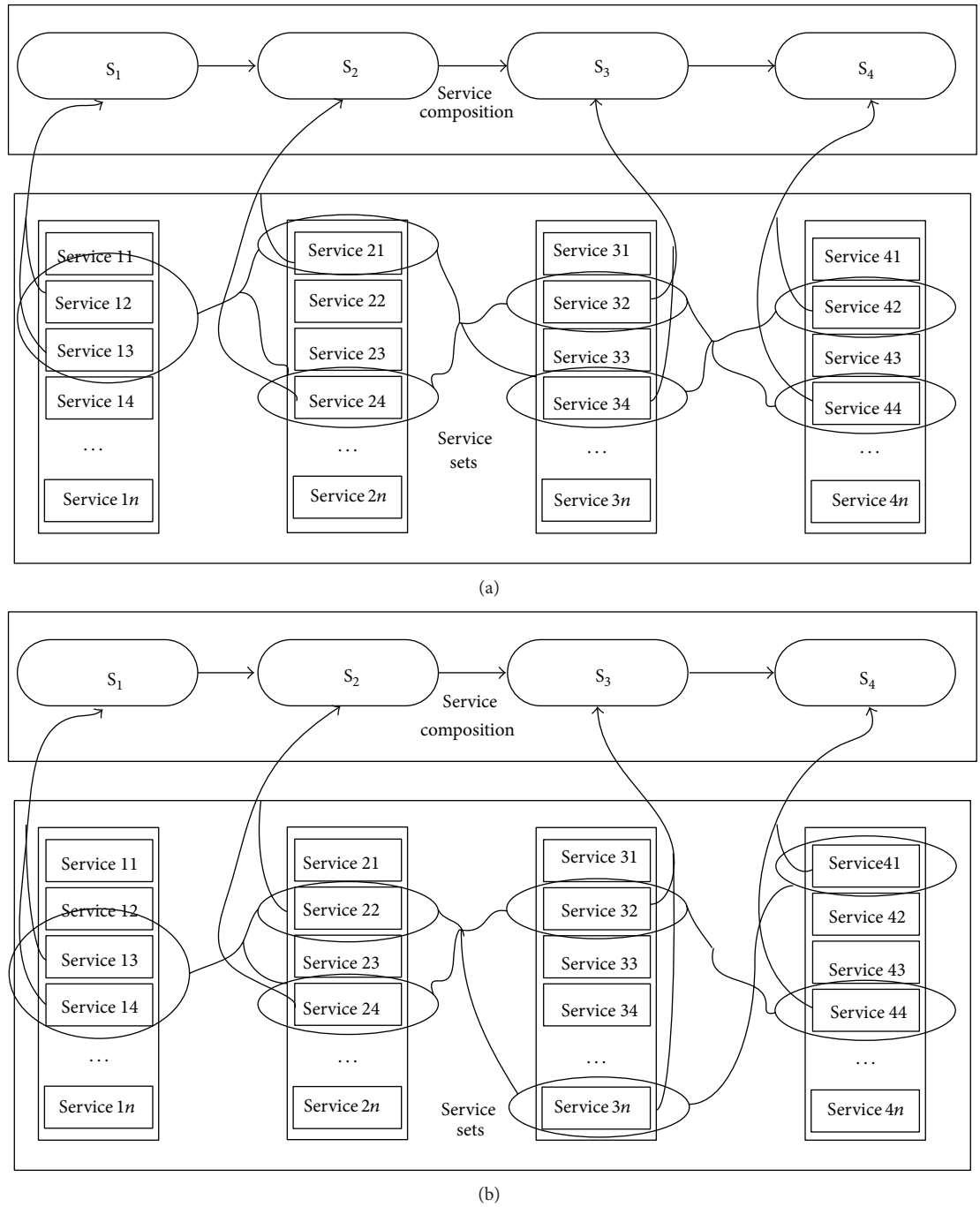


Figure 5: (a) The first path for multiple service parallel allocation. (b) The second path for multiple service parallel allocation.

Suppose n events according to n services in a service composition path. Let $m = \langle m_1, m_2, \ldots, m_n \rangle$, where m_i $(i = 1, 2, \ldots, n)$ represents the number of parallel called services in this node. Suppose simulation service bus could support A_i $(m_i \in A_i)$ service for the ith node, and the system expense in the parallel running m_i services can be shown by the following formula:

$$C_i(m_i) = e^{(m_i/A_i)^2}. \tag{4}$$

Therefore, the whole process of parallel service running expense can be shown as

$$C(m) = \sum_{i=1}^{n} e^{(m_i/A_i)^2}. \tag{5}$$

The ultimate optimization goal is to select proper m which is satisfied with constraint condition for service time and probability, and the service running expense is the lowest. Suppose the weighting for the ith service in the whole service

is W_i, $\sum_{i=1}^{n} W_i = 1$ and the ultimate optimization model can be represented as

$$\begin{aligned} \text{Min} \quad & (c(m)) = \text{Min} \sum_{i=1}^{n} W_i e^{(m_i/A_i)^2} \\ \text{s.t.} \quad & \text{Suc}T_j\left(\text{Suc}T_1\left(d_1, m_1\right), \text{Suc}T_2\left(d_2, m_2\right)\right. \\ & \left. \cdots \text{Suc}T_n\left(d_n, m_n\right)\right) > \text{SCP}_j \\ & 0 < T_j\left(d_1, d_2, \ldots, d_n\right) < \text{LET}_j. \end{aligned} \tag{6}$$

5. Service Scheduling Optimization Based on Quantum Optimization Algorithm

In fact, the problem of multiple services parallel allocation is service resource scheduling problem which is a NP problem. Many methods have been proposed to solve this problem: GA, PSO, and so forth. However, there always exist many disadvantages in partial convergence or slow optimization. Quantum computation is based on the principal concepts of the quantum theory [13, 14]. Numerous researchers have devoted increasing interests to quantum computation, a novel interdisciplinary field that covers quantum mechanics and information science [15–24]. This paper tries to use a quantum optimization algorithm with multiple chain coding schemes to solve this problem and mainly use four-chain quantum-inspired evolutionary algorithm (FCQIEA) to solve.

5.1. Expanded Encoding Method for Quantum Chromosome. From [24] we can know that multiple chains can be required in Figure 6.

From Figure 6, we can get (7) as follows:

$$\left|\varphi\right> = [\cos\theta, \cos\theta \sin\theta, \sin\theta \sin\theta]^T. \tag{7}$$

To describe the quantum dynamics behavior objectively, comprehensively, and unambiguously, we can use a new angle φ ($0 < \varphi < \pi$), which is called "supporting role," to replace θ and obtain vector $\sin\theta$ as follows:

$$\left[\sin\varphi \times \sin\theta, \cos\varphi \times \sin\theta, \cos\theta\right]^T. \tag{8}$$

Equation (8) also satisfies the normalization condition. In fact, (8) also corresponds to the three-chain encoding method:

$$p_i = \begin{vmatrix} \sin\varphi_{i1} \sin\theta_{i1} \\ \cos\varphi_{i1} \sin\theta_{i1} \\ \cos\varphi_{i1} \end{vmatrix} \cdots \begin{vmatrix} \sin\varphi_{\text{in}} \sin\theta_{ij} \\ \cos\varphi_{\text{in}} \sin\theta_{ij} \\ \cos\varphi_{ij} \end{vmatrix}. \tag{9}$$

Likewise, we can obtain vector $\sin\varphi \sin\theta$ by adding the "supporting role" β to form the four-chain encoding method as follows:

$$\left[\cos\beta \times \cos\varphi \times \sin\theta, \sin\beta \sin\varphi \times \sin\theta, \cos\varphi \sin\theta, \cos\theta\right]^T,$$

$$p_i = \begin{vmatrix} \cos\beta_{i1} \sin\varphi_{i1} \sin\theta_{i1} \\ \sin\beta_{i1} \sin\varphi_{i1} \sin\theta_{i1} \\ \cos\varphi_{i1} \sin\theta_{i1} \\ \cos\theta_{i1} \end{vmatrix} \cdots \begin{vmatrix} \cos\beta_{ij} \sin\varphi_{ij} \sin\theta_{ij} \\ \sin\beta_{ij} \sin\varphi_{ij} \sin\theta_{ij} \\ \cos\varphi_{ij} \sin\theta_{ij} \\ \cos\theta_{ij} \end{vmatrix}. \tag{10}$$

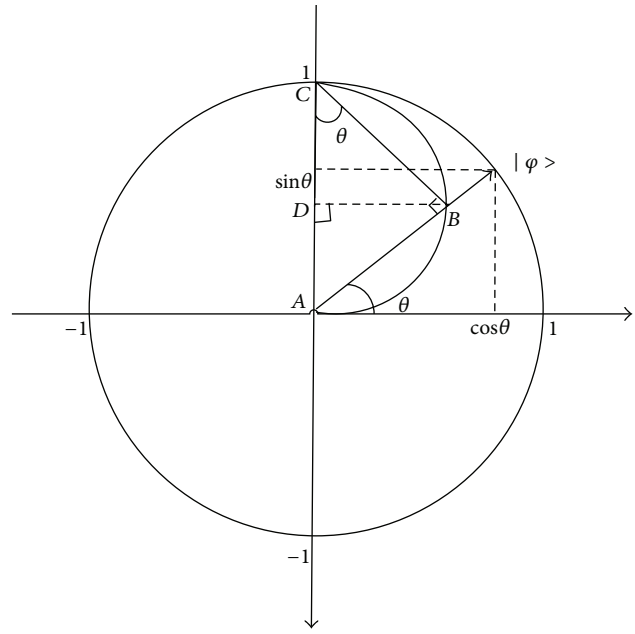


Figure 6: Schematic diagram of the qubit probability amplitude decomposition.

We can obtain four optimal solutions, which are expressed as follows:

$$\begin{aligned} p_{i1} &= \left(\cos\beta_{i1} \sin\varphi_{i1} \sin\theta_{i1}, \ldots, \cos\beta_{\text{in}} \sin\varphi_{\text{in}} \sin\theta_{\text{in}}\right) \\ p_{i2} &= \left(\sin\beta_{i1} \sin\varphi_{i1} \sin\theta_{i1}, \ldots, \sin\beta_{\text{in}} \sin\varphi_{\text{in}} \sin\theta_{\text{in}}\right) \\ p_{i3} &= \left(\cos\varphi_{i1} \sin\theta_{i1}, \ldots, \cos\varphi_{\text{in}} \sin\theta_{\text{in}}\right) \\ p_{i4} &= \left(\cos\theta_{i1}, \ldots, \cos\theta_{\text{in}}\right). \end{aligned} \tag{11}$$

With the same principle, we can get $N + 1$ chains' coding scheme as follows:

$$p_i = \begin{vmatrix} (\sin\theta_n)_{i1} (\sin\theta_{n-1})_{i1} \cdots (\sin\theta_2)_{i1} (\sin\theta_1)_{i1} \\ (\cos\theta_n)_{i1} (\sin\theta_{n-1})_{i1} \cdots (\sin\theta_2)_{i1} (\sin\theta_1)_{i1} \\ \vdots \\ (\cos\theta_3)_{i1} (\sin\theta_2)_{i1} (\sin\theta_1)_{i1} \\ (\cos\theta_2)_{i1} (\sin\theta_1)_{i1} \\ (\cos\theta_1)_{i1} \end{vmatrix} \cdots \begin{vmatrix} (\sin\theta_n)_{ij} (\sin\theta_{n-1})_{ij} \cdots (\sin\theta_2)_{ij} (\sin\theta_1)_{ij} \\ (\cos\theta_n)_{ij} (\sin\theta_{n-1})_{ij} \cdots (\sin\theta_2)_{ij} (\sin\theta_1)_{ij} \\ \vdots \\ (\cos\theta_3)_{ij} (\sin\theta_2)_{ij} (\sin\theta_1)_{ij} \\ (\cos\theta_2)_{ij} (\sin\theta_1)_{ij} \\ (\cos\theta_1)_{ij} \end{vmatrix}. \tag{12}$$

5.2. Solution Space Transformation. In the quantum evolution progress, all qubits have limited values within –1 to 1.

Thus, we need to transform all qubit values from unit space $I^n = [-1,1]^n$ to space Ω of the continuous optimization problem (1) by using linear transformation. Each gene value corresponds to an optimization variable in the solution space. If the jth qubit on chromosome p_i is $[x_{ij}^4, x_{ij}^3, x_{ij}^2, x_{ij}^1]$, then the corresponding variables in the solution space are computed as follows:

$$\begin{aligned} X_{i1}^j &= 0.5 \times \left[b_j\left(1 + x_{ij}\right) + a_j\left(1 - x_{ij}\right)\right] \\ X_{i2}^j &= 0.5 \times \left[b_j\left(1 + x_{ij}\right) + a_j\left(1 - x_{ij}\right)\right] \\ X_{i3}^j &= 0.5 \times \left[b_j\left(1 + x_{ij}\right) + a_j\left(1 - x_{ij}\right)\right] \\ X_{i4}^j &= 0.5 \times \left[b_j\left(1 + x_{ij}\right) + a_j\left(1 - x_{ij}\right)\right], \end{aligned} \tag{13}$$

where $i = 1, 2, 3, \ldots, m$ and $j = 1, 2, 3, \ldots, n$. Thus, each chromosome maps to four approximate solutions of the optimization problem.

5.3. Quantum Chromosome Update. Considering that m quantum chromosomes are present in the colony and we can obtain $4m$ approximate solutions by solution space transformation, we can then compute the fitness of these approximate solutions and define the solution with the maximum fitness as the current optimum solution in the quantum evolution progress. The chromosome corresponds to the current optimum solution called the optimum chromosome. By computing the fitness, we can obtain both optimum solution and optimum chromosome and subsequently update the colony by using the quantum rotation gate to obtain the optimal solution. In this updated process, the new optimum chromosome can be produced such that the colony can likely evolve. The present study proposes the quantum rotation gate U to update the individual qubit as follows:

$$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ u_{21} & u_{22} & u_{23} & u_{24} \\ u_{31} & u_{32} & u_{33} & u_{34} \\ u_{41} & u_{42} & u_{43} & u_{44} \end{bmatrix},$$

$$\begin{aligned} u_{11} &= \left[\cos\Delta\beta\cos\Delta\varphi\left(\cos\Delta\theta - \frac{\sin\Delta\theta\cos\theta}{\sin\theta}\right)\right]^T, \\ u_{12} &= \left[\sin\Delta\beta\cos\Delta\varphi\left(\frac{\sin\Delta\theta\cos\theta}{\sin\theta} - \cos\Delta\theta\right)\right]^T, \\ u_{13} &= \left[\sin\Delta\varphi\cos\Delta\theta\left(\sin\beta\sin\Delta\beta - \cos\beta\cos\Delta\beta\right)\right]^T, \\ u_{14} &= \left[\cos\varphi\sin\Delta\varphi\left(\cos\beta\cos\Delta\beta - \sin\beta\sin\Delta\beta\right)\right]^T, \\ u_{21} &= \left[-\sin\Delta\beta\sin\Delta\varphi\left(\cos\Delta\theta + \frac{\sin\Delta\theta\cos\theta}{\sin\theta}\right)\right]^T, \\ u_{22} &= \left[-\cos\Delta\beta\sin\Delta\varphi\left(\cos\Delta\theta + \frac{\sin\Delta\theta\cos\theta}{\sin\theta}\right)\right]^T, \\ u_{23} &= \left[\cos\Delta\varphi\cos\Delta\varphi\left(\sin\beta\cos\Delta\beta + \cos\beta\sin\Delta\beta\right)\right]^T, \\ u_{24} &= \left[\cos\varphi\cos\Delta\varphi\cos\Delta\varphi\left(\sin\beta\cos\Delta\beta + \cos\beta\sin\Delta\beta\right)\right]^T, \\ u_{31} &= -\frac{\sin\Delta\varphi}{\cos\beta}, \qquad u_{32} = 0, \qquad u_{33} = \cos\Delta\varphi\cos\Delta\theta, \\ u_{34} &= \left[\sin\Delta\theta\left(\cos\varphi\cos\Delta\varphi - \sin\varphi\sin\Delta\varphi\right)\right]^T, \\ u_{41} &= u_{42} = 0, \qquad u_{43} = \frac{\sin\Delta\theta}{\cos\varphi}, \qquad u_{44} = \cos\Delta\theta, \end{aligned}$$

$$\begin{aligned} U &\begin{bmatrix} \cos\beta\sin\varphi\sin\theta \\ \sin\beta\sin\varphi\sin\theta \\ \cos\varphi\sin\theta \\ \cos\theta \end{bmatrix} \\ &= \begin{bmatrix} \cos\left(\beta + \Delta\beta\right)\sin\left(\varphi + \Delta\varphi\right)\sin\left(\theta + \Delta\theta\right) \\ \sin\left(\beta + \Delta\beta\right)\sin\left(\varphi + \Delta\varphi\right)\sin\left(\theta + \Delta\theta\right) \\ \cos\left(\varphi + \Delta\varphi\right)\sin\left(\theta + \Delta\theta\right) \\ \cos\left(\theta + \Delta\theta\right) \end{bmatrix}. \end{aligned} \tag{14}$$

5.4. Mutation Operation. Quantum nongate is applied to exchange the probability amplitudes to avoid local optimal solution in a certain qubit as follows:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} = \begin{bmatrix} \sin\theta \\ \cos\theta \end{bmatrix}. \tag{15}$$

Such influence as expressed in (15) can be considered as the phase mutation of a qubit, in which θ is mutated to $(\pi/2) - \theta$. In this case, a quantum nongate V is proposed to mutate the quantum as follows:

$$V = \begin{bmatrix} \tan\beta\cot\varphi\cot\theta & 0 & 0 & 0 \\ 0 & \tan\beta\cot\varphi\cot\theta & 0 & 0 \\ 0 & 0 & \tan\varphi\cot\theta & 0 \\ 0 & 0 & 0 & \tan\theta \end{bmatrix}. \tag{16}$$

5.5. The Procedure of FCQIEA. It can be summarized as follows.

Step 1 (initialize the population). Let the current generation $t = 0$; generate an initial population $Q(t) = \{q_1^t, q_2^t, \ldots, q_m^t\}$, which has m individual qubits. Set the magnitude of the rotational angle $|\Delta\beta| = \beta_0$, $|\Delta\varphi| = \varphi_0$, and $|\Delta\theta| = \theta_0$, respectively. Set p_m as the mutation and Max_gen as the maximum generation.

Step 2 (transform the solution space). Four approximate solutions in each chromosome are transformed from the unit space $I^n = [-1,1]^n$ to the solution space Ω of the continuous optimization problem (1); thus, the set of approximate solution $X(t)$ can be obtained.

Step 3 (compute the fitness). By computing the fitness of $4m$ approximate solutions, obtain the best solution *BestX* in the current solution and the best chromosome *BestC* in the current chromosome. Store *BestX* as the global optimum solution *GX* and store *BestC* as the global optimum chromosome *GC*.

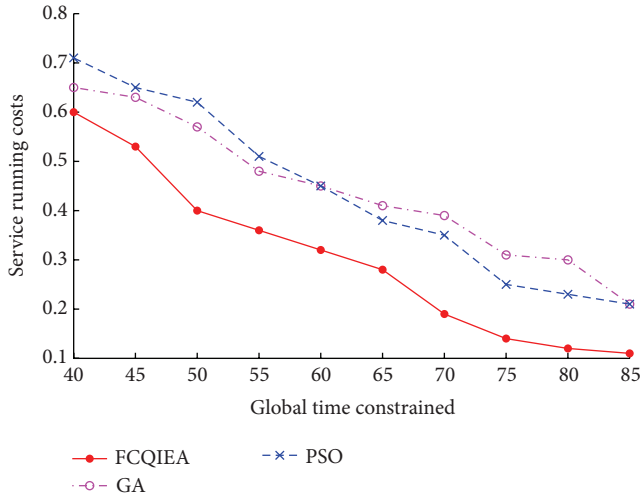


Figure 7: Comparison of the optimization result with different algorithms.

Step 4 (set $t = t + 1$). Update and mutate $Q(t-1)$. Calculate the new population $Q(t)$.

Step 5. Transform the solution space again and obtain a set of approximate solution $X(t)$.

Step 6. By computing the fitness of $Q(t)$, determine the current optimum solution $BestX$ and the current optimum chromosome $BestC$. If fit($BestX$) < fit(GX), then update the current optimum solution $BestX = GX$; at the same time, update the current optimum chromosome $BestC = GC$ to avoid population degradation. Otherwise, let $GX = BestX$ and $GC = BestC$ so that the algorithm approaches the global optimum solution.

Step 7. If the algorithm does not converge and if $t <$ Max_gen, then go back to Step 4 until the algorithm becomes convergent or until $t >$ Max_gen.

6. Simulation Experiment and Analysis

6.1. Comparisons of the Different Optimization Result. Under the hypothesis condition as Figure 2, four events are in accordance with service as S_1, S_2, S_3, and S_4, and the specific constraint conditions can be shown as in Table 1.

In order to compare the different optimization results between different optimization algorithms, FCQIEA, GA, and PSO, optimize the service scheduling, respectively. In both of FCQIEA and GA, the generation set is 100 and intersection probability and mutation set are 0.6 and 0.1, respectively. The number of coding chains is 4, and maximum rotate angle set $\Delta\theta = 0.04\pi$. In PSO, regulatory factor set $w = 0.5$ and study factor set $C_1 = 0.7$ and $C_2 = 1.2$. In order to avoid the influence of random factors, we performed the calculation 10 times in each algorithm and obtained the average value as the optimization result. The global constraint time of each service node is added with 5 seconds as base unit; the optimization result can be shown in Figure 7.

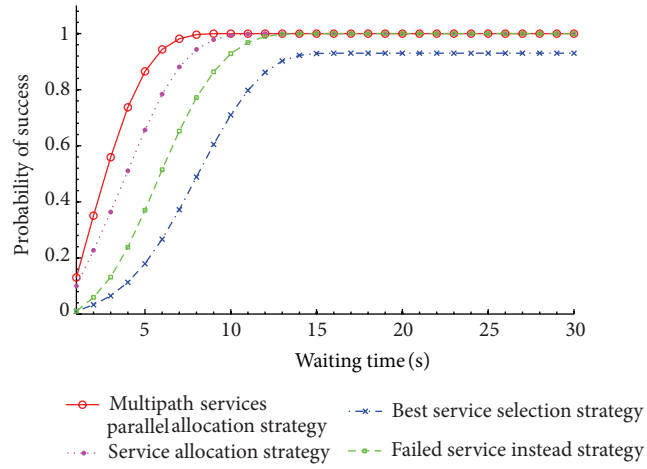


Figure 8: The probability of success of different service allocation strategy.

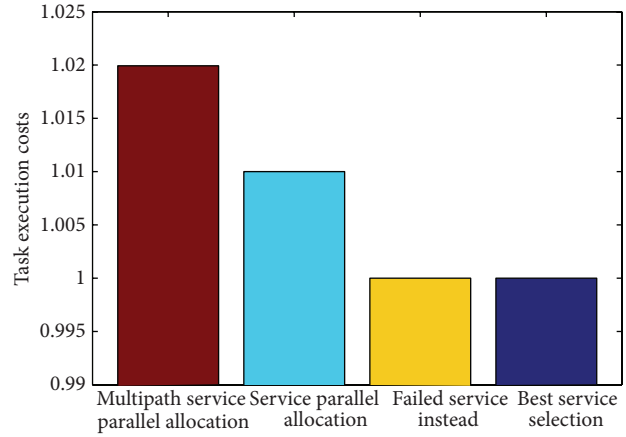


Figure 9: Comparison of task execution costs with different service allocation strategy.

From Figure 7, obviously, FCQIEA has best performance compared with GA and PSO and its service running costs most rapid decline than other algorithms.

6.2. Comparisons of the Service Allocation Strategy. The different service allocation strategy, that is, selection best service, failed service instead, services parallel allocation, and multipath services parallel allocation, is compared under the constraint condition in Table 1. The failed service is produced randomly with probability of 1%. The failed service instead strategy is that the service can be called failure when waiting time is beyond the 3 times of δ (3 times δ rule) service execution time and it should select another service to replace the failed service. In service parallel allocation strategy, 20 percent of service sets are randomly selected as parallel execution service. In the multipath service parallel allocation strategy, the number of path set 3 and 20 percent of service sets are randomly selected as parallel execution service. After long running time, the relation between the probability of success and execution of different service allocation strategy is investigated. It can be shown in Figure 8.

Table 1: Service constrained condition.

Constrained condition	Service node			
	S_1	S_2	S_3	S_4
Execution time distribution	[5 s, 15 s] uniform distribution	[10 s, 20 s] uniform distribution	[10 s, 25 s] uniform distribution	[2 s, 8 s] uniform distribution
Probability of success	A random number from 0.94 to 1	A random number from 0.94 to 1	A random number from 0.92 to 1	A random number from 0.93 to 1
The number of parallel services	30	20	30	20
The maximum waiting time	10	30 40		5
The minimum probability of success	98%	98% 98%		98%
W_i	0.2	0.3	0.1	0.4

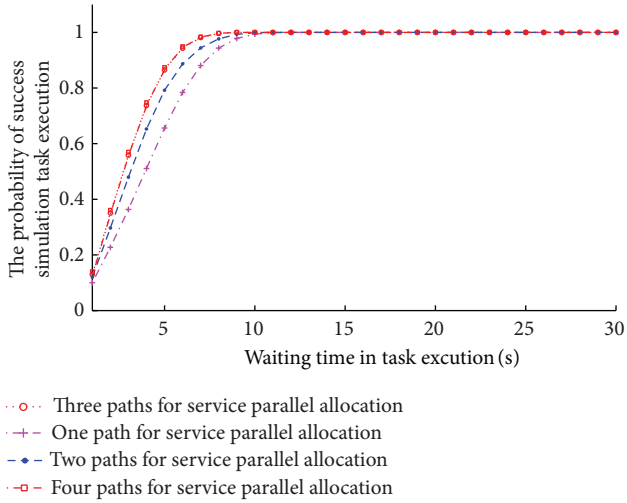


Figure 10: Comparison of the effect for the probability of success simulation task with different paths.

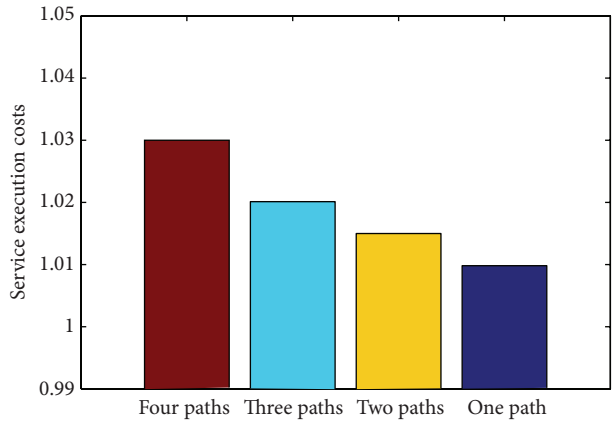


Figure 11: Comparison of simulation task execution costs among different paths.

Different service execution costs can be shown in Figure 9.

From analysis, it can be found that the probability of success of multipath service parallel allocation strategy is obviously higher than other strategies; however, it also results in extra system costs than other service allocation strategies.

6.3. The Effect of Probability of Success from the Number of Paths. Service parallel allocation and multipath services parallel allocation are used to execute simulation task and the number of paths is 1, 2, 3, and 4. The failed service is produced randomly with probability of 1%. After running in long time, the relation between the probability of success and execution time is investigated. It can be shown in Figures 10 and 11.

From Figures 10 and 11, we can get the conclusion that we can improve probability of success by adding the number of service paths in service parallel allocation, but it will not always get the best probability of success when increasing the number of service paths, and there exist the best value. From the simulation experiment, it can be seen that the probability of success for completion task is improved obviously with the numbers of 2 and 3, but when the number of path set is 4, the probability of success for completion task is not improved and the simulation execution costs are raised. Therefore, three paths are proper.

7. Conclusions

In military modeling and simulation field, it is important to improve the probability of success and timeliness in simulation task workflow. This paper established the multipath service parallel allocation optimization mathematical model to investigate service scheduling optimization algorithm in which FCQIEA is used. Through the simulation experiment, this paper investigates the effect for the probability of success in simulation task workflow from different optimization algorithm, service allocation strategy, and path number, and the simulation result shows that the optimization algorithm for multipath service resource parallel allocation is an effective

method to improve the probability of success and timeliness in simulation task workflow. In the next step, multipath service parallel allocation will be used in military modeling and simulation field, and this theory can be improved in practice.

Conflict of Interests

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

References

[1] D. König, N. Lohmann, S. Moser, C. Stahl, and K. Wolf, "Extending the compatibility notion for abstract WS-BPEL processes," in *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*, pp. 785–794, Beijing, China, April 2008.

[2] N. Salatge and J.-C. Fabre, "Fault tolerance connectors for unreliable Web Services," in *Proceedings of the 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN '07)*, pp. 51–60, June 2007.

[3] J. Harney and P. Doshi, "Speeding up adaptation of web service compositions using expiration times," in *Proceedings of the 16th International World Wide Web Conference (WWW '07)*, pp. 1023–1032, May 2007.

[4] E. Sirin, *Combining Description Logic Reasoning with AI Planning for Composition of Web Services*, University of Maryland, College Park, Md, USA, 2006.

[5] J. Rao and X. Su, "A survey of automated web service composition methods," in *Proceedings of the 1st International Workshop on Semantic Web Services and Web Process Composition (SWSWPC '05)*, vol. 3387, pp. 43–54, July 2005.

[6] S. Mcilraith and T. C. Son, "Adapting golog for composition of semantic web services," in *Proceedings of the 8th International Conference on Knowledge Representation and Reasoning*, pp. 482–496, Morgan Kaufmann, Toulouse, France, 2002.

[7] S. Narayanan and S. A. McIlraith, "Simulation, verification and automated composition of web services," in *Proceedings of the 11th International Conference on World Wide Web (WWW '02)*, pp. 77–88, ACM, New York, NY, USA, May 2002.

[8] S. Yajuan, *Research on the Approaches for Web Services Composition*, Jilin University, Changchun, China, 2011.

[9] X. Fan, C. Jiang, X. Fang, and Z. Ding, "Dynamic web service selection based on discrete particle swarm optimization," *Journal of Computer Research and Development*, vol. 47, no. 1, pp. 147–156, 2010.

[10] H. Zheng, W. Zhao, J. Yang, and A. Bouguettaya, "QoS analysis for Web service composition," in *Proceedings of the IEEE International Conference on Service Computing (SCC '09)*, pp. 235–242, Bangalore, India, September 2009.

[11] L. Zeng, B. Benatallah, A. H. H. Ngu, M. Dumas, J. Kalagnanam, and H. Chang, "QoS-aware middleware for Web services composition," *IEEE Transactions on Software Engineering*, vol. 30, no. 5, pp. 311–327, 2004.

[12] Y. Xiaohao, H. Dan, L. Xueshan, and L. Junxian, "Parallel optimization method of service resource in military information system," *Systems Engineering Theory & Practice*, vol. 32, no. 9, pp. 2078–2086, 2009.

[13] D. Deutsch, "Quantum theory, the Church-Turing principle and the universal quantum computer," *Proceedings of The Royal Society of London A: Mathematical and Physical Sciences*, vol. 400, no. 1818, pp. 97–117, 1985.

[14] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, 1st edition, 2000.

[15] S.-Y. Li and P.-C. Li, "Quantum genetic algorithm based on real encoding and gradient information of object function," *Journal of Harbin Institute of Technology*, vol. 38, no. 8, pp. 1216–1223, 2006.

[16] P. C. Li and S. Y. Li, "Quantum-inspired evolutionary algorithm for continous spaces optimization based on Bloch coordinates of qubits," *Neurocomputing*, vol. 72, pp. 581–591, 2008.

[17] X. Shaohua, X. Chen, H. Xing, W. Ying, and L. Panchi, "Improved quantum genetic algorithm with double chains and its application," *Application Research of Compute*, vol. 26, no. 7, pp. 2090–2092, 2010.

[18] S. Li and P. Li, *Quantum Computing and Quantum Optimization Computing Algorithm*, Harbin Polytechnical University, 2009.

[19] K.-H. Han and J.-H. Kim, "Quantum-inspired evolutionary algorithm for a class of combinatorial optimization," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 6, pp. 580–593, 2002.

[20] A. D. S. Nicolau, R. Schirru, and A. Alvarenga de Moura Meneses, "Quantum evolutionary algorithm applied to transient identification of a nuclear power plant," *Progress in Nuclear Energy*, vol. 53, no. 1, pp. 86–91, 2011.

[21] J. X. V. Neto, D. L. de Andrade Bernert, and L. dos Santos Coelho, "Improved quantum-inspired evolutionary algorithm with diversity information applied to economic dispatch problem with prohibited operating zones," *Energy Conversion and Management*, vol. 52, no. 1, pp. 8–14, 2011.

[22] J. Xiao, J. Xu, Z. Chen, K. Zhang, and L. Pan, "A hybrid quantum chaotic swarm evolutionary algorithm for DNA encoding," *Computers and Mathematics with Applications*, vol. 57, no. 11-12, pp. 1949–1958, 2009.

[23] J. Gu, M. Gu, C. Cao, and X. Gu, "A novel competitive co-evolutionary quantum genetic algorithm for stochastic job shop scheduling problem," *Computers and Operations Research*, vol. 37, no. 5, pp. 927–937, 2010.

[24] W. Zhiteng, Z. Hongjun, Z. Rui, X. Ying, and H. Jian, "Quantum genetic algorithm based on multi-chain coding scheme," *Computers & Operations Research*, vol. 29, no. 26, pp. 2061–2064, 2012.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 267403, 7 pages
http://dx.doi.org/10.1155/2014/267403

Research Article

Visual Quality Assessment after Network Transmission Incorporating NS2 and Evalvid

Zhengyou Wang,[1,2] Wan Wang,[1] Yanhui Xia,[3] Zheng Wan,[2] Jin Wang,[1] Liying Li,[1] and Cong Cai[1]

[1] *Shijiazhuang Tiedao University, School of Information Science and Technology, Shijiazhuang 050043, China*
[2] *Jiangxi University of Finance & Economics, School of Information Technology, Nanchang 330032, China*
[3] *Shijiazhuang Tiedao University, School of Economics and Management, Shijiazhuang 050043, China*

Correspondence should be addressed to Zhengyou Wang; zhengyouwang163@163.com

Received 9 October 2013; Accepted 4 December 2013; Published 15 April 2014

Academic Editors: W. Sun, G. Zhang, and J. Zhou

On the basis of Evalvid tool integrated in NS2 (Network Simulator version 2), the paper gets new set of tools, myEvalvid, to establish the simulation and evaluation platform for multimedia transmission. Then the paper investigates the effects of various influence factors when multimedia information is transmitted in the network and the relationships among these factors. Based on the analysis, the paper gets different evaluation models, respectively. In this paper, we study the impact on performance of several basic source and network parameters of video streams, namely, GOP (Group of Pictures) pattern, compression quantitative parameters, packet length, and packet error rate. Simulation results show that different parameters lead to different distortion levels which are calculated according to the reconstruction images at the receiver and the original images. The experimental results show that the video transmission and quality evaluation model we designed can evaluate multimedia transmission performance over complex environment very well.

1. Introduction

Image quality evaluation is one of the most important steps in the field of image processing. In all aspects of image processing, such as image compression, image transmission, and image deblur, image quality evaluation plays a very important role. Image quality evaluation methods can be divided into subjective evaluation method and objective evaluation method according to the subject of evaluation. The researchers often use mean opinion score method and the peak signal-to-noise ratio method. Over the past years, a lot of researches have been conducted towards the construction of objective video quality metrics. Applying NS2 (Network Simulator version 2) and Evalvid, video streams could be transmitted through the simulation network and the corresponding peak signal-to-noise ratio (PNSR) could be calculated. And the image transmission quality in different network environments, the influence factors on the quality of the image transmission can be analized of image transmission quality in different network environments and the influence factors on the quality of the image transmission.

Recently, more and more telecommunication systems support different types of real-time transmission, and video transmission is one of the most important applications. The growing needs lead to the support of the video quality evaluation. Although many papers have been committed to study the QoS (Quality of Service) mechanism which is supported through different types networks, many researchers are limited to the ranges of using packet loss rate, packet delay, and packet jitter as video transmission quality metrics. As known to us all, the above metrics cannot be easily transformed into video transmission quality. In fact, this relation may be due to coding scheme, hide loss scheme, and delay or jitter processing which are not the same leading to differences. To make detailed descriptions to these related parameters, we use NS2 and myEvalvid tools as simulation platform to obtain

the transmission results. Obtained correlations among these parameters are of great significance to the evaluation of video quality.

The paper creates the platform to simulate and evaluate the transmission image in the multimedia network. Various factors that affect the quality of video transmission are probed and the correlations among these factors are achieved. As for the issue of the simulation of video stream, the authors apply the method of introducing the traffic trace file of video stream to the network simulation circumstance, aiming to simulate the actual transmission of video stream in the network. Thus, the author can get the record file sd (send data) of sender and the record file rd (receive data) of receiver and use them to make an effect assessment. Different effects caused by various factors during video transmission can be also evaluated. Evaluation model could be established on the basis of these experimental findings. The results reveal that various value settings of GOP (Group of Pictures) pattern, quantization value, packet size, and packet error rate lead to different distortions between the original video and the reconstructed video.

The remainder of this paper is organized as follows. Section 2 discusses related work of image quality assessment. The experimental environment is described in detail in Section 3 and experimental PSNR values are obtained in Section 4. In Section 5 we build linear regression model of based network parameters and analyze and explain experimental results. And a general conclusion is presented in Section 6.

2. Related Work

Reference [1] used ANOVA (Analysis of Variance) to analyze the delay, jitter, packet loss, and their relationships in the video quality assessment and safely draw a conclusion that jitter and packet loss have significant influences on video quality. Furthermore, the influence degree of the jitter and packet loss is equal to the video quality. Reference [2] puts forward all the multimedia signal processing algorithm need appropriate Fidelity index verification results. It makes use of the human visual system (HVS)-image processing based on objective fidelity metrics. And this method is described for a set of special image which is called region of interest (ROI) of the measurement problem. Moreover, this paper has not built video quality model. ITU-TG-1070 (International Telecommunication Union Telecommunication-G) [2] provides video quality evaluation model which takes packet loss as variable for video phone. Reference [3] studied the video quality evaluation under the situation of burst packet loss based on [2]. The improved model (T-Model) draws into the complement factor B and comes to describe the burst degree of continuous packet loss. References [4, 5] aimed at using MPEG-2 (Moving Picture Experts Group) coding to research the relationship between the video quality evaluation and packet loss and the relationship between the video quality evaluation and coding rate. It takes the method of MPQM (Moving Pictures Quality Metric). Grading the video quality finally gets the grading value Q of the video quality. These video quality evaluation models only considered the influences of packet loss, while they do not consider the influences of other factors such as QoS factors (delay and jitter). In [6] the author presented a metric based multichannel model of human spatiotemporal vision that has been parameterized for video coding applications. Evaluation method of video evaluation QoE (Quality of Experience) can be divided into subjective evaluation method and objective evaluation method. Subjective evaluation methods are Mean Opinion Score (MOS), Distortion Mean Opinion Score (DMOS) [7], Single stimulate continuous Quality Score (SSCQS) [8], Double Stimulate Continuous Quality Score (DSCQS) [8], and so on. Subjective evaluation method's advantages are accurate and easy to understand, but it is hard to implement because steps are complex and it is hard to implement in the laboratory environment. Objective evaluation methods often use PSNR, MDI (Media Delivery Index) [9], NTIA (National Telecommunications and Information Administration) [10] model, and so on. In PSNR method, PSNR values are obtained through comparing source files and those after decoding pixel-by-pixel, the results can not accurately reflect the quality of the user's experience [11]. MDI consists of two parts: delay parameters and media loss rate. This evaluation method's premise is "if the transmission quality is good, then the video quality is good." MDI directly uses network parameters to show video quality because MDI provides no quantitative score evaluating network video quality. Thus the video quality researches rarely use this method. ITU-TJ.144 (International Telecommunication Union Telecommunication-J) compares and analyzes the different objective evaluation and subjective evaluation models fitting degree in [12], and the results show that comprehensive performance of NTIA model is the optimal in all chosen models (the Pearson correlation coefficient is 0.938; the mean square error is only 0.074). Furthermore, this study also used this model to evaluate the video after transmission. This model's algorithm has already been implemented in the video quality evaluation software BVQM (Batch Video Quality Metric) which was developed by telecommunications academy of sciences ITS (Institute for Telecommunication Sciences). Reference [13] proposed a content-adaptive packet-layer (CAPL) model for networked video quality assessment, by evaluating the distortions induced by both compression and packet loss. This paper took the regression model to get the mapping relationship between the QoS index and QoE. Reference [14] proposed a kind of network QoS index affecting Experience Quality (QoE) estimation based on WiMAX (Worldwide Interoperability for Microwave Access). This method is more accurate than the previous. Reference [15] illustrated the influences of the packet loss on the quality of video, and reference [13] put forward a kind of content adaptive model, considering video complexity in time domain and airspace views. Based on the above references, we can extract network parameters from kinds of aspects for our research.

3. Experiment Environment

Since in the real network environment the parameters of the network cannot be set to realize controllability and repeatability of the network damage, the paper presents the design of the simulation platform to simulate the process

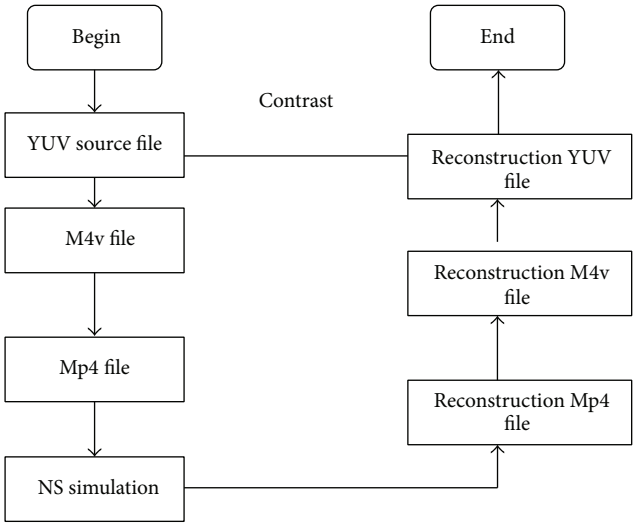


Figure 1: Image transmission quality evaluation process.

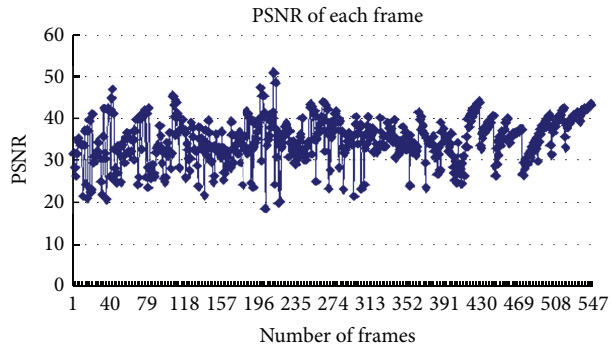


Figure 2: Each frame PSNR value of akiyo video stream.

when the users watch video. The network transmission module is used to set up different QoS parameters and control network damage. Video quality evaluation module is used to compare the source video files and target video files to get user experience quality.

In order to improve Evalvid and to enhance the function of simulation, we can make Evalvid through the myEvalvid, my_UDP, and myEvalvid_Sink, which are three interface programs (or we can say agent) communicated with NS2. The integration is myEvalvid [20]. Below we aim at giving a simple introduction to myEvalvid, my_UDP, and myEvalvid_Sink.

(1) myEvalvid: the interface program's main job is to learn to read VS (Video Sender) procedures from the film log files. The log files of each picture are cut into smaller segments, and the user in Tcl (Tool Command Language) Script can set in good time to the section which is the bottom of the UDP layer sending out.

(2) my_UDP: basically my_UDP agent is the extension of UDP agent. The new agent has the packet transmission time, packet identification, and packet load size recorded in the files which are set up by Tcl Script. Simply say, generally My_UDP's work is as Tcp-dump or Win-dump's work.

(3) myEvalvid_Sink: it is responsible for receiving packets which are issued by my_UDP. This agent records the receive time, packet identification, and packet load size in the files which are specified in Tcl Script.

4. Image Transmission Effect Analysis and Evaluation

4.1. Use MyEvalvid for PSNR Evaluation. Using myEvalvid to evaluate image transmission quality, Figure 1 shows the process of implementation. We firstly use the source YUV file to generate NS2 simulation file. After the NS2 network simulation, we reconstruct YUV file at the receiving end. By comparison and analysis, we can calculate the value of PSNR and evaluate the quality of video transmission.

This paper uses akiyo_cif as an example and calculates PSNR value of each frame. A plot curve of results is shown in Figure 2. Through the network transmission, the image's average PSNR value is 36.92.

We can apply YUV viewer software to observe video frame distortion situation.

When using myEvalvid to verify the multimedia network structure, we can calculate the PSNR of reconstructed video, which can be used to observe the difference between the received movie and the original movie. Taking 151st frame as example, the received PSNR value is 27.24. Compared with the 51st frame, the received PSNR value is 45.48, as shown in Figure 3.

As shown from Figure 3, the image on the left side is better than the image on the right side, owing to simulation transmission process in the above network. During packet transmission process, there will be some packet loss. Thus it will lead to some pictures becoming in unsolvable state. Therefore after arriving at the receiving end, the reconstructed image quality is worse than the original image.

In network simulation transmission process, there are packet losses in the network. Since packet loss probabilities of different pictures are distinct, some pictures are in unsolvable state and the others are in solvable state. Therefore, distortion level of individual frame of the video stream will vary. That is to say, packet loss rate has a great influence on the quality of image transmission.

5. Parameters' Influence of Image Transmission

5.1. Influence Factors of Image Transmission. For different multimedia streams, the quality of the transmission is distinct because of compression parameters, network parameters, and network states. In the network transmission, possible multimedia factors generally include the following: GOP pattern, quantization value, packet size, and packet error rate.

5.2. Parameters' Influence on the Quality of Image Transmission. Through the above analysis, we can change the settings of the parameters and conduct simulation experiment to

(a) Contrast figure of 151st frame image

(b) Contrast figure of 51st frame image

Figure 3: Contrast figure of original image and reconstruction image.

Table 1: The average PSNR value of simulation results.

Compression quantitative parameters (Q)	Average PSNR
31	26.83
20	28.82
10	32.28

Table 2: The simulation results of the average PSNR value.

MTU	Average PSNR
1000	31.06
800	29.98
600	28.58
400	27.82

obtain data. From the data analysis, we can discuss the influences of various kinds of factors on the quality of image transmission.

(1) The influences of quantization parameters on the quality of image transmission; quantization is a process that attempts to determine which information can be safely discarded without a significant loss in visual fidelity.

When the quantization parameters are set to 31, 20, and 10, the PSNR values of simulation results are shown in Table 1.

From Table 1 we can know that when performing compression, the larger the value of the quantization parameter is, the worse the image quality is. This is because, in the image compression process, when the setting of quantitative parameters is bigger, the image quality after compression will be worse. Therefore, the image quality at the receiving end after the network transmission is worse than that with a small quantitative parameters setting. When using smaller quantitative parameters, it will have a good image effect. At the same time, we can also find that the more quantization parameters the data compressed, the more necessary it is to transmit packets.

(2) The influences of packet length in the quality of image transmission: to explore the influences of packet length in the quality of image transmission, we can change the packet size and MTU (Maximum Transmission Unit) values and keep other parameters constant. After conducting simulation, we can get different packet sizes and their correspondent average PSNR values. The processes of the experiments are complied with compression quantitative parameters simulation in the quality of the same image transmission. The results are shown in Table 2.

From Table 2 we can find that, when the packet length is longer, the image quality will be higher. The reason of this phenomenon is, for the same image, if the packet length is long, on behalf of each picture it is required to split the packet number less. Since packet error rate is the same, compared to shorter packet length, the lost packet number will be less. In this case more pictures could be decoded, leading to higher image quality. When using shorter packet length, on behalf of each picture it is required to split the number of packets more, so the lost packet numbers will also be more. In this

Table 3: The simulation results of the average PSNR value.

GOP	Average PSNR
9	32.28
15	32.09

case, relatively few pictures can be decoded, resulting in poor image quality.

(3) The influences of GOP pattern in the quality of image transmission: to explore the influences of GOP pattern in the quality of image transmission, we can set parameters of the GOP and calculate PSNR value to get the connection between the GOP and the quality of image transmission. The simulation results are shown in Table 3.

We can find from Table 3 that the shorter the image GOP length is, the better its quality is. This is because I frame packet loss will lead to pictures of the same GOP becoming noncoding. So in the image with longer GOP length, I frame loss has to wait for a long time to wait the next I frame arrival. As for the image with shorter GOP length, it will be waiting for the next I frame whose coming time is short. So the recovery time is shorter, resulting in relatively good image quality.

5.3. Network Video Mapping Model of QoE and QoS

5.3.1. Establishing Evaluation Model. The two examples cited above are akiyo_cif and foreman_qcif because the two videos have different characteristics, considering the node, video dynamic, action, and complexity. Thus we make observation and comparison on different behaviors of each video and get the final results. It can be divided into the following steps to complete the evaluation model.

First of all, conduction video is streaming under the reservation condition, with Evalvid frame integrated into NS2 module. Each of the two selected videos is simulated fifty times. All of the schemes, namely, 1, 2, 3, and 4 nodes, include a total of 400 times transmission.

In the second step, we perform the statistical analysis. The goal is to get a more reliable and consistent database and eliminate the problems which may occur. Therefore, those who are away from the standard deviation value will be eliminated. one of these records contains the following QoS parameters: delay, jitter, and packet loss. This loss is calculated into total loss of a video. The loss refers to each frame's MPEG codec (I, P, and B).

In the third step, the original and decoded video file sequence are used to obtain QoS and QoE index. Using myEvalvid can obtain QoS index (jitter, delay, throughput, and packet loss), MOS, and PSNR but cannot obtain more stable index, such as structural similarity and VQM (Video quality Metric). However, we can combine packets which are transmitted through myEvalvid to get QoS index.

The last step is to multiple nonlinear regression equation to establish evaluation model and to evaluate its performance in the training sets and testing sets.

Finally, through fitting, we get the superiority of the evaluation model.

5.3.2. Packet Loss and Jitter Influence on the Video Quality. It can be seen from the front of the research that delay has nothing to do with video's user quality of experience, while the jitter and packet loss have great effects on the video's user quality of experience. Thus the research of network effect of multiple factors on video transmission draws the conclusion that it is mainly the combined effect of jitter and packet loss on the video of the user quality of experience. On the other hand, in the actual network conditions jitter and packet loss exist at the same time and have constant changes. So it is of more practical significance to research the jitter and packet loss's comprehensive influence on video's user quality of experience. It can be concluded that the influence of packet loss and jitter on the video quality is nonlinear relationship, so this paper adopts multiple nonlinear regression equation to establish nonparametric model. The regression equations are as follows:

$$\mathrm{MQ} = \alpha_1 * \mathrm{DP} + \alpha_2 * \mathrm{JT} + \mu_1, \quad (1)$$

$$\mathrm{VQ} = \beta_1 * \mathrm{DP} + \beta_2 * \mathrm{JT} + \mu_2, \quad (2)$$

where MQ is mean quality, the average quality index of video sequences; VQ is varquality, the quality of video sequence variance; DP is droprate, packet loss rate; JT is jitter, Jitter.

Considering DP and JT separately, we can calculate their influence of MQ and VQ respectively.

(a) The influence of DP on MQ: regression model is shown in

$$\widehat{\mathrm{MQ}} = 11.06955 - 0.0058637 * \mathrm{DP}. \quad (3)$$

(b) The influence of DP on VQ: regression model is shown in

$$\widehat{\mathrm{VQ}} == 18.15667 - 0.0356798 * \mathrm{DP}. \quad (4)$$

From Tables 4 and 5 we can find that packet loss rate DP for video quality variance VQ has more explanatory power relative to the average video quality MQ.

(c) The influence of JT on MQ: regression model is shown in

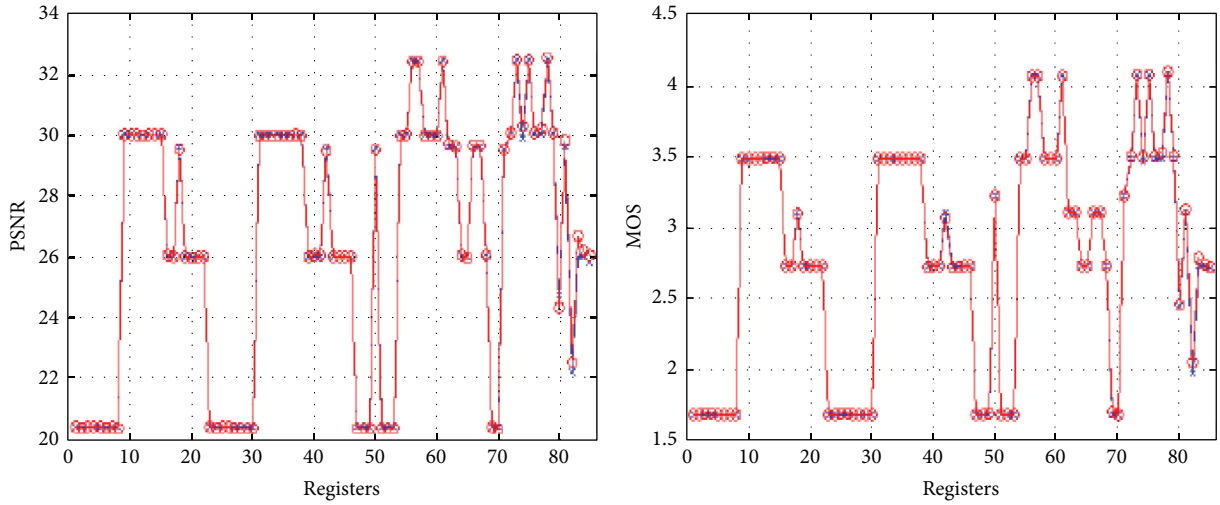
$$\widehat{\mathrm{MQ}} = 10.68654 - 4.67 \exp t\,(-6) * \mathrm{JT}. \quad (5)$$

(d) The influence of JT on VQ: regression model is shown in

$$\widehat{\mathrm{VQ}} = 15.85106 - 0.0000264 * \mathrm{DP}. \quad (6)$$

From Tables 6 and 7, we can find that explanatory power of JT is not strong for MQ and VQ. But in the 5% level of significance, it is statistically significant. And if making a fusion of these two network parameters, MQ and VQ can get a good degree of interpretation and estimation.

Taking first ninety frames of akiyo streaming video as example, we can draw PSNR value and MOS value of line charts. In Figures 4(a) and 4(b), the abscissas represent the



(a) Signal-to-noise ratio PSNR (b) MOS value

Figure 4: PSNR value and MOS value of first ninety frames.

Table 4: Statistical parameters of DP on MQ.

	t-statistics	R-square	Adjusted R-square	The 95% confidence interval
DP	−5.68	0.4021	0.3897	[−0.0079424, −0.0037906]

Table 5: Statistical parameters of DP on VQ.

	t-statistics	R-square	Adjusted R-square	The 95% confidence interval
DP	−6.73	0.4856	0.4749	[−0.0467363, −0.0252389]

Table 6: Statistical parameters of JT on MQ.

	t-statistics	R-square	Adjusted R-square	The 95% confidence interval
JT	−3.00	0.1578	0.1403	[$-8.25e-06$, $-1.63e-06$]

Table 7: Statistical parameters of JT on VQ.

	t-statistics	R-square	Adjusted R-square	The 95% confidence interval
JT	−3.52	0.2050	0.1885	[−0.0000494, −0.0000135]

frames, and the ordinates represent the PSNR and MOS values.

As shown in Figure 4, when observing the same processing to the video and the characteristics of the complexity, we can find that MOS values are not significant. There is even more and more competition that it has influences on video broadcast node. For akiyo streaming video which has a node in the competition, the average score is about 3.6 and the standard deviation is 0.25. The average of four nodes is close to 3.4, and the standard deviation is 0.19. For video foreman that has a higher level of complexity and movement, the results are worse. As expected, each node of the MOS reduces. The study uses a large number of experimental results (include experimental results used for modeling and not used for modeling) and compares to the MOS value which is obtained by model calculation. The result error is no more than 0.4 and accuracy is within the acceptable ranges.

6. Conclusion

The paper studied the video transmission and quality evaluation issue over NS2 and Evalvid based simulation networks. Contributions of this paper are summarized in the following.

(1) Verify the influences of all kinds of factors when investigating the quality of image transmission, by changing values of the parameters.

(2) Establish effective evaluation model on the basis of influence analysis of different factors, using linear regression model.

In the simulation platform, we implement the transmission and quality evaluation of video streams. The results show that the video transmission and quality evaluation model we designed can evaluate the multimedia transmission performance over complex environment very well, and it could be a powerful tool in multimedia transmission research.

As for future work, we plan to study the impacts of network QoS on QoE of video, based on the model proposed in this paper. We will establish evaluation model of QoS to QoE and apply it to the network monitoring system which uses QoE as evaluation standard.

Conflict of Interests

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

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (nos. 60963011, 61162009), the Jiangxi Natural Science Foundation of China (no. 2009GZS0022), and the Special Research Foundation of Shijiazhuang Tiedao University (nos. Z9901501, 20133007).

References

[1] B. Tulu and S. Chatterjee, "Internet-based telemedicine: an empirical investigation of objective and subjective video quality," *Decision Support Systems*, vol. 45, no. 4, pp. 681–696, 2008.

[2] K. Navas and M. Sasikumar, "Image fidelity metrics: future directions," *IETE Technical Review*, vol. 28, no. 1, pp. 50–56, 2011.

[3] F. You, W. Zhang, and J. Xiao, "Packet loss pattern and parametric video quality model for IPTV," in *Proceedings of the 8th IEEE/ACIS International Conference on Computer and Information Science (ICIS '09)*, pp. 824–828, Shanghai, China, June 2009.

[4] K. Yamanishi and J.-I. Takeuchi, "A unifying framework for detecting outliers and change points from non-stationary time series data," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02)*, pp. 676–681, Edmonton, Canada, July 2002.

[5] E. Keogh, J. Lin, and A. Fu, "HOT SAX: efficiently finding the most unusual time series subsequence," in *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM '05)*, pp. 226–233, Houston, Tex, USA, November 2005.

[6] C. J. B. Lambrecht and O. Verscheure, "Perceptual quality measure using a spatiotemporal model of the human visual system," in *Digital Video Compression: Algorithms and Technologies*, vol. 2668 of *Proceedings of SPIE*, pp. 450–461, February 1996.

[7] "Methods for subjective determination of transmission quality," ITU-T Recommendation P.800, International Telecommunication Union-Telecommunication Standardization Sector, 1996.

[8] "Methodology for the subjective assessment of the quality of television pictures," ITU-R Recommendation BT.500-11, 2002.

[9] J. Welch and J. Clark, "A proposed media delivery index (MDI)," RFC 4445, 2006.

[10] M. H. Pinson and S. A. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[11] S. Mohamed, *Automatic evaluation of real-time multimedia quality: a neural network approach [Ph.D. thesis]*, University of Rennes I, Rennes, France, 2003.

[12] "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," ITU-T J.144, International Telecommunication Union-Telecommunication Standardization Sector, 2007.

[13] F. Yang, J. Song, S. Wan, and H. R. Wu, "Content-adaptive packet-layer model for quality assessment of networked video services," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 672–683, 2012.

[14] V. A. MacHado, C. N. Silva, N. L. Vijaykumar et al., "A new proposal to provide estimation of QoS and QoE over WiMAX networks: an approach based on computational intelligence and discrete-event simulation," in *Proceedings of the 3rd IEEE Latin-American Conference on Communications (LATINCOM '11)*, Federal University of Para (UFPA), Belem, Brazil, October 2011.

[15] J. Zečić, M. Hadžialić, and A. Hasković, "An approach to estimate correlation between QoS and perceptual video quality in packet-switched networks," in *Proceedings of the 35th International Convention (MIPRO '12)*, pp. 573–578, Opatija, Croatia, May 2012.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 725210, 6 pages
http://dx.doi.org/10.1155/2014/725210

Research Article

Multichannel Broadcast Based on Home Channel for Cognitive Radio Sensor Networks

Fanzi Zeng,[1] Yuting Tang,[1] and Jianjie Pu[2]

[1] *Key Laboratory for Embedded and Netowork Computing of Hunan Province, Hunan University, Changsha, Hunan 410082, China*
[2] *Yunnan Electric Power Design Institute, Kunming, Yunnan 650051, China*

Correspondence should be addressed to Fanzi Zeng; zengfanzipaper@outlook.com

Received 5 November 2013; Accepted 11 December 2013; Published 9 April 2014

Academic Editors: W. Sun, G. Zhang, and J. Zhou

Considering the limited resources and the dynamic spectrum distribution in the cognitive radio sensor networks (CRSN), a half-duplex Multichannel broadcast protocol for CRSN is presented based on the home channel. This protocol maintains the networks topology only through the home channel, so there is no need for the public channel to transmit the control information and no need for the synchronization. After network initialization, node broadcasts data via home channel in half-duplex transmission way. The simulation results show that, compared with complete broadcast, the proposed protocol effectively reduces broadcast delay and overhead.

1. Introduction

Wireless sensor network (WSN) is composed of a large number of densely deployed sensor nodes and is widely used in environment monitoring and protection, medical care, and military field [1, 2]. The WSN operates in the crowded ISM frequency band and uses fixed spectrum allocation strategy. The interference from other wireless technologies (e.g., Wifi, Bluetooth, and ZigBee) operating in the same frequency band degrades the transmission reliability and causes the bulk of retransmissions that drains off the battery of the sensor node.

Cognitive radio (CR) [3] can dynamically sense the spectrum hole and adaptively adjust the radio parameter, so it can increase dramatically the spectrum utilization efficiency and reduce the interference. It is natural to integrate CR into WSN that brings forward the new network model: cognitive radio sensor networks (CRSN) [4]. With the CR, CRSN can increase working bandwidth, reduce the waiting time, conflict, and energy consumption which are caused by channel competition, and then reduce the delay and overhead.

Broadcast as a kind of basic network operation is used to maintain the routing and gather the data in WSN. The existing broadcast protocol in WSN and CR cannot directly apply to CRSN because the broadcast protocol in WSN does not consider the dynamic spectrum allocation, and those in CR do not consider the resource limitation. The next two paragraphs will give an overview of broadcast protocol in WSN and CR, respectively.

For the broadcast protocol in WSN, Zhao [5] had proposed Maximum Life-time Broadcast Protocol which reduced the redundancy and energy consumption incurred by rebroadcasting through the method of node self-delay. Chakraborty et al. [6] presented a reliable completely spanning tree method for WSN broadcasting; Hsu et al. [7] proposed two kinds of WSN basic broadcast model. Those traditional WSN broadcasting methods were based on fixed spectrum allocation, without considering the channel change, routing change, and topology change caused by primary cognitive radio users (PU) access. Thus, the additional redundancy and delay will be produced and therefore cannot be directly applied to CRSN network.

For the broadcast protocol in CR, Kondareddy and Agrawal [8] proposed a kind of CR selectivity broadcast algorithm to reduce the broadcast delay. This method assumed network topology and the available channel information was known, but this assumption was difficult to set up for CRSN. Song and Xie [9] proposed a multihop CR broadcast algorithm based on the QoS; it used the dividing retransmission method for radio, ensured the QoS, but

consumed a lot of energy and was not suitable for node energy limited CRSN. Wu et al. [10] proposed MAC protocols based on the reservation request, but it could not handle the situation where a large number of CRSN nodes try to transmit data in a short time. So and Vaidya [11] proposed a MAC technology based on time division, but it required strict time synchronization for whole network; it was not practical for CRSN. Choi et al. [12] proposed a MAC protocol based on the home channel (HCh) in a kind of multihop CR network; this method set an available channel as one node's home channel and assumed that each node had multiple signal transceivers; one of the transceivers was fixed in the home channel. When a node wanted to send data to another node, the transceiver without the fixed channel jumped to the destination node home channel to transmit data. However, because the sensor node had simple hardware and requires low energy consumption, a single node using multiple transceivers to carry out the full duplex communication was not practical [4] in CRSN. The above cognitive radio MAC protocols did not consider the limited resource of sensor nodes; if directly applied to CRSN, network lifetime would be shorten. In addition, global public channel is not available in CRSN to transmit control information [13], and local public channel is not easy to achieve. Moreover, the public control channel methods have no way to cope with denial of service attack (DoS, denial of service) [14]. Therefore, an effective CRSN broadcast protocol must be developed to solve the above contradictions.

This paper proposes a half-duplex broadcast protocol for CRSN based on home channel, referred to as home channel broadcast. This protocol maintains the networks topology only through the home channel, so there is no need for the public channel to transmit the control information and no need for the synchronization. Moreover, the node has only one transceiver that simplifies the hardware design and reduces the energy consumption. When broadcasting, node selects its neighbor's home channel one by one from the channel state table.

The remainder of this paper is organized as follows. In Section 2, a network model is presented. Section 3 discusses the network initialization and network topology generalization. Section 4 proposes half-duplex data transfer model based on the home channel and shows a broadcast example using the transfer model. Section 5 verifies the validity of this protocol through the simulation. In Section 6, we conclude the paper.

2. Network Model

Assuming that N CRSN nodes (SU) and M primary users (PU) are randomly distributed in an $L \times L$ area. Each SU has the same hardware and software settings. In particular, each SU has only one transceiver and two states, one for sending signal (ST) and the other for receiving signal (SR, state receive), and the two states cannot occur at the same time. This means that each CRSN node can only do half-duplex communication.

Table 1: Node i's channel table.

Channel number	Channel state
Channel 1	Available
Channel 2	Available
Channel 3	Noisy
Channel 4	Access-PU
Channel 5	HCh_i
Channel 6	HCh_i
Channel 7	Available
⋮	⋮

Assume that the K channels can be utilized and the whole network without time synchronization signal and the public channel does not exist.

Each SU is equipped with omnidirectional antenna with the propagation radius r_t, and the other CRSN nodes in the propagation range are the neighbor nodes. Each SU has a radius for r_s circular sense scope to sense the primary user access and noise interference. The probability of each channel occupied by PU is λ_p; the probability of noise or interference beyond threshold of each channel is λ_n. This paper assumes that $r_t = r_s$.

$P(n)$ denotes the probability that n SUs are randomly distributed in the sensing range with area πr_s^2 and is described as

$$P(n) = C_N^n \left(\frac{\pi r_s^2}{L^2}\right)^n \left(\frac{L^2 - \pi r_s^2}{L^2}\right)^{N-n}. \quad (1)$$

Then the expectation of n is $E(n) = \sum_{n=1}^{N} nP(n) = N \times \pi r_s^2 / L^2$; if $E(n) \geq 1$, then the expectation of SU's neighbor number is

$$E_{\text{nb}} = E(n) - 1 = N \times \frac{\pi r_s^2}{L^2} - 1. \quad (2)$$

3. Network Initialization

Initially, all nodes are in SR state and sense the usage of channel to maintain the channel state stable. If occupied by PU, the channel is marked as *access-PU*. If the channel is not occupied by PU, but the interference goes beyond the threshold, it is marked as *noisy*. The rest of the channels are marked as *available*. The node randomly selects an available channel to mark as its home channel HCh. Table 1 shows node i's channel state table.

After sensing, node randomly selects an available channel as home channel and transfers its state from SR to ST and then broadcasts its channel state stable once. When finishing transmission, this node returns to SR state. The node within the propagation range of this node and happening to be in SR state receives the broadcast and updates its own channel state stable according to the received packets. Next, we take example to illustrate the update steps.

Take node i and node j as an example.

Node i in state ST sends a packet on an *available* channel and, after completing, returns to SR state.

If the node j has received packet from i, then j compares its own channel state table with i's channel state table and updates j's channel state table according to the following steps:

(1) node j marks channel HCh_i as the node i's home channel; it also means that node i is node j's neighbor node;

(2) node j marks node i's *noisy* channel as node j's *noisy* channel;

(3) node j marks node i's *access-PU* channel as node j's *access-PU* channel;

(4) if node j finds that HCh_j channel is the same as node i's *noisy* channel, then node j randomly chooses an available channel as a new HCh_j and marks the previous HCh_j as *noisy*;

(5) if node j finds that HCh_j is the same as HCh_i, then node j randomly chooses an available channel as a new HCh_j.

After node j completed updating channel state table, node j turns SR state to ST state, transmits the new channel state table to node i through HCh_i immediately after finishing sending, and returns to SR state. Node i repeats the above steps to update its channel state table after receiving the reply packet from node j. This procedure is iterated until the two nodes state table is consistent.

Through the network initialization, all the nodes know its neighbors, neighbors' home channel, and channel state. So the network topology is constructed and the broadcast routing is built. If one node needs to send broadcast packet, it just checks its channel state table to get the neighbors' home channel and transmits it.

4. Broadcast Protocol Based on Home Channel

4.1. Half-Duplex Data Transfer Mode Based on Home Channel. Through the network initialization, each CRSN node chooses an available channel as its HCh and tunes its transceiver into SR state to receive data on HCh until node needs to send a packet. When node needs to send a packet, state is changed from SR to ST and data packets are transmitted on HCh to target node. After finishing transmission, the node returns to SR state.

When source node broadcasts the packet and the target node happens to be in SR state, the broadcast packet can be forwarded successfully. Otherwise, the packet is dropped. To deal with this situation, this paper utilizes the time-out retransmission mechanism: when receiving packets, the target node returns an acknowledgement on the HCh to source node. If source node receives the acknowledgement within the given interval, the packet transmits successfully. Otherwise, the packet fails to broadcast and the source node retransmits the packet. In this paper, we allow the retransmission only once. This packet is sent two times in the failure situation.

Let P_1 represent the probability that source node sends successfully data packets to its neighbor node with just one time transmission. Let P_2 denote the probability that source node sends data packets to its neighbor node with two times transmissions. Let P_0 denote the probability that the source node receives the replica of broadcast packet from neighbor node and no transmission is triggered for the source node. The expectation of the times that a node sends a packet to one neighbor node is as follows:

$$E_s = 0 \times P_0 + 1 \times P_1 + 2 \times P_2, \tag{3}$$

where $P_0 + P_1 + P_2 = 1$; thus $0 \le E_s \le 2$ and when $P_0 = 0$, $P_1 = 0$, and $P_2 = 1$, E_s has maximum value 2.

4.2. Home Channel Broadcast. When one node has data packets to broadcast, it checks the channel state table, tunes its transceiver into the neighbor's home channel, and sends the data packet and corresponds to neighbors in turn until the end of the channel state table. Neighbor node follows the update step in Section 3 to update its channel state table according to the received channel state table and then sends broadcast packet to its updated neighbor nodes (except for its upstream node) according to the updated channel state table. When receiving multiple broadcast packet replicas, the node no longer forwards this packet. So, according to (2) and (3), if $(n) \ge 1$, then the expectation of times which a node needs to forward a broadcast packet in a broadcast task is $E_{\mathrm{node}} = E_{\mathrm{nb}} \times E_s$; then the number of broadcast packets across the network can be calculated as follows:

$$\begin{aligned} E_{\mathrm{all}} &= E_{\mathrm{node}} \times N \\ &= N\left(\frac{\pi r_s^2}{L \times L} \times N - 1\right)\left(0 \times P_0 + 1 \times P_1 + 2 \times P_2\right) \\ &\le 2N\left(\frac{\pi r_s^2}{L \times L} \times N - 1\right). \end{aligned} \tag{4}$$

Set

$$E_{\max} = 2N\left(\frac{\pi r_s^2}{L \times L} \times N - 1\right). \tag{5}$$

$E_{\max}$ represents the maximum value of times that a broadcast packet transmitted in the whole network; in another word, in the worst situation of home channel broadcast, the times of a broadcast packet transmitted will not exceed $E_{\max}$.

As shown in Figure 1, black dot represents CRSN nodes; solid line represents broadcast path. Node i as source node sends broadcast packet through neighbor node a's home channel (represented as Ch n in Figure 1) Ch6, neighbor node b's home channel Ch4, neighbor node c's home channel Ch1, and neighbor node d's home channel Ch3 in sequence in the sensing range r_s. Then nodes a, b, c, and d receive broadcast packets and forward them through HCh of its own neighbor nodes, respectively. Nodes which constitute a connected graph with node i will receive broadcast packets after a while.

For example, the process of node g receiving broadcast packets from source node i is as follows. i sends broadcast

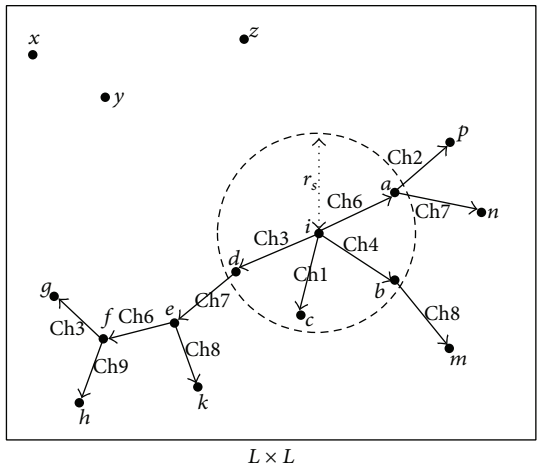


Figure 1: Nodes broadcast example.

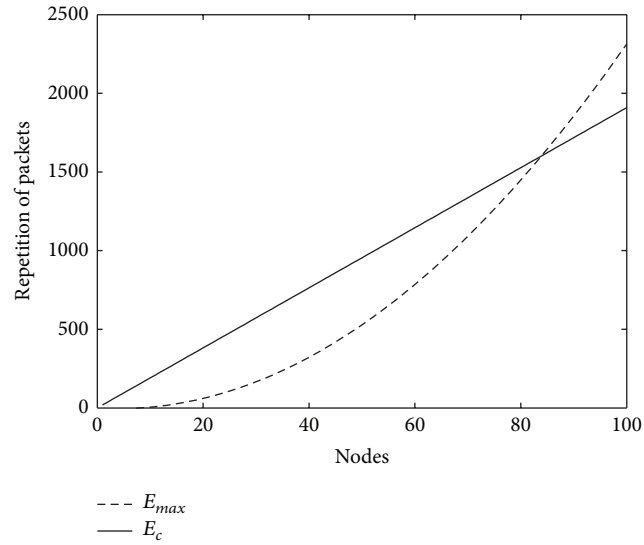


Figure 2: Number of a broadcast package forwarded in simulation region versus total number nodes.

packets to d through the d's home channel Ch3, d sends broadcast packets to e through the e's home channel Ch7, e sends broadcast packets to f through the f's home channel Ch6, and finally f sends broadcast packets to g through the g's home channel Ch3. Because the nodes are set randomly scattered in this paper, no connectivity path between two nodes may occur, as shown in Figure 1; the x, y, and z nodes will never receive broadcast packets from node i. In addition, two nodes such as k and m are not within the scope of mutual sensing range, so the same channel can be selected as the HCh by them which does not cause the interference.

5. Simulation Results and Analysis

To evaluate the performance of the proposed protocol, some simulations results are presented in this section. Nowadays, no reference is about the broadcast protocol in CRSN, so this protocol is compared to the traditional complete broadcast protocol.

Complete broadcast refers to the protocol where the source node is sending a broadcast packet on each available channel one by one; other nodes receive this broadcast packet and forward it on its all available channels one by one, and nodes are in half-duplex communication.

Expectation of the number of available channels is $E_a = K(1 - \lambda_p - \lambda_n)$; then the number of a broadcast packet sent across the network expects to be

$$E_c = N \times E_a = N \times K\left(1 - \lambda_p - \lambda_n\right). \quad (6)$$

This paper uses MATLAB to do Monte Carlo simulation. Simulation region is set to 1000 m × 1000 m; node number N changes from 1 to 100; all nodes are randomly distributed in the region. Sensing radius of each node r_s and transmitting radius r_t are equal to 200 m, and, within the range, the probability λ_p that channel is occupied by PU is equal to 0.025 and probability λ_n that channel is noisy is equal to 0.02. Total channels $K = 20$ can be utilized across the region; PU randomly selects one of the K channels and accesses this channel. All simulation results are an average of 10,000 simulations.

From the parameters given above and (5) and (6), we have $E_c = N \times 19.1$, when $N \geq 8$, $E_{\max} = 2(0.1257N^2 - N)$. E_c and $E_{\max}$ theoretical value is shown in Figure 2. It can be seen that $E_{\max}$ will surpass E_c only when the node is extremely dense. But in practice, due to no forwarding broadcast packets to the node which has sent the packets to it before, $P_0 > 0$. In addition, node stays in ST state only when it needs to transmit packets. After finishing transmission, node is back to SR state, so the probability $P_1 > 0$, so it can be estimated that E_{all} is a lot smaller than $E_{\max}$ in reality from (3) and (4).

In this paper, overhead is defined as the quantity of broadcast packets all nodes send and forward in the simulation region; E_c and E_{all} represent theoretical overhead value of the complete broadcast and home channel broadcast, respectively. Simulation results in Figures 3 and 4 show broadcast overhead with the number of nodes and the number of channels, respectively. As can be seen, compared with the full broadcast, the home channel broadcasts dramatically reduce overhead, such as when there are 40 and 100 nodes and when overhead is only about one-thirteenth and one-seventh compared with complete broadcast. Thereby, the proposed protocol significantly reduces energy consumption, which is essential for energy limited sensor nodes.

For home channel broadcast, the packets transmission uses the known and fixed channel, so extra overhead will be less. While in complete broadcast, every packet must be sent and forwarded across all available channels one by one, which produces a lot of extra overhead.

This paper tests average broadcast delay under the different number of total channels. As can be seen from Figure 5, average delay of the home channel broadcast is far less than complete broadcast's and has nothing to do with the total number of channels. This is because the home channel broadcast sending and forwarding through a fixed

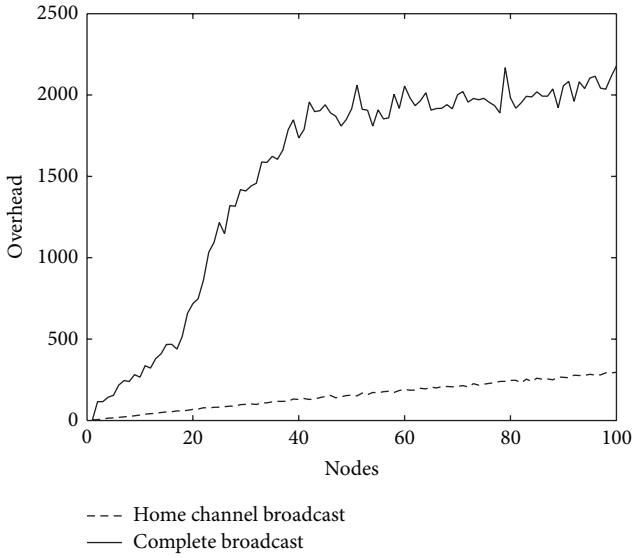


Figure 3: Average overhead in simulation region versus total number of nodes.

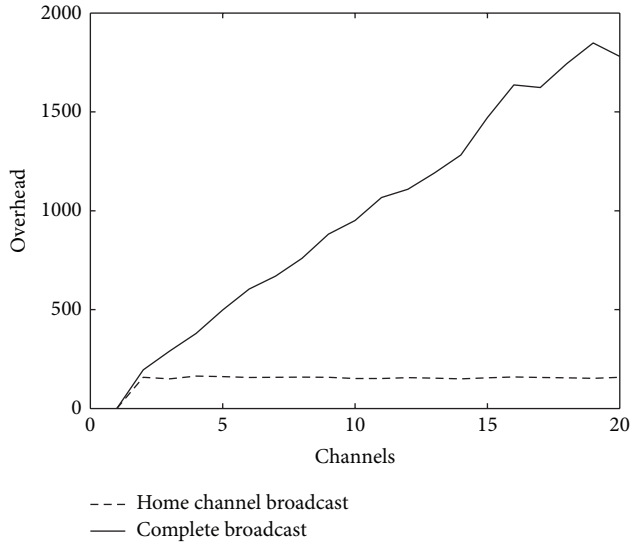


Figure 4: Average overhead versus channels.

channel at a time is not affected by the total number of channels, and the fixed channel transmission mode avoids the delays caused by blind transmission and retransmission. For complete broadcast, every broadcast packet must be sent and forwarded across all available channels, so the more delay occurs.

6. Conclusions

This paper proposed multichannel broadcast protocol for CRSN based on home channel. After network initialized, nodes form a topology through the home channel. Node broadcasts data via home channel in half-duplex transmission way. Simulation experiments show that the home channel broadcasts reduce delay and overhead compared with the complete broadcast.

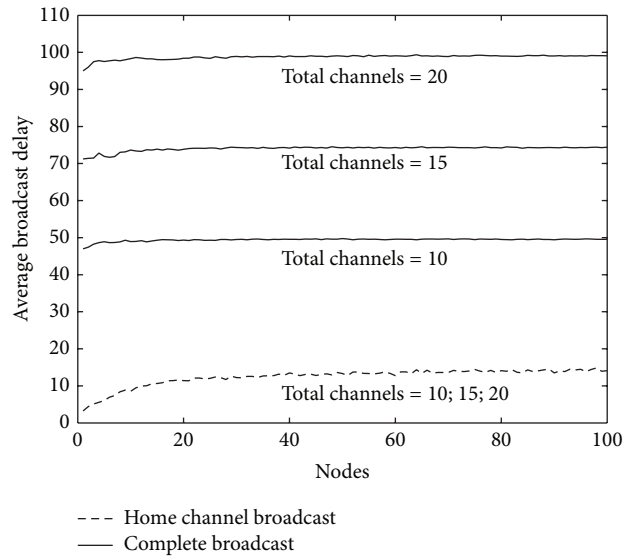


Figure 5: Average broadcast delay (ms) in simulation total region versus total number of channels.

Conflict of Interests

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

References

[1] A. Merentitis, N. Kranitis, A. Paschalis, and D. Gizopoulos, "Low energy online self-test of embedded processors in dependable WSN nodes," *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 1, pp. 86–100, 2012.

[2] J. Li and H. Gao, "Survey on sensor network research," *Computer Research and Development*, vol. 45, no. 1, pp. 1–15, 2008 (Chinese).

[3] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, 2005.

[4] O. B. Akan, O. B. Karli, and O. Ergul, "Cognitive radio sensor networks," *IEEE Network*, vol. 23, no. 4, pp. 34–40, 2009.

[5] R. Zhao, X. Shen, X. Zhang, and J. Hou, "Maximum lifetime broadcast protocol for Wireless Sensor Networks," in *Proceedings of the International Conference on Computer Application and System Modeling (ICCASM '10)*, pp. V11440–V11444, Taiyuan, China, October 2010.

[6] S. Chakraborty, S. Chakraborty, S. Nandi, and S. Karmakar, "A reliable and total order tree based broadcast in wireless sensor network," in *Proceedings of the 2nd International Conference on Computer and Communication Technology (ICCCT '11)*, vol. 1, pp. 618–623, Allahabad, India, September 2011.

[7] C.-S. Hsu, J.-P. Sheu, and Y.-J. Chang, "Efficient broadcasting protocols for regular wireless sensor networks," in *Proceedings of the International Conference on Parallel Processing (ICPP '03)*, vol. 1, pp. 213–220, Kaohsiung, Taiwan, 2003.

[8] Y. R. Kondareddy and P. Agrawal, "Selective broadcasting in multi-hop cognitive radio networks," in *Proceedings of the IEEE Sarnoff Symposium (SARNOFF '08)*, vol. 1, Princeton, NJ, USA, April 2008.

[9] Y. Song and J. Xie, "A QoS-based broadcast protocol for multi-hop cognitive radio ad hoc networks under blind information," in *Proceedings of the 54th Annual IEEE Global Telecommunications Conference (GLOBECOM '11)*, pp. 1–5, Houston, Tex, USA, December 2011.

[10] S. Wu, C. Lin, Y. Tseng, and J. Sheu, "A new multi-channel MAC protocol with on-demand channel assignment for multi-hop mobile ad hoc networks," in *Proceedings of theInternational Symposium on Parallel Architectures, Algorithms, and Networks*, vol. 1, pp. 232–237, Dallas, Tex, USA, 2000.

[11] J. So and N. Vaidya, "Multi-channel MAC for ad hoc networks: handling multi-channel hidden terminals using a single transceiver," in *Proceedings of the 5th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MoBiHoc '04)*, vol. 1, pp. 222–233, New York, NY, USA, May 2004.

[12] N. Choi, M. Patel, and S. Venkatesan, "A full duplex multi-channel MAC protocol for multi-hop cognitive radio networks," in *Proceedings of the 1st International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM '06)*, vol. 1, Mykonos Island, Greece, June 2006.

[13] J. Zhao, H. Zheng, and G.-H. Yang, "Distributed coordination in dynamic spectrum allocation networks," in *Proceedings of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '05)*, vol. 1, pp. 259–268, Baltimore, Md, USA, November 2005.

[14] Z. Li and Z. Huaibei, "Two types of attacks against cognitive radio network MAC protocols," in *Proceedings of the International Conference on Computer Science and Software Engineering (CSSE '08)*, pp. 1110–1113, Wuhan, China, December 2008.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 630986, 6 pages
http://dx.doi.org/10.1155/2014/630986

Research Article

An Improved Clustering Algorithm of Tunnel Monitoring Data for Cloud Computing

Luo Zhong,[1] KunHao Tang,[1,2] Lin Li,[1,3] Guang Yang,[1] and JingJing Ye[1]

[1] *Department of Computer Science and Technology, Wuhan University of Technology, Wuhan 4300702, China*
[2] *Department of Computer and Information Science, Hunan Institute of Technology School, Hunan 421002, China*
[3] *Collaborative Innovation Center of Hubei Province for Information Technology & Service of Elementary Education, Hubei 430070, China*

Correspondence should be addressed to KunHao Tang; kunhaotang@sina.com

Received 7 November 2013; Accepted 4 December 2013; Published 2 April 2014

Academic Editors: W. Sun, G. Zhang, and J. Zhou

With the rapid development of urban construction, the number of urban tunnels is increasing and the data they produce become more and more complex. It results in the fact that the traditional clustering algorithm cannot handle the mass data of the tunnel. To solve this problem, an improved parallel clustering algorithm based on k-means has been proposed. It is a clustering algorithm using the MapReduce within cloud computing that deals with data. It not only has the advantage of being used to deal with mass data but also is more efficient. Moreover, it is able to compute the average dissimilarity degree of each cluster in order to clean the abnormal data.

1. Introduction

At present, with the rapid development of municipal construction in our country, the tunnel of urban is developing progressively. The tunnel data is becoming more and more complex. The clustering algorithm based on partition is of simplicity and accuracy, and it has been widely applied in scientific researches and production practice. As a classical clustering algorithm, k-means, which is based on partition, is a hot topic all the time. However, tunnel data has already entered the level of massive amounts of data, taking the time, space, and data volume complexity [1] into account. The traditional clustering algorithm k-means has been unable to deal with these data.

According to the characteristics of the k-means, to use the parallel computing based on MapReduce in the cloud platform imperative. The probability of cloud computing is proposed by Google; it is a kind of calculation model that combines several technologies; it also has some characteristics, such as high reliability, high extensibility, and super large scale. After that, IBM also brought the cloud computing platform [2] so that clients could be ready to use them. Cloud platform, using Hadoop technology, can realize large-scale distributed computing and processing. It is mainly made up by Hadoop distributed file system and MapReduce programming model [3] which was widely used in cloud computing.

In addition, the MapReduce model is a distributed parallel programming model, which was proposed by Google Labs. It can deal with problems including large datasets with computer cluster, and that makes it become a mainstream of the parallel data processing model of cloud computing platform [4]. The MapReduce model also can easily solve the problems that K-means cannot handle, such as processing large data and demanding bigger serial overhead.

The city tunnel monitoring data include a variety of forms; they contain the instantaneous of traffic lane, speed, occupancy rate, wind speed, temperature, humidity, light intensity, and CO concentration [5]. They are multidimensional data from different equipment so that it can ensure the effectiveness of the experimental data [6]. With the continuous development of data processing technology, people begin to analyze huge amounts of data statistics [7] by using various data mining techniques and tools. The purpose of cluster data

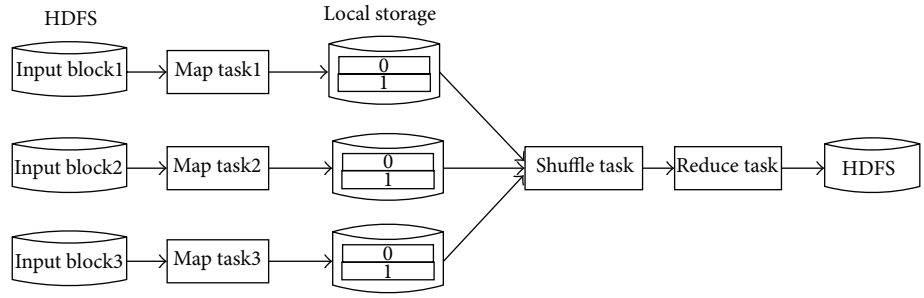


Figure 1: The interaction figure of HDFS and MapReduce.

is to use relevant technology to classify the extracted data from various data sources while finding the abnormal data.

However, the tunnel monitoring system demands real-time higher. At the same time, the collected data storage density is intensive while the data storage is also large [8]. It is such a large amount of data that it cannot use serializing approach to find out the abnormal data. Moreover, it is not able to cluster [9]. Therefore, we use the distributed processing technology to deal with data by cloud computing [10].

2. The Introduction of Cloud Platforms

The cloud platform we normally refer to is the platform which can develop easily and deal with processing of mass data in parallel Hadoop [11]. It is mainly composed of the distributed file system, which is called HDFS and the calculation model, which is called MapReduce [12].

2.1. Hadoop Distributed File System (HDFS). HDFS has adopted the master/slave structure [13], which is principally composed by Client, Name Node, Secondary, Name Node, and Data Node. Every HDFS cluster consists of a unique management node (the Name Node) and a certain number of data nodes (the Secondary Node), and every node is an ordinary PC. HDFS is very similar to the stand alone; it also can build directory, create, copy, delete files, check the file content [14], and so forth. HDFS is a distributed file system which possesses high fault-tolerant; it can provide high throughput data access [15], which is fairly suitable for the application in large datasets. The interaction process between HDFS and MapReduce is shown in Figure 1.

2.2. The Calculation Model of MapReduce. MapReduce is a highly efficient distributed programming model [16] which is developed by Google. It can abstract problems highly and make the problem become simple and it is mainly particularly used for large-scale (TB) data processing [17]. The work process of MapReduce is mainly composed of three parts, Map, Shuffle, and Reduce.

(1) The process of Map: Map reads data from the HDFS [18], and the data is divided into several independent shards (Split) and through the Map function iteration parsed into key/value pair ⟨key/value⟩.

(2) The process of Shuffle: Sorting provisional results is generated by the Map phase, according to the key value, dividing the map production temporary data (key/value) into several groups (partition), and, finally, every partition will give a the Reduce Task processing in the form of ⟨key, value list⟩.

(3) The process of Reduce: reading from the remote node ⟨key, value list⟩ and merging them through the reduce function, and in the end, transferring the final result into the HDFS [19].

3. Traditional Clustering Algorithm and Its Parallel

3.1. Traditional (Serializing) k-Means Clustering Algorithm. So far, the commonly used clustering analysis algorithms are composed of the following five categories: the algorithms based on classification, the algorithms based on the hierarchy, the algorithms based on density, the algorithms based on grid, and the algorithms based on model-based [20]. Studies show that tunnel data storage density is large, so algorithms based on data partition are more applicable to tunnel. But k-means is a typical representative of algorithms based on partition. k-means algorithm; a partition clustering method [21], which regards an average value as the center of the cluster, and the specific steps of serializing k-means clustering method are as follows:

(1) selecting several objects from the original dataset randomly, and treating them as the initial clustering center,

(2) to calculate the distances between the other objects and the center and assign them to the class of the closest center,

(3) for each class, working out the means of all the objects as a new center,

(4) repeat (2) and (3) until the standard measurement function begins to converge.

3.2. The Parallel k-Means Clustering Algorithm. Although the sequential k-means algorithm is simple and convenient, a lot of defects still exist. The time complexity of k-means is $O(n * k * t)$; in addition, n is the number of all the objects, k denotes the cluster number of clustering, and t represents the number of iterations [22]. Thus it can be seen that the time complexity of sequential algorithm is very high; the iteratively replaced

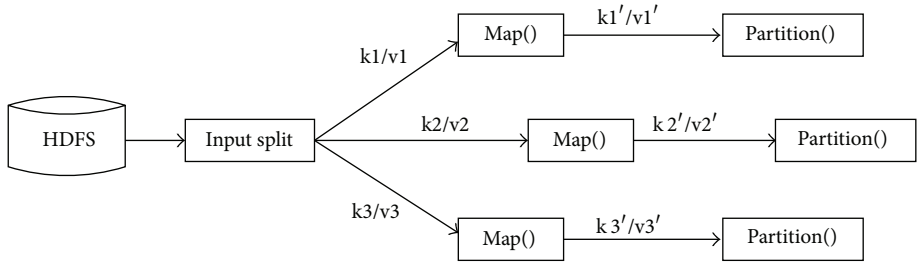


Figure 2: The execution figure of Map.

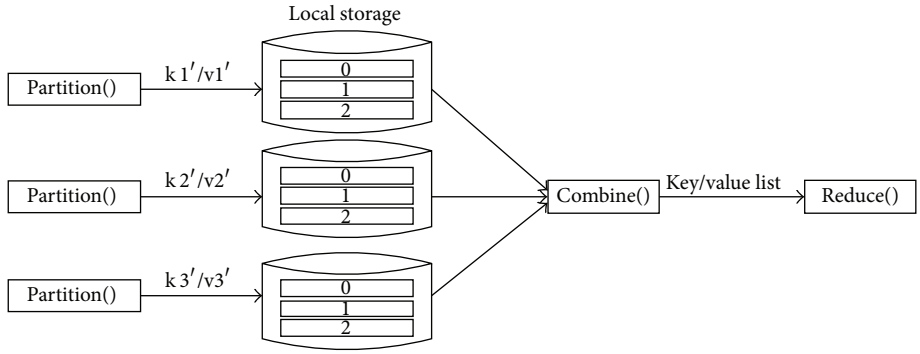


Figure 3: The execution figure of Shuffle.

part consumes the most time during this operation, and it is easier for parallel processing [23]. The MapReduce process of parallel k-means is as follows.

(1) The Process of Map. The Map phase of the MapReduce process is performed by several Map functions. According to the number of input splits, there will be Map function which has some inputs to calculate the distance between every data object and the current cluster center. Furthermore, we can tag the corresponding cluster number for data objects according to the distance. The cluster number is the number of cluster center which is closest to the object. Its input is that all the data objects waiting on clustering and the clustering center during the previous iteration (or the initial center); its input is in the form of (key/value). In key/value pairs, key is corresponding to the offset in current data samples which is relative to the starting point, while value is corresponding to the vector of strings in the current data sample which is composed of each dimension coordinate. After the Map phase, the result will be in the form of key/value pairs to output the clustering cluster number and recorded value.

Its execution process is shown in Figure 2.

(2) The Process of Shuffle. There are two main functions, Partition and Combine. Finishing it, the input during this phase is the key/value pair which was worked out during the Map phase, and it is mainly used before entering the Reduce phase, combining the data locally in order to reduce the consumption of data communication in the process of iteration. First of all, the Partition function sorts the provisional result which was produced during the Map phase and divides the temporary data (key/value pairs) which was produced by map into several groups (Partition). The key which was input is corresponding to the cluster number of clustering, while values are corresponding to vector of strings which are composed by dimensional coordinates. Then, the Combine function will combine group and count the number of data samples and the sum of all the data coordinates; in addition, the outputs which were in the form of key/value pair (key/value list) are given to different Reduce functions. Its execution process is shown in Figure 3.

(3) The Process of Reduce. Reduce function accepts the key/value pairs (key and value list) which were worked out by Combine function. Among them, corresponding to the key is the number of data samples, and the values list is corresponding to the vector of strings which are composed by dimensional coordinates. By adding string vectors of every dimensional coordinates, and combining with the number of the data, to work out the new cluster center which is input in the form of key/value pairs into the next MapReduce process.

4. The Good Center Parallel K-Means (GCPK) Algorithm

4.1. The Calculation Method of Dissimilarity Degree. This experiment uses Euclidean distance as a measurement standard to calculate the similarity degree among data in the cluster. In this article, the average dissimilarity degree mentioned refers to the average Euclidean distance between the data within the cluster and the cluster center. If the average dissimilarity degree is lower, the data within the cluster will

be more similar; in other words, the results of clustering are more ideal. The Euclidean computationalformula is

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}. \quad (1)$$

The "m" denotes the dimension value of the array.

4.2. The Defects of Parallel K-Means Clustering Algorithm. K-means parallel clustering algorithm solves the problem of the iteration time that k-means spends very well, but it does not handle the selection problem of k-means initial center. In order to choose appropriate initial center, reduce unnecessary iteration, and decrease the operation time and improve efficiency, in view of the tunnel data, we proposed the K-means parallel improved algorithm. Good center parallel k-means algorithm, whose abbreviation is GCPK algorithm. In the dataset, the algorithm can find out the algorithm firstly find out a point as the center, which can separate data on average basically; in other words, it is appropriate to be the good center for the initial center, and then to parallelize the k means clustering. Finally, based on the clustering results, the abnormal data will be analyzed.

4.3. The Implementation of GCPK Algorithm

(1) To Get the Initial Center through Relative Centroid. k-means clustering method is a typical method based on clustering; in the k-means clustering method, because of replacing the center constantly [24], the selection of initial center has a great influence on k-means clustering algorithm [25]. In order to decrease the times of replacement, and make the k-means clustering become more efficient, it is necessary to find the initial center which is relative to the center and roughly divide the dataset.

Because the tunnel data has the characteristics of nonnegativity, tunnel data can be regarded as multidimensional data graphics, which is composed of n-dimensional data. Regarding each data as their own dimensional coordinates, find out the point $Q1(i_1, i_2, \ldots, i_n)$ which is closest to the origin of coordinates and the point $Q2(j_1, j_2, \ldots, j_n)$ which is farthest away from the origin.

To connect $Q1$ and $Q2$ with a line segment, and divide the line segments into $k + 1$ parts (this k here is the same as the k in k-means), regard each data as a particle, and there are k equal diversion points on the line segment (not including the endpoint). Furthermore, the coordinate of the t $(0 < t < k + 1, t \in Z)$ equal diversion point is $(P_{t1}, P_{t2}, \ldots, P_{tn})$, and n denotes the number of dimensions. So $P_{tm} = (m/k)(i_m + j_m)$, $m = 1, 2, \ldots, n$. These k points obtained are the points in the center position which can average the multidimensional data graphics, and we regard the k points as the initial center points.

(2) The Specific Implementation of GCPK Parallel Clustering

(1) Upload the dataset to the HDFS.

(2) Upload the initial center to the HDFS.

(3) According to the original data and the previous iteration (or initial center), calculate the current iteration with the Map function.

(4) Calculate the sum of the points' number and weight in the same cluster and recount the clustering center of cluster.

(5) Do not control the iterative process until the maximum number of iterations is exceeding or all the clustering has been convergent, at this time, calculate the average dissimilarity degree and the maximum dissimilarity degree within each cluster of all clusters; in addition, count the average dissimilarity degree and the maximum dissimilarity degree in the whole cluster.

(3) The Analysis of Abnormal Data. Download the result of clustering to local through HDFS and convert binary serialization format file into a text file. Analyzing the results file, and getting the average dissimilarity degree and the maximum dissimilarity degree in every cluster, counting the radius of every dimension, combining the weight of each dimension data for cleaning abnormal data. Because the amounts of data are huge and the data quantity is too large, it is unable to complete artificial recognition and use the simple method to identify the abnormal data. It is necessary to use cloud computing technology to clean the abnormal clustering results.

(1) Read the average dissimilarity degree D_i in every cluster and the dimension radius R_j. Furthermore $i = 1, 2, \ldots, k$, and k are the cluster's numbers, $j = 1, 2, \ldots, n$, and n are the dimensions.

(2) Calculate the weight Q_j, which is within the cluster. $Q_j = t_j/n$. The "t" is the number of times in which the data of dimension j come out; the "n" is the total dimension.

(3) Work out the reasonable range of dissimilarity degree dis(i) in clusters; in addition

$$\mathrm{dis}(i) = \sqrt{D(i)^2 + \sum_{j=1}^{n} Q_j * R_j^2} \quad (2)$$

$i = 1, 2, \ldots, k$, k denotes the number of all clusters.

(4) Test the data object which is got from MapReduce according to the data objects. Euclidean distance of the cluster center (dissimilarity degree) and the reasonable range of dissimilarity degree within clusters, and to judge whether the data is abnormal or not. If dissimilarity degrees are beyond reasonable range, we can know it is the abnormal data and record the number.

5. Experimental Environment and Results Analysis

5.1. Experimental Environment

5.1.1. Single-Machine Environment. This paper uses MATLAB2010 single-machine environment; the operations of the

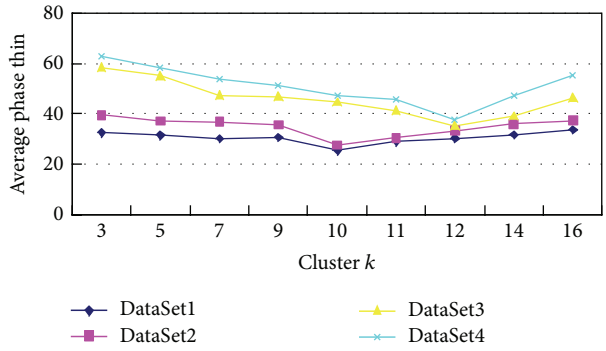


Figure 4: k values for experiments.

specific environment are the Intel(R) Core(TM) i3 CPU 2.53 GHz, 4 GB of memory, and a 64-bit operating system.

5.1.2. Distributed Environment. In this paper, the experiment of distributed environment uses open source cloud computing platform hadoop, whose version is 1.1.2. Specific cloud cluster is as follows: one machine runs as the Name Node and Job Tracker while the other 19 run as a Catano and Task Tracker. The experimental cloud platform uses the VM technology; specific software configuration is Cream ware workstation-full-9.0.2, a 64-bit Ubuntu-12.04.3, JDK version is jdk1.7.0_21, and eclipse version is Linux version.

5.1.3. Data Source. The data for clustering analysis are from the tunnel monitoring data, including flow, temperature, wind speed, illumination intensity, CO concentration, speed, wind speed, and so on.

5.2. The Experimental Data. This experiment has adopted the five datasets which were gathered from tunnel monitoring system of Wuhan Fruit Lake; respectively, they were DataSet1, DataSet2, DataSet3, DataSet4, and DataSet5. Furthermore, the seventh dimensions tunnel data record has been covered 42993 times by the DataSet1, which can be used for single experiment and distributed experiment. The thirtieth dimensions tunnel data record has been covered, respectively, 42775 times, article 65993 times, article 100 000 000 times, and article 2000 000 000 times by DataSet2–DataSet5.

5.3. Analysis of Experimental Result

Experiment 1. There are 42993 tunnel data whose dimension value is 7; there are 42993 tunnel data whose dimension value is 7; the ideal value of these data is 5, under the condition of single machine; MATLAB was used to complete k-means serial clustering algorithm [26], Clara clustering algorithm [27], and RICA clustering algorithm [28] in the case of the distributed cluster to accomplish k-means parallel clustering algorithm with hadoop [29] and to cluster the tunnel data with GCPK; the results are shown in Table 1 under the condition of two kinds of the clustering algorithm.

Experiment 2. For k-means clustering, the initial value of k has great influence on the clustering results, and the traditional clustering cannot analyze the k in the huge amounts of data. This experiment adopted GCPK algorithm, respectively. In terms of four datasets, which, respectively, are DataSet1, DataSet2, DataSet3, and DataSet4, we choose the different k values for them to conduct experiments. And the results are shown in Figure 4.

Table 1: Thin contrast of DataSet1.

	Serializing k-means	Clara	RICA	Parallel k-means	GCPK
Cluster (k)	5	5	5	5	5
Average phase thin	23.2027	21.8868	19.0924	16.0240	16.0226
Maximum phase thin	105.2276	113.2348	96.5631	78.2311	77.3419
Abnormal finding	83.22%	89.36%	89.53%	93.23%	93.3%

Note: the average phase thin and maximum phase thin are the mean values of each cluster.

Table 2: Clustering time.

Dataset	cluster k-means	Serializing k-means	Parallel k-means	GCPK
DataSet2	10	20 min 18 s	7 min 50 s	7 min 22 s
DataSet3	10	27 min 35 s	9 min 21 s	8 min 9 s
DataSet4	12	incalculable	18 min 33 s	16 min 20 s
DataSet5	12	incalculable	29 min 9 s	23 min 15 s

Experiment 3. According to the tunnel monitoring data collected from DataSet2 to DataSet5 (the dimension value is 30), under the distributed circumstances, clustering with k-means parallel algorithm and k-means parallel improved algorithm severally; the cluster number k would find ideal value through GCPK; the time cost in Clustering was shown at Table 2.

6. The Conclusion and Expectation

Experimental results show that GCPK, the improved algorithm of k-means parallel clustering algorithm, has a better clustering result than the traditional clustering algorithm and has a higher efficiency. It is very convenient to clean the abnormal data for the clustering results, in terms of massive data, such as tunnel data; it has shown its strong practicability. At present, due to the limited data form, we are only able to do preliminary clustering about structured document. However, after the relevant documents for clustering of data protection and the processing of unstructured documents, we will do further researches on the file data protection after clustering and the processing of unstructured documents in the future.

Conflict of Interests

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

Acknowledgments

This research was undertaken as part of Project 61003130 funded by National Natural Science Foundation of China, Collaborative Innovation Center of Hubei Province for Information Technology and Service of Elementary Education, Wuhan Technology Department, Innovation Project (2013070204005), Hengyang Technology Department (2012KG76), and Hunan Science and Technology Plan Project (2013GK3037).

References

[1] G. J. Li, "The scientific value in the study of big data," *Computer Science Communication*, vol. 8, no. 9, pp. 8–15, 2012.

[2] K. Chen and W.-M. Zheng, "Cloud computing: system instances and current research," *Journal of Software*, vol. 20, no. 5, pp. 1337–1348, 2009.

[3] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[4] X. P. Jing, C. H. Li, and W. Xiang, "The effectuation of Bayes book clustering algorithm under cloud platform," *Computer Application*, vol. 31, no. 9, pp. 2551–2554, 2011.

[5] R. Guo and X. H. Zhao, "The research of urban traffic information service system based on database technology," *Information Communication*, vol. 2, pp. 105–109, 2013.

[6] Q. J. Zhang, L. Zhong, J. L. Yuan, and Q. B. Wang, "Design and realization of a city tunnel environment monitoring system based on observer pattern," in *Proceedings of the 2nd International Conference on Electric Technology and Civil Engineering (ICETCE '12)*, no. 5, pp. 639–642, 2012.

[7] D. Ben-Arieh and D. K. Gullipalli, "Data envelopment analysis of clinics with sparse data: fuzzy clustering approach," *Computers & Industrial Engineering*, vol. 63, no. 1, pp. 13–21, 2012.

[8] L. Zhong, L. Li, J. L. Yuan, and H. Xia, "The storage method and system of urban tunnel monitoring data based of XML," patent, 201110458995.2, China.

[9] C. Liu, "No default categories for large amount of data clustering algorithm research," vol. 4, pp. 1–6, 2012.

[10] Q. Deng and Q. Chen, "Cloud computing and its key technologies," *Development and Application of High Performance Computing*, vol. 1, no. 2, p. 6, 2009.

[11] S. Owen and R. Anil, *Shadoof in Action*, Manning Publications, 2012.

[12] M. Armbrust and A. Fox, "Above the clouds: a Berkeley view of cloud computing," Tech. Rep., University of California at Berkeley, Berkeley, Calif, USA, 2009.

[13] L. Dong, "Shadoof technology insider," *Mechanical Industry Publication*, vol. 5, pp. 31–36, 2013.

[14] R. Chao, "Research and application of Shadoof based on cloud computing," Tech. Rep., Chongqing University, 2011.

[15] S. Thu, G. Bhang, and E. Li, "The improvement of k-means algorithm based on Hadoop platform oriented file data set," *Harbin Normal University Journal of Natural Science*, vol. 1, no. 28, 2012.

[16] J. Zhang, H. Huang, and J. Wang, "Manifold learning for visualizing and analyzing high-dimensional data," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 54–61, 2010.

[17] R. Duo, A. Ganglia, and D. Chen, "Spread neighbor clustering for large data sets," *Computer Engineering*, vol. 36, no. 23, pp. 22–24, 2010.

[18] E. Ca and Y. Chou, "Algorithm design and implementation based on graphs," *Computer Engineering*, vol. 38, no. 24, pp. 14–16, 2012.

[19] D. Y. Ha, "The new clustering algorithm design applied to large data sets," *Guiyang Teachers College Journal*, vol. 28, no. 1, pp. 67–71, 2011.

[20] H. Wang, "Clustering algorithm based on gird research," Tech. Rep., Nanjing Normal Unicameral, Nanjing, China, 2011.

[21] B. Chi, "k-means research and improvement of the clustering algorithm," Tech. Rep. 4, 2012.

[22] A. Chou and Y. Yu, "k-means clustering algorithm," *Computer Technology and Development*, vol. 21, no. 2, pp. 62–65, 2011.

[23] B. L. Han, K. Wang, G. F. Ganglia et al., "A improved initial clustering center selection algorithm based on k-means," *Computer Engineering and Application*, vol. 46, no. 17, pp. 151–152, 2010.

[24] W. Chao, H. F. Ma, L. Yang Cu et al., "The parallel k-means clustering algorithm design and research based on cloud computing platform Hadoop," *Computer Science*, vol. 10, no. 38, pp. 166–167, 2011.

[25] R. L. Ferreira Cordeiro, C. Traina Jr., A. J. Machado Traina, J. López, U. Kang, and C. Faloutsos, "Clustering very large multi-dimensional datasets with MapReduce," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*, pp. 690–698, New York, NY, USA, 2011.

[26] L. Zhang, Y. Chen, and J. Zhang, "A k-means algorithm based on density," *Application Research on Computer*, vol. 28, no. 11, pp. 4071–4085, 2011.

[27] M. X. Duan, "Clara clustering algorithm combined with the improvement," *Computer Engineer and Application*, vol. 46, no. 22, pp. 210–212, 2010.

[28] K. H. Tang, L. Zhong, L. Li, and G. Yang, "Urban tunnel clean up the dirtydata based on clara optimization clustering," *Journal of Applied Sciences*, vol. 13, pp. 1980–1983, 2013.

[29] M. H. Zhang, "Analysis and study of data mining algorithm based on Hadoop," Tech. Rep. 3, 2012.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 913046, 14 pages
http://dx.doi.org/10.1155/2014/913046

Research Article

MWAHCA: A Multimedia Wireless Ad Hoc Cluster Architecture

Juan R. Diaz, Jaime Lloret, Jose M. Jimenez, and Sandra Sendra

Universidad Politécnica de Valencia, Camino Vera, s/n 46022 Valencia, Spain

Correspondence should be addressed to Jaime Lloret; jlloret@dcom.upv.es

Received 27 October 2013; Accepted 15 January 2014; Published 11 March 2014

Academic Editors: W. Sun and J. Zhou

Wireless Ad hoc networks provide a flexible and adaptable infrastructure to transport data over a great variety of environments. Recently, real-time audio and video data transmission has been increased due to the appearance of many multimedia applications. One of the major challenges is to ensure the quality of multimedia streams when they have passed through a wireless ad hoc network. It requires adapting the network architecture to the multimedia QoS requirements. In this paper we propose a new architecture to organize and manage cluster-based ad hoc networks in order to provide multimedia streams. Proposed architecture adapts the network wireless topology in order to improve the quality of audio and video transmissions. In order to achieve this goal, the architecture uses some information such as each node's capacity and the QoS parameters (bandwidth, delay, jitter, and packet loss). The architecture splits the network into clusters which are specialized in specific multimedia traffic. The real system performance study provided at the end of the paper will demonstrate the feasibility of the proposal.

1. Introduction

Wireless ad hoc networks are formed by a set of distributed nodes inside a limited geographic area. Generally, the physical topology is given by the density, the placement of the nodes, and their mobility along the time in the area. Wireless nodes are able to communicate without a wired infrastructure by using an ad hoc communication. The way they are organized in the physical topology and the way they communicate with other nodes determine the logical network topology and the routes followed by the data in the ad hoc network [1]. Once the logical topology is built, end nodes have the needed transport infrastructure for the information exchange. This infrastructure allows monitoring, remote management, data gathering, and so forth. There could be a huge number of logical topologies depending on the criteria and algorithms followed by the nodes when the network is built [2]. The architecture should take into account the network traffic features with the objective of optimizing the performance and efficiency of the transmissions.

Cluster-based architectures are very common topologies in ad hoc networks. These architectures organize the nodes in small groups of nodes that work independently and autonomously [3]. Nodes in each cluster can communicate with other nodes establishing neighborhoods or through the neighborhoods of their neighbors. At the same time, a cluster can communicate with other clusters or with external networks through a higher hierarchical level, which is shared by all clusters [4]. We can find in the related literature many wireless ad hoc clustering algorithms and schemes. The network topology will depend on the neighbor selection criteria when building the cluster. Moreover, the motion of the nodes change the topology constantly, which increases the number of management messages [5]. Neighbor node selection has been widely researched in many network structures [6]. Depending on the purpose nodes can be organized taking into account the geographic distance [4], decrease the energy consumption [7], decrease convergence time, maximize the whole available bandwidth for data transmission, fault tolerance, or load distribution purposes, and so forth.

Due to the fast growth and development of wireless technologies, wireless ad hoc networks are becoming more and more common between people. There are many emergent applications and new environments to be used. New tendencies are user-oriented and service-oriented wireless ad hoc networks [8], where one of the main ones is the real-time audio and video streaming (sometimes their use in other types of networks is very complex or too expensive).

In order to deliver real time data traffic inside a wireless ad hoc network, it is essential to quantify and measure the network quality of service (QoS) parameters: delay, jitter, lost packets, and guaranteed bandwidth [9]. QoS should be analyzed from two complementary points of view. On one hand, it should be taken into account that for each multimedia stream, QoS parameters should be kept inside a range along the time. On the other hand, different multimedia streams (depending on whether audio or video, or depending on the used codec) will have different optimum QoS values and ranges.

Although there is some interest to manage the traffic flow in ad hoc networks [10], many researchers take into account QoS parameters in their proposals [11]. But existing cluster systems do not use QoS parameters as criteria to build the logical topology. So, the obtained QoS values could be outside of the required range. Therefore, the quality of the multimedia communications will not be guaranteed and the quality of the experience (QoE) of the end users will be affected. Moreover, QoS parameters should be monitored continuously because their values may vary considerably due to network changes (node joining and leavings, new concurrent audio, video, and data streams).

In this paper we propose a new architecture to build wireless ad hoc clusters based on QoS parameters criteria. The architecture allows structuring the network taking into account the features of the multimedia streaming, the number of streams delivered through the network, and the capacity of the nodes belonging to the network. The objective is to offer a guaranteed and differentiated service for each multimedia stream in order to optimize the communication between nodes and taking full advantage of the available bandwidth, but guaranteeing the required delay, jitter, and packet loss levels.

The remainder of the paper is structured as follows. Section 2 reviews the papers we have found related to the multimedia streaming in ad hoc networks. In Section 3, we detail the proposed architecture, introduce the multimedia init profile (MIP) concept, and describe the most important processes of the architecture. The state machine of the architecture is described in Section 4. Section 5 shows the performance study of the architecture in order to validate our proposal. Finally, Section 6 draws the conclusions and future work.

2. Related Works

In the recent years, the research on multimedia distribution over ad hoc networks has been increased. It has happened because of the improvement of the hardware capabilities and the appearance of new multimedia services. In order to support real-time multimedia applications, we should mainly take into account QoS constraints.

In [12], Zhang et al. propose several improvements over MAC protocols to solve some of the main problems as the stringent quality of service (QoS) requirements of video traffic, the limited wireless channel bandwidth, and the broadcast nature of wireless medium in ad hoc networks. Authors proposed two conflict avoidance strategies for reservation and contention interleaved wireless systems. With a dual buffer, the video packets being transmitted using contention or reservation-based channel access are separated and stored in two buffers. Authors also developed analytical models considering the interactions of reservation and contention periods. The performance tests were focused on the contention-based access. Simulation results showed that the backoff strategy can achieve higher throughput when the number of reserved periods in each superframe is large. Authors also checked the results of their proposal when transferring MPEG-4 video streams. The proposed hybrid approach with the two buffering architectures provided a considerably better performance, due to the higher reservation utilization and lower contention level.

Mehta and Narmawala used a video traffic model to generate video traffic frames in [13]. They observed that network coding performs well in lossy wireless ad hoc networks in both multicast and broadcast scenarios. Even in wireless ad hoc networks with low density of nodes network coding performs well using their multicopy packet transmission scheme. In their work, each sender node encodes the packet using a variant of network coding, which is random linear network coding (RLNC) with multigeneration mixing (MGM), with the aim to provide more protection to I (intraframe of MPEG 4 video traffic) frames in order to minimize the multiplicative loss by incurring slight delay in transmission. Mixing different types of packets increases the packet delivery fraction and reduces packet drop rate and block delay of multimedia transmission over wireless ad hoc networks.

Since the main weight of maintaining a fast multimedia delivery and an optimum path for the streams in an ad hoc network is carried out by the routing protocols [14], most authors have studied the routing protocols in ad hoc networks in order to know their features and which ones are the most appropriate to provide QoS [15]. Moreover, some authors have developed QoS-aware routing protocols for ad hoc networks such as the following ones.

Al Turki and Mehmood in [16] studied video streaming applications over ad hoc networks and analyzed the results obtained through simulations using the OPNET software. They also surveyed the main challenges in ad hoc network research and reviewed the QoS literature for ad hoc networks. They evaluated performance of video streaming applications over ad hoc networks by simulating few scenarios with 5 different routing protocols. The results show that it is possible to support multimedia applications over medium sized networks.

Jamali et al. demonstrated in [17] that ad hoc networks can support video streaming. In order to do it, they analyzed some routing protocols through simulations in OPNET environment in terms of multimedia and real-time application and QoS. In their study, they analyze AODV, DSR, OLSR, TORA, and GRP for multimedia streaming. Their results demonstrate that ad hoc networks can have good video streaming quality. They conclude the paper stating that designing a multimedia ad hoc network is difficult because

of the higher QoS requirement and the kind of network topology.

Abdrabou and Zhuang presented in [18] a model-based quality-of-service (QoS) routing scheme for IEEE 802.11 ad hoc networks. This proposal is based on a cross-layer design approach. The scheme proposed selects the routes based on a geographical on-demand ad hoc routing protocol and checks the availability of the network resources by using traffic source and link-layer channel modeling. The system also considers the IEEE 802.11 features and the node interactions. The protocol checks if the selected route is able to admit traffic flow without affecting other flows already in service. The simulation results show that the proposal is efficient in resource utilization while satisfying the delay bound probabilistically with a low overhead.

Kandris et al. present in [19] a dual scheme based on the combined use of an energy aware hierarchical routing protocol with an intelligent video packet scheduling algorithm for efficient video communication, which aims at both energy saving and high QoS attainment. PEMuR adopts a routing protocol which is able to select the most energy efficient routing paths while it manages the network load according to the energy residues of the nodes and prevents useless data transmissions through the proposed use of an energy threshold. In addition, this protocol is able to reduce the video transmission rate with the minimum possible increase of distortion. The simulations performed by authors showed that this proposal prolongs the node lifetime. It also enhances the metric of network performance in the case of nodes with nonuniform energy distribution while maintaining high levels of the perceived video quality (PSNR).

Taing et al. [20] propose a routing scheme for multimedia services, which selects the shortest path by using power level. This proposed scheme selects the shortest path for multimedia traffic by applying larger power level because the delay is sensitive to such kind of traffic. Moreover, for nonreal time traffic, this algorithm uses smaller power level longer path for non-real time. They conclude that their proposal provides the lower mean number of hops for multimedia traffic than the mean number of hops for non-real-time traffic. As a result, the transmission delay of multimedia traffic can be decreased. They also show that its proposal scheme can provide higher throughput for multimedia traffic.

In [21] we presented multimedia-oriented architecture and protocol for wireless ad hoc networks. This proposal takes into account the multimedia services offered by the nodes in the wireless ad hoc network in order to select the best multimedia service provider node at application layer. We designed a new protocol and the appropriate decision algorithms to provide the best multimedia QoE and QoS to the end users participating in the ad hoc network.

We have not found in the related literature any cluster-based ad hoc architecture focused on multimedia streaming.

3. Proposed Cluster-Based Architecture

In this section we detail the proposed architecture to build wireless ad hoc clusters for multimedia streaming with service guaranteed. First we will describe the initial state (Init State) of the architecture, which will be used as a starting point for our protocol. Then, we will define the multimedia init profile (MIP), which collects the multimedia information used by the network nodes. Then, we will detail the system processes for the proper operation of the architecture. Finally, the routing algorithm to estimate the most convenient paths for multimedia communications through the cluster will be explained.

3.1. Multimedia Init Profile (MIP). Let multimedia init profile (MIP) be a data structure which represents the multimedia streams delivered through an ad hoc cluster from a source to a destination node. MIP contains a single array with all the information needed to decide the route for each stream. It contains the information of the QoS requisites that should be guaranteed by each cluster node to transmit each type of multimedia stream. Network topological features and the capacity of the nodes in the ad hoc network will determine the most adequate number of nodes and the properties of the MIP available to be selected as an initial configuration by a node. The network topological features are the density of the nodes, their location, space distribution (these data are obtained by using GPS data), obstacles, and possible signal interferences (estimated by using geographical maps or building maps if it is indoors). Other features such as transmission power and coverage area could also be added in future works. From the multimedia streams we have added the type of multimedia stream (video, audio, or both), the used codec, and the QoS requisites (delay, jitter, lost packets, and bandwidth).

The cluster-based architecture uses MIP as a main feature to build the clusters. It groups in the same cluster the nodes with the same MIP under the coverage area. We can adapt the definition of the MIP to each particular case. Moreover, we can define several MIPs for a single cluster. For example, a network with low nodes density in the clusters can have both MIP one for audio transmission and the other for video transmission, but in a network with high density of nodes dedicated only for the video transmission using many types of codecs, they can use several MIPs that will allow them to transmit the streams with different codecs into different clusters. In order to simplify our explanation and the system deployment details, we will assume that all nodes in the ad hoc multimedia cluster share the same MIP, but it can be extended to several MIPs or to MIPs with range of values.

The number of defined MIPs, available to be selected in the system startup, should be wide enough to cover accurately the most common multimedia streams, but it should not be too much to facilitate new node joining and avoid having too many different clusters.

MIP has the following parameters inside: maximum bandwidth, minimum bandwidth, maximum delay, maximum jitter, maximum packet loss, and maximum number of hops. Each MIP has a one byte long hexadecimal code called HCode and an alphanumeric code called ACode, with variable size.

Table 1: Defined MIP list for the practical implementation of the architecture.

MIP	ACode	HCode	MinBW	MaxBW	MaxDelay	MaxJitter	MaxHops	MaxLoss
Audio 32 K	A1	0x01	8 Kbps	32 Kbps	50 ms	20 ms	6	0.5
Audio 64 K	A2	0x02	8 Kbps	64 Kbps	100 ms	40 ms	6	0.5
Audio 128 K	A3	0x03	16 Kbps	128 Kbps	150 ms	40 ms	5	0.5
Audio HQ	A4	0x04	32 Kbps	1024 Kbps	150 ms	40 ms	4	0.5
Video 256 K	V1	0x41	64 Kbps	256 Kbps	100 ms	20 ms	4	1
Video 1024 K	V2	0x42	128 Kbps	1024 Kbps	150 ms	40 ms	4	1
Video 2048 K	V3	0x43	256 Kbps	2048 Kbps	200 ms	40 ms	3	1
Video HQ	V4	0x44	1024 Kbps	20 Mbps	200 ms	40 ms	2	1
Default	Default	0xFF	56 Kbps	1 Mbps	200 ms	40 ms	4	1

(i) Maximum bandwidth (MaxBW): this parameter establishes the maximum bandwidth spent by all the multimedia flows processed by the node at the same time. This value represents the whole bandwidth provided by the node for multimedia transmissions with guaranteed service.

(ii) Minimum bandWidth (MinBW): this value describes the minimum bandwidth required to transmit just one multimedia flow. It represents the bandwidth requirements specified by a multimedia codec or a group of multimedia codecs with similar requirements for a single multimedia communication.

(iii) Maximum delay (MaxDelay): this parameter allows knowing the maximum latency value allowed for a multimedia packet across the cluster between the source node and the target node. It represents the maximum guaranteed quality for audio or video transmissions inside a cluster.

(iv) Maximum jitter (MaxJitter): this value indicates the maximum jitter that is considered tolerable for multimedia transmissions inside the cluster.

(v) Maximum packet loss (MaxLoss): this is the maximum percentage of acceptable lost packets. When this value is exceeded, the target node, the node at the end of the multimedia path, breaks the multimedia transmission and notifies the rest of the nodes in the path because the quality of communication cannot be guaranteed. It is calculated to each one individually. This parameter, together with the maximum delay and maximum jitter, represents the quality of service provided for a real-time communication.

(vi) Maximum hops (MaxHops): this value provides the maximum diameter of the cluster and it can be estimated through the routing table. When a new node joins the network, it will start the connection process trying to connect with a cluster using the same MIP. The system uses the MaxHops value to check that the cluster dimension is always kept under acceptable values for multimedia traffic transmission guaranteeing enough quality of service. A node belonging to a cluster will not accept new joining nodes if it has reached MaxHops.

Table 1 details the list of MIPs set in advance for audio and video transmission. We have tagged an alphanumeric code to each MIP, called ACode, in order to let the node set each MIP and use it in the protocol messages. It is easy to use by the user or by the network administrator when the initial configuration of the node is set. We have also associated a hexadecimal code called HCode, which has a byte size, which will be used in the protocol header when information is exchanged between nodes in the same cluster. Table 1 also shows the QoS parameters associated with each MIP maximum and minimum bandwidth, delay and jitter, the maximum number of hops (MaxHops), and the cluster diameter.

Figure 1 shows an example of a multimedia ad hoc network using a cluster-based architecture. All nodes share an area and all are reachable by the other nodes because they are under their wireless coverage area. They are distributed logically in clusters that are specialized in the transmission of similar multimedia streams with similar audio and video QoS parameters. Figure 1 shows how nodes are grouped in 4 clusters, 2 for audio transmission and 2 for video transmission. One video and one audio cluster are dedicated to the transmission of codecs with low bandwidth requirements; the other video and audio cluster are dedicated to the transmission of codecs with higher bandwidth requirements. The head node of each cluster can communicate with head nodes of other clusters in a higher hierarchical level that allows the communication between clusters.

Figure 2 shows the elements of the proposed topology and the relationship between them. The architecture has three levels of operation: hardware infrastructure, logic management, and admin interface. The hardware infrastructure level is formed by different types of nodes (regular cluster nodes, gateway nodes, and head nodes), which build the physical topology, and clusters, which build the logical topology. When new nodes join the network, they have the regular cluster node role. A regular cluster node cannot communicate with nodes from other clusters or with external devices, but with nodes of the same cluster. When a new regular cluster node tries to join the network, it searches nodes under its coverage area. When it receives replies from nodes having the same MIP, the developed protocol will let them exchange information in order to build clusters following the proposed architecture. Each node in the ad hoc network, despite its

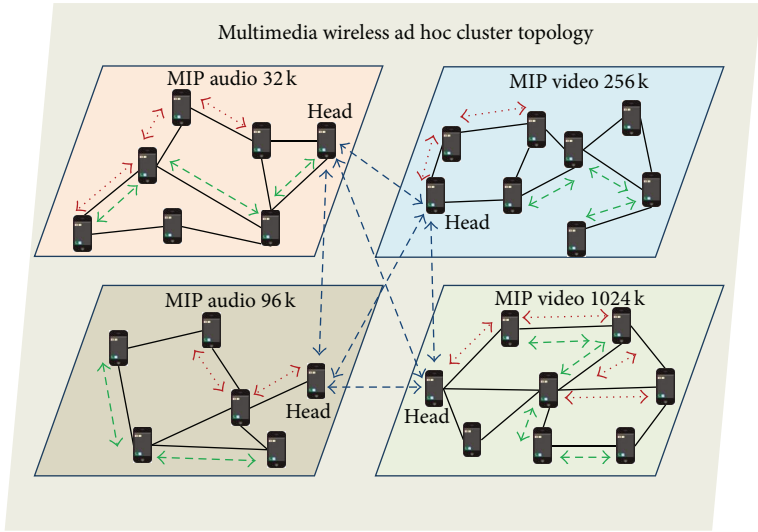


Figure 1: Multimedia ad hoc cluster topology.

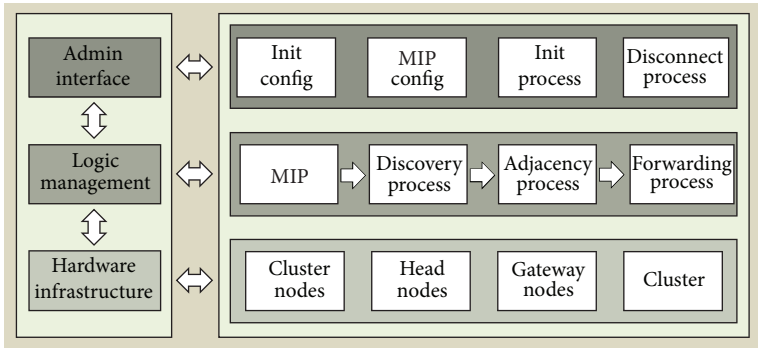


Figure 2: Multimedia ad hoc wireless network architecture elements.

role, can only belong to a single cluster. First node will be the head node, and it will be responsible for locating and communicating with the head nodes of the other clusters. Gateway nodes have two network interfaces. One interface will be used to connect with the nodes in the ad hoc network and the second interface will be used to connect two with an external network. A node can have both roles: head node and a gateway node.

Logic management level defines the elements of the protocol, which will be used to manage the hardware infrastructure elements by gathering the information obtained from the admin interface level. MIP will be used to group the nodes in clusters and assign the cluster to the new nodes. A new node can only be neighbor of a node with the same MIP. All nodes in the same MIP will always have similar features. Figure 3 shows the internal organization of a cluster. It is formed by a cluster node and a gateway node that use the same MIP. Multimedia streams can be initiated or ended in external multimedia networks like VoIP, IPTV, or ISPs. The connection to external networks is always made by gateway node.

Logic management level also defines the logical processes performed by the nodes automatically as a function of their states and the events given in the network: discovery process, adjacency process, and forwarding process. When a node starts up correctly, it executes the discovery process and seeks other nodes with the same MIP under its wireless coverage area. When it finds other nodes, the adjacency process starts in order to establish a neighborhood between both nodes. The process is repeated every time it finds a new node with the same MIP, allowing the system to build the network clusters. When a cluster is built, it has the capability and resources to retransmit the multimedia streams whose features meet the MIP of the nodes of the cluster. Forwarding process is started when a node starts a new stream query. The query can be started inside the cluster or can be started by another cluster node or by an external network (in this case the query comes from a gateway node). Forwarding process uses the routing algorithm to know the route that should follow the multimedia stream inside the cluster and requests the resource reservation to each node of the route. This process is responsible for establishing the connection between nodes and guaranteeing the QoS required by the MIP during the communication.

The third level is the admin interface level, which allows the interaction between the user and the device. By using

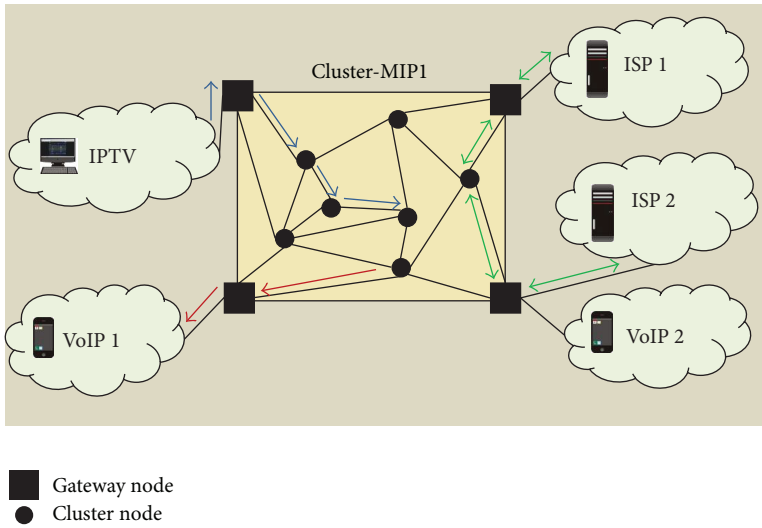


Figure 3: MIP multimedia cluster.

a graphical user interface (GUI), node init configuration can be modified, including IP addressing, the MIP to be used by the node, and in case of a gateway node, the communication between the ad hoc network and the external network. Admin interface level is used to manually control the init process and the disconnect process. The user can initialize the node and join or disconnect the node from the ad hoc network. The node can only be configured before the init process starts, so in order to make any change, it is necessary to stop the node, through the disconnect process, perform the appropriate changes, and restart the system with the init process.

3.2. System Processes. In order to design the architecture, we propose four basic processes, which correspond to the basic actions of a node inside the ad hoc network. Each process is associated with a set of states and transitions that will be detailed later when the system state machine is explained. Figure 4 shows the relationship between the processes of the system. Init/disconnect process is the start and end process of the system. It is the only process that requires the user intervention for executing it.

Init process starts the node when the user (or the system) has selected the appropriate MIP. Disconnect process allows the user to leave the network safely (or to restart with a new MIP configuration). Init/disconnect process brings the system to the discovery process, where the node will try to find the nodes in the network with the same MIP. When the node finds another node with the same MIP, and the cluster does not arrive to the maximum number of hops defined by the MIP, the system starts the adjacency process, in which both nodes exchange their network information and lets the new node join the cluster. When a node, belonging to a cluster, receives a query for multimedia stream transmission and checks that it is possible to guarantee MIP requirements, forward process is started.

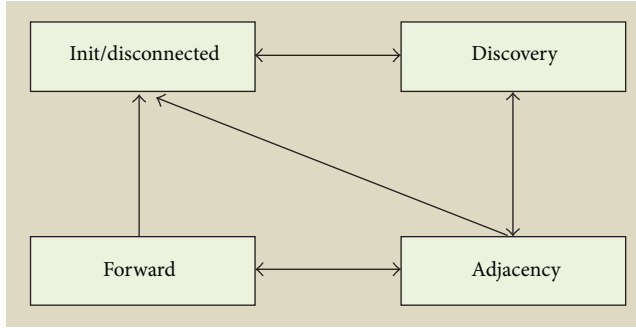


Figure 4: System processes of the multimedia wireless ad hoc cluster architecture.

System processes, with the states of each process, are described in detail next.

(1) Init/Disconnect Process. This process includes the subprocess executed by the node when it joins or leaves the network. This system process includes two states: init state and disconnect state. The node will be in the init state when it is running the init process, and it will be in the disconnect state when it is leaving the network. In the init process the node tries to access the physical network and obtain information about possible neighbors. The init process is divided into three different phases: MIP selection, unicast IP configuration, and group multicast IP configuration. At the first phase, the system allocates the node MIP according to characteristics and the available resources. The MIP of the node can be statically selected by the user, but there is MIP default profile, identified with the HCode value of 0xFF, if no MIP is selected. In the second phase of the init process, the IP configuration of the node is established, including unicast IP address, network mask, and gateway address. The use of a DNS server is optional and it is not required for the normal operation of the protocol. The IP configuration can

be manually configured by the user, or may be dynamically obtained by using IETF Zeroconf as defined in RFC 3927. In the next phase of the initialization process the node joins the multicast group matching its MIP. All nodes sharing the same MIP must be listening the same IP multicast group address; thus all nodes in the same multicast group belong to the same cluster. The range of IP multicast group addresses used by the system is 239.100.100.X/24. In this multicast address, the fourth and last byte matches the MIP HCode value. For instance, a node with the MIP 256 K video profile with ACode equal to "V1" and HCode value of 0x41 (decimal value 65) will join the IP multicast group 239.100.100.1.65. Using multicast addresses, each node will communicate exclusively with other nodes with the same profile, without interfering with other nodes in other clusters with different MIP. This system process is also in charge of making the node leave the ad hoc cluster it belongs to. This part of the process occurs when the node is in the disconnect state. The system can reach this state from the other 3 system processes, since the node can leave the system at any time regardless of the assigned state. The node uses the multicast group address of the cluster to notify its neighbors that it is leaving the cluster. Then the neighbors can update their status tables in order to reorganize their forwarding process.

(2) Discovery Process. Upon completion of the init process, the node is ready to make the transition to the discovery process. In this process, the node will try to detect the presence of a cluster with the same MIP in order to join it. There are two possible states in the discovery process: discovering state and stand-alone state. When the node accesses the discovery process for the first time, then the system changes to discovering state. This is an active state; while the node stays in this state, it keeps sending discovery messages to the IP multicast group of its MIP. The node waits 60 seconds for replies after sending each discovery message. If no reply is received, during this time interval, another discovery message is sent. Discovering state has a maximum duration of three minutes. If one or more reply messages are received during these 60 seconds after the discovery message is sent, the system changes to the adjacency process. After sending three discovery messages without any result, it changes to the stand-alone state, but still remains in the discovery process in passive mode; that is, the discovery process does not keep sending periodic discovery messages, but the node remains listening for new nodes trying to join the network. If a discovery reply from another node is received, it first checks if the MaxHops of the MIP is not exceeded. If MaxHops is not exceeded, the adjacency process starts.

(3) Adjacency Process. This process starts when the above discovery process has detected the presence of one or more nodes with the same MIP. The adjacency process includes join state, associated state, and established state. If a node in the discovery process receives replies from two or more nodes belonging to the same cluster, it will try to establish the adjacency with all detected nodes. If a node receives replies from two or more nodes belonging to different clusters, but all of them are using the same MIP, the system will choose the best cluster and reject other options. The best cluster choice is made by a three-step algorithm, which uses the information included in the discovery reply messages. First, the node estimates the diameter of the cluster if this neighbor is selected. The best selection is the smallest diameter. In case of a tie, the second step comes. The node checks the number of adjacencies of that neighbor and selects the node with the minimum number of established adjacencies in order to distribute the load between different clusters. Finally, if there is a tie in the previous step, it selects the source node of the first received reply. Once the node selects the best candidate, it sends a join message to the selected cluster nodes. When a reply message is received, the node changes to the join state. Then, the new node will receive the information about the cluster characteristics and the topology structure. When the node has the whole information about the cluster, then it changes to the associated state. Finally, the new node sends the information about its resources and availability to the other nodes in the cluster. When all nodes inside the cluster have the same information, the cluster has converged. Then, the new node changes to the established state. In this state, the node is fully integrated in the cluster and it is ready for multimedia transmissions. A node remains in the established state indefinitely until it receives a multimedia transmission request, until the user invokes the disconnect process or until the adjacency is broken. When there is a multimedia request, the forwarding process starts.

(4) Forwarding Process. This process is in charge of the multimedia traffic transmission. Inside the forwarding process we can find two different states: queued state and forwarding state. The forwarding state can be initiated only when the node has successfully completed at least one valid adjacency with a node. Multimedia requests could be originated by the node, for example, a request for audio or video communication performed by the user interface, other adjacent nodes, or a gateway node from external networks. When a multimedia request is processed, regardless of the origin, available bandwidth resources at the node are checked. If the node has enough available resources, the node changes to the forwarding state, makes a temporary reservation of resources for the transfer, and notifies the origin node that it is ready for transmission. When all nodes in the path from the source to the target node confirm they have made the resource reservation, the source sends a confirmation message to the nodes in the path to allocate permanently the reserved resources for the multimedia flow. Then, the nodes change their state to the forwarding state and the multimedia transmission takes place. When a node in the multimedia flow path does not have enough resources for the multimedia connection request (e.g., there is not enough available bandwidth) the node changes to the queued state. The node in queued state informs the origin of the multimedia request that it cannot process this request, but it will keep it queued. Then, the source node can wait until the bandwidth resources are released or, if there is some alternative route provided by the routing algorithm, it can cancel the current request and try to establish a new communication using a new path.

When the forwarding process for multimedia transmission ends successfully, the node changes to the established state.

3.3. Routing Algorithm. Source node (SN) is the node belonging to a cluster that requests a multimedia connection. The request can be performed by a user through the graphical user interface or from external networks (in this case the source node is a gateway node). Target node (TN) is the destination node of the multimedia connection, which will receive and process the multimedia streams. It can be a regular node or a gateway node.

Every node has its neighbors table, which is built and maintained through the adjacency processes, and the cluster topology database which is built using the topology information received from its neighbors. When a SN starts a multimedia transmission, the routing algorithm uses the multimedia streaming bandwidth requirements and the topology information of the nodes inside the cluster. The estimations to determine the route, including the nodes that will forward the multimedia streams inside the cluster, are performed by the SN. The routing algorithm selects as the first hop the node that is the closest (in terms of number of hops) to the TN. When there is a tie, the node with the oldest adjacency will be selected. Selected node is called forward node (FN). FN will estimate the path to the TN using the same process, so it will obtain the second hop in the route to the TN. This process is repeated till TN is achieved. This information is saved in the MEDIA_ROUTE parameter, which will be used by the SN in the resource reservation request in order to guarantee the transmission quality. The resource reservation request is firstly sent to the first FN, which will check if it has enough available resources. If it meets the requirements, it uses the information included in the MEDIA_ROUTE parameter of the message to forward it to the next hop. This process is repeated in each node of the route till it reaches the TN. If the TN receives the request, it means that the cluster has enough resources to perform the multimedia communication meeting MIP requisites, so it replies with a confirmation message that will follow the same route in order to confirm the resource reservation in each node. When the confirmation message reaches SN the multimedia communication starts.

In case of not having enough resources when a node belonging to a route does not have enough resources, the request is included in the queue of this node till it has enough available resources. If the SN receives neither a confirmation reply nor a queue request in 30 seconds (e.g., because a node left the cluster suddenly), it sends a message containing the route verification, which uses MEDIA_ROUTE parameter, to the TN.

Nodes keep updated their neighbor table by sending keepalive messages to all their neighbors and waiting for a reply in less than 10 seconds. If during this process a node detects a topology change, it will send an update message to the rest of the nodes in the cluster to let them update their tables. Both, SN and TN, are able to stop the multimedia streams by closing the communication. They will notify the rest of the nodes of the route that they have to liberate the reserved resources.

4. Finite-State Machine

Figure 5 shows the system finite-state machine. We can see its different states and the transitions between states. In this section we describe each state of the system and the conditions and events that will make the node change from one state to another inside a process.

The processes included in Figure 5 are the following ones.

(i) Init State. This is the initial state of the node during the init process. There are two possible ways to access the init state: first, when the node starts for the first time and, second, when the node is rebooting. There is only one possible transition from the init state to the discovering state. This transition is made when the node has initialized correctly; that is, when the whole information has been obtained from the MIP, the IP settings are correct, and the network connection is active. There are several events that may cause the init process to fail: an IP address conflict with another node in the wireless network, the wireless network connection being not enabled, or when it is not possible to join the IP multicast group. When an error event happens in the boot process, the system remains inactive in the init state for 120 seconds before it tries again to initialize the system.

(ii) Discovering State. In this state the node has not yet established any adjacency and it is looking for a neighbor by sending discovery messages. The first time the system makes the transition to the discovering state is when, being at the init state, the system initialization has been completed successfully. The node can also change to the discovering state from the stand-alone state. It happens when the system has remained in the stand-alone state for 12 minutes and no discovery message has been received from other nodes. Finally, there could be a transition to the discovering state from the join state when the adjacency fails in the adjacency process. While the node remains in the discovering state a discovery message is sent every 60 seconds to the IP multicast address of the MIP. The maximum number of discovery messages is set to 3. From the discovering state there is a transition to the join state when the node receives a discovery confirmation message. The waiting time for discovery confirmation messages is set to 60 seconds. Upon finishing 60 seconds the node gathers all received messages and processes them as explained before. Then, there is a transition to the join state. After three times of 60 seconds without receiving any discovery message, a transition to the stand-alone state is made.

(iii) Stand-Alone State. The node reaches this state when the discovery process has not found any valid node, and thus cluster, to join. Then, the node remains isolated from the remaining nodes and it does not establish any adjacency. There are two possible ways to arrive to the stand-alone state: first, when the node is in the discovering state, as described above, and, second, from the established state (it happens

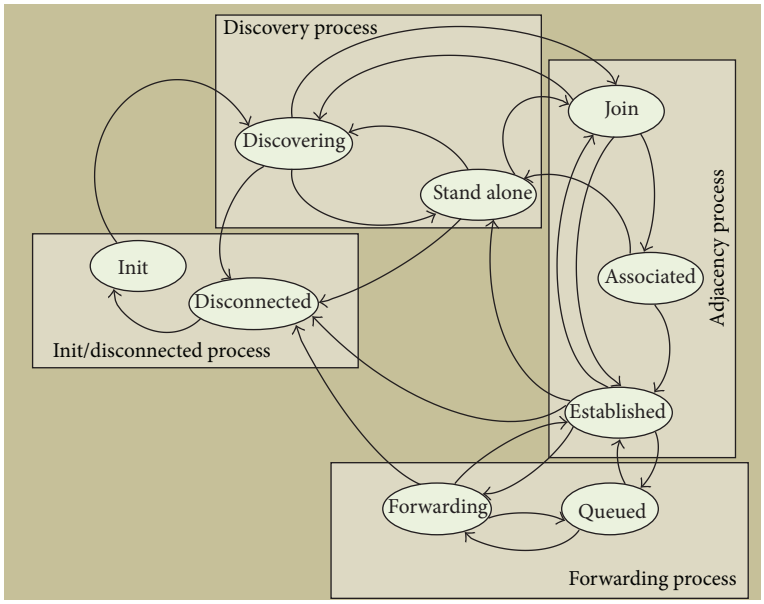


Figure 5: Finite-state machine for multimedia wireless ad hoc cluster architecture.

when the node has just one established adjacency and it is broken because the neighbor node has left the network). A node assumes that its neighbor is down when it receives a leaving notification or when it has not received any response message during 10 seconds (e.g., after a keepalive message has been sent or in the path verification subprocess that takes place in the forwarding process). There are three admissible transitions from the stand-alone state: discovering state, join state, and disconnecting state. If the node receives a valid discovery message from another node while it stays at the stand-alone state, then the system replies with a discovery confirmation message in order to offer a new adjacency. Then, if it receives a join request message, it will answer with a join acknowledgment message and the system automatically changes the status to join state and the adjacency process starts. If the node remains in the stand-alone state for 12 minutes and no message has been received from another node, it makes a transition to the discovering state. Then, the discovering process starts again an active search for neighbor nodes. Finally, through the intervention of the user, the system can make a transition to the disconnecting state in order to close the connection and leave the network or to restart because some of the values of the initialization process have been changed.

(iv) Join State. This state is the starting point of the adjacency process. The nodes have not yet shared any information from the neighbor tables but they want to build a new adjacency with the discovered node because it has the same MIP. The system can achieve the join state from three different states: discovering state, stand-alone state, and established state. A transition from the discovering state is made when the node has received at least a confirmation of the discovery message and the acknowledge join message. The transition from the stand-alone state takes place when the node has received a new discovery message and a join request message. Finally, the transition from the established state to join state occurs when the node has already one or more established adjacencies and it receives a new discovery message from a new node requesting a new adjacency. The regular next step from the join state is the associated state. It occurs when both nodes have exchanged the whole information in its neighbor tables and the routing database. If the transition to the associated state cannot be completed, because the received information is inconsistent or incomplete, the system will make a transition to the discovering state (if it is the first adjacency) or to the established state (if there are other established adjacencies). A node can establish adjacencies with two or more nodes.

(v) Associated State. This is a transient state. Both nodes have exchanged the neighbor tables and the routing database, but they have not yet confirmed the integration of the new node at the cluster. This state is reached from the join state as described above. From the associated state the node can make two transitions: towards the established state and to the stand-alone state. The transition to the established state will occur when the new node receives the cluster acceptance notification. The transition to the stand-alone state takes place when the node is not accepted and there are no other established adjacencies.

(vi) Established State. At this state the node has established at least a valid adjacency and it is integrated inside the cluster. This is the regular operation mode for a cluster node when no multimedia traffic stream is transmitted through the cluster. In the established state, the node holds a neighbor table with the information about the neighbors and routing database with the cluster topology. The node needs this information to reach other nodes in the cluster and to calculate the best

route based on hop count and multimedia available resources. The established state can be activated by a transition from the associated state when a new adjacency is established, from the join state when the an adjacency fails, but there are other active adjacencies in the node, from the forwarding state when a multimedia communication using that node finishes or from the queued state when the resource request remaining in queue is canceled. Possible transitions that can be made from the established state are to the join state, when the node receives a new discovery message, to the stand-alone state, when the last established adjacency in the node is broken, to the forwarding state, when a multimedia transmission request is received and there are enough resources to process it, to the queued state when the node receives a request for multimedia transmission but there are not enough resources to process it at that time, and, finally, to the disconnecting state due to the user intervention when he/she wants to disconnect or reboot the node.

(vii) Forwarding State. In this state the node is processing and transmitting multimedia packets for every received resource reservation request. This state is reached from the established state when the first request for resource reservation is received and completed successfully or, from queued state, when a queued resource request can be satisfied because the node has released enough resources. When the last active multimedia stream on the node finishes its transmission, the system makes a transition to the established state and it remains listening to new requests. If the node receives a new resource reservation request and the needed resources are not available, then the system changes to the queued state. Finally, if the user wants to abort the active multimedia connections in the node in order to reboot or to close the node, it makes a transition to the disconnecting state, but first it notifies it to the source node and target node of each active communication.

(viii) Queued State. The system uses this state when a node is working properly inside the cluster and receives a new multimedia request but it cannot be processed because it has exhausted their bandwidth resources. Queued state can be reached through a transition from the established state or the forwarding state when it receives a new stream request. The node leaves the queued state when it has released enough resources to process the request and it makes a transition to the forwarding state or established state. If the resource request is canceled and there are other active multimedia streams on the node, then a transition to the forwarding state takes place. But if there are no other multimedia streams processed at the same time, then the transition is made to the established state. User can close or restart the node from the queued state making a transition to the disconnecting state.

(ix) Disconnecting State. The node is in this state when the system is shutting down or rebooting, for example, to update the values of its initial configuration, such as the MIP or the IP settings. The system can change to the disconnecting state by the user intervention from several states: discovering state, stand-alone state, established state, queued state, and forwarding state. When the system changes to the disconnecting state the established adjacencies are checked. If there are adjacencies, a notification message is sent to every neighbor in order to let them update their neighbor tables and forward the information to the other nodes in the cluster. If there are active multimedia transmissions, the node notifies the source node and the target node in order to let them cancel the transmission. If the node is restarting, a transition is made from the disconnecting state to the init state.

5. System Performance Study

When multimedia streams are sent through ad hoc wireless networks, the bandwidth and the logical topology characteristic requirements should be adjusted as a function of the type of traffic, audio or video, and the codec used for the transmission.

We have deployed our architecture with the aim of measuring the delay and jitter parameters when several multimedia streams in different wireless ad hoc cluster topology configurations are set up. Obtained results will allow us to validate our protocol and architecture proposal, which groups the nodes in clusters based on the MIP. Nodes are classified and clustered based on their capacity to support different types of multimedia streams. Because we wanted to avoid any dependence with the devices characteristics, we used the same hardware configuration for all devices. They had Intel Core 2 Quad Processor working at 2.50 GHz with 2 GB RAM. These devices were connected through a wireless interface, which used IEEE 802.11 g standard. The wireless channel used to perform our test bench was 2.412 MHz.

The parameters of the cluster topology, such as the diameter, are limited based on the type of multimedia stream that is going to be used. The protocol allows several simultaneous multimedia streams guaranteeing the required resources for each one of them in their respective cluster.

We have selected the most appropriated MIPs taking into account the most used video codec characteristics. With the objective of maintaining equilibrium between the flexibility of the options and maintaining a reduced number of profiles, we have defined 3 MIPs for video in our test bench. The values assigned to each MIP are based on our studies previously performed in [21]. Each node in our test bench has any of these MIPs configured before joining the network.

5.1. Codecs Comparison. In order to compare the multimedia stream behaviour, we have selected three video codecs using 600 Kbps, 1800 Kbps, and 3600 Kbps bandwidth consumption. They correspond to the MIPs V2, V3, and V4 in Table 1. We have analyzed their behaviour when they are being streamed over the same cluster topology and with the same experimental conditions, so the differences in the results are only caused by the codecs characteristics used by each one of the multimedia streams. Figure 6 shows the delay obtained when those three video codecs are streamed during 30 seconds. In order to provide a graphical representation, we compute the average delay of the last 20 received packets,

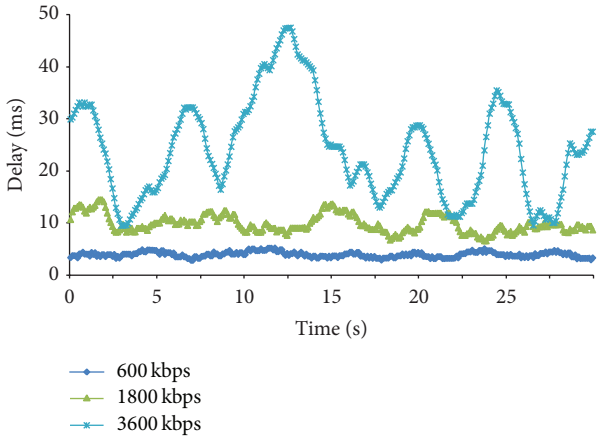


Figure 6: Delay of different streams using different codecs.

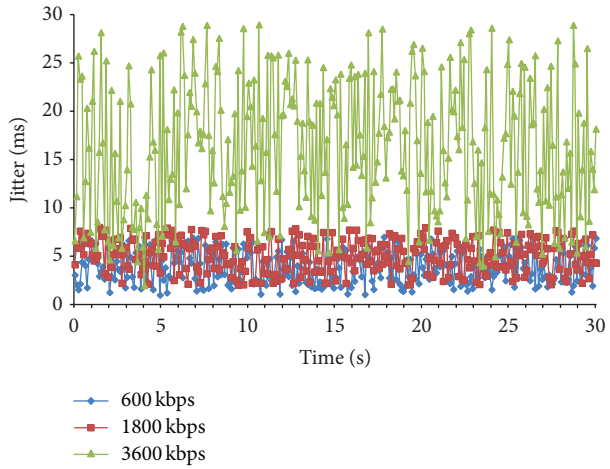


Figure 7: Jitter of different streams using different codecs.

estimating the value in 100 milliseconds intervals. With X_i being the delay of a single packet, our average delay is given by:

$$Y_i = \frac{\sum_{j=i}^{i+20} X_i}{20}. \quad (1)$$

Figure 6 provides Y as a function of the time. We can observe that 3600 Kbps has higher delay and has higher delay variation. 1800 Kbps and 600 Kbps are more stable. The one that provides lower values is 600 Kbps.

We have also performed a statistical analysis in order to interpret the results. In order to determine whether the observed differences in the delay are random or are caused by intrinsic characteristics of the codecs, we have defined the following null hypothesis H_0. There is no difference between the average delay obtained by the three codecs with bandwidths of 600 Kbps, 1800 Kbps, and 3600 Kbps. Table 2 shows the estimations performed for each codec. N is the number of samples, μ is the average score, σ is the standard deviation, Min is the minimum score, Max is the maximum score, and Conf. Int. is the confidence interval. In order to perform the statistical analysis, we have used a confidence level (α) of 0.01, with a confidence interval of 99%. The results show that the average delay value of each codec is outside of the confidence interval obtained for all codecs in all analyzed cases, so we can reject the null hypothesis with $P < 0.01$. The highest value has been obtained for 3600 Kbps in all cases, while the lowest value has been obtained for 600 Kbps in all cases. We can conclude that the behaviour of a multimedia stream when using the same cluster topology is different and depends on the bandwidth required by the codec, so we have to use a different treatment. We have also observed that lower bandwidth consumption provides lower delay values with higher confidence.

Figure 7 shows the results obtained when jitter is measured as a function of the used codec during 30 seconds. Jitter values are the average jitter values of the last received samples for the three multimedia streams using the same cluster topology. The three streams use the same number of hops (2 hops). We have observed that the jitter is quite higher for the codec with higher bandwidth consumption (3600 Kbps), while it remains quite stable and considerably lower for 1800 Kbps and 600 Kbps.

The statistical analysis provided in Table 3 shows that there is a significant difference between the codec with 3600 Kbps and the other codecs, obtaining μ and Max (ms) values 3 times higher. There is also a significant difference with a value of $\alpha = 1$, between 600 Kbps and 1800 Kbps codecs.

5.2. Hops Comparison. We performed the following test with the aim to show how a multimedia stream has different quality of service values as a function of the number of hops in the wireless ad hoc cluster. In order to perform this test we have selected a codec with an average of 600 Kbps and we have tested it in four topologies with different number of hops inside the cluster. Figure 8 shows the obtained delay as a function of the number of hops. We have observed that 1 and 2 hops do not increase the delay much, but it is considerably increased in three hops and hugely increased in 4 hops. Delay values are not increased proportionally with the number of hops.

We have also performed a statistical analysis based on the null hypothesis H_0. There is no difference in the delay average when a multimedia stream of 600 Kbps is being transmitted over several cluster ad hoc networks with diameters 1, 2, 3, and 4 hops. Table 4 details the estimated values for all obtained data. The estimated parameters are the same as the ones provided for Table 2. We have selected a confidence level (α) of 0.01, with a confidence interval of 99%. After obtaining these results we can discard the null hypothesis and affirm that the delay of a multimedia stream in a cluster ad hoc topology depends on the number of hops between the source node and the target node. We have also observed that the main difference is between 2 hops and 3 hops.

Figure 9 shows the measurements gathered for the jitter as a function of the number of hops in the cluster when 1800 Kbps multimedia stream is used. It shows a 30 seconds interval. We have observed that the highest values are

Table 2: Statistical values of the delay of different streams using different codecs.

Video codecs	Parameters						
	N	μ (ms)	σ (ms)	Min (ms)	Max (ms)	Conf. Int. (ms)	
600 Kbps	300	3.96	0.51	2.86	5.23	3.88	4.03
1800 Kbps	300	9.95	1.72	6.62	14.52	9.69	10.20
3600 Kbps	300	24.27	9.57	9.52	47.58	22.84	25.70

Table 3: Statistical values of the jitter of different streams using different codecs.

Video codecs	Pararameters						
	N	μ (ms)	σ (ms)	Min (ms)	Max (ms)	Conf. Int. (ms)	
600 Kbps	300	3.94	2.91	1	7	3.68	4.19
1800 Kbps	300	5.11	1.72	2	8	4.85	5.37
3600 Kbps	300	16.26	48.17	2	29	6.21	17.29

Table 4: Statistical values of the delay as a function of the diameter.

HOPS	Parameters						
	N	μ (ms)	σ (ms)	Min (ms)	Max (ms)	Conf. Int. (ms)	
1	100	3.96	0.51	2.86	5.23	3.88	4.03
2	100	4.40	0.52	3.02	6.21	4.32	4.47
3	100	96.66	5.60	85.02	106.11	95.83	97.47
4	100	165.38	72.162	158.08	1783.02	154.86	175.90

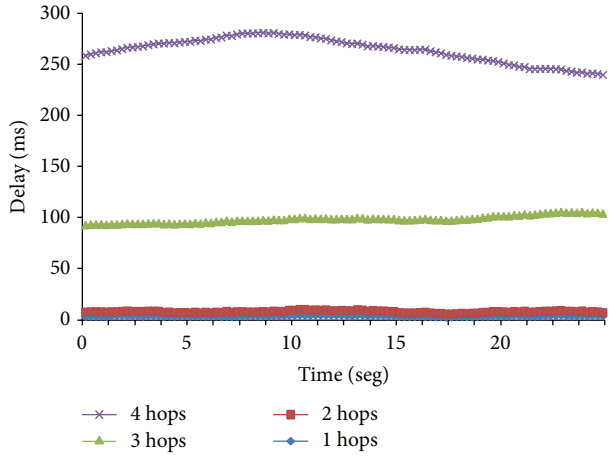


Figure 8: Delay of a multimedia stream of 600 kbps for different cluster diameters.

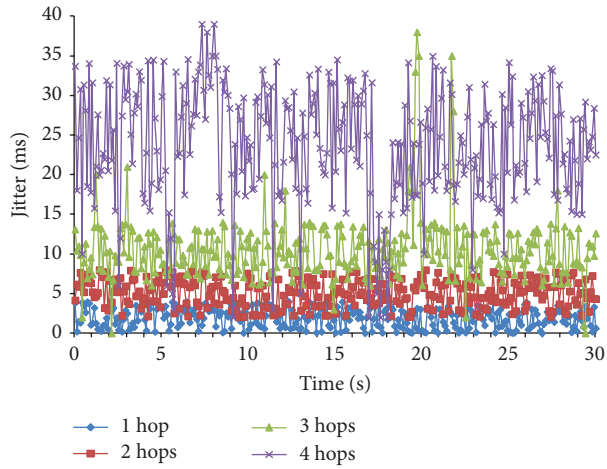


Figure 9: Jitter of a multimedia stream of 600 kbps for different cluster diameters.

obtained for 4 hops. The difference with the rest of cases is high. One hop has the lowest jitter values.

We have performed the statistical analysis of the results with $\alpha = 1$ (see Table 5). We can check that there is a significant difference when the number of hops is increased. Three hops doubles 2 hops values and 4 hops doubles 3 hops values. We can conclude that the jitter values directly depend on the number of hops in the cluster topology.

6. Conclusion

In this paper, a new architecture for ad hoc wireless networks has been proposed. It is a cluster-based architecture and it uses QoS profiles to optimize multimedia traffic. The architecture provides a flexible solution with the ability to guarantee the quality of multimedia communication over the ad hoc wireless network. It is able to adapt to many physical network configurations through the suitable selection of the multimedia init profiles (MIPs). The paper shows how QoS parameters and the multimedia codec characteristics affect the topology of the cluster. Moreover, the cluster diameter affects severely the delay and jitter. The proposed architecture provides a control mechanism to build the appropriate topology for each cluster. Furthermore, the system uses a

Table 5: Statistical values of the jitter as a function of the diameter.

HOPS	Parameters						
	N	μ (ms)	σ (ms)	Min (ms)	Max (ms)	Conf. Int. (ms)	
1	300	1.99	1.23	0	6	1.80	2.17
2	300	5.11	1.72	2	8	4.85	5.37
3	300	10.56	4.28	0	38	9.92	11.19
4	300	23.71	7.57	2	39	22.58	24.84

resource reservation scheme to guarantee the quality of the multimedia streams.

In future works we will integrate some mechanisms to allow the system to adapt very fast to spatial changes and node mobility. Moreover, we will add security to the communications through authentication integrity and confidentiality techniques. Our final purpose is to deploy the proposed architecture in a real environment to provide multimedia streaming in wireless sensor networks [22].

Conflict of Interests

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

References

[1] R. Lacuesta, J. Lloret, M. Garcia, and L. Peñalver, "A spontaneous ad hoc network to share WWW access," *EURASIP Journal on Wireless Communications and Networking*, vol. 2010, Article ID 232083, 2010.

[2] J. Lloret, M. Garcia, J. Tomás, and F. Boronat, "GBP-WAHSN: a group-based protocol for large wireless ad hoc and sensor networks," *Journal of Computer Science and Technology*, vol. 23, no. 3, pp. 461–480, 2008.

[3] J. Y. Yu and P. H. J. Chong, "A survey of clustering schemes for mobile ad hoc networks," *IEEE Communications Surveys & Tutorials*, vol. 7, no. 1, pp. 32–48, 2005.

[4] J. Lloret, M. Garcia, D. Bri, and J. R. Diaz, "A cluster-based architecture to structure the topology of parallel wireless sensor networks," *Sensors*, vol. 9, no. 12, pp. 10513–10544, 2009.

[5] D. Wei and H. A. Chan, "Clustering ad hoc networks: schemes and classifications," in *Proceedings of the 3rd Annual IEEE Communications Society on Sensor and Ad Hoc Communications and Networks (SECON '06)*, pp. 920–926, Reston, Va, USA, September 2006.

[6] M. Garcia, D. Bri, F. Boronat, and J. Lloret, "A new neighbour selection strategy for group-based wireless sensor networks," in *Proceedings of the 4th International Conference on Networking and Services (ICNS '08)*, pp. 109–114, Gosier, Guadeloupe, March 2008.

[7] M. Lehsaini, H. Guyennet, and M. Feham, "Cluster-based energy-efficient k-coverage for wireless sensor networks," *Network Protocols and Algorithms*, vol. 2, no. 2, pp. 89–106, 2010.

[8] J. Lloret, L. Shu, R. Lacuesta, and M. Chen, "User-oriented and service-oriented spontaneous ad hoc and sensor wireless networks," *Ad-Hoc & Sensor Wireless Networks*, vol. 14, no. 1-2, pp. 1–8, 2012.

[9] A. Canovas, D. Bri, S. Sendra, and J. Lloret, "Vertical WLAN handover algorithm and protocol to improve the IPTV QoS of the end user," in *Proceedings of the IEEE International Conference on Communications (ICC '12)*, pp. 10–15, Ottawa, Canada, June 2012.

[10] C. Zhou and N. Maxemchuk, "Distributed dottleneck flow control in mobile ad hoc networks," *Network Protocols and Algorithms*, vol. 3, no. 1, pp. 22–45, 2011.

[11] R. C. Suganthe and P. Balasubramanie, "Improving QoS in delay tolerant mobile ad hoc network using multiple message ferries," *Network Protocols and Algorithms*, vol. 3, no. 4, pp. 32–53, 2011.

[12] R. Zhang, L. Cai, J. Pan, and X. Shen, "Resource management for video streaming in ad hoc networks," *Ad Hoc Networks*, vol. 9, no. 4, pp. 623–634, 2011.

[13] T. Mehta and Z. Narmawala, "Performance enhancement of multimedia traffic over wireless ad hoc networks using network coding," in *Proceedings of the Nirma University International Conference on Engineering (NUiCONE '12)*, pp. 1–6, Ahmedabad, India, December 2012.

[14] M. Tarique, A. Hossain, R. Islam, and C. A. Hossain, "Issues of long-hop and short-hop routing in mobile ad hoc networks: a comprehensive study," *Network Protocols and Algorithms*, vol. 2, no. 2, pp. 107–131, 2010.

[15] R. Lacuesta, M. Garcia, J. Lloret, and G. Palacios, "Study and performance of ad hoc routing protocols," in *Mobile Ad Hoc Networks: Current Status and Future Trends*, chapter 4, pp. 71–101, CRC Press, Toronto, Canada, 2011.

[16] R. Al Turki and R. Mehmood, "Multimedia ad hoc networks: performance analysis," in *Proceedings of the 2nd UKSim European Symposium on Computer Modelling and Simulation (EMS '08)*, pp. 561–566, Liverpool, UK, September 2008.

[17] A. Jamali, N. Naja, and D. El Ouadghiri, "Comparative analysis of ad hoc networks routing protocols for multimedia streaming," in *Proceedings of the International Conference on Multimedia Computing and Systems (ICMCS '09)*, pp. 381–385, Ouarzazate, Morocco, April 2009.

[18] A. Abdrabou and W. Zhuang, "Statistical QoS routing for IEEE 802.11 multihop ad hoc networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 3, pp. 1542–1552, 2009.

[19] D. Kandris, M. Tsagkaropoulos, I. Politis, A. Tzes, and S. Kotsopoulos, "Energy efficient and perceived QoS aware video routing over wireless multimedia sensor networks," *Ad Hoc Networks*, vol. 9, no. 4, pp. 591–607, 2011.

[20] N. Taing, S. Thipchaksurat, R. Varakulsiripunth, and H. Ishii, "Routing scheme for multimedia services in mobile ad hoc network," in *Proceedings of the 5th International Conference on Information, Communications and Signal Processing*, pp. 11–15, Bangkok, Thailand, December 2005.

[21] J. R. Diaz, J. Lloret, J. M. Jiménez, and M. Hammoumi, "A new multimedia-oriented architecture and protocol for wireless ad

hoc networks," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 16, no. 1, 2014.

[22] J. R. Diaz, J. Lloret, J. M. Jimenez, and J. J. P. C. Rodrigues, "A QoS-based wireless multimedia sensor cluster protocol," *International Journal of Distributed Sensor Networks*. In press.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 601913, 9 pages
http://dx.doi.org/10.1155/2014/601913

Research Article

A Lightweight Neighbor-Info-Based Routing Protocol for No-Base-Station Taxi-Call System

Xudong Zhu, Jinhang Wang, and Yunchao Chen

School of Computer Science & Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China

Correspondence should be addressed to Xudong Zhu; zhuxd@zjgsu.edu.cn

Received 26 October 2013; Accepted 28 November 2013; Published 9 March 2014

Academic Editors: W. Sun, G. Zhang, and J. Zhou

Since the quick topology change and short connection duration, the VANET has had unstable routing and wireless signal quality. This paper proposes a kind of lightweight routing protocol-LNIB for call system without base station, which is applicable to the urban taxis. LNIB maintains and predicts neighbor information dynamically, thus finding the reliable path between the source and the target. This paper describes the protocol in detail and evaluates the performance of this protocol by simulating under different nodes density and speed. The result of evaluation shows that the performance of LNIB is better than AODV which is a classic protocol in taxi-call scene.

1. Introduction

The VANET has characteristics of mobile ad hoc network, such as self-organization, flexible structure, multihop routing, dynamic changes of network topology, and need of good scalability. Special application environment, such as narrow road, high density node distribution, high-speed node moving, and other factors, will directly influence the information transmission ability of VANET network, which leads to the increase of packet loss and delay. Meanwhile, the limited wireless channel and huge network scale challenge the data transmission of VANET.

The key to solve the above problems is how to design a routing protocol [1–7]. Since the nodes in VANET are of high mobility, it results in connection failure while applying routing protocol of the traditional MANET [8–10], such as AODV [11–13] and DSR [14, 15], into the VANET, thus leading to delay or even failure to transmit information to the target node.

Due to the special environment in VANET, it cannot have a widely applicable routing algorithm. The typical algorithm in VANET, such as GSR routing protocol, involves each node storing the neighbor lists, topology table, next hop table, and distance table. All of them maintain the state information of adjacent nodes and choose the appropriate router according to the location and topological information. In small network with high mobility and limited bandwidth, the performance of transmission is good However, it requires the node to maintain the network topology. And with the increase of network size, the routing information that needs to be exchanged will increase exponentially. Another type of algorithm, such as the GPSR [16, 17], is based on position. It depends on the overall geographical location information search system. It cannot work without GPS.

When applied to the urban taxi-calling scene, the VANET vehicle network has the following characteristics. (1) The density distribution of vehicle is uneven. Places such as the business center, theaters, and bus stations are dense. The traditional VANET routing protocols have serious information channel congestion problem. (2) The calling passenger has no special requirements to the not-taken taxis, so all the nearby taxis can serve. While the traditional VANET routing protocols always have designated transmission target, or specifies the routing service of all nodes in the target area, which will lead to huge routing overhead, further deteriorating the channel congestion problem. (3) The data transmission is time valid. Passengers' calls need to be responded to in a short time, while the routing connection time is fairly long in some of the traditional VANET routing protocols. Therefore, it cannot effectively find not-taken taxis with such application

and will have bad effects on passengers' experiences. We describe some of the work roughly in [18].

This paper aims to solve the quick topology change, short connection duration, and unstable wireless signal quality of the VANET network, thus proposing a kind of lightweight routing protocol applicable to the urban taxi-call without base stations. The protocol maintains and predicts neighbor information dynamically in order to find the reliable path between the source and the target. This paper describes the protocol in detail and evaluates the performance of this protocol by simulating under different nodes density and speed. The result of evaluation shows that the performance of LNIB is better than that of AODV which is a classic protocol in taxi-call scene.

The second chapter analyzes the application and proposes the model; the third chapter describes the routing protocol and its algorithm; the fourth chapter is about the experiment and data analyses; chapter five concerns other related research; the last part will draw a conclusion.

2. Transmission Mode and Analyses

In the taxi-calling scene, the vehicle network has two sets: taxi nodes $Q = \{q_1, q_2, \dots q_n\}$ and passenger nodes $P = \{p_1, p_2, \dots p_m\}$. From the start of the call to get on the car or the call to cancel, a connection should be set up in the network. Because the passenger always waits at the original place or moves slowly, suppose p is a static node. And p only acts as the initiator of data transmission or receiver, not as mediator in other routings. The node q can send, receive or act as a routing node to transmit data. Assume that the speed scalar of each node in Q is $|v|$.

When p initiates a call, the message transits through the taxi nodes, so p can quickly find nearby not-taken cars node q_s that can meet the requirements.

Supposing there are two nodes a and b in the network, the signal coverage radii for them are, respectively, R_a and R_b. If the distance between a and b is $D_{a,b} \le \min(R_a, R_b)$, it can be called "meet." To be more general, assuming that signal coverage radius of each node in the network is R, then "meet" in time t means $D_{a,b}(t) \le R$. Define H_{a,q_s} as the minimal number of hops to transmit message from a to q_s. Among all q_s that get the call request, H_a is the lowest of H_{a,q_s}.

2.1. Prediction Based on Neighbor Information. In order to percept the current meet nodes in real time, q needs to send broadcast packets to notice their own state information periodically. In the starting time t of every broadcast cycle, q measures their own motion state, including the speed, the current position, and other information. According to the above information, q predicts the coordinates at the time of $t + t_f$ and broadcasts the position. The broadcast cycle is t_b. If t_b is small, then the channel will be occupied by a large number of broadcast packets, thus influencing the normal data transmission. However, large t_b will lose the prediction value due to the frequent changes in actual situation. Therefore, t_b should be set properly, such as 10 seconds. A broadcast cycle can be divided into multiple correction cycles. The node in a correction period measures their movement state. While finding the coordinates' different Δs between the current forecast coordinate at $t + t_f$ and the last broadcast coordinates surpasses, it indicates that the vehicle's driving condition changes. If Δs is greater than a certain value, the node broadcasts the position again to correct the predict position of itself in neighbor nodes' state table, thus to ensure the prediction accuracy. Suppose the correct period is t_r.

Meanwhile, if q receives the position broadcast of the nearby nodes, it maintains an adjacent link (see Table 1) to record node information that meets with q. It can be seen from the table that q_2 measures its position and velocity at $13{:}11{:}12'36$ and broadcasts. q_1 receives the position broadcast. If $D_{q_1,q_2}(t) \le R$, then it can be predicted that, at $t + t_f$, q_1 and q_2 will keep meeting. If q_2 is not in the table, q_1 adds the information of q_2 to the table and sets q_2 to be active. If q_2 is already in the table, update the information of q_2. When q_1 receives the correction broadcast of q_2, the position information of q_2 also needs updating. If it can be predicted that q_1 and q_3 will not meet at $t + t_f$, set q_3 to be inactive. As to the inactive node in the table, if no broadcasts are received in more than one broadcast cycle, then delete the entry of this inactive node. Each time the passenger calls a taxi, several links will be produced. If q_1 and q_3 are neighbors and are both in one link, then they set up link ID in the entry for each other in their own table. Link ID is a large number that is randomly generated. It can be assumed that it is unique in a limited region.

When q_1 receives the broadcast of neighbors at time t, it will predict coordinates of itself and neighbor nodes at $t + t_f$, respectively. If the two cars can keep communication at that time (such as q_2), the communication is effective. Therefore q_1 stores the information of q_2 in the table otherwise discards it (such as the q_3).

In order to provide the position information and routing of not-taken taxis, the hop distance should be introduced when q node is broadcasting message (including position and correction broadcast). Periodically q_s sets H_{q_s} to 0 and sends broadcast. The q_s discards any broadcast information it received without processing. The q sets H_q value to ∞, when no neighbor nodes can reach q_s node. When q_1 receives the broadcast of q_2 with the active state, if $H_{q_2} + 1 < H_{q_1}$, q_1 modifies the H_{q_1} value to $H_{q_2} + 1$.

Once H_{q_1} changes, q_1 will spread to notify neighbor nodes to be updated by periodical breakfast. But each cycle can only update 1 hop on the link. If $H_{q_1,q_2} = n$, it needs n cycle to update H_{q_2}.

In order to improve the broadcast speed of H, the sending phase of each cycle is divided into two time slices. The node of H changes, but no broadcast will be broadcasted at the first time slice. If the H value has no change after last broadcast, the node broadcasts in the second time slice. Thus, the value of H may spread 2 hops in a cycle. For the length of link will be limited in 3-4 hops in the actual system, two cycles can complete the information disclosure of not-taken taxis by the improved broadcast speed.

2.2. Links Setting and Handshake. When p launches the call, it needs to find suitable not-taken taxis q_s as soon as possible and make an appointment. In this process, the following

Table 1: State of the q_1's Neighbor Nodes.

Neighbor	H_{q_n,q_s}	Position	Direction of speed	Link ID	Reliability	Measuring time	State
q_2	3	120.213476E: 30.321768N	d_{q_2}	182378627	0.94	13:11:12′36	Active
q_3	5	120.213484E: 30.321775N	d_{q_3}	452766732	0.83	13:10:37′34	Inactive

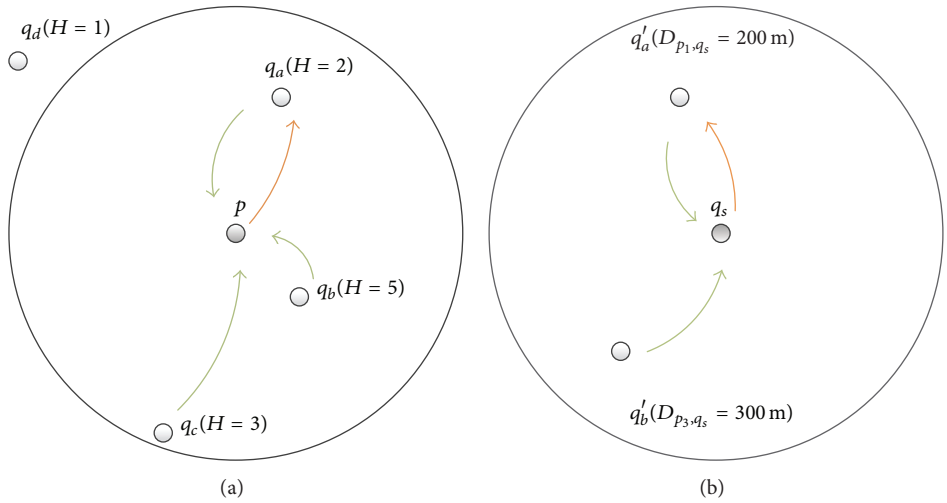


Figure 1: (a) If there are the q nodes within the single hop range of p, then they shake hands and choose one as q_a that has minimal hop distance to q_s to make an appointment. (b) If there are several q nodes around q_s and several p nodes, send the appointment requests, then they shake hands and choose the nearest p node to response.

questions need to be solved. First, how to find q_s in the shortest time and establish the link to q_s? Secondly, if there are several q_s nodes, how to choose one of the most suitable ones to make an appointment?

The simplest method is to use the broadcast to send out the call until a node of q_s receives the request and returns a response. The response will go backtracking to the p node. If p receives several responses, then choose one from them, after which appointment confirmation will be sent to the chosen node through the path just set up. Meanwhile, it will send negative responses to other not-taken taxis. This method is easy to be implemented, but in the real VANET scene, there are serious problems; the message must pass the whole link three times: request, response, and appointment. On the one hand, it will lead to serious delay and difficulties in keeping the link; on the other hand, the flooding of breakfast will deteriorate the network environment and influence the normal communication.

This paper proposes a two-end handshake agreement to great link (show as Figures 1 and 2) According to the neighbor nodes' information described in Section 2.1, each node can know the nearest (by hop distance) q_s and the shortest route to arrive at the node. If p sending the request broadcast, its neighbor q node which receives p's breakfast directly can judge whether it can reach q_s or not. In fact, it is supposed that the whole data transmission is a single route (the node on the link will transmit a message to only one neighbor

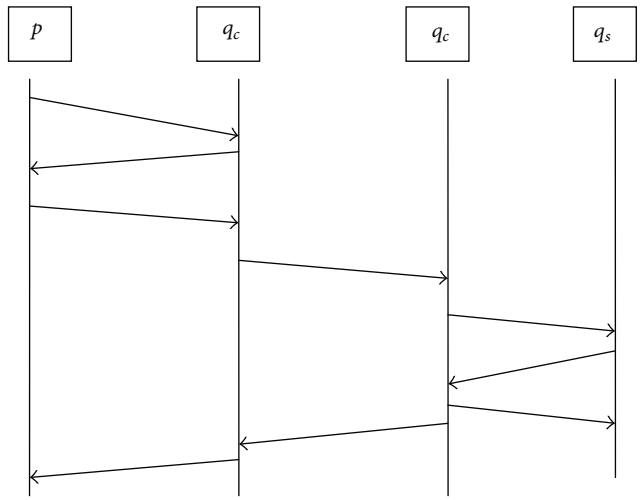


Figure 2: The two-end handshake agreement.

node). Once q is identified as the first hop, the routing from p to q_s will also be established. The q can directly send a request response to the p. If p receives responses from several q nodes, it chooses a q node as q_a to make appointment confirmation and send a veto appointment to other response q. After that q_a accepts the confirmation and sends it to the q_s

based on the routing information, that is, to transmit message to the neighbor q node in which $H_q = H_{q_a} + 1$. On the routing, node q' will record the information of last hop, while receiving the message. Meanwhile, transmit to the next hop which $H_q = H_{q'} + 1$ until sent to the node which $H_q = 0$, that is q_s.

The q_s end of the link also needs handshaking. If several neighbor q nodes, in which $H_q = 1$, are for the same q_s node, they will send appointment requests to the q_s, and the q_s sends a confirmation response to one of them and negative responses to others.

The two-end handshaking method can reduce the three times handshaking to two times, thus reducing the risk of delaying and disconnection. Meanwhile, the broadcast can be controlled in 1 hop and can avoid the flooding.

2.3. Disconnection. Suppose $H_{q_s} = 0$, $H_{q_1} = 1$, $H_{q_2} = 2$, and $H_{q_3} = 3$. If q_2 receives the position correction information from q_1, the adjusted position predictions indicate that q_1 will lose contact with q_2. Since H_{q_1} is the minimum H stored in q_2, after q_2 node deletes the q_1, it modifies its own H value to ∞ and triggers an extra position correction broadcast. Also, when q_3 receives the position correction by this broadcast, it continues the same operation.

In the communication process, the effectiveness of link should be guaranteed from the sending of p's request to q_s's response return to p. Once a pair of nodes on the link are disconnected, it should be processed correspondently. The relatively ideal way is to find the third node before disconnection and to maintain the link. Another simple method is to notify p and q_s to abandon the appointment on both ends. Thus, it can avoid feedback loss which will result in information consistency. Then the p initiates the request process again.

As the vehicle's speed is fast and unstable, the time of link maintains a relatively short time. The frequently broken links will lead to repeated reconnection. To avoid this, higher reliable links are required. So in the q's neighbor nodes state table, the reliability attribute is added for each neighbor node. The reliability refers to the prediction accuracy at $t + t_f$ and the help to effective datagram transmission on the link. It mainly depends on the following two conditions: (1) during t to $t + t_f$ period, the relative position between q' and q. If the distance is far, on the one hand, the weak signal can lead to transmission instability; on the other hand, since q' has been close to the edge of the q communication range, the slight change of the vehicles will make q' out of q's communication range and cause broken connection. If q' and q are close, the data transmission will take place in a short distance. The datagram will not move a lot in geographical location; thus, it influences the efficiency of data transmission. Thus supposing $U_{q,q'}(t') = 2|D_{q,q'}(t') - R|/R$ refers to the transmission reliability of q, and q'

$$U_{q,q'}\left(t, t + t_f\right) = \frac{\text{integration}}{\left(t, t + t_f\right)}. \quad (1)$$

Modify the condition of correcting H_q in Section 2.1. Given the reliable threshold $r \in [0, 1]$ and after q_1 node receives the broadcast of q_2 node, if state (q_2) = active, $H_{q_2} + 1 < H_{q_1}$, and $U_{q,q'}(t, t + t_f) > r$, then modify the H_{q_1} value to be $q_2 + 1$.

3. Protocol Implementation

In order to further introduce the implementation of the protocol, the pseudocode algorithms of taxi q node and passenger p node are given, respectively, in this section. To be simple, based on the discussion in the second section, the following several types of datagram in the taxi-call network will be referred to.

Broadcast (L, V, H, t_c) represents position broadcast. The taxi node sends periodically, and the nearby taxi node receives.

M_{1-6} represents the six kinds of datagrams in the two-end handshaking protocol after the passengers make a request.

M_e represents the cancelation of the reserved datagram while finding the abnormal disconnection.

Algorithm 1 is divided into two parts. The timer gets the position and speed information for a short time (e.g., 3 seconds). If the interval exceeds t_f, then send broadcast to the nodes around. According to the current measured data, the position at $t + t_f$ can be predicted. If there exists significant difference between the position and the position of the last broadcast, then send broadcast without considering whether the time intervals exceed t_f or not. The timer will also inspect the neighbor node that is not updated over time in the adjacency list and set these nodes as inactive. Since the time in the record is the position measuring time, to consider the communication delay, the overtime window will be slightly amplified to $1.1t_f$. If not updated for more than two broadcast cycles, then delete the node.

As for the inactive nodes, check whether they are in the current communication connection. If it is true, then send M_e to other connected nodes to cancel the conformation.

The other part of Algorithm 1 is the listener. After the listener receives a datagram, he/she judges data types. If it is the position broadcast datagram, then update the list. Once the predicted node position changes are identified in the current connection and in a short time will overpass the scope of communication, and then notify other nodes on the connection to cancel the appointment. When accepting other types of datagram, again transmit or respond based on the protocol.

Algorithm 2 mainly includes two timers and a listener. When initiating the request, timer_1 is set. At the same time, the listener will put the received response to the response queue. When timer_1 is triggered, it checks the queue. If the queue is empty, it means no response, and it sends the request again. Otherwise, it chooses the response that has the nearest transmission distance from passengers to not-taken taxis in order to make appointment. And it simultaneously sends veto to other response nodes. After sending the appointment datagram, set timer_2. Once no appointment conformation M_4 in a certain time M_4, then the timer_2 will be launched, and send request M_1 again. Attention here, once the listener receives an M_e, the datagram cancellation is caused by

```
Initialization t = 0;
Initiate the prediction position S = 0;
Initialization H = ∞;
Set timer;
Timer_task(){
    Obtain the current position L, speed V, time T_c
    H_new = H;
    If (no-load){
        H_new = 0;
    }
    Calculate prediction position S' at (t + t_f)
    If (t_c − t≥t_f){
        H = H_new
        broadcast(L, V, H, t_c);
        t = t_c;
        S = S'
    }else if (|S − S'| larger than threshold ‖ H − H_new⟨⟩0){
    H = H_new
        broadcast(L, V, H, t_c);
        t = t_c;
        S = S';
    }
    For each node q in table{
        get q's prediction time t_q
        if (state(q) == active && t_c − t_q ≥ 1.1t_f){
            state(q) = inactive;
            if (q have the link ID){
                send M_e to other node which have the same lnik ID as q;
            }
        }
        If (t_c − t_q ≥ 2t_f){
              Remove q from table;
         }
    }
    Update H;
}
Count = 0; // reset 0 after passengers getting off or appointment cancellation
Block_listen(Message){
      Switch typeof(Message) {
          Case M_b:
                update table
                update H
                if q node that sends M_b gets out of the communication range, and it has link ID in the table
                       send M_e to other node
                       which have the same link ID as q;
                break;
          Case M_1:
                If (H < ∞){send M_2 back}
                break;
          Case M_3:
                If (Message == Reserve){
                     If (H == 0 && count == 0){
                             Count = 1;
                             send M_4(Agree or Disagree) back;
                     }else if (H == 1){
                          Set link ID;
                          send M_3 to all q in table which H_q = 0;
                      }else{
```

Algorithm 1: Continued.

```
                    Set link ID;
                    send M3 to one q which Hq = H − 1;
                }
            }
        Case M4:
            If (Message == Agree){
                Send M4 forward to q which Hq = H + 1
                    and have the same link ID;
                delete link ID;
            }
            If (H == 1){
                Send Confirm back;
        }
        Default:
    }
}
```

Algorithm 1: Algorithm of taxi node.

```
Initiation queue = empty
Call_taxi{
    Get its position L
    Broadcast(M1)
    Set Timer1();
}
Timer1_task(){
    If (queue = empty){
        Call_taxi();
    }else{
      Choose the nearer hop q,
      Produce link ID;
      Send M3(Reserve, link ID) to q;
      Send M3(Negative) to other
    }
    Set timer2;
}
Timer2_task(){  Call_taxi(); }
Block_listen(Message){
  Switch typeof(Message) {
      Case M2
          Add q to queue;
          Break;
      Case M4:
          Cancel Timer2;
          Break;
      Case Me:
          Set Timer2 current time;
          Break;
      Default:
}
```

Algorithm 2: Algorithm of passenger node.

connection error. Then immediately start the timer_2 and send request again.

4. Performance Analysis

In order to evaluate the performance of the proposed routing protocol, the paper uses NS2 and VanetMobiSim to simulate.

The NS2's main parameters are shown in Table 1. The simulation scene size is 5×5 square kilometers. In VanetMobiSim simulation, the scene size is 5×5 square kilometers too. In the total area there are 20 traffic signals and the entire simulation area is divided into several regions. In every region the density of streets and the allowed maximum speed are different. The regions which have higher street density have lower allowed maximum speed. When vehicles get into the corresponding region, they choose the speed from street allowed maximum speed and self-allowed maximum speed with a low one. In all of our simulations, the empty taxis account for 40%.

We set the interval of route maintain message T_b to be equal to 10 s and then set the maximum speed of taxi nodes to be equal to 60 km/h. When we adjusted the density of taxi nodes, we got the probability of passenger node P to find empty taxi node T_D and the probability of node P to receive the message from T_D as shown in Table 3.

As shown in Table 3, when node density is too low, the transmission range will be so small that it is hard to maintain the information of nodes for a long time. Hence, when P appears, the number of nodes that P can communicate with is small and the time of link is short so that it is hard to communicate with T_D. As the node density becomes higher, situations are getting better, but some nodes have higher velocity, which leads to more failure. Because the links chosen are the links that have long duration, so the probability of node P receiving acknowledgement messages from node T_D is pretty high.

Then if we fix the density of taxi nodes as 5 per km^2 and fix the transmission range as 500 m, we will get the relation of the maximum speed and the successful ratio.

We called it successful calling when the passenger node finds the empty taxi node. In Figure 3, we could see that when the maximum speed of taxi is equal 60 km/h, it has the highest

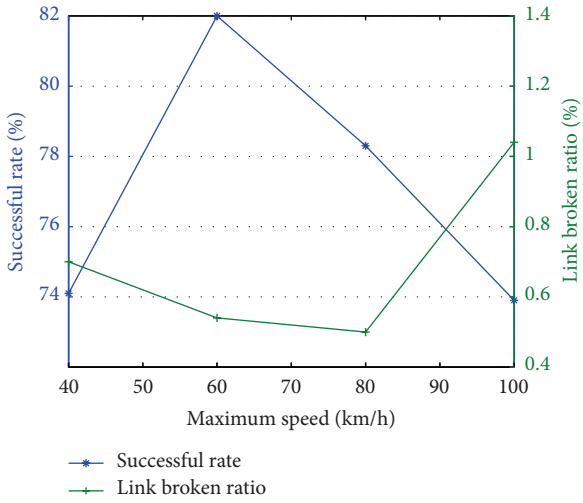


Figure 3: The relation of the maximum speed of taxi nodes and the ratio of passenger node contacting empty taxi nodes successfully.

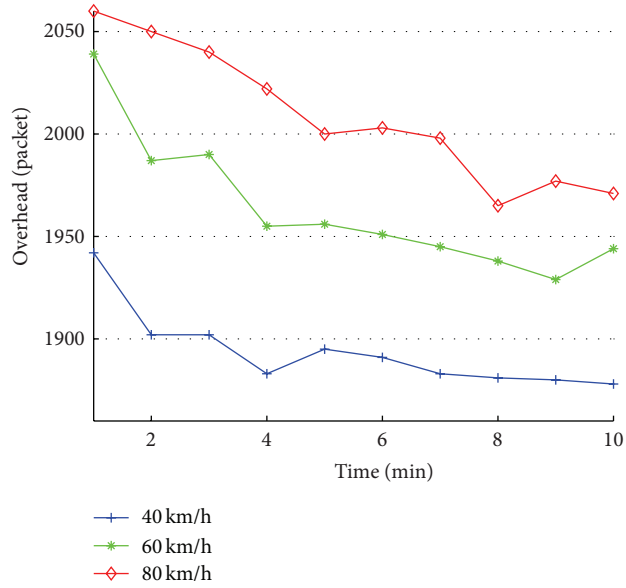


Figure 4: The impact of node's moving speed on the amount of switching packets of entire network.

successful ratio which enables the passenger node to find the empty taxi nodes. This is because when the node's moving speed is low, the node can contact with fewer other nodes. And when the node's moving speed is high, the network topology changes fast, which causes fewer available routing link and thus causes lower successful calling.

Figure 4 shows the impact of node's moving speed on the amount of switching packets of entire network.

As shown in Figure 4, the amount of switching packets grows up as the vehicle movement speed increases. This is because when the speed of node movement increases, it causes more correcting message, resulting in the increase of network data traffic. In addition, in one simulation, the routing overhead is higher at the beginning and gradually declines over time. This is because when the simulation starts, all taxi nodes are active at the same time and all taxi nodes broadcast at the same time, which causes more channel collision, leading to an increasing number of retransmissions. As time progresses, due to the correcting message, the taxi routing maintenance cycles gradually stagger and the entire network routing overhead declines.

Then we adjust the density of taxi nodes in the region. The average density of taxi nodes is set at 1–6 per square kilometer and the vehicle maximum speed is set at 60 km/h and then the results are shown in Figure 5.

As shown in Figure 5, when the taxi nodes are scarce, the rate of the success calling drastically reduces. While increasing the number of taxi nodes, the rate of success calling improves accordingly. This is because when the number of taxi nodes is too small, there is no taxi node around passenger node so that the passenger node cannot contact with any node, resulting in more calling failure. When the number of taxi nodes increase and, the passenger node sends an appointment request, the number of alternative routing paths increases, so the node can choose a more stable routing path to delivery packet; thereby the rate of success calling will increase.

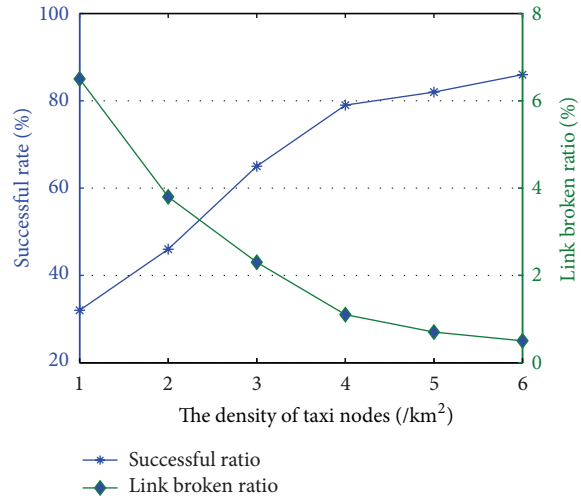


Figure 5: The impact of nodes density on successful ratio.

Table 2: The NS2 parameters.

Parameters	Values
Loss model	Propagation/TwoRayGround
Physical model	Phy/WirelessPhyExt
Antenna	Antenna/OmniAntenna
Transmission rate	2 Mbps
Network protocol	IEEE 802.11

At last, we control node velocity between 0 and 20 m/s; interval of T_b 10 s, transmission range is 500 m, and the density is 4 per square kilometers. Compared to AODV (show as Table 4), the results are shown in Table 2. We can see that the proposed routing protocol performs better than

Table 3: The impact of nodes density and transmission range on successful calling ratio.

Transmission range	Density of taxi			
	4		6	
	The probability to find T_D	The probability to receive from T_D	The probability to find T_D	The probability to receive from T_D
200 m	14.8%	78.5%	49.8%	95.5%
300 m	44.3%	91.2%	78.4%	94.2%
500 m	61.1%	94.1%	87.3%	94.6%

Table 4: Compare with AODV.

	CBLIGR	AODV
Delivery rate/(%)	61.1	52.3
Control message/packet	1206.8	1815.0

AODV. More importantly, our protocol communication data is scattered throughout the network, while AODV is relatively concentrated. As a result, our protocol performs better in the channel collision. It is because AODV requires a great deal of broadcast to maintain the entire link routing information and has no efficient methods to deal with the disconnection of links.

5. Conclusion

This paper proposes a routing protocol for taxi-calling application. Firstly, it analyzes scenes of taxi-calling application and studies the differences between the common vehicle application scene and the taxi-calling application scene. Secondly, it proposes the routing information maintaining algorithm based on position prediction which performs better reliability. Thirdly, a self-network forming method for taxi-calling application is drawn up.

Our protocol queries the destination node by nodes' keeping the information of the hop count to destination and decreases the routing cost. Meanwhile, the validity of routes is guaranteed by predicting the future position of themselves and their neighbors.

This protocol has also been evaluated by a traffic simulator. The results indicate that our protocol can efficiently find the target nodes in the area. It has also been found that the number of broadcasts is affected by nodes' velocity. And the fact that our protocol generates 30% less messages than AODV has been proven here.

As to future work, our protocol can probably be extended to solve the low delivery rate problem when the density of vehicle is low.

Conflict of Interests

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

Acknowledgment

This work is supported in part by the Natural Science Foundation of China "Research on the snapshot data security storage technology for authorization of release," no. 61100057.

References

[1] C. Rezende, R. W. Pazzi, and A. Boukerche, "An efficient neighborhoood prediction protocol to estimate link availability in VANETs," in *Proceedings of the 7th ACM International Symposium on Mobility Management and Wireless Access (MobiWAC '09)*, pp. 83–90, Tenerife, Spain, October 2009.

[2] T. Zahn, G. O'Shea, and A. Rowstron, "Feasibility of content dissemination between devices in moving vehicles," in *Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies (CoNEXT '09)*, pp. 97–108, Rome, Italy, December 2009.

[3] N. Wisitpongphan, F. Bai, P. Mudalige, and O. K. Tonguz, "On the routing problem in disconnected vehicular ad Hoc networks," in *Proceedings of the 26th IEEE International Conference on Computer Communications (IEEE INFOCOM '07)*, pp. 2291–2295, Anchorage, Alaska, USA, May 2007.

[4] A. Boukerche, H. A. B. F. Oliveira, E. F. Nakamura, and A. A. F. Loureiro, "Vehicular Ad Hoc networks: a new challenge for localization-based systems," *Computer Communications*, vol. 31, no. 12, pp. 2838–2849, 2008.

[5] C. E. Perkins and P. Bhagwat, "Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers," in *Proceedings of the Conference on Communications Architectures, Protocols and Applications (SIGCOMM '94)*, pp. 234–244, 1994.

[6] T. Spyropoulos, T. Turletti, and K. Obraczka, "Routing in delay-tolerant networks comprising heterogeneous node populations," *IEEE Transactions on Mobile Computing*, vol. 8, no. 8, pp. 1032–1047, 2009.

[7] C. Lochert, M. Mauve, H. Füßler, and H. Hartenstein, "Geographic routing in city scenarios," *Mobile Computing and Communication Review*, vol. 9, no. 1, pp. 69–72, 2005.

[8] M. Jerbi and S. M. Senouci, "Towards efficient routing in vehicular Ad Hoc networks," in *Proceedings of the 3rd IEEE International Workshop on Mobile Computing and Networking*, 2006.

[9] H. Menouar, M. Lenardi, and F. Filali, "An intelligent movement-based routing for VANETs," in *Proceedings of the ITS World Congress*, London, UK, October 2006.

[10] K. Shafiee and V. C. M. Leung, *Connectivity-Aware Minimum-Delay Geographic Routing with Vehicle Tracking in VANETs*, Elsevier B.V., 2010.

[11] J. Haerri, F. Filali, and C. Bonnet, "Performance comparison of AODV and OLSR in VANETs urban environments under realistic mobility patterns," in *Proceedings of the Med-Hoc-Net 2006, 5th Annual Mediterranean Ad Hoc Networking Workshop*, Lipari, Italy, June 2006.

[12] C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing," in *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications (WMCSA '99)*, pp. 90–100, New Orleans, La, USA, February 1999.

[13] T. W. Chen and M. Gerla, "Global state routing: a new routing scheme for Ad-hoc wireless networks," in *Proceedings of the IEEE International Conference on Communications (ICC '98)*, vol. 1, pp. 171–175, Atlanta, Ga, USA, June 1998.

[14] B. Karp and H. T. Kung, "GPSR: greedy perimeter stateless routing for wireless networks," in *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking (MOBICOM '00)*, pp. 243–254, Boston, Mass, USA, August 2000.

[15] D. B. Johnson, D. A. Maltz, and J. Broch, "DSR: the dynamic source routing protocol for multi-hop wireless Ad hoc networks," in *Ad Hoc Networking*, C. E. Perkins, Ed., chapter 5, pp. 139–172, Addison-Wesley, New York, NY, USA, 2001.

[16] D. B. Johnson, D. A. Maltz, and J. Broch, "DSR: the dynamic source routing protocol for multi-Hop wireless ad hoc networks," in *Ad Hoc Networking*, C. E. Perkins, Ed., chapter 5, pp. 139–172, Addison-Wesley, New York, NY, USA, 2001.

[17] Z. Wei and S. Weibin, "Research of GPSR routing protocol for wireless sensor networks," *Electronic Measurement Technology*, vol. 33, no. 9, 2010.

[18] G. Wei, Y. Chen, and X. Zhu, "A vanet-oriented routing protocol for intelligent taxi-call systems," in *Proceedings of the IEEE 14th International Conference on Communication Technology (ICCT '12)*, pp. 41–45, Chengdu, China, November 2012.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 685102, 15 pages
http://dx.doi.org/10.1155/2014/685102

Research Article

Uplink Scheduling and Adjacent-Channel Coupling Loss Analysis for TD-LTE Deployment

Woon-Young Yeo,[1] Sung Ho Moon,[2] and Jae-Hoon Kim[3]

[1] *Department of Information and Communication Engineering, Sejong University, 98 Gunja-dong, Gwangjin-gu, Seoul 143-747, Republic of Korea*
[2] *Network Technology R&D Center, SK Telecom, 9-1 Sunae-dong, Bundang-gu, Seongnam 463-838, Republic of Korea*
[3] *Department of Industrial Engineering, Ajou University, Yeongtong-gu, Suwon 443-749, Republic of Korea*

Correspondence should be addressed to Jae-Hoon Kim; jayhoon@ajou.ac.kr

Received 11 October 2013; Accepted 25 November 2013; Published 20 February 2014

Academic Editors: W. Sun and J. Zhou

TD-LTE, one of the two duplexing modes in LTE, operates in unpaired spectrum and has the advantages of TDD-based technologies. It is expected that TD-LTE will be more rapidly deployed in near future and most of WiMax operators will upgrade their networks to TD-LTE gradually. Before completely upgrading to TD-LTE, WiMax may coexist with TD-LTE in an adjacent frequency band. In addition, multiple TD-LTE operators may deploy their networks in adjacent bands. When more than one TDD network operates in adjacent frequency bands, severe interference may happen due to adjacent channel interference (ACI) and unsynchronized operations. In this paper, coexistence issues between TD-LTE and other systems are analyzed and coexistence requirements are provided. This paper has three research objectives. First, frame synchronization between TD-LTE and WiMax is discussed by investigating possible combinations of TD-LTE and WiMax configurations. Second, an uplink scheduling algorithm is proposed to utilize a leakage pattern of ACI in synchronized operations. Third, minimum requirements for coexistence in unsynchronized operations are analyzed by introducing a concept of adjacent-channel coupling loss. From the analysis and simulation results, we can see that coexistence of TD-LTE with other TDD systems is feasible if the two networks are synchronized. For the unsynchronized case, some special cell-site engineering techniques may be required to reduce the ACI.

1. Introduction

Long Term Evolution (LTE) is a radio platform that allows mobile operators to achieve much higher peak data rates and better spectral efficiency than those of the third generation (3G) networks (e.g., WCDMA and cdma2000) [1]. LTE was initiated by the Third Generation Partnership Project (3GPP) in 2004 and is now commercially deployed or in progress worldwide. The LTE standard has two duplexing modes (LTE-TDD and LTE-FDD) and the technical specification of the two modes is almost the same. The LTE-TDD, also known as TD-LTE, can operate in unpaired spectrum and supports asymmetric resource allocation between uplink and downlink. Even though both TD-LTE and LTE-FDD will be widely used, TD-LTE will be more rapidly deployed in near future due to a number of advantages of the TDD-based technology [2].

As a fourth generation (4G) technology competing with TD-LTE, WiMax provides a broadband wireless access to mobile subscribers based on IEEE 802.16 standards [3]. Starting from IEEE 802.16, the standard was amended many times and changed to 802.16a, 802.16c, 802.16d, 802.16e, and 802.16m. IEEE 802.16e is the most popular standard for WiMax today and 802.16m is the most advanced version approved in March 2010. WiMax networks have been commercially deployed in many countries, but most of these are niche deployments for specialized applications [4]. WiMax operators can upgrade 802.16e to 802.16m or migrate to TD-LTE directly in their unpaired spectrum. Considering the competitive 4G markets, it may be better to upgrade their network towards TD-LTE gradually instead of upgrading to 802.16m.

Upgrading to TD-LTE is a potentially difficult decision for WiMax operators because most WiMax operators have

recently launched their networks. Before completely upgrading to TD-LTE, WiMax may coexist with TD-LTE in the same frequency band because the two systems are technically very alike and the mobile operators want to minimize the extra cost for TD-LTE deployment. WiMax Forum has considered a network evolution path to accommodate harmonization and coexistence across multiple wireless access technologies including TD-LTE [5]. In addition, it is also possible that multiple TD-LTE operators deploy their networks in adjacent frequency bands. When two TDD networks operate in adjacent frequency bands and are deployed in the same area, severe interference may happen due to adjacent channel interference and unsynchronized transmission timing between the two networks.

A number of papers and technical reports have been published for coexistence analysis among 3G/4G wireless networks. In [5], network and air interface requirements for WiMax were specified to facilitate the coexistence of WiMax and TD-LTE. In [6], capacity loss of 3G/4G TDD systems (TD-SCDMA, TD-HSDPA, and TD-LTE) was presented when they were interfered by an LTE-Advanced system operating on an adjacent frequency band. In [7], coexistence of TD-LTE and LTE-FDD was analyzed, and intercell interference coordination was studied through simulation. In [8], requirements for coexistence between TD-SCDMA and TD-LTE were given by a simple interference calculation, and the required antenna installation was investigated to reduce mutual interference.

In this paper, coexistence between TD-LTE and other systems is analyzed and operational requirements are provided to maintain an acceptable level of adjacent channel interference. Most of examples and applications will be explained by assuming TD-LTE/WiMax coexistence, but the proposed methodology and analysis can be applied to other coexistence scenarios (e.g., TD-LTE vs. TD-LTE and TD-LTE vs. LTE-FDD). There are three research objectives in this paper. First, frame synchronization between TD-LTE and WiMax is discussed by investigating possible combinations of TD-LTE and WiMax configurations. Because an unsynchronized operation may cause serious interference, the synchronization feasibility between the two systems is checked and some additional techniques for synchronization are explained. Second, an uplink scheduling algorithm is proposed to make the uplink transmission robust against the adjacent channel interference when two networks are synchronized. Third, a concept of adjacent-channel coupling loss is introduced to estimate the minimum requirements for network coexistence when two networks are not synchronized. Because there has been little work on the interference analysis for the unsynchronized operation, the proposed analysis methodology can provide a guideline for cell-site engineering in the unsynchronized network.

The rest of this paper is organized as follows. The basic frame structures of TD-LTE and WiMax are explained in Section 2. In Section 3, the synchronization feasibility between TD-LTE and WiMax is analyzed by checking possible combinations of network configurations and by adjusting some system parameters. In Section 4, an uplink scheduling algorithm is proposed for robust operations in the synchronized operation, and a concept of adjacent-channel coupling loss is introduced to estimate the coexistence requirements in the unsynchronized operation. In Section 5, evaluation methodology and simulation assumptions are described in detail. The simulation and analysis results are presented and discussed in Section 6. Finally, Section 7 summarizes and concludes this paper.

2. TD-LTE and WiMax Frame Structure

Downlink (DL) and uplink (UL) transmissions in TD-LTE and WiMax are based on orthogonal frequency division multiple access (OFDMA). Specifically, OFDMA is used for DL of TD-LTE and for UL and DL of WiMax. The UL transmission for TD-LTE uses a technology called single carrier-frequency division multiple access (SC-FDMA), which is similar to OFDMA but more power efficient than OFDMA at a user equipment (UE). In this section, frame structures of the two systems are explained because the difference between them is critical to analyze the network synchronization.

2.1. TD-LTE Frame Structure. In the 3GPP specification, there is no operational difference between TD-LTE and LTE-FDD at higher layers or in the system architecture [2]. The design goal of the LTE physical layer is to achieve as much commonality as possible. As in LTE-FDD, a resource block (RB, 180 kHz) is a minimum allocation unit in frequency domain and defined as a group of 12 consecutive subcarriers with subcarrier spacing of 15 kHz. The minimum allocation unit in time domain is a subframe or transmission time interval (TTI), which has a duration of 1 ms. One subframe consists of two time slots and has 12 or 14 OFDM symbols depending on a length of cyclic prefix (CP).

The TD-LTE frame structure is shown in Figure 1. TD-LTE has a 10 ms TDD frame structure. A radio frame consists of 10 subframes. Subframes 0 and 5 contain synchronization signal and broadcast information necessary for a UE to perform synchronization and obtain relevant system information. Subframes 1 and 6 are special subframes that serve as a switching point between DL and UL transmission. The special subframe contains three fields: downlink pilot time slot (DwPTS), guard period (GP), and uplink pilot time slot (UpPTS).

Table 1 provides the special subframe configurations, where the length of each field is given in multiples of OFDM symbols and a normal CP is assumed. DwPTS can be viewed as an ordinary DL subframe and is used for DL data transmission. Unlike a normal subframe, where the control region can span up to three OFDM symbols, the maximum control region in DwPTS is limited to two OFDM symbols. UpPTS has a short duration of one or two OFDM symbols and can be used for transmission of sounding reference signals (SRS) and random access. SRS enables a base station (BS) to estimate UL channel quality. SRS can be also provided by normal subframes as in LTE-FDD. Random access typically uses one of normal subframes, enabling a relatively long random access preamble. GP is used to prevent overlap between transmission and reception, and the length of GP

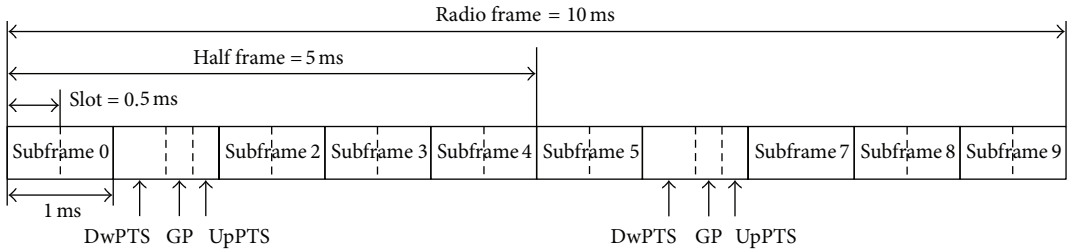


Figure 1: TD-LTE frame structure (5 ms periodicity).

Table 1: DwPTS/GP/UpPTS length in a special subframe (OFDM symbols, normal CP).

Format	0	1	2	3	4	5	6	7	8
DwPTS	3	9	10	11	12	3	9	10	11
GP	10	4	3	2	1	9	3	2	1
UpPTS	1	1	1	1	1	2	2	2	2

depends on the cell size. Because formats 0 and 5 have a large GP, they can support a large cell size, but they cannot provide additional DL capacity due to a small DwPTS.

In TD-LTE, two switching periodicities are supported: 5 ms and 10 ms. For the 5 ms periodicity, subframe 6 is a special subframe identical to subframe 1, whereas it is a regular DL subframe for the 10 ms periodicity. Table 2 shows the DL/UL allocation according to the switching periodicity. In this paper, TD-LTE configurations 0 to 2 are candidates for network coexistence because WiMax only supports a frame length of 5 ms.

2.2. Frame Structure of WiMax IEEE 802.16e-TDD. Figure 2 illustrates a mobile WiMax (IEEE 802.16e-TDD) frame structure [3]. Each 5 ms frame is time-division duplexed with DL and UL subframes. There are time gaps between DL and UL subframes, considering mobile transceiver turnaround time and a guard time avoiding interference between DL and UL signals. A time gap for transition from DL to UL subframe is called transmit time gap (TTG). A time gap for transition from UL to DL subframe is called receive time gap (RTG).

At the beginning of each frame, DL control information is transmitted and consists of a preamble, a frame control header (FCH), and MAP messages. The preamble can be used for synchronization and DL channel estimation. The subcarriers allocated to the preamble are uniformly distributed over the spectrum and occupy every third subcarrier. FCH provides information required to decode the subsequent DL-MAP message. The DL/UL-MAP messages indicate the resource allocation for DL/UL data and control transmission. UL control channels consist of ranging, channel quality indicator (CQI), and acknowledgment (ACK) channels. The ranging channel provides the random access for initial entry, timing adjustment, periodic synchronization, bandwidth request, and handover entry. The CQI or fast-feedback channel is used by a mobile station to report the measured signal quality back to an access point. The ACK channel reports ACK/NACK feedback for DL data transmission.

Table 2: TD-LTE DL/UL configuration (D: downlink, U: uplink, S: special subframe).

Configuration	Periodicity	0	1	2	3	4	5	6	7	8	9
0		D	S	U	U	U	D	S	U	U	U
1	5 ms	D	S	U	U	D	D	S	U	U	D
2		D	S	U	D	D	D	S	D	D	D
3		D	S	U	U	U	D	D	D	D	D
4	10 ms	D	S	U	U	D	D	D	D	D	D
5		D	S	U	D	D	D	D	D	D	D
6	5 ms	D	S	U	U	U	D	S	U	U	D

The WiMax system supports a scalable system bandwidth of 3.5, 5, 7, 8.75, and 10 MHz. If TTG and RTG are excluded, each 5 ms frame has 47 OFDM symbols for 5 and 10 MHz bandwidth, 42 symbols for 8.75 MHz bandwidth, and 33 symbols for 3.5 and 7 MHz bandwidth [9]. In addition, there are different numbers of DL/UL ratios depending on the system bandwidth: 10 different DL/UL ratios for 5 and 10 MHz bandwidth and 7 ratios for 3.5, 7, and 3.75 MHz bandwidth. The DL/UL ratio is adjustable to support asymmetric traffic, and the parameters for 10 MHz bandwidth are summarized in Table 3. Internationally, the 29 : 18 DL/UL ratio is very common and popular for WiMax operators. Note that the ratio of DL and UL may not dynamically change per cell but should change on a system-wide basis.

3. Frame Synchronization Analysis for TD-LTE and WiMax Coexistence

All cellular wireless systems suffer from interference in adjacent bands and TDD can have two more serious interference sources: from BS to BS and from UE to UE. Depending on frequency arrangements, the interference may occur between two or more TDD systems or between TDD and FDD systems. TDD performance is significantly degraded when DL and UL signals from two TDD systems operating in adjacent bands overlap in time. Special cell-site engineering techniques may be required to reduce the excessive interference. The simplest way is frame synchronization, where DL/UL frame boundaries across TDD networks are aligned in time. Frame start timings of all BSs can be synchronized through GPS, IEEE 1588 version 2, and "network listening" [9].

Considering frame structures of TD-LTE and WiMax, it is necessary to specify a frame offset to TD-LTE. The frame

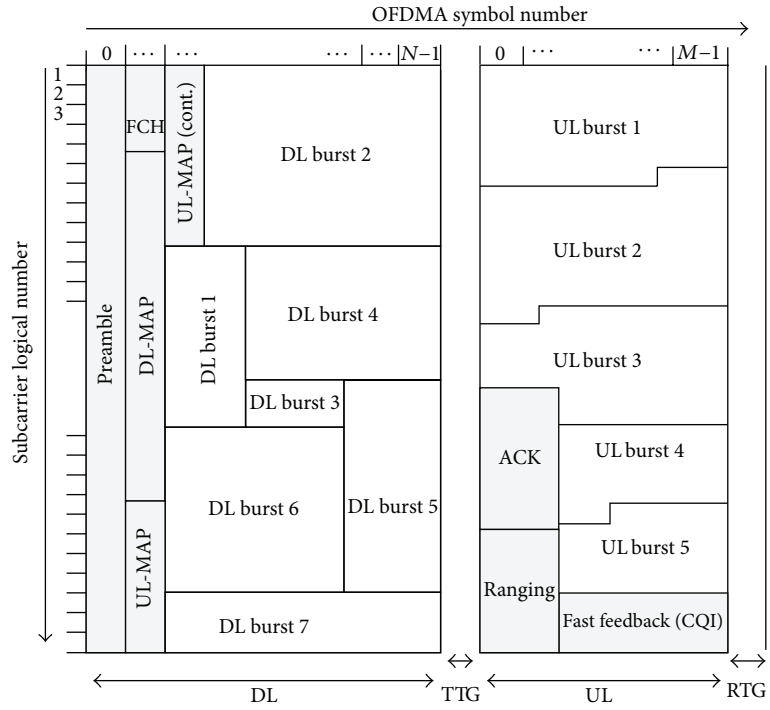


Figure 2: Frame structure of IEEE 802.16e-TDD.

Table 3: WiMax frame configuration (TDD mode, 10 MHz).

Configuration	DL symbols	UL symbols	DL duration (ms)	TTG (ms)	UL duration (ms)	RTG (ms)
(35, 12)	35	12	3.600000	0.105714	1.234286	0.06
(34, 13)	34	13	3.497143	0.105714	1.337143	0.06
(33, 14)	33	14	3.394286	0.105714	1.440000	0.06
(32, 15)	32	15	3.291424	0.105714	1.542855	0.06
(31, 16)	31	16	3.188571	0.105714	1.645714	0.06
(30, 17)	30	17	3.085714	0.105714	1.748571	0.06
(29, 18)	29	18	2.982857	0.105714	1.851429	0.06
(28, 19)	28	19	2.880000	0.105714	1.954286	0.06
(27, 20)	27	20	2.777143	0.105714	2.057143	0.06
(26, 21)	26	21	2.674286	0.105714	2.160000	0.06

offset ensures that the start of continuous DL subframes lines up with the WiMax DL subframe. Figure 3 shows an example of the frame synchronization between TD-LTE and WiMax, assuming TD-LTE with configuration 1 and WiMax with a default frame configuration of (29, 18). In this example, UpPTS of TD-LTE overlaps with the last DL symbol of WiMax, and thus one DL symbol of WiMax and UpPTS of TD-LTE may cause serious interference to each other if they operate in adjacent bands.

In this section, frame synchronization between TD-LTE and WiMax is analyzed for possible combinations of system configurations. In TD-LTE, configurations 0 to 2 in Table 2 are candidates for coexistence due to a frame length of WiMax. WiMax has up to 10 different DL/UL ratios and the number of OFDM symbols per frame is different depending on the system bandwidth. The 10 MHz WiMax system is assumed in this analysis, but it is possible to apply the same analysis technique to other bandwidths.

In TD-LTE configuration 0 (DL : UL = 1 : 3), one subframe is configured for DL transmission in a 5 ms frame. Because the WiMax DL has more OFDM symbols than UL, it is difficult to align the WiMax frame to TD-LTE. The last part of WiMax DL always overlaps with the UL subframe of TD-LTE. In all combinations, frame synchronization is not supported between the two systems, and thus TD-LTE configuration 0 is not suitable for coexistence in a normal network environment.

As for TD-LTE configuration 1 (DL : UL = 2 : 2), frame synchronization between the two TDD systems is possible in some combinations, but the default WiMax configuration of (29, 18) overlaps with UpPTS as shown in Figure 3. To solve the synchronization problem for the default configuration, it

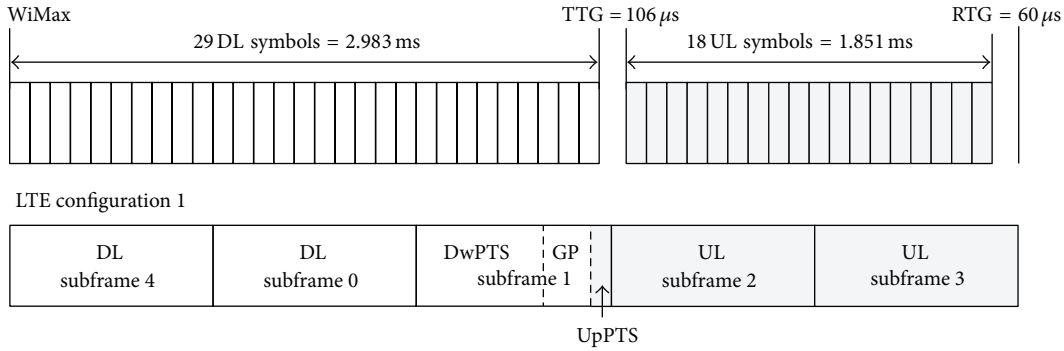


Figure 3: Example of frame synchronization between TD-LTE and WiMax.

has been suggested to disable some resources in one of the two systems, by disabling the last one or two symbols in WiMax or by disabling UpPTS [9]. Due to capacity loss of WiMax, it is desirable to disable UpPTS instead of WiMax DL symbols. Note that UpPTS is used to transmit RACH and SRS, and UL symbols in a normal subframe can be allocated to RACH and SRS. Tables 4(a) and 4(b) show the number of overlapped TD-LTE symbols when UpPTS is enabled and disabled, respectively. The overlap region occurs when the last DL symbols of WiMax overlap with subframe 2 (UL) of TD-LTE or the first UL symbols of WiMax overlap with DwPTS. If UpPTS is enabled, three WiMax configurations can support frame synchronization with TD-LTE, whereas four WiMax configurations can do by means of disabling UpPTS.

In TD-LTE configuration 2 (DL : UL = 3 : 1), frame synchronization is possible in some combinations, but there is no TD-LTE frame configuration that naturally aligns with the default WiMax configuration of (29, 18). The overlap region occurs mainly because WiMax UL overlaps with DwPTS. If DwPTS is disabled, more WiMax configurations can be synchronized with TD-LTE. Tables 5(a) and 5(b) show the number of overlapped TD-LTE symbols when DwPTS is enabled and disabled, respectively. With enabled DwPTS, frame synchronization is only supported in a few combinations. TD-LTE special subframe formats 0 and 5 have a large GP supporting a large cell size and WiMax configuration of (35, 12) allocates much more DL resources than UL. These network parameters are not commonly used in commercial TD-LTE and WiMax services. If DwPTS is disabled, the first seven WiMax configurations can support frame synchronization and the last three WiMax combinations partially overlap with TD-LTE subframe 1 (DL). However, because DwPTS can be used for DL data transmission in TD-LTE, disabling DwPTS can result in up to 22% performance loss assuming two control symbols in a subframe. Thus, TD-LTE configuration 2 is not strongly recommended for frame synchronization between TD-LTE and WiMax.

4. Uplink Scheduling and Adjacent-Channel Coupling Loss Analysis

To analyze coexistence issues, this paper focuses on adjacent channel interference from an aggressor system (WiMax or

Table 4: Number of TD-LTE symbols overlapped with 10 MHz WiMax (TD-LTE configuration 1).

(a) UpPTS enabled

WiMax DL : UL ratio	TD-LTE special subframe format								
	0	1	2	3	4	5	6	7	8
35 : 12	10	10	10	10	10	11	11	11	11
34 : 13	8	8	8	8	8	9	9	9	9
33 : 14	7	7	7	7	7	8	8	8	8
32 : 15	6	6	6	6	6	7	7	7	7
31 : 16	4	4	4	4	4	5	5	5	5
30 : 17	3	3	3	3	3	4	4	4	4
29 : 18	1	1	1	1	1	2	2	2	2
28 : 19	0	0	0	0	0	1	1	1	1
27 : 20	0	0	0	0	0	0	0	0	0
26 : 21	0	0	0	1	2	0	0	0	1

(b) UpPTS disabled

WiMax DL : UL ratio	TD-LTE special subframe format								
	0	1	2	3	4	5	6	7	8
35 : 12	9	9	9	9	9	9	9	9	9
34 : 13	7	7	7	7	7	7	7	7	7
33 : 14	6	6	6	6	6	6	6	6	6
32 : 15	5	5	5	5	5	5	5	5	5
31 : 16	3	3	3	3	3	3	3	3	3
30 : 17	2	2	2	2	2	2	2	2	2
29 : 18	0	0	0	0	0	0	0	0	0
28 : 19	0	0	0	0	0	0	0	0	0
27 : 20	0	0	0	0	0	0	0	0	0
26 : 21	0	0	0	1	2	0	0	0	1

Table 5: Number of TD-LTE symbols overlapped with 10 MHz WiMax (TD-LTE configuration 2).

(a) DwPTS enabled

WiMax DL:UL ratio	TD-LTE special subframe format								
	0	1	2	3	4	5	6	7	8
35:12	0	0	1	2	3	0	0	1	2
34:13	0	1	2	3	4	0	1	2	3
33:14	0	3	4	5	6	0	3	4	5
32:15	0	4	5	6	7	0	4	5	6
31:16	0	5	6	7	8	0	5	6	7
30:17	1	7	8	9	10	1	7	8	9
29:18	2	8	9	10	11	2	8	9	10
28:19	4	10	11	12	13	4	10	11	12
27:20	5	11	12	13	14	5	11	12	13
26:21	7	13	14	15	16	7	13	14	15

(b) DwPTS disabled

WiMax DL:UL ratio	TD-LTE special subframe format								
	0	1	2	3	4	5	6	7	8
35:12	0	0	0	0	0	0	0	0	0
34:13	0	0	0	0	0	0	0	0	0
33:14	0	0	0	0	0	0	0	0	0
32:15	0	0	0	0	0	0	0	0	0
31:16	0	0	0	0	0	0	0	0	0
30:17	0	0	0	0	0	0	0	0	0
29:18	0	0	0	0	0	0	0	0	0
28:19	1	1	1	1	1	1	1	1	1
27:20	2	2	2	2	2	2	2	2	2
26:21	4	4	4	4	4	4	4	4	4

TD-LTE) to a victim system (TD-LTE) in macrocellular scenarios. The adjacent channel interference (ACI) is the total interference from adjacent channels and mainly related to the amount of signal leakage from a transmitter, the amount of signal loss between two transceivers (i.e., coupling loss), and the ability of a receiver to suppress out of band interference [10]. The adjacent channel leakage is measured in terms of adjacent channel leakage ratio (ACLR), which is defined as the ratio of the desired signal power in its channel to the power measured in an adjacent channel. Similarly, adjacent channel selectivity (ACS) is a measure of the ability of a receiver to filter and reject the signal from adjacent channels. ACS is defined as the ratio of the receiver filter attenuation on the desired channel to the receiver filter attenuation on the adjacent channel(s). The coupling loss (CL) is simply the amount of signal attenuation between the transmitter and receiver. The CL is the sum of the path attenuation (or path loss), antenna gains, and any other cabling losses.

A parameter named ACIR (adjacent channel interference ratio) is used to measure the overall ACI and defined as the ratio of the total power transmitted from an aggressor transmitter to the total interference power affecting a victim receiver in the adjacent channel. The ACIR is expressed as a combination of ACLR and ACS according to the following equation (expressed in linear scale):

$$\mathrm{ACIR} = \frac{1}{(1/\mathrm{ACLR}) + (1/\mathrm{ACS})}. \quad (1)$$

In (1), the UL limiting factor is the UE transmitter because UE ACLR is much lower than BS ACS (i.e., UE ACLR $\ll$ BS ACS). Similarly, the DL limiting factor is the UE receiver because UE ACS is much lower than BS ACLR (i.e., UE ACS $\ll$ BS ACLR) [11, 12].

In this section, an uplink scheduling algorithm is proposed to make the UL transmission robust against the ACI when two different networks are synchronized. In addition, a concept of adjacent-channel coupling loss is introduced to analyze the minimum requirements for network coexistence when two networks are not synchronized.

4.1. Uplink Scheduling for Synchronized Operations. Considering UL ACIR in (1), the limiting factor for UL is UE ACLR, and thus the UL performance loss may depend on the ACLR pattern of aggressor UEs. Generally, the amount of ACI to a victim channel is related to the frequency location of the victim channel. This is also reflected in the UL interference model of 3GPP [12]. The typical UL ACI characteristics are shown in Figure 4 assuming that an aggressor frequency band is just below the victim band. If one of victim channels is located near the upper edge of the aggressor band, it can suffer from more ACI than other channels far from the aggressor band.

Because the ACI pattern is not constant over the victim spectrum, the performance loss can be related to UL frequency allocation policy. For example, if low-ACI RBs are allocated to UEs with low signal quality (i.e., edge UEs), the edge throughput degradation is not so significant even though the cell throughput can be reduced. Similarly, if the low-ACI RBs are allocated to UEs with high signal quality (i.e., central UEs), the cell throughput degradation is negligible by the sacrifice of the edge throughput.

In order to design a robust TDD system against the ACI from the aggressor, it is worthwhile to analyze the UL scheduling algorithms that take into account the ACI distribution over the victim spectrum. The proposed UL scheduling is based on the ACI distribution and is called ACI-based uplink scheduling in this paper. Depending on the allocation policy of the low-ACI RBs, the proposed scheduling has two types of operation: minimum CL first (minCL) and maximum CL first (maxCL). The ACI-based minCL scheduling allocates the low-ACI RBs to the UEs with a low CL (i.e., central UEs), whereas ACI-based maxCL scheduling allocates them to the UEs with a high CL (i.e., edge UEs).

The ACI-based minCL scheduling is expected to minimize the cell throughput loss caused by the ACI. If N UEs

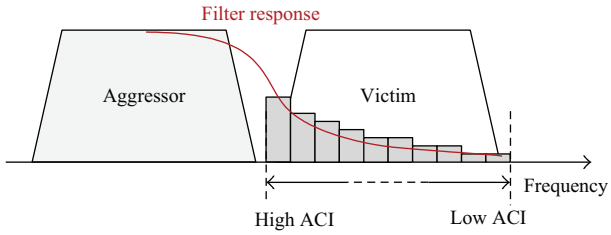


Figure 4: Conceptual example of adjacent channel interference caused by ACLR.

are selected for UL scheduling, the minCL procedures are described as follows.

Step 0: divide UL frequency resources into N consecutive RB sets to support SC-FDMA transmission. Let $L(k)$ be a CL between the BS and UE k ($k \in U = \{1, \ldots, N\}$), and let $I(k)$ be an ACI level for RB set k ($k \in R = \{1, \ldots, N\}$) at the BS receiver.

Step 1: find a UE with the minimum CL to the serving BS in U. Letting i be the corresponding UE identity,

$$i = \arg \min_{k \in U} L(k). \quad (2)$$

Step 2: find an RB set with the minimum ACI in R. Letting j be the corresponding RB set identity,

$$j = \arg \min_{k \in R} I(k). \quad (3)$$

Step 3: allocate RB set j to UE i.

Step 4: delete i in U and delete j in R.

Repeat Steps 1 to 4 until all RB sets are allocated.

Note that the ACI distribution over the victim spectrum may be different according to frequency planning of mobile operators. If a victim band is located in the middle of two aggressor bands, the ACI pattern may be different from Figure 4.

Another scheduling option, ACI-based maxCL scheduling, is expected to minimize the edge throughput loss because it allocates low-ACI RBs to edge UEs. Procedures for ACI-based maxCL scheduling are almost the same as those of minCL scheduling, except in Step 1. Step 1 and (2) are replaced by the following procedure.

Step 1: find a UE with the maximum CL to the serving BS in U. Letting i be the corresponding UE identity,

$$i = \arg \max_{k \in U} L(k). \quad (4)$$

4.2. Adjacent-Channel Coupling Loss Analysis for Unsynchronized Operations. It is generally assumed that two TDD systems in adjacent bands are perfectly synchronized, but it is necessary to analyze the unsynchronized network for considering the worst-case scenarios. The interference between a BS and a UE operating in adjacent bands is not significant because the CL between the BS and the UE is high [10]. In the unsynchronized network, TDD can give rise to two more interference mechanisms, between two BSs (BS-to-BS) or two UEs (UE-to-UE). In this section, a concept of adjacent-channel coupling loss is introduced to estimate the minimum requirements for coexistence in the unsynchronized network.

As mentioned before, the CL is the amount of signal attenuation between the transmitter and receiver and is the sum of the path loss (PL), antenna gains, and any other cabling losses. The minimum coupling loss (MCL) represents the lowest reasonable CL between any two transceivers. In this paper, the concept of CL is extended to effectively evaluate the performance of the unsynchronized network. The adjacent-channel coupling loss (ACCL) is defined as follows:

$$\text{ACCL} = L + \text{ACIR} - G_a - G_v, \quad (5)$$

where L is the PL between two BSs or two UEs and G_a and G_v are antenna gains of an aggressor transmitter and a victim receiver, respectively. (Cabling loss and other minor losses are ignored.) The transmitted signal from an aggressor is attenuated by the ACCL and acts as interference to a victim receiver.

In the unsynchronized operation, the nearest aggressor from a victim may cause the most serious interference to the victim receiver if it transmits signals in the opposite direction of the victim. Thus, the nearest interferer may dominate the interference on the victim receiver. In unsynchronized UL, the nearest aggressor BS may be a main source of BS-to-BS interference. Similarly, in unsynchronized DL, the nearest aggressor UE can be considered a main source of UE-to-UE interference.

In the proposed ACCL analysis, the system performance is evaluated by assuming a single dominant aggressor around a victim receiver. Because the ACCL in (5) encompasses all gains and losses, the performance evaluation does not need any detailed information about the aggressor, such as PL, filter characteristics, and antenna patterns. In the ACCL analysis, the performance evaluation procedures are as follows.

Step 1: assume a single aggressor around a victim receiver.

Step 2: evaluate the system performance for a specific ACCL.

Step 3: obtain the minimum ACCL that guarantees an acceptable system performance.

Step 4: design and optimize the system parameters by using the minimum ACCL.

In Step 4, the minimum ACCL can be used for configuring optimal system parameters. The ACCL in (5) consists of adjustable components. First, PL is related to the distance between two BSs or two UEs. Second, ACIR represents filter characteristics of the transmitter and receiver. Finally, G_a and G_v are related to beam patterns and antenna directions at the transmitter and receiver. To meet the minimum ACCL requirement, the following actions are possible: (1) increase a distance between the aggressor and the victim, (2) improve ACLR and ACS performance of the transmitter and receiver, and (3) adjust antenna direction to ensure the minimum interference.

If the ACIR and antenna gains are known, the minimum PL between the aggressor and victim can be calculated by (5). Then it is possible to approximately estimate the minimum distance by a proper PL formula. There are a lot of PL models in the literature [13] and they should be properly applied to the estimation according to environmental and operational factors. In Section 6, the entire ACCL procedures can be found with typical examples, and two PL models are used for BS-to-BS and UE-to-UE cases, respectively.

5. Evaluation Methodology and Simulation Assumptions

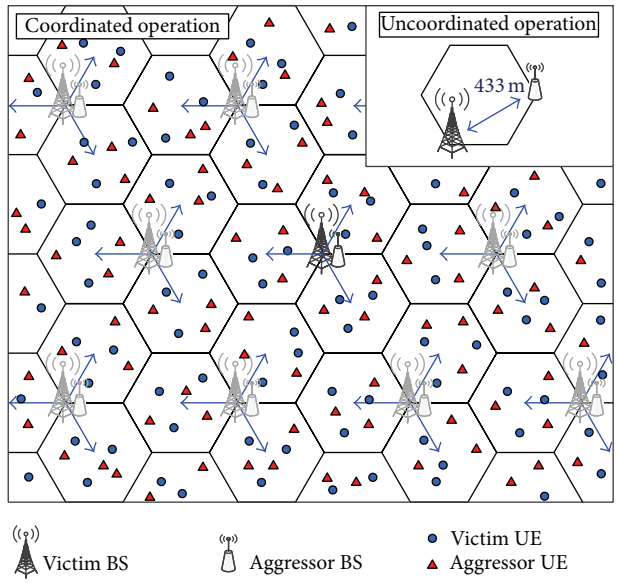
The simulation framework is based on the Vienna LTE system level simulator [14] and includes additional procedures and algorithms required for performance evaluation. Most of simulation parameters and assumptions are based on [12]. In this section, evaluation methodology and simulation assumptions are explained.

5.1. Network Layout. BSs are placed on a hexagonal grid with intersite distance of 750 m. Each BS has three sectors with directional antennas and there are 19 BS sites (57 sectors) in a single network. For a coordinated network, identical cell layouts are applied and aggressor/victim BSs are colocated at the same sites. For an uncoordinated network, identical cell layouts are applied, but with the worst-case shift between sites. Thus, every site in the aggressor network is located at the cell edge of the victim network. The distance between aggressor and victim BSs is 433 m in the uncoordinated layout.

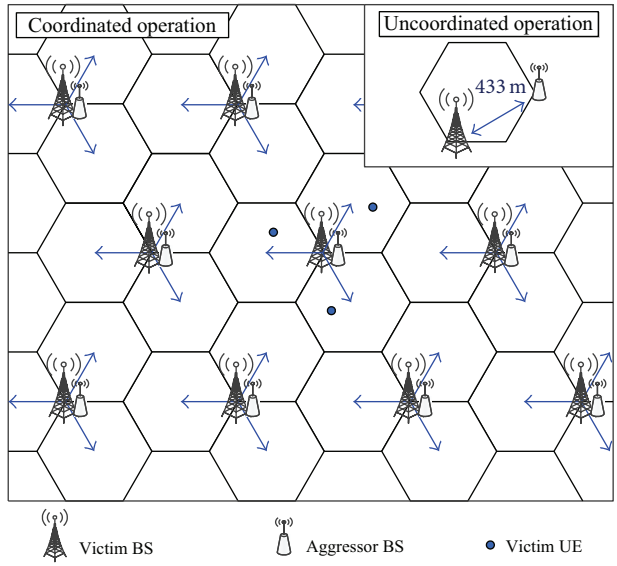
In UL simulation, the number of UEs per subframe might affect simulation results because UE transmit power depends on the number of UEs per subframe. In this simulation, three UEs per subframe are assumed and they are randomly distributed in each sector. Figure 5(a) shows the network layout for the UL simulation, with an example of UE distribution. Circles represent victim UEs in TD-LTE and triangles are aggressor UEs in WiMax or TD-LTE. 1,000 different UE distribution scenarios are applied to the simulation. All BSs of the two TDD systems can control the UE transmit power by a power control mechanism. Statistics are only collected from the central site.

As for DL simulation, the number of UEs per subframe does not affect simulation results because BS transmit power is constant. Only one UE is randomly distributed in each sector. Figure 5(b) shows the network layout for the DL simulation, with an example of UE distribution. Statistics are only collected from the central site, and three UEs are located in the central site. 1,000 different UE distribution scenarios are applied and all BSs of the two systems transmit at full power of 46 dBm.

5.2. Interference Model. In this simulation, the system bandwidth of the victim system (TD-LTE) is 10 MHz and there are 50 RBs in frequency domain. The aggressor system (TD-LTE or WiMax) has the same system bandwidth as the victim system. The frequency band of the aggressor system is adjacent to the victim and its location is just below the victim system in frequency. This interference model is based on [12].



(a) Network layout for uplink simulation



(b) Network layout for downlink simulation

FIGURE 5: Cell layout for performance evaluation.

For UL, the ACIR is dominated by UE ACLR and UL ACIR ≈ UE ACLR from (1). Three UEs are served simultaneously by round-robin scheduling. Among the 50 RBs, 16 RBs are allocated to each UE and the outer 2 RBs are not used for data transmission. According to [12], the UL ACLR model for 10 MHz bandwidth consists of three emission levels. The UL ACLR model and reference ACLR levels ($ACLR_1$, $ACLR_2$, and $ACLR_3$) are shown in Figure 6(a). If the aggressor UE is adjacent to the victim channel, the ACIR is $ACLR_1$ + X dB to the corresponding victim channel, where X adjusts the actual ACLR level in simulation. If the aggressor UE is 16 RBs away from the victim channel, the ACIR is $ACLR_2$ + X dB.

Otherwise, the ACIR is $ACLR_3$ + X dB. There are three aggressor UEs in the aggressor band and the resulting ACI is accumulated in each victim UL channel.

A common ACIR for all RBs is used for DL performance evaluation. The ACLR of the aggressor BS has negligible impact on ACIR in (1) and DL ACIR ≈ UE ACS. The DL ACIR model and reference ACS levels (ACS_1, ACS_2, and ACS_3) are shown in Figure 6(b). The common ACIR is mathematically described as ACIR = ACS_{avg} + X dB, where ACS_{avg} is the average ACS and X is an offset relative to ACS_{avg}. ACS_{avg} is calculated by the reference ACS values (expressed in dB scale):

$$10^{-ACS_{avg}/10} = \frac{10^{-ACS_1/10} + 10^{-ACS_2/10}}{2}, \quad (6)$$

where ACS_3 is not used in (6) because the aggressor bandwidth is 10 MHz and ACS_3 is out of range.

5.3. Uplink Power Control. In DL, no power control mechanism is applied and transmit power per RB is constant. In UL, the fractional transmit power control (TPC) scheme is adopted [12, 15]. This mechanism will attempt to control the UE transmit power to compensate the PL to the BS. The UE transmit power for a data channel in subframe i is set to as follows:

$$P(i) = \min\left(P_{max}, 10\log_{10}(M(i)) + P_0 + \alpha L\right), \quad (7)$$

where P_{max} is the maximum transmit power (23 dBm), L is the PL between the UE and its serving BS (including antenna gains), and $M(i)$ is the number of allocated RBs in subframe i. In this simulation, there are two TPC sets, 1 and 2, to adjust the UL interference level, and the related parameters are given in Table 6 [12]. Generally, TPC set 1 causes more interference than TPC set 2.

5.4. Simulation Parameters and Other Assumptions. The BS antenna radiation pattern to be used for each sector is given as follows:

$$A(\theta) = -\min\left[12\left(\frac{\theta}{\theta_{3\,dB}}\right)^2, A_m\right], \quad \text{where } -180 \le \theta \le 180, \quad (8)$$

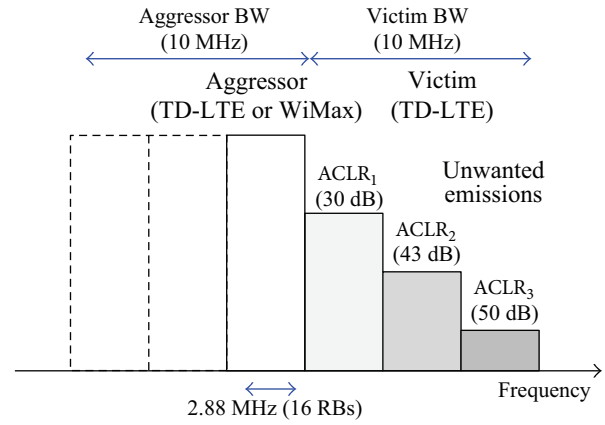
where $\theta_{3\,dB}$ is the 3 dB beam width that corresponds to 65 degrees and A_m = 30 dB is the maximum attenuation.

The PL (L) is given as follows according to the ITU-R model with high BS antennas [13]:

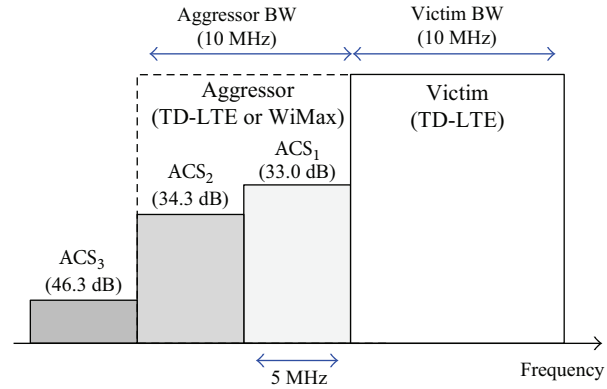
$$L = 40\left(1 - 4\times 10^{-3}\Delta h_b\right)\log_{10}(d) - 18\log_{10}(\Delta h_b) + 21\log_{10}(f_c) + 80, \quad (9)$$

where d is the BS-UE separation in kilometers, f_c is the carrier frequency in MHz, and Δh_b is the BS antenna height in meters measured from the average rooftop level. Considering a carrier frequency of 2,000 MHz and a BS height of 15 m above average rooftop level, the propagation model is given as follows [12]:

$$L = 128.1 + 37.6\log_{10}(d). \quad (10)$$



(a) Uplink ACIR model



(b) Downlink ACIR model

Figure 6: ACIR models for simulation.

Table 6: TPC parameters for TPC sets 1 and 2.

Parameter set	α	P_0 (dBm)
Set 1	1	−101
Set 2	0.8	−92.2

Other simulation parameters are summarized in Table 7. DwPTS can be used for DL data transmission and special subframe format 2 is used. Two OFDM symbols are assigned to DL control signaling in each subframe (including DwPTS). The link-level performance for UL is based on [16]. The retransmissions of UL and DL packets are not considered.

6. Results and Discussion

In this section, the performance of the proposed UL scheduling is compared with random scheduling in synchronized operations, and the minimum ACCL is obtained to estimate an acceptable level of interference in unsynchronized operations. The simulation results are presented in terms of throughput reduction relative to the reference throughput without ACI. The cell throughput is the average total data rate in a sector and the edge throughput is defined as the data rate that 5% of UEs cannot reach [12]. Generally, 5% throughput loss in the victim system is set as the evaluation criterion for the maximum allowable ACI.

Table 7: Simulation parameters.

Parameters	Values
Environment	Macro cell, urban area
Carrier frequency	2 GHz
System bandwidth	10 MHz
Noise power density	−174 dBm/Hz
BTS noise figure	5 dB
UE noise figure	9 dB
BTS antenna gain (including feeder loss)	15 dBi
UE antenna gain (omnidirectional antenna)	0 dBi
BS maximum power	46 dBm
UE maximum power	23 dBm
UE minimum power	−40 dBm
Scheduling algorithm	Round robin
Traffic model	Full buffer
TD-LTE configuration	Configuration 1 (DL : UL = 2 : 2)
DwPTS/GP/UpPTS length	10/3/1 symbols (format 2)
Microscale fading	Pedestrian A channel at 3 km/h
DL MIMO receiver modeling	Zero forcing

6.1. Uplink Scheduling in Synchronized Operations. In the UL simulation, MCL of 70 dB including antenna gains is additionally assumed for synchronized operations [12]. Table 8 summarizes reference cell throughput and edge throughput for all TPC sets and scheduling algorithms. The reference values are obtained by assuming no external interference from the aggressor system. The random scheduling allocates one of the three UL channels (16 RBs each) randomly to each UE without considering the ACI distribution. In Table 8, the ACI-based minCL and maxCL scheduling algorithms show higher cell throughput than the random scheduling, whereas the edge throughput is lower. Note that, in the proposed algorithms, a central UE in each sector is assigned the same UL channel as other central UEs in neighboring sectors. Because the central UEs generally transmit with low power, the interference to other central UEs is reduced and thus the cell throughput can be improved. Similarly, in the proposed algorithms, an edge UE in each sector is assigned the same UL channel as other edge UEs in neighboring sectors. The cell throughput of the minCL scheduling is almost the same as that of the maxCL. The edge throughput of the minCL scheduling is slightly higher due to the fast fading characteristics in a low frequency channel. The minCL scheduling allocates the low frequency channel to the edge UE, which can improve the edge UE performance slightly compared with allocation to the high frequency channel. The reference throughput with TPC set 1 is much higher than that with TPC set 2 because UEs with TPC set 1 generally transmit at higher power than those with TPC set 2. TPC set 1 can ensure a higher network throughput in a single system whereas TPC set 2 can reduce the ACI to other systems due to the reduced UE transmit power.

Table 8: Reference cell throughput and edge throughput.

TPC set	Scheduling	Cell throughput	Edge throughput
Set 1	Random scheduling	3.24 Mbps	597 kbps
	ACI-based minCL	3.57 Mbps	538 kbps
	ACI-based maxCL	3.59 Mbps	512 kbps
Set 2	Random scheduling	2.15 Mbps	397 kbps
	ACI-based minCL	2.22 Mbps	361 kbps
	ACI-based maxCL	2.23 Mbps	346 kbps

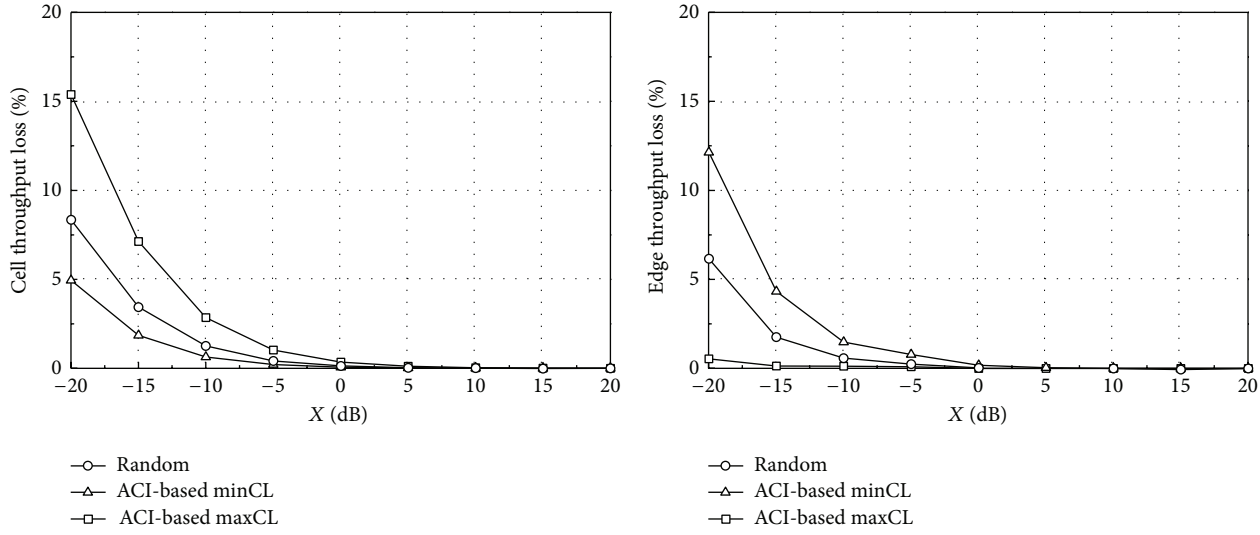
Figures 7 and 8 give the simulation results for the throughput loss in different network layouts when TPC set 1 is applied to both of the two TDD systems. As the ACI decreases (i.e., a higher value of X), the capacity loss decreases to zero. The capacity loss in coordinated layout is much lower than that in uncoordinated layout because the aggressor UEs give much higher interference to the victim BS in uncoordinated layout. Especially, the edge throughput loss is high in uncoordinated layout because more edge UEs in the aggressor system are likely to locate near the victim BS. These aggressor UEs will transmit at a higher power level according to the TPC mechanism, which results in severe interference and edge throughput degradation.

The edge throughput loss of the minCL scheduling is higher than that of other algorithms. The lower frequency band of the victim system is susceptible to interference from the aggressor system and the lower band is allocated to edge UEs in the minCL scheduling. On the other hand, the minCL scheduling has better performance on the cell throughput loss because low-ACI RBs are allocated to the central UEs.

Figures 9 and 10 show the throughput loss when TPC set 2 is applied. The performance patterns are similar to those of TPC set 1, but the overall throughput loss is much lower due to reduced interference from aggressor UEs. However, as shown in Table 8, TPC set 1 ensures a higher network throughput than set 2.

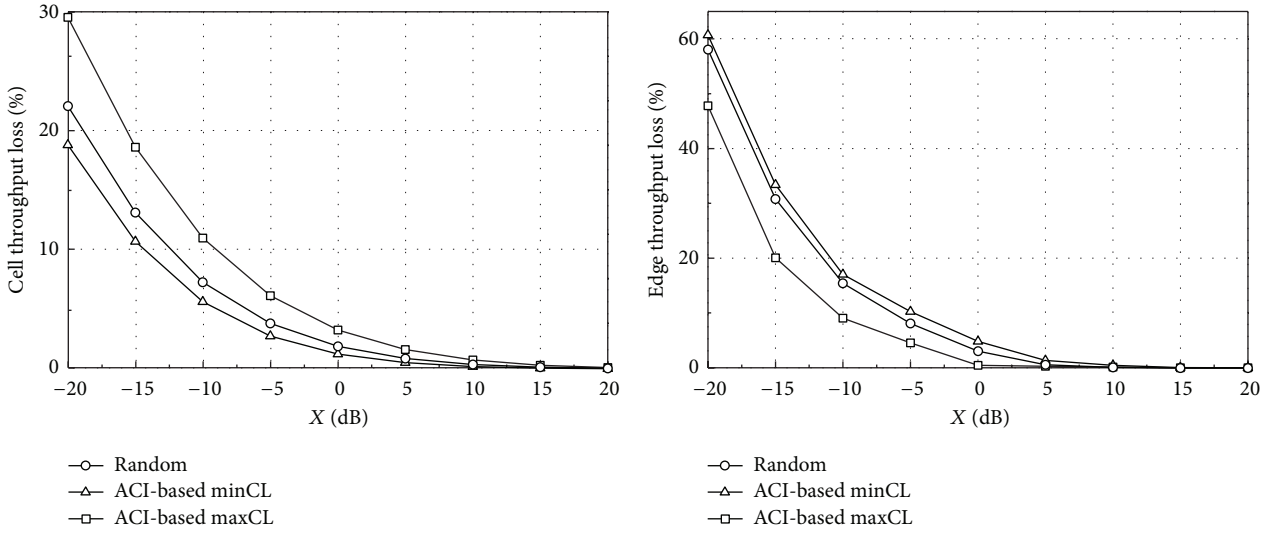
From Figures 7–10, we can see that each scheduling algorithm is a trade-off between the cell throughput loss and the edge throughput loss. If one algorithm has a higher cell throughput loss, it has a lower edge throughput loss, and vice versa. In coordinated layout, the capacity loss is in an acceptable level for X greater than −15 dB and thus there may be no specific preference on the UL scheduling algorithm, from the throughput loss point of view. In uncoordinated layout, the performance is much worse than coordinated layout. The ACI-based minCL and random scheduling algorithms have almost the same edge throughput loss, and the cell throughput loss of the minCL scheduling is lower than that of the random scheduling. The maxCL algorithm shows the highest cell throughput loss among the three algorithms, especially in uncoordinated layout with TPC set 1.

On the other hand, it is worthwhile to compare the absolute throughput performance, not the relative capacity loss. Although the exact values are not shown in this section, they can be easily calculated by combining Figures 7–10 with Table 8. The edge throughput of the random scheduling algorithm is higher than that of other two algorithms for most



(a) Cell throughput loss (b) Edge throughput loss

Figure 7: Throughput loss (TPC set 1, coordinated).



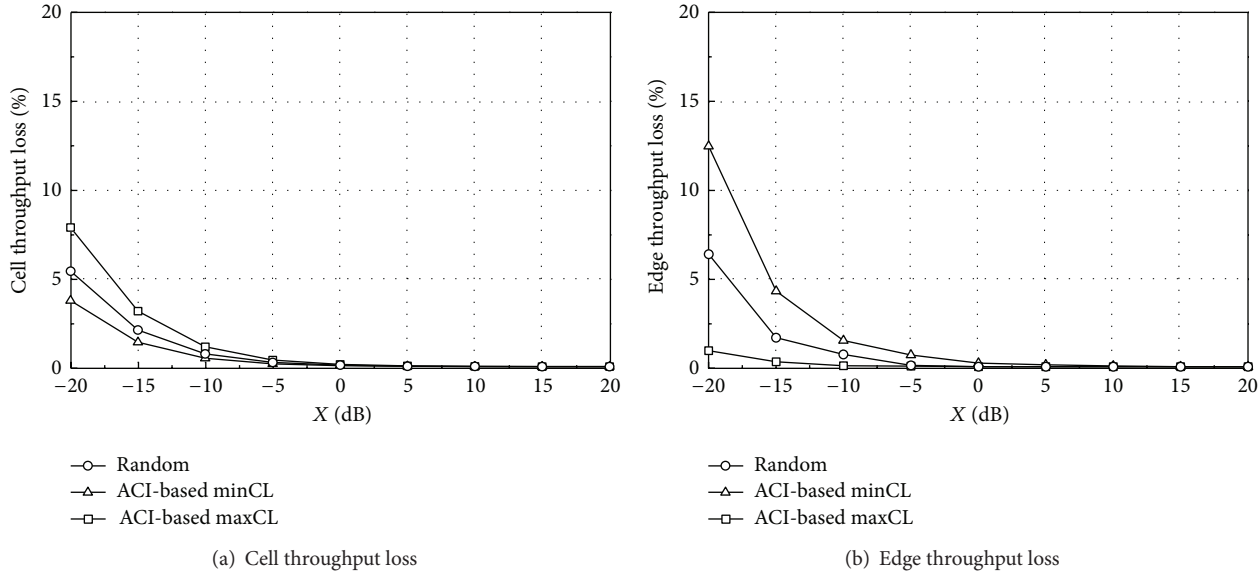
(a) Cell throughput loss (b) Edge throughput loss

Figure 8: Throughput loss (TPC set 1, uncoordinated).

of X. In addition, the minCL algorithm has the highest cell throughput, regardless of the ACIR. From the cell throughput point of view, we recommend the ACI-based minCL algorithm for UL scheduling because the cell throughput is the highest among the three algorithms and is robust against interference from the aggressor system.

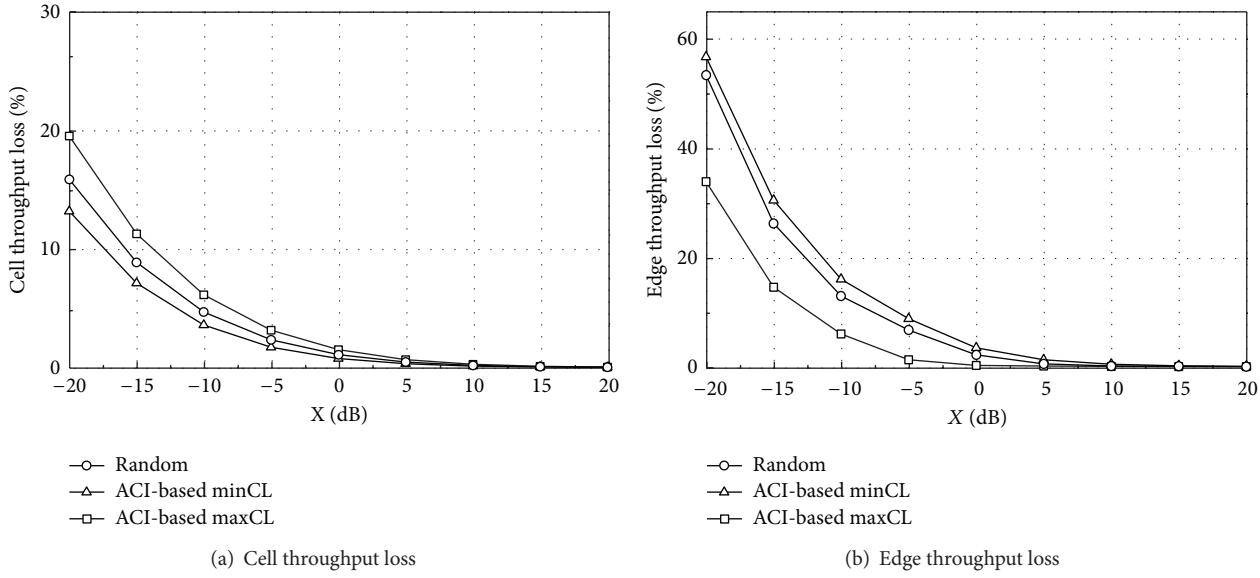
6.2. ACCL Analysis for Unsynchronized Operations. In this section, the minimum ACCL is obtained in the unsynchronized network and the operational requirements are discussed. The analysis and evaluation procedures for UL and DL are different. The UL interference to a victim BS is caused by a neighboring BS in an adjacent frequency band, whereas the DL interference to a victim UE is caused by a neighboring UE in an adjacent band.

6.2.1. Uplink ACCL Analysis. In the UL unsynchronized operation, the received signal from the nearest BS site can be a main source of interference. According to the proposed ACCL method, it is assumed that only one aggressor BS site with three sectors is located near a victim BS. The UL simulation conditions and parameters are almost the same as the synchronized operation, except the ACI model. The interference to the victim BS is modeled by the received power from the aggressor BS site attenuated by the ACCL. It is assumed that all UL subframes of TD-LTE are exposed on



(a) Cell throughput loss (b) Edge throughput loss

Figure 9: Throughput loss (TPC set 2, coordinated).

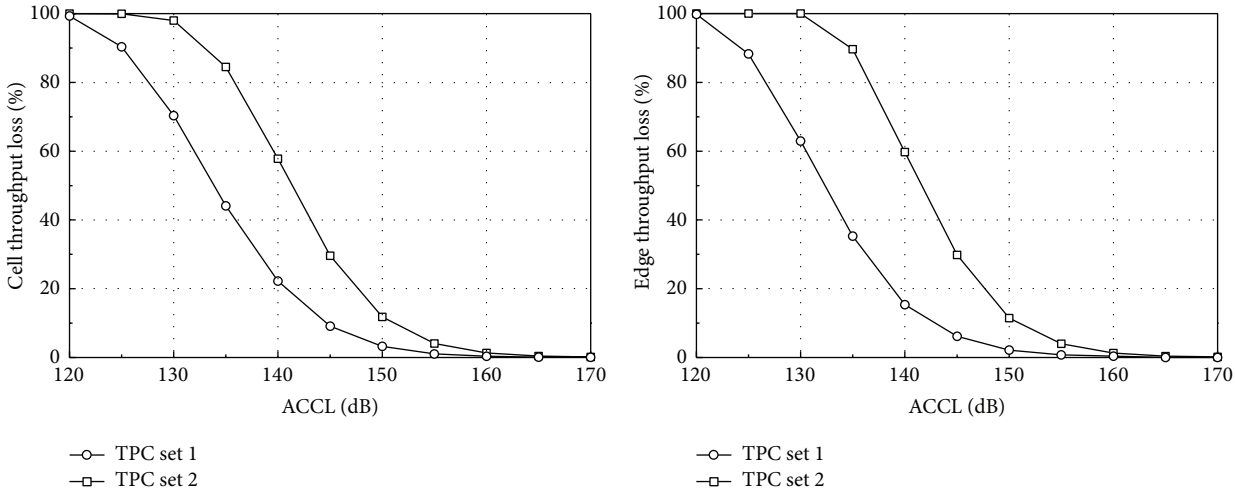


(a) Cell throughput loss (b) Edge throughput loss

Figure 10: Throughput loss (TPC set 2, uncoordinated).

interference from the aggressor BS site in order to take into account the worst-case scenario, where UL reception of the victim BS completely overlaps with DL transmission of the aggressor BS site. Other overlapping scenarios can be extended by utilizing this worst-case result because nonoverlapped subframes have a negligible impact on the throughput loss. 1,000 different UE distribution scenarios are applied to this UL simulation.

Figure 11 shows the UL throughput loss in the unsynchronized operation as a function of ACCL. For 5% capacity loss, the ACCL values of at least 148 dB and 154 dB are required for TPC sets 1 and 2, respectively. By using the minimum ACCL, the related operational requirements for the unsynchronized operation can be provided. First, the ACIR is calculated by the ACLR of the aggressor BS and the ACS of the victim BS. In [17], the recommended ACLR and ACS for BSs are 45 dB and 46.5 dB, respectively, for 10 MHz bandwidth. Thus, from (1), the ACIR is 42.7 dB. Next, the aggressor BS site has a transmit antenna gain of 7.8 to 15 dB considering a 3-sector site structure and antenna directions, whereas the victim BS has a receive antenna gain of 10.2 to 15 dB depending on antenna directions in its serving area. Then, a combined antenna gain has a range of 18 to 30 dB. We set 18 dB as the combined antenna gain assuming appropriate cell-site engineering techniques. Now, considering the ACIR and the combined antenna gain, the minimum PLs between the two BSs are 123.3 and 129.3 dB for TPC sets 1 and 2, respectively. For the BS-to-BS PL, the propagation environment is very



(a) Cell throughput loss

(b) Edge throughput loss

Figure 11: Throughput loss of unsynchronized uplink operations.

different from the ITU-R PL model in (9). In this section, we use the Stanford University Interim (SUI) PL model, which can be applied to BS-to-BS, BS-to-UE, and UE-to-UE cases, with three terrain categories common around the United States [18]. An intermediate PL category is represented by Category B in [18] and a BS antenna height for the transmitter and receiver site is set to 25 m. Then, the median PL for the SUI model can be expressed as follows:

$$L = 78.46 + 10\left(4 - \frac{65h_t}{10000} + \frac{17.1}{h_t}\right)\log_{10}\left(\frac{d}{100}\right) + 6\log_{10}\left(\frac{f_c}{2000}\right) - 10.8\log_{10}\left(\frac{h_r}{2}\right), \quad (11)$$

where d is the BS-BS separation in meters ($d > 100$ m), h_t is the transmit antenna height (m), and h_r is the receive antenna height (m). From (11), the PLs of 123.3 and 129.3 dB can be converted to the BS-BS separation of 2,970 m and 4,290 m, respectively.

If only one subframe of TD-LTE overlaps with the DL transmission of the aggressor BS, the overall capacity loss of 5% corresponds to 10% capacity loss in Figure 11. Note that Figure 11 is based on the completely overlapping scenario and TD-LTE is assumed to have two UL subframes. Thus, the minimum ACCL values become 145 and 151 dB for TPC sets 1 and 2, respectively. Assuming the same values of ACIR and the combined antenna gain, the minimum PLs are 120.3 and 126.3 dB, and they correspond to the BS-BS separation of 2,500 m and 3,590 m, respectively. The estimated distance between the two BSs may not be acceptable in an urban network environment, and some special cell-site engineering techniques may be required to reduce the interference. The estimated minimum distance can be different when other PL models and environments are considered.

6.2.2. Downlink ACCL Analysis. In the DL unsynchronized operation, the received signal from the nearest aggressor UE is considered a dominant DL interference source to a victim UE. According to the proposed ACCL method, it is assumed that only one aggressor UE is located near a victim UE and transmits at 23 dBm over entire bandwidth of 10 MHz bandwidth. The DL simulation conditions and parameters are almost the same as the synchronized operation, but the ACI is generated by the aggressor UE near the victim UE. The interference to the victim UE is modeled by the received power from the aggressor UE attenuated by the ACCL. The DL performance is related to a distance between the BS and UE in the victim system, and the same amount of ACI can cause a different impact on the DL performance depending on the distance. Thus, it is reasonable to evaluate the DL performance according to the distance between the BS and the victim UE. For simplicity, each victim UE is located in the direction of the antenna main lobe. In addition, it is assumed that all DL subframes of TD-LTE are exposed to the interference from the aggressor UE. Other overlapping scenarios can be extended by utilizing this worst-case result, as in the UL ACCL analysis.

Figure 12 shows the DL throughput loss of a victim UE in the unsynchronized operation as a function of ACCL at a different distance (D) from the BS. Because one UE is located at a fixed distance from the BS, only the average throughput loss is illustrated in Figure 12 after simulations with 500 different random seeds. The throughput loss decreases as the ACCL increases, and the victim UE far from the BS is very susceptible to the ACI. For D = 100, 200, and 500 m, the minimum ACCL values required for 5% throughput loss are 88, 90, and 96 dB, respectively.

By using the minimum ACCL, the related operational requirements can be obtained. First, the ACIR is calculated by the ACLR of the aggressor UE and the ACS of the victim UE. In [19], the recommended ACLR and ACS for UEs are 30 dB and 33 dB, respectively, for 10 MHz bandwidth. Thus the ACIR is 28.2 dB from (1). Because a UE has an omnidirectional antenna, $G_a = G_v = 0$ dBi. Considering the ACIR and the antenna gain, the minimum PLs for D = 100, 200, and

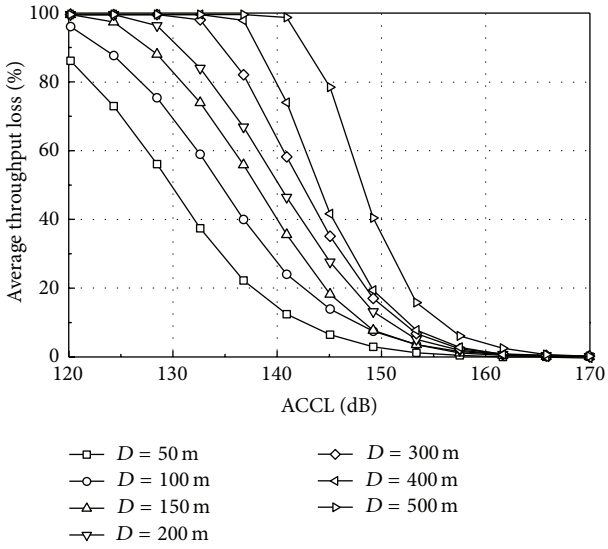


Figure 12: Throughput loss of unsynchronized downlink operations.

500 m are 59.8, 61.8, and 67.8 dB, respectively. The SUI PL model in (11) may not be accurate if the distance between the transmitter and receiver is less than 100 m. For the UE-to-UE PL, an empirical radio propagation model can be used for low height UE antennas [20]. The model considers line-of-sight (LOS) and nonline-of-sight (NLOS) propagation paths. In this analysis, the LOS model is adopted to consider a high interference scenario and given as follows:

$$L_{\mathrm{LOS}} = 4.62 + 20\log_{10}\left(\frac{4\pi}{\lambda}\right) - 2.24h_t - 4.9h_r + 29.6\log_{10}(d), \quad (12)$$

where d represents the UE-UE distance (m), λ is the wavelength, and h_t and h_r represent the transmitter and receiver height (m), respectively. This empirical model is applicable when both the transmitter and the receiver lie up to 3 m from the local ground. Assuming a UE height of 2 m, the PLs of 59.8, 61.8, and 67.8 dB can be converted to the UE-UE distances of 11, 13, and 21 m, respectively, from (12).

If the UL transmission from the aggressor UE partly overlaps with TD-LTE DL subframes, the required ACCL can be reduced. For example, DwPTS may overlap with WiMax UL as discussed in Section 3. If only DwPTS overlaps with the UL transmission of the aggressor UE, the overall capacity loss of 5% can be inferred from the completely overlapping scenario and it corresponds to 20% loss in Figure 12. Then, the minimum ACCL values are 77, 83, and 89 dB for $D = 100$, 200, and 500 m, respectively, and the minimum PLs are 48.8, 54.8, and 60.8 dB, respectively. They are converted to the distances of 5, 8, and 12 m, respectively.

When compared to the UL unsynchronized operation, the minimum PLs for the DL unsynchronized operation are relatively small. The reason is that the aggressor transmit power (23 dBm) is not as high as the BS (46 dBm), and the omnidirectional antenna is assumed for UEs whereas the BS has directional antennas with high gains. However, mobile operators cannot locate the UE at a specific position to guarantee a minimum distance between two UEs. On the other hand, cell-site engineering may be possible for the UL unsynchronized operation by cooperation between mobile operators.

7. Conclusions

In this paper, coexistence issues between TD-LTE and other systems in adjacent spectrum were analyzed and operational requirements were provided with the following research objectives. First, frame synchronization between TD-LTE and WiMax was discussed by investigating possible combinations of TD-LTE and WiMax configurations. TD-LTE configuration 1 can be a strong candidate for coexistence with WiMax because some frame configurations do not overlap with WiMax at all. If UpPTS or DwPTS is disabled, more coexistence candidates can be supported in TD-LTE configurations 1 and 2. Second, an uplink scheduling algorithm was proposed to make the UL transmission robust against the adjacent channel interference by utilizing the interference leakage pattern of transmitters. From the cell throughput point of view, the proposed minCL scheduling is recommendable for UL scheduling because the corresponding cell throughput is higher than that of other algorithms and is robust against interference from the aggressor system. Third, an adjacent-channel coupling loss (ACCL) method was introduced to estimate the minimum requirements for coexistence when two networks are not synchronized. The minimum ACCL can be used to optimize the network parameters and converted to the minimum path loss or the minimum distance between two BSs or two UEs. From the analysis and simulation results, we can see that coexistence of TD-LTE with other systems is feasible if the two networks are synchronized. For the unsynchronized case, the BS-to-BS interference may not be acceptable for a normal operation in an urban environment and some special cell-site engineering techniques may be required to reduce the interference. The UE-to-UE interference is not significant compared to the BS-to-BS interference.

Conflict of Interests

The authors declare that they have no conflict of interests.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2010496). This research was also supported by SK Telecom.

References

[1] E. Dahlman, S. Parkvall, and J. Skold, *4G LTE/LTE-Advanced For Mobile Broadband*, Elsevier Academic Press, 2011.

[2] R. Ratasuk, A. Ghosh, W. Xiao, R. Love, R. Nory, and B. Classon, "TDD design for UMTS long-term evolution," in *Proceedings of the IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '08)*, pp. 1–5, September 2008.

[3] F. Wang, A. Ghosh, C. Sankaran, P. J. Fleming, F. Hsieh, and S. J. Benes, "Mobile WiMAX systems: performance and evolution," *IEEE Communications Magazine*, vol. 46, no. 10, pp. 41–49, 2008.

[4] S. Shyam Wagle, M. Ade, and M. Ghazanfar Ullah, "Network Transition from WiMax to LTE," *Journal of Computing*, vol. 3, no. 1, pp. 66–70, 2011.

[5] WiMax Forum, "Requirements for WiMax Coexistence with LTE Network," WMF-T31-132-v01 13, 2011.

[6] Y. Tian, X. Zhang, J. Wang, G. Cao, and D. Yang, "Coexistence study on LTE-advanced system interfering 3G/B3G TDD systems," in *IEEE Wireless Communications and Networking Conference (WCNC '12)*, pp. 2729–2734, April 2012.

[7] Z. Ruiming, Z. Xin, L. Xi, P. Qun, F. Yinglong, and Y. Dacheng, "Performance evaluation on the coexistence scenario of two 3GPP LTE systems," in *Proceeedings of the IEEE 70th Vehicular Technology Conference Fall (VTC '09)*, Anchorage, Alaska, USA, September 2009.

[8] P. Chang, Y. Chang, J. Zhou, and D. Yang, "Interference analysis of TD-LTE and TD-SCDMA system coexistence," in *Proceedings of the 2nd International Conference on Networks Security, Wireless Communications and Trusted Computing (NSWCTC '10)*, pp. 77–80, Hubei, China, April 2010.

[9] ECC, "Practical guidance for TDD networks synchronization," ECC PT1(13)048, draft ECC Report, April 2013.

[10] P. Chitrapu, *Wideband TDD: WCDMA for the Unpaired Spectrum*, John Wiley & Sons, New York, NY, USA, 2004.

[11] S. Pike, "Liaison statement to ERC TG1," TSGR#3(99) 23, TSG-RAN, 1999.

[12] 3GPP TR 36.942, "Evolved Universal Terrestrial Radio Access (E-UTRA), Radio Frequency (RF) system scenarios (Release 10)," TR 36.942 v10.3.0, 2012.

[13] A. F. Molisch, *Wireless Communications*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2010.

[14] J. C. Ikuno, M. Wrulich, and M. Rupp, "System level simulation of LTE networks," in *Proceedings of the IEEE 71st Vehicular Technology Conference (VTC '10)*, pp. 1–5, Taipei, Taiwan, May 2010.

[15] 3GPP TS 36.213, "Evolved Universal Terrestrial Radio Access (E-UTRA), Physical layer procedures (Release 10)," TS 36.213 v10.10.0, 2013.

[16] A. Ubisse and N. Ventura, "Modeling a link level simulator for long term evolution uplink," in *The Southern Africa Telecommunication Networks and Applications Conference (SATNAC '11)*, East London, South Africa, September 2011.

[17] 3GPP TS 36.104, "Evolved Universal Terrestrial Radio Access (E-UTRA), Base Station (BS) radio transmission and reception (Release 10)," TS 26.104 v10.11.0, 2013.

[18] A. Sadek and S. Shellhammer, "SUI Path-Loss Model for Coexistence Study," IEEE 802.19-08/0010r0, IEEE P802.19 Wireless Coexistence, 2008.

[19] 3GPP TS 36.101, "Evolved Universal Terrestrial Radio Access (E-UTRA), User Equipment (UE) radio transmission and reception (Release 10)," TS 36.101 v10.11.0, 2013.

[20] K. Konstantinou, S. Kang, and C. Tzaras, "A measurement-based model for mobile-to-mobile UMTS links," in *PRoceedings of the IEEE 65th Vehicular Technology Conference (VTC '07)*, pp. 529–533, Dublin, Ireland, April 2007.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 617087, 9 pages
http://dx.doi.org/10.1155/2014/617087

Research Article

Reputation Revision Method for Selecting Cloud Services Based on Prior Knowledge and a Market Mechanism

Qingtao Wu, Xulong Zhang, Mingchuan Zhang, Ying Lou, Ruijuan Zheng, and Wangyang Wei

Information Engineering College, Henan University of Science and Technology, Luoyang 471023, China

Correspondence should be addressed to Qingtao Wu; wuqingtao.cn@hotmail.com

Received 12 October 2013; Accepted 4 November 2013; Published 17 February 2014

Academic Editors: W. Sun, G. Zhang, and J. Zhou

The trust levels of cloud services should be evaluated to ensure their reliability. The effectiveness of these evaluations has major effects on user satisfaction, which is increasingly important. However, it is difficult to provide objective evaluations in open and dynamic environments because of the possibilities of malicious evaluations, individual preferences, and intentional praise. In this study, we propose a novel unfair rating filtering method for a reputation revision system. This method uses prior knowledge as the basis of similarity when calculating the average rating, which facilitates the recognition and filtering of unfair ratings. In addition, the overall performance is increased by a market mechanism that allows users and service providers to adjust their choice of services and service configuration in a timely manner. The experimental results showed that this method filtered unfair ratings in an effective manner, which greatly improved the precision of the reputation revision system.

1. Introduction

The rapid developments of cloud computing means that cloud services have become the main computing mode on the Internet. Many services have been deployed to provide similar functionalities. However, the problem of identifying reliable services has attracted the attention of researchers [1, 2]. Thus, the concepts of trust and reputation [3] have been introduced to assess the reliability of cloud services.

Reputation is a subjective assessment of a cloud service, which is based on individual experience or the recommendations of other users. Reputation and trust are dynamic, which makes the construction of an evaluation standard a challenging task. In addition, the occurrence of malicious evaluations [4, 5], deliberate praise, and the personal preferences of users means that a standardized reputation value may differ from the true value.

Recently, various reputation revision systems have been proposed to address the challenges posed by open and dynamic cloud service environments [6]. Most of these systems are focused on the calculation of reputation ratings, reputation management, experience, and other features of dynamic environments [7–10] that might provide an appropriate reference for users. However, the existence of unfair ratings greatly affects the accuracy of trust evaluations. Currently, these reputation models are mainly tending to the accuracy of trust evaluations [11, 12]; however, these existing methods are limited by personality preference.

Based on historical user evaluations and preferences related to specific requirements, we propose a method that revises the reputation rating. This prior knowledge is combined with a filtering algorithm based on the similarities of evaluations, which can distinguish between unfair ratings in an effective manner. This algorithm also includes a market mechanism that allows users and services to act as buyers and sellers. Our method uses only the user feedback ratings data related to a service. Extensive experiments showed that our method distinguished and filtered unfair ratings correctly, while it could also recommend appropriate services for a specific user based on their preferences in a dynamic market environment.

The remainder of this paper is organized as follows. In Section 2, we provide an overview of related work on reputation revision methods. Section 3 describes the problem and

provides definitions related to a specific scenario. Section 4 explains our proposed reputation revision framework and Section 5 presents the experimental results. In Section 6, we conclude with the discussion and we summarize the contributions of this study.

2. Related Work

Reputation is a term that has different meanings in various domains. In our study, reputation is defined as an indicator of whether a user is willing to select a service based on the evaluations of other users. Thus, the result of a reputation evaluation will affect the decision about whether to interact with a service provider. Feedback related to previous interactions among users and service providers is collected by a reputation system to predict its future reliability [13, 14]. Reputation evaluations are essential parts of many recommendation systems [15]. Many systems are located in a central server, which can access, collect, and evaluate historical reputation scores from a large number of users [16–19]. Yang et al. [9] proposed a reputation management framework for service selection based on similarity theory. Different weights were set to compute the reputation based on the unique recommendations of users. However, the management framework was designed for a centrally controlled server, which was not suitable for a dynamic cloud service environment.

Several reputation evaluation approaches have been proposed for distributed systems. Ghaffarinejad and Akbari [10] introduced a distributed reputation mechanism, which was based on a number of special reputation centers. Each special reputation center collected reputation information for predetermined services offered by different service providers. However, this method was still somewhat centralized. Faniyi and Bahsoon [7] proposed a decentralized resource control mechanism, which introduced a market-oriented cloud computing architecture. However, the reputation system was vulnerable to whitewashing, incorrectly reported feedback, and collusion attacks (where several users coordinate their feedback to manipulate reputation information).

Kussul et al. [20] focused on the analysis of security threats in trust models and assessed the most important and critical security threats for a utility-based reputation model based on grids. Dong-Sheng et al. [21] proposed a distributed trust mechanism, which calculated two reputation values (for a seller and a buyer) for each node in an iterative manner based on the transaction history. This mechanism could rapidly reduce the reputation values of malicious nodes and prevent collusion attacks. To construct a trustworthy computing environment, Gui et al. [19] proposed a penalty-incentive mechanism based on a repeated game theory, which included a rule related to rewards and punishments. Hawa et al. [16] introduced enhanced reputation-based cooperation incentives, which facilitated better detection and control of free riders. This approach enhanced the scalability and fairness of the system.

Kim and Phalak [18] proposed a computational trust framework for predicting the degree of trust. In addition, Su et al. [12] developed a priority-based trust model, which determined the trustworthiness of a service provider based on designated referees and its historical performance. This method only used third-party evaluations of the previous overall performance of the service and did not consider the individual preferences of the current user. Gorner et al. [22] proposed several improvements for trust modeling, including limiting the size of the advisor network either by specifying the maximum size of a buyer's advisor network or by setting a minimum trustworthiness threshold for agents accepted into the advisor network.

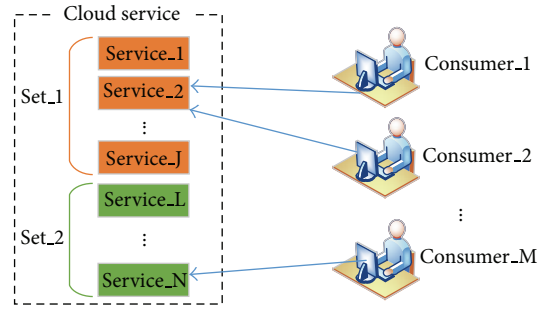


Figure 1: Cloud service selection model.

Another previous study [23] proposed a method for revising reputation values that calculated the reputation based on the difference between the advertised quality of service (QoS) provided by service providers and the evaluations made by consumers. Next, the consumers were sorted based on the reputation ratings they provided and those consumers that might be involved with collusion were mined using an association rules algorithm. Finally, the updated reputation was recalculated and saved in the reputation center.

3. Overview

3.1. Problem Description and Scenario Definition. Our method was developed to overcome some of the limitations of existing reputation mechanisms and is suitable for open and dynamic environments. Before providing the details of our proposed reputation revision mechanism, it is necessary to define the scope of our study and to explain some of the definitions used in this paper.

A cloud computing system provides services according to a third-party mechanism. Thus, users only need to be concerned with the service provided by the cloud. The cloud service selection model shown in Figure 1 contains two agent types, that is, consumer agents and service agents. Services that share the same functionality are placed in the same sets.

Users often select services from the same functional groups based on their own experiences and those of other consumers. However, unreliable evaluations may mean that the reputation of a service does not represent their actual reliability. Some service providers pay consumers to give them high scores for their services or to give low scores to their competitors. Furthermore, it may be difficult to satisfy users with specific requirements. These issues may mean that the service scores are not effective reference sources for users.

Definition 1. A service can be described as a 2-tuple; that is, Service = (Function, QoS). Function is a set of common properties, where different services are classified into separate sets. The main components of QoS include the response time (RT), cost (C), and reliability (R). In the present study, we define QoS as a 3-tuple; that is, QoS = (RT, C, R).

Definition 2. The rating score (RS) is derived from the distributed consumer agents. RT represents the degree of satisfaction with the service. RT is defined as a 3-tuple; that is, RS = (V_RT, V_C, V_R), where V_RT, V_C, and V_R are the values of the properties of the service given in Definition 1. We use percentile scores to distinguish between good and bad performance.

Definition 3. The consumer rating is the major focus, and the consumer is associated with multiple services from different providers. Thus, the consumer agent can be described as follows.

Consumer = (C_ID, $\bigcup_i$ Service_i, $\bigcup_j$ RS_j), where C_ID is the consumer's identifier and $\bigcup_i$ Service_i and $\bigcup_j$ RS_j are the services and rating scores associated with the consumer, respectively.

3.2. Overview of the Entire Reputation System. In this section, we briefly introduce our reputation system structure. The framework of the reputation system is shown in Figure 2.

In general, the framework is applicable to most of the ratings-based experiences shared on online platforms where users evaluate services with numerical ratings. Figure 2 shows that the users ratings are collected as feedback to reputation processing node and the feedback is quantified based on the QoS attributes in the knowledge repository. After normalizing the feedback data, the ratings are filtered based on similarity classification. Next, we set the preferences for abnormal users, who may have specific service requirements. The user preferences mean that the recommendation system provides the most relevant services. There is also a certain degree of punishment for collusive users, who make the results confusing when service recommendations are required. The core of the system is the reputation calculation, where we combine the result from filtering with the historical reputation to generate a reputation value that is credible and reliable. The final part of the reputation framework is reputation management. Using a market mechanism, the service provider can optimize his service configuration and the user can optimize his decision. Thus, the service quality is optimized for the overall environment using the reputation system. Filtering abnormal users also makes the reputation of the service more accurate.

4. Reputation Revision Mechanism

In this section, we describe the mechanism used to ensure a more accurate level of consumer trust in dynamic environments. As mentioned in Section 3, consumers evaluate services using numerical ratings. The proposed method estimates the degree of trust a consumer places in a service based on the consumer's preference and it filters the abnormal reputation ratings.

Table 1: Quality of service parameters for a cloud service.

	p_1	p_2	p_3	p_4
s_1	0.9	0.6	0.7	0.4
s_2	0.8	0.7	0.6	0.3
s_3	0.6	0.6	0.4	0.4
s_4	0.5	0.2	0.2	0.8

Table 2: Reputations of services evaluated by consumers.

	s_1	s_2	s_3	s_4
c_1	0.9	0.6	0.7	0.4
c_2	0.8	0.7	0.6	0.3
c_3	0.6	0.6	0.4	0.4
c_4	0.5	0.2	0.2	0.8

4.1. Filtering the Abnormal Reputation Ratings. The presence of inaccurate assessments affects the overall evaluation of a service to some extent. Thus, we use the similarity to distinguish between abnormal evaluations, which reduces the effects of abnormal reputation ratings. The Euclidean distance is the shortest length of a line in n-dimensional space, which is usually defined as the real distance between two points in n-dimensional space.

We use $S = \{s_1, s_2, \ldots, s_n\}$ to represent a service set where services share the same function. $C = \{c_1, c_2, \ldots, c_m\}$ is the set of consumers. $P = \{p_1, p_2, \ldots, p_s\}$ is the set of QoS parameters [24] for a service. We specify q_{ij} as the parameter values of p_j for s_i, as shown in Table 1. r_{ij} represents the reputation of s_j in c_i, where $0 \le r_{ij} \le 1$, as shown in Table 2.

Each row is regarded as a node. Thus, the similarity between two nodes can be represented by the Euclidean distance. If the distance between two nodes is high, the similarity will obviously be low. Thus, we use the following to compute the similarity between the service parameter and a user's evaluation:

$$\text{similarity} = \frac{1}{\left(\sum_1^n (X_i - Y_i)^2 / n\right)}. \quad (1)$$

In (1), X and Y are vectors, where X_i (or Y_i) represents the value of the i-dimension in the vector. To filter abnormal evaluations, we need to determine whether the parameters of the service configuration have changed. If the parameters of a service have changed, a distance will be generated. In general, there is no change in the configuration of the service parameters. However, if the parameters of the service configuration change, the historical reputation weighting will decline rapidly. If a consumer's evaluation is a considerable distance from the mean of all the other service evaluations made by consumers, we conclude that this consumer's evaluation is problematic and it must be filtered before further processing.

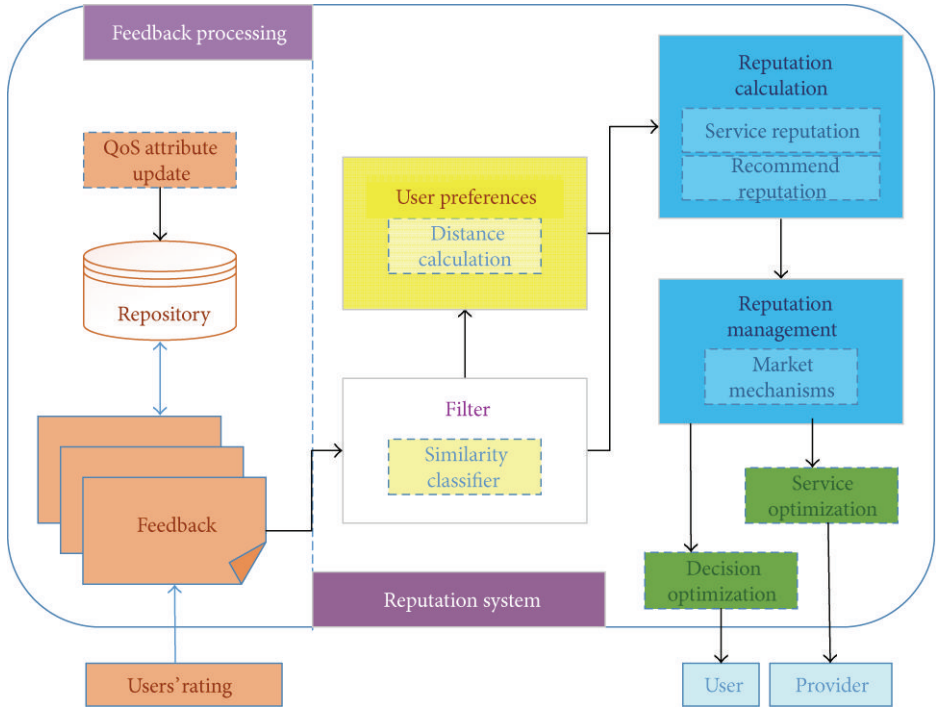


Figure 2: Reputation system framework.

4.2. Estimation of Consumer Preference. All consumer evaluations are biased to some extent, so we need to set all of the abnormal consumer preferences. In this case, bias refers to subjective evaluations of objective factors by consumers that deviate from the norm. Consumers who make frequent unfair evaluations of services have extreme preferences. Malicious evaluation refers to the intentional denial of the objective facts, undeserved praise, or unmerited negative feedback. Thus, setting preferences allows us to eliminate abnormal evaluations. However, consumers can change their preferences if they have unusual requirements.

The preference weight (PW) represents a consumer's personal service preferences. A higher PW indicates that the rating for a service differs greatly from that of most users. The PW is constructed using two factors, that is, C Value and Ref. C Value represents a consumer's rating score, while Ref describes the contribution to the PW made by the total consumer scores. PW is defined using the following formula:

$$\text{PW} = C\,\text{Value} - \text{Ref}. \tag{2}$$

C Value is the service's current reputation score given by a consumer. According to (3), Ref is the reference value for the overall score. The consumer rating scores have a normal distribution. Ref is the normal position parameter that describes the location of the central tendency of the normal distribution. rateScore = Ref is the normal to the axis of symmetry, which is completely symmetrical. In a normal distribution, the mean, median, and mode are the same; that is, they are equal to Ref. σ indicates the data distribution with a normal degree of dispersion:

$$f\,(\text{rateScore}) = \frac{1}{\sqrt{2\pi}\sigma} e^{(\text{rateScore}-\text{Ref})^2/2\sigma^2}, \tag{3}$$

$$\text{Ref} = \frac{1}{N}\sum_{i=1}^{N} \text{rateScore}_i, \tag{4}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{rateScore}_i - \text{Ref})^2}. \tag{5}$$

The rateScore distribution of a service is shown in Figure 3 (the data used to produce the figure were based on the evaluation of a service). The PWs are set for overall consumer but they just apply to a small minority of consumers who have obviously deviation to the true quality of service level. According to Figure 3, we simply need to set separate PWs for the reputations that are outside the confidence interval. The confidence interval here refers to the proportion of real evaluation,which can be used as the reliability estimation that is usually defined by a large amount of observation. Obviously, an excessive proportion may mislead the service evaluation while a too little one leads to an unauthentic conclusion.

If σ and Ref are known, the service rateScore follows a normal distribution N (Ref, σ). We define the main interval using (6). α is the width of the main interval and δ is the probability of the evaluation given by most users:

$$P\left\{|\text{rateScore} - \text{Ref}| \le f\left(\frac{\alpha}{2}\right)\right\} = \delta. \tag{6}$$

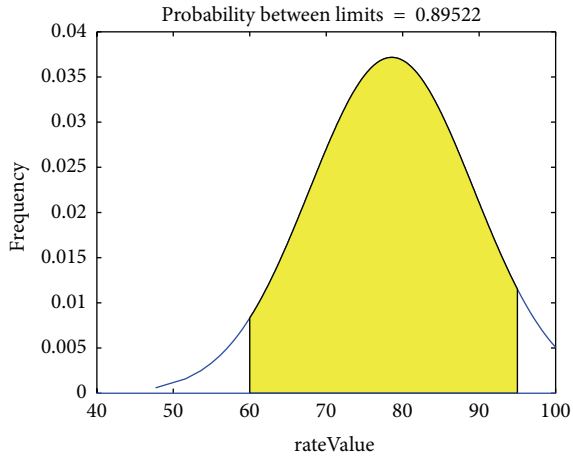


Figure 3: Distribution of a service's rateScore given by consumers.

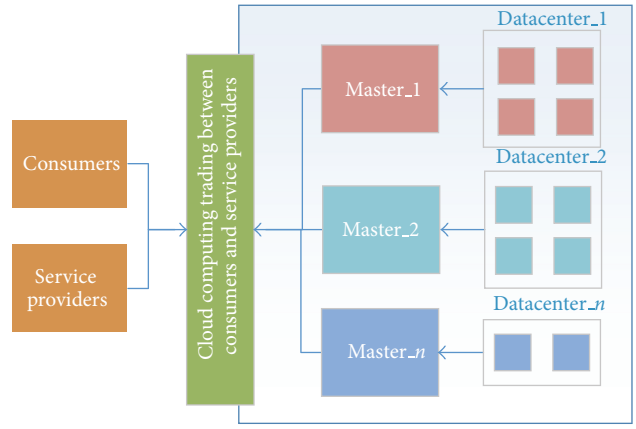


Figure 4: Cloud service market structure.

Thus, we only need to set an appropriate value for δ to distinguish between abnormal evaluations. So we can give the abnormal ratings malicious preference treatment and reduce the weight of them in the whole calculation of service reputation. Suitable recommendations of appropriate services can be made based on the global reputation of a service and the preferences of users, while the system also punishes malicious users.

4.3. Calculation of the Integrated Reputation. The reputation rating of a cloud service given by a specific user is the weighted average of the directly experienced reputation and the historical reputation, which is obtained using (7). RoS_{ij} is the reputation rating based on the direct experience of service j given by consumer i. C_{ij} is the current reputation rating of service j given by consumer i. HsR_{ijk} is the kth reputation rating of the historic reputation rating given by consumer i to service j, and l is the total number of times that consumer i rates service j. β is the weight factor of the current reputation, $0 \le \beta \le 1$

$$\mathrm{RoS}_{ij} = (1-\beta)\, C_{ij} + \frac{\beta}{l}\sum_{k=1}^{l} \mathrm{HsR}_{ijk}. \tag{7}$$

We need to integrate the evaluations of the same service to calculate a service's reputation using the method proposed earlier. After calculating the mean score and the variance (using (8) and (9), resp.) of a service, we can determine its normal distribution (using (10)):

$$\mathrm{Ref}_j = \frac{1}{r}\sum_{i=1}^{r} \mathrm{RoS}_{ij}, \tag{8}$$

$$\sigma^2 = \frac{1}{r}\sum_{i=1}^{r} \left(\mathrm{RoS}_{ij} - \mathrm{Ref}_j\right)^2, \tag{9}$$

$$\mathrm{RoS}_j \sim \left(\mathrm{Ref}_j, \sigma^2\right). \tag{10}$$

Ref_j represents the mean score of the service where the ID is j. r represents the total number of service ratings. RoS_j is the score for j's service. We need to set the value of δ to control the confidence interval. Filtering is more accurate if δ is a low value. However, this increases the complexity of processing the preferences. The specific PW is obtained using the following:

$$\mathrm{PW}_{ij} = \begin{cases} 0 & \left|\mathrm{RoS}_j - \mathrm{Ref}_j\right| \le f\left(\dfrac{\alpha}{2}\right) \\ C_{ij} - \mathrm{Ref}_j & \text{else.} \end{cases} \tag{11}$$

PW_{ij} represents the ith consumer's PW for the jth service. We obtain a more objective service reputation rating after eliminating the personal preferences. The reputation rating of service j is calculated using the following:

$$\mathrm{RoS}_j = \frac{1}{r}\sum_{i=1}^{r} \left(\mathrm{RoS}_{ij} - \mathrm{PW}_{ij}\right). \tag{12}$$

If a consumer wants to access a service, the trust system provides a recommendation based on the reputation using (13). The reputation rating of each consumer is not the same because they have specific preferences, which also facilitates the punishment of malicious evaluations. RR_{ji} represents the reputation rating of service j recommended to consumer i:

$$\mathrm{RR}_{ji} = \mathrm{RoS}_j + \mathrm{PW}_{ij}. \tag{13}$$

4.4. Dynamic Reputation Optimization Using a Market Mechanism. Cloud services are reliant on the dynamic and distributed cloud environment. Thus, a trust system needs to adapt to open and changing conditions. Service providers will change the configurations of their services to meet consumer demands, and the consumers will have variable service preferences. Different consumers use the same service, which runs in different data centers. We use a market mechanism [25] to construct a reputation optimization method that provides a better QoS. The cloud service market environment is shown in Figure 4. The data center, which is a distributed cloud resource provider, includes many nodes. The master node is the seller's agent, which is responsible for changing the service configuration and optimizing the service performance.

The market mechanism includes buyers and sellers. The buyers are consumers and the sellers are service providers in our method. We state the objectives of the buyers and sellers in the cloud market. The buyers' goal is to satisfy their personal demand for a service with a high reputation. The sellers' goal is to receive a better evaluation by optimizing their configuration to maximize the number of tasks completed successfully. The global objective of the market is to supply each buyer with a reliable service while minimizing the costs of the sellers.

The buyers and sellers have different demands, and the cloud service market is open and dynamic. We define the buyer's personal valuation using a utility function, which for buyer b using service i is defined as $P_b(i) = -\alpha\text{Cost}(i) - \beta\text{Time}(i) + \gamma\text{Profit}(i)$, where Cost($i$) and Time($i$) represent the price and response time for service i, respectively, and Profit(i) is the payoff derived from the usage of service i. The seller's utility function is defined by $P_s(i) = O(i) - C(i)$, where $O(i)$ represents the reputation and fees received for service i and $C(i)$ is the cost of service i. The seller's objective is to maximize $P_s(i)$.

After each transaction, the sellers are rated based on their performance in the task allocated to them. The buyer decides this rating by comparing the actual service completion time with the expected time, as well as with other services. The global objective of the market is to maximize the number of services that are supplied satisfactorily by service providers. This objective is defined using the following:

$$G(i, j) = \text{Max} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(R_{ij} - A_{ij}\right), \tag{14}$$

where R_{ij} is the reputation rating given by buyer i and A_{ij} is the actual performance of service j. Algorithm 1 shows how the market transactions operate between buyers and sellers in pseudocode.

The buyer's timely and correct feedback facilitates the continuous optimization of the service in a dynamic market. If new services and buyers join the market, the system needs to be updated and new items will be available for the buyer to choose. After the service has been updated, a recommendation will be given to the buyer, but the reputation continues to be accumulated.

The market contains many submarkets. Submarkets are synchronized regularly using older market information, which reduces the frequency of management. A service can be present in different submarkets, which are distributed and flexible.

```
(1)  buyer = {b_1, b_2, ..., b_k}; //the buyer's set
(2)  seller_service = {s_1, s_2, ..., s_l}; //the service provider's
(3)  services comprise the seller's set.
(4)  maxProfitofBuyer = 0; //buyer pursues their best
     interests
(5)  maxReputationRepay = {0, 0, ..., 0}; //the seller
(6)  receives a high payoff from the buyer, which forms
     a set
(7)  for i = 1 to k
(8)     for j = 1 to ldo
(9)        if b_i has used s_j
(10)           RS[ij] = the rating provided by b_i for s_j;
(11)                                    //collect the
     ratings
(12)       else RS[ij] = null;
(13)       end if
(14)    end for
(15) end for
(16) for j = 1 to l
(17)     for i = 1 to k do
(18)          RS[j] = comput_reputation (RS[i, j]);
(19)                          //compute service j's
     reputation
(20)    end for
(21) end for
(22) if same_function(service_set) //service_set includes
(23)                  //several services with the same function
(24)    compare(maxProfitofBuyer, p_b (service_set));
(25)    //return the buyer pursues a better quality service
(26) end if
(27) for j = 1 to l do
(28)    F_s[j] = check the recommendation frequency of s_j;
(29)    if F_s[j] < LOW
(30)       change the service provider to update the
(31) service configuration;
(32) //then after the update, the service will be renewed;
(33)    end if
(34) end for
```

Algorithm 1: Trading between buyers and sellers.

Table 3: Quality of service (QoS) parameters for cloud services.

rateID	consumerID	serviceID	rateValue
int	int	int	float (calculated from the user's evaluation of the QoS attributes)

5. Experiments and Analysis

We produced a simulation program in Java to validate our reputation revision method. We simulated several service providers, which had different services with the same functions, as well as transaction behaviors of consumers. Our experiment comprised 500 consumers and four services, where each consumer rated the services they used after each transaction. The results were saved in the format shown in Table 3. To filter abnormal evaluations, our reputation revision method was used to cluster and mine the data in the transaction records after each round of transactions.

In our experiments, we tested two main hypotheses: (1) the consumer ratings of services follow a normal distribution; (2) the use of the approach described in Section 4 improves the accuracy of reputation ratings.

5.1. Distribution of Consumer Ratings. To verify the service score distribution, the consumers were divided into three types (i.e., bad, right, and good) who gave different ratings to services in our experiment. The bad consumers gave a score

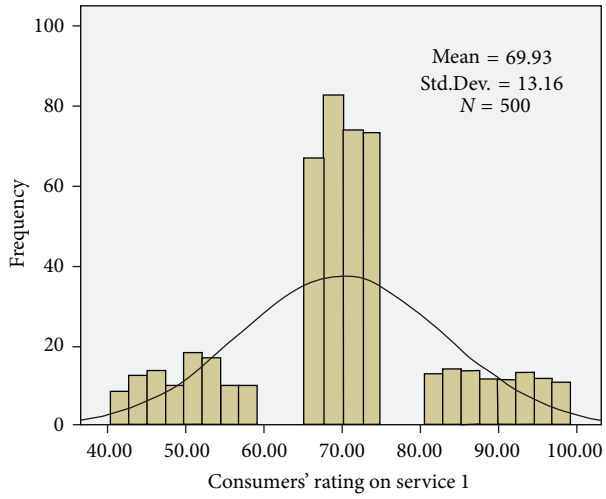


Figure 5: Distribution of rateValue for service 1 in the experiments (N = 500 users).

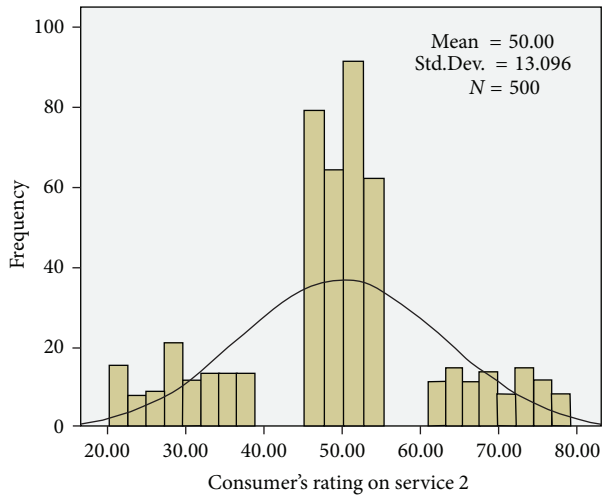


Figure 6: Distribution of rateValue for service 2 in the experiments (N = 500 users).

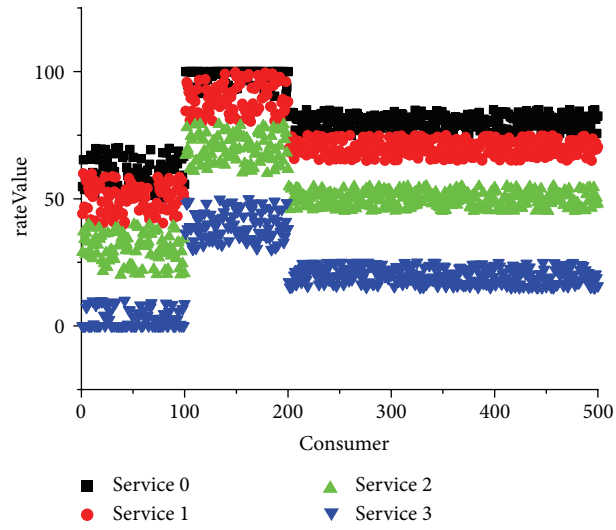


Figure 7: Ratings for four services given by 500 consumers.

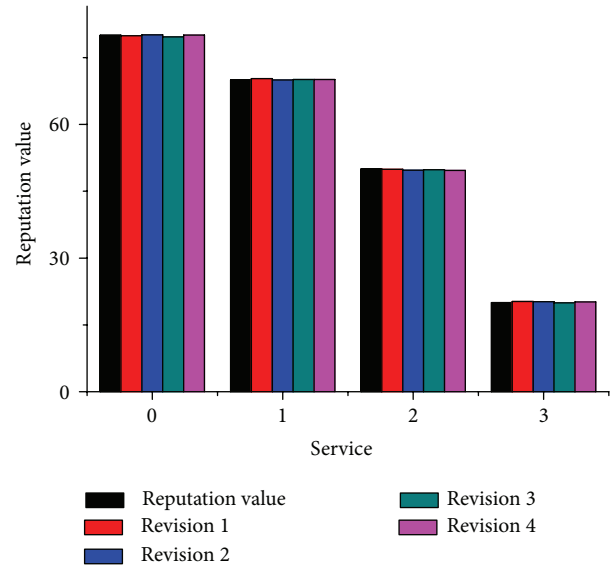


Figure 8: Four rounds of revised reputation.

below the service's QoS level and the good consumers gave a score higher than the QoS level. The right consumers gave a score around the QoS level. All of the consumers gave random scores from the corresponding interval.

The following QoS metrics were considered in the experiments: cost, response time, and execution time. The cost value was selected randomly from the range $[1, 10]$ \$ and the response time was selected randomly from the range $[1, 1000]$ s. The execution time was set as the functional unit/single unit. The experimental data were analyzed using SPSS. Figures 5 and 6 show the score distributions for the two services.

The results of the data analysis showed that the distribution of the service rating scores was not a strictly normal distribution. However, we used the confidence interval (the formula proposed in Section 4) to identify the right consumer ratings and to filter the bad and good ratings. Some value intervals were empty because the consumer ratings were based on the QoS level of the service and the consumers' preferences.

5.2. Reputation Revision. In this experiment, we validated whether our method improved the accuracy of the service reputation ratings. In this experiment, four services provided the same function, but the QoS value belonged to different classes. The actual reputation of each service equaled the QoS value. We considered the revised reputation results for the four services after four cycles of revisions had been calculated. Figure 7 shows the ratings made by 500 consumers for each service they used. Figure 8 shows the results after four cycles of reputation revision.

Figure 7 shows that the four services had different rateValue levels and the ratings also reflected the three different

types of consumers, who had distinct scoring trends. Figure 8 demonstrates the accuracy and stability of the reputation revision method. The experiment showed that the revision method could identify abnormal reputation ratings, which were filtered from the overall evaluations to improve the accuracy of the service.

6. Conclusions

In this study, we developed a reputation revision method for cloud services based on the confidence interval of a normal distribution of ratings and a market mechanism. To address the problem of abnormal evaluations, we used prior knowledge to distinguish between different types of consumers before filtering dishonest ratings and setting the preferences for consumers.

There are still some limitations in the dynamic provision of services and reputation management [26] for each distributed submarket. The dynamic provision of services will add more complexity to the processing of historical data. It will also be necessary to consider reputation management in the distributed submarkets.

Future research should investigate the development of a mechanism for the dynamic evolution of reputation ratings and the application of this method to large-scale service-oriented systems. The growing number of services and users means that the configuration parameters of services and the user base are changing constantly, so reputation evaluations should be capable of evolving. We also expect that this method could be deployed in a real cloud service application system. Further verification of this reputation revision mechanism will help to identify new problems.

Conflict of Interests

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

Acknowledgments

The authors would like to thank the reviewers for their detailed reviews and constructive comments, which have helped improve the quality of this paper. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant nos. 61003035, U1204614, and 61370221 and in part by the Plan for Scientific Innovation Talent of Henan Province under Grant no. 124100510006.

References

[1] Z. Yan, Y. Chen, and Y. Shen, "A practical reputation system for pervasive social chatting," *Journal of Computer and System Sciences*, vol. 79, no. 5, pp. 556–572, 2013.

[2] A. Satsiou and L. Tassiulas, "Reputation-based resource allocation in P2P systems of rational users," *IEEE Transactions on Parallel and Distributed Systems*, vol. 21, no. 4, pp. 466–479, 2010.

[3] T. Chen, A. Bansal, and S. Zhong, "A reputation system for wireless mesh networks using network coding," *Journal of Network and Computer Applications*, vol. 34, no. 2, pp. 535–541, 2011.

[4] B. Qureshi, G. Min, and D. Kouvatsos, "Countering the collusion attack with a multidimensional decentralized trust and reputation model in disconnected MANETs," *Multimedia Tools and Applications*, vol. 66, no. 2, pp. 303–323, 2013.

[5] Z. Li, H. Shen, and K. Sapra, "Leveraging social networks to combat collusion in reputation systems for peer-to-peer networks," in *Proceedings of the 25th IEEE International Parallel and Distributed Processing Symposium (IPDPS '11)*, pp. 532–543, May 2011.

[6] L. Jiang, L. Ding, J. Liu, and J. Chen, "Reputation rating modeling for open environment lack of communication by using online social cognition," *Journal of Network and Computer Applications*, vol. 35, no. 3, pp. 881–891, 2012.

[7] F. Faniyi and R. Bahsoon, "Self-managing SLA compliance in cloud architectures: a market-based approach," in *Proceedings of the 3rd international ACM SIGSOFT symposium on Architecting Critical Systems*, pp. 61–70, ACM, 2012.

[8] X. Wu, "A fuzzy reputation-based trust management scheme for cloud computing," *International Journal of Digital Content Technology and Its Applications*, vol. 6, no. 17, pp. 437–445, 2012.

[9] N. Yang, X. Chen, and H. Yu, "A reputation evaluation technique for web services," *International Journal of Security and Its Applications*, vol. 6, no. 2, pp. 329–334, 2012.

[10] A. Ghaffarinejad and M. K. Akbari, "An incentive compatible and distributed reputation mechanism based on context similarity for service oriented systems," *Future Generation Computer Systems*, vol. 29, no. 3, pp. 863–875, 2013.

[11] H. Zhao, X. Yang, and X. Li, "An incentive mechanism to reinforce truthful reports in reputation systems," *Journal of Network and Computer Applications*, vol. 35, no. 3, pp. 951–961, 2012.

[12] X. Su, M. Zhang, Y. Mu et al., "A robust trust model for service oriented systems," *Journal of Computer and System Sciences*, vol. 79, no. 5, pp. 596–608, 2013.

[13] T. Kaszuba, A. Hupa, and A. Wierzbicki, "Advanced feedback management for internet auction reputation systems," *IEEE Internet Computing*, vol. 14, no. 5, pp. 31–37, 2010.

[14] X. Li, F. Zhou, and X. Yang, "Scalable feedback aggregating (SFA) overlay for large-scale P2P trust management," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 10, pp. 1944–1957, 2012.

[15] Z. Yan, P. Zhang, and R. H. Deng, "TruBeRepec: a trust-behavior-based reputation and recommender system for mobile applications," *Personal and Ubiquitous Computing*, vol. 16, no. 5, pp. 485–506, 2012.

[16] M. Hawa, L. As-Sayid-Ahmad, and L. D. Khalaf, "On enhancing reputation management using Peer-to-Peer interaction history," *Peer-To-Peer Networking and Applications*, vol. 6, no. 1, pp. 101–113, 2013.

[17] K. Lu, H. Jiang, M. Li et al., "Resources collaborative scheduling model based on trust mechanism in cloud," in *Proceedings of the IEEE 11th International Conference on Security and Privacy in Computing and Communications (TrustCom '12)*, pp. 863–868, IEEE, 2012.

[18] Y. A. Kim and R. Phalak, "A trust prediction framework in rating-based experience sharing social networks without a Web of Trust," *Information Sciences*, vol. 191, pp. 128–145, 2012.

[19] C. Gui, Q. Jian, H. Wang, and Q. Wu, "Repeated game theory based penalty-incentive mechanism in internet-based virtual computing environment," *Journal of Software*, vol. 21, no. 12, pp. 3042–3055, 2010.

[20] O. Kussul, N. Kussul, and S. Skakun, "Assessing security threat scenarios for utility-based reputation model in grids," *Computers & Security*, vol. 34, pp. 1–15, 2013.

[21] P. Dong-Sheng, L. Chuang, and L. Wei-Dong, "A distributed trust mechanism directly evaluating reputation of nodes," *Journal of Software*, vol. 19, no. 4, pp. 946–955, 2008.

[22] J. Gorner, J. Zhang, and R. Cohen, "Improving trust modelling through the limit of advisor network size and use of referrals," *Electronic Commerce Research and Applications*, vol. 12, no. 2, pp. 112–123, 2013.

[23] S. Zhao, G. Wu, G. Chen, and H. Chen, "Reputation-aware service selection based on QOS similarity," *Journal of Networks*, vol. 6, no. 7, pp. 950–957, 2011.

[24] Y. Cui, M. Li, Y. Xiang, Y. Ren, and S. Cesare, "A QoS-based fine-grained reputation system in the grid environment," *Concurrency Computation Practice and Experience*, vol. 24, no. 17, pp. 1990–2006, 2012.

[25] M. Hussin, Y. C. Lee, and A. Y. Zomaya, "Reputation-based resource allocation in market-oriented distributed systems," in *Algorithms and Architectures for Parallel Processing*, pp. 443–452, Springer, Berlin, Germany, 2011.

[26] Z. Bankovic, D. Fraga, J. Manuel Moya et al., "Improving security in WMNs with reputation systems and self-organizing maps," *Journal of Network and Computer Applications*, vol. 34, no. 2, pp. 455–463, 2011.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 946249, 9 pages
http://dx.doi.org/10.1155/2014/946249

Research Article

An Empirical Evaluation of Lightweight Random Walk Based Routing Protocol in Duty Cycle Aware Wireless Sensor Networks

Adnan Noor Mian,[1] Mehwish Fatima,[1] Raees Khan,[1] and Ravi Prakash[2]

[1] *Department of Computer Science, National University of Computer and Emerging Sciences, Lahore 54700, Pakistan*
[2] *Department of Computer Science, University of Texas at Dallas, Dallas, TX 75080-3021, USA*

Correspondence should be addressed to Adnan Noor Mian; adnan.noor@nu.edu.pk

Received 31 August 2013; Accepted 5 December 2013; Published 13 February 2014

Academic Editors: W. Sun and G. Zhang

Energy efficiency is an important design paradigm in Wireless Sensor Networks (WSNs) and its consumption in dynamic environment is even more critical. Duty cycling of sensor nodes is used to address the energy consumption problem. However, along with advantages, duty cycle aware networks introduce some complexities like synchronization and latency. Due to their inherent characteristics, many traditional routing protocols show low performance in densely deployed WSNs with duty cycle awareness, when sensor nodes are supposed to have high mobility. In this paper we first present a three messages exchange Lightweight Random Walk Routing (LRWR) protocol and then evaluate its performance in WSNs for routing low data rate packets. Through NS-2 based simulations, we examine the LRWR protocol by comparing it with DYMO, a widely used WSN protocol, in both static and dynamic environments with varying duty cycles, assuming the standard IEEE 802.15.4 in lower layers. Results for the three metrics, that is, reliability, end-to-end delay, and energy consumption, show that LRWR protocol outperforms DYMO in scalability, mobility, and robustness, showing this protocol as a suitable choice in low duty cycle and dense WSNs.

1. Introduction

Wireless Sensor Networks (WSNs) are one of the emerging technologies that are making revolutionary developments in everyday scenarios. Their acceptance is rapidly increasing for a wide range of real life applications. Many applications like medical technologies, environment and habitat monitoring, law enforcement, and security are shifting their paradigm from manual system to automation by means of WSN [1, 2].

In WSN, small sensor motes are used to collect the data by means of sensing from their surrounding and the data is forwarded to a common repository or some dedicated device called sink. Sensor nodes are limited in resources and are mostly battery powered for which replacement is not realistic especially in remote applications. Generally WSNs have various design and implementation issues regarding their nature, but energy efficiency is the main design paradigm due to the limited source of energy available in sensor motes.

A sensor node can be a direct neighbor of sink or it can be a distant device out of the direct communication range of sink. In the later case, sensor nodes cooperate with each other in a multithop fashion to forward the data sensed by them and their neighbors to the sink nodes. WSNs are typically used in static environment; for example, in a factory for automated monitoring of processes in different systems, the low cost wireless nodes network would be preferred over wired sensing devices. Recently some new applications are emerging that are inherently dynamic in nature, for example, military surveillance, tracking of an object that changes its position [1], sensor nodes used in sea like dynamic water environments [2], and cattle farming. Besides generating data for its sensing activity, the repeated involvement of energy constrained sensor nodes, in forwarding packets coming from other nodes, causes its life time to finish soon. Moreover, the applications which need frequent mobility of sensor nodes can add to make this situation worse.

Deploying sensor in static and dynamic environments has its own advantages and disadvantages. Static sensors are easy to deploy and manage but can result in issues like uneven energy depletion of sensor nodes. For example,

sensor nodes near sink consume more energy as most of the traffic routed to sink passes from these nodes thus resulting in shorter battery life time of these nodes. Dynamic wireless sensor environments, on the other hand, are complex in designing and deployment but have some advantages over static scenarios. Mobility allows sensor nodes to cover more area as compared to other static nodes and mobility policies can be made for taking account of energy saving. So if we allow mobility in sensor nodes while the sink is static, sensor nodes may change their geographical location. The traffic burden is shared and distributed among nodes. Similarly if we allow sinks to be mobile, it allows energy saving in whole network by changing its position according to the energy consumed by its neighbors [3].

Duty cycling is a technique in which different sensor nodes use power saving inactive mode, that is, sleep mode at different time intervals [4]. In many applications of WSNs, packets are generated periodically at low rate and then routed to sink node or sink may send different queries to different sensor nodes. Duty cycling is very suitable in low data rate scenarios where nodes get involved in activity only at specific intervals of time.

There are different routing protocols for wireless networks at network layer. These can be classified generally in broadcast and unicast based protocols. For example, flooding, gossiping, and opportunistic routing are broadcast based protocols whereas AODV, DSR, AOMDV, and DYMO are unicast protocols [5, 6]. Generally broadcast protocols show poor performance due to excessive packets generation and collisions. For these reasons, unicast protocols are preferred. In such protocols duty cycling can be used to conserve energy.

Unicast based routing protocols can be proactive, reactive, hybrid, and probabilistic by nature. Proactive (table driven) protocols use routing tables for estimating the connection between source and destination node. Reactive (on-demand) protocols do not require any precalculation and estimation for route and hence these are much appealing in dynamic environments. Hybrid protocols are usually combination of proactive and reactive approaches and tend to adapt their behavior according to network changes. Probabilistic protocols also do not require any precalculation of routes and probabilistic formulation is made for selecting the next forwarding node, thus avoiding large overheads of forming routing tables. Due to these features, probabilistic protocols are useful especially for low data rate and in dense deployment of WSN.

Random walk based routing is a probabilistic protocol in which each node selects randomly from its neighbors nodes to forward the data packet. The path thus formed is a random walk (RW). RW has been widely studied; for instance, see [7, 8] for a quick overview. RW based routing protocol is often proposed for very small devices, in large and dynamic networks due to being extremely simple to implement, requiring small memory footprints, and not requiring topology information of the network and load balancing property of the RW [9–11]. On the other hand, reactive (on-demand) routing protocols are also considered to be useful in resource constrained and dynamic WSNs. However due to their inherent properties, increasing density of nodes badly affects performance of such protocols in terms of scalability. Furthermore, high mobility of sensor nodes and enabling low duty cycling make routing quite challenging. In such scenarios, random walk based routing has not been studied widely. In this paper, we have put before a Lightweight Random Walk based Routing (LRWR) protocol in which each step follows a three messages exchange not only to discover neighbors but also to randomly select and forward the packets to the selected neighbor. We call this protocol lightweight since the number of messages required to achieve one step of RW are bare minimum. We applied the LRWR protocol in WSN with IEEE 802.15.4 standard and duty cycle enabled environments. By comparing its performance for low data rate with DYMO, a widely used protocol for WSN, we find that LRWR protocol offers a better alternative for duty cycle enabled mobile WSNs.

Rest of the paper is structured as follows. Section 2 gives a brief overview of IEEE 802.15.4 and implementation of duty cycle in it. Section 3 briefly describes some of the important related works. Section 4 briefly describes the LRWR protocol. Section 5 details the simulation setup, section 6 discusses the results, and finally section 7 concludes the study.

2. A Brief Overview of IEEE 802.15.4 Protocol

IEEE has established a standard 802.15.4 for low cost Low Rate Wireless Personal Area Networks (LR-WPAN) since such networks have very different requirements from other wireless networks. There are many topologies for WSN but the two underlying topologies are star and peer-to-peer.

IEEE 802.15.4 provides physical and MAC layer implementation for WSN. On physical layer it provides a low range connectivity that may vary from 10 meter to 100 meter on a data rate of 250 kbits/sec [4]. On MAC layer it provides different protocols and frame structure for implementation of low cost, low rate communication. IEEE 802.15.4 MAC can be operated on two modes either beacon enabled (synchronized) or nonbeacon enabled (unsynchronized). Nonbeacon enabled mode uses unslotted CSMA/CA while beacon enabled mode uses slotted CSMA/CA and allows duty cycling. Duty cycling is implemented with the superframe structure that consists of beacon interval and superframe duration between two beacon intervals [4].

In beacon enabled mode, a PAN coordinator is responsible for synchronizing the nodes along with local coordinators as shown in Figure 1. These coordinators would be either synchronized or unsynchronized with respect to initialization time. If all coordinators are synchronized with respect to initialization time then nodes will have a symmetric awake and sleep behavior otherwise it will be asymmetric. In nonbeacon enabled mode, PAN coordinator is not responsible for synchronizing the PAN and due to unsynchronized environment there is no duty cycling.

Synchronization among nodes is achieved by a scanning process through all the channels for coordinator and orphan nodes. This scanning furnishes reliable synchronization; however, it increases the latency in the network. In mobile WSN, achieving synchronization is more difficult because

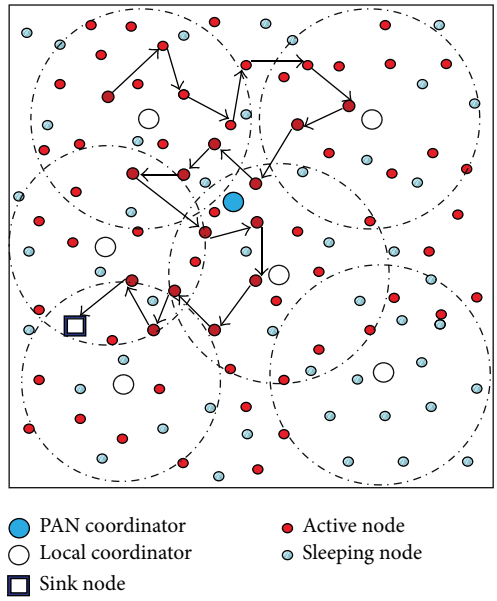


Figure 1: Random walk routing in a WSN with PAN Coordinators.

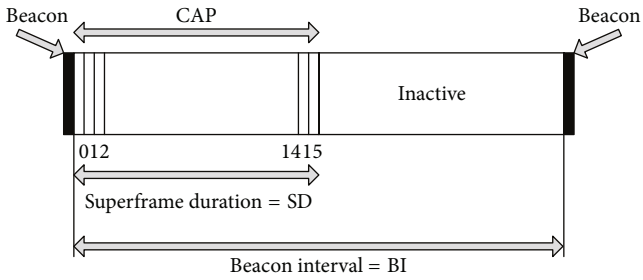


Figure 2: IEEE 802.15.4 superframe structure.

nodes are highly inclined to change their geographical positions due to which synchronization loss occurs. In such networks latency is comparatively greater than static WSN due to synchronization loss and reconnection scans.

In IEEE 802.15.4 MAC, duty cycle is implemented by superframe structure which is the time interval between two beacons sent by coordinator as described in Figure 2. "Beacon Order" (BO) and "Superframe Order" (SO) are two parameters that are responsible for duty cycling by changing the superframe structure [4]. BO and SO are used to formulate the values of beacon interval (BI) and superframe duration (SD), respectively. BI and SD determine the active (awake) period and inactive (sleep) period for nodes that are synchronized [4]. BO and SO values can be modified explicitly. The difference of BO and SO determines the duty cycle. Different values of BO and SO can result in the same duty cycle. The duty cycle percentage according to the difference between BO and SO is given by the formulae $DC = (1/2^{SO\text{-}BO}) * 100$ and its different values are shown in Table 1.

Duty cycling tends to increase the latency and drop ratio along with its energy saving characteristic. Although if a node has started its reception or transmission, it will go into inactive period only after completing its reception or transmission [4]. In spite of this, however, there may be packet drops due to collision of packets. These collisions may be due to ACK transmissions without backoff mechanism or due to the transmission of hidden nodes [12].

Table 1: Variation in duty cycle with BO and SO values [4].

BO-SO	0	1	2	3	4	5
DC	100	50	25	12	6.25	3.125
BO-SO	6	7	8	9	>10	—
DC	1.56	0.78	0.39	0.195	<0.1	—

3. Related Work

On the network layer, there are many routing protocols that are very well known in wireless ad hoc and sensor networks. The "Dynamic MANET On-demand" (DYMO) routing protocol is intended for mobile wireless networks. It adapts to changing network topology and determines unicast routes between nodes within the network [13]. The DYMO routing protocol is successor to AODV and can work as proactive or as a reactive routing protocol [13]. There are many simulation based studies on DYMO in MANET and WSN [14–17].

Random walk based routing is unicast based protocol which has been studied extensively from different perspectives. Some of the most important related works are as follows. In [11] the authors study a scenario in which the data source sends agents that walk randomly in the network and leave traces. Sink will send queries along some other random walks and hope to encounter the data traces. In [18] based on analysis and simulations Avin and Brito study query processing in dynamic environments using RW. The paper [19] motivates the need of RW in dynamic networks and then compare the efficiency of search using flooding and biased RW in such networks. Mabrouki et al. [20] study RW based routing protocol for WSN from theoretical perspective. These works study random walk routing at an abstract level using custom built simulators assuming simplified network model. These studies do not present detail level protocols that implement random walk routing.

Ahn et al. [21] proposed an implementation of random walk protocol considering 3-way handshake, like TCP. In this protocol RW packet is forwarded to one of the neighbors which has been selected through a list of neighbors. The list of all neighbor IDs are got from a distinct neighbor discovery phase. In such a case, the RW process will be suffered significantly in case the chosen node is in sleep mode. The distinct two phases may result in multiple message exchanges that may lead to large latencies. The proposed protocol on the other hand does not have distinct two phases. Instead just in one phase of the three, both the neighbor discovery and random forwarding are achieved. This results in efficient forwarding mechanism. Moreover, [21] assumes TDMA access mechanism for communicating with nodes, which has different characteristics from CSMA used in the widely used standard IEEE 802.15.4.

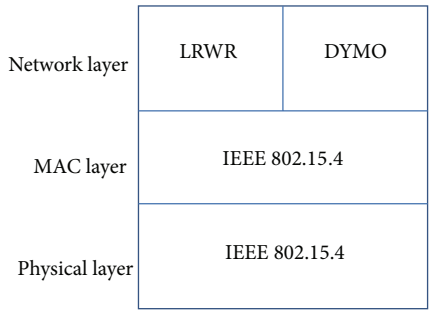


Figure 3: Protocol stack.

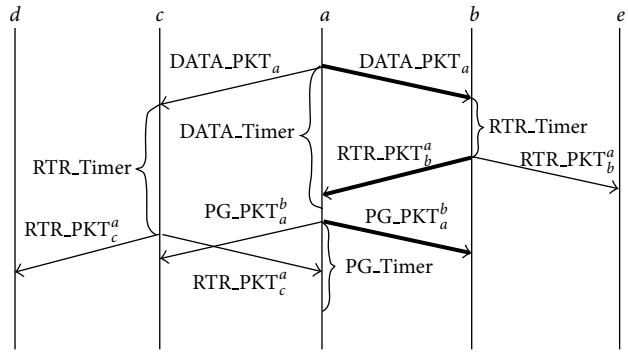


Figure 4: LRWR protocol sequence diagram.

The proposed protocol routes packets from source to the sink and guarantees that random walk does not branch off. The paper evaluates the performance of the protocol in duty cycle aware sensor networks and compares it with DYMO.

4. Lightweight Random Walk Based Routing Protocol

In random walk based routing the data packets are forwarded to one of the neighbors selected at random. This can be achieved in two phases. In the first phase of neighbor discovery a neighbor discovery packet is broadcasted and in return all neighbors reply with their IDs. After a specific time when all or most of the neighbors are known by the forwarding node, the node selects one of the neighbors randomly as the next hop node and then sends the packet to it. Such a simplistic approach may result in inefficiencies. Each of the phases would result in multiple message exchanges due to unicast with explicit acknowledgements and may result in large latencies due to carrier sensing mechanism. Moreover in case of dynamic environments, due to large latencies involved in neighbor discovery and then forwarding packet to one of the selected neighbors which may have moved out of the transmission range, the probability of an unexpected halting of RW increases. We present a network layer protocol, as shown in Figure 3, which not only discovers a random neighbor but also forwards the data to it in three messages exchange sequences. Due to low latencies involved in between messages, the probability of unexpected RW halting is very low.

4.1. Basic Assumptions. The system consists of mobile WSNs based on IEEE 802.15.4. The target node is present within the network. The data may be routed from a source sensor nodes to a sink or from sink to a sensor node. The network remains connected and there is a bidirectional connectivity between neighbor nodes at any time. The nodes can communicate only via a broadcast primitive. This primitive sends the packet using the local broadcast facility of the underlying MAC protocol, which uses the carrier sense and binary backoff mechanism to access the medium and is provided by the implementation of IEEE 802.15.4 standard.

4.2. Protocol Description. The presented protocol uses three messages exchange sequence DATA/RTR/PG at network layer. The sequence diagram of the protocol is shown in Figure 4. The protocol embeds lightweight efficient distributed selection logic at the network layer as described below.

Suppose node a is the forwarding node that generates data packet or forwards an existing data packet DATA_PKT$_a$ that has to be sent to some destination in the network using random walk routing. Such packets are "agents" carrying information about events or "queries" seeking such information [22]. To initiate the three messages sequence, node a broadcasts the packet DATA_PKT$_a$. All nodes in the transmission range of node a will receive the DATA_PKT$_a$. There is a possibility that some nodes may not receive this packet due to collisions or sleep mode. Note that it is not essential in this protocol that all of the neighbors should receive the broadcast packet. It is sufficient if some or just one of the neighbors receives it.

Now each neighbor that receives the data packet replies by broadcasting a special packet called Ready-To-Route (RTR_PKT), with a random delay RTR_Timer. This packet indicates the willingness of the neighbor node to become the new forwarding node. The random delay decreases the probability of collisions at the node a. If no RTR_PKT is received till the expiry of the DATA_Timer, the DATA_PKT$_a$ would be rebroadcasted. The node a will accept only the first RTR_PKT and after receiving the first one, say RTR_PKT^a_b packet from one of the neighbors node b, discards any other RTR_PKT from other neighbors. The packet RTR_PKT^a_b can only be processed by node a and all other nodes shall discard this packet. In the Figure 4 node e does not process RTR_PKT^a_b. Similarly RTR_PKT^a_c will be discarded by node d.

Node a, after accepting the request to route from node b, sends a special packet called Permission-Granted (PG_PKT^b_a). This packet is to grant permission to node b in response to its request to route. When other neighbor nodes receive this packet, they cancel their RTR_Timers to stop transmission of any RTR_PKT packets if not yet transmitted, thus reducing the number of unnecessary transmissions. Once node b gets the permission granted, it initiates its own three messages sequence by broadcasting the data packet,

which now also acts as an acknowledgement for node a that PG packet was received at node b. If the forwarding node a does not receive DATA_PKT from the new selected forwarding node within PG_Timer expiry, the PG_PKT_a^b is rebroadcasted. The newly selected forwarding node b only accepts the first PG_PKT_a^b and discards the remaining. Now if node b is a destination node then neighbors do not generate any further RTR_PKT packets and the algorithm terminates otherwise the same three messages exchange sequence is repeated.

Note that in the protocol, RTR_Timer is not the only reason that results in random selection of neighbor node. Other reasons which may lead to random selection are duty cycling and random delays due to MAC layer carrier sensing.

An important aspect of this protocol is that the random walk path does not branch off. The branching occurs if more than one RTR packet from different neighbor nodes are processed, as each processed RTR packet will trigger a PG packet to a different node thus starting a different RW. This is not possible because only the first RTR is processed and all other received RTR packets are discarded. The branching can also occur if a node receiving a PG packet triggers the broadcast of a data packet, which initiates a new RW. In such a case a node that receives a PG packet, which is not part of the current hop message exchange sequence or due to multiple transmissions of PG packets from the forwarding node, will trigger the broadcast of another data packet thus causing branching. But in the proposed protocol this will not be possible because the protocol makes sure that only the first PG packet received, which is part of the current hop 3-way message exchange sequence, is processed and all others are discarded.

In Figure 4 bold lines show the interaction between the forwarding node and the neighbor node selected as a new forwarding node.

The working mechanism of the proposed protocol can be summarized as below.

(i) When a node wants to forward data, it broadcasts the DATA_PKT.

(ii) All neighbor nodes receiving the DATA_PKT schedule their RTR_Timer timers randomly.

(iii) When RTR_Timer timers (of neighbor nodes) expire, RTR_PKT packets are unicasted to the data forwarding node.

(iv) The first RTR_PKT received on the forwarding node is accepted and all other RTR_PKT packets are discarded.

(v) If the forwarding node does not receive any RTR_PKT packet after waiting for a time set by DATA_Timer, it rebroadcasts the same packet. This is repeated for a specific number of times.

(vi) The forwarding node unicast a permission granted packet PG_PKT to the neighbor node whose RTR_PKT packet was accepted. If the selected neighbor node is the destination node then algorithm terminates.

(vii) All other neighbor nodes that receive this PG_PKT packet will halt their transmission of RTR_PKT if their RTR_timers not yet expired.

(viii) After sending the PG_PKT, if the forwarding node does not receive DATA_PKT from the selected neighbor node after time set by PG_Timer, the PG_PKT is sent again. This is repeated for a specific number of times.

(ix) Once a PG_PKT is received by the selected neighbor node, all other PG_PKT packets are discarded.

Table 2: Simulation parameters.

Parameter	Scenario	
	Static	Mobile
Simulation time	500 sec	500 sec
Area	200 m × 200 m	500 m × 500 m
Number of nodes	100, 200, 300, 400, 500	100
Trans. range	40 m	40 m
Duty cycle	100%, 50%, 25%, 12%	100%, 50%, 25%, 12%
PAN coordinator	1	1
Local coordinators	39, 78, 117, 156, 195	39
Mobility model	Random deployment	Random waypoint
Speed of nodes	0 m/s	10, 20, 30, 40, 50 m/s
Pause time	—	2 sec

5. Simulation Setup

We did all simulation experiments using NS-2 version 2.34. We used random waypoint mobility model for generating mobility pattern of nodes and performed simulations on multiple scenario files with the same initial parameters.

In case of static scenario we randomly deployed 100, 200, 300, 400, and 500 static nodes that form cluster head tree topology in 200 m × 200 m area. All sensor nodes had transmission range of 40 meters and same initial energy. The simulations were run for 500 seconds. On application layer, a data sending application was used that sent data to a sink in the network.

Similarly in case of dynamic nodes scenario, we assumed a cluster head tree topology of 100 nodes that are mobile in 500 m × 500 m area. All nodes had transmission range of 40 meters and same initial energy. The simulations were run for 500 seconds. On application layer, similar to the static case, a data packets generating application was used that sent data to a sink node in the network. Each value of the plot is calculated as mean of 200 runs.

Both of the simulation environments were unsynchronized with respect to time. This means that even when all nodes have the same duty cycle set, some nodes may be in sleep mode. This is due to the association of nodes with different coordinators. The synchronization of nodes depends upon the device initialization time and association time with coordinator. The parameters used in the simulations are summarized in Table 2.

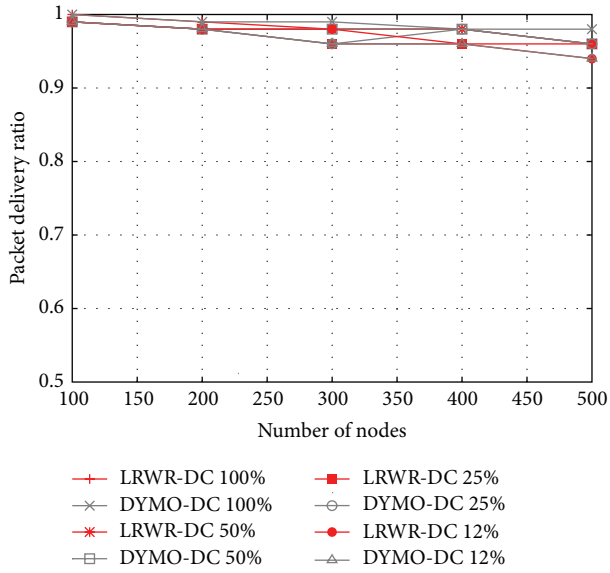


Figure 5: Packet delivery ratio versus total number of nodes.

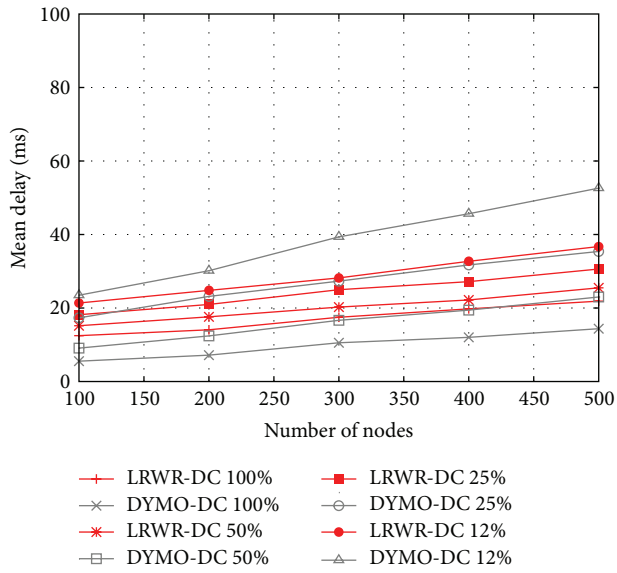


Figure 6: Mean delay versus total number of nodes.

6. Results and Discussion

Our simulation stack architecture is shown in Figure 3. We compared LRWR and DYMO protocols with respect to three metrics, that is, packet delivery ratio, end-to-end delay for data delivery from source to sink node within a specific time, and energy consumption for the routes. These results are evaluated using different duty cycles and in both static and dynamic scenarios. In all of these experiments, node 0 was taken as a source node which generates data and sends it to a sink node. The results are discussed below.

6.1. Scalability Results. In these simulations we evaluated scalability of the protocol, that is, the performance of the protocol when the number of nodes are increased in terms of packet delivery ratio, end-to-end delay, and energy consumption in the presence of different duty cycle.

Figure 5 shows the packet delivery ratio from source to sink as a function of the number of nodes. The packet delivery ratio of LRWR and DYMO is similar even with low duty cycle. Since nodes are static in the network, therefore duty cycling has very less effect on delivery ratio. However, with the increase in network density, there is a little decline in delivery ratio in both of the protocols.

Figure 6 shows the mean end-to-end delay from source to sink as a function of number of nodes in both LRWR and DYMO protocols for different duty cycles in static nodes. We see that the mean delay increases with increasing number of nodes. This is quite initiative since with the increase of number of nodes the number of hops in the path from source to sink increases. We also see that the mean delay increases with decreasing duty cycle. This is because in low duty cycle more nodes are inactive and forwarding of packets is delayed. Interestingly the mean delay in very low duty cycles increases at a faster rate in DYMO as compared with LRWR with increasing number of nodes. This is because in DYMO whenever a node is not approachable its entry is deleted from the forwarding node routing table and then a route discovery process is initiated again causing high delay. On the other hand LRWR immediately selects randomly any active node and forwards the packet to it resulting in comparatively low delays. Thus LRWR protocol performance with respect to mean end-to-end delay is better as compared with DYMO in low duty cycles and large number of nodes.

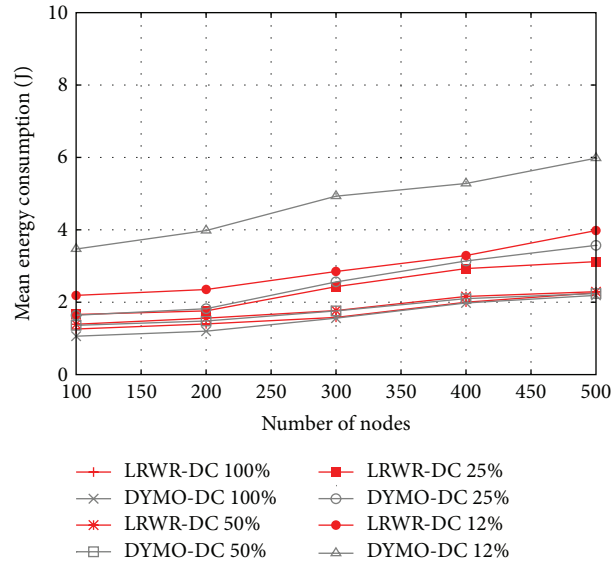


Figure 7: Mean energy consumed versus total number of nodes.

In Figure 7 mean energy consumption of nodes is shown as a function of the number of nodes. These results of LRWR and DYMO are evaluated under different duty cycles. Generally the energy consumption is high because more effort is required to achieve synchronization between coordinators and sensor nodes [23]. From the plots, we see in both

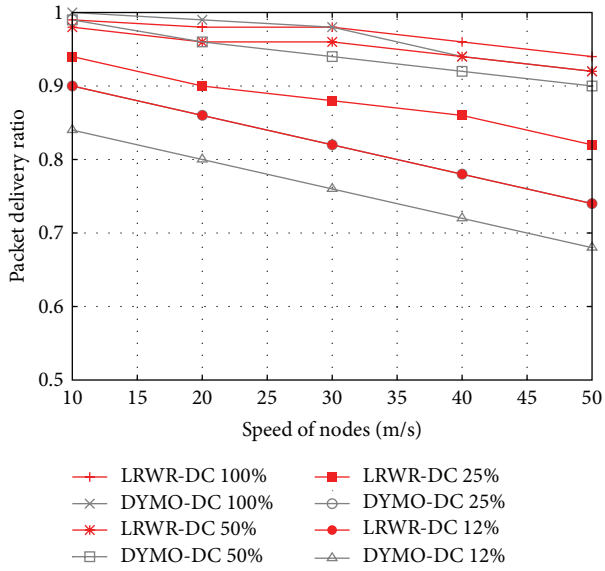


Figure 8: Packet delivery ratio versus speed of nodes.

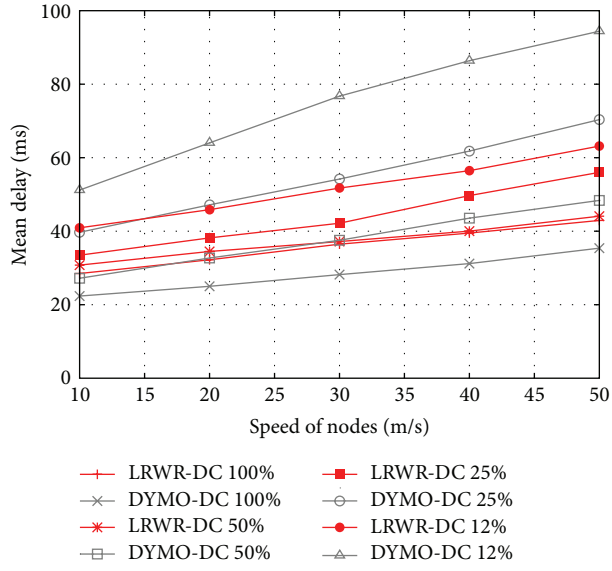


Figure 9: Mean delay versus speed of nodes.

protocols that with the increase in the number of nodes, the energy consumption also increases. This is because of the increase in the number of hops and thus transmissions. However it is observed that the energy consumption of LRWR is less than that of DYMO in low duty cycles. In DYMO routes from source to sink need to be discovered and in case of low duty cycles, since more nodes may be in sleep mode, more effort is required to find routes to sink.

6.2. Mobility Results. In this section we discuss the results of mobility scenario and compare both protocols in the presence of mobility. In the experiments we executed a data generation application that sent data from the source to sink node. We deployed 100 nodes. Node 0 was taken as source and node 99 was taken as sink node in all the experiments. We compared the results of both LRWR and DYMO protocols using three metrics, namely, packet delivery ratio, end-to-end delay for data delivery within a specific time, and energy consumption for the route.

Figure 8 shows the packet delivery ratio as a function of the speed of nodes. These results of LRWR and DYMO are evaluated using different duty cycles. The packet delivery ratios of LRWR and DYMO are similar at high duty cycle. Since nodes are mobile in the network, therefore, with the increase in speed of nodes, data delivery ratio decreases. The reason of this decline is that mobile nodes tend to lose connectivity and synchronization with the coordinators more often due to their change of geographical position and having to reassociate themselves with some coordinators. Due to this synchronization loss and reassociation, packet loss is more often. DYMO is deterministic protocol that maintains the route on-demand. These routes are fixed and packet delivery failure would be due to packet loss and unassociation of some nodes (called ophran nodes). While LRWR is a probabilistic protocol that finds the next forwarding node among the active neighbors and if some nodes are off or unsynchronized, these will not be considered in next forwarding node selection. Due to the similar reasons the low duty cycles also drastically affect DYMO as compared with LRWR protocol. In short we find that the proposed LRWR protocol performance is better in low duty cycles and high mobility.

Figure 9 shows the mean delay for data delivery as a function of the speed of nodes. These results of LRWR and DYMO are evaluated under different duty cycles. At 100% duty cycle with mobile nodes DYMO performs better than LRWR. At 50% duty cycle with mobile nodes LRWR protocol mean delay is similar to DYMO. However in low duty cycles 25% and 12% and with high speed of nodes LRWR is performing better than DYMO. The reason of this performance decline of DYMO is that mobile nodes tend to loose connectivity with coordinators more often, due to their change of geographical position. This causes loss of synchronization in the network. With the increase of speed, synchronization loss is even more often. Due to the deterministic nature and routing table based, DYMO has to forward packets according to its routing information. It might be possible that in a specific route some nodes are in sleep mode. In such a case a node has to wait for the next node to go into active mode. This increases latency. LRWR, on the other hand, takes the advantage of probabilistic nature and randomly selects one of the available active neighbors.

In Figure 10 mean energy consumption of nodes is shown as a function of the speed of nodes in the presence of different duty cycles. We see that in low duty cycles and high mobility LRWR protocol performs better than DYMO. The DYMO consumes more energy since it has to generate more control packets to maintain routes and synchronization in higher mobility and lower duty cycle. LRWR protocol, on the other hand, simply forwards the data packets to any of the available active nodes without having an overhead of maintaining routes and synchronization.

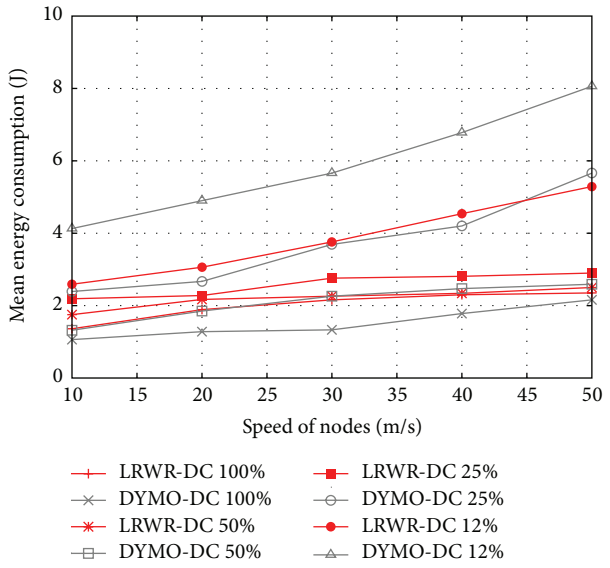


Figure 10: Mean energy consumed versus speed of nodes.

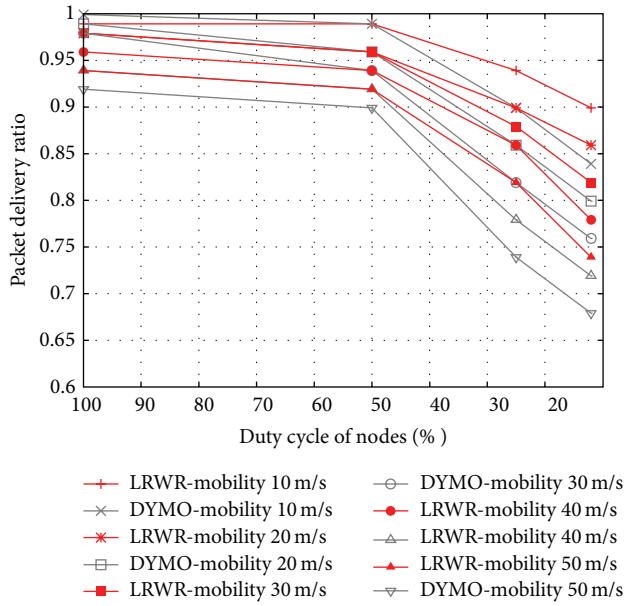


Figure 11: Packet delivery ratio versus duty cycle of nodes.

6.3. Stability Results. In this experiment we have compared the performance of both of the protocols in terms of packet delivery ratio under decreasing duty cycles with different speeds. In Figure 11 LRWR protocol shows more robust and stable behavior in duty cycle aware WSN. By reducing duty cycle of nodes, DYMO gets steep decline in packets delivery ratio whereas LRWR shows less decrease in its delivery ratio. When duty cycle is decreased, though the number of active nodes gets lesser, LRWR protocol still maintains its high data delivery ratio. This higher delivery ratio even with the support of less number of nodes shows the robustness of LRWR protocol.

7. Conclusion and Future Work

In this paper, we have presented a network layer unicast three-way messages exchange protocol LRWR and analyzed its performance based on packet delivery ratio, end-to-end delay, and energy consumption in the presence of duty cycles.

Based on extensive simulations of the proposed protocol with IEEE 802.15.4 lower layers, we found that the proposed LRWR protocol outperforms DYMO protocol especially in low duty cycles and in large number of nodes. We did simulation experiments for both static and dynamic environments. From the static simulation results, we find the scalability of both protocols. We see that the packet delivery ratio scale well in both protocols. However the mean energy and delay are higher in case of DYMO as the number of nodes increases and the duty cycle is reduced. From the dynamic environment experiments we observe that the packet delivery ratio decreases in both cases. LRWR performs better in high mobility and in low duty cycles. Mean delay and energy consumed increase in both cases but the increase is greater in case of DYMO with the increase of speed of nodes. Thus LRWR performs better than DYMO even under high mobility.

In this paper we studied network in synchronized environments. Synchronization has an overhead of the synchronization protocol which affects adversely. It would be interesting to see the effects when the network is unsynchronized and duty cycling is obtained at node level rather than at coordinator level. We plan to extend our experiments for such unsynchronized mobile WSN in the future.

Conflict of Interests

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

References

[1] Y. Miao, "Applications of sensor networks," Seminar on ireless Self-Organization Networks, http://www7.informatik.uni-erlangen.de/~dressler/lectures/seminar-sensornetze-ss05/paper-ying-miao.pdf.

[2] J. Luo, D. Wang, and Q. Zhang, "Double mobility: coverage of the sea surface with mobile sensor networks," in *Proceedings of the IEEE INFOCOM*, pp. 118–126, 2009.

[3] S. A. Munir, B. Ren, W. Jiao, B. Wang, D. Xie, and J. Ma, "Mobile wireless sensor network: architecture and enabling technologies for ubiquitous computing," in *Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW '07)*, pp. 113–120, IEEE Computer Society, Washington, DC, USA, 2007.

[4] N. P. Mahalik, *Sensor Networks and Configuration: Fundamentals, Standards, Platforms, and Applications*, Springer, New York, NY, USA, 2006.

[5] S. K. Singh, M. Singh, and D. K. Singh, "Routing protocols in wireless sensor networks—a survey," *International Journal of Computer Science & Engineering Survey*, vol. 1, no. 2, pp. 63–83, 2010.

[6] K. Akkaya and M. Younis, "A survey on routing protocols for wireless sensor networks," *Ad Hoc Networks*, vol. 3, no. 3, pp. 325–349, 2005.

[7] L. Lovász, "Random walks on graphs: a survey," in *Combinatorics Paul Erdös in Eighty*, vol. 2, pp. 1–46, J'anos Bolyai Mathematical Society, Budapest, Hungry, 1993.

[8] P. Erdos and S. J. Taylor, "Some problems concerning the structure of random walk paths," *Journal Acta Mathematica Hungarica*, vol. 11, no. 1-2, pp. 137–162, 1960.

[9] H. Tian, H. Shen, and T. Matsuzawa, "Randomwalk routing for wireless sensor networks," in *Proceedings of the 6th International Conference on Parallel and Distributed Computing Applications and Technologies (PDCAT '05)*, pp. 196–200, IEEE Computer Society, 2005.

[10] R. Beraldi, L. Querzoni, and R. Baldoni, "A hint-based probabilistic protocol for unicast communications in manets," *Ad Hoc Networks*, vol. 5, no. 4, pp. 547–566, 2006.

[11] D. Braginsky and D. Estrin, "Rumor routing algorthim for sensor networks," in *Proceedings of the 1st ACM international Workshop on Wireless Sensor Networks and Applications (WSNA '02)*, pp. 22–31, ACM, New York, NY, USA, 2002.

[12] D. Rohm, M. Goyal, H. Hosseini, A. Divjak, and Y. Bashir, "Configuring beaconless IEEE 802.15. 4 networks under different traffic loads," in *Proceedings of the International Conference on Advanced Information Networking and Applications (AINA '09)*, pp. 921–928, 2009.

[13] T. Watteyne, A. Molinaro, M. Richichi, and M. Dohler, "From MANET to IETF ROLL standardization: a paradigm shift in WSN routing protocols," *IEEE Communications Surveys Tutorials*, vol. 13, no. 4, pp. 688–707, 2011.

[14] V. Kumar, A. Raghuvansi, and S. Tiwari, "Performance study of beaconenabled IEEE 802.15.4 standard in WSNs with clustering," in *Proceedings of the International Conference on Power, Control and Embedded Systems (ICPCES '10)*, pp. 1–5, 2010.

[15] M. Hossen, A. F. M. Kabir, R. H. Khan, and A. Azfar, "Interconnection between 802.15.4 devices and IPv6: implications and existing approaches," *International Journal of Computer Science Issues*, vol. 7, no. 1, 2010.

[16] V. Kumar and S. Tiwari, "Performance of routing protocols for beaconenabled IEEE 802.15.4 WSNs with different duty cycle," in *Proceedings of the International Conference on Devices and Communications (ICDeCom '11)*, pp. 1–5, 2011.

[17] A. Raghuvanshi and S. Tiwari, "DYMO as routing protocol for IEEE- 802.15.4 enabled wireless sensor networks," in *Proceedings of the 6th International Conference on Wireless Communication and Sensor Networks (WCSN '10)*, pp. 1–6, 2010.

[18] C. Avin and C. Brito, "Efficient and robust query processing in dynamic environments using random walk techniques," in *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks (IPSN '04)*, pp. 277–286, ACM, New York, NY, USA, 2004.

[19] G. Hasslinger and T. Kunz, "Wireless systems and mobility in next generation internet," in *Efficiency of Search Methods in Dynamic Wireless Networks*, L. Cerdà-Alabern, Ed., pp. 142–156, Springer, Berlin, Germany, 2008.

[20] I. Mabrouki, X. Lagrange, and G. Froc, "Random walk based routing protocol for wireless sensor networks," in *Proceedings of the 2nd International Conference on Performance Evaluation Methodologies and Tools (ValueTools '07)*, pp. 1–10, 2007.

[21] J. Ahn, S. Kapadia, and S. Pattem, "Empirical evaluation of querying mechanisms for unstructured wireless sensor networks," *SIGCOMM Computer Communication Review*, vol. 38, no. 3, pp. 17–26, 2008.

[22] D. C. Dhanapala, A. P. Jayasumana, and Q. Han, "On random routing in wireless sensor grids: a mathematical model for rendezvous probability and performance optimization," *Journal of Parallel and Distributed Computing*, vol. 71, no. 3, pp. 369–380, 2011.

[23] A. Prince-Pike and J. Collins, "Power characterisation of zigbee wireless networks," in *Proceedings of the 16th Electronics New Zealand Conference (ENZCon '09)*, 2009.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 793171, 12 pages
http://dx.doi.org/10.1155/2014/793171

Research Article

Data Delivery Method Based on Neighbor Nodes' Information in a Mobile Ad Hoc Network

Shigeru Kashihara,[1] Takuma Hayashi,[1] Yuzo Taenaka,[2] Takeshi Okuda,[1] and Suguru Yamaguchi[1]

[1] *Graduate School of Information Science, Nara Institute of Science and Technology, Takayama, Ikoma, Nara 8916-5, Japan*
[2] *Information Technology Center, The University of Tokyo, Yayoi, Bunkyo, Tokyo 2-11-16, Japan*

Correspondence should be addressed to Shigeru Kashihara; shigeru@is.aist-nara.ac.jp

Received 11 October 2013; Accepted 19 December 2013; Published 4 February 2014

Academic Editors: W. Sun and G. Zhang

This paper proposes a data delivery method based on neighbor nodes' information to achieve reliable communication in a mobile ad hoc network (MANET). In a MANET, it is difficult to deliver data reliably due to instabilities in network topology and wireless network condition which result from node movement. To overcome such unstable communication, opportunistic routing and network coding schemes have lately attracted considerable attention. Although an existing method that employs such schemes, MAC-independent opportunistic routing and encoding (MORE), Chachulski et al. (2007), improves the efficiency of data delivery in an unstable wireless mesh network, it does not address node movement. To efficiently deliver data in a MANET, the method proposed in this paper thus first employs the same opportunistic routing and network coding used in MORE and also uses the location information and transmission probabilities of neighbor nodes to adapt to changeable network topology and wireless network condition. The simulation experiments showed that the proposed method can achieve efficient data delivery with low network load when the movement speed is relatively slow.

1. Introduction

Despite their widespread use, smart mobile devices (SMDs) still carry out communications only through existing wireless infrastructures, such as base stations. However, since SMD density is high in large cities, there are many more opportunities for SMDs to directly communicate with each other. Such communication has the potential to make a wireless network more resilient. In the near future, SMDs will directly and autonomously connect to each other to form a mobile ad hoc network (MANET). MANETs will also coexist with or replace current wireless infrastructures.

Compared with the existing wireless infrastructure, since a MANET requires no cost for network design, construction, and management, it furthermore has the potential to improve fault tolerance by having a distributed control. However, it is difficult for a MANET to make reliable communication among moving nodes due to the instability of the network topology and the wireless network condition. In particular, the movement of nodes characterizing such a wireless network exacerbates this difficulty. To provide reliable communication, a MANET thus requires a reliable data delivery method that can adapt to a changeable network topology and an unstable wireless network.

To improve data delivery over an unstable wireless network, Biswas and Morris [2] proposed opportunistic routing, which uses packet distribution via wireless broadcast. In this approach, multiple neighbor nodes (NNs) simultaneously receive a packet transmitted by a sender, and they then help to forward the packet to increase throughput. On the other hand, to enhance the efficiency of data delivery, network coding schemes are also receiving increasing attention [3]. In a network coding scheme, in addition to a conventional packet-forwarding process, some packets are coded by a sender or routers and transmitted as a single-coded packet. A receiver can then decode the coded packets when it receives enough of them. Consequently, since each coded packet randomly has a random linear combination of multiple packets,

the communication becomes more resilient against packet loss and the reliability of data delivery is improved. Based on the previous two approaches, and to further improve the efficiency of data delivery in a fixed wireless mesh network, Chachulski et al. proposed MAC-independent opportunistic routing and encoding (MORE) method [1]. MORE combines opportunistic routing and network coding to enhance data delivery and to reduce the number of redundant packets that arise during opportunistic routing. Although MORE performs well in an unstable wireless mesh network, it is difficult to adapt to the movement of nodes because each node needs to share all link conditions throughout the entire network. Delivery of data among moving nodes therefore remains a challenging issue, because a MANET can have no static route.

To improve reliable communication in a MANET, we now propose a data delivery method that employs opportunistic routing, network coding, and information about NNs. Opportunistic routing and network coding technologies are the same as those in MORE, since we use MORE as a base system. However, in a MANET, since each node moves freely, it is impossible to accurately grasp the entire network topology and wireless network condition, as is the case with MORE. Therefore, in the proposed method, the following new functions are added to MORE to adapt to the changeable network topology and wireless network condition. First, each node periodically broadcasts an information packet that includes its own and the NNs' location information, and it then obtains the packet error rate (PER) to the one-hop NN from the bit error rate (BER) of the received information packet. Next, the source node (SN) and relay nodes (RNs) efficiently distribute coded packets to the next RNs according to the location information and transmission probability of their NNs. Whereas in MORE an SN computes a forwarders list from the SN to the destination node (DN), in our proposed method each node makes a forwarders list for only the next RNs hop-by-hop. The primary contribution of the present study is thus efficient data distribution using network coding and opportunistic routing based on NNs' information, that is, location information and PER. Through the simulation experiments, we demonstrate the basic characteristics of the proposed method in a MANET.

The remainder of the present paper is organized as follows. Section 2 presents related work. Section 3 proposes a data delivery method using opportunistic routing, network coding, and the information of NNs for a MANET. Section 4 evaluates the communication performance of the proposed method. Finally, concluding remarks are presented in Section 5.

2. Related Work

As mentioned above, opportunistic routing and network coding have been proposed to improve the efficiency of data delivery in an unstable wireless mesh network. Since the method proposed herein is based on these concepts, we outline these approaches in Sections 2.1 and 2.2; Section 2.3 then describes issues related to data delivery in a MANET.

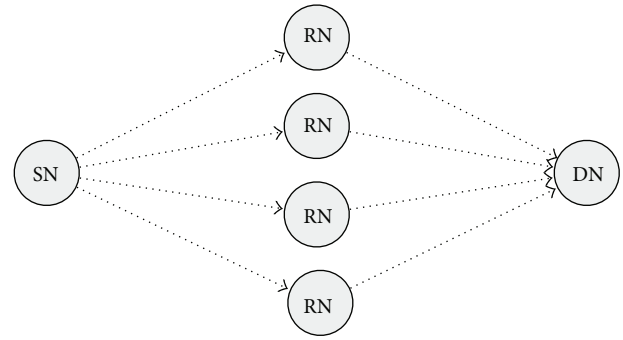


Figure 1: Advantage of opportunistic routing.

2.1. Opportunistic Routing. Various routing protocols have been proposed for a MANET [4–7]. In these protocols, data delivery to a DN is based on a routing table. This approach is satisfactory when the channel quality is very good and nodes are always on but may otherwise be very wasteful. Indeed, in a dynamic network such as a MANET, maintaining a reliable routing table is a particularly difficult task.

As a routing method suited to such an unstable wireless network, opportunistic routing has lately attracted considerable attention [1, 2, 8, 9]. Opportunistic routing has been introduced by means of the ExOR protocol [2]. Since opportunistic routing exploits wireless broadcasting, multiple NNs receive a packet transmitted by a sender, at the same time, and candidate forwarders specified in the packet rebroadcast the packet. As a result, ExOR can increase the transmission probability in a wireless mesh network.

Figure 1 illustrates the basic advantage of opportunistic routing. If the transmission probability of each link is p ($0 < p \leq 1$), then the transmission probability from the SN to the DN is p^2 in unicast transmission. On the other hand, if all four intermediate nodes, that is, RNs, can relay a packet by opportunistic routing, then the probability of transmission to the DN is improved to $1 - (1 - p^2)^4$.

2.2. Network Coding. Ahlswede et al. proposed a method to improve the transmission the transmission probability by coding some packets over a network, that is, network coding [3]. Figure 2 outlines the basic advantage of network coding over a wireless network. We assume that nodes 1 (N1) and 3 (N3) communicate with each other. If network coding is not employed, N1 first sends a packet (a) to node 2 (N2) and N2 then forwards the packet to N3. N3 also sends a packet (b) to N2, which is then forwarded to N1. Thus, the transmission count is four. If network coding is employed, after receiving packets (a) and (b), N2 broadcasts a coded packet ($a + b$) to N1 and N3. N1 and N3 can then extract the necessary packet from the coded packet by themselves. Therefore, the transmission count is reduced to three.

Based on this concept, a number of network coding schemes over wireless networks have been proposed [1, 10, 11]. In COPE [10], as described above, when two sender nodes transmit packets, the intermediate node XORs their packets and transmits the single-coded packet to them. If the two

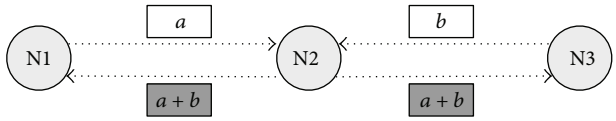


Figure 2: Advantage of network coding over a wireless network.

receiver nodes have sufficient information to decode the coded packet, the packet is then decoded. CodeCast [11] is a network-coding-based multicast protocol for an ad hoc network and employs a random network coding to implement both localized loss recovery and path diversity transparently. Simulation results demonstrated that CodeCast achieves an almost perfect packet delivery ratio while maintaining a lower overhead than conventional multicast. In the present study, to improve the data delivery in unicast communication, we focus on MORE [1], which improves the transmission probability by opportunistic routing and network coding. MORE codes and decodes together only packets belonging to the same flow, that is, an intraflow technique, to obtain higher throughput. In addition, MORE implements both opportunistic routing and intraflow network coding in a real system.

We now briefly explain the operation of MORE, which is a base system in the proposed method. As illustrated in Figure 3, to provide reliable file transfer, the SN generates batches from the file which are divided into m packets each. The SN then broadcasts a coded packet created by a random linear combination of m native packets (uncoded packets are referred to as native packets) in the current batch. If an RN receives an innovative packet (coded packets with new information are referred to as innovative packets), it also creates and broadcasts a random linear combination of the coded packets that were received in the RN's buffer. If the received coded packet is not an innovative packet, the RN discards it because the coded packet has no new information. On the other hand, when the DN receives a coded packet, it also checks whether the received coded packet is an innovative packet. If it is not, the DN discards the coded packet because it contains no new information to decode the batch. Conversely, if the DN receives m innovative packets, it can decode the entire batch; whereupon, to obtain the next batch, the DN sends back an ACK packet to the SN. Therefore, since it is more likely that coded packets broadcasted by the SN contain information that is different from the coded packets broadcasted by an RN, MORE can also alleviate the problem of redundant packets that arise due to opportunistic routing. Moreover, MORE can achieve reliable communication because it controls data transmission at every batch, such as a retransmission mechanism.

2.3. Data Delivery Issues in a MANET. Since the network topology and wireless network condition in a MANET are changeable due to the movement of nodes, the improvement of transmission efficiency is an important issue. Many routing protocols for MANET have been studied to improve data delivery [7]. However, they do not have reliable transmission controls such as a retransmission mechanism, so they cannot

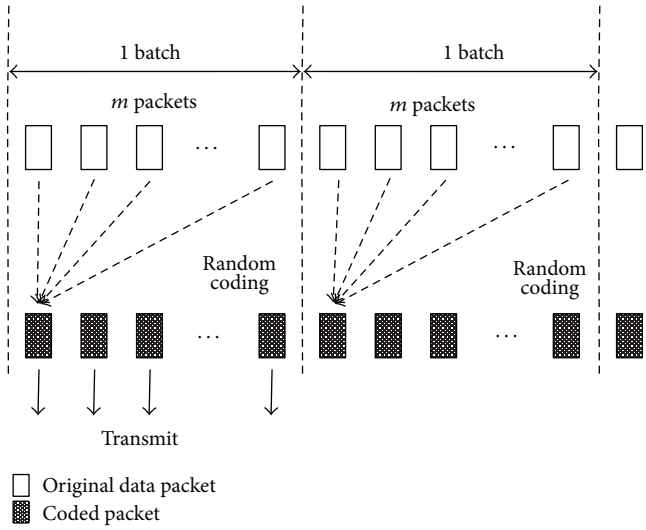


Figure 3: Random network coding in MORE.

achieve reliable communication. For instance, since traditional MANET routing protocols such as DSR [5] and AODV [4] typically need to discover a route before exchanging packets between communication nodes, the predetermined routes are susceptible to changeable link quality or link failure when nodes move freely. To alleviate such impacts on routes, position-based routings have been proposed [12]. Moreover, approaches that combine position-based routing with opportunistic routing [13, 14] can also enhance the delivery ratio of position-based routing. These approaches select nodes that are closest to the destination from among all the current node's neighbors as the next forwarding nodes, without considering transmission probability to the next forwarding nodes. However, if transmission probabilities to all the forwarding nodes are severely low, the sent data may be lost and the delivery ratio may be reduced, because opportunistic routing has no retransmission mechanism. To make reliable communication, we therefore need to have a retransmission mechanism.

As described in Section 2.2, since MORE controls data transmission at every batch, it can achieve a reliable communication over unstable wireless mesh networks. To optimize transmission efficiency, MORE employs expected transmission count (ETX) [15]; the ETX of a link is calculated from the forward and reverse delivery ratio of the link. Since each node measures the transmission probability to communicable NNs by periodically sending ICMP packets, each ETX can be calculated from the measurement results. All of the ETXs are then shared throughout the entire network. Before sending a packet, the SN uses the ETX values to make a forwarders list, which specifies the RNs that will forward the packet from the SN to the DN, to limit unnecessary forwarding of packets by RNs. Since MORE focuses on a fixed wireless mesh network, it is easy to share ETXs over the entire network because the transmission probability does not change frequently or abruptly. In addition, routing paths based on ETXs do not change drastically. However, in a MANET, the network

topology and ETX values change substantially due to the constant movement of nodes. Consequently, MORE cannot avoid communication degradation in a MANET because all of the ETXs in the entire network cannot be quickly updated due to frequent changes in the network topology and wireless network condition. Therefore, to achieve reliable communication over a MANET, we need to quickly update the wireless network condition of communicable NNs and select appropriate RNs.

3. Data Delivery Method

To enhance the efficiency of data delivery in a MANET, we first employ the same opportunistic routing and network coding used in MORE. However, as described in Section 2.3, MORE has difficulty in appropriately delivering packets based on ETXs due to the changeable network topology and wireless network condition. Therefore, in the present study, to adapt to the changeable network topology and wireless network condition, we propose a data delivery method based only on NNs' information, that is, location information and PER of NNs. Since the proposed method does not need to share ETXs over the entire network, it can quickly update the wireless network condition of NNs and appropriately select RNs hop-by-hop. In this section, we first outline the proposed method in Section 3.1 and then explain its operations in detail in Sections 3.2, 3.3, and 3.4.

3.1. Operation Outline. In a MANET, it is impossible to obtain a completely stable route from an SN to a DN based on ETXs due to the changeable network topology and wireless network condition. In addition, a long time is required to measure and share ETXs. We now consider a data delivery method based only on NNs' information, rather than sharing ETXs over the entire network.

In our proposed method, first of all, to obtain NNs' information, each node periodically broadcasts an information packet and collects its own NNs' information by receiving information packets that NNs broadcast. Figure 4 shows the information packet format. PACKET TYPE field indicates that the packet is an information packet, while SENDER IP and SENDER POSITION fields indicate the IP address and geographical position of the sender node, respectively. The packet also includes NN's information that the sender already knows, consisting of an IP address, a geographical position, and PER between the sender and the NN. We assume that each node can obtain location information using a positioning system, such as GPS. The receiver can also obtain the PER to the sender from the BER of the received information packet. Thus, as illustrated in Figure 5(a), each node can obtain the other nodes' location information and PER within two hops.

Next, to efficiently deliver data, the SN first generates a forwarders list, which only specifies the next RNs that forward a coded packet by broadcast. Note that the next RNs are selected based on node location and transmission probability for the one-hop NNs. The SN then broadcasts the coded packet including the forwarders list (see Figure 5(b)).

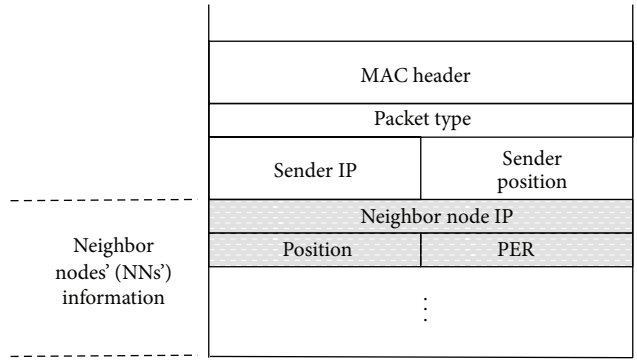


Figure 4: Format of an information packet.

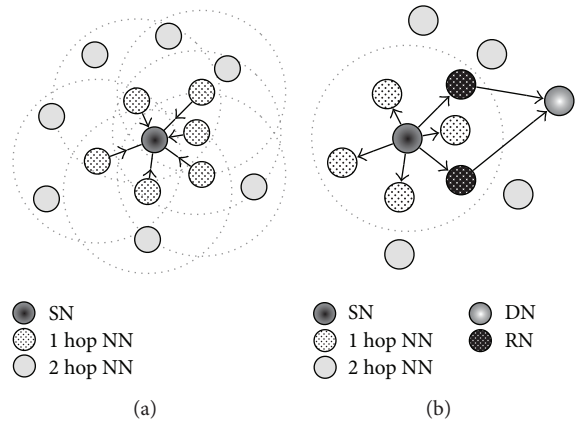


Figure 5: Outline of the data delivery process.

Figure 6 illustrates the coded packet format, which is based on a MORE coded packet; the grey fields indicate additional fields in our proposed method. PACKET TYPE field indicates that the packet is a coded packet. Since data delivery uses location information, the coded packet includes source and destination's geographical positions, that is, SN POSITION and DN POSITION fields. In MORE a coded packet holds all forwarder nodes' information from an SN to a DN in a forwarders list, while our proposed method inserts only the next RNs' information into the forwarders list because it is difficult to maintain a stable route from an SN to a DN. Thus, the forwarders list is updated in a hop-by-hop fashion, and its size depends on the number of NNs that forward the packet. In nodes that receive the coded packet, the nodes specified in the forwarders list become RNs. The RNs also use the same operation as the SN to select the next RNs and forward the coded packet after replacing the previous forwarders list with a new one. Thus, the operation is repeated until the coded packet reaches the DN. In the following sections, we explain in detail the operations used to achieve efficient data delivery in a MANET.

3.2. Neighbor Node Information. As noted above, it is difficult to maintain a stable transmission path in a MANET. The proposed method therefore forwards a coded packet by broadcasting it in a hop-by-hop fashion. To achieve this,

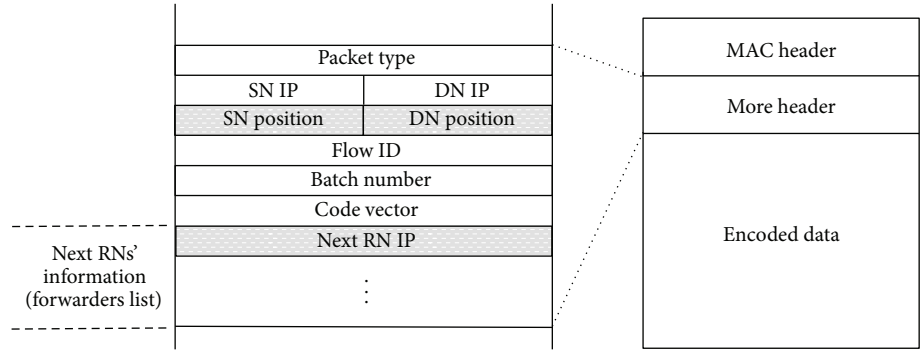


Figure 6: Format of a coded packet.

the information of NNs is indispensable. In the proposed method, to obtain the NNs' information, each node, say node i, periodically broadcasts an information packet that includes the location of node i, and the locations and the PER of the NNs that node i already knows, as described in the previous section. Each node can also obtain a PER calculated from the BER of the information packet as the transmission probability to a one-hop NN. The PER is calculated by the following equation:

$$\mathrm{PER} = 1 - (1 - \mathrm{BER})^{L}, \tag{1}$$

where L indicates the packet size. Thus, each node can quickly and reliably obtain the PER to one-hop NNs, and then each node makes a routing table containing the location information and the PER.

3.3. Selection of RNs. Since the proposed method forwards a coded packet in broadcast transmission, if we employ no transmission control, the packet would be distributed in all directions by every NN. This would lead to the consumption of unnecessary wireless resources, and the communication quality of the entire network would degrade. Therefore, to minimize the unnecessary consumption of wireless resources, each node that broadcasts a coded packet must generate a forwarders list based on location information and transmission probability for NNs.

We explain here the process of generating a forwarders list for an SN. Note that RNs also use the same process. To make a communication with the DN, the SN must first obtain the location information of the DN. However, in the present paper, we focus on data delivery because the problem is almost as difficult in either case. Thus, we assume, for example, that the SN can obtain the location information of the DN by broadcasting a message such as AODV [4]. After obtaining the location information of the DN (d_x, d_y), the SN calculates the distance to the DN (D_{sd}) by the following equation:

$$D_{sd} = \sqrt{\left(d_x - s_x\right)^2 + \left(d_y - s_y\right)^2}, \tag{2}$$

where (s_x, s_y) indicates the location information of the SN. Next, in order to make a forwarders list, the SN also calculates the distances between each one-hop NN (n_x, n_y) in its own routing table and the DN (D_{nd}). Then, based on the following equation, the SN selects NNs that are closer to the DN than the SN as candidate RNs:

$$N = \left\{n \mid n \in D_{sd} > D_{nd}\right\}, \tag{3}$$

where N is a set of candidate RNs. Thus, this process can provide the direction in order to deliver packets to the DN (see Figure 7). In this way, although packets can approach the DN, reachability is not guaranteed because no consideration is given to the communication quality between nodes. In addition, selecting more RNs leads to an increase in network load. The proposed method thus needs to restrict RNs that forward coded packets by considering the transmission probability to candidate RNs.

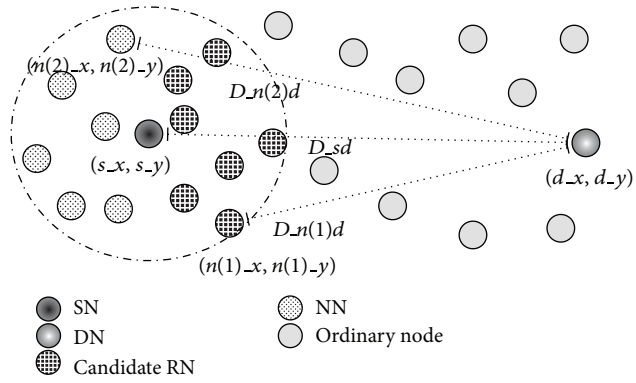


Figure 7: Selection of candidate RNs.

As described in Section 3.2, we employ PER as a metric of communication quality to NNs. A sender selects some RNs with high PER to enhance the transmission range and to conserve wireless resources. If the sender simply selects a number of RNs with low PER, wireless resources are unnecessarily consumed because the number of hops to the DN increases. Hence, since there is a trade-off between PER and the number of hops to the DN, the proposed method selects RNs with high PER.

From the candidate RNs (N), obtained from (3), we determine the appropriate RNs to actually forward a coded packet (see Figure 8). To obtain a transmission probability

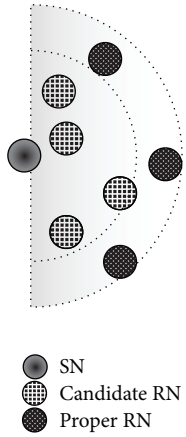


Figure 8: Selection of proper RNs.

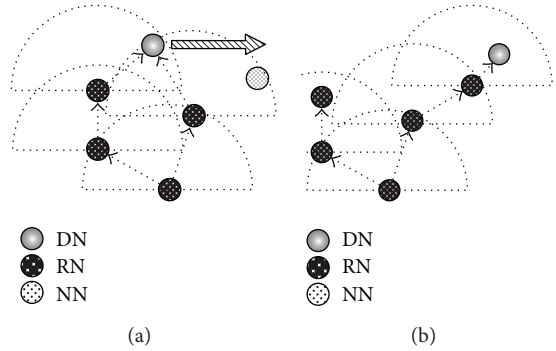


Figure 9: Updating of the DN location information by RNs.

greater than a certain value (α_t), we select RNs according to the following equation:

$$1 - \prod_{N} \mathrm{PER}_{sn} > \alpha_t, \tag{4}$$

where s indicates the SN and n is a candidate RN, selected in descending order of PER. Candidate RNs are selected until (4) is satisfied or all candidate RNs have been selected. By selecting multiple candidate RNs, the transmission probability is adjusted to α_t or higher. In this way, the SN specifies the next proper RNs in the forwarders list and then forwards a coded packet with the forwarders list. Therefore, when the NNs receive the coded packet that the SN broadcasted, only the NNs specified in the forwarders list become RNs. These RNs then broadcast the coded packet after selecting the next RNs, just as the SN did.

3.4. Updating of DN Location Information by RNs. In MORE, the only packet that the SN can receive from the DN is an ACK packet, which is sent after the DN has decoded an entire batch. In the proposed method, if the SN was able to update the location information of the DN only when the SN received an ACK packet from the DN, then the SN might hold the erroneous location information of the DN for a long time. In the worst situation, the SN might not be able to deliver packets to the DN if the DN had moved out of RNs' reach. Therefore, to enhance the data delivery to the DN, RNs within two hops from the DN update the location information of the DN to the latest location information based on the NNs' information.

The location information update of DN works as follows. As illustrated in Figure 9(a), when an RN receives a coded packet, it normally selects the next RNs based on its own NN's information and the location information of the DN in the packet header and forwards the coded packet. At this time, if the DN moves, the actual location information of the DN is different from that in the packet header. To alleviate that problem, in the proposed method, if an RN within two hops from the DN has the actual location information of the DN in its own routing table, the RN uses the information in the routing table to select the next RNs and also updates the location information of the DN in the packet header to the information in the routing table. Since each node broadcasts an information packet periodically (e.g., every second), NN's information is considered to contain the latest location information of the DN. Thus, as illustrated in Figure 9(b), even if the SN sends a coded packet including erroneous location information of the DN, the packet has a fair possibility of reaching the DN because RNs within two hops from the DN can update the location information of the DN. On the other hand, if an RN has not recorded the changed location information of the DN in its routing table, it deals with a coded packet based on the erroneous location information in the packet header. Consequently, if all RNs lose track of the DN, the SN may need to rediscover the DN.

4. Simulation Experiments

This section demonstrates the communication performance of the proposed method in a MANET through simulation experiments using Qualnet 4.5.1 [16]. Section 4.1 investigates the basic communication performance of the proposed method and MORE. Sections 4.2, 4.3, and 4.4, respectively, describe the communication performance of the proposed method for the cases in which the DN, the SN, and the RNs move. In addition, Section 4.5 investigates the communication performance for the case in which the SN, DN, and RNs move.

4.1. Communication Performance of The Proposed Method and MORE. This section provides the basic communication performance of the proposed method and MORE in the case when the DN moves at 1 m/s. We first consider the simulation model illustrated in Figure 10, in which 50 nodes are located at regular intervals of 20 m over a 140 meter squared area. All nodes are assumed to be equipped with an IEEE 802.11 g wireless interface of a fixed 54 Mb/s transmission rate. In the simulation experiments, we investigate the communication performance for one flow. In one flow, one batch contains 64 coded packets, each of which consists of 1,500 bytes, and a coded packet is sent every second. One simulation experiment ends when the DN successfully decodes five batches from the SN; that is, the DN must have received at least 320

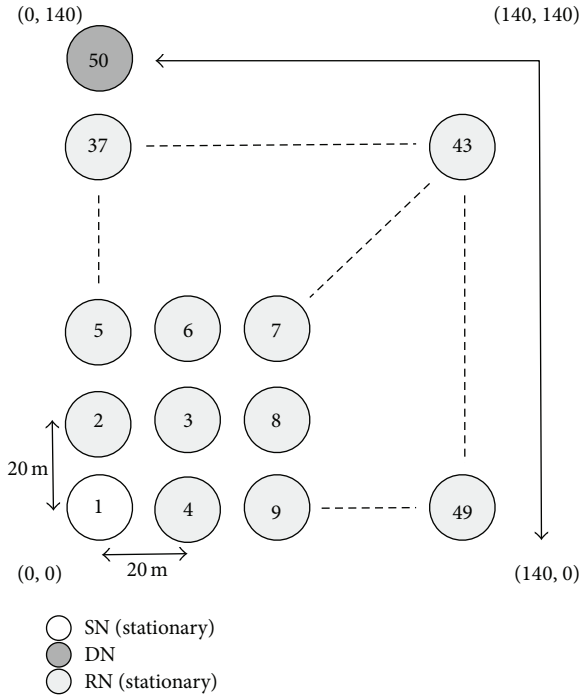


Figure 10: Simulation model 1 (movement of DN).

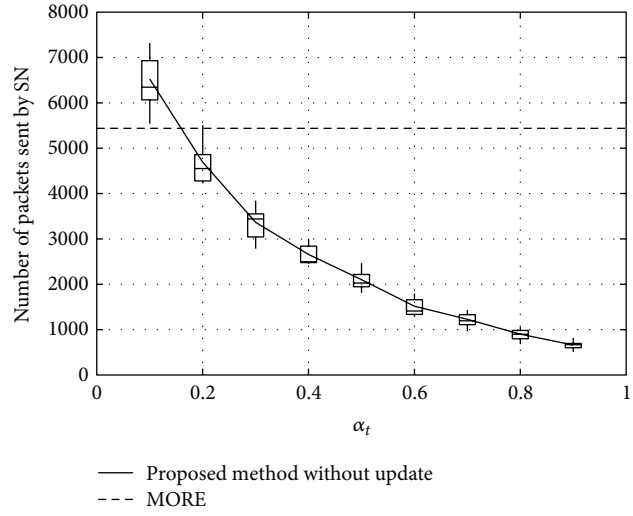


Figure 11: Number of packets sent by the SN.

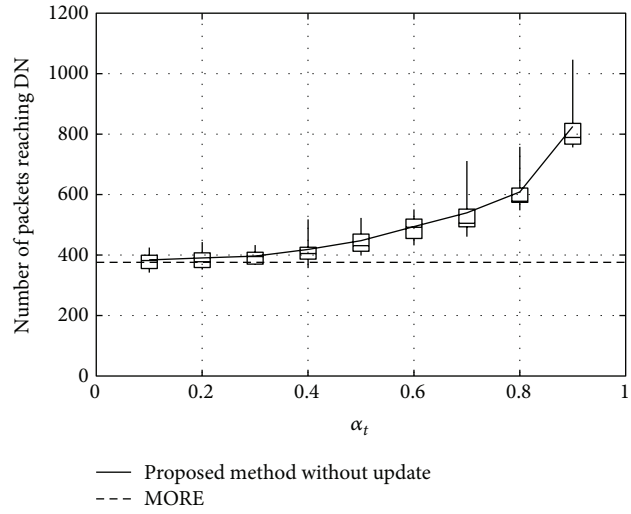


Figure 12: Number of packets reaching the DN.

coded packets. Also, during the experiment, each node also broadcasts an information packet every second.

We now evaluate the communication quality for cases in which the DN moves at 1 m/s. In Figure 10, the SN of node 1, situated at (0, 0), communicates with the DN of node 50, situated at (0, 140). The DN continues to move at 1 m/s according to the following directions: (0, 140) → (140, 140) → (140, 0) → (140, 140) → (0, 140). The remaining nodes are all stationary. To demonstrate the effectiveness of the NNs' information, we investigate the communication performance of the proposed method without location information updating of the DN by RNs. Note that, in the simulation, the SN can update the location information of the DN only when it receives an ACK packet from the DN.

Figures 11, 12, and 13 show the number of packets sent by the SN, the number of packets that reached the DN, and the number of packets forwarded over the entire network, respectively, for α_t, in the case when the DN moves at 1 m/s. Note that, in all the figures, the term "packets" means all the coded packets including innovative and noninnovative packets. In these figures, since the original MORE does not employ α_t, a value obtained for MORE in the simulation mode is employed for all α_t. These figures indicate that the number of packets sent by the SN decreases with increasing α_t, while the number of packets reaching the DN increases. This is because, as the SN and RNs select more RNs with increasing α_t, the number of packets sent by the SN decreases. Moreover, since the higher number of RNs enhances the transmission probability, the number of packets reaching the DN is also increased. Compared with MORE, which cannot appropriately update the NNs' information (i.e., ETXs), the proposed method provides better performance. In Figure 13, the number of forwarded packets tends to initially decrease with increasing α_t but then increases again when α_t exceeds 0.8. This is because the increase in the number of selected RNs, that is, the increased number of hops to the DN, impacts the network load. Particularly, since unnecessary transmissions consume limited wireless resources and lead to performance degradation in a wireless network, we need to reduce the network load as much as possible and make more packets reach the DN. Thus, since an optimal value for α_t is between 0.6 and 0.8, α_t is set to 0.7 in the present study.

4.2. Communication Performance for Movement of DN. In the previous section, we investigated how the movement of the DN impacts the proposed method when the location of the DN is not updated by RNs. This section evaluates the performance of the proposed method with location updating by RNs for various movement speeds (i.e., from 1 m/s to

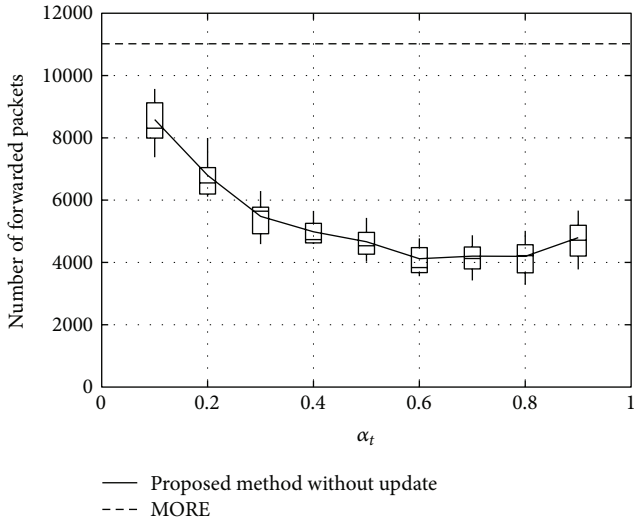


Figure 13: Number of packets forwarded over the network.

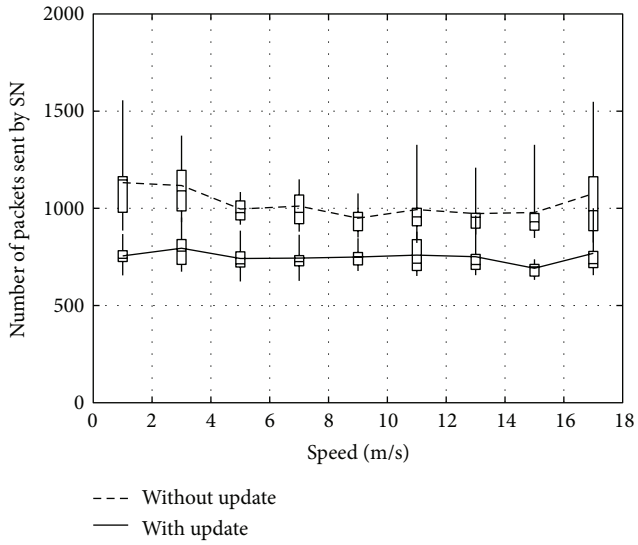


Figure 14: Number of packets sent by the SN.

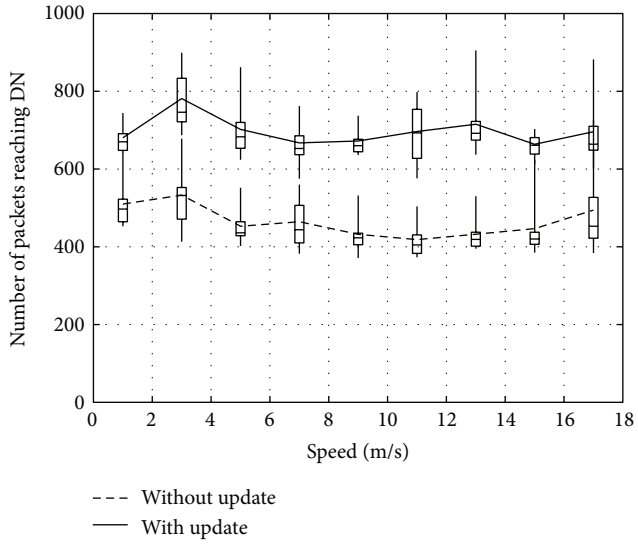


Figure 15: Number of packets reaching the DN.

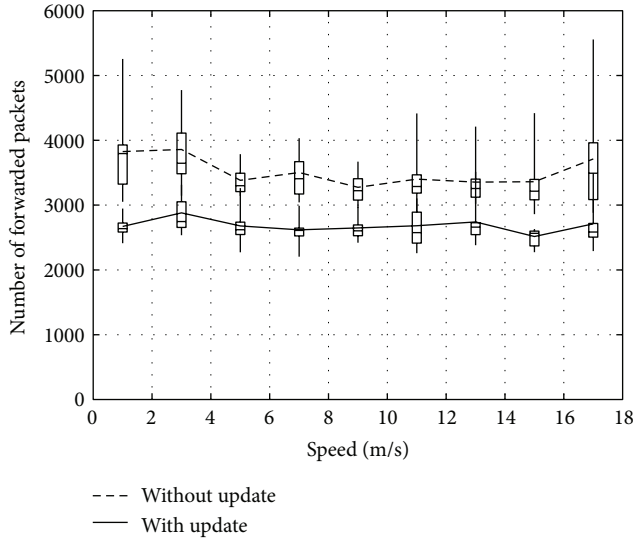


Figure 16: Number of packets forwarded over the network.

17 m/s) of the DN. The DN moves as described in Section 4.1, and the remaining nodes are all stationary.

Figures 14, 15, and 16 show the number of packets sent by the SN, the number of packets that reached the DN, and the number of packets forwarded over the entire network, respectively. The evaluation compares the performance of the proposed method with and without DN's location being updated by RNs. From the results, the method with DN's location being updated by RNs provides better performance irrespective of the movement speed of the DN. Therefore, the method reduces the number of packets sent by the SN, while the DN receives more packets. As a result, the network load is also improved substantially.

4.3. Communication Performance for Movement of SN. This section investigates how SN's movement speed (from 1 m/s to 17 m/s) affects the communication performance of the proposed method with location updating by RNs. As illustrated in Figure 17, the SN of node 50 sends five batches to the DN of node 1. The SN moves between (0, 140) and (140, 0), as did the DN in the previous section, and then the DN and RNs are all stationary.

Figures 18, 19, and 20 show the number of packets sent by the SN, the number of packets that reached the DN, and the number of packets forwarded over the entire network, respectively. As the speed of the SN increases, the number of packets it sends decreases, while the network load increases. On the other hand, the number of packets that reaches the DN is not significantly changed. Now, let us consider the reasons for these results. In the simulation, the positions of the DN and RNs do not change. Thus, the performance degradation with changes in speed is caused by the errors in the PER

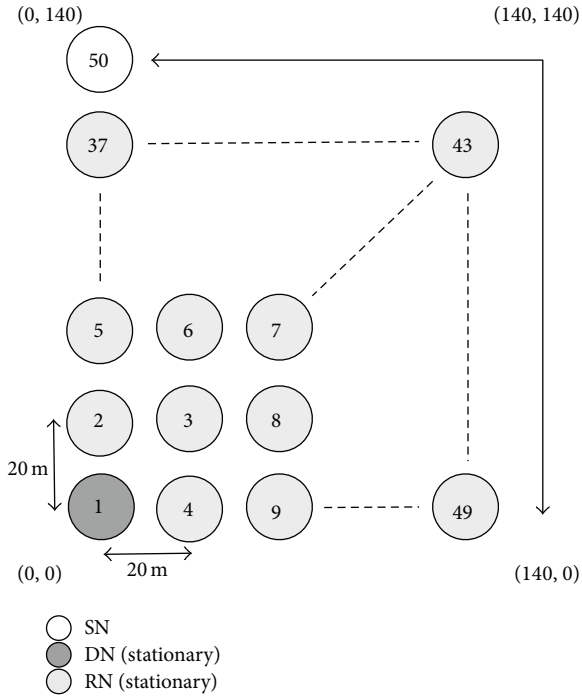


Figure 17: Simulation model 2 (movement of SN).

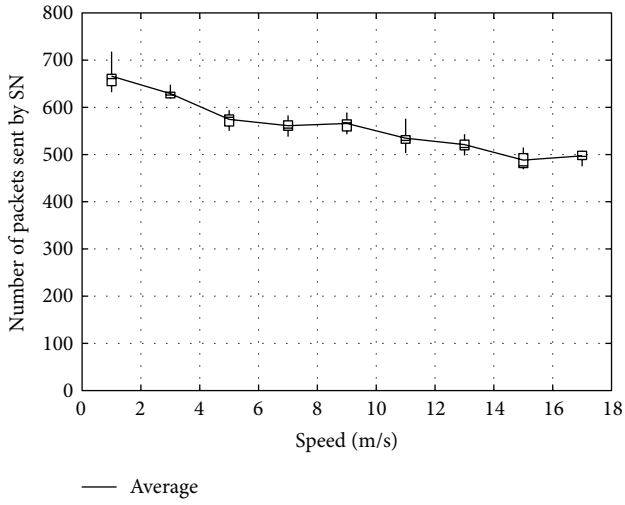


Figure 18: Number of packets sent by the SN.

of the RNs around the SN. Since the PER is calculated from the BER of a received information packet, the calculated and actual PERs for the RNs may differ somewhat due to the movement of the SN. In the proposed method, since RNs are selected in descending order of their PER, the PERs for some RNs are improved when the SN actually sends packets. Conversely, the PERs for some RNs may deteriorate. Consequently, more RNs are able to successfully receive a packet than expected. Therefore, the number of packets sent by the SN decreases, and the network load is increased due to the increase of RNs receiving a packet that the SN sends.

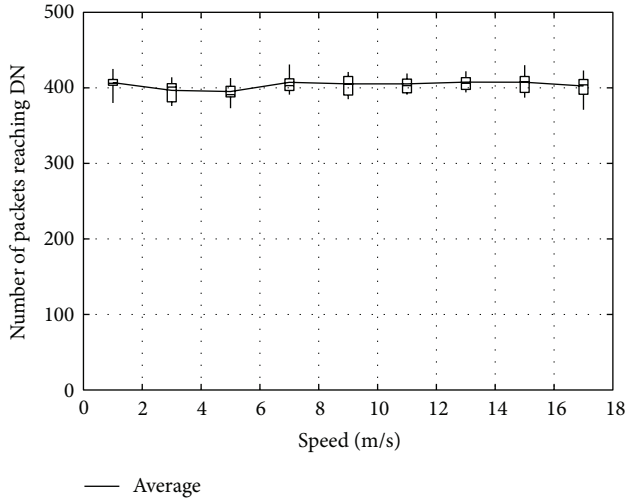


Figure 19: Number of packets reaching the DN.

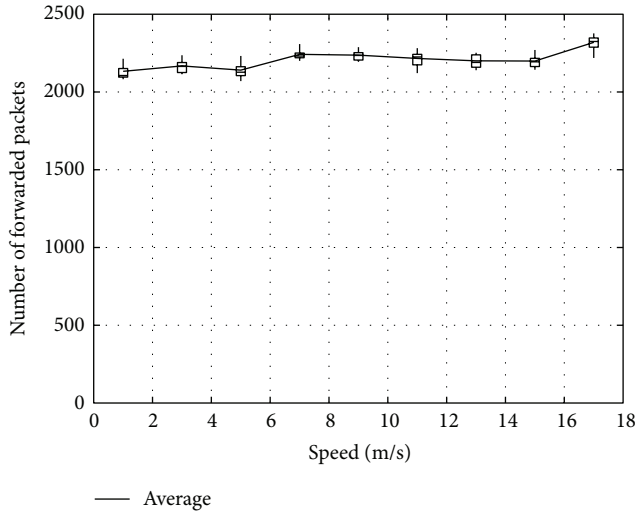


Figure 20: Number of packets forwarded over the network.

4.4. Communication Performance for Movement of RNs. This section evaluates the impact of movement and the number of RNs on communication performance. We here employ two simulation models to describe how the movement and the number of RNs impact the communication performance of the proposed method. In the first simulation, illustrated in Figure 21, odd-numbered RNs are all stationary, whereas the remaining even-numbered RNs move randomly within the simulation area. Figure 22 illustrates the second simulation model, where we employ more moving RNs than in the first simulation. However, when all RNs move randomly, the communication could possibly fail due to there being no RNs in the communicable range. To avoid this problem, we then employ the five situated RNs (whose node IDs are 10, 13, 16, 40, and 46). Therefore, the first simulation has 24 moving nodes, while the second simulation has 43. In both simulation models, the SN and the DN are situated at (0, 0) and (140, 140),

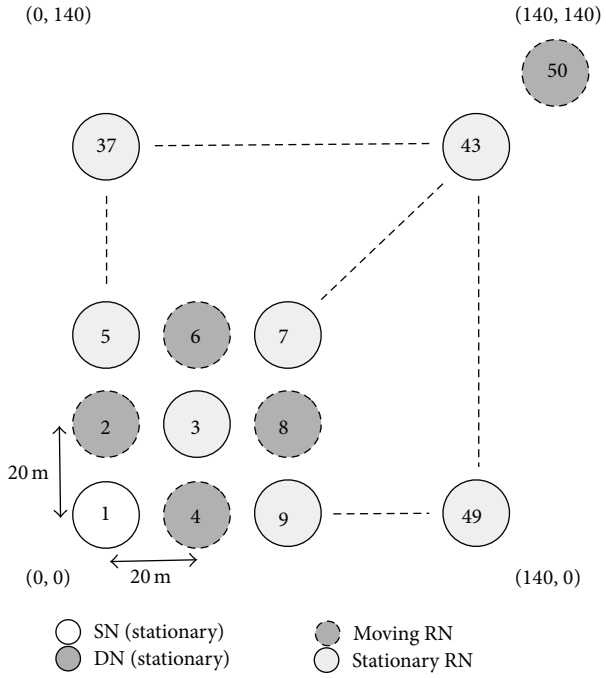


Figure 21: Simulation model 3 (movement of 24 RNs).

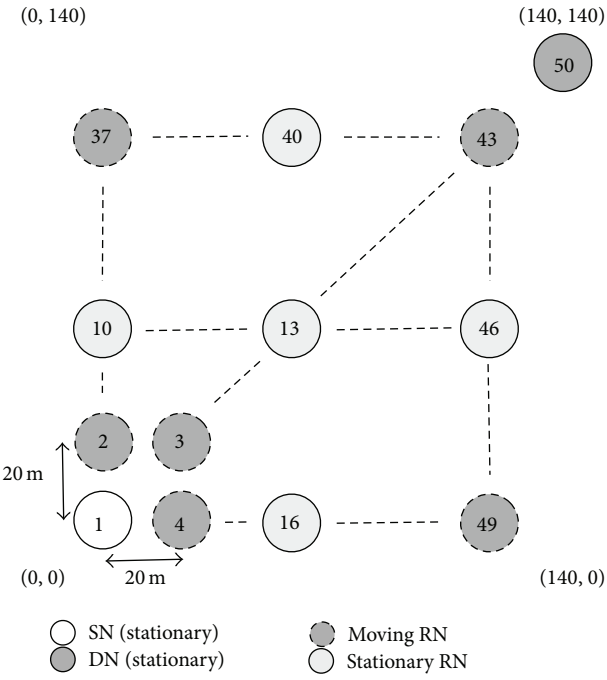


Figure 22: Simulation model 4 (movement of 43 RNs).

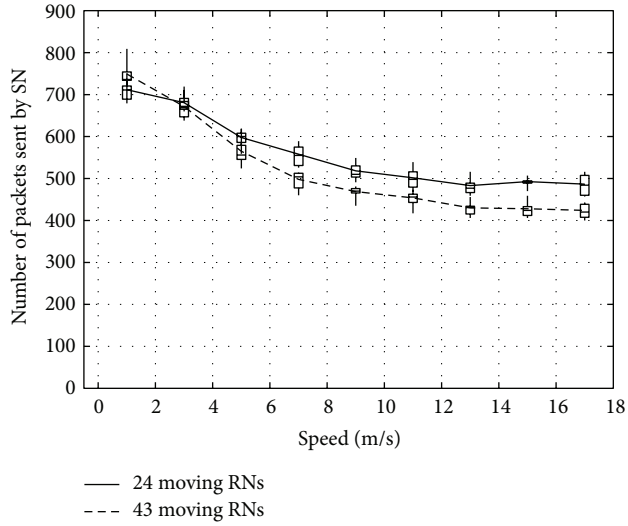


Figure 23: Number of packets sent by the SN.

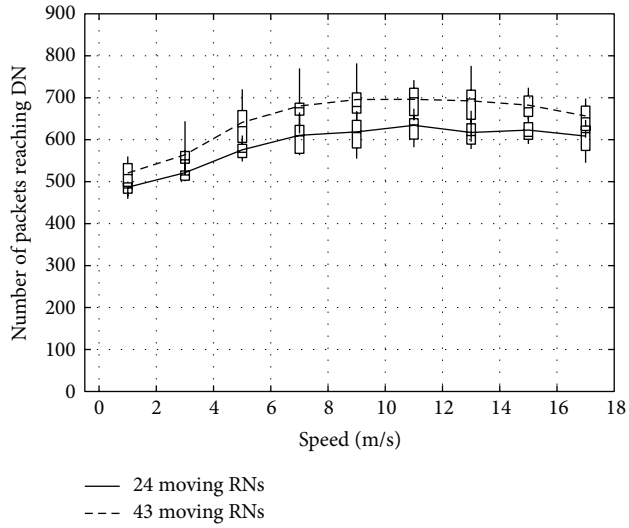


Figure 24: Number of packets reaching the DN.

respectively, and we vary the speed of the moving RNs from 1 m/s to 17 m/s.

Figures 23, 24, and 25 show the number of packets sent by the SN, the number of packets that reached the DN, and the number of packets forwarded over the entire network, respectively. The number of packets sent by the SN decreases with the increasing movement speed of the RNs, whereas the network load increases. On the other hand, with an increase of the forwarded packets, the number of packets that reached the DN also increases. This is because, since the communication quality of some RNs improves due to movement of RNs when a packet is actually sent, more RNs successfully receive coded packets. Moreover, since higher movement speed makes it difficult to obtain appropriate location information, the number of forwarded packets over the entire network increases when the number of moving nodes is larger and their movement speed is higher.

4.5. Communication Performance for Movement of SN, RNs, and DN. This section investigates how the movement of the SN, the DN, and the RNs impacts communication performance. As illustrated in Figure 26, the simulation model employs 51 nodes. The SN of node 51 communicates with the DN of node 50, and they continue to move at 1 m/s according

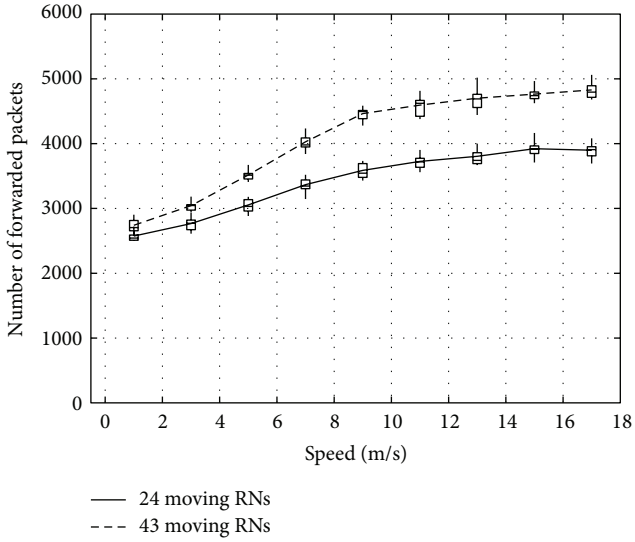


Figure 25: Number of packets forwarded over the network.

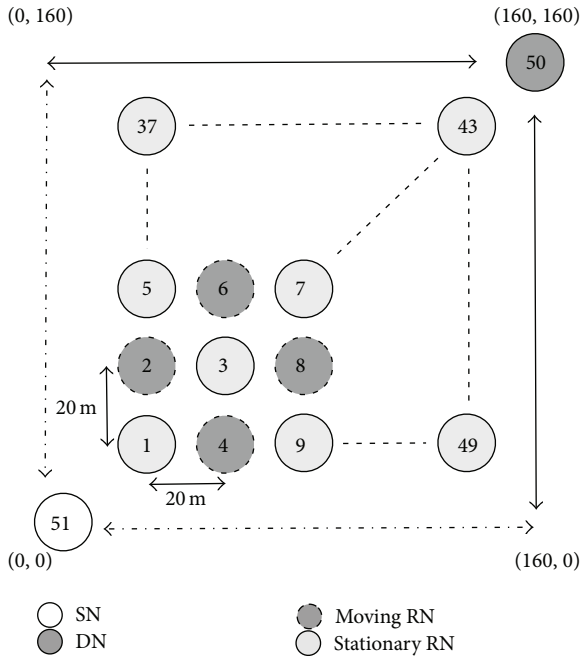


Figure 26: Simulation model 5 (movement of SN, DN, and RNs).

to the following directions (the SN: (0, 0) → (160, 0) → (0, 0) → (0, 160) → (0, 0); the DN: (160, 160) → (0, 160) → (160, 160) → (160, 0) → (160, 160)). Odd-numbered RNs are all stationary, whereas even-numbered RNs move randomly within an area of 160 m by 160 m. We then vary the speed of the moving RNs from 1 m/s to 17 m/s.

Figures 27, 28, and 29 show the number of packets sent by the SN, the number of packets that reached the DN, and the number of packets forwarded over the entire network, respectively. From the results, we can see that there results have almost the same features as the foregoing results.

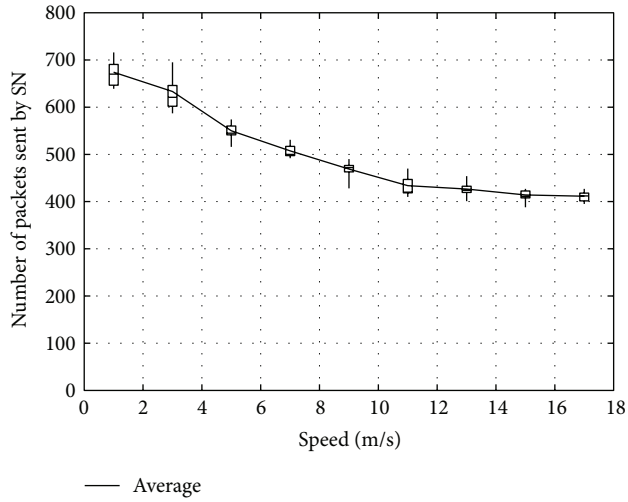


Figure 27: Number of packets sent by the SN.

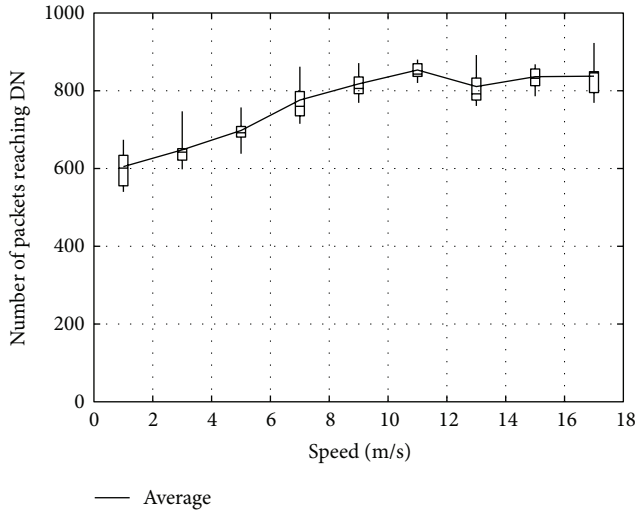


Figure 28: Number of packets reaching the DN.

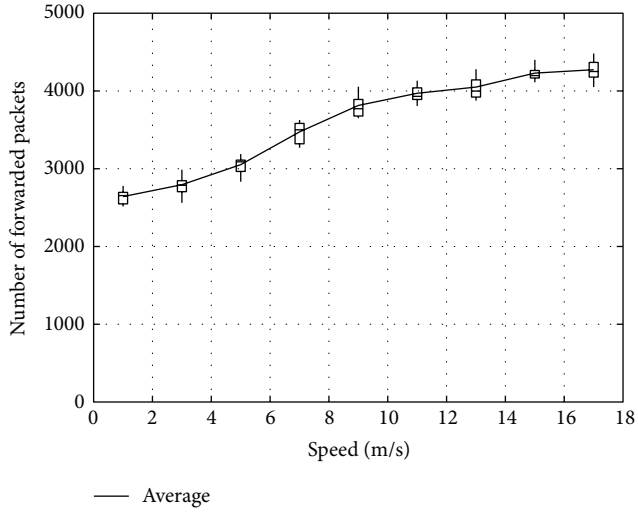


Figure 29: Number of packets forwarded over the network.

5. Conclusion

In the present paper, we proposed a data delivery method based on the NNs' information in a MANET. To enhance reliable communication, the proposed method first employs opportunistic routing and network coding, in the same manner that MORE does. In addition, the proposed method uses the location information and the transmission probability of its own NNs to efficiently deliver packets to the DN. That is, in the proposed method, all of the nodes can hold the latest NNs' information within two hops by updating periodically, and a packet can then be forwarded by RNs that are appropriately selected based on the NNs' information hop-by-hop. Moreover, even if an SN sends a coded packet including erroneous location information about the DN, the packet has a fair possibility of reaching the DN because RNs within two hops from the DN can update the location information of the DN in the coded packet. Thus, the proposed method can deliver data without having to share the communication quality of each node over the entire network. We further demonstrated the communication performance of the proposed method through simulation experiments. The results of these experiments revealed that the proposed method provides better performance than MORE and the proposed method without DN's location update by RNs in a MANET environment. Notably, the proposed method can achieve efficient delivery with low network load when the movement speed is relatively slow.

We demonstrated the basic communication performance of the proposed method in grid simulation models. However, to mimic a more realistic environment, additional evaluations with changing scenarios will be required in the future. We also need to study the case in which different multiple flows exist, that is, the combination of an intraflow network coding and an interflow network coding. Although in this study we did not consider how to discover the DN to start communication in a MANET, this will be investigated in a future study.

Conflict of Interests

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

References

[1] S. Chachulski, M. Jennings, S. Katti, and D. Katabi, "Trading structure for randomness in wireless opportunistic routing," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM '07)*, pp. 169–180, August 2007.

[2] S. Biswas and R. Morris, "ExOR: opportunistic multi-hop routing for wireless networks," *SIGCOMM Computer Communication Review*, vol. 35, no. 4, pp. 133–144, 2005.

[3] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[4] C. E. Perkins, E. M. Belding-Royer, and S. R. Das, "Ad hoc On-Demand Distance Vector (AODV) Routing," *IETF RFC3561*, 2003.

[5] D. B. Johnson, D. A. Maltz, and Y. C. Hu, "The Dynamic Source Routing Protocol (DSR) for Mobile Ad Hoc Networks for IPv4," *IETF RFC4728*, 2007.

[6] T. H. Clausen and P. Jacquet, "Optimized Link State Routing Protocol (OLSR)," *IETF RFC3626*, 2003.

[7] A. Boukerche, B. Turgut, N. Aydin, M. Z. Ahmad, L. Bölöni, and D. Turgut, "Routing protocols in ad hoc networks: a survey," *Computer Networks*, vol. 55, no. 13, pp. 3032–3080, 2011.

[8] Y. Yuan, H. Yang, S. H. Y. Wong, S. Lu, and W. Arbaugh, "ROMER: resilient opportunistic mesh routing for wireless mesh networks," in *Proceedings of the 1st IEEE Workshop on Wireless Mesh Networks (WiMesh '05)*, 2005.

[9] E. Rozner, J. Seshadri, Y. Mehta, and L. Qiu, "SOAR: simple opportunistic adaptive routing protocol for wireless mesh networks," *IEEE Transactions on Mobile Computing*, vol. 8, no. 12, pp. 1622–1635, 2009.

[10] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Medard, and J. Crowcroft, "XORs in the air: practical wireless network coding," *IEEE/ACM Transactions on Networking*, vol. 16, no. 3, pp. 497–510, 2008.

[11] J.-S. Park, M. Gerla, D. S. Lun, Y. Yi, and M. Médard, "CodeCast: a network-coding-based ad hoc multicast protocol," *IEEE Wireless Communications*, vol. 13, no. 5, pp. 76–81, 2006.

[12] M. Mauve, J. Widmer, and H. Hartenstein, "A survey on position-based routing in mobile ad hoc networks," *IEEE Network*, vol. 15, no. 6, pp. 30–39, 2001.

[13] N. Zhi, L. Jing, L. Botong, L. Hanchun, and X. Youyun, "A relay node selection technique for opportunistic routing in mobile ad hoc networks," in *Proceedings of the 15th Asia-Pacific Conference on Communications (APCC '09)*, pp. 665–668, October 2009.

[14] S. Yang, F. Zhong, C. K. Yeo, B. S. Lee, and J. Boleng, "Position based opportunistic routing for robust data delivery in MANETs," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '09)*, December 2009.

[15] D. S. J. De Couto, D. Aguayo, J. Bicket, and R. Morris, "A high-throughput path metric for multi-hop wireless routing," *Wireless Networks*, vol. 11, no. 4, pp. 419–434, 2005.

[16] Qualnet, http://www.scalable-networks.com/products/qualnet/.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 718698, 9 pages
http://dx.doi.org/10.1155/2014/718698

Research Article

QoS-Oriented High Dynamic Resource Allocation in Vehicular Communication Networks

Tarek Bejaoui

FSB, University of Carthage, 02 rue de Sfax, Bizerte 7000, Tunisia

Correspondence should be addressed to Tarek Bejaoui; tarek.bejaoui@yahoo.fr

Received 29 August 2013; Accepted 4 December 2013; Published 30 January 2014

Academic Editors: W. Sun and G. Zhang

Vehicular ad hoc networks (VANETs) are emerging as new research area and attracting an increasing attention from both industry and research communities. In this context, a dynamic resource allocation policy that maximizes the use of available resources and meets the quality of service (QoS) requirement of constraining applications is proposed. It is a combination of a fair packet scheduling policy and a new adaptive QoS oriented call admission control (CAC) scheme based on the vehicle density variation. This scheme decides whether the connection request is to be admitted into the system, while providing fair access and guaranteeing the desired throughput. The proposed algorithm showed good performance in testing in real world environment.

1. Introduction

Similar to other mobile ad hoc networks, vehicular ad hoc networks (VANETs) are self-organizing and are formed directly by a set of smart vehicles. They can interact without using fixed infrastructure or centralized administration. They have, moreover, several characteristics that distinguish them from the other networks. The mobility of vehicles in different traffic conditions, such as during traffic jams, accidents, traffic lights, and rush hours, results in the dynamic change in the network topology. In these networks, vehicles move only on predetermined roads, and they do not have the problem of resources limitation in terms of data storage and power.

In high dynamic vehicular network topology like in rural highways or during late night, the vehicles move with high speed and thus the vehicle-to-vehicle communication link remains active for short-time duration. Increasing the transmission range and then the transmission power is one of the solutions that could be considered to prolong the vehicle-to-vehicle or the vehicle to infrastructure roadside connection time. In vehicular networks the high transmission power results high interference and high network overhead in highly dense traffic (e.g., area with high penetration ratio or urban area or traffic jams). Therefore, in order to address these problems, dynamic adaptation of transmission range/power is crucial. It is required to decrease the transmission power for high vehicle density or high penetration ratio and increase the transmission power for less vehicle density or low penetration ratio. Thus, the vehicle density becomes a single metric describing the nodes' mobility in VANETs [1] that should be taken into account in any resource allocation algorithm that could be developed for these networks.

In this paper, we propose therefore a novel vehicle density-based call admission control scheme for vehicular networks that dynamically adapts vehicles' transmission powers and provides desired throughput guarantees to users performing communication within a real world-like vehicular-to-roadside IEEE 802.11p communication networks.

The paper is organized as follows. In the next section, we present an overview of the 802.11p standard and recent works devoted to CAC in VANETs, and then we highlight the impact of vehicle density on network performance. In Sections 3 and 4 we present, respectively, the new CAC function and the novel scheduling policy that we propose for VANETs. We describe our simulation platform in Section 5, in which some quantitative results are reported. Concluding remarks are presented in Section 6.

2. An Overview of 802.11p and CAC Implementation in VANET

In this decade, the smart vehicles become part of the Intelligent Transportation System (ITS). The MAC and physical layers of this system are supported by IEEE 802.11p Wireless Access in Vehicular Environments (WAVE) standard [2]. In this section we will briefly outline the design and concept of the WAVE standard [3] relevant to our work. The physical layer can rely on seven channels of 10 MHz bandwidth each. The spectrum of WAVE is allocated in the upper 5 GHz range.

The MAC layer in WAVE is equivalent to the IEEE 802.11e Enhanced Distributed Channel Access (EDCA) quality of service (QoS) extension. The IEEE 802.11e standard is often used for VANET prototyping implementations to support quality of service and the EDCA is the preferred channel access of this standard. It is used for service differentiation and has four access categories (AC) that support priority-based service [4]. Within the MAC layer a packet queue exists for each AC.

There is a set of EDCA parameters associated with each AC. Those parameters include arbitration interframe space (AIFS[AC]), contention window (CW) with its minimum and maximum values CWmin[AC] and CWmax[AC], and the transmission opportunity (TXOP). A backoff mechanism is defined for each ACi ($i = 0, 1, 2, 3$).

Each AC from every station starts independently a backoff timer after detecting that the channel is idle for an AIFS[AC] interval and competes with other ACs to gain a transmission opportunity. For each AC, the backoff period is selected from a uniform distribution over [0, CW[AC]]. In 802.11p, the CW size is initially assigned "CWmin" value and becomes equal to "$CW = 2*(CW+1) - 1$" when transmission fails up to the CWmax. The smaller the AIFS[AC] and the CWmin[AC] or the larger the TXOP, the shorter the channel access delay for the corresponding priority and hence the better chance to access the media for a given traffic condition.

2.1. Admission Control: Related Work. The VANET MAC policies proposed in the literature often take less into consideration power constraints or time synchronization problems. However, they care about the fast topology changing and the different kinds of applications for which the transmission will be established.

EDCA is very likely to be the dominant channel access mechanism in such networks because it is a distributed MAC scheme and easy to implement. Many research works are focusing on admission control in EDCA. It was proposed to administer policy or regulate the available bandwidth resources. Basically, the existing EDCA admission control schemes can be classified into two categories: measurement-based admission control and model-based admission control.

In the measurement-based schemes, admission control decisions are made on the basis of the continuously measured work conditions such as throughput and delay. In [5, 6], for example, the Distributed Admission Control policy (DAC) was proposed and for which the transmission budget for each AC is computed by subtracting the occupied time from the transmission limit of this AC.

The DAC policy can only protect existing flows when the traffic load is not very heavy. In addition, this scheme does not provide direct relationships between TXOP parameters and the QoS requirements from applications.

In [7] authors presented the Two-Based Protection and Guarantee mechanism, which is based on the DAC scheme. The purpose of the first level protection is to protect each existing voice or video flow from new and other existing QoS flows, while the purpose of the second level protection is to protect the existing QoS flows from best-effort traffic. However, this scheme has the problems of performance oscillation and lack of direct QoS relationships with applications.

The authors in [8] propose the Virtual MAC and Virtual Source Algorithms that virtually run the applications and the MAC processes in order to measure the achievable service qualities. The Threshold-Based Admission Control protocol presented in [9] takes into account the traffic condition on the wireless link. Depending on how the traffic condition is computed and measured, it can be implemented while using the relative occupied bandwidth or using the average collision ratio. In the HARMONICA scheme [10] the link-layer quality indicator (LQI) parameters are periodically sampled. They include drop rate, link layer end-to-end, and throughput for each traffic class. Two adaptation algorithms over different time scales are employed to select the channel access parameters, which can best match the QoS requirements of each traffic class and the current channel contention level. Whenever a new real-time application requires admission, HARMONICA will select a traffic class i that best matches its QoS requirement and then execute an admission control process. The decision of admission control is then based on the throughput requirement of the flow and the monitored LQI parameters. This is done while checking whether it is possible to squeeze some bandwidth out of the current throughput for the best effort class on the condition of guaranteeing a minimal bandwidth for this class.

On the other hand, many model-based schemes were proposed for EDCA. They define some performance metrics to evaluate the status of the network. One of such protocols is the Markov Chain Model-Based Admission control [11] in which the admission control is performed on the basis of the predicted achievable throughput for each flow. Authors in [12] propose an algorithm for providing throughput guarantee services in 802.11e EDCF wireless LANs and a Contention Window Based Admission control. Its key idea is to adjust the CW values for different stations so that the goals of admission control can be fulfilled. The main disadvantage of this scheme is to not consider the nonsaturation conditions and virtual collision, and this is the case of the Markov Chain Model Based Admission control scheme presented before.

Very few works dealing with the optimization of the minimum contention window values CWmin were presented in the literature. In [13], for example, authors consider the case where stations can have different weights corresponding to different throughput classes. Their test-bed evaluation considers both the long-term throughput achieved by wireless nodes and the short-term fairness. When all the nodes have

the same transmission rate, optimality is achieved when a station throughput is proportional to its weight factor, and the optimal minimum contention windows maximize the aggregate throughput. When stations have different transmission rates, the optimal minimum contention window for high rate stations is smaller than for low rate stations.

2.2. Vehicle Density Impact in VANET. Common mobility models of vehicular ad hoc networks allow station mobility to be considered independently of their density in admission control protocols development and performance evaluation. In contrast, car-following models [14] show that the average speed of vehicles is a function of the vehicle density in the area. This would motivate the use of "density" as a single metric describing the nodes' mobility in VANETs.

Nodes' density has a great impact on the performance of ad hoc networks by influencing factors such as capacity, routing efficiency, robustness, and delay. Waves of traffic jams, whether caused by constraints in the transportations network, driving fluctuations, or traffic controls, cause the network density to vary from one location to another, thus disturbing the homogenous distribution of nodes. Moreover, the abrupt and frequent change in density creates a highly dynamic topology that would cause severe degradation to the network performance (increased collisions and interference, excessive broadcasts, too many routing paths, etc.) if protocols in VANETs were not designed to handle such conditions. Controlling the communication range by adjusting the transmission power can be used to mitigate the adverse effects of high density condition. The choice of the communication range has a direct impact on the connectivity which represents a fundamental property of an ad hoc network. In a VANET, a static transmission range cannot maintain the network connectivity due to the non-homogenous vehicles distribution and rapid change of traffic conditions. In [15], authors estimate the lower and upper bounds for the transmission range and in [16] they provide a probability for gap existence among nodes. In networks of infinite size, the transmission range is related to vehicle density rather than the line length [17]. Connectivity in infinite networks is limited to short range communications, and a large-scale ad hoc network is not feasible because it is almost surely divided into an infinite number of partitions.

In [1], authors provide a relationship that allows to estimate the local density and distinguish between two phases of traffic, free-flow and congested traffic. The density estimate is used to develop an algorithm that sets a vehicle transmission range dynamically according to local traffic conditions.

On the basis of some policies presented above, we have developed a novel resource allocation scheme in the context of highly dynamic vehicular communication networks architecture. Its features are presented in the following.

3. Throughput Guarantee and Vehicle Density-Based CAC Algorithm: Description and Architecture

The standards 802.11e EDCA and the 802.11p that could be considered for communications in vehicular networks define a number of parameters that can be used to achieve service differentiation. However, they do not define how these parameters should depend on the network conditions like the load and the traffic characteristics in order to efficiently utilize the shared wireless channel. Moreover, common mobility models of vehicular ad hoc networks allow station mobility to be considered independently of their density in admission control protocols development and performance evaluation. However, nodes' density could have a great impact on the performance of ad hoc networks by influencing factors such as capacity, routing efficiency, robustness, and delay. This provides then the motivation to use "density" as a single metric to describe the nodes' mobility in VANETs.

In this paper, we propose then a novel CAC algorithm that provides the desired throughput guarantees on the basis of the vehicle density and the nodes' transmission range in 802.11p vehicular ad hoc networks. We consider vehicle-to-roadside (V2R) communications which are essential to properly manage traffic situations [18]. This scheme uses a cross layer approach as it adapts the transmission power (in PHY layer) and optimizes the contention window size (in MAC-layer) on the basis of the vehicle density estimation, to enhance the performance of vehicular communications.

The throughput guarantee investigation will be based on the algorithm proposed by Banchs et al. presented in [12] and that was originally performed for EDCF. It was proven to be efficient and we will adapt it in the context of 802.11e EDCA. The novelty is to take into consideration the vehicle density, as well as the transmission range of each vehicle operating within the coverage area of a road infrastructure access point (AP).

Let r_i be the throughput experienced by station i. According to Banchs et al. in [12], r_i is represented by the following expression:

$$r_i = \frac{w_i}{\sum_i w_i} \times \frac{l}{T_s - T_c + \left(P_e \cdot \left(T_e - T_c\right) + T_c\right)/P_s}, \quad (1)$$

where l is the average payload length, T_s is the average duration of a successful transmission, T_c is the average duration of a collision, T_e is the duration of an idle time slot, P_s is the probability of a successful packet transmission, and P_e is the probability of an empty time slot. Consider the following:

$$w_i \text{ was defined as : } \quad w_i = \frac{\tau_i}{\tau_1}, \quad (2)$$

where "τ_i" is the probability that station i transmits in a generic time slot and "τ_1" refers to the "*reference station*" from which the AP receives the highest signal power. And since l, T_s, and T_c are constants, the maximum throughput experienced by all r_i is already defined in [12] as the following:

$$\widehat{r} = \frac{P_s}{P_e \cdot \left(T_e - T_c\right) + T_c} = \frac{\sum_i w_i \tau_i \prod_{j \neq i} \left(1 - w_j \tau_1\right)}{\prod_i \left(1 - w_i \tau_1\right)\left(T_e - T_c\right) + T_c}. \quad (3)$$

The experienced throughputs given by (1) define the set of contention window $\{CW_1, \ldots, CW_n\}$ that meets the throughput requirement set $\{R_1, \ldots, R_n\}$ for n stations and guarantee the throughput provision to them; that is,

$$\forall i, \in \{1, \ldots, n\}, \quad r_i \geq R_i, \quad (4)$$

where r_i represents the throughput actually experienced by station i and R_i is its throughput requirement.

In this paper, instead of defining one reference node for all vehicles connected to the infrastructure Access Point, on the basis of its highest throughput and highest delivered signal level, we refer to each vehicle "i" competing for resources within this AP coverage area (Figure 1), a reference node "ref_i" [19].

The "ref_i" will be then selected from the pool of the vehicles within the transmission range of the current node i. This would be more adequate and fairer.

The transmission range of each station operating in the network should be then computed. This is done in order to define the set of vehicles operating within its coverage area and from which the reference node will be chosen. The reference node "ref_i" is always the node that uses the highest throughput and delivers the highest signal level as well.

Therefore, we consider the relationship that allows a vehicle to estimate the local density. This relationship, defined in [1], is used to set the vehicles' transmission range (TR_i) dynamically and then the transmission power as depicted in Figure 1. It is a function of the vehicle local density and given by

$$\text{TR}_i = \min\left(L \times (1-K), \sqrt{\left(L \times \frac{\ln(L)}{K} + a \times L\right)}\right), \quad (5)$$

where "L" is the length of the road segment, "a" is a traffic constant obtained from traffic flow theory [1, 20], and K is the estimated vehicle density given by

$$K = \left[\frac{u_f K_j}{\lambda'}(1-F_s)^{n+1} + 1\right]^{-1}, \quad (6)$$

where n is a parameter that indicates the quality of service in the transportation network. Both n and λ' reflect the traffic service level of the road and can be determined statistically [1].

F_S is the average fraction of vehicles stopped in traffic during a time window size T. The choice of this parameter depends highly on the rate of change in traffic conditions.

$(1-F_S)^{n+1}$ represents then the normalized average traffic speed and $1/K_j$ is the average distance from front-bumper to front-bumper between vehicles.

u_f is the average speed of vehicles including the stopped ones.

The vehicle density surrounding a node i will be the main parameter that will define its transmission range. And then its own transmission power will be calculated and dynamically adjusted accordingly.

Note that, in free-flow traffic where there are no interactions between vehicles, equation (5) cannot provide a right estimate of the vehicles' transmission range.

The throughputs "r_i" that all communicating vehicles will receive will be optimized as well as the number of nodes to be admitted into the system. τ_1 will be then replaced by τ_{ref_i} in (2) and (3).

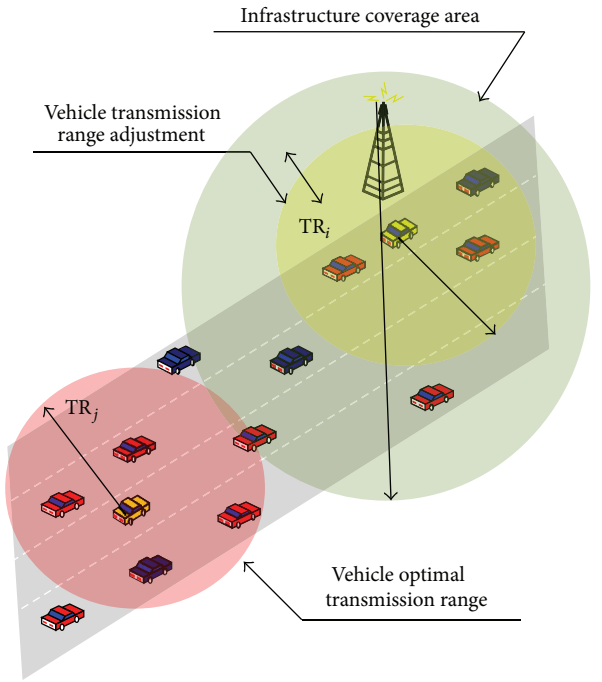


Figure 1: Transmission range/power adaptation based on local density.

In this paper we consider that no exponential backoff is considered. According to Bianchi in [21], "τ_i" is given by

$$\tau_i = \frac{2}{\text{CW}_i + 1}, \quad (7)$$

where CW_i is the minimum contention window of the station i.

Since $\tau_{\text{ref}_i} \ll 1$, the throughput experienced by a station i could be approximated as the following:

$$\hat{r}_i^{\text{opt}} \approx \frac{a\tau_{\text{ref}_i} - b\tau_{\text{ref}_i}^2}{c\tau_{\text{ref}_i} + T_e}, \quad (8)$$

where

$$a = \sum_i w_i, \quad b = \sum_i \sum_{j \neq i} w_i w_j, \quad c = \sum_i w_i \cdot (T_c - T_e), \quad (9)$$

with

$$w_i = \frac{\tau_i}{\tau_{\text{ref}_i}}. \quad (10)$$

Let $\tau_{\text{ref}_i}^*$ be the optimal value of τ_{ref_i} that maximizes $\hat{r}_i^{\text{opt}}$. Thus, for $d\hat{r}_i^{\text{opt}}/d\tau_{\text{ref}_i}|_{\tau_{\text{ref}_i}=\tau_{\text{ref}_i}^*} = 0$ we obtain $bc \cdot (\tau_{\text{ref}_i}^*)^2 + 2bT_e \cdot \tau_{\text{ref}_i}^* - aT_e = 0$, and then

$$\tau_{\text{ref}_i}^* = \frac{\sqrt{(bT_e)^2 + abcT_e - bT_e} - bT_e}{bc}. \quad (11)$$

The optimal contention window set that we propose and that maximizes simultaneously all throughputs r_i experienced by each vehicle within the transmission range of the AP will be defined then as

$$\mathrm{CW}_i^{\mathrm{opt}} = \frac{2}{\tau_i^*} - 1 = \frac{2}{w_i \tau_{\mathrm{ref}_i}^*} - 1 = \frac{2}{(\tau_i/\tau_{\mathrm{ref}}) \cdot \tau_{\mathrm{ref}_i}^*} - 1. \quad (12)$$

When a new station $(n + 1)$ that requires a throughput R_{n+1} would like to join the road infrastructure, the AP first computes a new contention window set $\{\mathrm{CW}_1^{\mathrm{opt}}, \ldots, \mathrm{CW}_n^{\mathrm{opt}}, \mathrm{CW}_{n+1}^{\mathrm{opt}}\}$ using (12) and taking into account the new incoming vehicle to compute the reference node for each station. Then, it uses (8) to calculate the new throughputs that the $(n + 1)$ stations would receive with this new contention window set.

If the resulting throughputs meet the requirements, that is,

$$\forall i, \in \{1, \ldots, n+1\}, \quad \hat{r}_i^{\mathrm{opt}} \geq R_i, \quad (13)$$

then the station $(n + 1)$ is accepted within the network and the new contention window set is distributed to all stations. Otherwise, the station $(n + 1)$ is rejected.

Finally, it is to remind that the function of the vehicle density estimation that we have took into account is not usable in the free-flow traffic. So, the CAC scheme that we have proposed in this paper and the adaptation of the contention window are valid only in the case of dense traffic.

4. Priority Level Computing Function

In this section we propose a set of rules to prioritize traffic under certain conditions in the context of VANET. They are governed by an appropriately designed utility objective priority function that we have developed. This function, for which we give the acronym "PLCF" for priority level computing function, is used to make the resource allocation decision fairer. Requests' packets will be queued according to their access categories defined in 802.11p and, after, they will be served according to the EDCA scheduler after being adjusted with PLCF. Because of their importance, we choose the following network parameters for PLCF.

(i) Security (S): when the information being exchanged is confidential, packets with high encryption will be served firstly.

(ii) Link quality (LQ): packets sent with the strongest signal strength have the highest priorities.

(iii) Roadside conditions (D): available bandwidth is used to indicate the access point conditions and is a major factor especially for services belonging to access categories AC[2] and AC[3] as voice and video traffic.

(iv) Precedence class (P): packets of users that have negotiated with the service provider a premium service will be served in first followed by those of medium and low priorities, when the system is close to congestion.

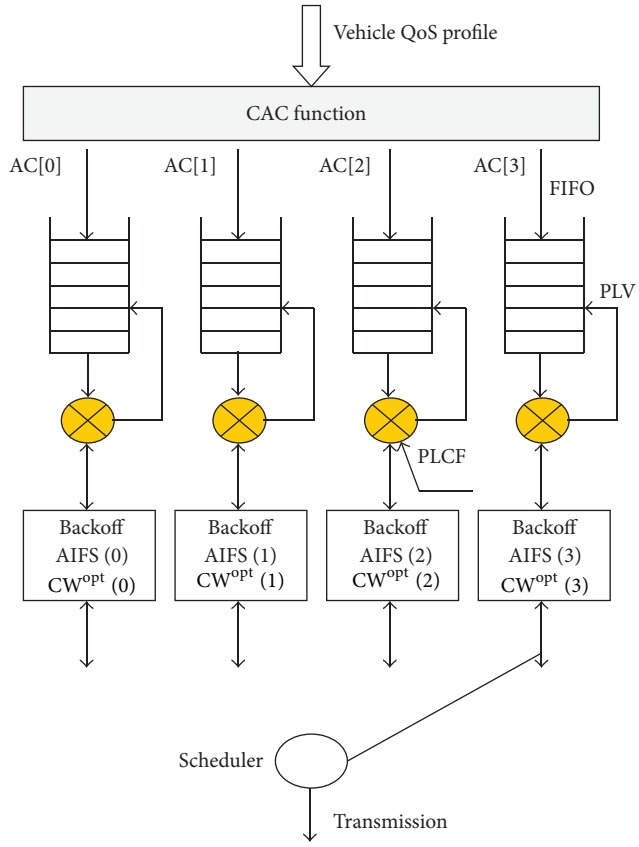


Figure 2: Structure of the fair packet scheduler.

(v) Waiting time (WT): as real time services are very sensitive to access delay, packet losses are due to excessive waiting times and the higher priority is therefore given to packets having waited for the longest average time.

PLCF is evaluated for all packets queued in buffers reserved to each vehicle access category (cf. Figure 2). Packets with the highest calculated value for PLCF will be the first to be scheduled.

The priority level value (PLV), which provides a measure of the packet priority, is measured via the function

$$\mathrm{PLV} = \mathrm{PLCF}\left(S_i, D_i, P_i, \mathrm{LQ}_i, \mathrm{WT}_i\right). \quad (14)$$

In order to allow for different circumstances, there is an apparent necessity to weight each factor relative to the magnitude it endows upon the priority computing function. Therefore a different weight is introduced as follows:

$$\mathrm{PLV} = \mathrm{PLCF}\left(w_s S_i, w_d D_i, w_p P_i, w_{\mathrm{lq}} \mathrm{LQ}_i, w_{\mathrm{wt}} \mathrm{WT}_i\right), \quad (15)$$

where $w_s, w_d, w_p, w_{\mathrm{lq}}$, and w_{wt} are weights for each of the network parameters. The values of these weights are fractions; that is, they range from 0 to 1. Furthermore all five weights add up to 1.0. Each weight is proportional to the significance of a parameter to the priority level computing. The larger the weight of a specific parameter, the more important that

parameter is to the user and vice versa. These weights could be estimated and monitored by the access point of the roadside infrastructure according to the vehicle density in its coverage area; the weights that refer to the precedence class and the packets waiting time parameters will have, for example, the highest values when the system is close to congestion.

Even though we could add the different parameters in the priority level computing function to obtain packets priority level values, each vehicular network parameter has a different unit which leads to the necessity of normalization. The final normalized equation for n packets is

$$\begin{aligned} \mathrm{PLV} = {} & \frac{w_s S_i}{\max(S_1, \ldots, S_n)} + \frac{w_d D_i}{\max(D_1, \ldots, D_n)} \\ & + \frac{w_p P_i}{\max(P_1, \ldots, P_n)} + \frac{w_{\mathrm{lq}} \mathrm{LQ}_i}{\max(\mathrm{LQ}_1, \ldots, \mathrm{LQ}_n)} \\ & + \frac{w_{\mathrm{wt}} \mathrm{WT}_i}{\max(\mathrm{WT}_1, \ldots, \mathrm{WT}_n)}. \end{aligned} \tag{16}$$

For each access category (AC[0], AC[1], AC[2], AC[3]) we may have two buffers so that eight buffers serving different types of traffic could be allowed to ongoing packets of each user priority defined in 802.1D.

The backoff mechanism used in the contention-based channel access is independent of the priority level computing function that we have proposed. The PLCF is then developed to provide fairness among competing 802.11p users belonging to the same access category.

5. Performance Evaluation

5.1. Simulation Platform. We investigate the adaptation of the contention window on the basis of the vehicle density in the IEEE 802.11p vehicular networks. The investigation is performed using the same traffic conditions presented in [3], while considering both time and spatial dimensions of the traffic variation. The time variation of traffic is represented as arrival process, call duration, or packet length for various types of services. Spatial variations characterize the vehicle mobility in the coverage area of 500 m of an 802.11p access point of the road infrastructure which uses an omnidirectional antenna. In the real world, an 802.11p access point is expected to cover a distance that can vary between 300 and 1000 m.

In this paper, we consider that the edges of the network system are wrapped around such that the "border effect" is suppressed, as in a real world environment. As we consider dense traffic, the vehicles' speeds vary randomly between 3 km/h and 50 Km/h. These vehicles are characterized by their positions that vary each 0.5 s in the coverage area of the access point. In addition, the power transmissions vary from one terminal to another, according to their transmission range which depends on the vehicle density in the network.

We consider that signals' strength received by vehicles to define the node reference is computed while considering the three stage propagation model [22]. It takes into account the path loss, the Rayleigh fading, and the shadowing as well.

In this work, we consider that generated vehicles move in straight highway lines (cf. Figure 1) and do not change their direction as their movement is constrained by the road. In these conditions, the estimated vehicle density given by (6) will be computed using the following parameters values: $n = 0$ and $1/\lambda' = 2$ s [1].

The average fraction "F_S" of vehicles stopped in traffic is defined during a time window size $T = 10$ s. It results in high correlation between the actual density and its estimate.

We neglect the interference produced by each vehicle on other stations in the network and we chose a packet size of 1500 bytes, since this is a reasonable average packet size including data and security information.

We assume that all stations are greedy; that is, they always have packets to transmit. We consider two groups of stations with different throughput requirements, $R_1 = 100$ Kbps and $R_2 = 200$ Kbps. They are competing to access to 7 channels as defined in IEEE 802.11p standard.

In this paper, we consider that the traffic generated by each communicating vehicle consists of periodic broadcast message (PBM) as it roughly accounts for 95% of the traffic in VANET. It is not time-critical but loss-sensitive service.

Since we use a discrete time simulator, the interarrival of vehicles sending PBM is Poissonly distributed with a mean value ranging between 0 and 120 s.

The value of traffic constant used in (5) is assumed to be $a = 0.25$ as per the traffic flow theory [1], and $L = 500$ m is the road segment length that is assumed to be equal to the transmission range of the 802.11p access point.

5.2. Simulation Results. The simulations were performed while using the Network Simulator tool NS2, to evaluate the effectiveness of the proposed resource allocation policy.

Figure 3 reports the impact of the proposed CAC scheme on the overall blocking probability. At a load of 8 Erlangs, for example, the blocking probability decreases from about 6% when the reference node is chosen by the AP (CAC_reference AP) to about 5.4% when it is chosen from the pool of vehicles in the transmission range of the current station.

When no CAC scheme is applied, the overall blocking probability at 8 Erlangs is about 6.7%.

Figure 4 exhibits the average normalized throughput, and since PBM is a loss-sensitive service, we report in Figure 5 the packet loss probability. These figures clearly show that the proposed CAC policy achieves a good performance and capacity gain.

We can intuitively conclude that, when the reference node is chosen by a vehicle from the pool of nodes in its transmission range rather than choosing it by the AP from the pool of all accepted nodes, CWmin will be smaller and then the channel access delay will be shorter for the corresponding service. This provides better chance to access the media for a given traffic condition.

In Figure 6 we compare the performance of the proposed CAC scheme while packets are queued according to FIFO scheduling policy or using PLCF function that adjusts the PBM packets priority level values. To show the effectiveness of PLCF, the simulations were performed by giving to all users

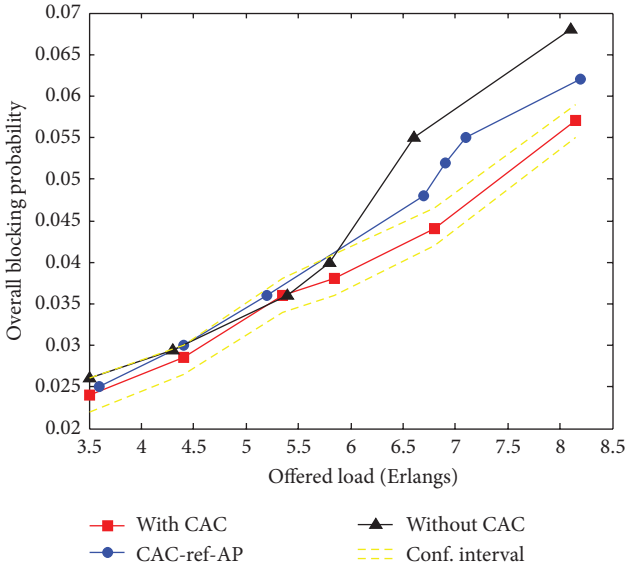


Figure 3: Overall blocking probability versus offered load.

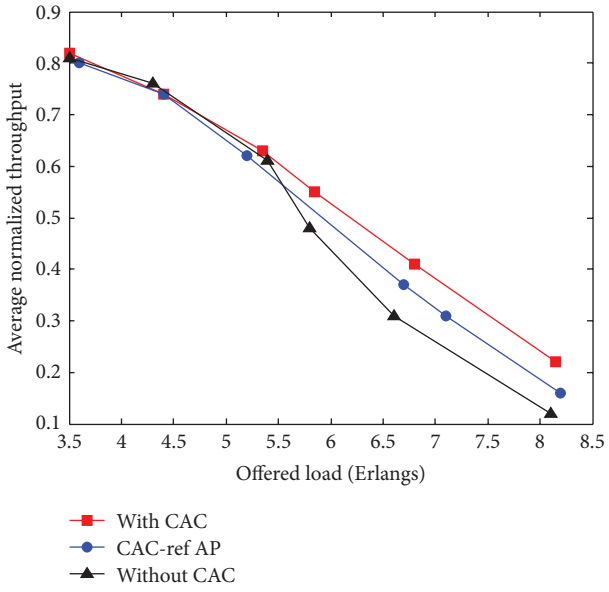


Figure 4: Average normalized throughput versus offered load.

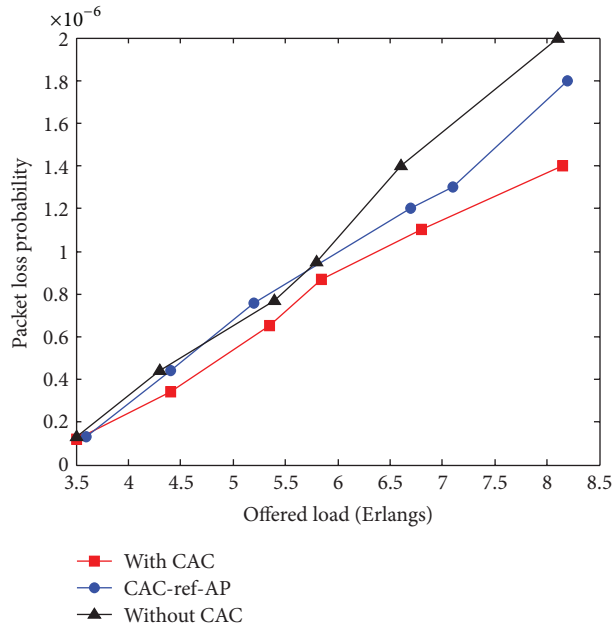
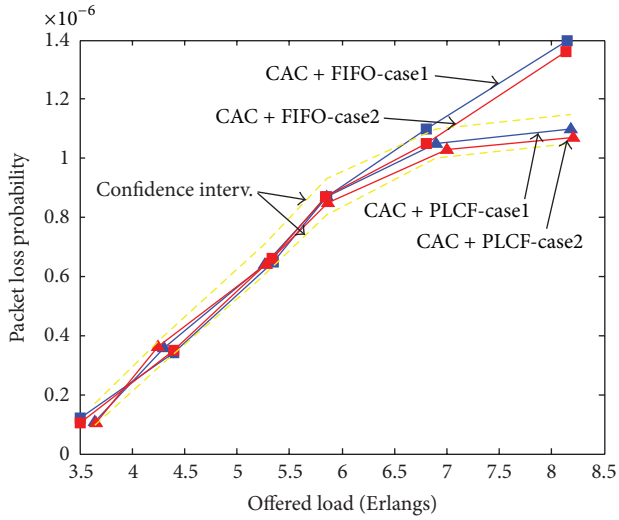


Figure 5: Packet loss probability versus Offered load.



Case 1:
$w_d = 0.1$
$w_{lq} = 0.6$
$w_p = 0.3$
$w_s = w_{wt} = 0$

Case 2:
$w_d = 0.3$
$w_{lq} = 0.4$
$w_p = 0.3$
$w_s = w_{wt} = 0$

Figure 6: Performance gain with PLCF function.

the same weights for the security parameter, and the waiting time factor since PBM is not a time-critical service. This is as to consider $w_s = w_{wt} = 0$. The three remaining parameters in the PLCF function are then the link quality, the roadside conditions, and the precedence class. PBM is not bandwidth hungry; that is, bandwidth could not be the most significant parameter, and thus, users of this type of service should be given low weights. In this work, simulations were performed considering two cases, where $w_d = 0.1$ and $w_d = 0.3$. The weight w_{lq} relative to the signal strength is equal to 0.6 and 0.4, respectively. The weight w_p for the precedence class is admitted to be constant and then equal to 0.3. The results show that, in case 2, the packet loss probability is almost equal or slightly lower than in case 1 at high traffic load. This is because stations using higher throughputs are served in first and the most of their packets will not wait beyond the limit of time.

By implementing the PLCF, the system manages to decrease the packet loss probability by up to 22%. For example, for loss-sensitive services like PBM using PLCF-based scheme, it shows a significant improvement in performance in VANETs. This improvement is made clear for high traffic load (from about 7 Erlangs).

The validation of results by analytical model is very difficult since the full details of a scheduling policy and a dynamic CAC function with users' mobility and propagation environment cannot be described by formulas usable in practice.

The simulation run takes duration of about ($10^6\times$ simulation time unit) to achieve a confidence level of 95%.

6. Concluding Remarks

In this paper we presented a cross-layered admission control algorithm based on throughput guarantee and dynamic adaptation of joint transmission range and contention window according to the vehicle density and network traffic conditions. A priority function based on significant vehicular network parameters to make the scheduling decision fairer is proposed as well.

Results have shown that our scheme proposed for the IEEE 802.11p vehicular networks improves the performance of the vehicular communication and achieved capacity gains.

Future work includes the performance analysis of the proposed scheme for each type service independently while taking into consideration more constrained movement. The EDCA scheduler will be substituted by other schedulers like those based on WRR and WFQ scheduling policies. A comparison investigation could be provided for all classes of services.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

References

[1] M. M. Artimy, W. Robertson, and W. J. Phillips, "Assignment of dynamic transmission range based on estimation of vehicle density," in *Proceedings of the 2nd ACM International Workshop on Vehicular Ad Hoc Networks (VANET '05)*, pp. 40–48, Cologne, Germany, September 2005.

[2] "Part 11: wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications amendment 6: wireless access in vehicular environments," IEEE 802. 11p published standard, IEEE, 2010.

[3] S. Eichler, "Performance evaluation of the IEEE 802.11p WAVE communication standard," in *Proceedings of the IEEE 66th Vehicular Technology Conference (VTC '07)*, pp. 2199–2203, Baltimore, Md, USA, October 2007.

[4] Standards Committee, *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements*, IEEE Computer Society, 2005.

[5] Y. Xiao and H. Li, "Evaluation of distributed admission control for the IEEE 802.11e EDCA," *IEEE Communications Magazine*, vol. 42, no. 9, pp. S20–S24, 2004.

[6] Y. Xiao and H. Li, "Voice and video transmissions with global data parameter control for the IEEE 802.11e enhance distributed channel access," *IEEE Transactions on Parallel and Distributed Systems*, vol. 15, no. 11, pp. 1041–1053, 2004.

[7] D. Gao, J. Cai, and K. N. Ngan, "Admission control in IEEE 802.11e wireless LANs," *IEEE Network*, vol. 19, no. 4, pp. 6–13, 2005.

[8] A. Veres, A. T. Campbell, M. Barry, and L.-H. Sun, "Supporting service differentiation in wireless packet networks using distributed control," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 2081–2093, 2001.

[9] D. Gu and J. Zhang, "A new measurement-based admission control method for IEEE 802.11 wireless local area networks," Tech. Rep. TR-2003 122, Mitsubishi Electric Research Laboratory, 2003.

[10] L. Zhang and S. Zeadally, "HARMONICA: enhanced QoS support with admission control for IEEE 802.11 contention-based access," in *Proceedings of the 10th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS '04)*, pp. 64–71, Toronto, Canada, May 2004.

[11] D. Pong and T. Moors, "Call admission control for IEEE 802.11 contention access mechanism," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '03)*, vol. 1, pp. 174–178, San Francisco, Calif, USA, December 2003.

[12] A. Banchs, X. Perez-Costa, and D. Qiao, "Providing throughput guarantees in IEEE 802.11e wireless LANs," in *Proceedings of the 18th International Teletraffic Congress (ITC-18)*, Berlin, Germany, September 2003.

[13] V. A. Siris and G. Stamatakis, "Optimal CWmin selection for achieving proportional fairness in multi-rate 802.11e WLANs: test-bed implementation and evaluation," in *Proceedings of the 1st ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation and Characterization (WiNTECH '06)*, pp. 41–48, Los Angeles, Calif, USA, September 2006.

[14] R. W. Rothery, "Car-following models," in *Traffic Flow Theory: A State of the Art Report*, N. H. Gartner, C. Messer, and A. K. Rathi, Eds., Monograph on Traffic Flow Theory, Oak Ridge National Laboratory, Oak Ridge, Tenn, USA, 1997.

[15] P. Santi and D. M. Blough, "An evaluation of connectivity in mobile wireless ad hoc networks," in *Proceedings of the International Conference on Dependable Systems and Networks (DNS '02)*, pp. 89–98, June 2002.

[16] M. Desai and D. Manjunath, "On the connectivity in finite ad hoc networks," *IEEE Communications Letters*, vol. 6, no. 10, pp. 437–439, 2002.

[17] Y. Cheng and T. G. Robertazzi, "Critical connectivity phenomena in multihop radio models," *IEEE Transactions on Communications*, vol. 37, no. 7, pp. 770–777, 1989.

[18] V. Milanes, J. Villagra, J. Godoy, J. Simo, J. Perez, and E. Onieva, "An intelligent V2I-based traffic management dystem," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 1, pp. 49–58, 2012.

[19] T. Bejaoui, "Vehicle density-based resource allocation policy for QoS provisioning in vehicular communication networks," in *Proceedings of the International Conference on Innovations in Information Technology (IIT '12)*, pp. 310–315, Abu Dhabi, UAE, March 2012.

[20] R. P. Roess, E. S. Prassas, and W. R. McShane, *Traffic Engineering*, Pearson Prentice Hall, Erewhon, NC, USA, 3rd edition, 2004.

[21] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.

[22] J. D. Parsons, *The Mobile Radio Propagation Channel*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2001.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 815720, 10 pages
http://dx.doi.org/10.1155/2014/815720

Research Article

Coverage Extension and Balancing the Transmitted Power of the Moving Relay Node at LTE-A Cellular Network

Jaafar A. Aldhaibani,[1,2] Abid Yahya,[1] and R. Badlishah Ahmad[1]

[1] *School of Computer and Communication Engineering, University Malaysia Perlis (UniMAP), Perlis, 01000 Kangar, Malaysia*
[2] *Department of Space Technology and Communication, Ministry of Science and Technology, Baghdad, Iraq*

Correspondence should be addressed to Jaafar A. Aldhaibani; jaffar_athab@yahoo.com

Received 21 August 2013; Accepted 27 October 2013; Published 29 January 2014

Academic Editors: G. Zhang and J. Zhou

The poor capacity at cell boundaries is not enough to meet the growing demand and stringent design which required high capacity and throughput irrespective of user's location in the cellular network. In this paper, we propose new schemes for an optimum fixed relay node (RN) placement in LTE-A cellular network to enhance throughput and coverage extension at cell edge region. The proposed approach mitigates interferences between all nodes and ensures optimum utilization with the optimization of transmitted power. Moreover, we proposed a new algorithm to balance the transmitted power of moving relay node (MR) over cell size and providing required SNR and throughput at the users inside vehicle along with reducing the transmitted power consumption by MR. The numerical analysis along with the simulation results indicates that an improvement in capacity for users is 40% increment at downlink transmission from cell capacity. Furthermore, the results revealed that there is saving nearly 75% from transmitted power in MR after using proposed balancing algorithm. ATDI simulator was used to verify the numerical results, which deals with real digital cartographic and standard formats for terrain.

1. Introduction

Long-term evolution-advanced (LTE-A) is the enhancing of the 3rd generation partnership project (3GPP) LTE, which improves LTE features in terms of coverage and throughput [1]. The relay is one of the major innovations of LTE-A, to meet growing demand for coverage extension, throughput, capacity enhancement, and saving the high deployment cost whether if deploying small size BS as solution to increase the coverage. The basic idea of relaying is that the relay received the signals from source and forwarded these signal after amplification to the destination node.

On relaying scenarios, there are two types of relaying architectures: fixed relay node (RN) and moving relay node (MR), where RNs are deployed near cell edge to increase the coverage and enhancing the throughput at the users in this region [2]. However, this improvement in coverage and throughput is based on the relay placement which provides fairness distribution of coverage within cell size as shown in Figure 1.

MR is the same kind of functionality as the RN but with the difference that they offer it while moving with the users. MR is new innovation to improve the throughput for vehicular users at LTE-A networks where it can be deployed flexibly to increase the throughput for passengers in buses or trains over rural area in cases where RNs are not available or not economically justifiable and the weak received signal from BSs [3].

MR is installed on vehicle and connected wirelessly with the BS via relay link (RL) and with passengers via access links (AL), so the MR and passenger are called group mobility [4] as shown in Figure 1. In fact, group mobility can be provided anywhere a large number of users are moving together during is using cellular network services.

The MR makes these services more reliable, with the assumption that the RL has a much better channel than regular UEs [5]. MR is connected to external power source via a battery charger or has its own power supply unit. This allows MRs to have a relatively high access to processing capabilities and to constant higher transmission powers.

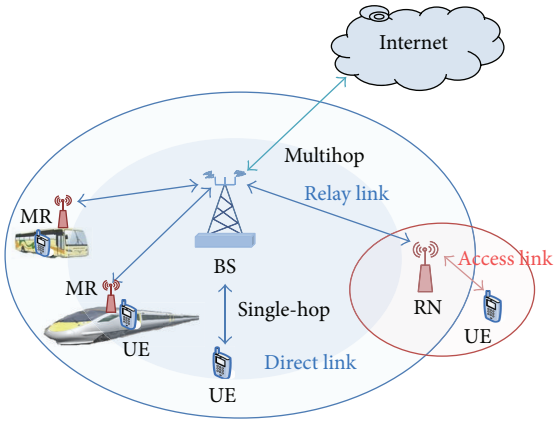


Figure 1: Relay nodes scenarios.

Using MRN in cellular systems is still under discussion in the 3GPP LTE [6]. Studies have shown that through deploying symmetrical and cooperative relays on top of trains, the quality-of-service (QoS) of a UE inside the vehicle can be significantly improved.

The main contributions of this paper are deriving the optimal relay node by considering the saturation throughput distance near the nodes locations, which is estimated from 200 to 500 m according to stations design and antenna configuration. This distance yields more accurate results in order to provide maximum achievable rate to users and increasing the number of active users at cell edge region. Furthermore the second contribution in this paper is proposing the balancing power algorithm which is reduced the transmitted power of moving relay within vehicle along with enhancing throughput for passengers.

2. System Model Description

Half-duplex mode is proposed in this work, where the relay cannot transmit and receive simultaneously. In general, while UE moves away from the cell-center, SINR degrades due to two factors. Firstly, the received signal strength goes down as the path loss increases with distance from the BS. Secondly, the intercell-interference rises because UE moves away from one BS and approach another BS.

At the cellular network without relaying with assuming that UE is connected to BS_i and moving away towards BS_j. The signal transmitted from BS_j appears as interference to the UE. The received signal at the downlink for each user k without relaying can be represented as the following equation:

$$Y_{i,k} = \sqrt{P_i} H_{i,k} X_{i,k} + \sum_{j=0}^{N_{\text{cell}}} \sqrt{P_j} H_{j,k} X_{j,k} + N_k, \quad (1)$$

where $j = 0 \rightarrow N_{\text{cell}}$, N_{cell} is the number of neighboring cell; P_i and P_j are the transmit power of donor BS and neighboring BSs, respectively; $H_{i,k}$ and $H_{j,k}$ are the fading channel gain for

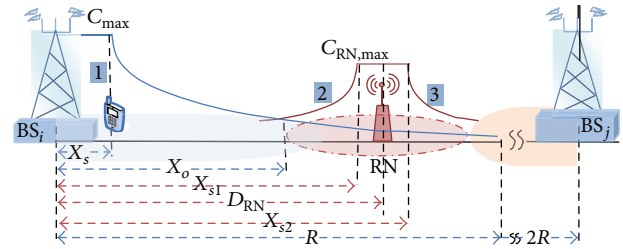


Figure 2: Scheme of proposed node locations.

donor and neighboring cell, respectively; and N_k is AWGN for user k [6, 7]:

$$\rho_{i,k} = \frac{P_i |H_{i,k}|^2}{N_k + \sum_{j=0}^{N_{\text{cell}}} P_j |H_{j,k}|^2}, \quad (2)$$

where $\rho_{i,k}$ is the SINR from the ith link in each single subcarrier (k) UEs.

In a severely limited interference scenario, the background noise N_k can be ignored to simplify the calculations and (2) can be written as

$$\rho_{i,k} = \frac{P_i G_{r,k} G_{t,i} (\lambda/4\pi)^2 D_{i,k}^{-\alpha}}{\sum_{j=1}^{N} P_j G_{r,k} G_{t,j} (\lambda/4\pi)^2 d_{j,k}^{-\alpha}},$$
$$\rho_{i,k} = \frac{P_i L D_{i,k}^{-\alpha}}{\sum_{j=1}^{N} P_j L_j d_{j,k}^{-\alpha}}. \quad (3)$$

The channel H is the function of path loss; therefore,

$$|H|^2 = L D^{-\alpha}, \quad (4)$$

where $L = G_r G_t (\lambda/4\pi)^2$ is constant depending on the infrastructure of sender and receiver, G_r, G_t is the antenna gains of the transmitter and receiver, respectively, $D_{i,k}$ and $d_{j,k}$ are the distances from user to donor BS_i and neighbour BS_j, respectively, and α is the path loss exponent [8].

2.1. Capacity of Cell without Relay. Using adaptive modulation and coding AMC is one of the basic enabling techniques in the standards for 3G wireless networks that have been developed to achieve high spectral efficiency on fading channels [9]. Typically the quality of the signal received by a UE depends on channel quality from BS, level of interference from neighboring cells, and noise level. For a given modulation, the code rate can be chosen depending on the radio link conditions. At the downlink data transmissions in LTE-A, the BS usually selects the code of modulation scheme according to the channel quality indicator (CQI) feedback transmitted by the UE in the uplink [10].

This work splits the down link capacity from BS into two regions according to modulation and coding scheme (MCS) to provide realistic transmission schemes, as shown in Figure 2.

Based on [11, 12] the capacity in a single-input single-output LTE system can be estimated by

$$C_i = \min\left\{C_{\max}, \mathrm{BW}_{\mathrm{eff}} \log_2\left(1 + \frac{\rho_i}{\rho_{\mathrm{eff}}}\right)\right\}, \quad (5)$$

where C_i is the estimated spectral efficiency in bps/Hz and $C_{\max}$ is the upper limit based on the hard spectral efficiency given by 64-quadrature amplitude modulation with the coding rate of 0.753 equal to 4.32 bps/Hz [10]. ρ_i is the SINR for each user in the cell, $\mathrm{BW}_{\mathrm{eff}}$ is the adjustment for the system bandwidth efficiency, and ρ_{eff} is the adjustment for the SINR implementation efficiency.

$(\mathrm{BW}_{\mathrm{eff}}, \rho_{\mathrm{eff}})$ has the value of $(0.56, 2.0)$ in the downlink and $(0.52, 2.34)$ in the uplink [13]. The proposed system used two regions in cell capacity distribution. The first region around the BS is known as the saturation throughput region which is specified from $0 \rightarrow X_s$, in which the level capacity is always steady based on the used modulation scheme, while the other region is determined from $X_s \rightarrow R$, where in this region, the cell capacity is never steady based on the Shannon theory, as shown in Figure 2. According to these concepts the system performance can be described as the following equations. The received signal at UE in location (1) can be written as

$$Y_{i,X_s} = \sqrt{P_i} H_{i,X_s} X_{i,X_s} + \sum_{j=0}^{N_{\mathrm{cell}}} \sqrt{P_j} H_{j,X_s} X_{j,X_s} + N_{X_s}. \quad (6)$$

The ideal SINR at X_s location is

$$\rho_{\mathrm{ideal}} = \frac{P_i L X_s^{-\alpha}}{P_j L_1 (2R - X_s)^{-\alpha}}, \quad (7)$$

where L_1 is constant depending on the infrastructure of neighbor BSs. The error vector magnitude (EVM) is a measure of the difference between the ideal symbols and the measured symbols after the equalization [14]. This difference is called the error vector magnitude. For 64QAM modulation in LTE-A the SINR (ρ_{i,X_s}) at X_s location is explained as [14, 15]

$$\frac{1}{\rho_{i,X_s}} = \frac{1}{1/\rho_{\max} + 1/\rho_{\mathrm{ideal}}}, \quad (8)$$

where

$$\rho_{\mathrm{ideal}} = \frac{P_i L X_s^{-\alpha}}{P_j L_1 (2R - X_s)^{-\alpha}},$$

$$\rho_{i,X_s} = \frac{\rho_{\mathrm{ideal}} \rho_{\max}}{\rho_{\mathrm{ideal}} + \rho_{\max}} = \frac{\rho_{\max} P_i L X_s^{-\alpha} / P_j L_1 (2R - X_s)}{\rho_{\max} + P_i L X_s^{-\alpha} / P_j L_1 (2R - X_s)^{-\alpha}}. \quad (9)$$

Through simple mathematical processes we get

$$X_s = \frac{2R}{1 + \left(P_i L / P_j L_1\right)^{-1/\alpha} \left(\rho_{i,X_s}\right)^{1/\alpha} - \left(\rho_{\max}\right)^{1/\alpha}}. \quad (10)$$

For downlink LTE network the $\rho_{\max} = 0.08$ with 64QAM [14, 16]. The distance X_s depends on the infrastructure of the sender and the effect of interference from other cells. In the baseline outdoor capacity analysis, this work considers a three-sector omnidirectional antenna, with each sector having 120-degree diversity to avoid the interference and applying frequency reuse scheme. If all BSs have the same characteristics in LTE networks, then the constant region distance from the BS is

$$X_s = \frac{2R}{1 + \left(\rho_{i,X_s}\right)^{1/\alpha} - \left(\rho_{\max}\right)^{1/\alpha}}, \quad (11)$$

$$C_{X_s} = \mathrm{BW}_{\mathrm{eff}} \log_2\left(1 + \frac{\rho_{i,X_s}}{\rho_{\mathrm{eff}}}\right). \quad (12)$$

From (5), logically, the total capacity over cell is equal or more than the $C_{\max}$ capacity:

$$\mathrm{BW}_{\mathrm{eff}} \log_2\left(1 + \frac{\rho_{i,X_s}}{\rho_{\mathrm{eff}}}\right) \geq C_{\max},$$

$$\rho_{i,X_s} \geq \rho_{\mathrm{eff}} \left(2^{(C_{\max}/\mathrm{BW}_{\mathrm{eff}})} - 1\right). \quad (13)$$

By substation in (11),

$$X_s \leq \frac{2R}{1 + \left(\rho_{\mathrm{eff}} \left(2^{(C_{\max}/\mathrm{BW}_{\mathrm{eff}})} - 1\right)\right)^{1/\alpha} - \left(\rho_{\max}\right)^{1/\alpha}}. \quad (14)$$

According to (14), it is easy to evaluate the capacity saturation distance, for example, X_s equals 313.5 m if $\alpha = 2.2$, $C_{\max} = 4.32$ bps/Hz, $\rho_{\max} = 0.08$, and 2500 m radius.

2.2. Handover Analysis. In multihop relaying the handover process becomes more important and a difficult importance within any cellular network. It is necessary to guarantee that it can be implemented reliably and without disruption to any wireless service. A fixed relay-based LTE-architecture that fits perfectly to improve the SINR at the cell boundary, thereby increasing capacity, can possibly increase the number of accepted users. However, the relay node and base station are located at a certain distance from each other, in which the SINR at UE from the relay link is equal to the SINR for direct link, where the user is in a handover case [17]. Therefore the distance from the BS to the said location is X_o which is a known handover distance as shown in Figure 3.

To evaluate the X_o, first evaluating the received signals from both BS and RN at UE in X_o location putting the distance from relay location (D_{RN}) to X_o is $D_{\mathrm{RN}} - X_o$. Thus, the received signal from the BS at the UE in X_o point can be expressed as

$$Y_{i,X_o} = \sqrt{P_i} H_{i,X_o} X_{i,X_o} + \sqrt{P_{\mathrm{RN}}} H_{\mathrm{RN},X_o} X_{\mathrm{RN},X_o} + N_{X_o}. \quad (15)$$

SINR at X_o through direct link is

$$\rho_{i,X_o} = \frac{P_i \left|H_{i,X_o}\right|^2}{P_{\mathrm{RN}} \left|H_{\mathrm{RN},X_o}\right|^2}, \quad (16)$$

$$\rho_{i,X_o} = \frac{P_i L X_o^{-\alpha}}{P_{\mathrm{RN}} L_r (D_{\mathrm{RN}} - X_o)^{-\alpha}}, \quad (17)$$

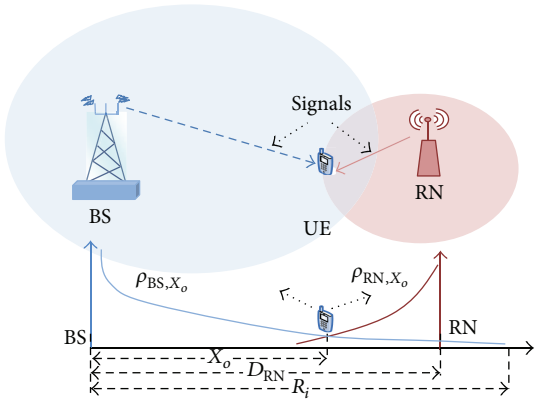


FIGURE 3: Hand over procedure.

where $L_r = G_r G_t (\lambda/4\pi)^2$ is the relay node characteristic and $G_r G_t$ are antenna gains of transmitter and receiver, respectively. The received signal from the relay link can be expressed as

$$Y_{RN,X_o} = \sqrt{P_{RN}} H_{RN,X_o} X_{RN,X_o} + \sqrt{P_i} H_{i,X_o} X_{i,X_o} + N_{X_o}, \quad (18)$$

$$\rho_{RN,X_o} = \frac{P_{RN} \left| H_{RN,X_o} \right|^2}{P_i \left| H_{i,X_o} \right|^2}, \quad (19)$$

$$\rho_{RN,X_o} = \frac{P_{RN} L_r (D_{RN} - X_o)^{-\alpha}}{P_i L X_o^{-\alpha}}. \quad (20)$$

At the X_o the received signal at UE from BS is equal to the signal at UE rather than RN. This location is known as the handover point, as shown in Figure 3. Using (17) and (20) to evaluate X_o through equal received SINR from both BS and RN within handover location is as follows:

$$\frac{P_i L X_o^{-\alpha}}{P_{RN} L_r (D_{RN} - X_o)^{-\alpha}} = \frac{P_{RN} L_r (D_{RN} - X_o)^{-\alpha}}{P_i L X_o^{-\alpha}},$$

$$\left(P_{BS} X_o^{-\alpha}\right)\left(P_{BS} X_o^{-\alpha}\right) = \left(P_{RN} (D_{RN} - X_o)^{-\alpha}\right)\left(P_{RN} (D_{RN} - X_o)^{-\alpha}\right), \quad (21)$$

$$P_{RN}^{-1/\alpha} L_r D_{RN} = X_o \left(L P_i^{-1/\alpha} + L_r P_{RN}^{-1/\alpha}\right),$$

$$X_o = \frac{D_{RN}}{\left((L_r P_{RN}/L P_i)^{1/\alpha} + 1\right)}.$$

Based on (21), X_o is a distance dependent on the relay location, the node characteristics, and path loss exponent.

The equations of relay coverage are calculated, which are determined by the locations of UE and presented in Figure 2. Therefore the received signals at UE in locations 2 and 3 can be represented as

$$Y_{RN,2} = \sqrt{P_{RN}} B_{RN,2} X_{RN,2} + \sqrt{P_i} B_{i,2} X_{i,2} + N_2,$$
$$Y_{RN,3} = \sqrt{P_{RN}} B_{RN,3} X_{RN,3} + \sqrt{P_i} B_{i,3} X_{i,3} + N_3. \quad (22)$$

The SINR for locations 2 and 3 in Figure 2 is

$$\rho_{RN,2} = \frac{L_r P_{RN} \left| B_{RN,2} \right|^2}{P_i L_i \left| B_{i,2} \right|^2}, \qquad \rho_{RN,3} = \frac{P_{RN} \left| B_{RN,3} \right|^2}{P_i \left| B_{i,3} \right|^2},$$

$$\rho_{RN,2} = \frac{P_{RN} L_r (D_{RN} - D_i)^{-\alpha}}{P_i L D_i^{-\alpha}}, \quad (23)$$

$$\rho_{RN,3} = \frac{P_{RN} L_r (D_i - D_{RN})^{-\alpha}}{P_i L D_i^{-\alpha}}.$$

Then the spectral efficiency of relaying system has been divided into four regions to explain the coverage distribution over cell edges according to RN location. Therefore the analytical expression of the rising and decline of spectral efficiency level, taking on the account of the saturation region, is explained with the following equations:

$$C_{RN,2} = \min \left\{ \mathrm{BW}_{\mathrm{eff}} \log_2 \left(1 + \frac{P_{RN} L_r (D_{RN} - D_i)^{-\alpha}}{P_i L \rho_{\mathrm{eff}} D_i^{-\alpha}}\right), C_{R\max} \right\},$$

$$C_{RN,3} = \min \left\{ C_{R\max}, \mathrm{BW}_{\mathrm{eff}} \log_2 \left(1 + \frac{P_{RN} L_r (D_i - D_{RN})^{-\alpha}}{P_i L \rho_{\mathrm{eff}} D_i^{-\alpha}}\right) \right\}. \quad (24)$$

$C_{R\max}$ is the maximum spectral efficiency of relay that depends on the transmitted power relay and antenna infrastructure. Where the domains of D_i are $X_o < D_i < X_{s1}$, $X_{s2} < D_i < R$ for $C_{RN,2}, C_{RN,3}$, respectively.

2.3. Optima Relay Location for LTE-A Cellular Networks. In this section, the issue of the optimum placement of the relay node deployment in a dual-hop network over LTE-A cellular networks will be addressed as well as the maximum throughput and limited interference between all in-band and out-band stations.

The optimum location of relay at the cell edge between two locations of the user when $D_i = X_o$ and $D_i = R$ is used as the relay capacity equations can be obtained through the following:

$$\frac{(D_{RN} - X_o)^{-\alpha}}{X_o^{-\alpha}} = \frac{(R - D_{RN})^{-\alpha}}{R^{-\alpha}},$$
$$X_o (R - D_{RN}) = R (D_{RN} - X_o). \quad (25)$$

Substituting (21) in (25), we obtain

$$D_{RN} = R \left(1 - \left(\frac{L_r P_{RN}}{L P_i}\right)^{1/\alpha}\right). \quad (26)$$

From (26) the relay location depends on the cell radius, the properties of the relay node and the BS, and α. This equation limits the relay location between handover point and the boundaries of cell.

3. Group Mobility Analysis

In this section, a mobility model is proposed, where all UEs in vehicle are moving across of cell at different velocities. According to the derived formulas of instantaneous SNR at direct and relay links in [1], the group mobility for MR and user has been derived in order to evaluate the system performance and reducing the transmitted power of MR. Therefore both UEs and MR are moving as group mobility across BS.

Typically, the instantaneous SNR changes according to environment of channel, such as the distance between the transmitter and receiver, and fading state of the channel.

The user in vehicle receives two signals: via direct link and from MR via relay link; therefore the combined SNR at UE is

$$\rho_{\mathrm{UE}}^{\mathrm{MRC}} = \rho_{\mathrm{UE}}^{\mathrm{DL}} + \rho_{\mathrm{UE}}^{\mathrm{MR}}, \tag{27}$$

where the $\rho_{\mathrm{UE}}^{\mathrm{DL}}$, $\rho_{\mathrm{UE}}^{\mathrm{MR}}$ is SNR at UE via direct and relay links, while $\rho_{\mathrm{UE}}^{\mathrm{MRC}}$ is the combination of SNR at UE via both the direct and relay links.

By inserting (4) the result is

$$\begin{aligned} \rho_{\mathrm{UE}}^{\mathrm{DL}} &= \frac{P_i L\left(d_{\mathrm{DL}}\right)^{-\alpha}}{N_o}, \\ \rho_{\mathrm{UE}}^{\mathrm{MR}} &= \frac{P_i P_{\mathrm{MR}} L\left(d_{\mathrm{MR}}\right)^{-\alpha} \left|H_{\mathrm{AL}}\right|^2}{\left[P_i L\left(d_{\mathrm{MR}}\right)^{-\alpha} + 2P_{\mathrm{MR}}\left|H_{\mathrm{AL}}\right|^2 + N_o\right] N_o}. \end{aligned} \tag{28}$$

d_{MR} is the distance between the BS and MR, P_{MR} is the transmitted power by MR, and H_{AL} is the channel gain of access link between the user and MR inside vehicle as shown in Figure 4.

Intuitively the distance is function of velocity and time so

$$\rho_{\mathrm{UE}}^{\mathrm{DL}} = \frac{P_i L\left(v_{\mathrm{UE}} T_{\mathrm{UE}}\right)^{-\alpha}}{N_o}, \tag{29}$$

$$\rho_{\mathrm{UE}}^{\mathrm{MR}} = \frac{P_i P_{\mathrm{MR}} L\left(v_{\mathrm{MR}} T_{\mathrm{MR}}\right)^{-\alpha} \left|H_{\mathrm{AL}}\right|^2}{\left[P_i L\left(v_{\mathrm{MR}} T_{\mathrm{MR}}\right)^{-\alpha} + 2P_{\mathrm{MR}}\left|H_{\mathrm{AL}}\right|^2 + N_o\right] N_o}, \tag{30}$$

$$C_{\mathrm{UE}}^{\mathrm{MRC}} = \mathrm{BW}_{\mathrm{eff}} \log_2\left(1 + \frac{\rho_{\mathrm{UE}}^{\mathrm{MRC}}}{\rho_{\mathrm{eff}}}\right). \tag{31}$$

3.1. Balancing Power Algorithm (BPA) of MR. The SNR at (29) and (30) depends on the transmitted power and path loss between both transmitter and receiver, so the proposed BPA is balanced and controlled on the transmitted power of MR over cell radius to achieve the required SNR and throughput at the users with mitigation of the consumption in transmitted relay power.

Typically the coverage distribution close to BS is better than boundaries and therefore does not require consumption additional power when the vehicle (i.e., train, bus, metro) passes near BS where there is a good SNR. The proposed BPA depends on this idea as explained in Figure 4.

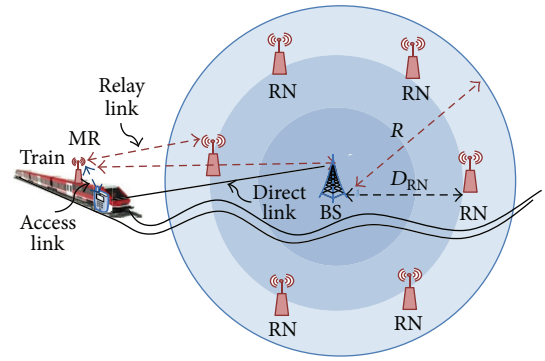


Figure 4: Vehicle travelling across a one cell size.

Two constraints are proposed in algorithm:

$$\begin{aligned} \text{minimize} \quad & P_{\mathrm{RN}} \\ \text{subject} \quad & 0 < P_{\mathrm{MR}} \le P_{\max} \\ & \rho_{\mathrm{th}} < \rho_{\mathrm{UE}}^{\mathrm{MR}} \le \rho_{\max} \\ & \rho_{\mathrm{UE}}^{\mathrm{DL}} \le \rho_{\mathrm{th}}, \end{aligned} \tag{32}$$

where ρ_{th}, $\rho_{\max}$ is the threshold and maximum required SNR at the UE.

Inputs for Algorithm 1 are that ρ_{th}, $\rho_{\max}$ is the threshold and maximum required SNR at the UE. $P_{\max}$, $P_{\min}$, is maximum and minimum level of power transmitted by MR. Q is the number of users at the vehicle, while V_{RN} is velocity of vehicle. $Q_{\mathrm{MR\text{-}UE}}$ and $Q_{\mathrm{BS\text{-}UE}}$ are the number of users which are attached with MR and BS, respectively. P_{MR} is the power transmitted by MR. $\rho_{\mathrm{UE},q}^{\mathrm{Direct}}$, ρ_{MR}, and $\rho_{\mathrm{UE},q}^{\mathrm{Access}}$ is the SNR at direct, MR, and access links, respectively.

The main body of balancing algorithm is described line by line as the following steps in Algorithm 1.

Line 1. Beginning of algorithm.

Line 2. Selection of the number of users which are attached with MR or BS.

Line 3. Calculating the SNRs at the direct, relay, and access links for each user.

Line 4. Comparison between the SNRs at direct and access links and then determination of the better link in order to enable it and disable the other.

Line 5. In case $\rho_{\mathrm{UE},q}^{\mathrm{Direct}}$ is better than the user attached directly with BS, the number of users that are attached with BS will increase by one.

Line 6. Enable the transmitted power of relay equal minimum chosen value.

Line 7–12. Algorithm proposed a second comparison between the SNR at relay link and required threshold SNR of system. If $\rho_{\mathrm{MR}} \ge \rho_{\mathrm{th}}$ the number of users that are attached with

Require:
V_{RN}, ρ_{th}, P_{max}, P_{min}, Q
Ensure:
P_{MR}
(1) BEGIN
(2) $Q_{MR\text{-}UE} = Q_{BS\text{-}UE} = 0$
(3) **for** $i = 1$ to Q **do**
Calculate $\rho_{UE,q}^{Direct}$, ρ_{MR}, and $\rho_{UE,q}^{Access}$
(4) **if** $\rho_{UE,q}^{Direct} \geq \rho_{UE,q}^{Access}$ **then**
(5) Increment $Q_{BS\text{-}UE}$ by 1
(6) $P_{MR} = P_{min}$
(7) **else**
(8) **if** $\rho_{MR} \geq \rho_{th}$ **then**
(9) Increment $Q_{MR\text{-}UE}$ by 1
(10) $P_{MR} = P_{min}$
(11) **else**
(12) $P_{MR} = P_{max}$
(13) **for** $j = 1$ to $(P_{MR}/0.1)$ **do**
(14) Calculate ρ_{MR}
(15) **if** $\rho_{MR} \geq \rho_{max}$ **then**
(16) Decrement P_{MR} by 0.1
(17) **else**
(18) Increment P_{MR} by 0.1
(19) **end if**
(20) **end for**
(21) **end if**
(21) **end if**
(22) **end for**
END

Algorithm 1: Balancing power algorithm of reducing and balancing of transmitted power consumption at MR.

MR will increase by one and enable the transmitted power of MR equal minimum chosen value. Otherwise enable the transmitted power of MR equal maximum chosen value.

Line 13-14. In this line there is counter to calculate the instantaneous value of ρ_{MR} at each user according to the distance between the MR and BS. Typically when the MR close to BS the ρ_{MR} is high, this link will degrade when MR is away from BS.

Line 15–18. Comparison between the ρ_{MR} and maximum chosen SNR in order to balance and save the transmitted power by MR and reducing the power consumption of MR. This step limits the transmitted power according to quality of received signal strength at the users in vehicle.

Line 19–22. Closed if and for statements.

4. Simulation Setup

The signal strength in the service area must be measured to design a more accurate coverage of modern LTE networks. The propagation of a radio wave is a complicated and less predictable process if the transmitter and receiver properties are considered in channel environment calculations. The process is governed by reflection, diffraction, and scattering; the intensities of which vary under different environments at different instances.

The ATDI simulator, used to approve the mathematical model for optimum relay placement. The propagation model for this simulator between the nodes can be expressed as the following equation:

$$P_r = P_t + G_t + G_r - L_{prop} - L_t - L_{re}\,[\mathrm{dB}], \tag{33}$$

where P_t indicates the power at the transmitter and P_r is the power at the receiver; G_t and G_r are the transmitter and receiver antenna gains, respectively; L_t and L_{re} express the feeder losses; and L_{prop} is the total propagation loss [18], formulated as

$$L_{prop} = L_{fsd} + L_d + L_{sp} + L_{gas} + L_{rain} + L_{clut}, \tag{34}$$

where L_{fsd} is the free space distance loss, L_d is the diffraction loss, L_{sp} is the sub path loss, L_{gas} is the attenuation caused by atmospheric gas, L_{rain} is the attenuation caused by hydrometeor scatter, and L_{clut} is the cutter attenuation.

This equation describes the link budget. A link budget describes the extent to which the transmitted signal weakens in the link before it is received by the receiver. The link budget depends on all the gains and losses in the path, which is facing the transmitted signal to reach the receiver. A link is created by three related communication entities:

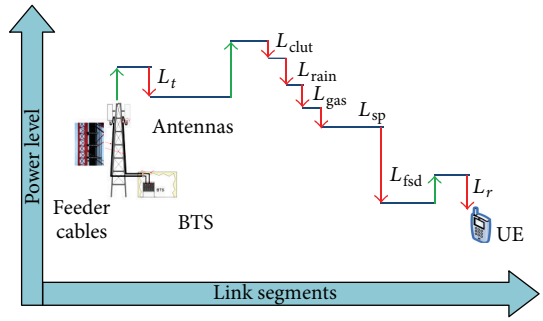


Figure 5: Link budget scheme.

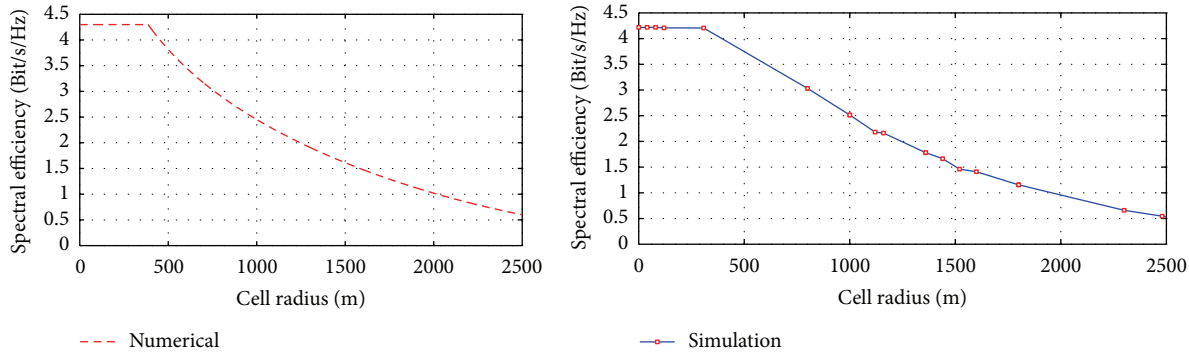


Figure 6: Spectral efficiency versus the cell radius: (a) numerical results and (b) ATDI simulation results.

transmitter, receiver, and a channel (medium) between them. The medium introduces losses caused by suction in the received power, as shown in Figure 5.

The SINR at the user equipment over the simulation test can be explained by using the following equation:

$$\mathrm{SINR}_{\mathrm{sim}} = \frac{P_r}{N_o + \sum_{j=1}^{j=N} P_{rj}}, \tag{35}$$

where the $\mathrm{SINR}_{\mathrm{sim}}$ is the received SINR by the user and calculated by the simulator; P_{rj} is the received signal from the neighbouring cell; and $j = \{1, \ldots, N\}$, where N is the number of neighbouring cells. For simplicity, we suggested the use of the first tier (six cells around the centralized cell) in planning for an urban area, with N_o as the background noise at the receiver:

$$\mathrm{SINR}_{\mathrm{sim}} = \frac{P_t G_t G_r / L_{\mathrm{prop}} L_t L_{\mathrm{re}}}{N_o + \sum_{j=1}^{j=N} P_j G_{t,j} G_r / L_{\mathrm{prop},j} L_j L_{\mathrm{re}}}, \tag{36}$$

where L_t, L_j, and L_{re} are the feeder loss for senders (central BS and the surrounding BS_j) and destination:

$$C_{\mathrm{sim}} = 0.5 \log_2 \left(1 + \mathrm{SINR}_{\mathrm{sim}}\right),$$

$$C_{\mathrm{sim}} = 0.5 \log_2 \left(1 + \frac{P_t G_t G_r / L_{\mathrm{prop}} L_t L_{\mathrm{re}}}{N_o + \sum_{j=1}^{j=N} P_j G_{t,j} G_r / L_{\mathrm{prop},j} L_j L_{\mathrm{re}}}\right). \tag{37}$$

5. Results and Discussion

In this section, the numerical results for the proposed mathematical model are explained and compared with results using the ATDI simulator, which uses a real digital cartographic representation of an urban area.

Figure 6(a) explains the downlink spectral efficiency versus cell radius without relaying scenario with consideration of the throughput saturation distance near station placement which is estimated here 306 m from BS as shown in Figure 6(a). Figure 6(b) shows the simulation analysis which considered the interferences for the first tier (six cells around the main cell) by using ATDI simulator.

These figures show the system performance degradation at the cell edge region when the user is away from BS.

Figure 7 displays the numerical and simulation curves are the results of the proposed mathematical model for optimal location (1660 m from the BS) by using (26). This model aims to improve the capacity and signal strength at the cell edge while mitigating interference between the stations. According to the proposed model the spectral efficiency at cell edge is improved from 0.6 bps/Hz to 1.45 bps/Hz for the proposed optimal location. Figure 7(a) describes the enhancement in spectral efficiency which is obtained by numerical analysis, while Figure 7(b) represents the simulation results by using ATDI simulator.

The difference between numerical and simulation results is due to the simulator since the ATDI simulator deals with

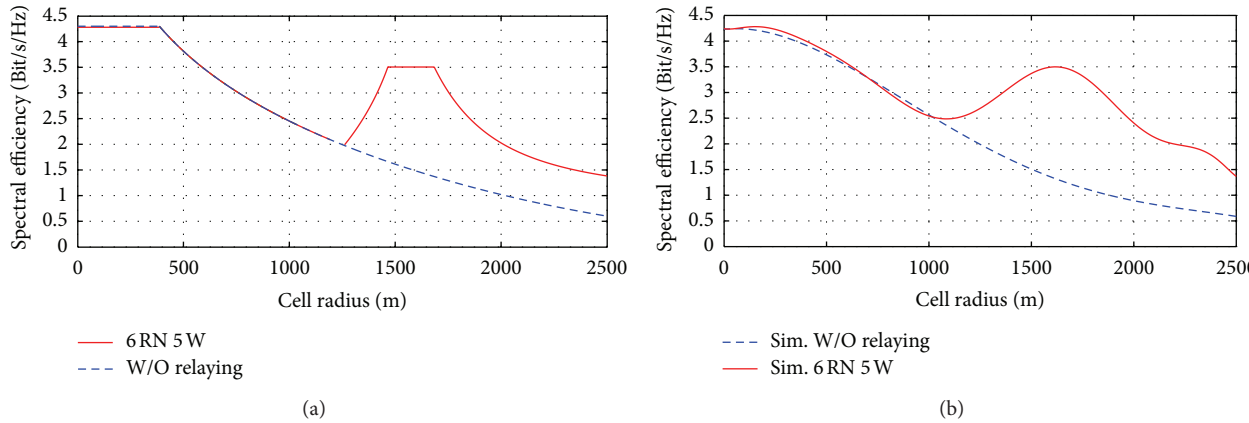


Figure 7: Spectral efficiency enhancement by deploying six RN at optimal location: (a) numerical and (b) simulation.

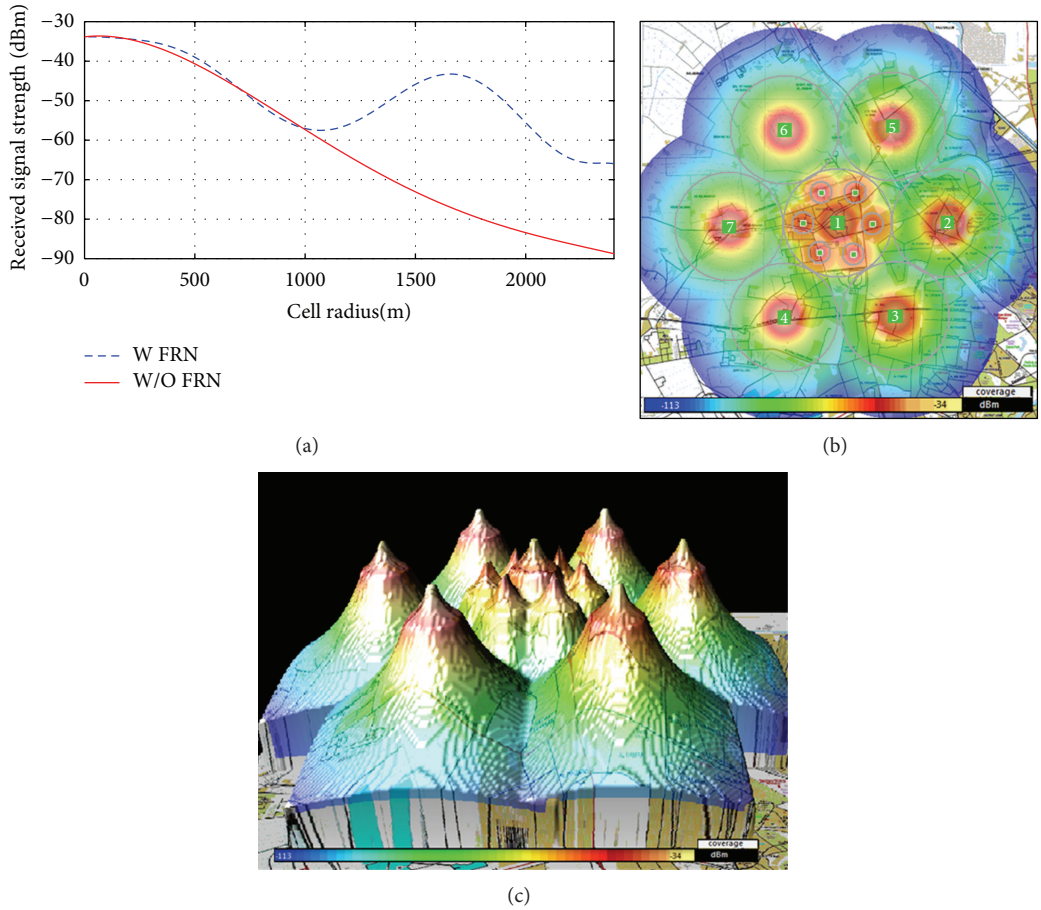


Figure 8: Received signal strength versus the distance with a real digital cartographic of an urban city for 6RN 5W, (a) received signal strength versus the distance, (b) 2-dimansion chromatic scheme of coverage area distribution, and (c) 3-dimansion chromatic scheme of coverage area distribution.

a real digital cartographic that contains several types of clutters for path loss conditions based on (34).

Figure 8 demonstrates the numerical and simulation results in received signal strength at the UE for proposed optimal location. Figure 8(a) shows the enhancement of the received signal strength after deploying six RN around BS in proposed location which is 40% for cell edge region according to simulated parameters as indicated in Table 1.

Figures 8(b) and 8(c) show chromatic scheme of coverage area distribution for 2-dimansion and 3-dimansion schemes, respectively, after deploying six relays each one 5 watts installed at optimal location (1660 m). These figures explain

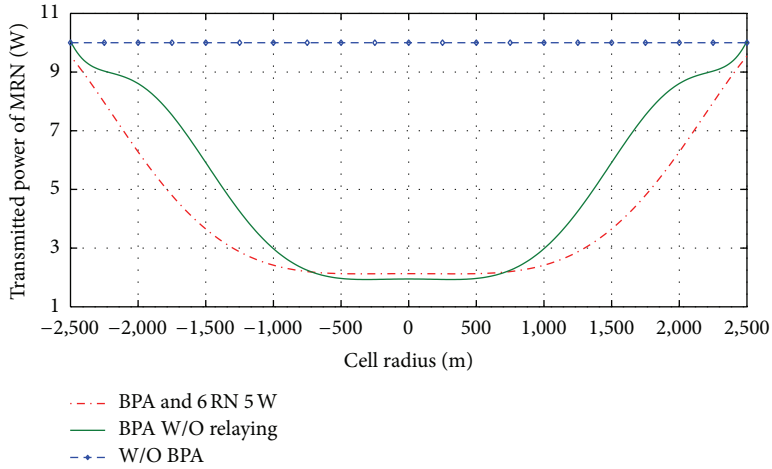


Figure 9: Reducing power consumption by using BPA.

Table 1: Simulation parameters.

Carrier frequency GHz	2
Bandwidth	1.4 MHz
Number of BS	7
Antenna height of BS	25 (m)
Antenna gain	17 dBi
Type of antenna	Omidirectional
Transmitted power of BS	40 W
Radius of cell	2500 m
Antenna height of RN	25 m
Antenna gain of RN	5 dBi
Transmitted power of RN	5 Watt
Number of UE	1
Antenna height of UE	1.5 m
Antenna gain	0 dBm
Coverage threshold	−30 dBm

the coverage extension along with mitigating the interference between them. It should be noted here that ATDI simulator is based on a real digital cartographic representation of an urban area that is approximately 176.7 km^2.

Figure 9 shows the reducing in transmitted power of MR together with providing the required throughput and SNR at each user inside vehicle. The saving in transmitted power of relay is approximately 60% from transmitted power in MR after proposed BPA and 75% saving in transmitted power with deploying six RN at proposed optimal relay location. BPA balanced the power along changing the distance between the vehicle and BS in order to insure maximum SNR at users and saving an extra power from MR as shown in Figure 9.

The number of active user is increased by installing MR node above the vehicle moreover the throughput for each user enhanced too compared with throughput at user in vehicle did not include MR as shown in Figure 9. This enhancement in throughput could be more with deploying six RN around BS at optimal proposed location as shown in Figure 9.

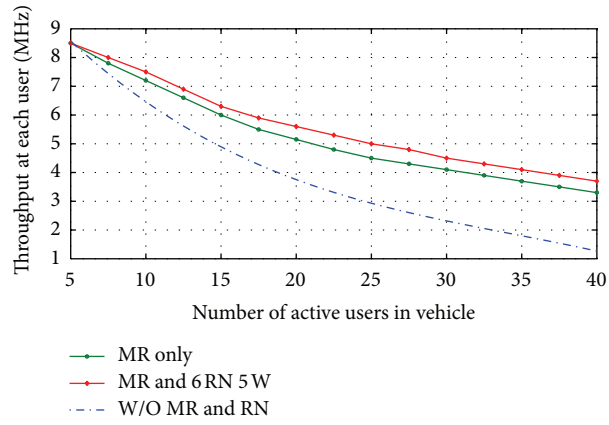


Figure 10: Throughput enhancement at users inside vehicle.

6. Conclusion

In this paper, we discussed two issues: the first is deriving the optimal relay placement in order to increase the capacity and coverage extension at LTE-A cellular network. This deriving of optimal RN placement was dependent on mathematical analysis and taken on account mitigating the interference among the nodes. The second is proposed BPA which saved the extra transmitted power of MR along with providing the required SNR and throughput at each user inside vehicle.

The numerical results together with simulation results indicated that there is 40% enhancement of capacity and received signal at users in cell edge region (Figure 10). Moreover there is 75% saving in transmitted power of MR achieved by proposing BPA with enhancing throughput for passengers.

Conflict of Interests

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

References

[1] J. A. Aldhaibani, A. Yahya, R. B. Ahmed, N. Omar, and Z. G. Ali, "Effect of relay location on two-way DF and AF relay for milt-users system in LTE-A cellular networks," in *Proceedings of the BEIAC IEEE Conference*, Langkawi, Malaysia, 2013.

[2] Y. Sui, J. Vihriala, A. Papadogiannis, M. Sternad, W. Yang, and T. Svensson, "Moving cells: a promising solution to boost performance for vehicular users," *IEEE Communications Magazine*, vol. 51, no. 6, 2013.

[3] O. Bulakci, "Multi-hop movingrelays for IMT-advanced and beyond," http://arxiv.org/abs/1202.0207.

[4] S. W. Peters and R. W. Heath, "The future of WiMAX: multihop relaying with IEEE 802.16j," *IEEE Communications Magazine*, vol. 47, pp. 104–111, 2009.

[5] Z. Ding, I. Krikidis, J. Thompson, and K. K. Leung, "Physical layer network coding and precoding for the two-way relay channel in cellular systems," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 696–712, 2011.

[6] C. Kosta, B. Hunt, A. U. Quddus, and R. Tafazolli, "A distributed method of inter-cell interference coordination (ICIC) based on dual decomposition for interference-limited cellular networks," *IEEE Communications Letters*, vol. 17, no. 6, pp. 1144–1147, 2013.

[7] H. Mei, J. Bigham, P. Jiang, and E. Bodanese, "Distributed dynamic frequency allocation in fractional frequency reused relay based cellular networks," *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1327–1336, 2013.

[8] R. Simon Saunders, *Antennas and Propagation for Wireless Communication Systems*, John Wiley & Sons, England, UK, 2nd edition, 2007.

[9] J. Yang, A. K. Khandani, and N. Tin, "Statistical decision making in adaptive modulation and coding for 3G wireless systems," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 6, pp. 2066–2073, 2005.

[10] K.-B. Song, A. Ekbal, S. T. Chung, and J. M. Cioffi, "Adaptive modulation and coding (AMC) for bit-interleaved coded OFDM (BIC-OFDM)," *IEEE Transactions on Wireless Communications*, vol. 5, no. 7, pp. 1685–1694, 2006.

[11] Y. Wang, W. Na, I. Kovács et al., "Fixed frequency reuse for LTE-advanced systems in local area scenarios," in *Proceedings of the IEEE 69th Vehicular Technology Conference (VTC '09)*, pp. 1–5, 2009.

[12] J. A. Aldhaibani, A. Yahya, R. B. Ahmad, A. S. Md Zain, M. K. Salman, and R. Edan, "On coverage analysis for LTE-A cellular networks," *International Journal of Engineering and Technology*, vol. 5, p. 6, 2013.

[13] Y. Wang, S. Kumar, L. Garcia et al., "Fixed frequency reuse for LTE-advanced systems in local area scenarios," in *Proceedings of the IEEE 69th Vehicular Technology Conference (VTC '09)*, pp. 1–5, Barcelona, Spain, April 2009.

[14] S. Sesia, B. Matthew, and T. Issam, *LTE, The UMTS Long Term Evolution: From Theory to Practice*, John Wiley & Sons, 2nd edition, 2011.

[15] Y. Wang, *System level analysis of LTE-advanced: with emphasis on multi-component carrier management [Ph.D. thesis]*, Department of Electronic Systems, The Faculty of Engineering and Science, Aalborg University, Aalborg, Denmark, 2010.

[16] 3GPP, "Evolved universal terrestrial radio access (E-UTRA), base station (BS) radio transmission and reception (release 9)," Tech. Spec. 36. 104 V9. 4. 0, Jun 2010.

[17] F. Khan, *LTE for 4G Mobile Broadband Air Interface Technologies and Performance*, Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[18] L. Korowajczuk, *LTE, WiMAX and WLAN Network Design, Optimization and Performance Analysis*, John Wiley & Sons, 2011.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 178937, 12 pages
http://dx.doi.org/10.1155/2014/178937

Research Article

Hybrid Network Defense Model Based on Fuzzy Evaluation

Ying-Chiang Cho and Jen-Yi Pan

Department of Electrical Engineering, National Chung Cheng University, Chiayi 62102, Taiwan

Correspondence should be addressed to Ying-Chiang Cho; silvergun@mail2000.com.tw

Received 29 October 2013; Accepted 25 November 2013; Published 16 January 2014

Academic Editors: W. Sun, G. Zhang, and J. Zhou

With sustained and rapid developments in the field of information technology, the issue of network security has become increasingly prominent. The theme of this study is network data security, with the test subject being a classified and sensitive network laboratory that belongs to the academic network. The analysis is based on the deficiencies and potential risks of the network's existing defense technology, characteristics of cyber attacks, and network security technologies. Subsequently, a distributed network security architecture using the technology of an intrusion prevention system is designed and implemented. In this paper, first, the overall design approach is presented. This design is used as the basis to establish a network defense model, an improvement over the traditional single-technology model that addresses the latter's inadequacies. Next, a distributed network security architecture is implemented, comprising a hybrid firewall, intrusion detection, virtual honeynet projects, and connectivity and interactivity between these three components. Finally, the proposed security system is tested. A statistical analysis of the test results verifies the feasibility and reliability of the proposed architecture. The findings of this study will potentially provide new ideas and stimuli for future designs of network security architecture.

1. Introduction

Network and information security is a critical link in a country's overall national security system. Once a war breaks out, the network will become a part of the battlefield. When a cyber war starts, no one—from governments to private enterprises—will be spared, similar to a war in real life [1, 2]. Cyber terrorism may not cause human casualties or fatalities, but the amount of damages that it can bring will definitely result in a wider scope. The disaster that it will cause is also likely to be a more devastating one [3, 4]. Hence, the security of networks and information systems is as important as military security. Along with the growing popularity of the Internet, the importance of network security has become increasingly prominent. Users' requirements and expectations of network security have also become more sophisticated, leading to the development and growth of network security technologies.

2. Network Defense Technologies

Firewalls can be used for segregating networks with different levels of security requirements. They also have the ability to control the interactions (if any) between various networks. A firewall is essentially a two-way mechanism of security management: it blocks external intrusions and limits outgoing communications by the internal host [5, 6]. In particular, it allows only certain designated information to pass through, irrespective of whether this information is external information trying to enter the intranet or that being transmitted by the intranet to the outside world. In either case, the process cannot proceed until the authentication procedures of the firewall have been carried out by the host. These procedures are carried out on the basis of the settings implemented in accordance with an organization's security rules and policies. Depending on their evaluation of the degree of risk that the network is facing, system administrators have the flexibility to adjust the security controls accordingly [7, 8].

An intrusion detection system (IDS) [9, 10] can be set up to facilitate a network's resistance against external attacks. The IDS collects information on its own, as well as from other systems and the Internet. It then matches the collected information against its database of attack signatures, analyzes potential attack scenarios, and takes the corresponding actions. The main functions of the IDS include the following:

(i) identification of the intruder(s); (ii) identification of the intrusion(s); (iii) monitoring and surveillance of security breaches; and (iv) provision of important information in a timely manner during the fight against an intrusion, so as to prevent its occurrence or escalation [11–13].

A honeynet works like an intelligence collection system, whereas a honeypot is deliberately set up as a target to lure hackers. After an intrusion, it will provide an analysis of the hacker's modus operandi, keep one abreast of the latest attack on the server, and highlight the existing network vulnerabilities. One of its other functions is to tap the communications between hackers in order to compile a list of the various tools that they employ [14–16]. A virtual computer network operates from the server of a virtual computer system that runs VMware or User-Mode Linux. A virtual system can comprise several virtual computers operating from a single host system. A virtual honeynet helps to reduce costs, machinery space, and the difficulty of managing a honeypot(s). Further, a virtual system typically allows the "freeze" and "resume" functions, which terminate operations when the computer is being threated; thus, one can analyze the mode of attack and take the necessary action(s) to deal with the situation [17, 18].

The core design concept for an intrusion prevention system (IPS) [19] revolves around immediate detection and active defense. In order to realize this concept, many technological breakthroughs have been achieved by IPS, leading to advantages that cannot be underestimated [20–22]. These can be elaborated as follows.

(i) In-line installation: The technologies and functions of instant detection under an IDS have been retained, with the additional feature of firewall-protected online installation (i.e., data are directly embedded in the network traffic). Incoming data packets from external systems are passed through a network port. After being checked to ensure the absence of unusual activities or suspicious contents, the data packets are then passed through another port before being sent to the intranet.

(ii) Real-time interdiction: An IPS has powerful, real-time blocking capabilities, allowing it to take preemptive action to block intrusions and intercept intrusive network traffic, thereby avoiding potential damages.

(iii) Advanced detection technology: All data packets that pass through an IPS must be preprocessed and restructured by the hardware to authenticate their specific application protocols [23, 24]. Next, the restructured data packets are screened by the IPS and matched against the identifying features and attack modes of various application protocols. Suspicious data packets are sent to a specialized signature library for comparison, thus improving the quality and efficiency of detection.

(iv) Built-in special rule: An IPS allows the implementation of special rules to block malicious programs from running. It can also assist in the implementation of an acceptable use policy (AUP), including the prohibition of the use of applications that consume large amounts of bandwidth, such as peer-to-peer file sharing and free Voice over Internet Protocol (VoIP) phone calls [22, 25].

(v) Self-study and self-adaptation abilities: An IPS must have the self-study and self-adaptation capabilities of artificial intelligence to respond to the myriad forms of attacks by hackers. It has to analyze and extract new attack signatures on the basis of the network's communication environment and intrusion statuses, before updating its signature library and formulating a new security and defense strategy [22, 26, 27].

3. Network Defense Model

The test subject for our study is a classified and sensitive network laboratory that belongs to the academic network. Its network security model was developed from scratch and gradually built into a complete system. The attempt was to establish a truly effective network defense method. The related procedure is as follows.

(i) Design a hybrid firewall module. This serves to segregate the intranet from the main gateway to the external network and imposes strict control over access to the intranet resources by external users.

(ii) Install an IDS at a critical node of the network (e.g., the server farms). The purpose of this step is to conduct real-time monitoring and detection of a variety of network activities and create appropriate records and issue early warnings when attacks occur.

(iii) Adopt the honeynet technology to link up the network's hybrid firewall, IDS, and virtual honeynet, and then enable the three components to interact with one another. This creates an early warning system for network security. The system administrators will be promptly alerted when there are intrusions or system vulnerabilities; thus, timely repairs and maintenance can be carried out.

The topology of the network defense model adopted in this study is shown in Figure 1.

4. Research Method and Theory

The module for screening data packets consists of the screening program (which can be operated at the gateway to the firewall) and two backend programs. This module functions as the security router between the network and the data link layers. It intercepts data packets that pass through the gateway of the firewall and verifies whether these contain application protocols that match the preinstalled security regulations, before deciding whether to forward, block, or discard the packets.

The firewall system uses the Netfilter/Iptables framework [28, 29] of the Linux environment. Appropriate modifications were made so that the system meets the requirements of connectivity and interactivity between the internal components of the computer. In this study, we used Iptables, packets

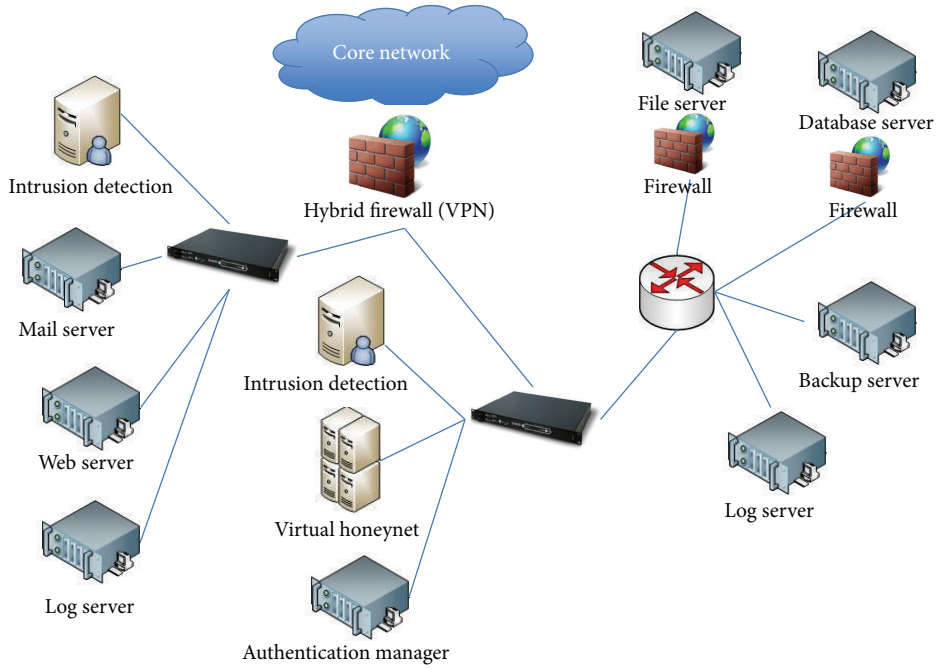


Figure 1: Network defense model.

selection tool based on the Netfilter framework, to develop a firewall subsystem with various functions. These Iptables include network address translation (dynamic NAT) during the screening of data packets, proxy servers, and others.

Further, the implementation of the firewall system under the Linux environment consists of two aspects. First, Netfilter provides a scalable and structured underlying framework, on top of which Iptables are implemented. The latter is a selection tool, responsible for the filtering and management of incoming and outgoing data packets. Second, Netfilter and Iptables jointly form the main Linux firewall system [30–32].

4.1. Interception and Filtering of Data Packets. The program flow for this process is shown in Algorithm 1.

4.2. Dynamic NAT. Its role is to conceal the Internet Protocol (IP) address of the internal host so that the security of system can be improved further. A Linux IP masquerading technology was used, making it necessary for the firewall to maintain a dynamic mapping table and update it all the time.

Files related to the masquerading function are included in Algorithm 2.

The standard headers include Algorithm 3.

Among the aforementioned files, the most important is ip_masq.c. It defines the interface of the application layer and the actual process of masquerading the IP address. The other files are extensions of the application.

4.3. Fuzzy Representation of the IDS Features. The fuzzy theory is suited for use in intrusion detection because it can easily combine input data from various sources. Since many types of intrusions cannot be clearly defined [33–35], the advance warnings that they trigger are usually vague too.

Fuzzy mathematics is used for describing, researching, and managing the mathematical relationships found in things with fuzzy characteristics. A comprehensive fuzzy evaluation is an important application of fuzzy mathematics. When the circumstances involve very complex factors, it can be used for selecting the best program for execution or making a choice after ranking the system detection results after the evaluation [36–38].

The main steps of the fuzzy evaluation method are as follows: (i) determine the factors and comments sets for evaluation, and then establish the fuzzy sets of the various factors (membership function); (ii) establish the fuzzy relationship between the evaluation factors and the comments, and then determine the weight that the respective factors have during evaluation; and (iii) derive a conclusion on the basis of calculations using a specific operand. Flexibility in the handling of attacks and the use of reasonable judgment are required for identifying a strict boundary between the normal and the abnormal [39, 40].

We have used the fuzzy sets technique in this study. The fuzzy sets of basic variables are represented by the following quintuple:

$$\text{Fuzzyset} ::= \langle \text{Object, Attribute, FC, Domain, ML} \rangle . \quad (1)$$

Here, Object refers to the item being described; Attribute, a particular property of the object; FC, the fuzzy concept; Domain, the location of the attribute; and ML, the membership list.

The procedure for conducting a fuzzy evaluation is as follows [41, 42].

```
Char * device;
/*declaring variable, name of the network interface being used for intercepting data packets*/
pcap_t*p;
/*declaring variable, control code of the intercepted data packets (the most important data structure)*/
Struct bpf_program fcode;
/* Berkeley Packet Filter (BPF) code structure involving the use of struct*/
Step 1. Locate the device that can intercept data packets;
device = pcap_lookupdev(errbuf);
Step 2. Create a control program for interception, and then prepare to intercept;
p = pcap_open_live (device, 8000, 1, 500, errbuf);
Step 3. If the user has set the screening criteria, proceed to compile and install the screening program;
pcap_compile(p, &fcode, filter_string, 0, netmask):
pcap_setfilter(p, &fcode):
Step 4. Enter (dead) loop, and then repeatedly intercept data packets by setting them to NULL;
for(;;)
{while
((ptr = (char*)(pcap_next(p,&hdr))) == NULL);
Step 5. Convert intercepted data to Ethernet data packet type;
eth = (struct libnet_ethernet_hdr*)ptr;
Step 6. Analyze Ethernet data packets, determine the type of data packets contained within,
and carry out further processing;
if(eth->ether_type == ntohs(ETHERTYPE_IP))
   ...
if(eth->ether_type == ntohs(ETHERTYPE_ARP))
   ...}
Step 7. Terminate the control program for interception. Increase the number of signal handlers
at program initialization so that the last iteration of this program can be executed prior to exiting the program.
pcap_close(p);
```

Algorithm 1

```
ip_masq.c          ip_masq_app.c
ip_masq_autofw.c   ip_masq_cuseeme.c
ip_masq_ftp.c      ip_masq_irc.c
ip_masq_mfw.c      ip_masq_mod.c
ip_masq_portfw.c   ip_masq_quake.c
ip_masq_raudio.c   ip_masq_user.c
```

Algorithm 2

```
#include ⟨net/ip_masq.h⟩
#include ⟨linux/ip_masq.h}
#ifdef CONFIG_IP_MASQUERADE_MOD
#include ⟨net/ip_masq_mod.h⟩
#endif
```

Algorithm 3

Step 1. Determine the factors and comments sets for evaluation, and then establish the fuzzy sets of the various factors. Internet access can be described using various characteristics such as the duration of the connection, communication volume, source and destination addresses, and types of service (i.e., the target port number). A compilation of these characteristics is known as the factors set. The evaluation vector is the bituple $E = \langle U, W \rangle$, where U denotes the factors set $U = \{u_1, u_2, \ldots, u_n\}$ and W represents the weight vector. Every component of W corresponds to the degree of importance of a factor during evaluation and can be represented as follows: $W = \int_u w/u$. Corresponding to the factors set is the comments set, which refers to the set of linguistic variables of the condition "degree of abnormality." The method of describing each factor is consistent. Therefore, the density distribution function of these factors can be treated as their membership function. During this step, the task is to calculate the density distribution function of each factor using the existing data.

Step 2. Evaluate the fuzzy relation between the factors and comments sets, and then determine the weight to be ascribed to the various factors during evaluation. This is the most important step in intrusion detection based on fuzzy evaluation. The detection model can be established once the fuzzy relation between the two sets has been determined. The fuzzy relation between the factors u_i and comments indicates the degree of membership that the respective factors have with the various degrees of abnormality. The determination of the fuzzy relation between the factors u_i and comments e_j is based on $f(u_i)$, which is the density distribution function of u_i. If the comments set is $\{e_1, e_2, \ldots, e_m\}$, then the density distribution function of u_i will be mapped onto m number of fuzzy relations. The relationship between the membership

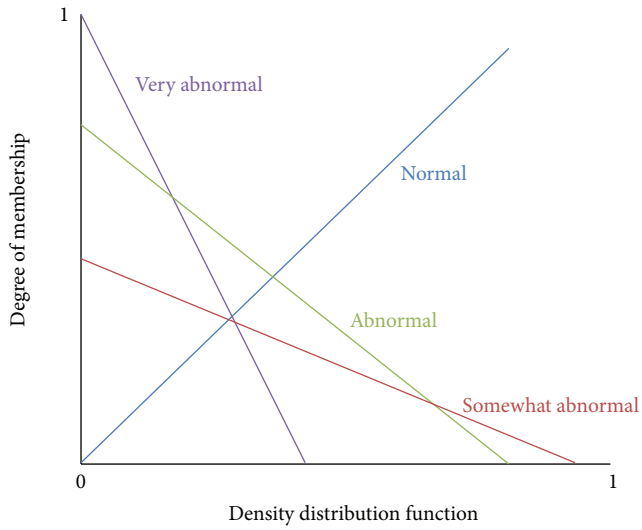


FIGURE 2: Membership function of the various comments.

functions of a fuzzy relation is shown in Figure 2. The following two characteristics of the fuzzy relation between the factors and the comments can be identified from Figure 2.

(i) The smaller the density of a particular eigenvalue is, the greater the degree of membership of the comment is to a higher degree of abnormality.

(ii) The higher the degree of abnormality of a comment is, the larger the membership function slope is.

In order to determine the weight of each factor, it is necessary to assess and rank the importance of all the factors. In this study, a judgment matrix established through the expert evaluation method (EEM) was used. The EEM is an important fuzzy mathematics tool used for creating fuzzy sets, fuzzy relations, and other mathematical models. It relies mainly on the experience of experts in the related fields. The sequence to establish a judgment matrix using EEM is as follows.

(i) Invite n number of experts to establish a comparative judgment matrix $A_1, A_2, \ldots, A_n$ for a particular type of intrusion, on the basis of their own experiences and the concept of fuzzy relations.

(ii) Set up a group of weights $W_1, W_2, \ldots, W_n$, $W_1 + W_2 + \cdots W_n = 1$ in accordance with the authority ranking of the experts, where W_i represents the authority ranking of expert number i, $i \in 1, 2, \ldots, n$.

(iii) Represent the final judgment matrix as $A = W_1 \times A_1 + \cdots + W_n \times A_n$.

Step 3. The conclusion from the evaluation and calculations carried out using a particular operand is derived as follows.

(i) Use the comments set to assess each eigenvalue that was determined by the aforementioned fuzzy relations, and then compose the evaluation matrix.

(ii) Carry out a compositional operation of the fuzzy matrix using the weight vector of the factors list and the evaluation matrix, thereby deriving a comprehensive evaluation vector.

(iii) Determine the comments for this particular set of eigenvalues on the basis of the principle of the maximum degree of membership.

4.4. Proactive and Early Security Warning Mechanism. The warning mechanism used in this study is a structurally implemented network based on a closed-ended virtual honeynet. Except for the managing platform, the virtual honeynet does not carry out any interactive data transmission with any external host or device. The closed-ended virtual honeynet comprises a virtual intruder, a virtual honeynet gateway, and two virtual honeypot systems. Its network topology is shown in Figure 3.

Within the virtual honeynet, the LANl, LAN2, VMnet0, and VMnetl switches are all virtual Layer 2 switching equipment. The LANl and LAN2 switches control the exchange of data between the virtual devices. The VMnet0 and VMnetl switches exercise similar controls but between the virtual devices and the host (managing platform). The VMnetl switch uses the host mode to ensure that the homed host (managing platform) can manage the honeynet gateway; that is, information from the homed host can be transferred to Interface Number 2 (eth2) of the honeynet gateway. On the other hand, the VMnet0 switch uses the bridging mode (normally not used in a closed-ended virtual honeynet) to ensure that the data pass directly through the physical network interface of the homed host (managing platform) to the real network.

There is a bridge between Interface Numbers 0 (eth0) and 1 (ethl); hence, these do not have any IP address. When data packets pass through the gateway, their time-to-live (TTL) values are not reduced. As such, the honeynet gateway is not visible to virtual intruders. A virtual honeynet system can provide system administrators with the ability to monitor, defend, and document the security of all segments of the network and can play a significant role in enhancing the security management of network systems.

Further, data control is also a very important concept. The main purpose of data control is to prevent intruders from using the honeynet as a springboard to send illegal information outside or attack other machines, after having obtained the administrative rights to the honeypot. Whenever intruders initiate any actions to scan, probe, or connect to the honeynet from the outside, these items must be captured. All scans, probes, and connections being made from the honeynet to the outside must also be strictly controlled and released subject to conditions. Data packets found with abnormalities must be blocked. At the same time, intruders must not realize that their behaviors are being monitored.

The main tool used for achieving the objective of data control at the honeynet gateway is the hybrid firewall, whose design was mentioned earlier. The firewall limits the frequency of outgoing connections. Since the data packets that must be controlled are being forwarded, the m-limit feature of Iptables can be used for implementing this function. The unit of time for counting purposes can also be specified, for

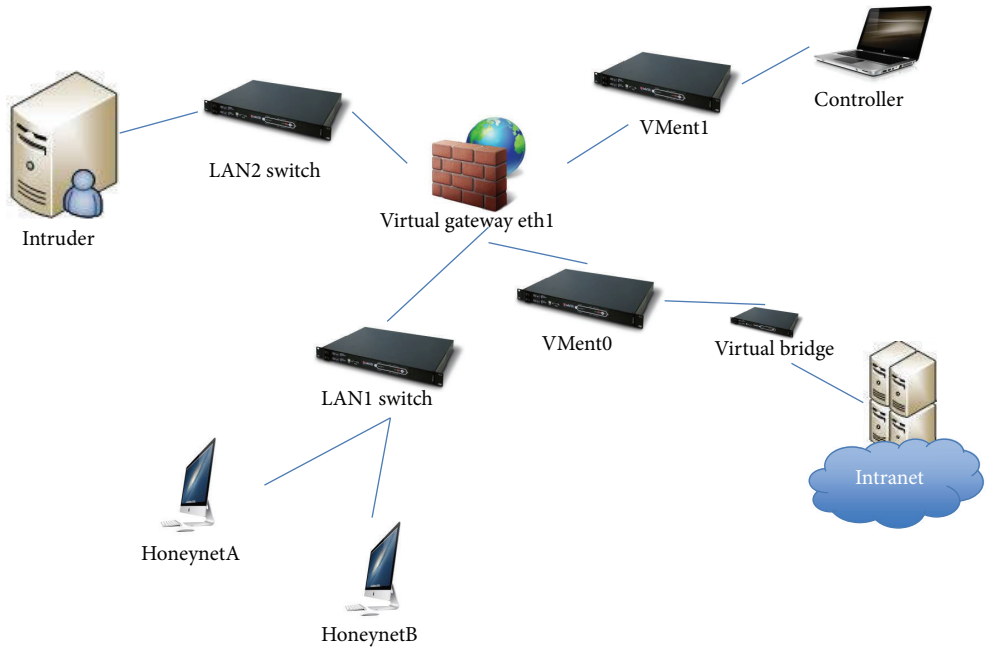


Figure 3: Membership function of the various comments.

example, by the seconds, minutes, or hours. If the number of outgoing data packets reaches the upper limit within a designated time period, the feature will record the relevant information and alert the system administrators accordingly. The data packets to be monitored include various types of information. In addition to the Transmission Control Protocol (TCP), User Datagram Protocol (UDP), and Internet Control Messaging Protocol (ICMP) packets, other data packets with unknown protocols must also be monitored. Further, specific rules can be set using the rc.firewall script. The actions to take against suspicious data packets include deletion, prohibition, or disposal.

4.5. Blocks Extensible Exchange Protocol (BEEP). We introduced the BEEP [43] concept and conducted a more in-depth study of its applications [44, 45] while implementing connectivity and interactivity between the firewall system, IDS, and the virtual honeynet system. BEEP concretizes the concept of messages into useful communication units. These messages, which form a part of the application protocol conversation, can be "The temperature of my CPU is 70°," or "This is a JPEG image." Since MIME is used as an envelope, these messages can be in any form or type. There is no actual limit on the size of the messages either. Rather, the message size is determined by the specific applications involved in the transmission. Therefore, BEEP is implemented by frames, communication units that are smaller than the messages. Although a message can often be reasonably sent out via one frame, it can also be split into multiple frames when necessary. A frame contains information that identifies the channel to which it belongs, the header of the message, and its sequential order within the entire message.

The existing implementation plans for the BEEP framework include Beepcore-c, Beepcore-j, and RoadRunner. Beepcore-c was adopted as the implementation method of BEEP in this study. The features of Beepcore-c are as follows.

(i) BEEP is implemented using C/C++.

(ii) It can achieve NULL/ECHO, NULLSINK, SASL/Anonymous, SASIJOTP, and TLS profile.

(iii) It has a hierarchical structure. The software structure is divided into four layers from top to bottom: Core, Wrapper, Profile Implementations, and Application.

Next, we used the OpenSSL library [46] to implement SSL and TLS. OpenSSL has the following characteristics.

(i) It is open source.

(ii) It supports SSL v2/v3 and TLS v1.

(iii) It is stable and supports the SSL and TLS properties completely.

After OpenSSL is fully compiled, two dynamic link library (DLL) files are generated: libeay32.dll and ssleay32.dll. The TLS profile of Beepcore-c contains the commands to call out these dynamic link library files.

Next is the implementation of the BEEP protocol over TCP. RFC3081 defines the implementation method to map BEEP onto the TCP protocol. Even though TCP provides the flow control for each session, a BEEP session may contain multiple simultaneous BEEP channels. As such, BEEP must provide a solution to avoid deadlocks. Hence, when BEEP is introduced to the TCP flow management mechanism, each channel's window size must be dynamic. Window sizes are exchanged between peers through SEQ.

The next issue to address is that of SMIP configuration. Under the pcore-c framework, profile realization is based on the DLL format. DLL calls out the pro init function, which is used by the SMIP Listener and Initiator to complete profile registration. During the registration process, in order to take into account the PROFILE_REGISTRATION structure, URI, Initiator_modes, listener_modes, the other values must be initialized and be sent to the corresponding callback function at the same time (Wrapper will call back these functions in due course).

The SMIP profile carries out its functions and roles via two aspects: Listener and Initiator. The former implements the monitoring function. When it has completed profile registration and receives a peer connection request from BEEP, Wrapper first uses pro_connection_init to complete profile initialization. Next, it uses the pro_session_init function to complete the initialization of the BEEP conversation. Thereafter, bpc start response is used for completing the establishment of the BEEP channel. SMIP will then go into the monitoring mode. On the other hand, after Initiator completes initialization using pro_connection_init and pro_session_init, it uses bp-start request to complete the connection request. At this time, profile negotiation between the peers is completed and MSG is sent to exchange SMIP-greetings. After the completion of profile negotiation, a data exchange can be carried out between the peers through the SMIP profile. Depending on the options in the configuration file, the SMIP profile can complete user authentication by using either TLS, SASL/Anonymous, or SASI/OTP profile. TLS then completes the encryption process for EMEF message transmission.

Now, we will discuss the issue regarding IMEF data format. In the program, a Clmefmessage class is used for completing the packaging of the IMEF message. Within the ClmefMessage class, an IMEF message structure is defined and used for storing the original unformatted data of the IMEF message. The IMEF message structure is converted by the CIamefMessage::MakeXMLImefMessageQ member function into an XML format that complies with the LMEF XML DTD definition to facilitate the connecting and interacting functions.

The managing components for connectivity and interactivity include the event engine and analysis modules, as well as the strategic module for connectivity and interactivity. The details are as follows.

(i) The event engine module is supported by the BEEP protocol. It is used for facilitating communication between the firewall and other security systems, as well as for receiving event information sent by various types of security systems. It then adds the received event information to the predefined events queue.

(ii) The event analysis module extracts events from the events queue and matches these against the library of strategies for connectivity and interactivity. Depending on the category of security event identified, it then selects the appropriate response strategy.

(iii) The strategic module for connectivity and interactivity is mainly used for effectively responding to the various intrusion behaviors existing in the actual network environment. This module can be used for the configuration and management of the entire library of strategies, thus ensuring that connective and interactive responses are carried out smoothly.

We would like to illustrate the process with a typical case of network intrusion (Figure 4). After the IDS detects an intrusion in the network or host by a hacker, it communicates with the firewall and the virtual honeynet interface through the BEEP protocol to lure the network intruder into the trap host. It then analyzes the characteristics of the intruder and notifies the firewall to generate dynamic rules to control and block the intrusion. Depending on the nature of the intrusion and the level of risk, the firewall can produce many different dynamic rules, each with a specific time period of effectiveness. This facilitates a variety of controlled operations, including early warning, termination of current conversation, and blocking of all connections from a particular source. After the firewall has implemented the measures, it reports the results to the system administrators and generates a log file.

5. System Implementation Test

Out of safety considerations, we only carried out tests targeting the laboratory's internal setup. Some common hacking tools were used for simulating an intrusion of the entire model. Two detection engines were installed prior to system intrusion. One was a network engine that received data packets from the intranet, and the other was a host engine that received data packets from the database server. Both engines collected the intrusions that each had respectively analyzed and then sent the information to the system console located in the intranet.

The operating system for the system console was Windows 2008 Server. The firewall and the virtual honeynet system used Linux, and the database used MySQL. The two machines that initiated the intrusion were located outside the network. The IP address of Attacking Hosts 1 and 2 was 122.116.76.141 and 122.116.76.147, respectively. The targets of the attack were Hosts 1 and 2 and the database server of the 192.168.0.0 subnet. The IP address of Hosts 1 and 2 was 192.168.0.9 and 192.168.0.11, respectively. The IP address of the database server was 192.168.1.10 (Windows 2008 Server was the operating system installed).

5.1. Tests for DoS Attacks. The log information of the hybrid firewall indicated that the attacker initiated the following connection:

> 2013-03-06, 09:48:05, 192.168.1.10, 2, Blocking host 122.116.76.147 completely for 600 seconds
>
> 2013-03-06, 09:48:05, -, l, iptables, Info: Blocking ip 122.116.76.147.

The data captured by Snort indicated that this was a typical DoS test attack.The records in the system log files were

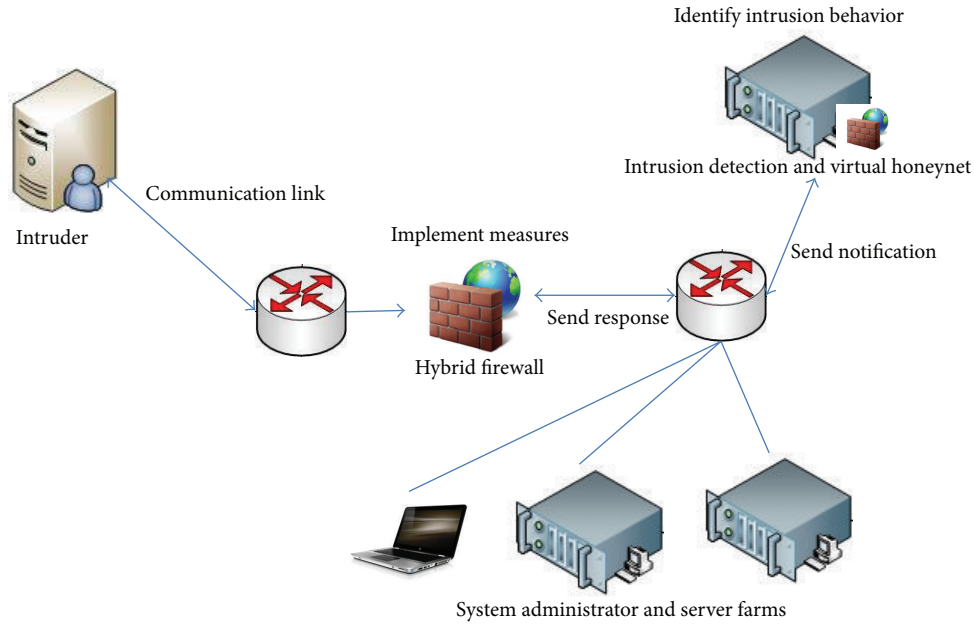


Figure 4: Schematic representation of a network intrusion.

as follows:

2013-03-06 09:48 snort[1852]: [1 : 474 : 1] Dos [Classification: Attempted Information Leak] [Priority: 2]:: Dos 192.168.0.10

2013-03-16 09:48 in.rlogind[1316]: connect from 192 .168.0.10

2013-03-16 09:48 inetd[413]: pid 1318: exit status 1

2013-03-16 09:48 in.rshd[1318]: connect from 192 .168.0.10

2013-03-16 09:48 in.fingerd[1315]: connect from 192 .168.0.10

2013-03-16 09:48 in.telnetd[1313]: connect from 192 .168.0.10

2013-03-16 09:48 rshd[1318]: Connection from 192 .168.0.10 on illegal port

2013-03-16 09:48 telnetd[1313]: ttloop: peer died: EOF

2013-03-16 09:48 inetd[413]: pid 1316: exit status 1

2013-03-16 09:48 inetd[413]: pid 1313: exit status 1

2013-03-16 09:48 sendmail[1314]: NOQUEUE: Null connection from [192.192.0.10]

2013-03-16 09:48 in.fingerd[1319]: connect from 192 .168.0.10

2013-03-16 09:48 in.telnetd[1320]: connect from 192 .168.0.10.

In response, measures were taken by the firewall to block the attack. The records in the system log were as follows:

2013-03-16, 09:48 : 10, -, l, iptables, Info: Command/ sbin/iptables -I FORWARD -i eth0 -s 122.116.76.147 -j REJECT Executed Successfully

The list of rules generated by the firewall included the following:

Chain INPUT (policy ACCEPT)

target port opt source destination

Chain FORWARD (policy ACCEPT)

target port opt source destination

REJECT all –122.116.76.147 anywhere reject-with-icmp-port-unreachable

Chain OUTPUT (policy ACCEPT)

target port opt source destination.

The above records showed that the DoS attack launched by the IP address 122.116.76.147 was promptly arrested by the firewall. The measure adopted in response was to block this IP address.

5.2. Tests for Different Types of Intrusion. The types of network intrusion used for the experiment are shown in Table 1.

Set the features set as F, where $F = \{f_1, f_2, \ldots, f_n\}$, and the comments set as E, where $E = \{e_1, e_2, \ldots, e_m\}$. Assess the individual factors to obtain the evaluation vector $R_i = (r_{ij})$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, followed by the evaluation matrix $R^T = [R_1, R_2, \ldots, R_n]$. The fuzzy comprehensive evaluation matrix (with the weight of the features being set

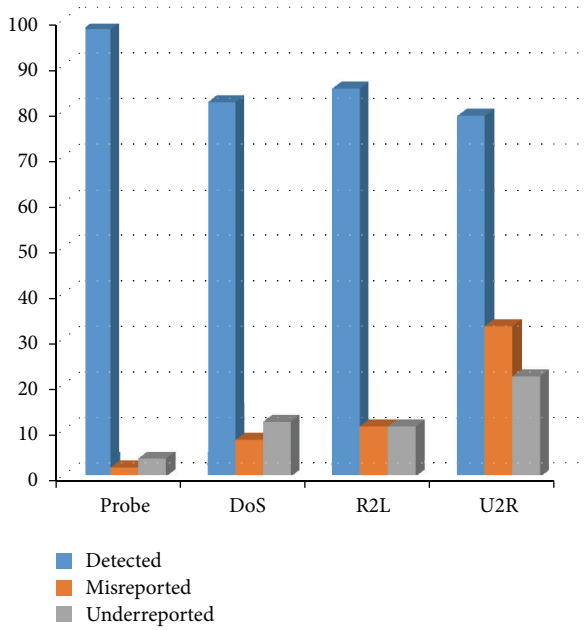


Figure 5: System performance indicators.

to $W_E = (W_{f_1}, W_{f_2}, \ldots, W_{f_n})$ can then be obtained via fuzzy composition as follows:

$$S = (S_1, S_2, \ldots, S_n) = \left(W_{f_1}, W_{f_2}, \ldots, W_{f_n}\right) \times \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix}. \quad (2)$$

Set $S_{K_0} = \max(S_j)$, $i = 1, 2, \ldots, n$, so that this particular factors set belongs to level K_0. The intrusion event is a subset of the comments set. The comments set created for the experiment comprises four comments, namely, "Normal," "Somewhat abnormal," "Abnormal," and "Very abnormal".

All comments assessed as "Abnormal" and above are classified as intrusions. At the start of the experiment, the tcp dump within the gateway was activated to collect the network data. The outputs of this process were multiple records of network communication. These records were divided into four groups. Group 1 was the baseline, which contained network data that did not relate to intrusion activities. In this group, 80% of the data were treated as training data, while the remaining 20% were used for testing the misreported rate. The data for Groups 2–4 were contained in network1, network2, and network3, respectively. Each group was subjected to three different types of attacks. After processing, the four groups of data used for testing the detection rates were stored in the database. The names of the network connection tables were "Normal," "Intrusion1," "Intrusion2," and "Intrusion3," respectively.

During the experiment, network connections were divided into three categories: (i) outgoing network connections from the local network; (ii) incoming network connections from the extranet; and (iii) connections within the local area network (LAN). The results of the experiment are shown in Table 2.

Table 1: Types of intrusion.

Category	Type	Activities
0	Normal	Normal
1	Probe	Probes on system vulnerabilities, for example, port scans
2	DoS (Denial-of-Service)	DoS attacks, for example, SYN flooding
3	R2L (Remote-to-Local)	Unauthorized access by remote machine, for example, password guessing
4	U2R (User-to-Root)	Unauthorized access by locally managed accounts, for example, buffer overflow attacks

Selective data were used for testing. The statistics related to the intrusion detection subsystem are shown in Tables 3 and 4. The former is based on the information collected and the latter on the intrusion type.

The system performance indicators of the network security early warning system were derived through further calculations and analysis. The data are shown in Table 5, and the graphical representation is presented in Figure 5.

Computed using the above data, successful detection rate was 87.15%. The requirements that we set for the experiment were met to a certain extent. The firewall system and IDS were stable in their operations and fully functional. The analysis of the two sets of experimental data indicated that the security protection of the system was further enhanced through the connectivity and interactivity between the firewall, IDS, and virtual honeynet. As long as the proposed network security architecture study initiated a connection, the system responded in a timely manner, irrespective of the mode of attack. The connection was also recorded and documented in the system log. This provided a comparative analysis to the system administrators and enabled them to take appropriate measures promptly.

With the above notwithstanding, the experiment reflected a number of shortcomings:

(i) although the IPSec protocol of the firewall could protect the security of the data packets, it reduced their transmission speeds;

(ii) there was room for further strengthening of the system's self-adaptability.

6. Conclusion

In this study, we used an actual implementation to validate the performance of the designed network defense model. The experimental data indicated that the entire system had an average successful detection rate of 87.15%, which met the

Table 2: Intrusion ratio of normal and abnormal data.

Network connection table	Ratio of network connections evaluated as intrusions		
	Outgoing connections	Incoming connections	Connections within the LAN
Normal	0.62	0.23	0.88
Intrusion 1	3.09	15.44	18.96
Intrusion 2	3.81	12.83	12.42
Intrusion 3	2.41	20.03	9.51

Table 3: Statistics based on information collected.

Data	Number of cases			
	Intrusion	Misreported	Underreported	Detection rate (%)
Dataset 1	94	13	8	92%
Dataset 2	90	8	10	89%
Dataset 3	88	7	8	91%
Dataset 4	96	12	5	95%

Table 4: Statistics based on intrusion type.

Type	Number of cases			
	Detected	Misreported	Under-reported	Total
Probe	226	4	7	233
DoS	131	14	23	155
R2L	77	12	12	89
U2R	29	12	8	36

Table 5: System performance indicators.

Performance indicator %	Type of attack				
	Probe	DoS	R2L	U2R	Mean
Detected	96.99	84.51	86.52	80.56	87.15
Misreported	1.71	9.03	13.48	33.33	14.39
Under-reported	3.01	14.84	13.48	22.22	12.36

design requirements. However, a certain margin of error still existed. Further, the firewall system, IDS, and virtual honeynet system had stable operations, were fully functional, and could fulfill the design requirements.

The contributions of this study could be summarized as follows.

(i) The concept of a network defense model was proposed shortly after a systems analysis of a sensitive and classified network was carried out. This systems analysis was an important prerequisite of and the basis for systems design. In this study, a model for an intelligent early warning system was designed. Based on the systems methodology and a combination of the theories of network security and the principles of automatic control, the proposed model was self-adaptive and could respond to network security issues in a dynamic manner.

(ii) IPS technology was used for establishing a distributed network security architecture comprising the following components.

(a) Hybrid firewall: A hybrid firewall system was designed on the basis of packet filtering and proxy and VPN technologies.

(b) IDS: Snort tools were used for creating a network intrusion detection system that used a fuzzy comprehensive evaluation to determine the intrusion detection eigenvalues.

(c) Virtual honeynet system: The approach of a virtual network trap was adopted, together with the implementation of a close-ended virtual honeynet, to give a proactive and early security warning to the network.

(d) Connectivity and interactivity were established between the firewall, intrusion detection, and virtual honeynet, which further proved the practicality and usability of the system.

Distributed network security architecture [47–49] can effectively prevent network intrusions and provide direct protection for key data. It plays an important supporting role in the construction of a network security system. Not only can it be applied to the development of a academic network, but it can also be used for constructing and improving the networks of private corporations and governmental network. The results of this study can provide new ideas and solutions, as well as serve as a reference for future network security topology design and related studies.

7. Future Research Direction

Given the multiple topics under the umbrella of network security, the following four aspects in this study were focused on: firewall, intrusion detection, virtual honeynet, and connectivity and interactivity between these three components. After a long period of research and actual implementation and on the basis of an extensive analysis, we could successfully establish and verify the entire architecture. However, because of the limitations of various factors, including knowledge, team experience, time, and individuals' abilities, we could not completely resolve all the issues related to the network security architecture.

The issues that require further research include the following:

(i) intrusion detection technologies: to expand the scope of research by using this study as the basis and to

continuously improve and refine data detection technologies in order to improve data detection efficiency,

(ii) distributed security communications: to strengthen related research such that data transfers between security components can be safer, more effective, and immediate,

(iii) data security: to carry out further research on encryption and transmission,

(iv) information on early warnings, storage of such information, and the operating mode of the information database: To conduct appropriate research in the related areas.

Conflict of Interests

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

References

[1] D. P. Fidler, "Was Stuxnet an act of war? Decoding a cyberattack," *IEEE Security and Privacy*, vol. 9, no. 4, pp. 56–59, 2011.

[2] M. Taddeo, "An analysis for a just cyber warfare," in *Proceedings of the 4th IEEE International Conference on Cyber Conflict (CYCON '12)*, pp. 1–10, 2012.

[3] R. A. Clarke and K. Robert, *Cyber War: The Next Threat to National SecurIty and What to Do about It*, HarperCollins, 2010.

[4] J. A. Lewis, *Assessing the Risks of Cyber Terrorism, Cyber War and Other Cyber Threats*, Center for Strategic and International Studies, 2002.

[5] J. Moscola, J. Lockwood, R. P. Loui, and M. Pachos, "Implementation of a content-scanning module for an Internet firewall," in *Proceedings of the 11th Annual IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM '03)*, pp. 31–38, 2003.

[6] P. Gupta and N. McKeown, "Algorithms for packet classification," *IEEE Network*, vol. 15, no. 2, pp. 24–32, 2001.

[7] J. Moscola, J. Lockwood, R. P. Loui, and M. Pachos, "Implementation of a content-scanning module for an Internet firewall," in *Proceedings of the 11th Annual IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM '03)*, pp. 31–38, 2003.

[8] A. Razzaq, A. Hur, S. Shahbaz, M. Masood, and H. F. Ahmad, "Critical analysis on web application firewall solutions," in *Proceedings of the IEEE 11th International Symposium on Autonomous Decentralized Systems (ISADS '13)*, pp. 1–6.

[9] C. H. Rowland, "Intrusion detection system," U.S. Patent No. 6, 405, 318. 11 Jun. 2002.

[10] J. F. Maddox and M. Arlington, "Intrusion detection system," U.S. Patent No. 4, 772, 875. 20 Sep. 1988.

[11] B. Mukherjee, L. T. Heberlein, and K. N. Levitt, "Network intrusion detection," *IEEE Network*, vol. 8, no. 3, pp. 26–41, 1994.

[12] R. P. Lippmann, D. J. Fried, I. Graf et al., "Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection evaluation," in *Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX '00)*, vol. 2, pp. 12–26, 2000.

[13] J. McHugh, A. Christie, and J. Allen, "Defending yourself: the role of intrusion detection systems," *IEEE Software*, vol. 17, no. 5, pp. 42–51, 2000.

[14] L. Spitzner, "The honeynet project: trapping the hackers," *IEEE Security and Privacy*, vol. 1, no. 2, pp. 15–23, 2003.

[15] C. Hecker and B. Hay, "Automated honeynet deployment for dynamic network environment," in *Proceedings of the 46th Hawaii International Conference on System Sciences (HICSS '13)*, pp. 4880–4889, 2013.

[16] N. Krawetz, "Anti-honeypot technology," *IEEE Security and Privacy*, vol. 2, no. 1, pp. 76–79, 2004.

[17] A. Lanoy and G. W. Romney, "A virtual honey net as a teaching resource," in *Proceedings of the 7th International Conference on Information Technology Based Higher Education and Training (ITHET '06)*, pp. 666–669, July 2006.

[18] F. H. Abbasi and R. J. Harris, "Experiences with a generation III virtual honeynet," in *Proceedings of the Australasian Telecommunication Networks and Applications Conference (ATNAC '09)*, pp. 1–6, November 2009.

[19] X. Zhang, C. Li, and W. Zheng, "Intrusion prevention system design," in *Proceedings of the IEEE 4th International Conference on Computer and Information Technology (CIT '04)*, pp. 386–390, September 2004.

[20] N. Ye, S. M. Emran, Q. Chen, and S. Vilbert, "Multivariate statistical analysis of audit trails for host-based intrusion detection," *IEEE Transactions on Computers*, vol. 51, no. 7, pp. 810–820, 2002.

[21] L. Tan and T. Sherwood, "A high throughput string matching architecture for intrusion detection and prevention," in *Proceedings of the 32nd Interntional Symposium on Computer Architecture (ISCA '05)*, pp. 112–122, June 2005.

[22] D. Stiawan, A. H. Abdullah, and M. Y. Idris, "The trends of intrusion prevention system network," in *Proceedings of the 2nd International Conference on Education Technology and Computer (ICETC '10)*, vol. 4, pp. V4217–V4221, June 2010.

[23] R. Janakiraman, M. Waldvogel, and Z. Qi, "Indra: a peer-to-peer approach to network intrusion detection and prevention," in *Proceedings of the 12th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE '03)*, pp. 226–231, 2003.

[24] K. Xinidis, I. Charitakis, S. Antonatos, K. G. Anagnostakis, and E. P. Markatos, "An active splitter architecture for intrusion detection and prevention," *IEEE Transactions on Dependable and Secure Computing*, vol. 3, no. 1, pp. 31–44, 2006.

[25] S. Vasanthi and S. Chandrasekar, "A study on network intrusion detection and prevention system current status and challenging issues," in *Proceedings of the 3rd International Conference on Advances in Recent Technologies in Communication and Computing*, pp. 181–183, IET Conference Publications, 2011.

[26] F. Colace, M. De Santo, and S. Ferrandino, "A slow intelligent approach for the improvement of intrusion detection and prevention system," in *Proceedings of the 6th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS '12)*, pp. 130–137, IEEE Conference Publications, 2012.

[27] N. Wattanapongsakorn, S. Srakaew, E. Wonghirunsombat et al., "A practical network-based intrusion detection and prevention system," in *Proceedings of the IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom '12)*, pp. 209–214, 2012.

[28] X. Chun-Tao, D. Xue-Hui, C. Li-Feng, and C. Hua-Cheng, "An algorithm of detecting and defending CC attack in real time," in *Proceedings of the International Conference on Industrial Control and Electronics Engineering (ICICEE '12)*, pp. 1804–1806, 2012.

[29] S. Chen, J. Xu, R. K. Iyer, and K. Whisnant, "Evaluating the security threat of firewall data corruption caused by instruction transient errors," in *Proceedings of the International Conference on Dependable Systems and Networks (DNS '02)*, pp. 495–504, June 2002.

[30] S. Mirzaie, A. K. Elyato, and M. A. Sarram, "Preventing of SYN flood attack with iptables firewall," in *Proceedings of the 2nd International Conference on Communication Software and Networks (ICCSN '10)*, pp. 532–535, February 2010.

[31] H. Salehi, H. Shirazi, and R. A. Moghadam, "Increasing overall network security by integrating signature-based NIDS with packet filtering firewall," in *Proceedings of the 1st IITA International Joint Conference on Artificial Intelligence (JCAI '09)*, pp. 357–362, April 2009.

[32] M. Marchi, R. Penzo, and A. Provetti, "Policy-based parametric firewall configuration: a real-case application," in *Proceedings of the 8th IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY '07)*, p. 276, June 2007.

[33] S. M. Bridges and M. U. Rayford, "Fuzzy data mining and genetic algorithms pplied to intrusion detection," in *Proceedings of the 23rd National Information Systems Security Conference*, National Institute of Standards and Technology, 2000.

[34] M. Botha and R. von Solms, "Utilising fuzzy logic and trend analysis for effective intrusion detection," *Computers and Security*, vol. 22, no. 5, pp. 423–434, 2003.

[35] L. Zhang and Q. Wang, "A network security evaluation method based on FUZZY and RST," in *Proceedings of the 2nd International Conference on Education Technology and Computer (ICETC '10)*, vol. 2, pp. V240–V244, June 2010.

[36] L. Rui and Y. Yi, "Network security assessment based on fuzzy sets and rough sets," in *Proceedings of the 5th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '09)*, pp. 1–4, September 2009.

[37] J. E. Dickerson, J. Juslin, O. Koukousoula, and J. A. Dickerson, "Fuzzy intrusion detection," in *Proceedings of the 9th IFSA World Congress and 20th NAFIPS International Conference*, vol. 3, pp. 1506–1510, July 2001.

[38] G. Florez, S. M. Bridges, and R. B. Vaughn, "An improved algorithm for fuzzy data mining for intrusion detection," in *Proceedings of the Annual Meeting of the North American on Fuzzy Information Processing Society (NAFIPS '02)*, pp. 457–462, 2002.

[39] H. Jin, J. Sun, H. Chen, and Z. Han, "A fuzzy data mining based intrusion detection model," in *Proceedings of the 10th IEEE International Workshop on Future Trends of Distributed Computing Systems (FTDCS '04)*, pp. 191–197, May 2004.

[40] A. Jadhav, A. Jadhav, P. Jadhav, and P. Kulkarni, "A novel approach for the design of network intrusion detection system (NIDS)," in *Proceedings of the International Conference on Sensor Network Security Technology and Privacy Communication System (SNS & PCS '13)*, pp. 22–27, 2013.

[41] G. J. Klir and Y. Bo, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, NJ, USA, 1995.

[42] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[43] The Blocks Extensible Exchange Protocol Core, http://tools.ietf.org/html/rfc3080.

[44] Using the NETCONF Protocol over the Blocks Extensible Exchange Protocol (BEEP), http://tools.ietf.org/html/rfc4744.

[45] The Intrusion Detection Exchange Protocol (IDXP), http://tools.ietf.org/html/rfc4767.

[46] J. Viega, M. Matt, and C. Pravir, *Network Security with OpenSSL: Cryptography for Secure Communications*, O'Reilly, 2009.

[47] E. Frécon and M. Stenius, "DIVE: a scaleable network architecture for distributed virtual environments," *Distributed Systems Engineering*, vol. 5, no. 3, pp. 91–100, 1998.

[48] O. Bidaud, "Distributed network architecture security system," U.S. Patent Application 10/118, 632.

[49] C.-F. Yu, "Distributed security procedure for intelligent networks," U.S. Patent No. 4, 919, 545. 24 Apr. 1990.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 349065, 8 pages
http://dx.doi.org/10.1155/2014/349065

Research Article

Social Network Supported Process Recommender System

Yanming Ye,[1,2] Jianwei Yin,[1] and Yueshen Xu[1]

[1] *College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*
[2] *College of Information Engineering, Hangzhou Dianzi University, Hangzhou 310018, China*

Correspondence should be addressed to Yanming Ye; yanmingye_com@sina.com

Received 22 October 2013; Accepted 25 November 2013; Published 15 January 2014

Academic Editors: W. Sun and J. Zhou

Process recommendation technologies have gained more and more attention in the field of intelligent business process modeling to assist the process modeling. However, most of the existing technologies only use the process structure analysis and do not take the social features of processes into account, while the process modeling is complex and comprehensive in most situations. This paper studies the feasibility of social network research technologies on process recommendation and builds a social network system of processes based on the features similarities. Then, three process matching degree measurements are presented and the system implementation is discussed subsequently. Finally, experimental evaluations and future works are introduced.

1. Introduction

Workflow technology has gained further application and development with the rapid growth of modern business environment. And process models are being widely used in the development of organizational structures [1], information systems [2], service-oriented architectures [3], and web services [4]. However, business process modeling is either complex or time-consuming, which often involves in selecting concrete activities to be performed, determining their execution order, dealing with the exceptions that may occur, and so forth. Besides, in modern commerce, both frequent changes of custom demands and the specialization of the business process necessitate the ability of modeling business processes in an effective and efficient way for enterprises. Thus, many business intelligence (BI) based techniques have been adopted to improve the business process modeling work, such as process mining and process retrieval [5–7]. Both process mining and process retrieval are complex or need more manual works. For improving the efficiency, some process modeling technologies so-called business process recommendation are proposed recently. However, most of the existing process recommendation technologies are only based on the static structure analysis of the process and the other properties of a process, such as the performer behaviors and the editor intensions, are not considered. In practice, when creating a new process, a modeler is inclined to refer to the modeling conduct of the familiar users or the processes with the same modeling intension or the process fragments that are used more frequently by certain users. This paper presents a social network supported method that can recommend the processes or process parts using the social features of process modeling, such as intension, activity performers, usage frequency, and modeling history.

This paper is organized as follows. After this introduction and the related works in Section 2, Section 3 gives some definitions with related basic instructions, and at the end of this section how to calculate process matching degree is highlighted. In Section 4, we discuss the implementation of the social network based process recommender system. In Section 5, some experiments on the system are discussed. Finally, Section 6 is devoted to research perspective.

2. Related Work

A process model is often in form of some graphical notations and describes how a certain process is composed out of different tasks, in which resources are involved in carrying out these tasks and objects are manipulated [8, 9]. Therefore, most of the existing process mining methods recur to graph minings, especially graphic structure mining. As for the

process recommendation technologies in process modeling, most of the methods are described as how to find certain process fragments as a subgraph that is familiar with the given process part. In [10], a workflow recommendation technique called FlowRecommender was proposed, which was announced that it can leverage provenance of workflows to provide recommendation for the best node that needs to be chosen to complete the workflow. Paper [11] used a NMSF (near neighbor and maximal subgraph first) process matching method to recommend the next nodes for the reference process fragment.

Social network has been widely researched in the last decade and various social network systems have been built, some of which also produce great commercial success. Social network analysis (SNA) has become an important branch of scientific studies. The related studies cover a wide range of different fields, including sociology, psychology, economy, and computer science [12]. SNA can effectively mine the implied information in social networks. Therefore, if it can use the SNA methods, the process modeling will be more effective. At first, we must construct a social network from a process repository. As far as we know, only a very small quantity of researches [12, 13] are directly related. In [12], three types of social networks can be generated from a process repository and a recommendation system, including (1) social network from process models, (2) social network from user history, and (3) social network from insertion history. Unfortunately, the three social networks are constructed for only having the network structure but more social features of the network cannot be used in the system. For example, the social network (1) is described to be more like an overlap of a processes graph, of which the nodes refer to the performers in the process models and the arcs show the transfer of work between the performers. In the social network (2), the authors build a matrix of users and processes in which each cell contains the number of uses of a process model by a user then calculate the distance between any two users by several measures, such as *Minkowski distance, Hamming distance,* and *Pearson's correlation coefficient*. By setting a certain threshold value, unimportant arcs could be removed and the left directly connected users are seen in different cliques, respectively. But these distance measures employ Euclidean distance or do not consider for the process correlation. Moreover, there are no other social features that can be used in this social network (2). The social network (3) is built on the relationship of process creators and is more suitable for change propagation process than process modeling.

From the discussion above, the greatest contribution of paper [12] is to show that the modeling support system can be extended with capabilities to take social information into account. But the three social networks that paper [12] proposed are insufficient in the SNA. This paper presents a method to build a unique social network that involves in more social features.

3. Preliminaries

Business process is a series of activities that are performed by different performers to specific targets, respectively, and the order of activities represents collaboration of these performers. In workflow systems, the process reflects the actual business process, and the activity node represents business operation in enterprise. Information or operations will flow or be conducted in turn according to the nodes sequence (or by the arcs). Therefore, the business process model can be abstracted as a directed graph, of which the nodes stand for activities and the arcs stand for orders.

3.1. Social Network from Business Process

Definition 1 (business process model). Let (P, T, F) be a WF-net (as defined in [14]), and a process model, PM, is a 5-tuple (P, T, F, U, Ψ) where

(i) R is a set of performers,

(ii) $\Psi : T \rightarrow R$ is the function of performer (U) assignment for a task (T).

In a workflow net, P, T, and F refer to places, transitions, and flow relations, respectively. The models in the process repository are represented in terms of a Petri net, in which activities are modeled by transitions, and casual dependencies are modeled by places and arcs, while performers related to activities are specified in the above transitions.

Next, we give a formal definition of social network as the following.

Definition 2 (social network model). A social network, SN, is a graph denoted by a 5-tuple $\mathrm{SN} = (U, A, E, R, \Phi)$, where

(i) U is the finite set of nodes and each node stands for a user,

(ii) A is a set of attributes of users,

(iii) $E \subseteq U \times U$ is the finite set of edges,

(iv) $R = \{R_1, \ldots, R_m\}$ is a finite set of relationships,

(v) $\Phi : U \times U \rightarrow R$ is the relationship function.

As the main manifestation of social network, SNS (social network software/social network site) has gained huge success in business and become the most concerned topic in related research fields. In [15], SNS is defined as follows:

> *. . .web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system.*

A more recent definition is proposed by Kwon and Wen [16] who define SNS as "*websites that allow building relationships online between persons by means of collecting useful information and sharing it with people. Also, they can create groups which allow interacting amongst users with similar interests.*"

SNS is more and more applied or studied in business domain and also attracts research attention as marketing

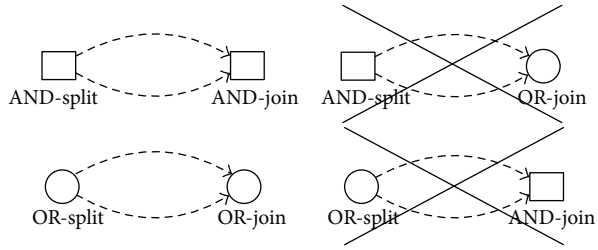


Figure 1: Good and bad constructions.

instruments [17–22], while the globalization and virtualization of enterprise business make efficient business process management (BPM) increasingly important. This inspires us to study how to apply SNS and the related researches in BPM. Note that the activities in business process are modeled by transitions and the transitions and places are both defined as nodes in WF-net. At the same time, performers related to activities are specified in the transitions. Therefore, the performers can be specified as users in the SNS and the order by which performers execute the tasks can be specified as the friend relationship in the SNS.

Almost in all cases, a process has four main structures, including AND-split, OR-split, and AND-join, OR-join. In short, the type "*AND*" means enforcement and the type "*OR*" means "select" or "condition execution." Intuitively, the combination of the *split* and *join* will be a total of four kinds of forms. However, a "good" workflow can balance AND/OR-splits and AND/OR-joins [23]; that is, two parallel flows initiated by an AND-split should not be joined by an OR-join. Two alternative flows created via an OR-split should not be synchronized by an AND-join. As shown in Figure 1, an AND-split should be complemented by an AND-join and an OR-split should be complemented by an OR-join [23]. So, in fact, there are only two kinds of forms (shown in the left of Figure 1) that own well structures. The other two kinds of forms (shown in the right of Figure 1) are defined as deadlocks and lack of synchronization, respectively. And various methods can be used to detect these structure errors [24–26].

Given that every process is verified not to have bad constructions before it is added into the process repository, it is enough to consider the good constructions only. Therefore, we can only consider the AND-split or OR-split and the AND-join or OR-join will be implied in the good constructions.

For simplicity's sake, the places and transitions can be combined as nodes, and the tasks, performers, split type, join type, and other properties of places and transitions can be added into the nodes' properties. Under this idea, business process is a series of activities that is performed by different performers to specific targets, respectively, and the order of activities represents collaboration of these performers. In a workflow system, the process reflects the actual business process, and the activity node represents business operation in enterprises. Information or operations will flow or be conducted in turn according to the nodes sequence. Therefore, the business process model can be abstracted as the directed graph, of which the nodes stand for activities, with a corresponding label (implicating activity type, content, the serial number, etc.), and the edges stand for orders.

Definition 3 (business process graph). Let T be a set of node types and let L be a finite alphabet of labels for nodes. A business process is a connected graph denoted by a 6-tuple $P = (N, E, \alpha, \beta, s, e)$, where

(i) N is the finite set of nodes,

(ii) $E \subseteq N \times N$ is the finite set of edges,

(iii) $\alpha : N \to T$ is the node typing function,

(iv) $\beta : N \to L$ is the node labeling function,

(v) $s \in N$ is the start node,

(vi) $e \in N$ is the end node.

Node $x \in N$ is the input of node $y \in N$ if and only if there exists a directed edge connecting x to y (i.e., $(x, y) \in E$). Node $x \in N$ is the output of node $y \in N$ if and only if there exists a directed edge connecting y to x (i.e., $(y, x) \in E$). The node that has no input node is called *start* node and the node that has no output node is called *end* node. In a workflow repository, any business process can be modeled as a process graph having only one *start* node and only one *end* node. So the business process graph P in Definition 3 can also be simplified as $P = (N, E, \alpha, \beta)$.

However, in a social network site, for example, twitter or Sina Weibo, there exist several interexchange methods, such as tweet, @, reply, and retweet. Tweet is the conduct of publishing a message, retweet is to forward a message that others tweeted without any change, @ is used to push the message to others by force, and reply tends to accept message self-selectively with no or more additional information. Thus, when every performer in a social network site is a user, their relationships that are implied in process structure can be regarded as information exchange in the social network. For example, if there is a node "*A*" with AND-split type connecting to nodes "*B*" and "*C*" in a process, the performer of "*A*" will be a user "*User_A*" in corresponding social network who replies a message that is received from upper user (that is performer of parent node) with the additional information about the node label "label_*A*" firstly. Then @ the message to users "*User_B*" and "*User_C*" who will reply the message in response. Similarly, if there is a node "*A*" with OR-split type connecting to nodes "*B*" and "*C*" in a process, the performer of "*A*" will be a user "*User_A*" in corresponding social network who only replies a message that is received from upper user with the additional information about the node label, and then the users "*User_B*" and "*User_C*" will reply the message. Such a proceeding makes the generated social network retain most socialization properties, and the study on message interaction can be used to analyse the dynamic features of business process. Compared to (1), social network from process models are more about static features based on graph algorithms that are used in network research.

It is important to note that the users being performers in the same process only depend on whether they @ or reply the same topic (the message series) when reconstructing a

process from social network. But what is the first message in a message series? We define a message series code for each message series and the code is also the only representative of the response process. So the message series code can also be called as process code. The first message in a message series is the series code that is published by the process creator. So a whole process can be found from its creator in the social network. In actual applications, a user can search for suitable process models and generate a new model by combining them; for example, user "U1" created a process P1 and tweeted the process code C1, and user "U2" created a process P2 and tweeted the process code C2. When user "U3" creates a new process P3, he may use the P1 and P2 as parts of the P3. Then in the social network, "U3" must retweet C1 and C2 in turn before he tweets the process code C3 and the C3 must contain the C1 and C2. Next, the performer of the first node of P3 will reply the message C3, . . ., and so forth. Thus, study on social network (2) from user history can be transformed to relationship mining from users that @ or reply the same or related topic. Likewise, study on social network (3) from insertion history can be transformed to relationship mining from users that retweet related topics.

Finally, three types of social networks defined in [12] can be replaced by the only social network constructed by the method that this paper presents. And the social network retains as much social features as possible. However, how to define message series code as the sole representative of a process is a problem that has to be solved. To facilitate reading, the message series code will be called as process code in the following, and the definition will be given.

3.2. Process Code. DFS (depth first search) and BFS (breadth first search) are widely used in graph mining algorithms. Compared with BFS, DFS consumes less memory but runs slower due to the stack utilization. On the contrary, the BFS uses more memory but runs faster than DFS. The facts in construction practice of process repositories and process patterns show that the vast majority of business processes are simply structural and with less nodes (in process repository used in this paper, there are 86% processes containing less than 20 nodes, and of which nearly 71% processes contain less than 10). Thus, the problem of memory usage is no longer an issue. On the contrary, because the number of processes may be huge, the time performance is more important. Therefore, this paper uses BFS to define the process code.

Breadth first search by different node orders of the same hierarchy may lead to different BFS sequences. Therefore, this paper presents the standard BFS sequence to ensure that the BFS sequence of the same process is unique. And we can reconstruct unique process with the standard BFS sequence. Next, we give the definitions about BFS code based on BFS sequence.

Definition 4 (BFS code). The breadth first traversal order on a process graph P is a linear order. As Definition 3 shows that the directed edge of P can be labeled by ordered nodes that the edge is connected to; therefore, BFS code of P can be represented as the following:

$$\begin{aligned}&\text{BFSsequence}\\&= s\#\left\{sn_i : sn_i \in E\right\}\#\left\{n_in_j : n_in_j \in E\right\}\#\cdots\left\{n_te : n_te \in E\right\}\#e,\end{aligned} \tag{1}$$

where the symbol # is the separator that divides the different traversal hierarchies.

For example, as shown in Figure 2, BFS code of the process sample P is

$$\text{BFSsequence}\,(P) = s\#sX\#XY,\ XZ\#YW,\ ZW\#We\#e. \tag{2}$$

Definition 5 (standard BFS code). The BFS code of process P is called standard BFS code if the labels of the same hierarchies are by lexicographic order denoted as BFSsequence(P).

Except for (2), the BFS code of process that is shown in Figure 1 can also be $s\#sX\#XZ, XY\#ZW, YW\#We\#e$, or others. But only (2) follows the Definition 5, so only (2) can be called standard BFS code of the process.

Next, we give the definition of process node label that is used in the process code.

Definition 6 (process node label). A process node label is a value of label function $l : F \cup E \cup C \rightarrow \Omega$, where

(i) F is a finite set of functions;

(ii) E is a finite set of events;

(iii) C is a finite set of connectors, and that is {and, or};

(iv) Ω is a set of text labels.

The process code is composed of each node's label, while the message that a user replies or @ comprises the relevant functions, events, and connector.

3.3. Process Matching Degree. Note that if a process is a fragment of the other process, then the first process matches the other, but the opposite is not always true. However, for simplicity, we do not consider the difference and regard them matching with each other.

In different intention, process matching degree is related to user matching degree, structural matching degree, and behavioral matching degree.

User matching degree between two processes can be measured based on their creators, performers, and reusers, and thus the user matching degree can be divided into three parts: creator matching degree, performer matching degree, and reuser matching degree. Let UC be a set of creators of a process and the processes that it reuses, and |UC| is the length of the UC; then the creator matching degree between two processes $p1$ and $p2$ is

$$Mat\_creator\,(p1, p2) = \frac{\left|\mathrm{UC}\,(p1) \cap \mathrm{UC}\,(p2)\right|}{\min\left(\left|\mathrm{UC}\,(p1)\right|, \left|\mathrm{UC}\,(p2)\right|\right)}. \tag{3}$$

Let UP be a set of performers of a process, and |UP| is the length of the UP; then the performer matching degree between two processes $p1$ and $p2$ is

$$Mat\_performer\,(p1, p2) = \frac{|\mathrm{UP}\,(p1) \cap \mathrm{UP}\,(p2)|}{\min\left(|\mathrm{UP}\,(p1)|, |\mathrm{UP}\,(p2)|\right)}. \quad (4)$$

Let UR be a set of creators of processes that a certain process is reused, and UR is the length of the UR; then the reuser matching degree between two processes $p1$ and $p2$ is

$$Mat\_reuser\,(p1, p2) = \frac{|\mathrm{UR}\,(p1) \cap \mathrm{UR}\,(p2)|}{\min\left(|\mathrm{UR}\,(p1)|, |\mathrm{UR}\,(p2)|\right)}. \quad (5)$$

The process user matching degree is a combined metric that can be represented as the following:

$$\begin{aligned} Mat\_user\,(p1, p2) = {} & \alpha \cdot Mat\_creator\,(p1, p2) \\ & + \beta \cdot Mat\_performer\,(p1, p2) \\ & + \lambda \cdot Mat\_\mathrm{reuser}\,(p1, p2), \end{aligned} \quad (6)$$

where α, β, and λ are the adjustment coefficients that are set to 0.2, 0.7, and 0.1, respectively, in this paper.

The process structural matching degree can be calculated according to process matching degree that is based on BFS code [27]. In this paper, we present a novel metric that takes the information propagation path into account. Firstly, we define the social network schema and metapath as follows.

Definition 7 (social network schema). A social network schema from social network model is $S = (O, R)$, where

(i) O is the finite set of roles and each user in the social network corresponds to a role;

(ii) $R = \{\mathrm{and}, \mathrm{or}\}$ is relationships between the roles that different users to.

Definition 8 (Metapath). A metapath of information propagation is a path defined on the network schema $S = (O, R)$ and is denoted in the form of $O_1 \xrightarrow{R_1} O_2 \cdots O_n \xrightarrow{R_n} O_{n+1}$.

For example, in the process sample p in Figure 2, if the users of node belong to roles of *staff*, *accountant*, *manager* and *treasurer,* respectively, and the split type of X is AND-split, then the metapath may be $staff \xrightarrow{and} manager \xrightarrow{and} treasurer$. Meanwhile, if the users belong to roles of *staff*, *manager*, *accountant*, and *treasurer*, respectively, and the split type of X is OR-split, then path $staff \xrightarrow{or} manager \xrightarrow{and} treasurer$ and path $staff \xrightarrow{or} accountant \xrightarrow{and} treasurer$ are the two metapaths. If the roles are labeled with unique symbol (as shown in Table 1) and the relationships *and* and *or* are labeled as R_1 and R_2, then the two metapaths can be represented as string $O_1R_2O_2R_1O_4$ and $O_1R_2O_3R_1O_4$.

As we discussed above, a process may have one or more metapaths in the social network. The matching degree of the two metapaths can be measured by the SED [28] of the labeled metapaths.

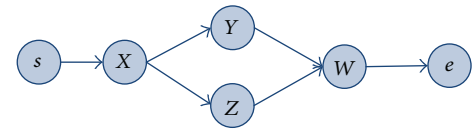


Figure 2: Process sample P.

Table 1: Label of role.

Role	*Staff*	*Manager*	*Accountant*	*Treasurer*
Label	O_1	O_2	O_3	O_4

Definition 9 (string edit distance (SED)). Given two strings x and y, the string edit distance of x and y, denoted as SED(x, y), is the minimum number of insertions, deletions, and substitutions to transform x into y. And the matching degree between metapaths mp_1 and mp_2 (labeled as lmp_1 and lmp_2, resp.) is

$$\begin{aligned} & Mat\_lmp\,(mp_1, mp_2) \\ & = 1 - \frac{\mathrm{SED}\,(lmp_1, lmp_2) - \mathrm{abs}\left(|lmp_1| - |lmp_2|\right)}{\min\left(|lmp_1|, |lmp_2|\right)}, \end{aligned} \quad (7)$$

where $|lmp_1|$ and $|lmp_2|$ are the length of lmp_1 and lmp_2, respectively.

Secondly, we give the definition of process structural matching degree.

Definition 10 (structural matching degree). Given two processes $p1$ and $p2$ and that $p1$ has m metapaths (from $p1\_mp_1$ to $p1\_mp_m$) and $p2$ has n metapaths (from $p2\_mp_1$ to $p2\_mp_n$), if $m > n$, then the structural matching degree is

$$\begin{aligned} & Mat\_structure\,(p1, p2) \\ & = \frac{\sum_{i=1}^{n} \max\left(Mat\_lmp\left(p1\_mp_i, p2\_mp_j\right)\right) \mid j \in \{1, \ldots, m\}}{\min(m, n)}. \end{aligned} \quad (8)$$

The behavioral matching degree mainly measures the dynamic features about functions and events. Generally, functions and events are named by the unified naming conventions (if not, it can use the ETL tool and semantic analysis to standardize the names). Given processes $p1$ and $p2$, if $p1$has nodes with functions of $p1\_f_1, p1\_f_2, \ldots, p1\_f_m$ and events of $p1\_e_1, p1\_e_2, \ldots, p1\_e_m$, and $p2$has nodes with functions of $p2\_f_1, p2\_f_2, \ldots, p2\_f_n$ and events of $p2\_e_1, p2\_e_2, \ldots, p2\_e_n$, then the process behavioral matching

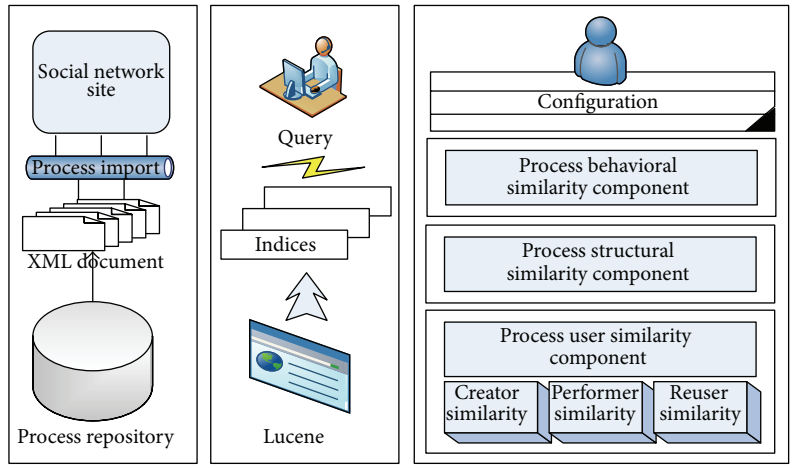


Figure 3: Structure of recommender prototype.

degree between $p1$ and $p2$ can be calculated by the following formula:

$$Mat\_behavioral\,(p1, p2) = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} \left(IsSame\left(p1_{f_i}, p2_{f_j}\right) + IsSame\left(p1_{e_i}, p2_{e_j}\right) \right) \right) \times (2 \times \min(m, n))^{-1}, \quad (9)$$

where $IsSame\,(\cdot)$ is the function that whether two parameters are equal in semantic sense and it returns 1 if they are equal and 0 if not. For simplicity, the behavioral matching degree can only be measured by functions and the formula will be as the following:

$$Mat_{behavioral(p1,p2)} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} IsSame\left(p1_{f_i}, p2_{f_j}\right)}{\min(m, n)}. \quad (10)$$

4. Implementation

Social network supported process recommender system turns to help process builders fitting processes to achieve a modeling intention with regard to the other builders' modeling intention and modeling behavior.

This recommendation system implementation consists of three parts (as shown in Figure 3) as follows.

(1) A simple social network site meets our research needs by providing the basic functions of tweeting, retweeting, @ing, replying and searching, and so forth. A process import method is also provided in this part and each process, in the form of a well-formed XML document, can be added into the SNS database automatically.

(2) A query interface allows users to request process models or process model parts that are of interest to them. In the system, we use open-source MySQL database to provide index capability, but not well in large dataset task. Therefore, we use Lucene to build indices on the objects of processes and network and provide query on creators, performers, reusers, functions, events, connectors, and so on through the query parser syntax.

(3) A series of recommender components proposes appropriate process models which fit to a business process model that is currently being edited under different considerations. The parameters configuration is also provided for the users to determine the execution sequence of three types of process matching degrees and the adjustment coefficients for the process user matching degree. The thresholds of three matching degrees in certain recommending proceeding can be set through parameter configuration.

Different parameters values may lead to different efficiencies, even different results. In the next section, the difference will be demonstrated and discussed.

5. Empirical Evaluation

In this section, a comprehensive study is conducted in our experiments on synthetic dataset. The synthetic data generator that we use is similar to [10], and we also improve the generator with the ability to generate the processes with different kinds of structures (i.e., AND-split, OR-split, etc.). The experimental dataset contains 500 processes in total of which each activity is assigned to a performer and a function at random. And there totally contain 1000 virtual functions that can be semantically transformed to 200 different functions and 1000 virtual users that can be distinguished into 50 different roles.

In the experiment, 5 process fragments (identified as (1), (2), (3), (4), and (5) in Figure 4) are extracted from the experimental dataset as the candidate processes and the three types of matching degrees between every candidate process

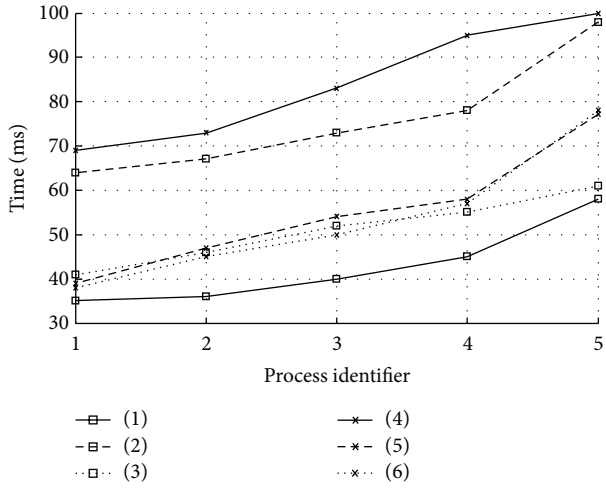


Figure 4: Experiment performance.

Table 2: Experiment result.

	Parameters configuration	Average time (ms)	The average number of recommended processes
(1)	$\theta_u = 0.8$ $\theta_s = 0.8$ $\theta_b = 0.8$	43	5
(2)	$\theta_s = 0.8$ $\theta_b = 0.8$ $\theta_u = 0.8$	76	5
(3)	$\theta_b = 0.8$ $\theta_u = 0.8$ $\theta_s = 0.8$	51	5
(4)	$\theta_s = 0.6$ $\theta_b = 0.8$ $\theta_u = 0.8$	84	8
(5)	$\theta_s = 0.8$ $\theta_b = 0.6$ $\theta_u = 0.8$	55	11
(6)	$\theta_s = 0.8$ $\theta_b = 0.8$ $\theta_u = 0.6$	53	17

and each process in experiment dataset are calculated under different thresholds groups (identified as (1), (2), (3), (4), (5), and (6) in Figure 4) in different orders successively. As it can be seen from Figure 4, the execution performance may vary considerably in different orders even under the same thresholds.

The average execution time and the average number of recommended processes are listed in Table 2.

In the same execution order, adjusting every threshold in turns may lead to big change in the results. From Table 2, the decrease of the threshold of structural matching degree makes the average execution time increase much more and this may be due to poor efficiency of the calculation of user matching degree and behavioral matching degree. Moreover, the decrease of the threshold of user matching degree makes the average number of recommended processes increase much more and this may be due to less overlapping between users of different processes in the process dataset.

6. Conclusion

We propose a social network supported process recommender and related techniques which can improve process modeling by providing reference processes from different perspectives to extend or complete the business process under construction, especially in some social considerations.

The similarities between process features and social network features are discussed in detail. And three types of process matching degrees between two processes are proposed by the consideration of contexts in different situations; then we implement the system prototype. The experimental evaluation shows that our method is efficient and effective for practical use.

The greatest contribution of this paper is that we show how to add social features to a recommendation-based process modeling support system. And the unique social network constructed from the process repository is first proposed which can assist the researchers in understanding social features in process modeling. The three types of process matching degrees are the initial attempt that benefits from the social network. At a more abstract level, this paper strengthens a research stream into process modeling that combines social software and social computing.

Still, much work has to be carried out in the future. For example, the user community with the similar interest may be more important for the modeling intention of process. Furthermore, process mining can benefit from the social network analysis. We hope that this paper can inspire more researchers to join in the newly ascendant field and many more efficient and effective methods or technologies can come to the fore.

Conflict of Interests

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

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant (no. 61272129), the National Science and Technology Supporting Program of China (no. 2012BAH06F02), the Research Foundation for the Doctoral Program by Ministry of Education of China (no. 20110101110066), the New-Century Excellent Talents Program by Ministry of Education of China (no. NCET-12-0491), and the Zhejiang Provincial Natural Science Foundation of China (no. LR13F020002).

References

[1] D. G. Wastell, T. McMaster, and P. Kawalek, "The rise of the phoenix: methodological innovation as a discourse of renewal," *Journal of Information Technology*, vol. 22, no. 1, pp. 59–68, 2007.

[2] M. Dumas, W. M. van der Aalst, and A. H. ter Hofstede, *Process-Aware Information Systems: Bridging People and Software through Process Technology*, Wiley-Interscience, Hoboken, NJ, USA, 2005.

[3] T. Erl, *Service-Oriented Architecture*, vol. 8, Prentice Hall, New York, NY, USA, 2005.

[4] G. Alonso, F. Casati, H. Kuno, and V. Machiraju, *Web Services*, Springer, Berlin, Germany, 2004.

[5] M. Havey, *Essential Business Process Modeling*, O'Reilly, Sebastopol, Calif, USA, 2009.

[6] M. Indulska, J. Recker, M. Rosemann, and P. Green, "Business process modeling: current issues and future challenges," in *Advanced Information Systems Engineering*, vol. 5565 of *Lecture Notes in Computer Science*, pp. 501–514, Springer, Berlin, Germany, 2009.

[7] M. Laguna and J. Marklund, *Business Process Modeling, Simulation and Design*, CRC Press, 2013.

[8] B. Curtis, M. I. Kellner, and J. Over, "Process modeling," *Communications of the ACM*, vol. 35, no. 9, pp. 75–90, 1992.

[9] G. M. Giaglis, "A taxonomy of business process modeling and information systems modeling techniques," *International Journal of Flexible Manufacturing Systems*, vol. 13, no. 2, pp. 209–228, 2001.

[10] J. Zhang, Q. Liu, and K. Xu, "FlowRecommender: a workflow recommendation technique for process provenance," in *Proceedings of the 8th Australasian Data Mining Conference*, vol. 101, Australian Computer Society, 2009.

[11] B. Cao, J. Yin, S. Deng, D. Wang, and Y. Xu, "A near neighbour and maximal subgraph first based business process recommendation technique," *Chinese Journal of Computers*, vol. 36, no. 2, pp. 263–274, 2013.

[12] A. Koschmider, M. Song, and H. A. Reijers, "Social software for business process modeling," *Journal of Information Technology*, vol. 25, no. 3, pp. 308–322, 2010.

[13] A. Koschmider, M. Song, and H. A. Reijers, "Advanced social features in a recommendation system for process modeling," in *Business Information Systems*, vol. 21 of *Lecture Notes in Business Information Processing*, pp. 109–120, Springer, Berlin, Germany, 2009.

[14] W. M. P. van der Aalst, "The application of Petri nets to workflow management," *Journal of Circuits, Systems and Computers*, vol. 8, no. 1, pp. 21–66, 1998.

[15] D. M. Boyd and N. B. Ellison, "Social network sites: definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[16] O. Kwon and Y. Wen, "An empirical study of the factors affecting social network service use," *Computers in Human Behavior*, vol. 26, no. 2, pp. 254–263, 2010.

[17] E. Constantinides and S. J. Fountain, "Web 2.0: conceptual foundations and marketing issues," *Journal of Direct, Data and Digital Marketing Practice*, vol. 9, no. 3, pp. 231–244, 2008.

[18] R. D. Waters, E. Burnett, A. Lamm, and J. Lucas, "Engaging stakeholders through social networking: how nonprofit organizations are using Facebook," *Public Relations Review*, vol. 35, no. 2, pp. 102–106, 2009.

[19] H. Tikkanen, J. Hietanen, T. Henttonen, and J. Rokka, "Exploring virtual worlds: success factors in virtual world marketing," *Management Decision*, vol. 47, no. 8, pp. 1357–1381, 2009.

[20] T. Hogg, "Inferring preference correlations from social networks," *Electronic Commerce Research and Applications*, vol. 9, no. 1, pp. 29–37, 2010.

[21] T. J. Spaulding, "How can virtual communities create value for business?" *Electronic Commerce Research and Applications*, vol. 9, no. 1, pp. 38–49, 2010.

[22] E. Constantinides, C. Lorenzo-Romero, and M. Alarcón-del-Amo, "Social networking sites as business tool: a study of user behavior," in *Business Process Management*, vol. 444 of *Studies in Computational Intelligence*, pp. 221–240, Springer, Berlin, Germany, 2013.

[23] W. M. P. van der Aalst, "Workflow verification: finding control-flow errors using petri-net-based techniques," in *Business Process Management*, vol. 1806 of *Lecture Notes in Computer Science*, pp. 161–183, Springer, Berlin, Germany, 2000.

[24] Y. Ye, J. Yin, Z. Feng, and B. Cao, "Wolf-pack algorithm for business process model syntactic and semantic structure verification in the workflow management environment," in *Proceedings of the IEEE Asia-Pacific Services Computing Conference (APSCC '10)*, pp. 694–699, December 2010.

[25] F. Rabbi, H. Wang, and W. MacCaull, "YAWL2DVE: an automated translator for workflow verification," in *Proceedings of the 4th IEEE International Conference on Secure Software Integration and Reliability Improvement (SSIRI '10)*, pp. 53–59, June 2010.

[26] H. Pranevičius and R. Misevičienė, "Verification of business process workflows," *Technological and Economic Development of Economy*, vol. 18, no. 4, pp. 623–635, 2012.

[27] Y. Ye, J. Yin, and B. Cao, "Process warping matrix based business process recommendation technique," in *Proceedings of the 3rd China Conference on Business Process Management (CBPM '13)*, 2013.

[28] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 172014, 6 pages
http://dx.doi.org/10.1155/2014/172014

Research Article

Proactive Schema Based Link Lifetime Estimation and Connectivity Ratio

Bouamoud Bachir,[1] Ouacha Ali,[2] Habbani Ahmed,[1,2] and Elkoutbi Mohamed[1]

[1] *SIME Lab, MIS Team, ENSIAS, University of Mohammed V SOUISSI, Madinat Al Irfane, BP 713, Agdal, Rabat, Morocco*
[2] *LEC Lab, MIS Team, EMI, University of Mohammed V AGDAL, Rabat, Morocco*

Correspondence should be addressed to Bouamoud Bachir; bachir.bouamoud@gmail.com

Received 16 August 2013; Accepted 11 September 2013; Published 9 January 2014

Academic Editors: W. Sun and J. Zhou

The radio link between a pair of wireless nodes is affected by a set of random factors such as transmission range, node mobility, and environment conditions. The properties of such radio links are continually experienced when nodes status balances between being reachable and being unreachable; thereby on completion of each experience the statistical distribution of link lifetime is updated. This aspect is emphasized in mobile ad hoc network especially when it is deployed in some fields that require intelligent processing of data information such as aerospace domain.

1. Introduction

Ad hoc networks have given numerous innovative possibilities in using computer networks in several situations. In fact mobile ad hoc networks are based on wireless communication; for that reason the requirement of many of services developed for wired networks does not match with resource availability in mobile networks.

In Manet, nodes forming a mobile network are not joined to any infrastructure and form networks on the fly during various time spans. When forming a mobile network there are many factors of complexity such as the shared wireless medium, the mobility of the nodes, the energy constraints, and the agitation of nodes that may join or leave the carrier range anytime during the network lifetime.

In context of high mobility, the routes in the network usually have a short life span. A route may or may not exist for the entire duration of a data communication session. Due to the dynamic changes of the factors that affect the performance in a mobile ad hoc network, it would be convenient that any proposed optimizations should consider this dynamics that acts on nodes and links which interconnect them. In this context, knowledge of network must also have the same character in terms of taking into account factors such as available bandwidth, delays, and the lifetime of nodes in the process of selection of multipoint relays.

In proactive routing mechanism, the use of relays aims to reduce the broadcast messages senders and then the number of flooded messages; here we highlight the importance of relay in OLSR protocol case, since they become the only responsible of broadcasting topology control messages. While this approach is pleasing to the eye, besides some control functions that are necessary to prevent an eternal duplication of broadcast messages, it is required to select relays in reliable manner; indeed defects in the reception of broadcast packets from MPR nodes can greatly affect the rate of delivery of packets across the network.

Thus all applications in ad hoc network depend on reliable and efficient routing of packets. Hence, it is extremely important to design routing protocols that can work within several constraints particularly those of aerospace applications, namely, mobility and lack of energy resources, and provide support for all higher level applications.

There have been many existing routing protocols for ad hoc networks emphasizing different implementation scenarios. However, the basic goals have always been to devise a routing protocol that minimizes control overhead, packet loss ratio, and energy usage while maximizing the throughput.

Because these types of network can be used in a variety of situations (disaster recovery, battlefields, conferences, and so on), they differ in terms of their requirements and complexities.

The key idea of this paper is to apply our link reliability estimation model to the multipoint relays selection and routing calculation of the standard OLSR protocol in order to guarantee reliable data transmission. In order to judge the reliability of the links in our link reliability estimation model, we use a statistic inference method that infers the reliability of links by accumulated past information from experimental results. That is, reliability of the links can be estimated by accumulated past information of link lifetime value.

There are two contributions in this paper. First one is the establishment of a link lifetime model to estimate the reliability of links based on a statistical distribution, which is constructed by the value that a given node observes. Second one is the introduction of the later model to optimize the selection of multipoint relays in the context of a proactive routing.

The paper is organized as follows. In Section 2, we present some previous works whose aim is to improve routing reliability based on efficient MPR selection. In Section 3, we introduce a new composite metric based on connectivity ratio and link lifetime prediction. Section 4 contains a set of simulations and results which concern the evaluation of our proposal. In Section 5, we evaluate the performance of the proposed scheme. Lastly in Section 6 we conclude and discuss future works.

2. Related Work

In the context of Proactive routing, the multipoint relay techniques perform very well to desseminate the broadcasted packet into the network. Indeed, many works present in the literature [1–8] were interested in this topic. This puts in evidence that multipoint relaying schemes still arouse the researchers interest.

In fact multipoint relay (MPR) nodes play an important role to deliver broadcast packets to their next MPR nodes as well as their neighbour nodes; failure to receive broadcast packets from MPR nodes can greatly affect the network performance. Then selecting reliable MPR based on link lifetime and ratio of connectivity present promising issues to improve routing performances

On one hand, reliability is a major issue in routing and some previous studies [9, 10] focused on it such that the shortest paths are usually preferred to route packets. Such as we assume the shortest path may fail quickly, because some of wireless links on the shortest path may be broken shortly after the establishment of the path, due to velocity and direction of nodes.

The velocity of the mobile nodes within a MANET is not unchanging. At the same time as there is no speed control of the wireless devices a desirable protocol for MANET should perform well both in relatively static and in totally vibrant network circumstances. The direction of a node's mobility is not known in advance. One common incident is that a node travels to a way where the density of nodes is low or there is no neighbor node in the carrier range.

Abundant research contributions to wireless routing are based on online nodes measurements in order to categorize nodes or links which are preferentially used in the established paths whereas their common weakness is inability to prevent possible change in links status occurring in future. Indeed link qualified as reliable based on past or current measurements may become unreliable with time because of dynamic nature of mobile environments. These drawbacks conducted others in [11, 12], to investigate some prediction method to expect the link availability in a continuous manner during a time span, based on link lifetime.

The residual link lifetime estimation is based on an offline generalized link lifetime distribution, which is unavailable in runtime. However, knowing that the link lifetime distribution is approximately a normal distribution, the trend of residual link lifetime changes can be deduced in advance.

Since the environment conditions of network are changing and velocity and direction of node are not known in advance, the parameters of links lifetime distribution are significantly correlated with the latter assumption.

3. Metric Proposal

3.1. Link Lifetime. Beginning from a typical example of a mobile ad hoc network consisting of a set of nodes among which a dynamic establishment of links such as $G(V, E)$ is a direct graph and V is the set nodes and E is the set of links $l = (x, y)$, where the node y is within the transmission range of x. In general cases routing metrics have attributed values which we call weights associated with each edge:

$\text{hop}(x, y)$: number of hops between node x and node y.

$N_1(x) = \{y \mid \text{hop}(x, y) = 1\}$, set of 1-hop adjacent nodes.

$N_2(x) = \{y \mid \text{hop}(x, y) = 2\}$, set of 2-hop adjacent nodes.

If we focus on the link state change, it varies according to the distance between a pair of nodes such as whether they are in the career range of each other. From statistical point of view this network could be assimilated into a statistical population of nodes.

Each node in the network, for example, (Figure 1) has a changing set of $N_1(x)$ that can be considered as a random and repeated sample of the population, each time using various sample size n. From each of these repeated samples we calculate independent sample means to form a sampling distribution mean.

Our concern in the current article is to seek for link lifetime measurement method based on what each node observes on the sample whose size is $|N_1(x)|$. The distribution of link lifetime was the concern of many research works such as [13] where the authors illustrate that links lifetime has a normal distribution.

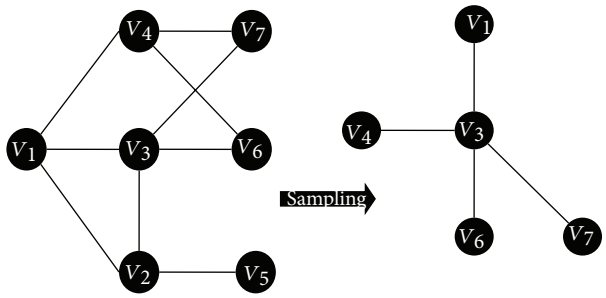


Figure 1: Typical example of nodes sampling.

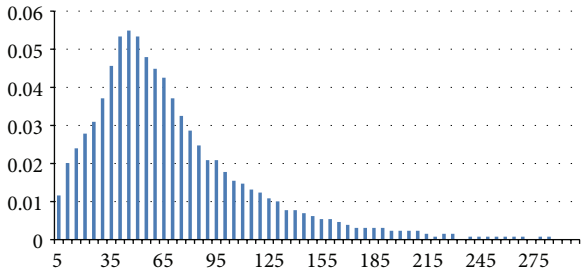


Figure 2: Distribution of link durations in a Gauss-Markov scenario.

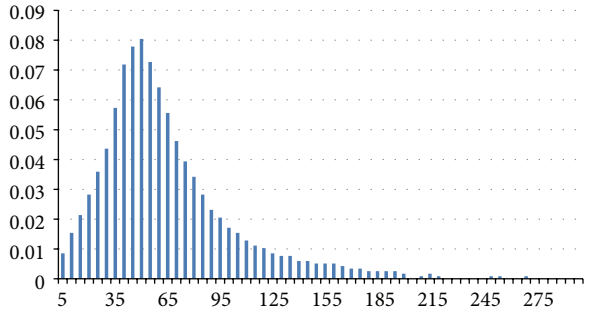


Figure 3: Distribution of link durations in a Random Waypoint scenario.

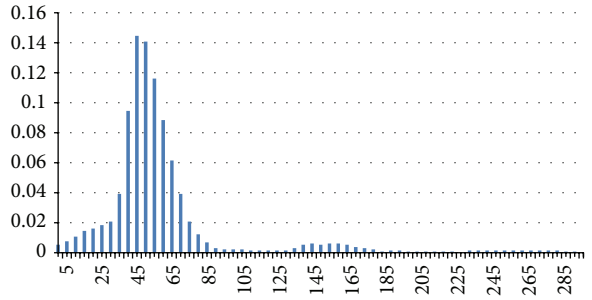


Figure 4: Distribution of link durations in a Manhattan Grid scenario.

The main works present in the recent literature interested to link lifetime prediction are based on an offline generalized link lifetime distribution, which is unavailable in runtime. The principal weakness of this approach is being biased; in other term it is dependent on the environment conditions which affect the latter distribution parameters such as its mean m and variance σ^2; then the only way to be efficient is by experiencing the network and generating the distribution to know its parameter on which the routing algorithms are based. So the main idea of our proposal is to use the results obtained from samples of link lifetime to describe the population and process this statistics to estimate the parameters of the latter distribution dynamically.

Hereafter we present some interesting results that appeared in [13]; Figures 2, 3, and 4 illustrate the distribution of link lifetime in several mobility scenarios, namely, Gauss-Markov, Random Waypoint, and Manhattan Grid. The main feature of this graph is that it has almost a normal distribution.

Also we conducted our own simulation to validate this hypothesis, as it's shown in (Figure 5).

Definition 1 (Random variable of link lifetime). We considered the set $\mathbb{V}$ of nodes forming the network as population whose elements possess a measurable character which is the realization of a random variable (link lifetime) that follows a normal or Gaussian distribution $(m; \sigma^2)X$. We assume a finite population and $|\mathbb{V}| \gg N_1(V_i)$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)((x-m)/\sigma)^2}. \quad (1)$$

Each node V_i samples a random set of n nodes (links); then we obtain the following series of values $x_{i,1}, x_{i,2}, \ldots x_{i,n}$.

The completions of the law X as in Figure 6 is given by

$$P\left(x \geq x_{i,j}\right) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_{i,j}}^{+\infty} e^{-(1/2)((x-m)/\sigma)^2} dx. \quad (2)$$

3.2. 2-Hop Connectivity Ratio. Assuming that every node $y \in N_1(V_i)$ we define a ratio of reachability such as

$$r(y) = \frac{\left|N_2(V_i) \cap N_1(y)\right|}{N_2(V_i)}. \quad (3)$$

3.3. New Decision Process in Multipoint Relay Selection Procedure. Our proposal is to calculate the reliability of a candidate node V_j to be in the MPR set of V_i, by being aware of the connectivity and the link lifetime mean emanating from that node to its neighbour. Consider the following:

$$\mathcal{R}\left(V_j\right) = \frac{\left|N_2(V_i) \cap N_1\left(V_j\right)\right|}{N_1(V_i)} \times \frac{1}{\sigma\sqrt{2\pi}} \int_{x_{i,j}}^{+\infty} e^{-(1/2)((x-m)/\sigma)^2} dx. \quad (4)$$

See Procedure 1.

4. Simulations Environments

Our proposal is experienced under network simulator NS2 [14] (Network Simulator) 2.35 version in which we have

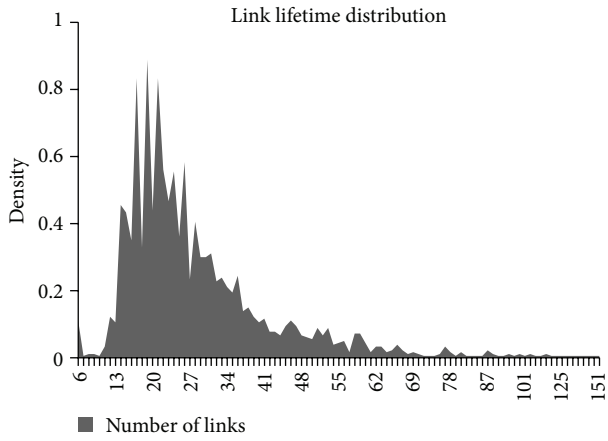


Figure 5: Distribution of link durations in a Random Waypoint scenario.

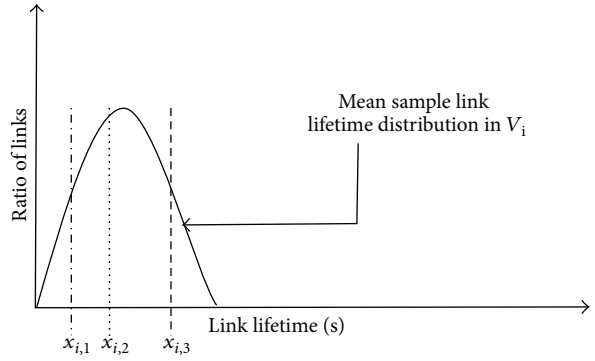


Figure 6: Calculation of remaining link lifetime probability.

integrated a standard version of OLSR (UM-OLSR-0.8.8 [15, 16]), which is developed by MASIMUM (MANET Simulation and Implementation at the University of Murcia).

Our simulation parameters are as follow. For all simulations, our network is consisted of a maximum number of mobile nodes (70) whose radio scoop is 100 m, moving in an area of $1000 \times 1000\,m^2$. Each node moves according to the RWP (Random Waypoint) mobility model [17] with pause time fixed to 0 second and maximum speed that varies between 5 and 30 meter/second with step of 5. The scenario that defines the nodes movement is regenerated at the beginning of each simulation. To generate traffic in the network, in each simulation, 1/5 of nodes are randomly selected to be a source of CBR (Constant Bit Rate) traffic. And these selected nodes use UDP (User Datagram Protocol) connections to send Packets with 1024 bytes of size such that one packet every 2.5 second is sent. Table 1 summarizes all the parameters used during simulations.

5. Results and Discussions

In the simulations work set all the possible attributes of network could be highlighted, such as node mobility, network load, the node density, and the number of connections.

Table 1: Simulation parameters.

Simulation environment	Option and parameter
Flat size	1000 m × 1000 m
Max number of nodes	70 nodes
Radio scoop	250 m
MAC layer	IEEE.802.11.peer to peer mode
Transport layer	User Datagram Protocol (UDP)
Traffic model used	CBR
Package size	1024 bytes
Rate	0.4
The number of connections	1/5 of the number of nodes
Mobility model	RWP (Random Waypoint)
Pause time	0 second
Maximum speed of nodes	5, 10, 15, 20, and 30 m/s
Simulation time	300 sec

In our context we care about link lifetime in mobile environment; indeed we vary the mobility of nodes: we start with a mobility scenario in which the nodes have a low velocity of 5 m/s (18 km/s). We then increase the node velocity up to 25 m/s (90 km/h). Our intention is to investigate the behavior of protocols in networks with varied mobility. We keep a constant data rate of 10 packets/s (40.960 kbps) and a constant number of connections.

Hereafter we compare the resulting performance of our proposal referred to as correlated lifetime and connectivity estimator (CLCE) and the standard OLSR referred as OLSR-STD.

In Figure 7 we observe that, at the high speed, the performance of end to end delay (elapsed time from sending to reception) of both protocols decreases. However the decreased performance is large, mainly due to link breakage, but it still makes the comparison meaningful.

Generally CLCE presents 16% lower delay the OLSR-STD, especially when velocity reaches higher values; this can be explained by the efficiency that CLCE has in selecting MPR which has more probability to be in the direct neighbourhood. In this way the routes continue to be valid and packet retransmission is decreased. As a consequence end to end delay is lower.

6. Conclusions

The key idea of this paper was to apply new link reliability estimation model to the multipoint relays selection and indirectly to routes calculation in the context of a proactive routing protocol, in order to guarantee reliable data transmission. Based on an updated estimation model of link lifetime correlated with a connectivity ratio by this proposal we have considerably optimized the end to end delay from sending to reception. Indeed this new approach could be subject of many application, where delay of packets delivery is critical, namely, in aerospace domain.

Initial: A node k, $N_1(k)$, $N_2(k)$.
Return: MPR_k; MPR set of k
Begin
(1) Add to MPR_k the node in $N_1(k)$ which is the only one to reach a node in $N_2(k)$
(2) Remove the nodes from $N_2(k)$ which are covered by a node in MPR_k
(3) While ($N_2(k)$ not empty) do
(i) For each node V_j in $N_1(k)$,
Calculate $\mathscr{R}(V_j)$
(ii) Add to MPR_k the nodes which has the highest $\mathscr{R}(V_j)$, If multiple choices.If multiple choices, select node with highest degree, that is, number of 1-hop neighbors.
(iii) Remove the nodes from $N_2(k)$ which are covered by a node in MPR_k
End

Procedure 1: Optimized procedure of MPR selection.

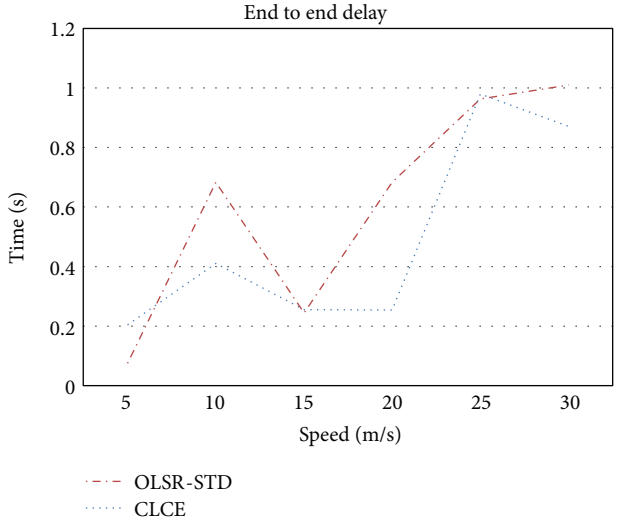


Figure 7: End to end delay comparison.

Conflict of Interests

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

Authors' Contribution

All authors have a significant scientific input into the work.

References

[1] D. Nguyen and P. Minet, "Analysis of MPR selection in the OLSR protocol," in *Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops/Symposia (AINAW '07)*, pp. 887–892, May 2007.

[2] K. Yamada, T. Itokawa, T. Kitasuka, and M. Aritsugi, "Cooperative MPR selection to reduce topology control packets in OLSR," in *Proceedings of the 2010 IEEE Region 10 Conference (TENCON '10)*, pp. 293–298, November 2010.

[3] C. Gomez, D. Garcia, and J. Paradells, "Improving performance of a real ad-hoc network by tuning OLSR parameters," in *Proceedings of the 10th IEEE Symposium on Computers and Communications (ISCC '05)*, pp. 16–21, June 2005.

[4] W. Wei and J. Cao, "A multi-metric QoS routing method for ad hoc network," in *Proceedings of the 4th International Conference on Mobile Ad-hoc and Sensor Networks (MSN '08)*, pp. 99–102, December 2008.

[5] G. Cervera, M. Barbeau, J. Garcia-Alfaro, and E. Kranakis, "Mitigation of topology control traffic attacks in OLSR networks," in *Proceedings of the 5th International Conference on Risks and Security of Internet and Systems (CRiSIS '10)*, pp. 1–8, October 2010.

[6] J. H. Ahn and T.-J. Lee, "A multipoint relay selection method for reliable broadcast in ad hoc networks," in *Proceedings of the International Conference on ICT Convergence (ICTC '11)*, pp. 616–617, September 2011.

[7] Z. Guo, S. Malakooti, S. Sheikh, C. Al-Najjar, and B. Malakooti, "Multi-objective OLSR for proactive routing in MANET with delay, energy, and link lifetime predictions," *Applied Mathematical Modelling*, vol. 35, no. 3, pp. 1413–1426, 2011.

[8] J. Toutouh, J. Garcia-Nieto, and E. Alba, "Intelligent OLSR routing protocol optimization for VANETs," *IEEE Transactions on Vehicular Technology*, vol. 61, pp. 1884–1894, 2012.

[9] K.-I. Kim and S.-H. Kim, "Effectiveness of reliable routing protocols in mobile ad hoc networks," *Wireless Personal Communications*, vol. 38, no. 3, pp. 377–390, 2006.

[10] P. Loh, "A scalable, efficient and reliable routing protocol for wireless sensor networks," in *Ubiquitous Intelligence and Computing*, J. Ma, H. Jin, L. Yang, and J. Tsai, Eds., vol. 4159, pp. 409–418, Springer, Berlin, Germany, 2006.

[11] S. Jiang, "An enhanced prediction-based link availability estimation for MANETs," *IEEE Transactions on Communications*, vol. 52, no. 2, pp. 183–186, 2004.

[12] S. Jiang, D. He, and J. Rao, "A prediction-based link availability estimation for routing metrics in MANETs," *IEEE/ACM Transactions on Networking*, vol. 13, no. 6, pp. 1302–1312, 2005.

[13] M. Gerharz, C. d. Waal, M. Frank, and P. Martini, "Link stability in mobile wireless Ad Hoc networks," in *Proceedings of the 27th Annual IEEE Conference on Local Computer Networks*, 2002.

[14] H. Alwan and A. Agarwal, "Multi-objective reliable multipath routing for wireless sensor networks," in *Proceedings of the IEEE Globecom Workshops (GC '10)*, pp. 1227–1231, December 2010.

[15] R. L. Gomes, W. A. Moreira, J. J. H. Ferreira, and A. J. G. Abele, "LatinCon14: providing QoE and QoS in wireless mesh networks through dynamic choice of routing metrics," *IEEE Latin America Transactions*, vol. 8, no. 4, pp. 454–462, 2010.

[16] G. Koloniari and E. Pitoura, "A game-theoretic approach to the formation of clustered overlay networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 4, pp. 589–597, 2012.

[17] D. Mahjoub and H. El-Rewini, "Adaptive constraint-based multi-objective routing for wireless sensor networks," in *Proceedings of the IEEE International Conference on Pervasive Services (ICPS '07)*, pp. 72–75, July 2007.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 464895, 5 pages
http://dx.doi.org/10.1155/2014/464895

Research Article

Compressed Wideband Spectrum Sensing Based on Discrete Cosine Transform

Yulin Wang and Gengxin Zhang

Institute of Communication Engineering, PLA University of Science and Technology, Yudao Street 14, Nanjing 210007, China

Correspondence should be addressed to Yulin Wang; wang_yulinsci@126.com

Received 6 November 2013; Accepted 25 November 2013; Published 8 January 2014

Academic Editors: W. Sun and J. Zhou

Discrete cosine transform (DCT) is a special type of transform which is widely used for compression of speech and image. However, its use for spectrum sensing has not yet received widespread attention. This paper aims to alleviate the sampling requirements of wideband spectrum sensing by utilizing the compressive sampling (CS) principle and exploiting the unique sparsity structure in the DCT domain. Compared with discrete Fourier transform (DFT), wideband communication signal has much sparser representation and easier implementation in DCT domain. Simulation result shows that the proposed DCT-CSS scheme outperforms the conventional DFT-CSS scheme in terms of MSE of reconstruction signal, detection probability, and computational complexity.

1. Introduction

In cognitive radio networks (CRNs), secondary CR users should fleetly and accurately sense the wideband spectrum, so that they can detect the unused spectrum holes, reconfigure their parameters to utilize the spectrum available, and avoid interference to primary users (PUs) [1, 2]. In practice, only a small part of the wideband spectrum is occupied by the PUs. Too high sampling rate required for scanning the wideband spectrum can cause immense computational costs and sensing problems. In compressed sampling (CS) [3], the sampling and compression operations are combined into a low complexity compressed sampling. For current CRNs, the CS has been used to alleviate the sampling burden, which aims at depressing the sampling rates for the acquisition of wideband signals [4].

To implement compressed wideband spectrum sensing, CRs need to exploit sparsity of signal in frequency domain. In the literature, [4] firstly applied CS for acquiring wideband signals using sub-Nyquist sampling rates, [5] exploited a structured compressed sensing, and [6] studied a power spectrum blind sampling (PSBS) algorithm trying to reconstruct the power spectrum. All of these systems belonged to the class of discrete Fourier transform (DFT) based compressed spectrum acquisition, which employed the complex exponential functions set as orthogonal sparse basis.

However, the DFT complex exponential matrix is not the only orthogonal basis that can be used to reconstruct wideband communication signals. The paper [7] proposed a cyclic spectrum based wideband spectrum estimating scheme considering the 2D sparse signal of interest in the cyclic spectrum domain. The paper [8] considered using an adaptive tree structured dictionary of orthogonal bases to optimize the compressive sensing recovery of image and audio signal. The paper [9] applied several time-frequency transforms including DFT, DCT, and discrete sine transform (DST) to spectrum sensing for cognitive power line communication (PLC) systems. However, there were few studies combining the CSS and DCT for simultaneous estimation of the spectrum occupancy states over a wide band.

A set of cosinusoidal functions can be used as an orthogonal sparse basis to implement the compressed spectrum sensing (CSS) scheme. Hence, we will synthesize the scheme as discrete cosine transform DCT-CSS and the conventional discrete Fourier transform (DFT) based CSS system as DFT-CSS in this paper. Although DFT-CSS algorithm could quite effectively reduce the sampling requirement, high computation complexity replaced in signal processing is

a serious disadvantage. Furthermore, the CRNs require the sensing time to be short to ensure network responsiveness and efficiency. In this paper, our main contribution is to reconstruct a wideband spectrum signal from sub-Nyquist-rate compressive samples by DCT-CSS. The novel algorithm provides more accurate recovery and lower computation complexity.

On the one hand, as another way of time-frequency transfer, the concentration capability of DCT is superior to that of DFT. In this regard, [10] has shown that the DCT is close to optimal in terms of energy-compaction capabilities. Our simulation results indicate that for the same wideband communication signal, representation of signal in DCT domain is much sparser than that in DFT domain. Higher sparsity not only means lower computation complexity and shorter processing time but also means lower minimal sampling rate (compression rate) and much more accurate reconstruction. On the other hand, the DCT uses only real arithmetic, as opposed to the complex-valued DFT. This further reduces the signal-processing complexity/power consumption, especially for real-valued signal samples. As a result of the aforementioned properties, same reconstruction accuracy at a lower implementation complexity and compression rate can be achieved by DCT-CSS algorithm.

The remainder of this paper is organized as follows. In Section 2, the system model is given. Then we introduce comparisons between DCT-CSS and DFT-CSS schemes through coefficients analysis in Section 3. Section 4 proposes the DCT-CSS scheme. Simulation results are presented in Section 5. Finally, we draw our conclusions in Section 6.

2. System Model

2.1. Signal Model. We assume that $r(t)$ is a band-limited signal spanning in a wide spectrum, as shown in Figure 1.

(1) The wideband is divided into m subbands $\{B_1, B_2, \ldots, B_m\}$, and the frequency boundaries are known to the CR. The bandwidth of the spectrum bands occupied by each PU is equally B.

(2) The signal power spectrum density (PSD) over each spectrum subband B_i is smooth; however, the PSD of PUs over two neighboring subbands is independent.

(3) The number of active PU subbands Q and their locations are unknown to the CR nodes.

(4) During the spectrum sensing period, all CRs keep quiet as enforced by protocols, for example, at the media access control layer.

(5) In a sensing period, the locations and the number of active subbands Q keep unchanged but may vary for different sensing period.

2.2. Compressed Sampling. In practice, a signal can always be sparsely or near sparsely represented on a transform domain. For a time window as $t \in [0, \tau]$, $r(t)$ have discrete form as an N-length signal r_t, which can easily be described as

$$r_t = Fr_f, \tag{1}$$

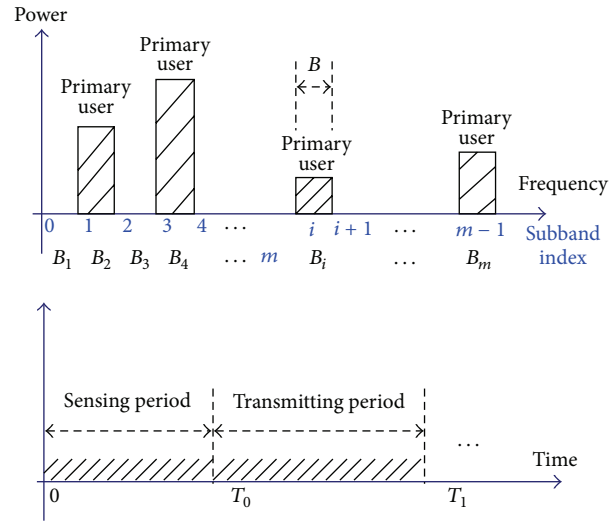


Figure 1: Wideband signal model.

where the $N \times 1$ vector r_f is the K-sparse frequency representation of r_t and K is the number of nonzero elements of r_f ($K \ll N$). F is an $N \times N$ DFT matrix, and ψ_i ($i = 1, 2, \ldots N$) is the similarly sampled basis function.

For Nyquist theorem N samples are necessary to exactly reconstruct the power spectrum density. As mentioned above, the CS is able to accurately reconstruct signals only with a small portion of samples with size of M ($M \ll N$)

$$y = \Phi r_t \tag{2}$$

in which y denotes an M-length measurement vector, and Φ is the measurement matrix. The spectrum of $r(t)$ can accurately be reconstructed when the measurement $\mathbf{y}$ is available. We aim at developing a spectrum sensing scheme with fewer nonadaptive measurements. Here, we use a modulated wideband converter (MWC) [11], which aims at sampling wideband sparse signal at a rate lower than that of Nyquist.

2.3. Reconstruction. Substituting (1) into (2), we can obtain

$$y = \Phi F^{-1} r_f. \tag{3}$$

The reconstruction of r_t could be resorted to the reconstruction of r_f

$$\hat{r}_f = \arg\min_{r_f} \|r_f\|_0 \qquad \text{s.t.} \left(\Phi F^{-1}\right) r_f = y. \tag{4}$$

It can be seen that (4) is a nonconvex problem. Equation (4) has a unique solution when the following holds:

$$\hat{r}_f = \arg\min_{r_f} \|r_f\|_1 \qquad \text{s.t.} \left(\Phi F^{-1}\right) r_f = y. \tag{5}$$

Actually, (5) is a second-order cone program. On the other hand, some variants of LASSO algorithm have been developed to deal with the noisy signals by minimizing the usual sum of squared errors:

$$\hat{r}_f = \arg\min_{r_f} \|r_f\|_1 \qquad \text{s.t.} \left\|\left(\Phi F^{-1}\right) r_f - y\right\|_2 < \varepsilon, \tag{6}$$

where bounds the noise in signals. A number of convex optimization software packages have been developed to solve the LASSO problem, for example, [12].

In our work, we use (6) to solve the reconstruction problem.

3. Comparison between DCT and DFT

For a normal signal, it is not difficult to find a sparse representation in a certain space, where ε bounds the noise in signals. Actually, signals involved in CRNs have been proved sparse in the frequency domain. So, it is possible to find the unoccupied spectrum in CRNs with compressed spectrum sensing with a rate lower than Nyquist.

For the consideration of better performance of energy concentration, we present an algorithm of DCT-based compressed spectrum sensing for the wideband frequency sparse signal.

The DCT sequence is represented by

$$V(k) = 2\sum_{n=0}^{N-1} x(n)\cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right], \quad k = 0, 1, \ldots, N-1. \tag{7}$$

The DFT sequence is represented by

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j(2\pi nk/N)} = \sum_{n=0}^{N-1} x(n)\, W_N^{kn}, \quad k = 0, 1, \ldots, N-1, \tag{8}$$

where by definition $W_N^{kn} = e^{-j2\pi/N}$.

Let $s(n)$ be a $2N$ point even symmetry extension of $x(n)$defined by

$$s(n) = \begin{cases} x(n), & 0 \le n \le N-1 \\ x(2N-n-1), & N \le n \le 2N-1. \end{cases} \tag{9}$$

The $2N$-point DFT of $s(n)$ is given by

$$S(k) = \sum_{n=0}^{2N-1} s(n)\, W_{2N}^{kn}, \quad k = 0, 1, \ldots, 2N-1. \tag{10}$$

Substituting (9) in (10) yields

$$S(k) = \sum_{n=0}^{N-1} x(n)\, W_{2N}^{kn} + \sum_{n=N}^{2N-1} x(2N-n-1)\, W_{2N}^{kn}. \tag{11}$$

If we change the second index of summation using $n = 2N-1-m$, we recall that $W_{2N}^{2mN} = 1$ for integer m, we factor out $W_{2N}^{-k/2}$, and we obtain

$$S(k) = W_{2N}^{-k/2}\sum_{n=0}^{N-1} x(n)\left[W_{2N}^{kn}W_{2N}^{k/2} + W_{2N}^{-kn}W_{2N}^{-k/2}\right], \quad k = 0, 1, \ldots, 2N-1. \tag{12}$$

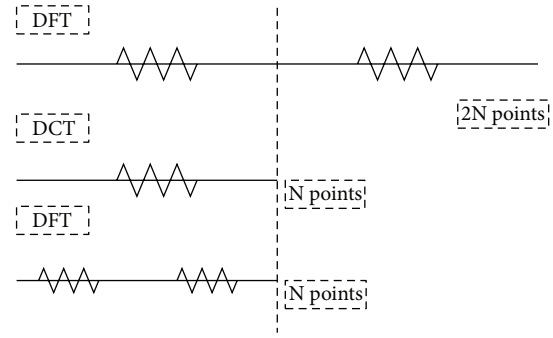


Figure 2: Comparison of coefficients of DCT and DFT.

The last expression may be written as

$$S(k) = W_{2N}^{-k/2} 2\sum_{n=0}^{N-1} x(n)\cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right], \quad k = 0, 1, \ldots, 2N-1. \tag{13}$$

Or equivalently

$$S(k) = W_{2N}^{-k/2} 2\,\mathrm{Re}\left[W_{2N}^{k/2}\sum_{n=0}^{N-1} x(n)\, W_{2N}^{kn}\right], \quad k = 0, 1, \ldots, 2N-1. \tag{14}$$

Substituting (7) in (13) yields,

$$S(k) = W_{2N}^{-k/2} V(k), \quad k = 0, 1, \ldots, N-1 \quad \text{or} \quad V(k) = W_{2N}^{k/2} S(k), \quad k = 0, 1, \ldots, N-1, \tag{15}$$

$$V(k) = 2\,\mathrm{Re}\left[W_{2N}^{k/2}\sum_{n=0}^{N-1} x(n)\, W_{2N}^{kn}\right], \quad k = 0, 1, \ldots, N-1. \tag{16}$$

Re[□] implies the real part of the term enclosed.

DCT of $x(n)$ can be computed by taking the $2N$-point DFT of $s(n)$, as in (9), and multiplying the result by $W_{2N}^{k/2}$, as in (15). Another approach is to take the $2N$-point DFT of $x(n)$ with N zeros appended to it, multiply the result by $W_{2N}^{k/2}$, and then take twice the real part. We note that $V(k)$ is real and $S(k)$ is complex.

As shown in Figure 2, we can easily find that the sparsity of DCT is half of the DFT.

As shown in Figure 3, we also compare the sparsity of our wideband signal between DCT and DFT. The wideband is only occupied by four channels, each of which is modulated by BPSK mode. From Figure 3, we apparently prove the validity of the conclusion we obtained above.

As discussed above, DCT has more superiorities than DFT in three aspects.

(1) Transform energy compaction capability means the capability of the transform to redistribute signal energy into small number of transform coefficients.

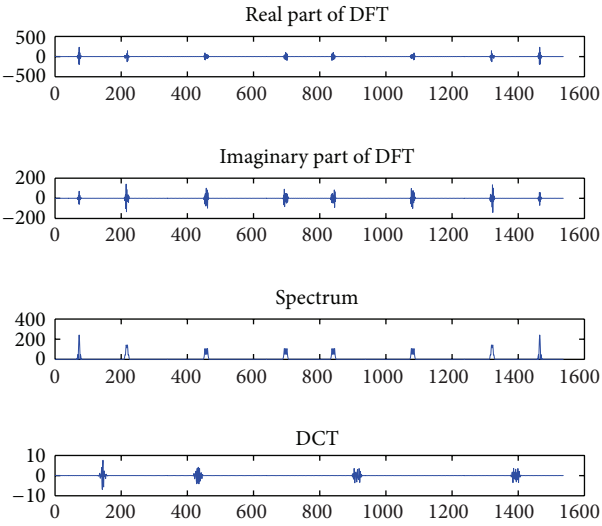


Figure 3: Comparison of coefficients of DCT and DFT for wideband signal.

The DCT basis has better spectral compaction and energy concentration properties than DFT. That is, signal representation is much sparser in DCT domain than in DFT. This, in turn, leads to improved performance with reconstruction accuracy and can result in released computation complexity.

(2) On the other hand, the DCT uses only real arithmetic, as opposed to the complex-valued DFT. Times of multiplication needed in real arithmetic account at most half of complex arithmetic. This reduces the signal-processing complexity/power consumption, especially for real-valued signal samples.

(3) Both of the two points mentioned above reduced computation complexity, which in other words means less processing time. This is an important parameter in scenarios with strict time limitation, such as dynamic spectrum access.

4. DCT Based Compressed Spectrum Sensing

As discussed in the preceding section, the signal response is sparse in DCT domain, so the DCT-CSS problem can be solved with a three-step scheme: (1) use compressed measurements $\mathbf{y}$ to estimate the sparse sequence $\hat{r}_d$, (2) reconstruct signal $\hat{r}_t$ according to $\hat{r}_d$, which can be done by an inverse DCT transfer, and (3) get frequency response $\hat{r}_f$ from $\hat{r}_t$ via a fast Fourier transform (FFT).

The mathematical description of the DCT-CSS scheme is similar to the DFT-CSS scheme with Fourier matrices being replaced by the DCT matrix:

$$y = \Phi D^{-1} r_d, \tag{17}$$

where $r_d = D r_t$ is the representation of r_t in DCT domain.

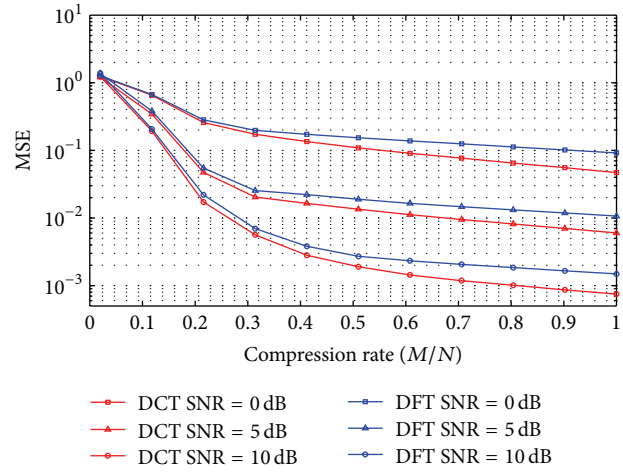


Figure 4: MSE performance of DCT-CSS and DFT-CSS.

Similar to (6), we can get the estimate response $\hat{r}_d$ from

$$\hat{r}_d = \arg\min_{r_d} \|r_d\|_1 \qquad \text{s.t.} \left\| \left(\Phi F^{-1}\right) r_d - y \right\|_2 < \varepsilon. \tag{18}$$

We can easily get $\hat{r}_t$ from the inverse transform $\hat{r}_t = D^{-1}\hat{r}_d$ and finally get frequency estimate

$$\hat{r}_f = \mathrm{FFT}\left(\hat{r}_t\right). \tag{19}$$

5. Simulation

In this section, we evaluate the performance of the proposed DCT-CSS scheme. We consider that a wideband spectrum occupies 512 MHz band, which is divided into 16 subbands. We use 8 MHz BPSK modulated signal to be active signal in each band. The received signal is corrupted by additive white Gaussian noise (AWGN). On average, four subbands are occupied simultaneously, while the locations and amplitudes change for several time bursts.

We compare the normalized MSE of the estimated spectrum, which is defined as

$$\mathrm{MSE} = E\left\{ \frac{\left\|\hat{r}_f - r_f\right\|_2^2}{\left\|r_f\right\|_2^2} \right\}. \tag{20}$$

From Figure 4, we can see that DCT-CSS outperforms DFT-CSS in terms of MSE of recovery as the compression rate varies from 0.05 to 1. As compression rate increases, recovery accuracy improves. On the other hand, different SNR leads to different recovery accuracy. Higher SNR means higher accuracy.

Figure 5 compares probability of detection as the compression rate varying from 0.02 to 0.3. As compression rate increases above 0.25, probability of detection of both the two schemes reaches 1 and goes to balance. However, during the region 0.02 to 0.22, DCT-CSS performs always better than DFT-CSS.

As shown in Figure 6, we compare the computation complexity of the two schemes. Processing time increases

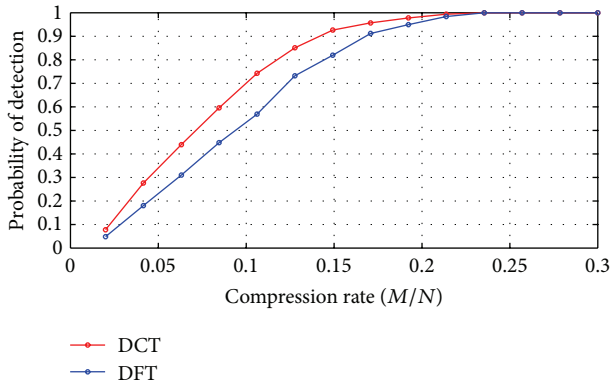


Figure 5: Probability of detection for DCT-CSS and DFT-CSS.

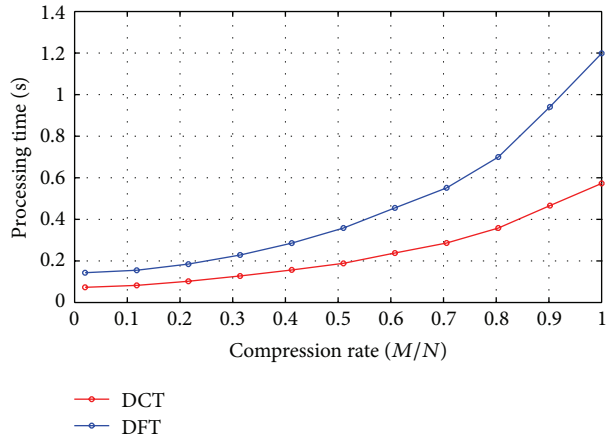


Figure 6: Processing time of DCT-CSS and DFT-CSS.

as the compression rate varies from 0.02 to 1. However, the processing time of DCT-CSS is approximately half of the DFT-CSS.

6. Conclusion

In this paper, we have proposed a novel DCT-CSS scheme for wideband spectrum sensing. Analysis has verified that wideband spectrum signal is sparser in DCT domain than in DFT. Simulation results have shown that DCT-CSS can not only improve the reconstruction accuracy and probability of detection but also save processing time. Our future work will focus on DCT-CSS scheme implemented by other reconstruction algorithms.

Conflict of Interests

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

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61032004) and the National High Technology Research and Development Program of China ("863" Program) (2012AA121605, 2012AA01A503, and 2012AA01A510).

References

[1] FCC, "Spectrum policy task force report," Tech. Rep. ET Docket No. 02-155, 2002.

[2] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey," *Computer Networks*, vol. 50, no. 13, pp. 2127–2159, 2006.

[3] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[4] Z. Tian and G. B. Giannakis, "Compressed sensing for wideband cognitive radios," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, pp. IV1357–IV1360, Honolulu, Hawaii, USA, April 2007.

[5] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: from theory to applications," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4053–4085, 2011.

[6] G. Leus and D. D. Ariananda, "Power spectrum blind sampling," *IEEE Signal Processing Letters*, vol. 18, no. 8, pp. 443–446, 2011.

[7] Z. Tian, Y. Tafesse, and B. M. Sadler, "Cyclic feature detection with sub-nyquist sampling for wideband spectrum sensing," *IEEE Journal on Selected Topics in Signal Processing*, vol. 6, no. 1, pp. 58–69, 2012.

[8] G. Peyré, "Best basis compressed sensing," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2613–2622, 2010.

[9] L. R. Amado, E. S. C. Losqui, F. P. V. de Campos, A. A. M. de Medeiros, and M. V. Ribeiro, "Spectrum sensing for powering power line communications," *Simpósio Brasileiro de Telecomunicaçoes*. In press.

[10] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*, Academic Press, New York, NY, USA, 1990.

[11] M. Mishali and Y. C. Eldar, "From theory to practice: sub-Nyquist sampling of sparse wideband analog signals," *IEEE Journal on Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 375–391, 2010.

[12] "SparseLab toolbox," http://sparselab.stanford.edu.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 984375, 6 pages
http://dx.doi.org/10.1155/2014/984375

Research Article

Classification Based on Pruning and Double Covered Rule Sets for the Internet of Things Applications

Shasha Li, Zhongmei Zhou, and Weiping Wang

Department of Computer Science and Engineering, Minnan Normal University, Zhangzhou 363000, China

Correspondence should be addressed to Shasha Li; shasha6320@163.com

Received 14 October 2013; Accepted 25 November 2013; Published 5 January 2014

Academic Editors: W. Sun, G. Zhang, and J. Zhou

The Internet of things (IOT) is a hot issue in recent years. It accumulates large amounts of data by IOT users, which is a great challenge to mining useful knowledge from IOT. Classification is an effective strategy which can predict the need of users in IOT. However, many traditional rule-based classifiers cannot guarantee that all instances can be covered by at least two classification rules. Thus, these algorithms cannot achieve high accuracy in some datasets. In this paper, we propose a new rule-based classification, CDCR-P (Classification based on the Pruning and Double Covered Rule sets). CDCR-P can induce two different rule sets A and B. Every instance in training set can be covered by at least one rule not only in rule set A, but also in rule set B. In order to improve the quality of rule set B, we take measure to prune the length of rules in rule set B. Our experimental results indicate that, CDCR-P not only is feasible, but also it can achieve high accuracy.

1. Introduction

The Internet of things is one of the hot topics in recent years. It has integrated many kinds of modern technology. By these kinds of technology, it produces large-scale data in IOT. In order to handle these large data, it requires techniques and methods of data mining and machine learning [1–6].

As one of the most important tasks of data mining, classification has been widely applied in IOT. The main idea of classification is that builds classification rules. According these rules, we can predict the class label for unknown objects.

Traditional rule-based classifications usually use greedy approach, such as FOIL [7], CPAR [8], and CMER [9]. These methods repeatedly search for the current best one rule or best-k rules and remove examples covered by the rules. They cannot guarantee that all instances can be covered by at least two classification rules. As a result, some traditional classifiers have less classification rules. Their accuracy may not be high. Decision tree classifiers produce classification rules by constructing classification trees, such as ID3 [10], C4.5 [11], and TASC [12]. The process of building a decision tree does not need to delete any examples. All examples can find only one matching rule in the classification rule set. That is why decision trees often generate small rule sets and cannot achieve high accuracy in some data.

Aiming at these weaknesses, we propose a novel Double Covered Rule sets classifier called CDCR-P (Classification based on the Pruning and Double Covered Rule sets). CDCR-P generates two different rule sets A and B and then prunes the rule set B. Each instance can be covered by at least one rule from rule set A. At the same time, each instance can be covered by at least one rule from rule set B. CDCR-P has four aspects. First, CDCR-P generates rule set A. We select several best values which can just cover the training set to construct a candidate set. CDCR-P employs candidate set to produce rule set A. Second, in order to induce rule set B, we remove the values of candidate set in training data and select other several best values to induce rule set B. Rule set A is fully different from rule set B. Third, each instance can find at least two matching rules. One of the rules is from rule set A, and another is from rule set B. Forth, we prune the length of rules in rule set B, so as to improve the quality of rule set B. Our method has the following advantages.

(1) CDCR-P can produce two rule sets. Thus, CDCR-P can generate large number of classification rules.

```
Input: Training data T = {t1, t2, ..., tn}
Output: T1, candidate value CV1
Method:
(1) T1 = φ, CV1 = φ, Length = 0;
(2) compute the information gain of each sample s;
(3) sort s according to the information gain in descending order. If two sample have the
    same information gain, sort the two sample according to the support;
(4) Length = T.Count × 1/3;
(5) While T ≠ φ
(6)     S = S − {a}, where a is the first element of S;
(7)     for each tuple t in T
(8)         While t.Contains(a) && T1.Count <= Length
(9)             T1 = T1 ∪ {t};
(10)            CV1 = CV1 ∪ {a};
(11)            T = T − {t};
(12)        end while
(13)    end for
(14) end while
(15) return T1, CV1
```

Algorithm 1: Dividing training set into three smalldatasets.

(2) All instances in training set can be matched by at least two classification rules.

(3) CDCR-P can achieve high accuracy by combining rule set A with rule set B.

The paper is organized as follows. In Section 2, we introduce the method of CVCR (Classification based on Value Covered Rules). In Section 3, we propose a new classifier CDCR-P and discuss how to use CDCR-P to classify new objects. We report our experimental results in Section 4. We finally conclude our study in Section 5.

2. Classification Based on Value Covered Rules

In this section, we introduce the method of value covered classifier; this method is called CVCR (Classification based on Value Covered Rules).

Suppose $T = \{t_1, t_2, \ldots, t_n\}$ is a set of tuples. Each tuple t has m attributes $\{A_1, A_2, \ldots, A_m\}$. Let C be a finite set of class labels $\{C_1, C_2, \ldots, C_k\}$ and S be a set consisting of s data samples. A rule r consists of several samples s and a class label c, which takes the form of $s_1 \wedge s_2 \wedge \cdots \wedge s_l \rightarrow c$. One rule set is formed by a lot of rules which are extracted from one classifier. If tuple t satisfies $s_1 \wedge s_2 \wedge \cdots \wedge s_l$ from rule r, the t is matched by r. r predicts that t belongs to class c.

Definition 1 (information gain). Let s_i be the number of samples of S in class C_i. The information gain of an attribute value is denoted by $I(s_1, s_2, \ldots, s_k)$ and is defined as follows:

$$I(s_1, s_2, \ldots, s_k) = -\sum_{i=1}^{k} p_i \log_2(p_i), \quad (1)$$

where p_i is the probability that a literal belongs to class label c_i. p_i is estimated by s_i/s.

CVCR finds a set of values which can cover all the training set. The process of constructing CVCR is as follows.

First, CVCR sorts all literals according to the information gain in a descending order and selects several best attribute values $v_1, v_2, \ldots, v_i$ which can just cover the training set T. $v_1, v_2, \ldots, v_i$ construct a candidate set. Let these values split T to subdatasets $t_1, t_2, \ldots, t_i$, respectively. Second, CVCR connects v_i with attribute values $v_{i1}, v_{i2}, \ldots, v_{ij}$ which can just cover dataset t_i to produce patterns. Finally, repeat the above steps until the information gain of each pattern is equal to 0.

The experimental results of CVCR are shown in Table 2. The experimental results show that CVCR can achieve higher accuracy than ID3 and FOIL. Because CVCR contains the global optimal attribute values, CVCR is more feasible than ID3. However, CVCR still produces less classification rules, which cannot guarantee that each instance can be matched by at least two rules.

3. Classification Based on Pruning and Double Covered Rule Sets

In this section, we produce a new method CDCR-P. First, we show the process of how to induce rule sets A and B. Second, we describe the method of how to prune rule set B. Finally, we give the way of how to use the two rule sets A and B to classify new objects.

3.1. Constructing Rule Sets A and B. Based on the idea of CVCR, we continue mining knowledge in-depth. This approach divides the training set T into three small datasets T_1, T_2, and T_3 according to candidate set. The method contains four steps. Step 1, we select several best attribute values $v_1, v_2, \ldots, v_k$ from candidate set which can just cover one-third of training set. Attribute values $v_1, v_2, \ldots, v_k$ have less information gain. The tuples which contain one of $v_1, v_2, \ldots, v_k$ form the small dataset T_1. The process is shown as Algorithm 1. We form T_2, T_3 using the same way as T_1.

```
Input: Dataset D
Output: A set of cover value cv
Method: find a set of cover value cv that can cover D
(1) data set d = D; cover value cv = φ
(2) compute the information gain of each sample s in D;
(3) sort s in according to the information gain in descending order. If two sample have
      same information gain, sort the two sample according to the support;
(4) while d ≠ φ
(5)     S = S − {a}; where a is the first element of S;
(6)     cv = cv ∪ {a};
(7)     remove from d that all examples contain a;
(8) end while
(9) return cv
```

Algorithm 2: Finding covered values in dataset.

```
Input: Training data T_i, CV_i
Output: Rule set A
Method:
(1) Rule set R = φ, dataset = φ, cover value cv = φ, IniteQueue Q;
(2) while CV_i = φ
(3)     Q . push(a) where a is first element of CV_i;
(4)     CV_i = CV_i − {a};
(5) end while
(6) while !Q . empty()
(7)     pattern x = Q . front(); Q . pop();
(8)     if x . length > max _rule_length continue;
(9)     compute the information gain of x;
(10)    if x . information gain == 0
(11)        R = R ∪ {x};
(12)    else
(13)        dataset = dataset ∪ each t . contains(x);
(14)        found cv in dataset;
(15)        connect x with cv;
(16)        Q . push();
(17)        dataset = φ, cv = φ;
(18)    end if
(19) end while
(20) return R
```

Algorithm 3: Inducing rule set A.

Step 2, according to $v_1, v_2, \ldots, v_k$, T_1 is split into datasets $t_1, t_2, \ldots, t_n$. We find cv_i (a set of cover values) from t_i on the basis of information gain. The measure of cv_i is the same as CVCR, shown as Algorithm 2. CDCR connects v_i with cv_i to produce patterns. If the information gain of pattern is equal to 0, $X \rightarrow C$ belongs to rule set A, shown as Algorithm 3. Step 3, CDCR recalculates the CV (cover values) in T_1 excluding $v_1, v_2, \ldots, v_k$. CV splits T_1 into some datasets and connects covered value in each dataset to produce new rules. These rules belong to rule set B, shown as Algorithm 4. Finally, we remove T_1 from T and iterate the process until T_2, T_3 are trained. Rule set A is the same as CVCR. Both rule sets A and B belong to CDCR.

3.2. Pruning Rule Set B. In order to improve the quality of rule set B, we introduce a new method CDCR-P (Classification based on the Pruning and Double Covered Rule sets).

Definition 2 (confidence). The confidence of sample X is defined as follows:

$$\text{conf}(X) = \frac{\text{count}(X_c)}{\text{count}(X)} \times 100\%, \tag{2}$$

where $\text{count}(X_c)$ means the number of tuples which contain sample X in class c.

The confidence of rules that CDCR generated is equal to 100%. We modify the length of rule set B. The rules are

```
Input: Training data T_i, CV_i
Output: Rule set B
Method:
(1) Rule set R = φ, dataset = φ, cover value cv = φ, candidate set cs = φ, IniteQueue Q;
(2) add cover value which can cover T_i to cs, and cs ∉ CV_i
(3) while cs ≠ φ
(4)       Q . push(a) where a is first element of cs;
(5)       cs = cs − {a};
(6) end while
(7) while !Q . empty()
(8)        x = Q . front(); Q . pop()
(9)        If x . length > max _rule_length continue;
(10)      compute the information gain of x;
(11)      if x . information gain == 0
(12)         R = R ∪ {x};
(13)      else
(14)         dataset = dataset ∪ each t . Contains(x);
(15)         Found cv in dataset, and cv ∉ CV_i;
(16)         connect x with cv;
(17)         Q . push();
(18)         dataset = φ, cv = φ;
(19)      end if
(20) end while
(21) return R
```

Algorithm 4: Inducing rule set B.

generated when the confidence is 100% in the small dataset T_i instead of in the whole training set T. Each rule is marked with the confidence in T. Thus, rule set B in CDCR-P is shorter than rule set B in CDCR.

3.3. Classifying Unknown Examples. In this part, we give the method of how to use CDCR and CDCR-P to classify unknown instances.

Definition 3 (support). The support of sample X is denoted by

$$\sup(X) = \frac{\text{count}(X)}{|T|} \times 100\%, \tag{3}$$

where count(X) means the number of tuples which contain sample X. $|T|$ is the number of tuples in training data.

When testing unknown examples, CDCR selects the matched rule with the highest support. If some rules have the same support, we select the maximum number of matched rules in each class.

CDCR-P first considers the rule with the highest confidence. If two rules have the same confidence, CDCR-P sorts the two rules according to the support.

Definition 4 (missing match rate). If the test instance cannot find any match rule, this unclassified instance is considered mismatch. The missing match rate is defined as

$$\frac{\text{count (unclassified instance)}}{|T|} \times 100\%, \tag{4}$$

where count (unclassified instance) means the number of tuples which cannot be matched by rules.

Table 1: Characteristics of UCI datasets.

Dataset	No. of instances	No. of attributes	No. of classes
Balance	625	5	3
Breast	699	10	2
Car	1728	7	4
Lymph	148	18	4
Monks	432	7	2
Mushroom	400	23	2
Soybean	307	36	19
SPECT	267	23	2
Tic-tac	958	10	2
Zoo	101	16	7
Cleve	303	13	2
Heart	270	13	2
Iris	150	4	3
Wine	178	14	3

4. Experiments

We show the experimental results in 14 UCI datasets. The character of each data is shown in Table 1. All the experiments are performed on a 2.2 GHz PC with 2.84 G main memory, running Microsoft Windows XP. Experiments run tenfold cross validation method for each data.

Table 2: The accuracy of ID3, FOIL, CVCR, CDCR, and CDCR-P.

Dataset	ID3	FOIL	CVCR	CDCR	CDCR-P
Balance	0.3716	0.4929	0.7810	0.7985	0.8289
Breast	0.9042	0.9342	0.9571	0.9556	0.9585
Car	0.7298	0.7714	0.8837	0.9248	0.8767
Lymph	0.7148	0.7424	0.8181	0.81	0.81
Monks	0.9448	0.8146	0.8959	0.9514	0.9537
Mushroom	0.985	0.995	0.99	0.99	0.99
Soybean	0.4102	0.4172	0.8180	0.8276	0.8601
SPECT	0.7181	0.752	0.7303	0.7453	0.8053
Tic-tac	0.8215	0.9875	0.8684	0.9541	0.9530
Zoo	0.97	0.9409	0.9709	0.9609	0.8918
Cleve	0.7426	0.7423	0.8152	0.8216	0.8482
Heart	0.8148	0.8148	0.7556	0.7852	0.7963
Iris	0.7733	0.9533	0.8133	0.8133	0.8533
Wine	0.96	0.9379	0.983	0.9712	0.9882
Average	0.7758	0.8069	0.8629	0.8793	0.8867

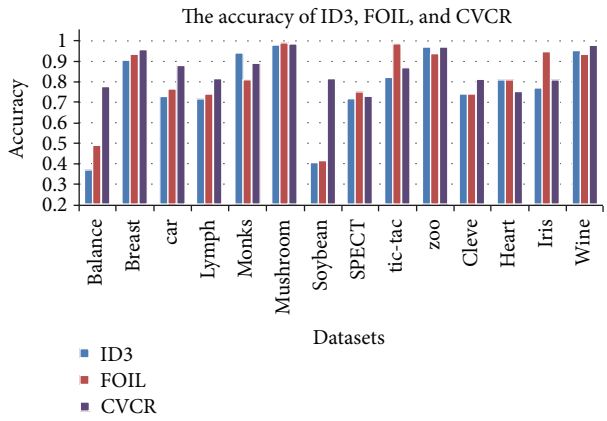


Figure 1: The accuracy of ID3, FOIL, and CVCR.

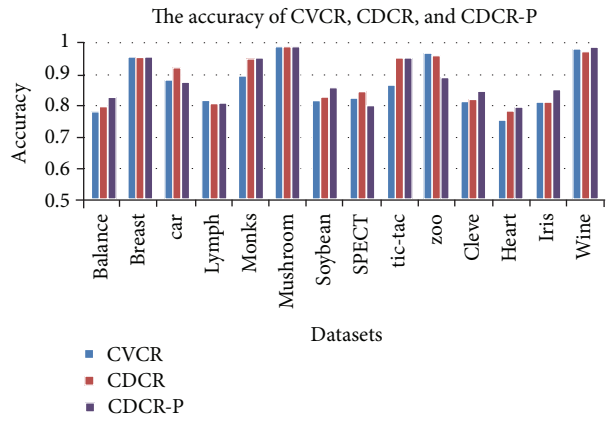


Figure 2: The accuracy of CVCR, CDCR, and CDCR-P.

In Table 2, we give the accuracy of ID3, FOIL, CVCR, CDCR, and CDCR-P. Figure 1 gives the accuracy of ID3, FOIL, and CVCR. CVCR employs the idea of covered values; these cover values are the global optimal attribute values in training data. From Figure 1 and Table 2 we can see that CVCR can achieve higher accuracy than ID3 and FOIL. Figure 2 gives the accuracy of CVCR, CDCR, and CDCR-P. CDCR not only uses the method of covered values, but also produces two rule sets A and B. Each instance can be matched at least by one rule from rule set A and rule set B. From Figure 2 and Table 2 we can see that CDCR can achieve higher accuracy than CVCR. Based on all advantages of CVCR and CDCR, CDCR-P take measure to prune the length of rule set B. The experimental results show that CDCR-P has the highest accuracy.

Table 3 displays the missing match rate of ID3, FOIL, CVCR, CDCR, and CDCR-P. CVCR can produce more rules than ID3 and FOIL. From Table 3 we can see that the missing match rate is decreased obviously by CVCR. CDCR produces two rule sets. Therefore, CDCR produces more rules than CVCR. From Table 3 we can see that the missing match rate of CDCR is lower than CVCR. CDCR-P modifies the length of rule set B; the quality of rules in CDCR-P is higher than CDCR. The experiments indicate that the mismatch rate of CDCR-P is the lowest.

Through all the above experimental results, we can conclude the following. (1) It is necessary for us to construct two rule sets. (2) It is necessary to prune rule set B. (3) CDCR-P can achieve high accuracy and has an excellent result in missing match rate.

5. Conclusions

Classification has been widely applied in IOT. The accuracy of classification is an important factor in classification task. The traditional rule-based classifications cannot guarantee that all test cases can be matched by two rules. They usually generate less classification rules. Thus, the accuracy of these algorithms may be low in some data. In this paper, a novel approach CDCR-P is proposed. CDCR-P generates two rule

Table 3: The missing match rate of ID3, FOIL, CVCR, CDCR, and CDCR-P.

Dataset	ID3	FOIL	CVCR	CDCR	CDCR-P
Balance	0.422	0.3518	0.0910	0.0799	0.0016
Breast	0.0529	0.0443	0.0029	0	0
Car	0.2165	0.1945	0.0231	0.0156	0
Lymph	0.0624	0.1014	0.0267	0	0
Monks	0	0.1023	0.0486	0.0486	0
Mushroom	0.005	0.005	0	0	0
Soybean	0.0559	0.2787	0.0097	0	0
SPECT	0.0566	0.0634	0.1387	0.1313	0
Tic-tac	0.0366	0.0094	0.0052	0	0
Zoo	0.01	0.0591	0	0	0
Cleve	0.0231	0.0795	0.0298	0.0265	0
Heart	0	0.0148	0.0704	0.0444	0.0111
Iris	0.1733	0.0067	0.1533	0.1533	0.06
Wine	0	0.0454	0.0056	0	0
Average	0.0796	0.0969	0.0432	0.0357	0.0052

sets: rule set A and rule set B. All instances can be matched by at least one rule not only in rule set A, but also in rule set B. This method greatly increases the number of extracted rules. Thus, it gets more information from training data. Our experimental results show that the methods of CDCR-P can produce more rules and achieve high accuracy. In future research, we will perform an in-depth study on combining distributed data mining with IOT in order to improve the efficiency of CDCR-P.

Conflict of Interests

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

Acknowledgments

This work is funded by China NFS program (no. 61170129), Fujian Province NSF Program (no. 2013J01259), and Minnan Normal University Postgraduate Education Project (no. 1300-1314).

References

[1] L. Hu, Z. Zhang, F. Wang, and K. Zhao, "Optimization of the deployment of temperature nodes based on linear programing in the internet of things," *Tsinghua Science and Technology*, pp. 250–258, 2013.

[2] M. A. Feki, F. Kawsar, M. Boussard, and L. Trappeniers, "The internet of things: the next technological revolution," *IEEE Computer*, vol. 46, no. 2, pp. 24–25, 2013.

[3] W. Zhu, J. Yu, and T. Wang, "A Security and privacy model for mobile RFID systems in the internet of things," in *Proceedings of the IEEE 14th International Conference on Communication Technology (ICCT '12)*, pp. 726–732, November 2012.

[4] C. K. Tham and T. Luo, "Sensing-driven energy purchasing in smart grid cyber-physical system," *IEEE Transactions in Systems, Man and Cybernetics A*, vol. 43, no. 4, pp. 773–784, 2013.

[5] X. Xing, J. Wang, and M. Li, "Services and key technologies of the internet of things," *ZTE Communications*, no. 2, pp. 26–29, 2010.

[6] Z. Qureshi, J. Bansal, and S. Bansal, "A survey on association rule mining in cloud computing," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 4, pp. 318–321, 2013.

[7] J. R. Quinlan and R. M. Cameron-Jones, "FOIL: a midtern report," in *Proceedings of the European Conference Machine Learning*, pp. 3–20, Vienna, Austria.

[8] X. Yin and J. Han, "CPAR: classification based on predictive association rules," in *Data Mining, The SIAM (Society for Industrial and Applied Mathematics) International Conference*, May 2003.

[9] X. Wang, Z. Zhou, and G. Pan, "CMER: classification based on multiple excellent rules," *Journal of Theoretical and Applied Information Technology*, pp. 661–665, 2013.

[10] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[11] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Series in Machine Learning, Kaufmann Academic, 1993.

[12] Y.-L. Chen, W.-H. Hsu, and Y.-H. Lee, "TASC: two-attribute-set clustering through decision tree construction," *European Journal of Operational Research*, vol. 174, no. 2, pp. 930–944, 2006.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 378083, 7 pages
http://dx.doi.org/10.1155/2013/378083

Research Article

Optimal Transport on Weighted Networks for Different Node Delivery Capability Schemes

Fei Shao[1,2]

[1] *Jiangsu Information Analysis Engineering Laboratory, Jinling Institute of Technology, Nanjing 211169, China*
[2] *School of Information Technology, Jinling Institute of Technology, Nanjing 211169, China*

Correspondence should be addressed to Fei Shao; shaofei@jit.edu.cn

Received 5 September 2013; Accepted 9 December 2013

Academic Editors: W. Sun, G. Zhang, and J. Zhou

Many real networks can be best described by weighted networks with a diversity of interactions between nodes measured by the weights of the edges. It is of great importance to improve the overall capacity of these real-world networks. In this paper, the traffic capacity of weighted network is investigated based on three different node delivery capability schemes: the delivery capacity of each node is constant in the first scheme while in the second and third schemes it is proportional to its node degree and node strength. It is shown by simulations that the network transfer capacity depends strongly on the tunable parameter. And different tunable parameter is suitable for different node delivery capability.

1. Introduction

The past few years have witnessed the emergence of interest in the network topological structure since the seminal study on small world networks by Watts and Strogatz [1] and on scale-free networks by Barabási and Albert [2]. Lots of biological, social, and communication systems can be viewed as complex networks while nodes represent individuals and edges represent the relationships between them. The previous studies on networks have been primarily focused on unweighted networks, edges between nodes are either present or not, represented as binary states. However, the gathering complete data of real networks display the variation of the strength of the edges between nodes, providing a more complete representation of real network structures. The analyses of the mobile communication networks [3], the scientific collaboration networks [4], the cellular metabolism [5], the world-wide airport network [6], and the Internet [7] have revealed that networks are not only specified by their topology but also by the dynamics of weight (such as the capacity and the intensity) taking place along the edges.

Lots of models have been presented to describe those real-world networks. The YJBT model [8] is a model for weighted scale-free network whose topology is the same as that of the BA model. Moreover, both the topology and the weights are driven by the connectivity according to the preferential attachment rule. The ZTZH model [9] is a generalization of the YJBT model incorporating a stochastic scheme for weight assignments based on both the degree and the fitness of node. And the AK model [10], in which the structural growth of the network is coupled with the edge weights, focus on a strength driven attachment instead of degree preferential attachment. In those models discussed above, the weight of edge is assigned when the edge is added and remains fixed thereafter. But actually, the addition of new nodes or edges will affect the weight of the existing edges. The BBV model [11] is proposed to describe the weights' reinforcement phenomenon which is triggered only by new added nodes. Wang model [12] is a traffic-driven evolution model where weights of edges between the existed nodes will also be rearranged.

Recently, the study of the network overall transfer capacity is becoming increasingly important due to the constantly growing significance of large communication networks such as the Internet. Finding optimal routing strategies to improve the transfer capacity is gaining increasing concern. Some are based on global information: the shortest path routing strategy [13], the efficient path routing strategy [14], and the generalized minimum information path routing strategy [15]; some are based on local information [16–20].

In this paper, we propose a novel routing strategy in which packets are transferred through the path based on the weight of edges with a tunable parameter α. To maximize the overall network transfer capacity which can be measured by the critical packet generating rate R_c, the optimal tunable parameter α is achieved.

This paper is organized as follows. In Section 2 we describe the model and our routing strategy, followed by the experimental evaluations on BBV weighted networks and real world network in Section 3. The conclusions are given in Section 4.

2. Model

In BBV networks, the topological as well as weighted properties can be completely described by a weighted adjacency matrix **W**, whose elements w_{ij} denote the weight of the edge between node i and j. The definition of the BBV network is based on two coupled mechanisms: the topological growth and the weight dynamics, which is same as BA network.

(i) Growth. Starting from an initial small number of N_0 nodes connected by edges with assigned weight w_0, a new node is added at every time step. The new added node is connected to m different previously existing nodes with equal weight w_0 for every edge and chooses preferentially nodes with large strength according to the probability $\prod_{n\to i} = s_i / \sum_l s_l$, where s_i is the node strength described as $s_i = \sum_j w_{ij}$.

(ii) Weight dynamics. The weight of each new add edge is initially set to a given value w_0 which is often set to 1 for simplicity. But the adding of edge connecting to node i will result in increasing the weight of the other edges linked to node i which is proportional to the edge weights. If the total increase is δ (we will focus on the simplest form: $\delta_i = \delta$), we can get

$$w_{ij} = w_{ij} + \Delta w_{ij} = w_{ij} + \delta * \frac{w_{ij}}{s_i}. \tag{1}$$

This will yield the strength increase of node i as:

$$s_i = s_i + \delta + w_0. \tag{2}$$

The degree distribution of BBV network $P(k) \propto k^{-\gamma_k}$ and the strength distribution $P(s) \propto s^{-\gamma_s}$ yield scale-free properties with the same exponent [6, 11, 21, 22]:

$$\gamma_k = \gamma_s = \frac{4\delta + 3}{2\delta + 1} = 2 + \frac{1}{2\delta + 1}. \tag{3}$$

Denoting $P_{i\to j}$ as the path between nodes i and j which pass through the nodes sequence $x_0(= i), x_1, x_2, \ldots, x_{n-1}, x_n(= j)$, we define

$$F\left(P_{i\to j}, \alpha\right) = \sum_{i=0}^{n-1} {w_{ij}}^{\alpha}. \tag{4}$$

In our routing strategy, we specify the routing path between i and j as the one that makes $F(P_{i\to j}, \alpha)$ minimum under a given tunable parameter α.

In this paper, the simplest transfer model can be described as follows.

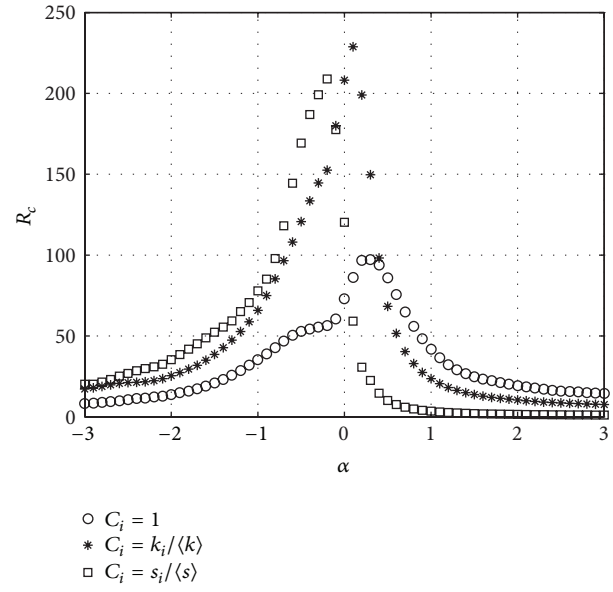


Figure 1: R_c versus α. BBV network with $n = 1000$, $\delta = 6$, $m = 6$, and $\omega_0 = 1$.

(1) All the nodes are treated as both hosts and routers. A host can create packets with addresses of destination and receive packets from other hosts while a router routes the data packets to their destinations.

(2) At each time step t, there are R packets generated in the whole network with randomly chosen sources and destinations. Once a packet is created, it is placed at the end of the queue if this node already has several packets waiting to be delivered to their destinations. The existing packets may be created at some previous time steps or they are transmitted from other nodes.

(3) At each time step, the first C_i packets at the top of the queue of each node i, if it has more than C_i packets in its queue, are forwarded one step toward their destinations and placed at the end of the queues of the selected nodes. Otherwise, all packets in the queue are forwarded one step. This procedure applies to every node at the same time.

(4) A packet, upon reaching its destination, is removed from the system.

In our model, three node delivery capability schemes are considered: (i) each node has the same packet delivery capability (C_i = const, CONC stands for this scheme); (ii) the node delivery capacity is considered to be proportional to the node degree k_i ($C_i \sim k_i$, DEGC stands for this scheme); (iii) the node delivery capacity is considered to be proportional to the node strength s_i ($C_i \sim s_i$, STRC stands for this scheme). To compare the overall transfer capacity, we assign the equal value to the total node delivery capability in three situations. In the last two schemes, we normalize the delivery capability of each node to set the total delivery capability of the whole network to be equal to the node number n, which is the same as the first case.

When R increases from zero to ∞, two phases will be observed: free flow for small R and congested phase for large

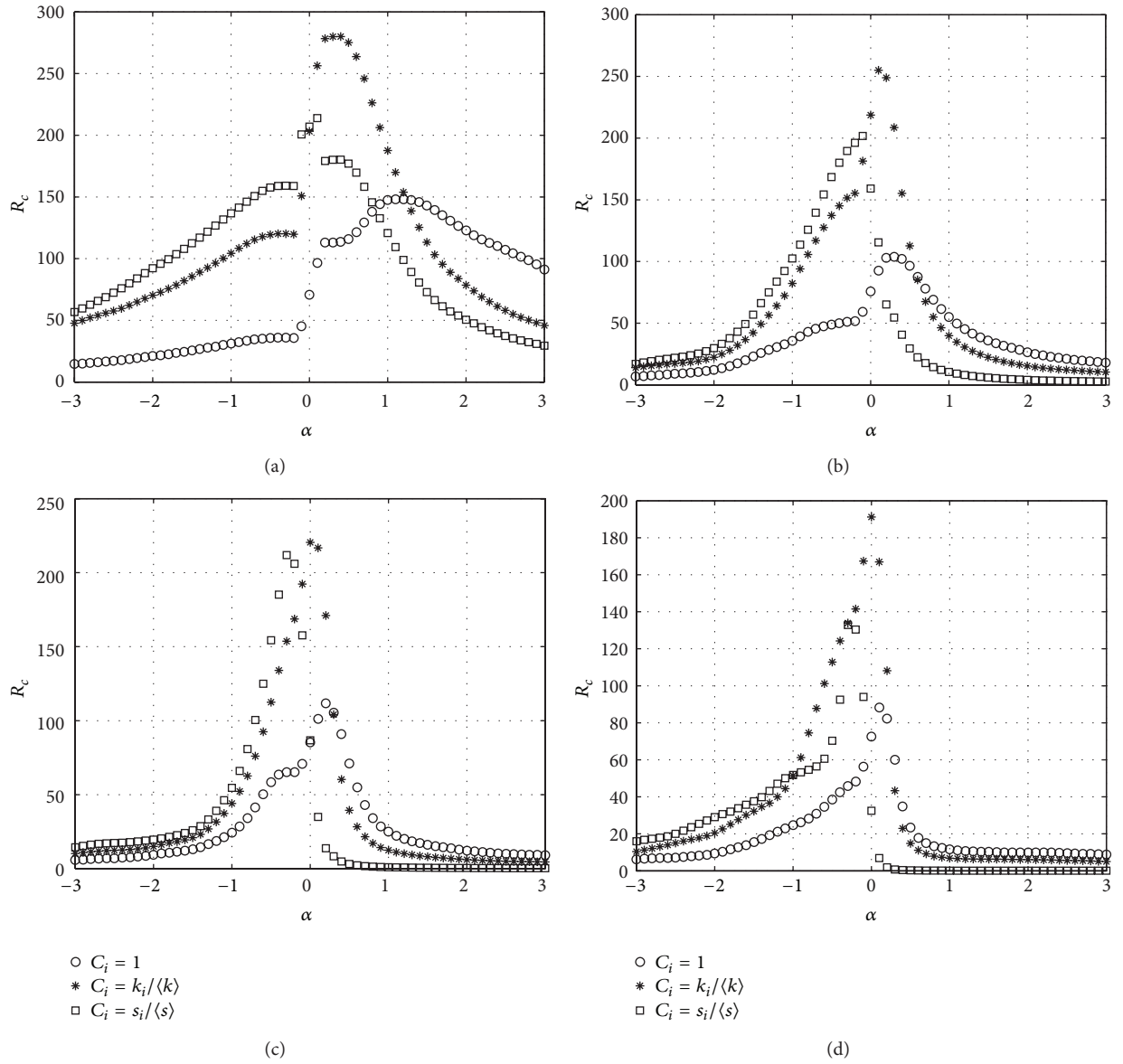


Figure 2: R_c versus α. BBV network with $n = 1000$, $m = 6$, and $\omega_0 = 1$. (a) $\delta = 0.6$, (b) $\delta = 3$, (c) $\delta = 12$, and (d) $\delta = 60$.

R. We focus on the critical value R_c where phase transit from the former to the latter which can best reflect the maximum information transfer capacity of a system. For $R < R_c$, the numbers of created and delivered packets are balanced, resulting in a steady free flow of traffic. For $R > R_c$, traffic congestion occurs as the number of accumulated packets increases with time, due to the fact that the capacities of nodes for delivering packets are limited. We are interested in determining critical value R_c in order to address which kind of routing strategy is more susceptible to phase transition and therefore traffic congestion.

We introduce the betweenness b_i to estimate the possible packet passing through a node i under a given routing strategy which is defined as

$$b_i = \sum_{s,t} \frac{\sigma(s,i,t)}{\sigma(s,t)}, \tag{5}$$

where $\sigma(s,i,t)$ is the number of paths under the given routing strategy between nodes s and t that pass through node i and $\sigma(s,t)$ is the total number of paths under the given routing strategy between nodes s and t and the sum is over all pairs s,t of all distinct nodes. The probability that a certain packet will pass through the node i is $b_i/\sum_{j=1}^{n} b_j$ (n is the total node number in the network). The average number of packets that the node i receives at each time step is $R * b_i/(n * (n-1))$. Congestion occurs when the number of incoming packets is equal to or larger than the outgoing packets that the node i can transfer at one time step; that is, $R * b_i/(n * (n-1)) \geq C_i$. So the critical packet generating rate R_c is

$$R_c = \min\left(C_i * n * \frac{n-1}{b_i}\right). \tag{6}$$

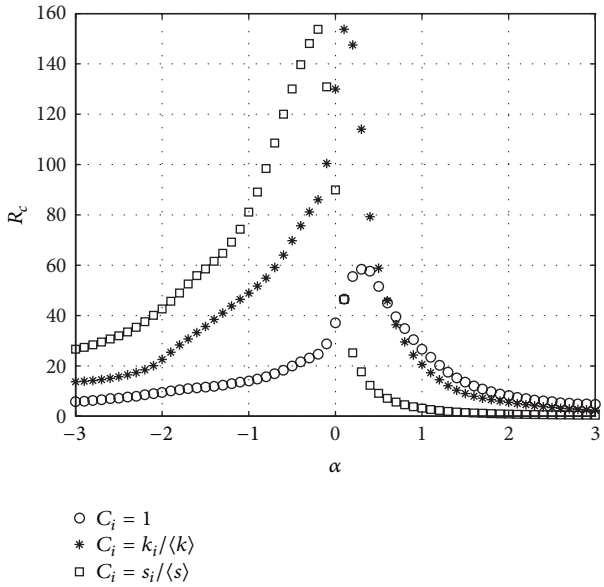


Figure 3: R_c versus α. BBV network with $n = 1000$, $\delta = 6$, $m = 3$, and $\omega_0 = 1$.

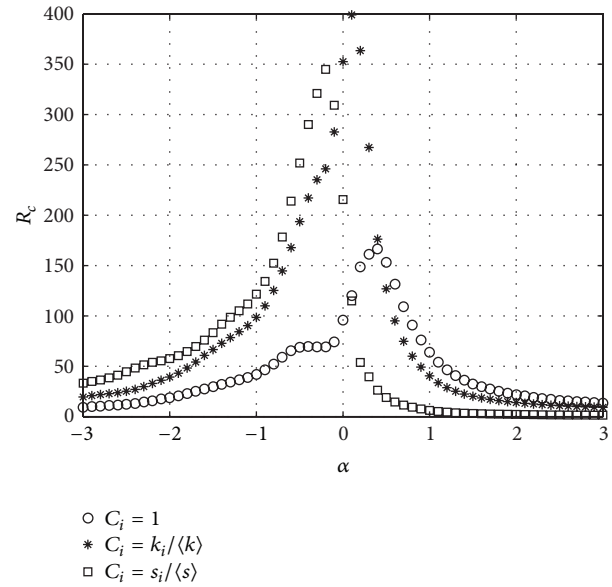


Figure 4: R_c versus α. BBV network with $n = 2000$, $\delta = 6$, $m = 6$, and $\omega_0 = 1$.

In order to characterize the phase transition from free flow to congested phase, we use the order parameter introduced in [23]:

$$\eta = \lim_{t\to\infty} \frac{\langle \Delta\Theta \rangle}{R * \Delta t}, \tag{7}$$

where $\Delta\Theta = \Theta(t + \Delta t) - \Theta(t)$, with $\langle \cdots \rangle$ indicating average over time windows of width Δt, and $\Theta(t)$ is the total number of packets in the network at time t. Therefore, in our simulation we can determine R_c as the phase transition point where η deviates from zero.

3. Simulation and Analysis

Since BBV networks have the same properties (i.e., the power-law distribution of degree, strength, and weight) as lots of real-world networks (the scientists collaboration networks, the Internet, and the WWW), we use BBV networks to investigate the network overall capacity.

In all simulations, we normalize the critical packet generating rate R_c: we plot the normalized R_c and label $C_i = 1$ for the CONC scheme, $C_i = k_i/\langle k\rangle$ for the DEGC scheme, and $C_i = s_i/\langle s\rangle$ for the STRC scheme.

In Figure 1, we plot the critical packet generating rate R_c versus different parameter α in a BBV network with $n = 1000$, $\delta = 6$, $m = 6$, and $\omega_0 = 1$. (For every network, 20 instances are generated and for each instance, we run 20 simulations. The results are the average over all the simulations.)

From Figure 1, we can see that in the three schemes, the critical packet generating rate R_c varies with the tunable parameter α. In the CONC scheme, R_c reaches the peak when α is 0.3 while in the DEGC scheme and STRC scheme α is 0.1 and −0.2 correspondingly. The DEGC scheme where node delivery capability is proportional to its degree has the largest transfer capacity which is 2.35 times as the CONC scheme where node delivery capability is constant. And the transfer capacity of the STRC scheme is 2.15 times as the CONC scheme. Figure 1 show that when the node delivery capability is different, the BBV network has different transfer capacity which varies with the tunable parameter. The BBV network has the highest transfer capacity when the node delivery capability is proportional to its degree and the tunable parameter is 0.1.

Then, we check the impact of δ on the R_c. We set $\delta = 0.6$, 3, 12, and 60 to get different simulation results in Figure 2.

As shown in Figure 2, all three schemes get their highest transfer capacity at different tunable parameter α because of different parameter δ. From formula (3), we can recognize that both the degree distribution and the strength distribution of BBV network are described by an exponent γ which depends on the parameter δ. When the parameter $\delta = 0$, the BBV network is similar to BA network with $\gamma = 3$. And when the parameter δ increases, the distributions become broader with $\gamma = 2$ when the parameter $\delta \to \infty$ which results in different schemes obtaining peak transfer capacity at different tunable parameter α.

Then we check the influence of the new add edges number m and the node number n on three different schemes. Simulation results are shown in Figures 3 and 4 correspondingly.

From Figures 3 and 4, we can come to the conclusion that the new add edges number m and the node number n have a little effect on the impact of tunable parameter α on the transfer capacity of three different schemes. They only affect the absolute value of the transfer capacity.

By investigating the betweenness distribution on the network, a heuristic explanation for the optimal tunable parameter α corresponding to the highest transfer capacity is presented in Figure 5.

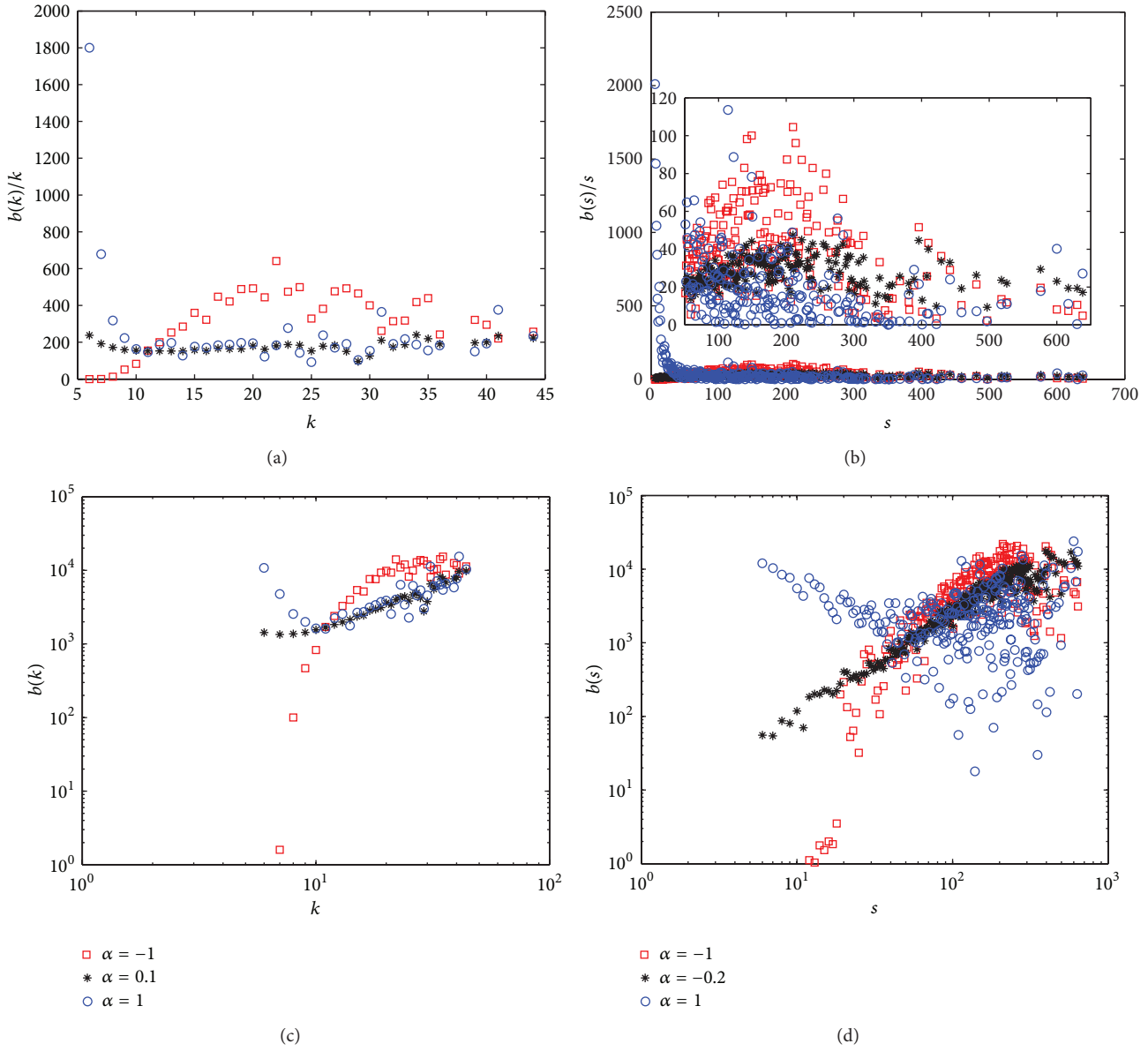


FIGURE 5: BBV network with $n = 1000$, $\delta = 6$, $m = 6$, and $\omega_0 = 1$. (a) DEGC, betweenness per node. (b) STRC, betweenness per node. (c) DEGC, betweenness distribution versus node degree. (d) STRC, betweenness distribution versus node strength.

The results of DEGC scheme are shown in Figures 5(a) and 5(c). In both figures, when the tunable parameter α is 0.1, the load is distributed more evenly than the other two. In Figure 5(a), the betweenness divided by node degree is relatively flat which means that the node with higher degree forward more packets. And in Figure 5(c), the linear characteristic in the log-log plot expresses the same meaning. The results of STRC scheme shown in Figures 5(b) and 5(d) are of the same meaning while the tunable parameter α is -0.2.

In Figure 6, we show the relationship between the critical packet generating rate R_c and the node number n. Figure 6(a) indicates that the network capacity of the DEGC scheme is always much larger than those of the STRC and CONC schemes when the tunable parameter α is 0.1. And Figure 6(b) indicates the situation of the STRC scheme with $\alpha = -0.2$.

The average weighted average length [24] L_{AVE} versus the node number n is reported in Figure 7. Although the weighted average length of DEGC scheme with $\alpha = 0.1$ and the STRC scheme with $\alpha = -0.2$ are higher than that of the traditional shortest path with $\alpha = -1$, the small-world phenomenon, that is, $L_{\text{AVE}} \propto \ln n$, is still maintained.

Finally, we test the three schemes on real-world networks. We choose the scientific collaboration network [25] which has a giant component of 5835 nodes. Simulation results are shown in Figure 8.

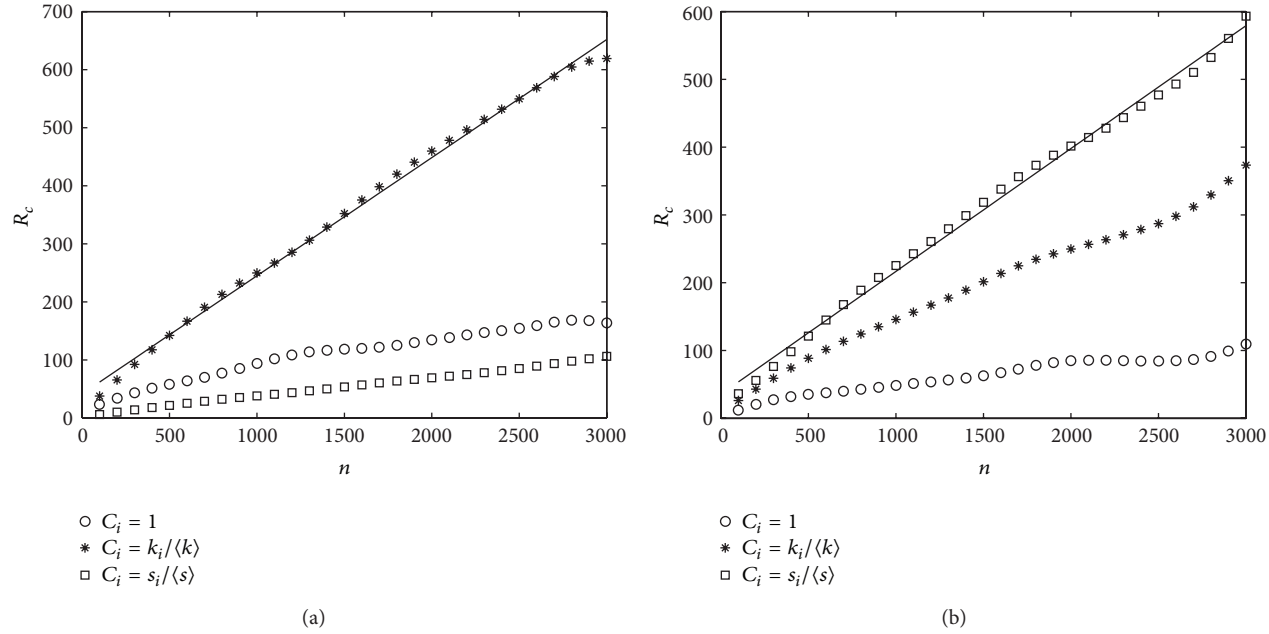


Figure 6: R_c versus n. BBV network with $\delta = 6$, $m = 6$, and $\omega_0 = 1$. (a) $\alpha = 0.1$, (b) $\alpha = -0.2$.

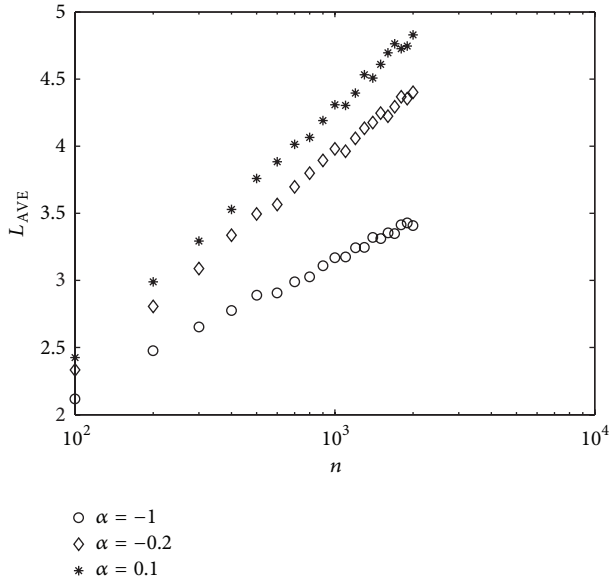


Figure 7: L_{AVE} versus n. BBV network with $\delta = 6$, $m = 6$, and $\omega_0 = 1$.

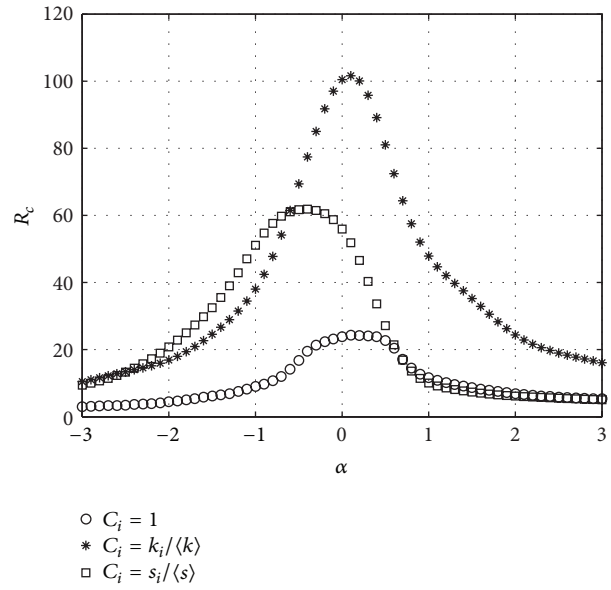


Figure 8: R_c versus α. Real-world network.

From Figure 8, we can see that in CONC scheme the network has the maximum transfer capacity, 24.2, when the tunable parameter α is near 0.3. In DEGC scheme the peak value is 101.6 with $\alpha = 0.1$ and in STRC scheme it is 60.5 with $\alpha = -0.2$. It means our strategy also works well in the real world network.

4. Conclusion

Considering the different node transfer capability, this paper has proposed a new routing strategy to enhance the network transfer capacity in weighted networks. The characteristic of our strategy is to select the optimal routing path according to three kinds of different schemes. The simulation yields some results different from those of previous studies. In most cases, the optimal value of the tunable parameter is 0.3 in the scheme in which each node has the same packet delivery capability (CONC), 0.1 in the scheme in which the node delivery capacity is considered to be proportional to the node degree (DEGC), and −0.2 in the scheme in which the node delivery capacity is considered to be proportional to the node strength (STRC). It is worth mentioning that in some

weighted network the optimal value fluctuate around the mentioned value. And the scheme in which the node delivery capacity is proportional to the node degree has the highest transfer capacity when the tunable parameter is −0.2. At last, we apply our routing strategy on the scientific collaboration network to show the validity of the strategy on real-world networks. Moreover, the above-mentioned research may be practically useful for designing communication protocols.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant no. 60874091), the Six Projects Sponsoring Talent Summits of Jiangsu Province, China (Grant no. SJ209006), and the Natural Science Foundation of Jiangsu Province, China (Grant no. BK2012082) and sponsored by Qing Lan Project.

References

[1] D. J. Watts and S. H. Strogatz, "Collective dynamics of "small-world" networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[2] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[3] J.-P. Onnela, J. Saramäki, J. Hyvönen et al., "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 18, pp. 7332–7336, 2007.

[4] M. E. J. Newman, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," *Physical Review E*, vol. 64, no. 1, Article ID 016132, 7 pages, 2001.

[5] E. Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai, and A.-L. Barabási, "Global organization of metabolic fluxes in the bacterium Escherichia coli," *Nature*, vol. 427, no. 6977, pp. 839–843, 2004.

[6] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, pp. 3747–3752, 2004.

[7] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach*, Cambridge University Press, Cambridge, UK, 2007.

[8] S. H. Yook, H. Jeong, A.-L. Barabási, and Y. Tu, "Weighted evolving networks," *Physical Review Letters*, vol. 86, no. 25, pp. 5835–5838, 2001.

[9] D. Zheng, S. Trimper, B. Zheng, and P. M. Hui, "Weighted scale-free networks with stochastic weight assignments," *Physical Review E*, vol. 67, no. 4, Article ID 040102(R), 4 pages, 2003.

[10] T. Antal and P. L. Krapivsy, "Weight-driven growing networks," *Physical Review E*, vol. 71, no. 2, Article ID 026103, 2005.

[11] A. Barrat, M. Barthélemy, and A. Vespignani, "Weighted evolving networks: coupling topology and weight dynamics," *Physical Review Letters*, vol. 92, no. 22, pp. 228701–1, 2004.

[12] W.-X. Wang, B.-H. Wang, B. Hu, G. Yan, and Q. Ou, "General dynamics of topology and traffic on weighted technological networks," *Physical Review Letters*, vol. 94, no. 18, Article ID 188702, 2005.

[13] T. Zhou, "Mixing navigation on networks," *Physica A*, vol. 387, no. 12, pp. 3025–3032, 2008.

[14] G. Yan, T. Zhou, B. Hu, Z.-Q. Fu, and B.-H. Wang, "Efficient routing on complex networks," *Physical Review E*, vol. 73, no. 4, Article ID 046108, 2006.

[15] S.-Y. Zhou, K. Wang, Y.-F. Zhang, W.-J. Pei, C.-L. Pu, and W. Li, "Generalized minimum information path routing strategy on scale-free networks," *Chinese Physics B*, vol. 20, no. 8, Article ID 080501, 2011.

[16] C.-Y. Yin, B.-H. Wang, W.-X. Wang, G. Yan, and H.-J. Yang, "Traffic dynamics based on an efficient routing strategy on scale free networks," *European Physical Journal B*, vol. 49, no. 2, pp. 205–211, 2006.

[17] C.-Y. Yin, B.-H. Wang, W.-X. Wang, T. Zhou, and H.-J. Yang, "Efficient routing on scale-free networks based on local information," *Physics Letters A*, vol. 351, no. 4-5, pp. 220–224, 2006.

[18] W.-X. Wang, B.-H. Wang, C.-Y. Yin, Y.-B. Xie, and T. Zhou, "Traffic dynamics based on local routing protocol on a scale-free network," *Physical Review E*, vol. 73, no. 2, Article ID 026111, 2006.

[19] M.-B. Hu, R. Jiang, Y.-H. Wu, W.-X. Wang, and Q.-S. Wu, "Routing on a weighted scale-free network," *Physica A*, vol. 387, no. 19-20, pp. 4967–4972, 2008.

[20] R. Yang, W.-X. Wang, Y.-C. Lai, and G. Chen, "Optimal weighting scheme for suppressing cascades and traffic congestion in complex networks," *Physical Review E*, vol. 79, no. 2, Article ID 026112, 2009.

[21] A. Barrat, M. Barthélemy, and A. Vespignan, "Modeling the evolution of weighted networks," *Physical Review E*, vol. 70, no. 6, Article ID 066149, 12 pages, 2004.

[22] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, "Characterization and modeling of weighted networks," *Physica A*, vol. 346, no. 1-2, pp. 34–43, 2005.

[23] A. Arenas, A. Díaz-Guilera, and R. Guimerà, "Communication in networks with hierarchical branching," *Physical Review Letters*, vol. 86, no. 14, pp. 3196–3199, 2001.

[24] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: generalizing degree and shortest paths," *Social Networks*, vol. 32, no. 3, pp. 245–251, 2010.

[25] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 404–409, 2001.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 654536, 13 pages
http://dx.doi.org/10.1155/2013/654536

Research Article

A Novel Coordinated Edge Caching with Request Filtration in Radio Access Network

Yang Li,[1] Yuemei Xu,[1] Tao Lin,[1] Xiaohui Wang,[2] and Song Ci[1,3]

[1] *High Performance Network Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China*
[2] *School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, China*
[3] *Department of Computer and Electronics Engineering, University of Nebraska-Lincoln, Omaha, NE 68182, USA*

Correspondence should be addressed to Yang Li; liy@hpnl.ac.cn

Received 7 October 2013; Accepted 25 November 2013

Academic Editors: W. Sun, G. Zhang, and J. Zhou

Content caching at the base station of the Radio Access Network (RAN) is a way to reduce backhaul transmission and improve the quality of experience. So it is crucial to manage such massive microcaches to store the contents in a coordinated manner, in order to increase the overall mobile network capacity to support more number of requests. We achieve this goal in this paper with a novel caching scheme, which reduces the repeating traffic by request filtration and asynchronous multicast in a RAN. Request filtration can make the best use of the limited bandwidth and in turn ensure the good performance of the coordinated caching. Moreover, the storage at the mobile devices is also considered to be used to further reduce the backhaul traffic and improve the users' experience. In addition, we drive the optimal cache division in this paper with the aim of reducing the average latency user perceived. The simulation results show that the proposed scheme outperforms existing algorithms.

1. Introduction

With the worldwide growth in the adoption of smart phones and tablets, access to Internet from mobile devices (MDs) is projected to grow very significantly [1]. When Internet is accessed by an MD, the content has to be fetched from the original servers outside of the mobile networks. The CDNs [2, 3] help reduce Internet bandwidth consumption and associated delay/jitter, but the content must additionally travel through the wireless carrier Core Network (CN) and Radio Access Network (RAN) before reaching the MD. Bringing each of the requested contents from the CDNs can put significant strain on the carrier's CN and RAN backhaul, leading to congestion, significant delay, and constraint on the network's capacity to serve large number of concurrent content requests.

Concurrently, the emergence of massive content delivery involves repeated wireless transmission of contents that are required multiple times by different users in a completely asynchronous way. Since the relative delay at which two users may require the same content is generally much larger than the duration of the file download, a conventional network architecture treats each session as independent data, and it is therefore incapable of exploiting the intrinsic multicasting capability of the wireless medium. Traditional methods for increasing data throughput in wireless networks have relied on the following three approaches [4]: (i) increase of spectrum usage, (ii) increase of the per-link spectral efficiency, and (iii) increase of spatial reuse. Increasing the amount of spectrum is limited by the fact that spectrum is a finite resource, and allocating new bands to cellular services is a long and expensive process. Increasing the spectral efficiency per link is also approaching its limits: fourth generation cellular systems such as LTE have a near-optimal physical layer, using OFDM together with capacity approaching codes and multiple antenna elements. While further improvements, such as cooperative multipoint and interference alignment, are expensive and/or require redesign of the physical layer. In addition, all three approaches require every base station (BS) to have a high-speed backhaul, whose quality must be better than the aggregate data rate of all its served users. It is not practical since the infrastructure promotion is bound to lag behind Internet traffic growth.

To facilitate the tremendous growth of mobile Internet traffic consumption without the above associated limitations, in this paper we are describing a radically new approach that is based on the following two key observations: (i) a large amount of content delivery traffic is caused by a few, popular files and (ii) disk storage is a quantity that increases faster than any other component in communications or processing systems. The capabilities of storage devices have been further enhanced by improved coding for storage, in particular excellent distributed storage codes. The essential idea of our approach thus is to trade off backhaul capacity with caching of contents at local BSs at the edge of the RAN. In other words, those BSs obtain the most popular video files by downloading through their weak backhaul links, such that most requests can be served from the RAN caches, instead of having to be fetched from the Internet CDNs.

Now the mobile network is undergoing an evolution from 3G to LTE, as shown in Figure 1 [5]. In such flattened network architecture, there are more eNodeBs than before under a centralized SAE-GW control directly. Since the size of edge caches in RAN is much smaller than that of caches used in Internet CDN, it is preferable to store contents at these massive BSs in a coordinated manner, which allows more diverse contents to be efficiently cached in a RAN. The controller SAE-GW, therefore, can leverage these storage resources to serve domestic request in a managed peer-to-peer fashion, thus improving the overall content delivery performance. Ideally, one would like to *optimize the design of this coordinated caching in order to maximize the amount of the supported request in the RAN*. There are many coordinated caching algorithms proposed based on optimization methods, where each content request arriving rate is required to be known in advance [6–8]. However, in practice, *the content request rate demanded by users is varying over time with its popularity change*, making the optimal scheme derived from the past statistic period not applicable anymore. Furthermore, it is complex and not cost effective to collect all content request rates at SAE-GW.

In addition, *due to the independent peer-to-peer unicast transmission mechanism, massive repeating traffics flood the network, consuming the precious link bandwidth resource*. For example, two MDs may fetch the same content from the same neighbor BS concurrently. In contrast, *merging these two traffic flows into one can greatly reduce the total network traffic delivered*. Therefore, we propose a novel caching strategy, based on new concepts introduced in the paper: *request filtration* and *asynchronous multicast*. As a result, one copy of the cached content is able to serve all users within one RAN. The bandwidth requirement, thereby, is derived from the shared cache size accordingly.

Our concept is pushed even further by introducing the notion of MDs themselves as cache nodes. Recent years have seen an enormous proliferation of smartphones and tablets that have anywhere between 10 and 64 GByte of storage (even 500 GByte on typical laptop hard disks). *By enabling communications among devices through base station in a p2p manner, the ensemble of MDs can become a distributed cache that allows much more efficient downloading*. The advantage of using MDs lies in the small deployment costs and automatic upscaling of the capacity as the density of such devices increases. The drawback lies in the necessity to motivate users to participate in the cache and the more random nature of the available throughput. For MDs as caches, some of the fundamental questions are (i) what files should be stored by the MDs? (ii) How can they most efficiently acquire those files? And (iii) does a centralized control of the communications through base station have significant advantages?

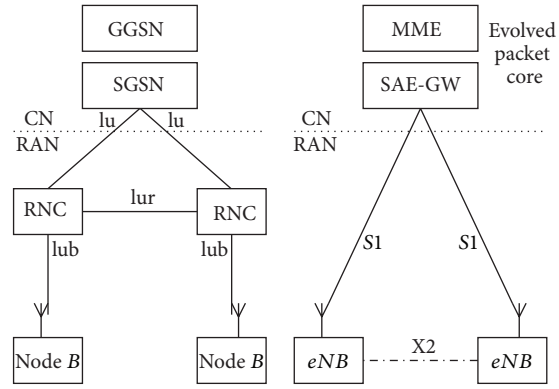


Figure 1: The RAN evolution from 3G to LTE.

We summarize the main contributions of this paper as follows.

(i) We divide the original cooperation problem into subproblems that focus on the direct or indirect cache connection according to the practical topology.

(ii) Based on a deep analysis on how the request filtration can reduce the total network traffic, we investigate the relationship between the shared cache size and the link bandwidth.

(iii) Combined with the caches provided by the MDs and the caches at the base station, we propose our complete coordinated edge caching strategy.

(iv) With the aim of reducing the average latency user perceived, we derive the optimal shared cache size for the proposed coordinated edge caching strategy, without knowing the content request rate in advance.

(v) We evaluate the effectiveness of the proposed scheme by extensive simulation experiments. The results show that our scheme outperforms existing schemes.

The remainder of this paper is organized as follows. Sections 2 and 3 present the related works and motivation, respectively. Based on the request filtration analysis, Section 4 illustrates the relationship between the shared cache size and link bandwidth; then proposes the indirect coordinated edge caching strategy Section 5. Section 6 derives the optimal shared cache size under the proposed strategy. Scheme evaluation results through simulation are presented in Section 7. Finally, Section 8 offers some concluding remarks.

2. Related Work

Content caching has been a key component of Internet based services for many years [3], and there have been many studies in the literature on content caching techniques [9–11]. In particular, coordinated content caching has been studied extensively [12–14]. Researchers have investigated the effectiveness of collaborative caching and proposed numerous collaborative caching schemes [15–17]. Additionally, there has been some research in content caching in wireless networks [18, 19], such as hoc networks [20–22], mobile broadcast environment [23], and MDs [4, 24–26]. Especially with the Information Centric Networking (ICN) resurgence in recent years [27, 28], the in-network caching attracts more research attention [29–33].

A promising evolution of the content delivery architecture consists of extending the caches to the "last mile" by incorporating small servers close to the edge of the network. For instance, this approach leverages devices at the AP and BS, or within users' homes, such as advocated by the Nanodatacenters [34] consortium, network attached storage (NAS) such as Boxee [35], or appliances promoted by business initiatives such as Apple TV. A scalable and adaptive mechanism designed for such "diffuse cloud" of nanoservers is proposed in [36], where part of or the entire storage and bandwidth capacities of set-top devices can be leased to a CDN or a content provider. The latter leverage these resources to store content and serve download requests effectively in a managed peer-to-peer fashion. Moreover, a new system for a neighborhood-assisted video-on-demand service, in [37], is designed to reduce access link traffic by carefully placing VoD data across the neighborhood. So the access network load can be relieved by local connectivity and storage in residential environments. Especially in [38], Fayazbakhsh et al. argue that the simple edge-based caching architecture can achieve a competitive good performance, compared to a full-fledged ICN architecture (i.e., with pervasive caches and nearest-replica routing) with respect to response time, network congestion, and origin server load. Using sensitivity analysis on a range of configuration parameters, they find that the optimistic best-case improvement that ICN can provide is 17% over the simple edge caching architecture (on all metrics).

In addition, AT&T [19] provides a simple cost model for cellular networks, which the third parties can easily use to determine the cost-benefit tradeoffs for their own cellular network settings. It is a first large scale caching analysis for cellular networks, which explored the potential of forward caching in 3G cellular networks by using traffic traces generated by millions of users from one of the world's largest 3G cellular networks. This cost model shows the tradeoffs between deploying forward caching at different levels in the 3G network hierarchy. They also draw a conclusion that caching at regional data centers is the most beneficial with a 26.7% savings in cost.

Our work differs from the above studies in three ways. First, we address the request filtration in order to reduce the total network traffic. To the best of our knowledge, our work is the first attempt to formally investigate and provide insights in addressing these issues. Secondly, we derive a model to define the relationship between the shared cache size and the upstream bandwidth requirement. Thirdly, we investigate the cache space provided by the users in our caching coordinated strategy in order to maximize the local supported traffic.

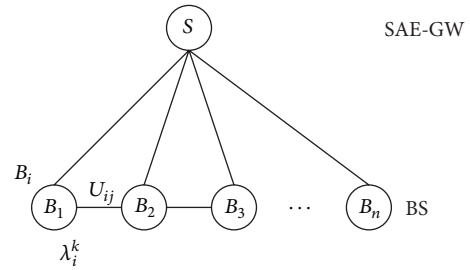


Figure 2: Direct coordinated edge caching model.

3. Coordinated Edge Caching Mechanism

Based on the topology derived from the LTE system, as shown in Figure 1, we divide the original problem into subproblems that focus on the cooperation in different cache levels. We first discuss the coordination among interconnected caches at the edge BSs. This is derived from the practical scenario where neighbor edge BSs are directly connected by high capacity links (known as X2 traffic). The coordination among edge BSs connected via the SAE-GW will then be investigated. Such a form of cooperation is achieved through the unique downstream content retrieval with dynamic routing decisions.

3.1. Direct Edge Caches Coordination. As denoted by Figure 2, fully interconnected caches at the edge BSs are able to disseminate contents via direct links. We then investigate the subproblem that aims to maximize the amount of supported traffic at this level by utilizing the capacity of direct links.

We use S and B_i to present SAE-GW and BS, respectively, where each B_i, $i \in n$ has storage capacity c_i. The link capacity between a pair of BSs at the edge caches is presented by U_{ij}. We define the average arrival rate of content requests at edge cache C_i as λ_i. The proportion of requests that attempt to access content $k \in K$ is represented by p_k, and we assume that such a distribution pattern is uniform for all fully connected caches colocated in the same domain. Thus the rate of a request to content k at cache is represented by $\lambda_i^k = \lambda_i p_k$. Combined with concerns about the segment size, we treat $\lambda_i^k s_k$ as the bandwidth requirement of requests to content k in one time unit. Without loss of generality, we assume that the link capacity U_{ij} is consistently larger than $\sum_{k \in K} \lambda_i^k s_k$, $i \in n$. The assumption is practical based on our observations in the real world system. This leads to the following theorem, which is used in cache cooperation of interconnected edge caches.

Theorem 1. *Given $\sum_{i \in n} c_i \leq \sum_{k \in K} s_k$ and $U_{ij} \geq \sum_{k \in K} \lambda_i^k s_k$, for any content k, there is at most one copy of the content k in the total n caches, in order to maximize the satisfied requests.*

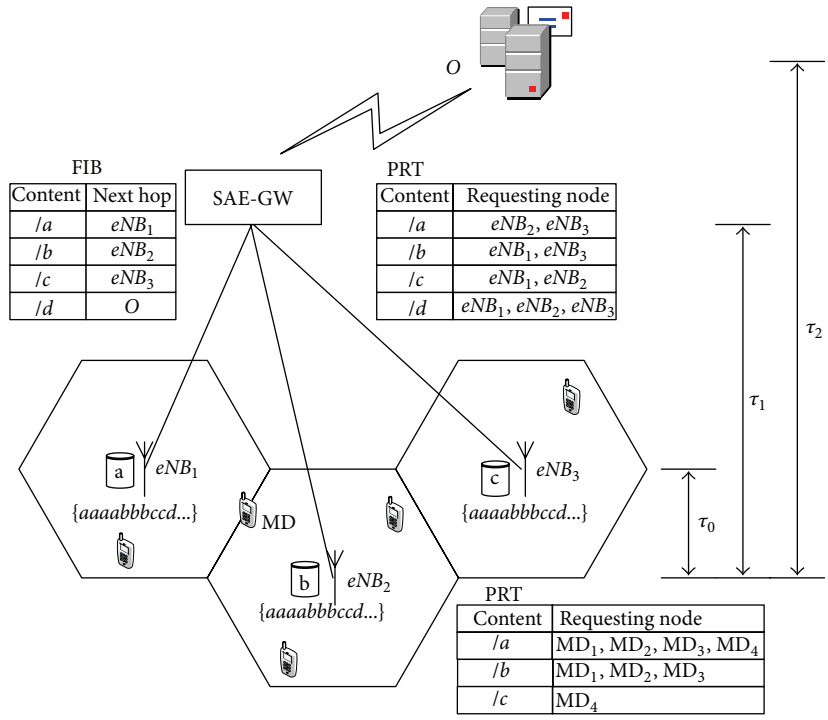


Figure 3: A motivating example.

Proof. Suppose there are two copies of a content k in caches at edges B_i and B_j, $i \neq j$. We delete one copy stored in the cache at edge B_j, which generates an assignment not worse than the optimal solution. Given $\sum_{i \in n} c_i \leq \sum_{k \in K} s_k$, we can store another content k' in the empty cache room at edge B_j. Since $U_{ij} \geq \sum_{k \in K} \lambda_i^k s_k$, the new cached content k' can provide service in the domain, which yields an assignment that surpasses the optimal solution. This is a contradiction. □

Theorem 1 suggests that all duplicated contents in the interconnected caches set need to be removed and replaced by other unavailable contents. Moreover, all directly connected caches can be treated as one single cache. There is no need to consider the strategic placement of contents with such a combination, since the traffic can be easily routed through direct links at the edge BSs.

3.2. Motivation for Indirect Edge Caches Coordination. If the direct link between the edge BSs is not available or the direct link is not used for data (only for signaling), the coordination among these caches can be realized with the aid of SAE-GW. If the content is available at the accessed BS, the request could be immediately served by the attached cache. When the request is not fulfilled, it will then be sent to the corresponding upper level SAE-GW. Whether the request is routed to the other bottom caches or the original server depends on the cached content category and the reverse link capacity from BS to SAE-GW. The efficiency of the indirect edge caches coordination is complicate, so we first give an *illustrate example* in order to make the problem easily understood. Then our proposed indirect cache coordination strategy is presented in the next following two sections.

Figure 3 shows a RAN consisting of three BSs, eNB_1, eNB_2, and eNB_3, and one original server O serving four content objects, a, b, c, and d. Moreover, all BSs have storage capacity to store one single content object only, whereas SAE-GW does not have any available capacity for storing contents. In addition, the capacity of each upstream link from the BS to the SAE-GW is only capable of fetching one single content object simultaneously.

We assume that there are sets of MDs sending request flows to their BSs at each cell. The Request flows are identical, represented by a repeating sequence {*aaaabbbccd*}. We assume that the performance (e.g., latency) of fetching contents from a peer BS in the same RAN is much better than from the original server O, since in real network, the origin may reside quite far away from the network. Then, apparently, storing distinct contents at the BSs would reduce the overall delay and improve the network performance that MDs experience. However, the performance improvement is subject to the BS's upstream link capacity when we consider the bandwidth consumed during the delivery of the massive contents. Due to the limited storage and link capacities, the problem is *how to select contents to store at each router so as to improve the network performance*.

We consider the following three edge caching strategies. (1) *Noncoordinated caching*: all BSs work independently, where they both adopt the canonical caching policy based on frequency or historical usage. Assume that the content popularity distribution is consistent, and all BSs have already cumulated the information that content a is requested more

often than b, c, and d. In this case, all BSs store a. (2) *Coordinated caching*: all BSs work jointly and always prefer each other over the original server whenever possible. In this case, each BS may store different content, respectively. Without loss of generality, we assume that eNB_1 stores a, eNB_2 stores b, and eNB_3 stores c. Then, on cache misses, a requested content will always be retrieved from peer BSs and then the original server O, sequentially. (3) *Request filtration*: the content caching and searching are the same as in the coordinated caching strategy. However, at each node (BS and SAE-GW) in a RAN, all unsatisfied requests for the same content are filtered into one request only (i.e., except the first request being sent out, all the following unsatisfied requests for the same content, before the reply returns, are filtered by the pending request table (PRT)). Through adding the requesting nodes to the corresponding content entry in the PRT built by its ancestor, the returned content is multicast to all record requesters, as shown in Figure 3.

We compare the above three edge caching strategies by using two metrics: the *load on origin* and the *traffic delivered*. We summarize the comparison results in Table 1.

First of all, the load on origin is measured by the percentage of all requests served directly by the original server O. With the noncoordinated strategy, the requests for content a will be directly served by the accessed BSs (recall that all BSs store a in this case), while the requests for b, c, and d will have to be served by the original server. This means a total 6/10 of all Requests from three flows incur the traffic load on the original server. However, with the coordinated and request filtration strategies, since a, b, and c are stored locally, requests from the MDs can be served by all BSs.

The performance difference between coordinated caching and request filtration is determined by the average latency of fetching contents and the request arriving rate. As shown in Figure 3, we denote by τ_1 the average latency of serving requests from a peer BS in the given RAN and by λ_i^k the average request arriving rate for content k in one cell. The case of $\lambda_i^k \cdot \tau_1 \geq 1$ means that there are some requests for content k arrived at BS before the previous Reply for the same content returns back. In this case, with the request filtration strategy, the BS filters the following requests and records the requesting nodes to the PRT (the specific operation is explained in Section 3.1), while with the coordinated strategy, all requests are sent out either to the peer BS or to the original server if the upstream link of the peer BS is busy.

Therefore, with the coordinated strategy, in the case of requests for the a, b, and c being all served in the RAN (*best case*) due to their lower request rate, the total 3/30 for d of all requests from three flows incurs the traffic load on the original server. However, in the case of all the upstream link being busy if all requests arrive simultaneously (*worst case*), the link capacity can serve one request only, so other requests for a, b, and c are still delivered to the original server. Hence, the load on the origin is $(4 + 5 + 6)/30 = 0.2$.

In contrast, using the request filtration strategy, during the above worst case, all the Requests for same content are filtered at the BSs (e.g., requests for the a from all MDs are filtered at the eNB_2 PRT). Moreover, the SAE-GW also filters the request from its BSs as shown in Figure 3 (e.g., requests for the a from eNB_1 and eNB_2 are filtered at the SAE-GW PRT). Hence, only one request for each content is sent out, where one Request for d is delivered to the origin (i.e., 1/30 load origin). If each request arrives after its previous Reply data returns back (the above best case), the request filtration strategy performs as well as the coordinated strategy since no request filtration happens.

Table 1: Comparing the three edge caching strategies.

	Non-coordinated caching	Coordinated caching	Request filtration
Load on origin	0.6	0.1~0.2	0.033~0.1
Traffic delivered	1.8	1.5~1.9	0.56~1.5

In addition, the traffic delivered is measured by the product of average number of network hops and content's size. Based on the assumption of unit-size contents, the traffic delivered can be directly expressed as the number of network hops traversed when fetching contents. Using the noncoordinated strategy, MDs requesting for a can directly be satisfied from their accessed BSs without going through any peer BS, while requests for b, c, and d have to go to the origin which is three hops away via SAE-GW (i.e., assume that the hop count is 2 from SAE-GW to origin). Therefore, the average traffic delivered for noncoordinated strategy is $(3 \cdot 6 \cdot 3)/30 = 1.8$ per request. Using the coordinated strategy, under best case where the requests for a, b, and c are all served in the RAN, only requests for d are sent to origin. Hence, the average traffic delivered is $[(2 \cdot 5 + 3 \cdot 1) + (2 \cdot 6 + 3 \cdot 1) + (2 \cdot 7 + 3 \cdot 1)]/30 = 1.5$ per request. While under the worst case when all requests arrive concurrently, the average traffic delivered is $[(2 \cdot 2 + 3 \cdot 4) + (2 \cdot 2 + 3 \cdot 5) + (2 \cdot 2 + 3 \cdot 6)]/30 = 1.9$ per request since the unsatisfied a, b, and c have to go to the origin. In contrast, the request filtration strategy perform best in worst case, since all repeating requests are filtered at BSs and SAE-GW, only generating $[(2 \cdot 2 + 3 \cdot 1) + (2 \cdot 2 + 1 \cdot 1) + (2 \cdot 2 + 1 \cdot 1)]/30 = 0.56$ per request. When no request filtration happens (the above best case), the request filtration strategy performs as well as the coordinated strategy.

In this example, the request filtration strategy leads to a lower load on original server and a lower traffic delivered than both the noncoordinated and coordinated caching strategies. This is not surprising since most repeating traffics are filtered by request filtration. Moreover, if no request filtration happens (i.e., $\lambda_i^k \cdot \tau_1 \prec 1$), the request filtration strategy performs as well as the coordinated strategy. Furthermore, normally, the coordinated strategy can achieve better performance than the noncoordinated strategy. However, in the case of the above worst case, the coordinated strategy causes more total traffic than the noncoordinated strategy. The reason is that the cached content cannot serve the requests from the peer BSs due to the upstream bandwidth limitation. This key discovery implies that *sometimes the simple noncoordinated strategy can perform better than the elaborative coordinated strategy, especially in the case of heavy request and limited link capacity.* Hence, it is important to investigate the relationship between storage and link capacity to design a proper

coordinated caching strategy so as to improve the network performance.

4. Bandwidth Allocation under Request Filtration

We next propose a coordinated edge caching strategy based on the request filtration in a RAN, which investigates the cache storage and link capacity cooperatively.

4.1. Request Filtration. As explained in the above section, the request filtration can greatly reduce the network traffic delivered. Here we first describe the specific request filtration operation in a RAN. Note that contents are segmented into smaller pieces in this paper, each of which is treated as an individually named content object, to allow flexible distribution and flow control.

First of all, a PRT which records all unsatisfied requests is maintained at each BS and SAE-GW. The PRT has a similar function to that of the Pending Interest Table (PIT) in CCN [27]. An entry in PRT is set up by the first new coming content request. When a request arrives at a node, a longest-match lookup is done on its content name. A PRT match means the content was solicited by request(s) sent by this node. So, if there is an exact-match PRT entry the requesting node (or arrival face) will be added to the PRT entry's "requesting node" list and the request will be discarded (filter the following same request). When the Reply data packet solicited arrives, a copy of that packet will be multicast to all the nodes that the PRT records. We called this procedure asynchronous multicast. Then PRT entries are erased as soon as they have been used to forward a matching Reply data packet (the Reply consumes the request).

An example illustrating the request filtration is shown in Figure 3. When four MDs send requests of content a to eNB_2 concurrently, the eNB_2 sends the request to SAE-GW only once, while adding all MDs to its PRT. Furthermore, the requests for content a arriving at SAE-GW (from eNB_2 and eNB_3) are also filtered into one request (see the entry for a at PRT of SAE-GW), which is sent to the eNB_1 according to the routing table. Then the returning content a will be multicast to eNB_2 and eNB_3 at SAE-GW and then multicast to all MDs at each BS instead.

In addition, note that the above asynchronous multicast can be implemented through intercepting the *HTTP GETs* by the aggregating nodes [39] or some other methods such as the content routing in CCN [27], which is out of the research scope of this paper.

4.2. Shared Cache Size and Link Bandwidth. As we know, the unsatisfied requests at the access BS always prefer fetching contents from the other caching nodes in the same RAN over the original server. Based on the above request filtration scheme, all these requests will filter again at the SAE-GW. Hence, there is at most one request being sent out at the SAE-GW, to fetch the cached content simultaneously. Therefore, *one copy of the cached content in the same RAN is enough to serve all requests from the peer BSs with the request filtration scheme*.

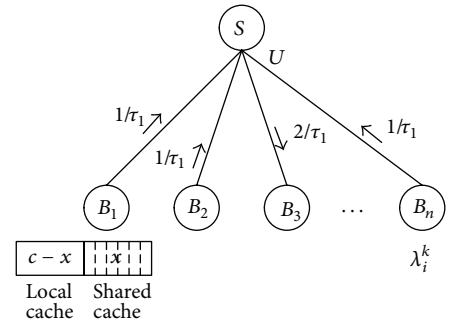


Figure 4: Indirect coordinated edge caching model.

However, the content cached at the BS does not mean that it can serve the peer requests. How percentage of the cached content can be shared within the RAN is determined by the upstream link capacity. Here we assume the downstream link will not become the bottleneck, compared to the upstream link, which is practical in the real mobile network. Hence, *the shared cache size is tied up with the upstream link bandwidth*.

Therefore, we divide the cache storage at the BS into two parts, the noncoordinated *local cache* $c - x$ and the coordinated *shared cache* x, as shown in Figure 4. Here we derive the relationship between the shared cache size x and its upstream link bandwidth U. As explained in Section 3.2, the request filtration happens when requests for the same content arrive at an aggregating node before the previous Reply data for the same content returns back (e.g., $\lambda_i^k \cdot \tau_1 \geq 1$ at BS). In this case, only one request is sent out during the latency of fetching the content, so the Request rate sent out from the aggregating BS is reduced to $1/\tau_1$ no matter how fast the λ_i^k is. Similarly, when all these requests filter again at the SAE-GW, the largest request rate for one content stressed on the link is only $2/\tau_1$ no matter how many child nodes request this content (we assume the round latency between SAE-GW and BS is $\tau_1/2$). Therefore, based on the assumption of unit-size contents, the maximum required upstream link bandwidth is expressed as

$$U_{\max} = x \cdot \frac{2}{\tau_1}. \tag{1}$$

Equation (1) gives the relationship between the shared cache size x and its upstream link bandwidth U; namely, the $U_{\max}$ *can satisfy all requests for the contents in the shared cache in a RAN, without caring about the request density,* with request filtration.

5. Cache Content Placement and Request Routing

In this section, we describe the details of the proposed indirect coordinated edge caching strategy, where the MD's storage capacity is additionally considered as an improvement to the RAN caching system.

5.1. Caches on the Users. Since smartphones have large hard disks built in, these devices can effectively act as caches to serve the other users in the same cell. Of course, users need to be incentivized to help other users: in other words, there has to be a compelling answer to the question: why should I spend my battery to provide you with faster video download? Since network operators benefit from offloading of traffic to end users, such incentives can be provided by network operators. They could take on, for example, the form of discounts or increased data caps for participating users, which is out of the scope of this paper.

We assume in the following that the content caching and delivery on the users are controlled by the BS, so that there is a central control unit that has knowledge about which station has what files in its cache and also knows the channel state information (CSI) between the users. This allows a more efficient scheduling of the communications and ensures that there is no interference between the two types of traffic; this is essential for network operators. Therefore, the content, if not cached at the storage in the BS, can be cached in the required user controlled by its access BS. In this way, the total caches in a cell can maximize the supported requests and reduce the backhaul traffic without contents redundancy. The specific caching strategy is explained in the following paragraph.

5.2. Content Placement and Request Routing. Based on the above analysis, we allocate the bandwidth in our proposed coordinated edge caching strategy according to (1). To simplify the analysis and without loss of generality, we assume that there are n BSs which has the equal cache size c and the equal upstream link capacity. Each BS stores in its local cache (i.e., the $c - x$ portion) the top ranked contents in a noncoordinated manner, and all BSs collaboratively store $n \cdot x$ contents that are ranked from $c - x + 1$ to $c - x + nx$. In order to manage the coordinated caching across the RAN, the SAE-GW has to collect the information of content popularity and disseminate contents to the corresponding BSs periodically. Hence, the SAE-GW keeps the catalog recording the cached content location.

The entire operation procedure of coordinated edge caching strategy is then summarized in Figure 5, where the requests are handled following request filtration rule at each node without specific explanation.

Upon receiving a request, the access BS first checks whether it has a copy in its cache or not. The cached copy can serve the MD directly; otherwise, the access BS will check whether its other MDs have a copy in their own caches or not. If another MD has cached the request content and is willing to serve others, the BS will arrange a communication between these two MDs. In the case that there is no copy of the request content cached in the total cell, the request will be sent up to the SAE-GW. Upon receiving a request, the SAE-GW checks whether any child BSs have a copy in their cache or not. If find, fetch the content from its child BS to serve the MD. If not find, fetch the content from the original server.

Upon receiving a Reply, the SAE-GW multicasts it to its child BSs and then the BS multicasts it to its MDs according to PRT. Additionally, if an MD agrees to share its storage with the others, when receiving a content, the BS can inform the required MD to cache it in case of the Reply content coming from the original server. So the BS keeps the MD catalog recording the cached content location. Note that the content caching at the MD's storage is pulled by the user's request, which is different from the content caching at the BS's storage (where content is pushed by the network periodically). Moreover, if the MD moves out of the cell control of the BS, the BS should delete the corresponding content items in its MD catalog through some detections or MD's notification.

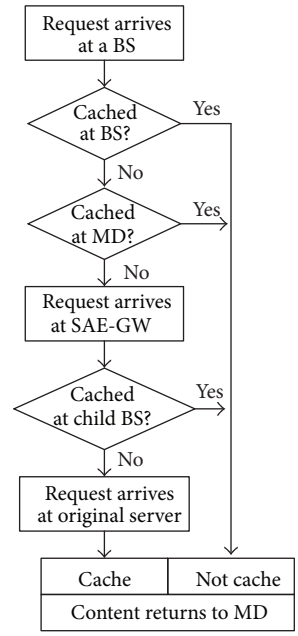


Figure 5: The operation procedure of coordinated edge caching strategy.

With this coordinated edge caching, the next work is to find the optimal x in order to improve the network performance, which is addressed in the following section.

6. Problem Formulation and Analysis

In this section, we formulate the problem of edge cache division as an optimization problem and study the optimal solution.

6.1. Problem Formulation. We consider a simple RAN model, where a set of BSs with storage capability serve content requests originated from its cell, as shown in Figure 3. The original server O stores all content objects, referred to as the "origin"; therefore, requests for any content object can always be satisfied by O. The other detail model notation is given in Table 2. To simplify the analysis, we also normalize the content size to one unit with respect to nodes' storage capacity. Our objective is to explore the optimal shared cache size x for the proposed coordinated edge caching strategy, in order to reduce the average latency user perceived.

Table 2: Model notation.

Symbol	Meaning
n	Number of BSs
N	Number of contents
c	Storage capacity of each BS
x	Shared cache size at each BS
α	Zipf exponent
τ_0	The average latency of serving requests from access BS
τ_1	The average latency of serving requests from peer BS
τ_2	The average latency of serving requests from original server O

Many studies have shown that the content popularity follows the Zipf distribution [40–42]. The Zipf's law predicts that out of a population of N elements, the frequency of elements of rank i, denoted by $f(i, \alpha, N)$, is

$$f(i, \alpha, N) = \frac{1/i^{\alpha}}{\sum_{j=1}^{N} (1/j^{\alpha})} = \frac{1/i^{\alpha}}{H_N^{\alpha}}, \quad i = 1, 2, \ldots. \tag{2}$$

Moreover, we compute the overall probability of requesting for the top k contents by

$$F(k, \alpha, N) = \sum_{i=1}^{k} f(i, \alpha, N) = \frac{H_k^{\alpha}}{H_N^{\alpha}}, \quad k = 1, 2, \ldots, \tag{3}$$

where H_k^{α} and H_N^{α} are the kth and Nth harmonic numbers of order α. In order to ease the analysis and derive meaningful results, we assume that N is sufficiently large and approximate $F(i, \alpha, N)$ using a continuous function

$$F(k, \alpha, N) \approx \frac{\int_1^k t^{-\alpha} dt}{\int_1^N t^{-\alpha} dt} = \frac{k^{1-\alpha} - 1}{N^{1-\alpha} - 1}. \tag{4}$$

Therefore, the average latency of serving a content request is expressed as [43]

$$\begin{aligned} L(x, \alpha, \tau) = {} & F(c - x, \alpha, N)\, \tau_0 + [F(c - x + xn, \alpha, N) \\ & - F(c - x, \alpha, N)]\, \tau_1 \\ & + [1 - F(c - x + xn, \alpha, N)]\, \tau_2. \end{aligned} \tag{5}$$

We can derive the average latency by adding (4) into (5)

$$\begin{aligned} L(x, \alpha, \tau) = {} & \tau_0 \cdot \frac{(c - x)^{1-\alpha} - 1}{N^{1-\alpha} - 1} \\ & + \tau_1 \cdot \frac{(c - x + nx)^{1-\alpha} - (c - x)^{1-\alpha}}{N^{1-\alpha} - 1} \\ & + \tau_2 \cdot \left(1 - \frac{(c - x + nx)^{1-\alpha} - 1}{N^{1-\alpha} - 1}\right) \\ = {} & \frac{1}{N^{1-\alpha} - 1} \left((\tau_0 - \tau_1)(c - x)^{1-\alpha} \right. \\ & + (\tau_1 - \tau_2)(c - x + nx)^{1-\alpha} \\ & \left. + \tau_2 N^{1-\alpha} - \tau_0\right). \end{aligned} \tag{6}$$

6.2. Optimal Strategy. Li et al. provide a rigorous proof of the existence and uniqueness of the optimal strategy in [43]. We solve the optimal solution by letting the first-order derivative of $L(x, \alpha, \tau)$ equal to zero; that is, $\partial L(x, \alpha, \tau)/\partial x = 0$. So we have

$$\begin{aligned} & (\tau_0 - \tau_1)(1 - \alpha)(-1)(c - x)^{-\alpha} \\ & \quad + (\tau_1 - \tau_2)(1 - \alpha)(n - 1)(c - x + nx)^{-\alpha} = 0. \end{aligned} \tag{7}$$

Then we can get

$$\frac{(c - x + nx)^{\alpha}}{(c - x)^{\alpha}} = \frac{(\tau_2 - \tau_1)}{(\tau_1 - \tau_0)}(n - 1). \tag{8}$$

Here, we introduce a parameter $\gamma = (\tau_2 - \tau_1)/(\tau_1 - \tau_0)$ as the ratio of layer latency. Then we can get the optimal shared cache size as

$$x^* = c \cdot \left(1 + \frac{n}{\gamma^{-\alpha}(n - 1)^{-\alpha} - 1}\right)^{-1}. \tag{9}$$

If we use parameter $\eta = x/c$ to represent the percentage of the local cache size used for shared cache, the optimal solution is

$$\eta^* = \left(1 + \frac{n}{\gamma^{-\alpha}(n - 1)^{-\alpha} - 1}\right)^{-1}. \tag{10}$$

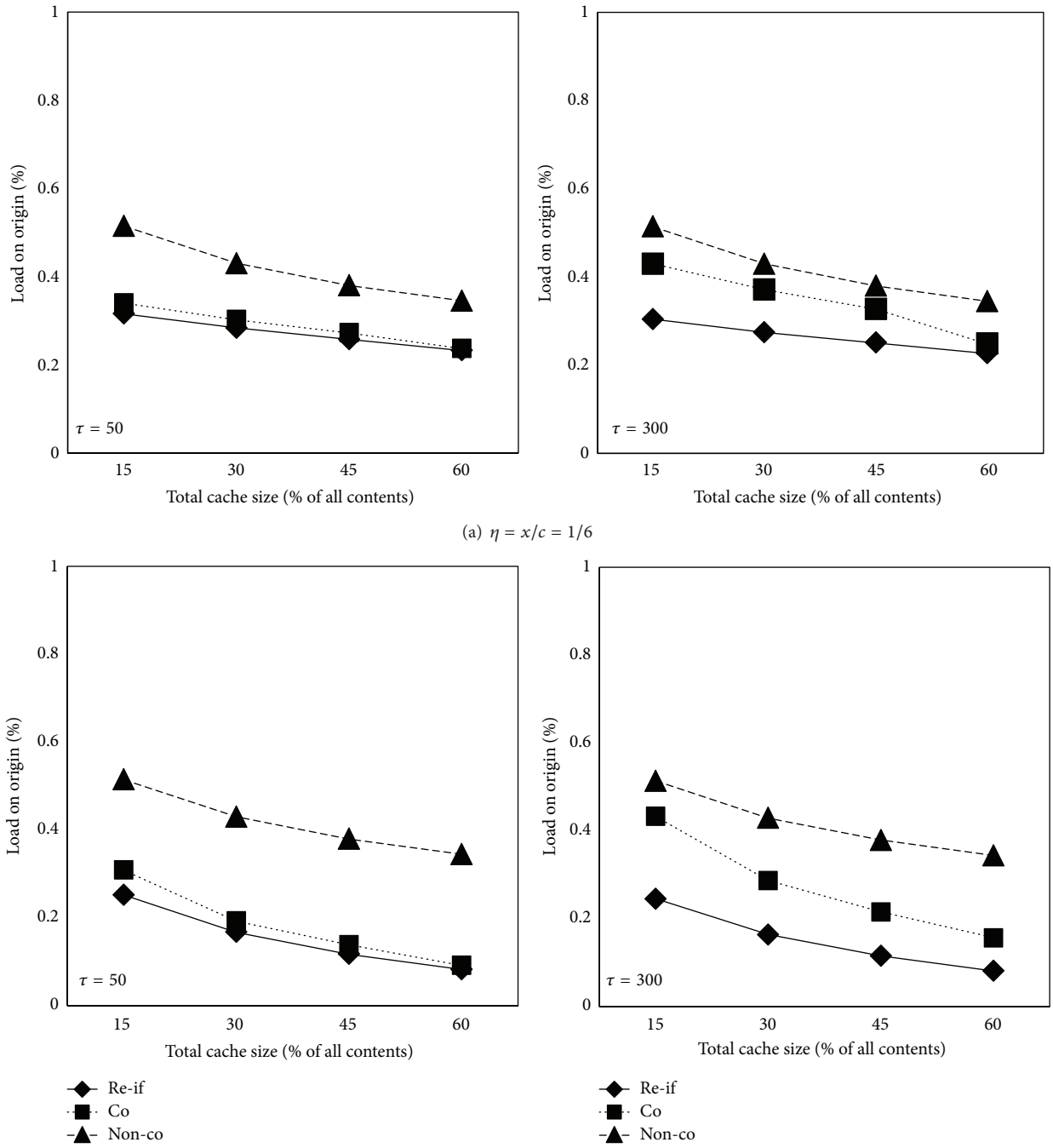
It is important to note that η^* is a function of the deliver latency, rather than the content request rate λ_i^k. Therefore, this scheme is applicable due to low complexity and cost efficiency.

7. Performance Evaluation

We test our scheme in a Java-built simulation environment for RAN, by comparing with noncoordinated caching (*non-co*) and coordinated caching (*co*) illustrated in Section 3. In a 3-level topology similar to that in Figure 3, the root node represents the original server node and there are 10 BSs with edge cache under a SAE-GW control.

The original server holds all the contents, and the total size of these edge caches ranged from 15% to 60% of the whole contents. We consider 2000 contents with the same unit size, fetched by 50000 requests, and the content popularity distribution is assumed to be Zipf (α) with $\alpha = 0.9$. The bigger α indicates that few distinct contents attract the majority of the requests while the smaller ones indicate almost uniform document popularities [40]. Note that we do not consider the cache size provided by the MDs in our simulation, since the users' contribution depends mainly on the number of the MDs and fluctuates according to the users' movement.

Additionally, we give two sets of results derived from different delivery latency τ_1, in order to show the effect of request filtration. However, since our test cannot simulate the real mobile network environment, we only collect data for different filtration interval (i.e., filter per 50 requests and per 300 requests) to analog the different fetching latency for simplicity.



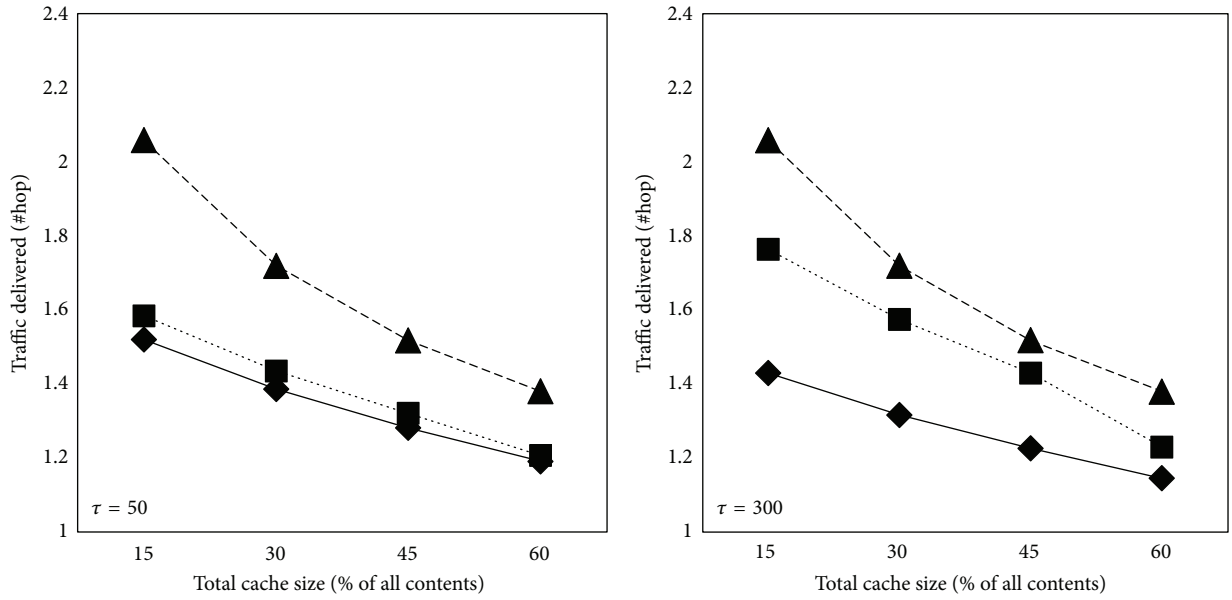
(a) $\eta = x/c = 1/6$

(b) $\eta = x/c = 5/6$

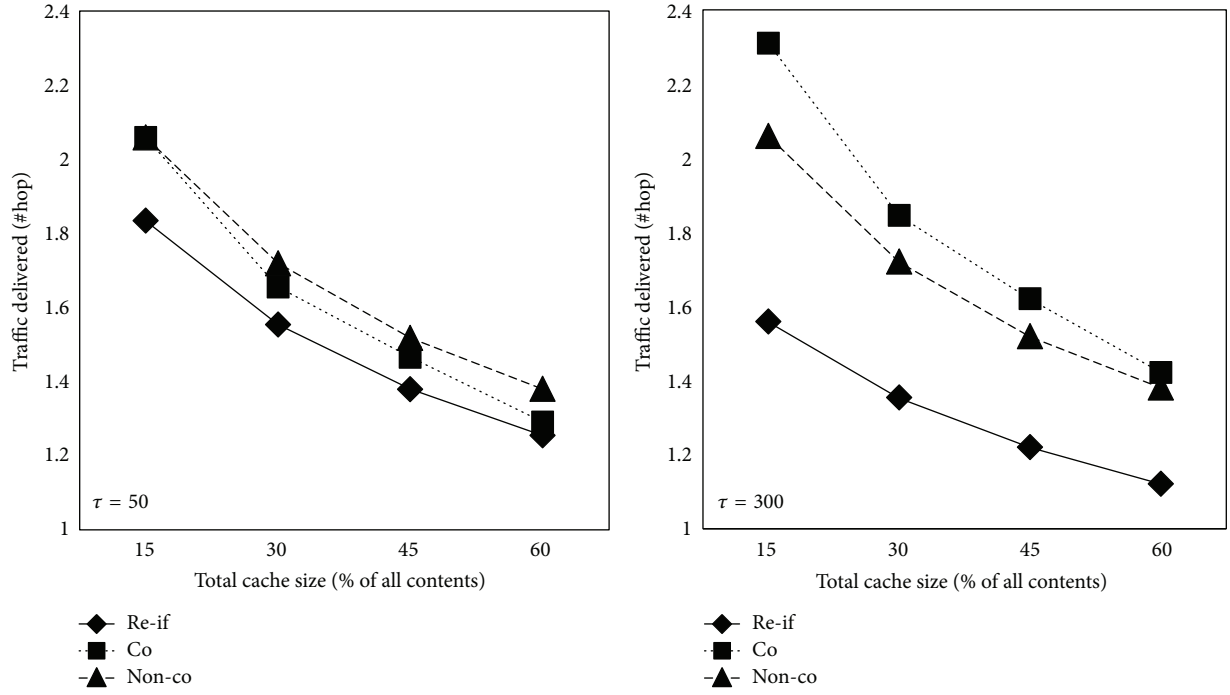
Figure 6: Load on origin versus cache size.

Furthermore, we evaluate how various optimal η^* affects the caching performance by changing the γ value. We derive the results from two η^* values of 1/6 and 5/6, respectively, as the higher the η^* is, the larger the shared cache size and the total cached contents at the BSs are.

Figure 6 plots the load on original server curves as a function of cache size for different caching schemes. The load on origin is defined as the percentage of all requests served directly by the original server. The proposed request filtration scheme (*re-fi*) improves the load on origin over the other schemes. For example, at cache size 15% ($\tau = 50$, $\eta^* = 5/6$), the load on origin is 25%, 31%, and 51% for *re-fi*, *co*, and *non-co* respectively. In addition, it can be clearly seen that the proposed scheme under the scenario of $\tau = 300$ shows a little better performance than that of $\tau = 50$. This is because more requests are filtered during the long content fetching latency, while the coordinated caching scheme sends more requests to the origin due to limited link bandwidth.



(a) $\eta = x/c = 1/6$



(b) $\eta = x/c = 5/6$

Figure 7: Traffic delivered versus cache size.

Moreover, we observe that the load on the origin decreases as η^* increases, because the larger total shared cache size is, the more requests are satisfied at the edge BS in a RAN.

Figure 7 shows the average traffic delivered under each scheme, where the traffic is measured in number of hops travelled when fetching the unit-size content. We assume the hop count is 4 when acquiring content from the original server and 2 from the peer BS. The lower the average traffic delivered, the better the performance. The proposed request filtration scheme reduces the total traffic compared to the other schemes. This is not surprising because the proposed scheme can greatly reduce the traffic delivered by request filtration and asynchronous multicast at all nodes in the RAN. Furthermore, we observe the traffic delivered decreases a little as η^* increases at $\tau = 300$, because the larger total shared cache size is, the more requests are satisfied at the edge BS in a RAN.

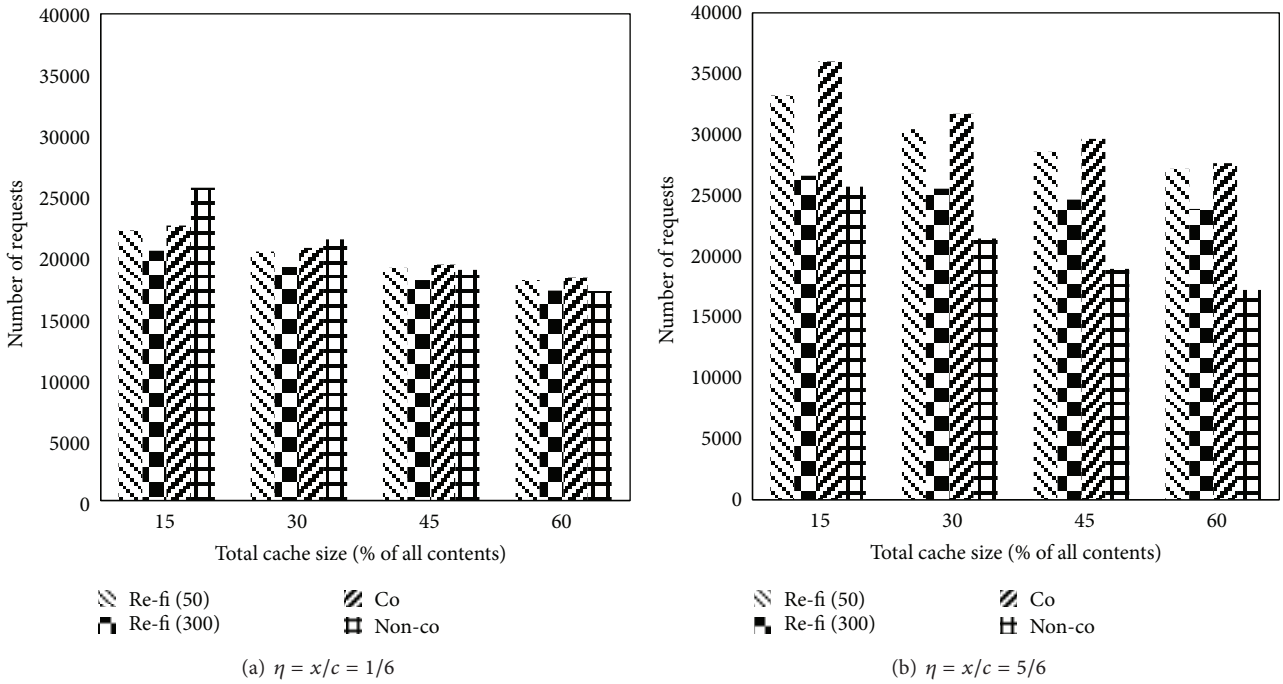


Figure 8: Number of requests versus cache size.

Note that under the scenario of $\tau = 300$, $\eta^* = 5/6$, the noncoordinated caching performs better than coordinated caching, as the worst case stated in Section 3.2. This observation reveals that the request rate and the upstream link capacity can seriously impact the desired effect of elaborative coordinated strategy.

In addition, Figure 8 shows the number of requests sent out by the SAE-GW, either to the original server or to its child BSs. These data reflect how many requests are filtered for the proposed scheme in contrast to the coordinated caching strategy. Hence, fewer requests are sent out in *re-fi* than that in *co*, especially under the scenario of $\tau = 300$. Additionally, it is obvious that the number of requests increases as η^* decreases, since the larger local cache size can satisfy more requests locally.

However, there are least requests sent out from the SAE-GW in non-co (no request is sent to child BSs in this case), since more requests are satisfied locally due to large local cache size without sharing with others. We also note that the proposed scheme achieves the above merits at a cost of maintaining PRT at all nodes in the RAN.

8. Conclusion

Content caching at the BSs is an effective method to improve the mobile network performance. We argued that massive repeating traffics flood the network and impact the coordinated caching effect in this paper. The proposed coordinated caching scheme could greatly reduce the traffic delivered in the network by request filtration and asynchronous multicast. The optimal edge cache size division is also derived in order to reduce the average latency user perceived. The experimental results showed that the proposed scheme effectively reduces the load on the original server and the total traffic delivered through the CN and RAN compared to the existing caching algorithms.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants no. 61202419, no. 61100178, and no. 61303243, the National High Technology Research and Development Program (863 Program) under Grant no. 2013AA013503-2; the Strategic Pilot Project of Chinese Academy of Sciences under Grant no. XDA06010302 and the National Science and Technology Support Program under Grant no. 2012BAH06B02.

References

[1] White Paper, "Cisco Visual Networking Index: Global Mobile Data," 2010–2015.

[2] G. Pallis and A. Vakali, "Insight and perspectives for content delivery networks," *Communications of the ACM*, vol. 49, no. 1, pp. 101–106, 2006.

[3] Akamai Technologies, http://www.akamai.com/.

[4] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: a new architecture for wireless video distribution," *IEEE Communications Magazine*, vol. 51, pp. 142–149, 2013.

[5] PDF3GPP TR 25. 913 V7. 3. 0: Requirements for Evolved E-UTRA and E-UTRAN.

[6] X. Tang and S. T. Chanson, "Coordinated en-route web caching," *IEEE Transactions on Computers*, vol. 51, no. 6, pp. 595–607, 2002.

[7] E. J. Rosensweig and J. Kurose, "Breadcrumbs: efficient, best-effort content location in cache networks," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '09)*, pp. 2631–2635, April 2009.

[8] J. Dai, Z. Hu, B. Li, J. Liu, and B. Li, "Collaborative hierarchical caching with dynamic request routing for massive content distribution," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '12)*, 2012.

[9] P. Krishnan, D. Raz, and Y. Shavitt, "The cache location problem," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 568–582, 2000.

[10] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '10)*, March 2010.

[11] Y. Li, T. Lin, H. Tang, and P. Sun, "A chunk caching location and searching scheme in content centric networking," in *Proceedings of the IEEE International Conference on Communications (ICC '12)*, 2012.

[12] L. Fan, P. Cao, J. Almeida, and A. Z. Broder, "Summary cache: a scalable wide-area Web cache sharing protocol," *IEEE/ACM Transactions on Networking*, vol. 8, no. 3, pp. 281–293, 2000.

[13] J. Dilley, B. Maggs, J. Parikh, H. Prokop, R. Sitaraman, and B. Weihl, "Globally distributed content delivery," *IEEE Internet Computing*, vol. 6, no. 5, pp. 50–58, 2002.

[14] A. Wolman, G. M. Voelker, N. Sharma, N. Cardwell, A. Karlin, and H. M. Levy, "On the scale and performance of cooperative web proxy caching," *ACM SIGOPS Operating Systems Review*, vol. 33, no. 5, pp. 16–31, 1999.

[15] M. R. Korupolu and M. Dahlin, "Coordinated placement and replacement for large-scale distributed caches," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 6, pp. 1317–1329, 2002.

[16] A. X. Jiang and J. Bruck, "Optimal content placement for en-route web caching," in *Proceedings of the 2nd IEEE International symposium on network computing and applications*, pp. 9–16, 2003.

[17] H. Shen and S. Xu, "Coordinated en-route web caching in multiserver networks," *IEEE Transactions on Computers*, vol. 58, no. 5, pp. 605–619, 2009.

[18] H. Ahlehagh and S. Dey, "Video caching in radio access network: impact on delay and capacity," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '12)*, pp. 2276–2281, 2012.

[19] J. Erman, A. Gerber, M. Hajiaghayi, D. Pei, S. Sen, and O. Spatscheck, "To cache or not to cache: the 3G case," *IEEE Internet Computing*, vol. 15, no. 2, pp. 27–34, 2011.

[20] L. Yin and G. Cao, "Supporting cooperative caching in ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 5, no. 1, pp. 77–89, 2006.

[21] T. Hara, "Effective replica allocation in ad hoc networks for improving data accessibility," in *Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '01)*, pp. 1568–1576, April 2001.

[22] B. Tang, H. Gupta, and R. D. Samir, "Benefit-based data caching in ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 3, pp. 285–300, 2008.

[23] C. Chow, H. V. Leong, and A. T. S. Chan, "Distributed group-based cooperative caching in a mobile broadcast environment," in *Proceedings of the ACM 6th International Conference on Mobile Data Management (MDM '05)*, pp. 97–106, May 2005.

[24] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: wireless video content delivery through distributed caching helpers," in *Proceedings of the Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '12)*, 2012.

[25] N. Golrezai, A. F. Dimakis, and A. F. Molisch, "Device-to-device collaboration through distributed storage," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM '12)*, 2012.

[26] N. Golrezai, A. F. Dimakis, and A. F. Molisch, "Base-station assisted device-to-device communications for high-throughput wireless video networks," in *Proceedings of the IEEE ICC Workshop on Video-Aware Wireless Networks*, 2012.

[27] V. Jacobson, D. Smetters, J. Thornton, M. Plass, N. Briggs, and R. Braynard, "Networking named content," in *Proceedings of the ACM CoNEXT Student Workshop (CoNEXT '09)*, December 2009.

[28] A. Ghodsi and J. Wilcox, "Information-centric networking: seeing the forest for the trees," in *Proceedings of the 10th ACM SIGCOMM Workshop on Hot Topics in Networks*, November 2011.

[29] Y. Li, Y. Xu, T. Lin, G. Zhang, Y. Liu, and S. Ci, "Self assembly caching with dynamic request routing for information-centric networking," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM '13)*, 2013.

[30] G. Zhang, Y. Li, and T. Lin, "Caching in information centric networking: a survey," *Computer Networks*, vol. 57, pp. 3128–3141, 2013.

[31] I. Psaras, W. K. Chai, and G. Pavlou, "Probabilistic in-network caching for information-centric networks," in *Proceedings of the 2nd Edition of the ICN Workshop on Information-Centric Networking, SIGCOMM Workshop*, pp. 55–60, ACM, 2012.

[32] G. Carofiglio, L. Muscariello, and M. Gallo, "Bandwidth and storage sharing performance in information centric networking," in *Proceedings of the ACM Sigcomm Workshop on Information-Centric Networking (ICN '11)*, pp. 26–31, August 2011.

[33] S. Guo, H. Xie, and G. Shi, "Collaborative forwarding and caching in content centric networks," in *Proceedings of the 11th International IFIP Networking Conference*, 2012.

[34] Nanodatacenters, http://www.nanodatacenters.eu/.

[35] Boxee, http://www.boxee.tv/.

[36] W. Jiang, S. Ioannidis, L. Massouli, and F. Picconi, "Orchestrating massively distributed CDNs," in *Proceedings of the ACM CoNEXT Student Workshop (CoNEXT '12)*, 2012.

[37] D. Han, D. Andersen, M. Kaminsky, D. Papagiannaki, and S. Seshan, "Hulu in the neighborhood," in *Proceedings of the 3rd International Conference on Communication Systems and Networks (COMSNETS '11)*, January 2011.

[38] S. K. Fayazbakhsh, Y. Lin, A. Tootoonchian, and A. Ghodsi, "Less pain, most of the gain: incrementally deployable ICN," in *Proceedings of the ACM SIGCOMM Conference*, 2013.

[39] J. Kim, G. Lee, and J. Choi, "Efficient multicast schemes using in-network caching for optimal content delivery," *IEEE Communications Letters*, vol. 17, no. 5, pp. 1048–1051, 2013.

[40] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," in *Proceedings of the 18th Annual Joint Conference of the IEEE Computer and Communications Societie (INFOCOM '99)*, pp. 126–134, March 1999.

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

[42] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube traffic characterization: a view from the edge," in *Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference (IMC '07)*, pp. 15–28, October 2007.

[43] Y. Li, H. Xie, Y. Wen, and Z. Zhang, "Coordinating in-network caching in content-centric networks: model and analysis," in *Proceedings of the 33rd International Conference on Distributed Computing Systems (ICDCS '13)*, 2013.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 283852, 6 pages
http://dx.doi.org/10.1155/2013/283852

Research Article

A Source-Initiated On-Demand Routing Algorithm Based on the Thorup-Zwick Theory for Mobile Wireless Sensor Networks

Yuxin Mao and Ping Zhu

School of Computer and Information Engineering, Zhejiang Gongshang University, Xuezheng Street No. 18, Hangzhou, Zhejiang 310018, China

Correspondence should be addressed to Yuxin Mao; maoyuxin@zjgsu.edu.cn

Received 5 September 2013; Accepted 20 November 2013

Academic Editors: W. Sun, G. Zhang, and J. Zhou

The unreliability and dynamics of mobile wireless sensor networks make it hard to perform end-to-end communications. This paper presents a novel source-initiated on-demand routing mechanism for efficient data transmission in mobile wireless sensor networks. It explores the Thorup-Zwick theory to achieve source-initiated on-demand routing with time efficiency. It is able to find out shortest routing path between source and target in a network and transfer data in linear time. The algorithm is easy to be implemented and performed in resource-constrained mobile wireless sensor networks. We also evaluate the approach by analyzing its cost in detail. It can be seen that the approach is efficient to support data transmission in mobile wireless sensor networks.

1. Introduction

As a kind of wireless technology, wireless sensor networks (WSNs) [1, 2] are systems that comprise large numbers (usually hundreds or thousands) of wirelessly connected heterogeneous sensor nodes that are spatially distributed across a large field of interest. There is a wide range of applications where the WSNs are extensively used, and their development in other applications is still growing. Mobile wireless sensor networks (MWSNs) are a particular class of WSN in which mobility plays a key role in the execution of the application. In many cases, MWSNs suffer from link breakages and frequent changes of network topology. For example, a sensor node with a limited battery life may sleep periodically in order to reserve energy. A sensor node may also be blocked by data packets from its neighbours at some time or jammed by malicious nodes. Hence, a normal sensor node will lead to denial of service in those situations. Moreover, intermediate nodes are often required to carry out end-to-end communications since the transmission range of sensor nodes is also limited. Therefore, the intrinsic features of MWSNs make it hard to perform end-to-end communications, especially for large-scale data transmission [3, 4].

In this paper, we propose a novel approach of source-initiated on-demand routing [5–7] for MWSNs. We explore the Thorup-Zwick theory [8] to achieve efficient end-to-end communications in MWSNs. The remaining of the paper is organized as follows. Section 2 illustrates the network model and problem statement for the approach. In Section 3, we present an efficient algorithm for source-initiated on-demand routing in MWSNs. We evaluate the algorithm from the point of cost and complexity in Section 4 and discuss the simulation results in detail. Section 5 gives an overview of the related works. Section 6 concludes the paper with an outlook to future research directions.

2. Network Model and Problem Statement

In this paper, we consider a relatively simple MWSN model. $I = \{1, 2, \ldots, n\}$ and assume a MWSN with k nodes. We assume the whole network consists of three tiers (see Figure 1). In the bottom, there are a number of sensor nodes. Each node has a unique identity i ($i \in I$) in the network. Each node in the network is battery-powered and has limited computation and wireless communication capabilities. We assume that the locations of the sensor nodes are relatively static, rather than moving. Without confusion, we will also use n_i to denote the location of a sensor node n_i, $i \in I$.

There is a sink node in the top level. We assume that the sink is a center equipped with sufficient computation and

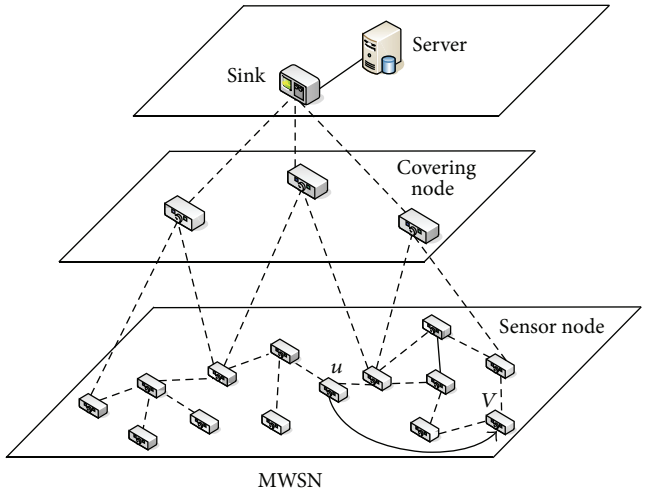


Figure 1: The three-tier network structure for MWSN.

```
Algorithm SIODRouting(u, v, s)
p ← checkCache(u)
If p is null
   routingReq(u, s)
   p ← pathQuery(u, v)
   notification(u, s, p)
end if
send data from u to v by p
r ← −1
if sending data success
   r ← 0
end if
return r
```

Algorithm 1: The source-initiated on-demand routing algorithm for MWSNs.

storage capabilities. Although the sink is able to communicate with each sensor node directly, direct communication between the sink and a sensor node is time consuming and energy consuming. For example, if a sensor node sends a large file (e.g., video file) to the sink, the energy of the sensor node will soon be exhausted. Therefore, we will decrease this kind of direct communication. Instead, the sink will allocate information from the covering nodes timely.

There are several *covering nodes* in the middle level, which are similar with cluster heads in clustering hierarchy or relay nodes in flat hierarchy. These nodes are only used to collect status information from sensor nodes, without any further processing or computation. Each covering node covers a part of the network with a number of sensor nodes. The placement of the covering nodes will ensure that all the sensor nodes in the network are covered. Also we will ensure that there are no more than η hops between a covering node and a sensor node that it covers (usually $\eta \leq 3$). Covering nodes are able to communicate with the sink directly.

Let x and y be two points in the Euclidean plane, then $[x, y]$ denotes the line segment connecting x and y, and $|x, y|$ denotes the Euclidean distance between x and y. Two sensor nodes n_i and n_j can communicate with each other if $|n_i, n_j| < R$, where R is the communication range of a sensor node in the MWSN.

The major task of a sensor node in the network is to communicate with other nodes and transmit data to others by routing. As we have mentioned before, intermediate nodes are often required to carry out end-to-end communications. Therefore, routing in this network should provide a *path* from source to destination and the path itself should be as short as possible. In this work, we do not consider the situation when there are selfish nodes in the network. We assume that each node is willing to cooperate with its neighbours.

3. Source-Initiated On-Demand Routing Algorithm

We attempt to find a short path for a source node in a MWSN by using some routing algorithm. The major difficulty of designing the routing algorithm is the cost in path construction. Due to the link breakages and frequent changes of network topology, a source node has to update its routing paths frequently, which is obviously time consuming. The situation becomes worse when the network grows up in size. In this work, we explore the Thorup-Zwick theory to solve the problem and achieve efficient routing in MWSNs.

3.1. Algorithm Overview. An overview of the source-initiated on-demand routing algorithm is given in Algorithm 1. Given a time t, the network topology of a MWSN can be denoted by a weighted undirected graph $G = (V, E)$. V is the sets of sensor nodes in the network. E is the sets of connection status among the nodes. When node u attempts to send data to node v, it first checks its local cache. If there is no existing routing path between them, u will send a routing request to the sink (s). Then the sink will query its local database that contains the data structure. Here the data structure is generated by preprocessing G. The sink will send back a notification to u, which contains the shortest path from u to v. Finally, u sends data to v by using the feedback path.

The algorithm is straightforward. The key point is how to perform path query from u to v. We will give the detailed explanation in Sections 3.3 and 3.4.

3.2. Status Allocation. As we have mentioned in Section 2, sensor nodes are not encouraged to communicate with the sink directly. However, the sink requires some basic information from sensor nodes in order to support source-initiated on-demand routing. This information contains the location of a sensor node as well as its connection status currently. The algorithm for this process of status allocation is illustrated in Algorithm 2.

The algorithm in Algorithm 2 is trivial. Each sensor sends a status vector to its covering node. The status vector contains the factors that have impacts on data communication. The status vector of a sensor node n_i can be formally represented by $v_i = \langle E_i, \mathrm{PRR}_i, L_i, C_i \rangle$, where E_i is the value of available energy of n_i, PRR_i denotes the packet reception ratio (PRR) at n_i, which is a metric for evaluating link quality, L_i is the load

```
Algorithm allocStatus(s)
for i ← 1 to p
   c ← c_i
   get covered sensor nodes of c as U_i = {u_1, u_2, ..., u_q}
   for j ← 1 to q
      u ← u_j
      v ← getStatusVector(u)
      info ← ⟨j, v⟩
      sendInfo(u, info, c)
   end loop
   forwardInfo(c, s)
end loop
```

Algorithm 2: Allocating status information from sensor nodes via covering nodes.

of n_i, and C_i denotes the connection status of n_i (the direct neighbors of the node).

For each window of Δw received packets at n_i, PRR_i is computed as follows:

$$\mathrm{PRR}_i(\Delta w) = \frac{\mathrm{Num}_{\mathrm{rp}}}{\mathrm{Num}_{\mathrm{sp}}}, \tag{1}$$

where $\mathrm{Num}_{\mathrm{rp}}$ denotes the number of successfully received packets, while $\mathrm{Num}_{\mathrm{sp}}$ the number of transmitted packets. For a given timeframe Δt, the load L_i is computed as follows:

$$L_i(\Delta t) = \frac{\mathrm{Num}_{\mathrm{rdp}}}{\mathrm{Num}_{\mathrm{lgp}}}, \tag{2}$$

where $\mathrm{Num}_{\mathrm{rdp}}$ denotes the number of relayed data packets (not locally generated), while $\mathrm{Num}_{\mathrm{lgp}}$ the number of locally generated packets.

After collecting the information, the covering node then forwards it to the sink together.

Moreover, there are different cases for status allocation in a MWSN. If we set η to 3, there are three kinds of cases for status allocation. Take the subnetwork shown in Figure 2 for example: (a) to the sensor nodes (n_3 and n_5) directly adjacent to a covering node (C), they are able to send status vectors to it. (b) To the sensor nodes (n_2, n_4 and n_6) directly adjacent to the ones in the first case, they could send status vectors to the covering node by two hops. (c) To the rest of the sensor nodes (n_1), they have to send status vectors by three hops.

3.3. Graph Construction. After allocating status from distributed sensor nodes, the sink is able to get the overall information of the network. Given a time t, the network topology of a MWSN can be denoted by a weighted undirected graph $G = (V, E)$. Assume $|V| = n$ and $|E| = m$. Each element in V denotes a sensor node in the MWSN and each element in E denotes a link between two nodes. For all $e_{ij} \in E$, we have the following equation:

$$e_{ij} = \begin{cases} w_{ij} \cdot \left|n_i, n_j\right|, & \left|n_i, n_j\right| < R, \\ \infty, & \left|n_i, n_j\right| \geq R. \end{cases} \tag{3}$$

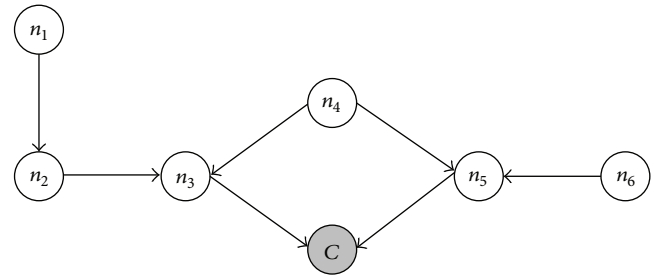


Figure 2: An example of status allocation by covering node.

It means that the distance between any two nodes in the graph is a weighted value. If the Euclidean distance between two nodes is greater than the communication rage R, we just set the distance value to be ∞ in the graph. The key to the graph construction is to fix the weight values for each edge in the graph. Weight is formally defined as follows:

$$w_{ij} = \alpha \cdot E_{ij} + \beta \cdot Q_{ij} + \gamma \cdot L_{ij}. \tag{4}$$

The weight depends on several factors. E_{ij} denotes the energy status for the two nodes. The value of E_{ij} is calculated by

$$E_{ij} = \frac{E_i - E_t}{E_i} \times \frac{E_j - E_t}{E_j}, \tag{5}$$

where E_i and E_j are the values of available energy for n_i and n_j and E_t is the energy required for an operation of data transmission. Q_{ij} denotes the link quality between n_i and n_j. Here we try to use software-based link quality estimators [9–13] to evaluate the link quality. We integrate the ETX estimator [14] to get an estimate of the link quality. Q_{ij} is calculated as follows:

$$Q_{ij} = \frac{1}{\mathrm{PRR}_i \times \mathrm{PRR}_j}, \tag{6}$$

where PRR_i reflects the uplink quality from n_i to n_j, while PRR_j the downlink quality from n_j to n_i. L_{ij} denotes the load status for the two nodes. The value of L_{ij} is calculated by

$$L_{ij} = \frac{1}{L_i \times L_j}. \tag{7}$$

α, β, and γ are coefficients for the weight and we have $\alpha + \beta + \gamma = 1$.

3.4. Graph Preprocessing. In order to perform efficient path query in the graph for a MWSN, we need to preprocess the weighted undirected graph G first. Assume $|V| = n$ and $|E| = m$. Thorup and Zwick in [8] have proposed an approach of preprocessing G in $O(knm^{1/k})$ expected time and constructing a data structure with size $O(kn^{1+1/k})$. Any subsequent path query can be answered approximately in $O(k)$ time. The approximate distance returned is of stretch at most $2k - 1$. Here k is an integer and $k \geq 1$. After allocating

```
Algorithm preproGraph(s)
given a integer k ≥ 1
given a time t, represent the network topology as G = (V, E)
prepro_Thorup-Zwick(V, E, k)
store the data structure in the local database of s
```

Algorithm 3: Preprocessing the graph structure of a MWSN.

```
Algorithm pathQuery(u, v)
given a integer k ≥ 1
δ(u, v) ← dist_Thorup-Zwick(u, v, k)
get the path p from u to v
output the edges on p
```

Algorithm 4: Path query on the graph structure of a MWSN.

status information, the sink is able to get the topology of the MWSN. Therefore, we could use the Thorup-Zwick theory directly to preprocess the graph structure of the MWSN (see Algorithm 3).

3.5. Path Query. After preprocessing the graph structure of the MWSN, the sink is able to answer a path query in linear time. The algorithm of path query is given in Algorithm 4. Here we make use of the Thorup-Zwick theory to perform path query in the database structure returned by the preprocessing algorithm.

4. Evaluation

In this section, we mainly evaluate the performance of the proposed algorithm by analyzing its complexity and cost, against existing routing algorithms for MWSNs.

The cost for the proposed algorithm is mainly generated from four activities: status allocation, graph construction, graph preprocessing, and path query. The first three activities are preprocessing ones. We try to evaluate the cost for these four activities by analyzing the time complexity.

We evaluate the cost for status allocation at first. According to Section 3.2, we have the following equation:

$$n = N_1 + N_2 + N_3. \tag{8}$$

Here N_1, N_2, and N_3 denote the number of sensor nodes in the three cases in status allocation, respectively. Assume that the one hop (sensor node to sensor node or sensor node to covering node) cost for status allocation is Δc_1, and the one hop between covering node and sink is Δc_2. Then the total cost ΔC for status allocation is as follows:

$$\Delta C = N_1 \Delta c_1 + N_2 2\Delta c_1 + N_3 3\Delta c_1 + n\Delta c_2. \tag{9}$$

We could reduce (9) into

$$\Delta C = \left(2\Delta c_1 + \Delta c_2\right) n + \Delta c_1 \left(N_3 - N_1\right). \tag{10}$$

As $|N_3 - N_1| < n$, then we have:

$$\begin{aligned} \Delta C &= \left(2\Delta c_1 + \Delta c_2\right) n + \Delta c_1 \left(N_3 - N_1\right) \\ &< \left(2\Delta c_1 + \Delta c_2\right) n + \Delta c_1 n \\ &= \left(3\Delta c_1 + \Delta c_2\right) n. \end{aligned} \tag{11}$$

Therefore, the cost for status allocation is $O(n)$.

We have to compute the weight for each edge in graph construction. Therefore, the cost for graph construction is $O(m)$. According to [8], the cost for graph preprocessing is $O(knm^{1/k})$ and the approximate cost for answering path query is $O(k)$. Finally, we can get the preprocessing cost as $O(n) + O(n) + O(m) + O(knm^{1/k}) = O(2n + m + knm^{1/k}) = O(n+m+knm^{1/k})$ and the query cost as $O(k)$. We can see that the activities in our algorithm have linear cost except graph preprocessing. If we set k to be a large integer, then the cost for graph preprocessing is also not very high and acceptable to MWSNs.

5. Related Works

Generally, existing routing protocols for WSNs fall into two categories: table-driven and on-demand routing [7] based on when and how the routes are discovered. For the table-driven routing protocols, consistent and up-to-date routing information for all the sensor nodes are maintained at each mobile host. It has been shown in [15] and stated in [16] that on-demand routing protocols can perform better than table-driven protocols in WSNs.

There have been many on-going research efforts in on-demand routing for WSNs or wireless networks. For example, the ad hoc on-demand distance vector routing (AODV) [17] is an improvement of the destination-sequenced distance-vector (DSDV) algorithm. AODV minimizes the number of broadcasts by creating routes on-demand as opposed to the DSDV that maintains a list of all the routes. The dynamic source routing protocol (DSR) [18] is another on-demand routing protocol. A sensor node maintains the route caches containing the source routes that it is aware of. The mobile host updates the entries in the route cache as soon as it learns about new routes. The temporally ordered routing algorithm (TORA) [19] is a highly adaptive, efficient, and scalable distributed routing algorithm based on the concept of link reversal. TORA is proposed for highly dynamic mobile and multihop wireless networks. It is a source-initiated on-demand routing protocol. However, none of these algorithms take the case that how to perform efficient routing in a WSN with link breakages and frequent changes of network topology into consideration.

Some routing algorithms just enhance the abovementioned ones with fault-tolerant or energy-balancing mechanism [20–23]. However, these algorithms are not able to provide short path for routing or they do not provide an efficient way for constructing short path. Moreover, there are also a few algorithms about shortest path routing [24, 25]; however, these algorithms fall short of efficiency due to high cost or large complexity.

Compared with existing works in this field, our approach uses a novel graph-based mechanism that makes full use of the Thorup-Zwick theory to improve the end-to-end communication in MWSNs. The algorithm is of time efficiency and the overhead is acceptable to large-scale MWSNs. The advantage of our approach is that we can still achieve efficient routing even the number of the nodes in a MWSN grows up.

6. Conclusion

In this study, we mainly present a novel source-initiated on-demand routing algorithm for efficient data transmission in MWSNs. We explore the Thorup-Zwick theory to achieve source-initiated on-demand routing with time efficiency. With this algorithm, we are able to find out shortest routing path between source and target in a network and transfer data in linear time. The algorithm is also easy to be implemented and performed in resource-constrained MWSNs. We also evaluate the algorithm by analyzing its time complexity in detail. It can be seen that the approach is efficient to support end-to-end data communication in MWSNs. Compared with existing works in this field, our approach is of time efficiency and the overhead is acceptable to large-scale MWSNs. The advantage of our approach is that we can still achieve efficient routing even the number of the nodes in a MWSN grows up.

Future works may include: (1) improving the efficiency of the algorithms to reduce the operations of graph preprocessing; (2) considering a more complex MWSN model to implement and evaluate the approach; and (3) considering the security problem of routing in MWSNs.

Acknowledgments

This work is partially supported by grants from a NSFC Program (no. NSFC61003309), Science and Technology Department of Zhejiang Province Program (no. 2010C13005), and Key Science and Technology Innovation Team of Zhejiang Province Program (2010R50041-03).

References

[1] S. Corson and J. Macker, "Wireless sensor networking (WSN): routing protocol performance issues and evaluation considerations," RFC, 2501, 1999.

[2] S. Basagni, M. Conti, S. Giordano, and I. Stojmenovic, Eds., *Wireless Sensor Networking*, IEEE Press/Wiley, 2004.

[3] K. Yang, H. H. Chen, and Y. Wu, "QoS-aware routing in emerging heterogeneous wireless networks," *IEEE Communications Magazine*, vol. 45, no. 2, pp. 74–80, 2007.

[4] Q. Zhang and Y. Q. Zheng, "Cross-layer design for qos support in multihop wireless networks," *Proceedings of the IEEE*, vol. 96, no. 1, pp. 64–76, 2008.

[5] V. O. K. Li and Z. X. Lu, "Ad hoc network routing," in *Proceeding of the IEEE International Conference on Networking, Sensing and Control*, pp. 100–105, March 2004.

[6] R. Ramanathan and J. Redi, "A brief overview of ad hoc networks: challenges and directions," *IEEE Communications Magazine*, vol. 40, no. 5, pp. 20–22, 2002.

[7] E. M. Royer and C. K. Toh, "A review of current routing protocols for ad hoc mobile wireless networks," *IEEE Personal Communications*, vol. 6, no. 2, pp. 46–55, 1999.

[8] M. Thorup and U. Zwick, "Approximate distance oracles," *Journal of the ACM*, vol. 52, no. 1, pp. 1–24, 2005.

[9] A. Woo and D. Culler, "Evaluation of efficient link reliability estimators for low-power wireless networks," Tech. Rep. UCB/CSD-03-1270, EECS Department, University of California, Berkeley, Calif, USA, 2003, http://www.eecs.berkeley.edu/Pubs/TechRpts/2003/6239.html.

[10] D. Lal, A. Manjeshwar, F. Herrmann, E. Uysal-Biyikoglu, and A. Keshavarzian, "Measurement and characterization of link quality metrics in energy constrained wireless sensor networks," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM'03)*, pp. 446–452, December 2003.

[11] R. Fonseca, O. Gnawali, K. Jamieson, and P. Levis, "Four bit wireless link estimation," in *Proceedings of the 6th Workshop on Hot Topics in Networks*, 2007.

[12] A. Cerpa, M. Potkonjak, J. L. Wong, and D. Estrin, "Temporal properties of low power wireless links: modeling and implications on multi-hop routing," in *Proceedings of the 6th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC '05)*, pp. 414–425, May 2005.

[13] M. Senel, K. Chintalapudi, D. Lal, A. Keshavarzian, and E. J. Coyle, "A Kalman Filter based link quality estimation scheme for wireless sensor networks," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '07)*, pp. 875–880, Washington, DC, USA, November 2007.

[14] D. S. J. de Couto, D. Aguayo, J. Bicket, and R. Morris, "A high-throughput path metric for multi-hop wireless routing," in *Proceedings of the 9th Annual International Conference on Mobile Computing and Networking (MobiCom '03)*, pp. 134–146, ACM, New York, NY, USA, September 2003.

[15] A. H. Altalhi and G. G. Richard III, "Load-balanced routing through virtual paths: highly adaptive and efficient routing scheme for ad hoc wireless networks," in *Proceedings of the 23rd IEEE International Performance, Computing, and Communications Conference (IPCCC '04)*, pp. 407–413, April 2004.

[16] P. Johansson, T. Larsson, N. Hedman, B. Mielczarek, and M. Degermark, "Scenario-based performance analysis of routing protocols for wireless sensor networks," in *Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom '99)*, pp. 195–206, August 1999.

[17] C. E. Perkins, E. M. Belding-Royer, and S. R. Das, "Ad hoc on-demand distance vector (AODV) routing," RFC, 3561, 2003.

[18] D. Johnson, Y. Hu, and D. Maltz, "The dynamic source routing protocol (DSR) for wireless sensor networks for IPv4," RFC, 4728, 2007.

[19] V. Park and S. Corson, "Temporally-ordered routing algorithm (TORA)," IETF internet draft, draft-ietf-WSN-tora-spec-05.txt, 2002, http://www.ietf.org/proceedings/53/I-D/draft-ietf-manet-tora-spec-04.txt.

[20] Z. Che-Aron, W. F. M. Al-Khateeb, and F. Anwar, "The enhanced fault-tolerant AODV routing protocol for wireless sensor network," in *Proceedings of the 2nd International Conference on Computer Research and Development (ICCRD '10)*, pp. 105–109, Kuala Lumpur, Malaysia, May 2010.

[21] Z. Che-Aron, W. F. M. Al-Khateeb, and F. Anwar, "An enhancement of fault-tolerant routing protocol for wireless sensor network," in *Proceedings of the International Conference on Computer and Communication Engineering (ICCCE '10)*, pp. 1–3, Kuala Lumpur, Malaysia, May 2010.

[22] W. B. Pöttner, O. Wellnitz, and L. Wolf, "QoS-AODV6E: an energy-balancing QoS routing scheme for WSNs," in *Proceedings of the 7th International Conference on Networked Sensing Systems (INSS '10)*, pp. 81–84, Kassel, Germany, June 2010.

[23] S. Ren, H. Han, B. Li, J. Lu, C. Peng, and W. Dou, "An improved wireless sensor networks routing protocol based on AODV," in *Proceedings of the IEEE 12th International Conference on Computer and Information Technology*, pp. 742–746, Chengdu, China, 2012.

[24] B. Musznicki, M. Tomczak, and P. Zwierzykowski, "Dijkstra-based localized multicast routing in wireless sensor networks," in *Proceedings of the 8th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP '12)*, pp. 1–6, Poznan, Poland, 2012.

[25] K. S. Shivaprakasha and M. Kulkarni, "Energy efficient shortest path routing protocol for wireless sensor networks," in *Proceedings of the International Conference on Computational Intelligence and Communication Systems (CICN '11)*, pp. 333–337, Gwalior, India, October 2011.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 704865, 11 pages
http://dx.doi.org/10.1155/2013/704865

Research Article

Multiple-Feature Extracting Modules Based Leak Mining System Design

Ying-Chiang Cho and Jen-Yi Pan

Department of Electrical Engineering, CCU University, Chia-Yi 62102, Taiwan

Correspondence should be addressed to Ying-Chiang Cho; yingchiang44@hotmail.com

Received 23 September 2013; Accepted 27 October 2013

Academic Editors: W. Sun, G. Zhang, and J. Zhou

Over the years, human dependence on the Internet has increased dramatically. A large amount of information is placed on the Internet and retrieved from it daily, which makes web security in terms of online information a major concern. In recent years, the most problematic issues in web security have been e-mail address leakage and SQL injection attacks. There are many possible causes of information leakage, such as inadequate precautions during the programming process, which lead to the leakage of e-mail addresses entered online or insufficient protection of database information, a loophole that enables malicious users to steal online content. In this paper, we implement a crawler mining system that is equipped with SQL injection vulnerability detection, by means of an algorithm developed for the web crawler. In addition, we analyze portal sites of the governments of various countries or regions in order to investigate the information leaking status of each site. Subsequently, we analyze the database structure and content of each site, using the data collected. Thus, we make use of practical verification in order to focus on information security and privacy through black-box testing.

1. Introduction

This paper discusses the issues of e-mail address leakage [1] and SQL injection [2] in online information security. Hackers commonly operate by means of social networks, APT (advanced persistent threat) and spam, and all of these methods begin with the acquisition of the e-mail addresses of their victims [3, 4]. Possible reasons for e-mail addresses being detected are (1) the web programmer not filtering the symbol "@", the most significant character of an e-mail address, which can be captured by the 24-hour running crawler mining system, and (2) website programmers not being sufficiently rigorous in implementing the "robots.txt" file that prevents crawler programs from accessing the site. Before such automated programs access a page, they determine whether the site contains the robots.txt file that limits their access to specific sites [5]. All normal and well-intentioned programs follow the command of this file [6]; however, the robots.txt is not always enforceable, which means that some spammers and other malicious programs are still able to ignore it [7]. Therefore, it is necessary for certain website programmers to protect confidential online information with a password or an improved program writing style [8, 9].

The Internet environment today is very different from how it was 20 years ago. Firstly, different kinds of automated programs for data collection are continually being developed. Malicious users are able to spread hazardous e-mails after having gathered e-mail address information, accompanied by automatic sending programs, which is a serious information security problem [1, 3, 10]. Secondly, the database that stores the core information for each site requires sufficient protection, and therefore, website programmers need to attempt to prevent malicious pervasive injection attacks by means of improved programming styles [11].

To solve above problems, this study implements a crawler mining system equipped with black-box testing thinking [12–14]. This system consists of three main modules: dynamic scanning module, static mining module, manual operating module, and six functions: data syntax analysis function, hidden page exploring function, multidomain searching on one IP function, specific file searching function, a search

engine assistant function, and a website vulnerability updating function. This study analyzes governmental web portals as measured objects, with 24 hours as a unit, where the number of e-mail addresses collected and injectable URLs detected in each is calculated in order to compare the degree of concern around information security in different countries or regions.

This paper proceeds as follows. Section 2 discusses the previous algorithms. Section 3 clarifies the system design analysis. Section 4 presents experiment and result analysis, and Section 5 concludes.

2. Previous Algorithms

There are currently several searching strategies commonly used by web crawlers, which are discussed below.

Algorithm Based on IP Address. The web crawler obtains a starting IP address, and then proceeds to search the file document of each website included in this IP address one at a time. This method does not take into consideration the URL addresses from pages that lead to other websites. The advantage of this method is the ability to search the pages that cannot be hyperlinked; the disadvantage is that it cannot crawl the entire website. Some modules of our proposed system adopt this strategy.

Algorithm Based on Breadth-First Searching. In the crawling process, after completing the searching of the first level, we shall move on to the next level, and so on, until the searching is completed. In order to cover as many pages as possible, we generally use breadth-first search. Other research has applied breadth-first search to focus the crawler [15, 16]. The main idea behind this method is that there is a large possibility of the starting URL having a theme connection with the pages that are within a certain linking distance [15, 17, 18]. The crawler of the breath-first search gathers all the pages of the starting page and chooses one link from which to take all the pages. Owing to this, the web crawler is capable of speeding up parallel processing. Some modules of our system make use of this strategy.

Algorithm Based on Depth-First Searching. The aim of such an algorithm is to reach the end of searched websites, including those websites that do not contain any hyperlinks. When the hyperlink of a page is chosen, the pages connected to it will undergo a depth-first search (DFS). Before searching the rest of the links, the crawler must completely search this single chain. The crawler will move to the end, along the hyperlink from the page, and then return to another page, continually selecting hyperlinks. When there are no further hyperlinks to choose, the search comes to an end [19, 20]. The advantage of this method is the ability to search through an entire website. The disadvantage is that the crawler may easily become trapped, since the structure of some websites is rather complex and deep.

Algorithm Based on Best-First Searching. This method makes use of the specific webpage analysis algorithm to set a standard value against which the target directed by the URL is evaluated, gathering those URLs that are determined to be above the standard value. The algorithm is limited to the pages whose relation is larger than the standard value set by the webpage analysis algorithm. The disadvantage of this method is that it is likely to ignore many pages during the crawling process, as it is a partial searching method [21]. Therefore, it should be used alongside other application methods in order to achieve optimum performance [22, 23].

Algorithm Based on Reinforcement Learning. Rennie and McCallum incorporated reinforcement learning into the web crawler's learning process [24]. This method is based on machine learning and is divided into the training phase and the crawling phase. It makes use of reinforcement learning to calculate the future return value, the link for which is provided to form the training module. Thereafter, it predicts the actual future return value by using the training module in the crawling phase, in order to determine the visiting priority of the link.

Algorithm Based on the Context Graph. The algorithm for the web crawler based on the context graph is able to determine the searching direction by means of the Q value, but it is incapable of estimating the distance to the target page. Therefore, Diligentim et al. proposed a new searching strategy based on the context graph [22, 25, 26] that estimates the distance to the target page by forming a typical webpage under the context graph. The closer the page, the sooner it is crawled, with the closest being crawled first.

Algorithms Based on Content Evaluation or Link Structure Evaluation. Representative methods based on content evaluation include the Fish-search algorithm [27], the shark-search algorithm [28], the neural network search algorithm, and the web hyperlink analysis algorithm [29–32], while Page Rank [30] and Hits [33] are algorithms based on link structure evaluation. The Page Rank algorithm attempts to determine the importance of a page based on other associated pages of a good quality [34, 35]. The basic idea behind the Hits algorithm is to use the reference link among pages to obtain useful information hidden within the pages through two weight evaluations, Authority and Hub, in order to evaluate the quality of pages [36]. The results obtained from Hits are highly dependent on the chosen starting figures. It is easy to generate theme drift, since the theme authority differs between pages, which aims at different themes. In order to resolve the atmosphere above, the ARC algorithm [37] as well as the Clever algorithm [38] have been developed, which take into consideration the performance of the theme search.

Algorithm Based on Classifier-Guided Crawlers. Although the algorithm based on classifier-guided crawlers is an effective method, there has been little research relating to it [39]. Chakrabarti et al. proposed the focused crawling system [25], which is based on the Naive Bayes classifier. Menczer and Belew presented the Infospiders system [40], which is based on the agent formulae of a genetic algorithm and reinforcement learning. This model is composed of collateral crawls, each of which has its own prediction classifier. Johnson et al.

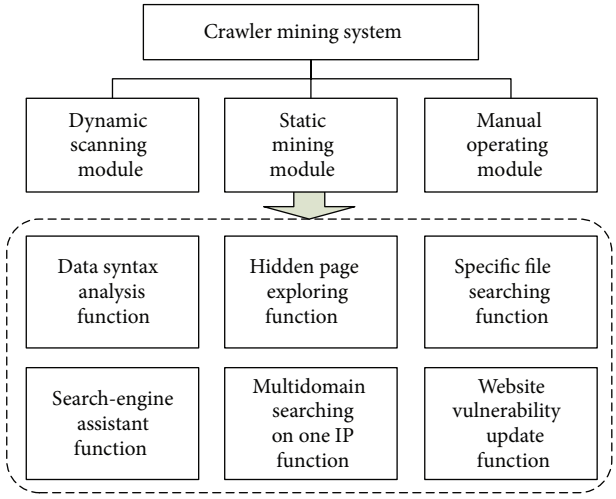


Figure 1: Crawler mining system framework.

Figure 2: Crawler mining system interface.

proposed theme crawls based on a support vector machine [41].

Other Algorithms. In addition to the above, many other web crawler algorithms have been developed. For example, [42] described the case of certain page contents being connected to irrelevant pages. Therefore, an algorithm is needed for crawlers to bypass irrelevant pages in order to reach the related ones. There is also an algorithm that limits the scope of themes that crawlers focus on, such as DNS, link class, IP, or national language. Some modules of our system adopt this strategy.

3. System Design

This research implements the crawler mining system, which was developed according to the algorithms above, using the JAVA SE7 version with more than 70,000 rows of coding. It can be installed normally in JRE (Java Runtime Environment) on WinXP/Vista/Win7. This system contains three modules and six functions, as shown in Figure 1. Figure 2 shows the main interface of the system, which is divided into three parts:

Figure 3: Official websites detection.

Figure 4: Official website stores database password in plain text.

the browser is in the middle, on the right is the URL queue to be examined, and the list of vulnerable information appears at the bottom.

The crawler mining system is capable of carrying out injections to penetrate databases such as MS-SQL, MySQL, Oracle, PostgreSQL, SQLite, and Access, as well as web languages such as ASP, ASPX, PHP, JSP, and CGI.

3.1. Design Principle Analysis

3.1.1. Dynamic Scanning Module. During the browsing process, the system automatically detects the current browsers one at a time, which makes it easy for the user to be aware of leaks that may occur in the website. Attacks that disclose information from websites can easily take place. Figure 3 provides a basic example, showing how the crawler mining system searches in Google for official websites. The system will immediately filter all the URLs that appear in the browser. When the system obtains a new URL, it places it into the queue list, and vulnerable URLs are placed into the injection list for further analysis. Figure 4 shows the results of the further analysis of the vulnerable websites, where the crawler mining system checks the database lists of the websites [43]. The database password for this website can be found as it is saved in plain text, which does not provide adequate security. When malicious attackers obtain this information, they are able to infiltrate this website immediately.

3.1.2. Static Mining Module. This type of module is used to conduct deep digging in specific websites, with its purpose being classified as static searching. The sensitive information function decides whether it is followed by the robots.txt of the website. Robots.txt is a document coded by ASCII that is stored in the website's root directory and tells the web crawler which content can or cannot be accessed [7]. Other hackers filter the e-mail accounts that are leaked from

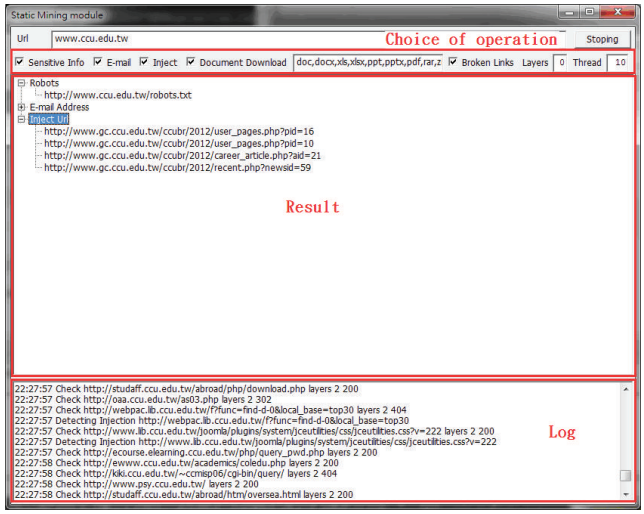


Figure 5: Static mining module.

Figure 6: E-mail address leak from the website.

the website, inspecting risky injection URLs, downloading documents with various filename extensions, or picking up the broken links containing private information. Figure 5 illustrates the static mining module. The crawler mining system is able to filter the e-mail addresses that are leaked from websites; this process is illustrated in Figure 6. The system is capable of determining injection weaknesses or the major loopholes of the website, as shown in Figure 7. These injection URLs can be used in conjunction with the built-in data programming analysis function in order to inject into the database of the website to obtain the usernames and e-mail accounts stored within it. Figure 8 demonstrates that, when malicious users obtain this information, they may perform various malicious attacks, such as spreading e-mails with viruses, conducting social engineering attacks to gain users' passwords, or obtaining personal information by infiltrating the database of the website [3, 10]. The more injection URLs that are detected by the crawler mining system, the more logical programming errors this website contains, leading to the invasion risk being relatively high.

3.1.3. Manual Operating Module. All of the functions that constitute the dynamic scanning module and the static mining module are automatic; however, human logic is the most important factor when it comes to web security [44]. The various functions that can be conducted by humans form the manual operation module.

3.2. Function Analysis

3.2.1. Data Syntax Analysis Function. Based on SQL injection [45, 46], we use the system to conduct an examination of the database of a website [11]. When an issue in the programming of the website causes the system to be incapable of determining whether the figures injected are attached to SQL orders, these attached orders will run normally, since the database does not detect the injection. In this way, illegal users are able to obtain database information, which causes serious

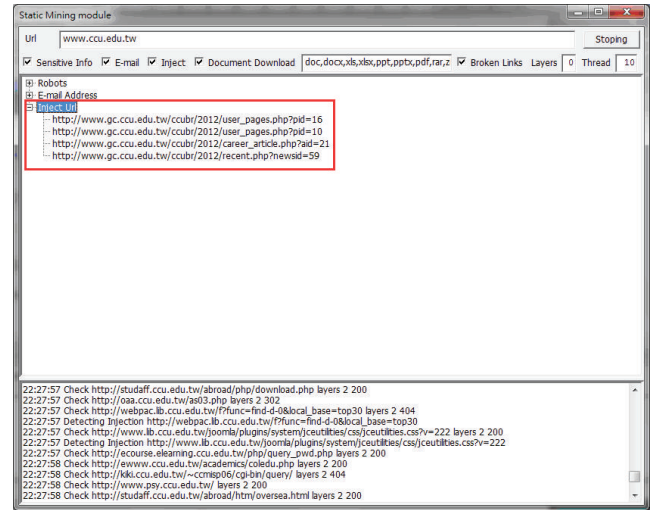


Figure 7: List of SQL injection URLs.

problems in information security [47]. Users may then be hacked because their information is leaked, the website structure is invaded illegally, or the administrator's account is altered to gain higher authority in order for malicious links or cross-site scripting to be added to the website [48]. When the system conducts the examination, the first step is to search for the injection point, to judge whether the website contains any harmful designs. Thereafter, a syntax dictionary is added to this system through the URL in order to block information and automatically judge from the feedback of the website whether anything will be achieved by conducting digging. The following shows the basic syntax injection:

(i) http://www.xxxx.gov.tw/fwdt/fw/list_show.asp?catei d=457 and 1=1

(ii) http://wj.xxxx.gov.tw/xxqk/InfoPublic/PublicView. asp?id=60 and user=0 and '8'='8'.

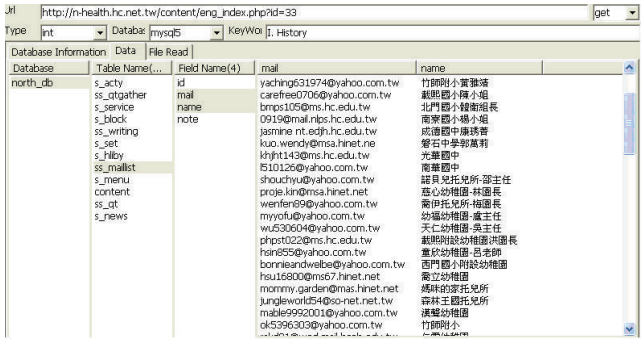


Figure 8: E-mail address leak in the database.

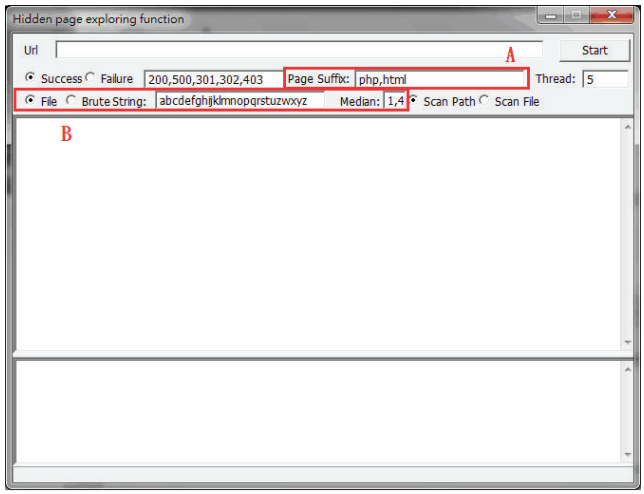


Figure 9: Interface for hidden page scanning.

The information leaks are evaluated according to the string of the syntax dictionary. When the system indicates that there is value in continuing digging, the type of database this website makes use of will be determined by means of the different functions defined by the database. For example, MS-SQL and MySQL use len( ) to calculate objective length, while Oracle does this according to length. Therefore, when using a len("s") = 1 for examination, normal feedback is obtained, which means the target website uses a MS-SOL or MySQL database; however, when this is not the case, the website may use Oracle. Additional functions defined by the database are also able to perform these evaluations. Once the database type has been established, the form fields of the database are analyzed, and finally, the figures and information of the database are obtained [43, 45, 46].

3.2.2. Hidden Page Exploring Function. There is some website information that is not easily detected or obtained by crawlers [49], such as paid journal articles saved in libraries or membership accounts of commercial websites, and this content is known as the Deep Web [50]. Some website designers believe that the website administration page not being obviously shown prevents easy browsing by users, and thus provides information security. However, the administration page may still be found by using specific filter methods. This system employs such methods to conduct hidden page exploring for examination and analysis of information. The first method is that of using syntax dictionary to explore the administration page according to the page naming rules that are used most, while the second one is exhaustive searching, using characters set up by users to perform permutation combinations.

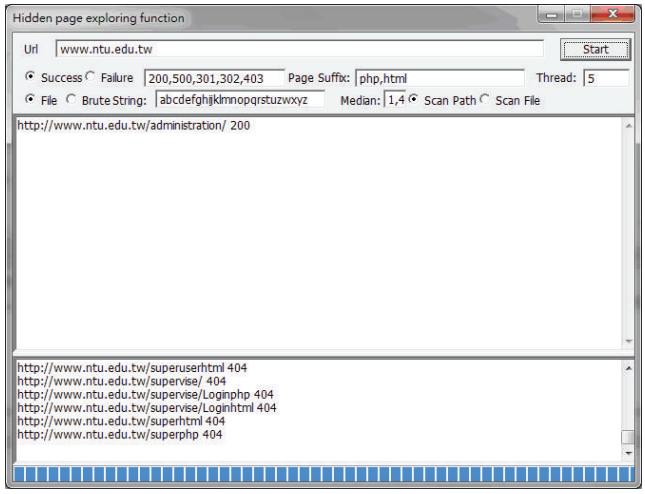


Figure 10: Category page.

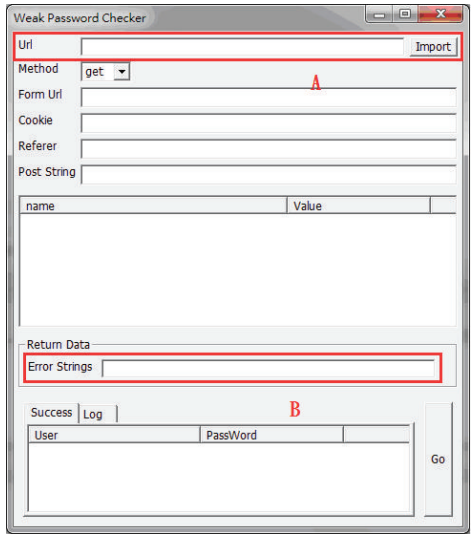


Figure 11: Interface for detecting a weak password.

Figure 9 shows the interface for the scanning of hidden pages. Area A is an addition to the interface for entering a web language, as there is an ever-increasing amount of web languages, such as HTML, ASP, PHP, CGI, and JSP, being used. In area B, when searching for hidden pages, the page name defined by the syntax dictionary is selected for scanning or exhaustive searching, Exhaustive searching examines the given characters and length one at a time; therefore, it takes more time, but has a wider coverage. The page names defined in the syntax dictionary can be added or deleted according to the naming rules that are used the most at any time.

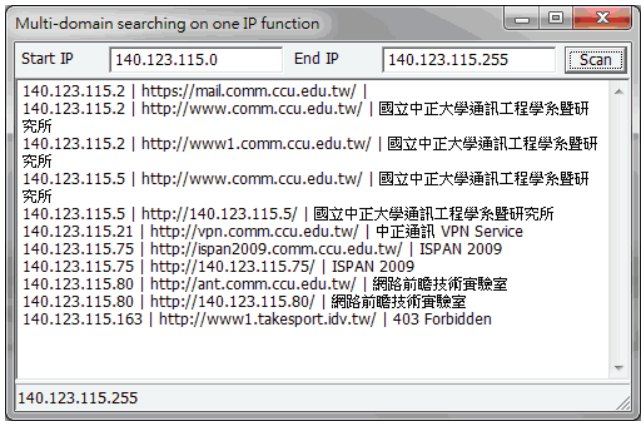


Figure 12: Multidomain searching on one IP function.

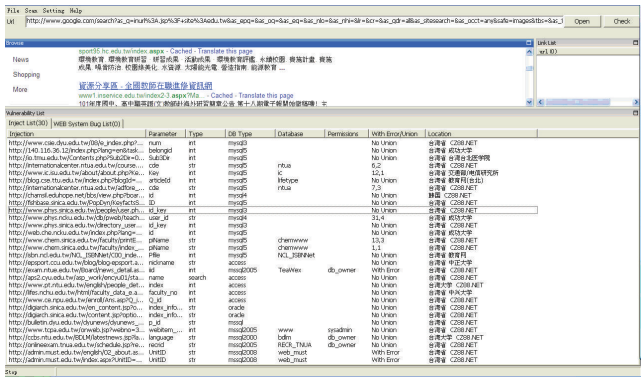


Figure 13: Possibility of leaking in databases.

The category page for the National Taiwan University website is provided as an example in Figure 10. The administration category page link can be obtained using the hidden page exploring function, by scanning the built-in syntax dictionary. In this example, if the administration interface is not adequately protected, it can easily be attacked by malicious users.

Many modern websites request an account name and password for authority control to be carried out. The proposed system contains a function aimed at the Deep Web that provides enterprises or persons with a feature for testing whether any password used is weak.

The URL provided in area A will be analyzed according to the variables and cookies related to input on the webpage. When incorrect account names and passwords are entered into area B, the webpage will show the corresponding string, which can be tested using the given ID/PWD dictionary. If a password in the database is automatically detected as being vulnerable to web crawlers logging on, we shall inform the personnel to change the password in order to protect personal information. Figure 11 shows the interface for detecting a weak password.

3.2.3. Multidomain Searching on One IP Function. Multidomain searching on one IP function built into the web crawler

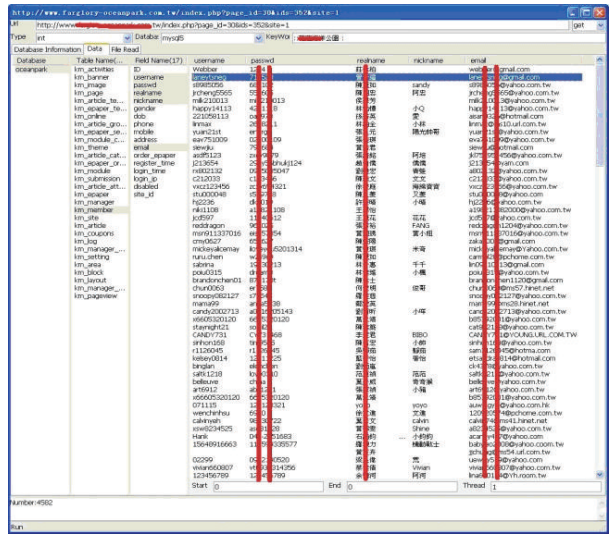


Figure 14: Account names and passwords of a database, stored using plain text.

system can be used to explore all the websites in a particular network segment. However, we must know which pages are under service and we must be able to avoid turning on more than two website services from certain servers or IPs; that is, when the TCP PORT is turned on it will not necessarily be in port 80. It can be seen in Figure 12 that there are 11 websites in this network segment in total, with their respective domain names being shown. There are three websites with the IP address 140.123.115.

3.2.4. Specific File Searching Function. In this study, we conduct a DFS on file types of which there are rich resources on the internet. If we wish to search for the file types of PDF, MS-Word, or MS-Excel, for example, we can use the specific file searching function to limit the scope. We are then able to conduct resources searching through the preferred need setup in order to find the document files that are associated with the preferred conditions.

3.2.5. Search Engine Assistant Function. Modern search engines, such as Google, YAHOO, Bing, and Baidu, are extremely powerful and convenient. We are therefore able to test the government-related websites in Taiwan by means of keyword searching and this system's dynamic scanning module. We do this by entering the keywords "inurl:php? site:gov.tw XX events" in Google to find out XX events relating to recently published material on government-related websites in Taiwan. In this way, we are able to conduct scanning of a specific issue or topic. The question is how should we order the keywords in the Google search engine? We provide an example. We enter the words "Inurl:.php?|.asp?|.jsp?|.apsx?site:edu.tw" into the Google search engine, because we wish to focus on dynamic-type pages for examination, such as PHP, ASP, JSP, or ASPX. In our experiment, we examined 800 samples of data from academic webpage databases of Taiwan; therefore, 200 data items for

```
<?xml version="1.0" encoding="utf-8"?>
<config>
<name> Name of Vulnerability </name>
    <date> Releasing Date of Vulnerability </date>
    <author> Author </author>
    <version> Version Number </version>
<type> Type of Vulnerability </type>
<description> Description of Vulnerability </description>
<file> File Causing Vulnerability </file>
<bugfile> The URL used for Vulnerability Testing </bugfile>
    <successkeyword> Successful keyword shown on the page after an error appears </successkeyword>
</config>
```

Algorithm 1

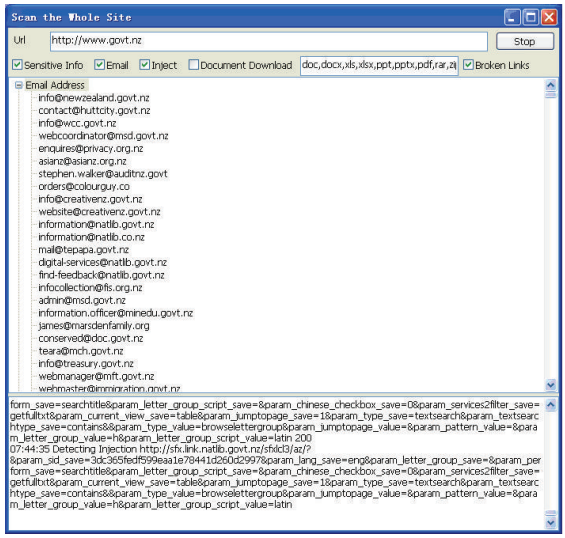


Figure 15: Leaked e-mail addresses in portal website of New Zealand.

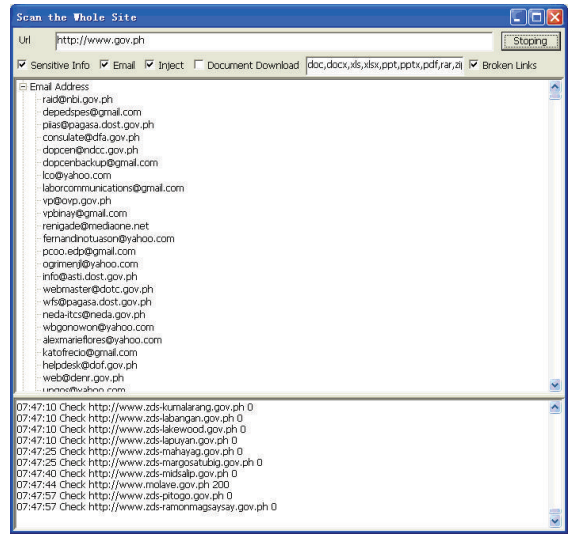


Figure 16: Leaked e-mail addresses in portal website of Philippines.

each type of page. Of the 800 data samples, 76 showed a high possibility of leaking occurring, with 31 being found in PHP; 26 in ASP; 12 in ASPX, and 7 in JSP. These figures are displayed in Figure 13.

Using the information of the injection URL, we can apply the data syntax analysis function in our web crawler system, as well as the SQL injection technique, in order to obtain the database contents. Within the Taiwan academic network, 5% of websites still store the account names and passwords of their databases using plain text, as shown in Figure 14. Thus, users' account names and passwords combined are likely to be obtained due to flaws in the design logic of the database.

3.2.6. Website Vulnerability Updating Function. New leaks on the internet are researched by experts regularly. We use the loophole published in the Exploit Database [51] as an example and convert this latest loophole into the forms that are supported by our web crawler system. The forms we apply are shown in Algorithm 1. Note that we ignore the encryption algorithm section.

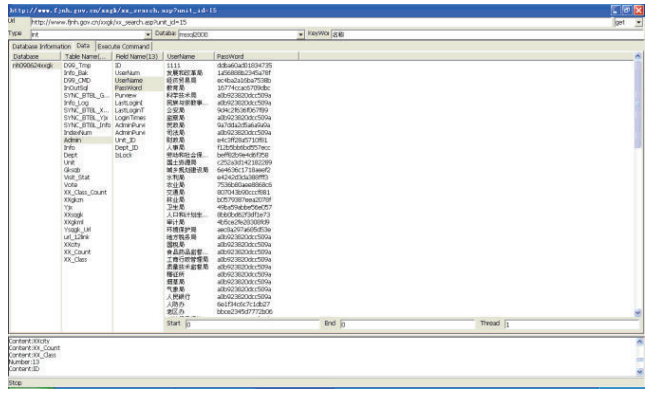


Figure 17: Content of the database obtained from URLs.

We are able to verify this loophole example by means of the instantaneous updating of the loophole database. We do not rely on the injection technique alone, as this is very similar to instantaneously updating virus codes.

Table 1: Testing results of 24 countries or regions.

Country or region	Government portal website address	Number of leaked e-mail addresses	Number of SQL injections
Canada	http://www.canada.gc.ca	1054	1295
PRC	http://www.gov.cn	3287	121
Germany	http://www.bundesregierung.de/	0	0
UK	http://www.direct.gov.uk/en/index.htm	1177	0
US	http://www.usa.gov	748	0
Taiwan	http://www.gov.tw	779	15
Singapore	http://www.gov.sg	0	0
Japan	http://www.e-gov.go.jp	0	0
Philippines	http://www.gov.ph	2352	6
France	http://www.premier-ministre.gouv.fr/en	31	0
New Zealand	http://www.govt.nz	8962	3
Australia	http://www.australia.gov.au	1727	0
South Korea	http://www.korea.go.kr	1	0
India	http://india.gov.in	0	0
Argentina	http://www.argentina.gob.ar	217	0
Brazil	http://www.brasil.gov.br	511	0
Portugal	http://www.portugal.gov.pt/pt.aspx	260	0
Spain	http://www.lamoncloa.gob.es/PaginaNoEncontrada.html	1032	0
Italy	http://www.italia.gov.it/itagov2	0	0
Norway	http://www.norway.no	1683	0
Iceland	http://www.stjornarrad.is	4	0
Finland	http://government.fi/etusivu/en.jsp	0	0
Sweden	http://www.sweden.gov.se	3	0
Russia	http://www.gov.ru	3485	12

Table 2: Statistics of permanent membership of the United Nations Security Council.

United Nations Security Council	Number of e-mail addresses leaked	Number of SQL injection URLs obtained
US	1776	0
PRC	3287	121
Russia	3485	12
UK	1267	0
France	31	0

4. Experiment and Result Analysis

In our experiment, we were able to infer the degree of importance that every country places on website security by investigating the level of security implemented in the portal websites. The experimental hardware used were Intel Core I5-2500 CPU, with 8 GB of RAM, using the Microsoft Windows 7 operating system. The experiment time was 24 hours for each county. The figures given in the experimental results were obtained by gathering the follow-up statistics.

4.1. Assessment of Government Portal Websites of 24 Countries or Regions. We selected 24 countries or regions to conduct experiments on using the online crawler mining system, by filtering the e-mail addresses and URLs that were leaked by SQL injections into the countries' or regions' government portal websites, as shown in Table 1.

Among these 24 countries or regions, the highest levels of protection in the portal websites were found to be those of Germany, Singapore, Italy, Finland, Japan, and India, as no e-mail addresses or URLs were obtained from the injections. The lowest degree of security was found in the government portal websites of New Zealand and the Philippines, with the number of leaks in these being 8972 and 2352, respectively, as shown in Figures 15 and 16. The largest number of URLs obtained from injections was in the government portal of PRC. Although Canada's figure is the highest, these results were obtained from the same database, with only 5 to 10 websites being weak. Furthermore, the crawler mining system analyzed the URLs obtained from the injections, revealing the usernames and passwords stored inside the database, as displayed in Figure 17. When it is possible to dig or permeate the database, malicious users will drill deeply into the system in order to carry out serial malicious attacks.

4.2. Comparison of Permanent Memberships of the United Nations Security Council. We analyzed the permanent membership of the United Nations Security Council in 24 countries or regions, as shown in Table 2. The most severe case of leaked e-mail addresses and URLs obtained was in PRC, followed by Russia, while France was found to have the best

results. Since the United Nations Security Council is responsible for resolutions within the UN, any opinions raised among factions may affect the outcome of the resolution of an important issue. Therefore, without sufficient information security, hackers may obtain significant information through attacks, which could potentially lead to the manipulation of information, and therefore major problems, within the UN.

5. Conclusion

Our experimental results indicate that the most severe leakage problems in government portal websites occur in PRC, with 121 injections, which means that there is a weakness in the website programming, making it relatively easy for many malicious hackers to obtain information. The lowest number of leaks was found in the government portal websites of India, Italy, and Finland, meaning that, in general, these sites are less likely to be attacked. Furthermore, the government portal websites of Germany, Singapore, and Japan could not even be tested or analyzed, which means that they are completely protected against any attacks by the online crawler mining system. It is thus clear that the portal websites of these countries or regions are developed with a much higher level of information security.

On December 21, 2011, a series of leakages of user information occurred in various famous websites in Mainland China, affecting 50 million online users, and was allegedly the largest known online information leakage incident. It is a known fact that no online information can be completely secure; however, research in this area has established four major aspects that should be taken into account in order to improve the protection of online information [2, 52], as follows. (1) Website contents that prohibit the research of reptile programs should be regulated. (2) During the programming of websites, it is necessary to filter the input values and limit the variable modes. (3) Unnecessary extensive processes for database management should be removed, such as xp_cmdshell and xp_regaddmultistring. Passwords stored in the database must be encoded, but not using plain text. (4) A good writing style of robot.txt and network management mechanisms should be implemented. In dealing with attacks, forged files and categories may be added into the robot.txt, programming annotation or error websites. Such mechanism may reveal the intention from attackers or malicious software, which is considered as a Honeypot approach.

SQL injection is not a new technique, and various methods for preventing leakage in databases have been extensively studied [53, 54]. However, over the past decade, this issue has become a major concern in the field of information security. The habits of front users and the concepts of backward development are the main reasons for this concern [11, 55]. We will continue to improve the functions of our system, from an online reptile angle, by comparing our research with that of related fields [47, 48, 56–60], in order to develop a more powerful system.

With the increasing use of computers and the rapid development of the internet, society is experiencing fast-changing and different lifestyles. Information security has become a major concern as easily available online information continues to grow. Hence, measures must be taken for improving information security. Only under more secure conditions will we be able to fully enjoy the convenience of online information, which is necessary for living in the information age and being capable of facing greater challenges.

References

[1] B. Hoanga, "How good are our weapons in the spam wars?" *IEEE Technology and Society Magazine*, vol. 25, no. 1, pp. 22–30, 2006.

[2] T. Feng, "The security mechanism of network database," in *Proceedings of the 3rd International Conference on Measuring Technology and Mechatronics Automation (ICMTMA '11)*, pp. 939–940, January 2011.

[3] D. Geer, "Malicious bots threaten network security," *Computer*, vol. 38, no. 1, pp. 18–20, 2005.

[4] P. O. Boykin and V. P. Roychowdhury, "Leveraging social networks to fight spam," *Computer*, vol. 38, no. 4, pp. 61–68, 2005.

[5] C. Castillo, "Effective web crawling," in *Computer Science*, The University of Chile in fulfillment: ACM SIGIR Forum, 2004.

[6] W. Tong and X. Xie, "A research on a defending policy against the Webcrawler's attack," in *Proceedings of the 3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication (ASID '09)*, August 2009.

[7] Y. Sun, I. G. Councill, and C. L. Giles, "The ethicality of web crawlers," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI '10)*, pp. 668–675, September 2010.

[8] M. MeiJunjin, "An approach for SQL injection vulnerability detection," in *Proceedings of the 6th International Conference on Information Technology: New Generations (ITNG '09)*, pp. 1411–1414, April 2009.

[9] G. Jawaheer and P. Kostkova, "Web crawlers on a health related portal: detection, characterisation and implications," in *Proceedings of the 4th International Conference on Developments in eSystems Engineering (DeSE '11)*, pp. 24–29, December 2011.

[10] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Fighting spam on social web sites: a survey of approaches and future challenges," *IEEE Internet Computing*, vol. 11, no. 6, pp. 36–45, 2007.

[11] L. Shar and H. Tan, "Defeating SQL injection," *Computer*, no. 99, p. 1, 2012.

[12] J. Bau, E. Bursztein, D. Gupta, and J. Mitchell, "State of the art: automated black-box web application vulnerability testing," in *Proceedings of the 31st IEEE Symposium on Security and Privacy (SP '10)*, pp. 332–345, May 2010.

[13] N. Khoury, P. Zavarsky, D. Lindskog, and R. Ruhl, "An analysis of black-box web application security scanners against stored SQL injection," in *Proceedings of the IEEE International Conference on Privacy, Security, Risk and Trust, and IEEE International Conference on Social Computing (SocialCom '11)*, pp. 1095–1101, October 2011.

[14] J. Bau, E. Bursztein, D. Gupta, and J. Mitchell, "State of the art: automated black-box web application vulnerability testing," in *Proceedings of the 31st IEEE Symposium on Security and Privacy (SP '10)*, pp. 332–345, May 2010.

[15] J. Cho, H. Garcia-Molina, and L. Page, "Efficient crawling through URL ordering," in *Proceedings of the 7th International WWW Conference*, pp. 161–172, 1998.

[16] A. Z. Broder, M. Najork, and J. L. Wiener, "Efficient URL caching for world wide web crawling," in *Proceedings of the 12th International WWW Conference*, 2003.

[17] V. Shkapenyuk and T. Suel, "Design and implementation of a high-performance distributed web crawler," in *Proceedings of the 18th International Conference on Data Engineering*, pp. 357–368, March 2002.

[18] M. Najork, "Breadth-first search crawling yields high-quality pages," in *Proceedings of the 10th International World Wide Web Conference*, 2001.

[19] J. B. Anderson and S. Mohan, "Sequential coding algorithms: a survey and cost analysis," *IEEE Transactions on Communications*, vol. 32, no. 2, pp. 169–176, 1984.

[20] L. Zhao, H. Xu, J. Li, and Q. Cai, "A kind of map coloring algorithm considering of depth first search," in *International Conference on Industrial Control and Electronics Engineering (ICICEE '12)*, pp. 648–651, 2012.

[21] R. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez, "Crawling a country: better strategies than breadth-first for web page ordering," in *Proceedings of the 14th International World Wide Web Conference (WWW '05)*, pp. 864–872, May 2005.

[22] C. F. Diligentim and L. S. Focused, "Crawling using context graphs," in *Proceedings of the 26th International Conference on Very Large Data Bases*, 2000.

[23] J. Qin, Y. Zhou, and M. Chau, "Building domain-specific web collections for scientific digital libraries: a meta-search enhanced focused crawling method," in *Proceedings of the 4th ACM/IEEE Joint Conference on Digital Libraries; Global reach and Diverse Impact (JCDL '04)*, pp. 135–141, June 2004.

[24] J. Rennie and A. K. McCallum, "Using reinforcement learning to spider the web efficiently," in *Proceedings of the 16th International Conference on Machine Learning*, 1999.

[25] S. Chakrabarti, M. Berg, and B. Dom, "Focused crawling: a new approach to topic-specific web resource discovery," in *Proceedings of the 8th International WWW Conference*, 1999.

[26] S. Chakrabarti, D. A. Gibson, and K. S. McCurley, "Surfing the web backwards," *Computer Networks*, vol. 31, no. 11, pp. 1679–1693, 1999.

[27] G. J. Houben, Y. Kornatzky, and R. Post, "Information retrieval in distributed hypertexts," in *Proceedings of the 4th RIAO Conference*, 1994.

[28] M. Jacovi, "The shark-search algorithm. An application: tailored web site mapping," *Computer Networks*, vol. 30, no. 1–7, pp. 317–326, 1998.

[29] S. B. L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the 7th International WWW Conference*, 1998.

[30] P. Lawrence, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bringing order to the Web," in *Proceedings of the 7th International WWW Conference*, 1999.

[31] T. H. Haveliwala, "Topic-sensitive PageRank," in *Proceedings of the 11th International Conference on World Wide Web (WWW '02)*, pp. 517–526, May 2002.

[32] Y. Ding, F. Scharffe, A. Harth, and A. Hogan, "AuthorRank: ranking improvement for the web," in *Proceedings of the International Conference on Semantic Web and Web Services*, 2006.

[33] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[34] W. Xing and A. Ghorbani, "Weighted PageRank algorithm," in *Proceedings of the 2nd Annual Conference on Communication Networks and Services Research*, pp. 305–314, May 2004.

[35] Y. Zhang, C. Yin, and F. Yuan, "An application of improved PageRank in focused crawler," in *Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD '07)*, pp. 331–335, August 2007.

[36] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, "Automatic resource compilation by analyzing hyperlink structure and associated text," in *Proceedings of the 7th International WWW Conference*, pp. 65–74, 1998.

[37] P. Gawrysiak, "Using data mining methodology for text retrieval," in *Proceedings of the International Information Science and Education Conference*, 1999.

[38] IBM, "Clever System (HITS)—a page ranking algorithm developed," 2004, http://www.ecsl.cs.sunysb.edu/~chiueh/cse646/cn4/cn4.html.

[39] G. Pant and P. Srinivasan, "Link contexts in classifier-guided topical crawlers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 107–122, 2006.

[40] F. Menczer and R. K. Belew, "Adaptive retrieval agents: internalizing local context and scaling up to the Web," *Machine Learning*, vol. 39, no. 2, pp. 203–242, 2000.

[41] J. Johnson, K. Tsioutsiouliklis, and C. Lee Giles, "Evolving strategies for focused web crawling," in *Proceedings of the 20th International Conference on Machine Learning*, pp. 298–305, August 2003.

[42] Y. Chun and Z. Hang, "Research on tunnel phenomenon in focused crawler," in *Proceedings of the International Conference on Internet Technology and Applications (ITAP '10)*, August 2010.

[43] C. Anley, *More Advanced SQL Injection*, An NGSSoftware Insight Security Research (NISR) Publication, 2002.

[44] Y. J. Yang and Y. X. Yu, "Behavioral science-based information security research," in *Proceedings of the 1st International Workshop on Education Technology and Computer Science (ETCS '09)*, pp. 428–432, March 2009.

[45] C. Anley, *Advanced SQL injection in SQL Server Applications*, An NGSSoftware Insight Security Research (NISR) Publication, 2002.

[46] V. Chapela, *Advanced SQL Injection*, 2005.

[47] D. A. Kindy and A.-S. K. Pathan, "A survey on SQL injection: vulnerabilities, attacks, and prevention techniques," in *Proceedings of the 15th IEEE International Symposium on Consumer Electronics (ISCE '11)*, pp. 468–471, June 2011.

[48] R. Johari and P. Sharma, "A survey on web application vulnerabilities (SQLIA, XSS) exploitation and security engine for SQL injection," in *Proceedings of the International Conference on Communication Systems and Network Technologies (CSNT '12)*, 2012.

[49] S. Bondar, "Search engine indexing limits: where do the bots stop," 2006, http://www.sitepoint.com/indexing-limits-where-bots-stop/.

[50] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang, "Accessing the deep Web," *Communications of the ACM*, vol. 50, no. 5, pp. 94–101, 2007.

[51] O. Security, "The Exploit Database," 2012, http://www.exploit-db.com/.

[52] S. Imran and I. Hyder, "Security issues in databases," in *Proceedings of the 2nd International Conference on Future Information Technology and Management Engineering (FITME '09)*, pp. 541–545, December 2009.

[53] S. W. Boyd and A. D. Keromytis, "SQLrand: preventing SQL injection attacks," in *Proceedings of the 2nd Conference on Applied Cryptography and Network Security*, pp. 292–302, 2004.

[54] Y.-W. Huang, F. Yu, C. Hang, C.-H. Tsai, D. T. Lee, and S.-Y. Kuo, "Securing web application code by static analysis and runtime protection," in *Proceedings of the 13th International World Wide Web Conference Proceedings (WWW '04)*, pp. 40–52, May 2004.

[55] R. Overstreet, *Protecting Yourself from SQL Injection Attacks*, 2006.

[56] M. Ficco, L. Coppolino, and L. Romano, "A weight-based symptom correlation approach to SQL injection attacks," in *Proceedings of the 4th Latin-American Symposium on Dependable Computing (LADC '09)*, pp. 9–16, September 2009.

[57] A. Mohosina and M. Zulkernine, "DESERVE: a framework for detecting program security vulnerability exploitations," in *Proceedings of the 6th IEEE International Conference on Software Security and Reliability (SERE '12)*, 2012.

[58] C. Mainka, J. Somorovsky, and J. Schwenk, "Penetration testing tool for web services security," in *Proceedings of the 8th IEEE World Congress on Services (SERVICES '12)*, 2012.

[59] R. Xu, J. Guo, and L. Deng, "A database security gateway to the detection of SQL attacks," in *Proceedings of the 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE '10)*, pp. 3537–3540, August 2010.

[60] G. Agosta, A. Barenghi, A. Parata, and G. Pelosi, "Automated security analysis of dynamic web applications through symbolic code execution," in *Proceedings of the 9th International Conference on Information Technology: New Generations (ITNG '12)*, pp. 189–194, 2012.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 403978, 11 pages
http://dx.doi.org/10.1155/2013/403978

Research Article

An MBS-Assisted Femtocell Transmit Power Control Scheme with Mobile User QoS Guarantee in 2-Tier Heterogeneous Femtocell Networks

Jenhui Chen,[1] Chih-Cheng Yang,[2] and Shiann-Tsong Sheu[2]

[1] *Department of Computer Science and Information Engineering, School of Electrical and Computer Engineering, College of Engineering, Chang Gung University, Kweishan, Taoyuan 33302, Taiwan*
[2] *Department of Communication Engineering, National Central University, Chung-Li 32001, Taiwan*

Correspondence should be addressed to Jenhui Chen; jhchen@mail.cgu.edu.tw

Received 10 October 2013; Accepted 4 November 2013

Academic Editors: W. Sun, G. Zhang, and J. Zhou

This study investigates how to adjust the transmit power of femto base station (FBS) to mitigate interference problems between the FBSs and mobile users (MUs) in the 2-tier heterogeneous femtocell networks. A common baseline of deploying the FBS to increase the indoor access bandwidth requires that the FBS operation will not affect outdoor MUs operation with their quality-of-service (QoS) requirements. To tackle this technical problem, an *FBS transmit power adjustment* (FTPA) algorithm is proposed to adjust the FBS transmit power (FTP) to avoid unwanted cochannel interference (CCI) with the neighboring MUs in downlink transmission. FTPA reduces the FTP to serve its femto users (FUs) according to the QoS requirements of the nearest neighboring MUs to the FBS so that the MU QoS requirement is guaranteed. Simulation results demonstrate that FTPA can achieve a low MU outage probability as well as serve FUs without violating the MU QoS requirements. Simulation results also reveal that FTPA has better performance on voice and video services which are the major trend of future multimedia communication in the NGN.

1. Introduction

Next generation communication networks are expected to provide a function of pervasive network access as well as quality-of-service (QoS) guarantee. The fourth generation mobile communication standard, for example, 3GPP long term evolution-advanced (LTE-A) [1], proposed a h-tier heterogeneous network architecture which integrates h types of networks for coexisting to achieve this goal. The 2-tier heterogeneous femtocell network is one type of the h-tier heterogeneous network architecture, which is composed of one macrocell and several femtocells within the radio coverage of macrocell. The femtocells formed by femto base stations (FBSs) are set by subscribers for the purpose of increasing data access rate in indoor environment [2, 3]. If FBSs are set privately, they are the type of closed subscriber group (CSG) FBSs [4] with which only allow authorized femto users (FUs) to connect. Otherwise, the FBSs belong to the type of full or partial open subscriber group (OSG) FBSs.

Because the macrocell and femtocell networks coexist in the same frequency band, the interference avoidance and mitigation problem dominate the key point of the performance of network coexistence. A femtocell network survey [4] and a femto forum report [5] showed that most used services of mobile users (MUs) are voice calls; more than 50% of voice calls and more than 70% of data traffic occur in indoor environment. Articles [6–8] indicated that if conventional power-control schemes are applied for multimedia traffic without any modification, the system capacity is limited by the traffic with the lowest bit error rate (BER) requirement. Important research studies [6, 7] indicated that voice packets can typically tolerate a BER up to 10^{-3} but are delay sensitive while data packets require a BER below 10^{-9} but are delay insensitive. Wang [8] showed that the medium access control (MAC) protocol design for wireless multimedia network is challenging because multimedia services have heterogeneous BERs and quality-of-service (QoS) requirements. He established the relation between target

Table 1: Parameters of different traffic types [8].

Index	Service type	BER	SINR	Timeout	Characteristic
1	Voice	10^{-3}	5.31	2 frames	Delay sensitive Talkspurt and silence
2	Audio	10^{-4}	7.31	6 frames	Delay sensitive Stream
3	CBR video	10^{-5}	9.32	5 frames	Delay insensitive Constant bit-rate
4	VBR video	10^{-6}	11.34	4 frames	Delay insensitive Variable bit-rate
5	Data	0	2.94	∞	Delay insensitive Variable size

signal-to-interference-plus-noise ratio (SINR) values and BER requirements by considering the error control schemes and channel fading. Table 1 shows the BER requirements of different traffic types and their corresponding SINR values.

In femtocell networks, most FBSs are used in indoor environment. In this scenario, several communication properties are specified as follows. First, because most FBSs are usually set indoors and surrounded by concrete walls, the leakage of electromagnetic energy to the outside is degraded significantly. Second, the probability of MU interfered by FBS transmit power (FTP) can be reduced if the FTP is controlled by taking the neighboring MUs (if any) into account. Third, the BER can be improved in indoor environment by adopting an adaptive modulation and coding (AMC) scheme according to radio channel quality, which can be used for heterogeneous femtocell networks deployments [9]. These factors motivate us to mitigate the interference between the macrocell and femtocells by adjusting the FTP and using the AMC scheme with the prerequisite of guaranteeing QoS services [10] of MUs and FUs, which are around FBSs.

The interference problems of the 2-tier heterogeneous femtocell networks are classified into two types: the downlink and uplink interference problems. Many interference management mechanisms were broadly studied such as cochannel femtocell deployment for interference-limited coverage area (ILCA) [11], power control and beamforming disjoint mechanism [12], interference reducing by minimizing the transmit power [13], and self-optimized coverage coordination mechanism [14]. Several spectrum allocation mechanisms were used for spectrum splitting such as preplanned frequency assignment approaches [15], gaming approaches [16, 17], cognitive radio approaches [18, 19], dynamic spectrum allocation and cell association mechanisms [20, 21], and cooperative spectrum allocation mechanism for intercell fairness [22]. In [23, 24], they gave comprehensive studies of analysis and simulation on downlink interference based on a SINR viewpoint because the impact of downlink interference on communications is more serious as compared with the uplink interference. However, none of them considered the QoS properties of service types when they dealt with the problem of mitigating interference.

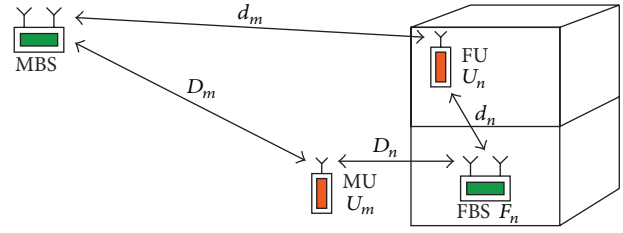


Figure 1: The system model of 2-tier heterogenous femtocell networks.

There are two technical challenges on the interference mitigation. The first challenge is how to adjust the FTP to avoid interfering with the ongoing transmission of neighboring MUs with corresponding QoS parameters. The second challenge is how to choose an appropriate modulation and coding scheme based on the upper bound of adjusted FTP to meet QoS parameters expected by FUs for the services requested.

To conquer these challenges, an FBS transmit power adjustment (FTPA) algorithm which considers MU's QoS requirements is proposed to dynamically adjust FTP for avoiding cochannel interference (CCI) with MUs. FTPA chooses the received SINR of users as the QoS index. The macro-BS (MBS) uses the location information (LI) of neighboring MUs around FBSs to notify corresponding FBSs through the S1 interface (backhaul networks) for CCI mitigation. The overheads of LI forwarding are light because the MBS only notify the nearest MU's LI to the corresponding FBS, and the bandwidth of backhaul is large enough to afford the needed traffic load.

The rest of this paper is organized as follows. The system model of macro- and femto-coexisting networks is introduced in Section 2. A detailed FTPA algorithm is given in Section 3. A simulation scenario and results for evaluation of FTPA algorithm are given in Section 4. Finally, some remarks and future research topics are in Section 5.

2. System Model

Figure 1 illustrates a scenario of 2-tier heterogeneous femtocell networks where one MBS and MUs are located in outdoor environment, and the FBSs and FUs are located in indoor environment. Suppose that N_m MUs and N_f FBSs are uniformly distributed in a 3D urban environment [25]. Assume that the probability density function (pdf) of the x-nearest FBSs to an MU denoted as U_m (assume that U_m is located at the center of a sphere) follows a homogeneous Poisson point process (HPPP) [26], and is given by

$$P_X(x, \lambda, V) = \frac{(\lambda V)^x}{x!} e^{-\lambda V}, \quad V = \frac{4\pi r^3}{3}, \tag{1}$$

where x is the number of FBSs in the sphere, λ is the density of FBSs, and r and V are the radius and volume of the sphere. As illustrated in Figure 1, the distance (in meters) between the MBS and the MU is denoted by D_m, the distance between the nearest FBS and the the MU is denoted by D_n, the distance

Table 2: Path loss parameters [28].

Path loss	δ_m or δ_f	η_m or η_f
$L_o(D_m)$	$30 \log f_c - 71$	4
$L_i(d_n)$	$37 + 18.3h^{((h+2)/(h+1)-0.46)}$	3
$L_x(D_n)$	$37 + 18.3h^{((h+2)/(h+1)-0.46)} + \delta_p$	3

between the MBS and the FU is denoted by d_m, and the distance between the FU and the serving FBS is denoted by d_n.

Haenggi [27] showed that the pdf of an Euclidean distance D_n (in meters) between an MU and its nth nearest neighboring FBS F_n, denoted by $P_{D_n}(r)$, is distributed according to the generalized gamma distribution as follows:

$$P_{D_n}(r) = \frac{3\left(4\pi\lambda r^3/3\right)^n}{r\Gamma(n)} e^{-4\pi\lambda r^3/3}, \tag{2}$$

where $\Gamma(n)$ is the gamma function.

2.1. Path Loss. The path loss between a transmitter and receiver in indoor environment is quite different from that in outdoor environment. Based on ITU-R M.1225 slow fading path-loss model [28], the outdoor and pedestrian path-loss, denoted by $L_o(D_m)$ in dB, from the MBS to an MU U_m is expressed as

$$\begin{aligned} L_o(D_m) &= \delta_m + 10\eta_m \log_{10}(D_m) \\ &= 30 \log f_c - 71 + 40 \log_{10}(D_m), \end{aligned} \tag{3}$$

where δ_m and η_m are the outdoor path-loss constant and exponent of macrocell as shown in Table 2, respectively, and f_c is the central frequency of operating frequency in MHz.

The indoor path loss, denoted by $L_i(d_n)$, between an FU and its serving FBS F_n with a distance d_n is

$$\begin{aligned} L_i(d_n) &= \delta_f + 10\eta_f \log_{10}(d_n) \\ &= 37 + 18.3h^{((h+2)/(h+1)-0.46)} + 30 \log_{10}(d_n), \end{aligned} \tag{4}$$

where δ_f and η_f are the indoor path-loss constant and exponent of femtocell and h is the number of floors between F_n and FU in a building. If the FBS and FUs are in the same floor, the value of h is equal to zero.

Assume that the FBS is placed inside the house and the radio wave to the MU crosses an external wall of the house. Thus, the indoor to outdoor path loss, denoted by $L_x(D_n)$, between F_n and U_m is given as

$$\begin{aligned} L_x(D_n) &= \delta_f + 10\eta_f \log_{10}(D_n) \\ &= 37 + 18.3h^{((h+2)/(h+1)-0.46)} + \delta_p + 30 \log_{10}(D_n), \end{aligned} \tag{5}$$

where δ_p is the penetration loss when the radio wave crosses the wall of house. δ_p varies depending on different materials of the wall and we assume that $\delta_p = 10$, 15, and 20 dB in this paper. The interference among femtocells is not considered here because they are separated by concrete walls or obstacles and are set by subscribers in distance [3].

2.2. Power Adjustment. Let ψ_m^q be the minimum required SINR threshold (the target SINR) of an MU for achieving one of service types with a QoS index q, $q = 1, 2, \ldots, k$, where k is the total number of service types that the MBS provides. Taking Table 1 for example, ($k = 5$), $\psi_m^1 = 5.31$ stands for the minimal required SINR to achieve voice service and $\psi_m^5 = 2.94$ stands for the minimal required SINR to achieve data service, and so forth. Because the received SINR of U_m, denoted by $\psi_m(D_m)$, can be simplified by the ratio of the received signal strength from the MBS to its first nearest neighboring FBS F_1 plus noise power [29], we have

$$\begin{aligned} \psi_m(D_m) &\approx K_m - L_o(D_m) - I_{F_1} \\ &= K_m - \delta_m - 10\eta_m \log_{10}(D_m) - I_{F_1}, \end{aligned} \tag{6}$$

where K_m is the transmit power of the MBS in dB and I_{F_1} is the interference power from F_1 and is calculated by

$$\begin{aligned} I_{F_1} &= K_{F_1} - L_x(D_1) \\ &= K_{F_1} - \delta_f - 10\eta_f \log_{10}(D_1), \end{aligned} \tag{7}$$

where K_{F_1} is the FTP of F_1 in dB. In this paper, only the first nearest FBS is considered as the dominating interference source because the difference of signal strength between the first nearest FBS and the tenth nearest FBS is about 15 dB [25]. Finally, considering the background noise to U_m, (6) is finalized as

$$\begin{aligned} \psi_m(D_m) \approx\ & K_m - \delta_m - 10\eta_m \log_{10}(D_m) - 10 \log_{10} \\ & \cdot \left(10^{(K_{F_1} - \delta_{f1} - 10\eta_f \log_{10}(D_1))/10} + N_m\right), \end{aligned} \tag{8}$$

where N_m is the noise floor to U_m.

Theorem 1. *To satisfy the QoS requirement of U_m, $\psi_m(D_m) \geq \psi_m^q$, the maximum allowable FTP of FBS F_1 (i.e., K_{F_1}) follows the inequality*

$$\begin{aligned} K_{F_1} \leq\ & K_m - \delta_m + \delta_f - 10\eta_m \log_{10}(D_m) \\ & + 10\eta_f \log_{10}(D_n) - \psi_m + 10 \log_{10} \\ & \cdot \left(1 - \frac{N_m}{10^{(K_m - \delta_m - 10\eta_m \log_{10}(D_m) - \psi_m^q)/10}}\right). \end{aligned} \tag{9}$$

Proof. To satisfy the QoS requirement of the connection which belongs to U_m, the MBS selects a modulation and coding rate M_m to transmit packets to U_m. Based on the assumption and (8), the required SINR of U_m for M_m must satisfy the condition $\psi_m(D_m) \geq \psi_m^q$; then

$$\begin{aligned} \psi_m^q \leq\ & K_m - \delta_m - 10\eta_m \log_{10}(D_m) - 10 \log_{10} \\ & \cdot \left(10^{(K_{F_1} - \delta_f - 10\eta_f \log_{10}(D_1))/10} + N_m\right). \end{aligned} \tag{10}$$

Rearranging (10) for K_{F_1}, we get

$$\begin{aligned} K_{F_1} \leq\ & K_m - \delta_m + \delta_f - 10\eta_m \log_{10}(D_m) \\ & + 10\eta_f \log_{10}(D_1) - \psi_m^q + 10 \log_{10} \\ & \cdot \left(1 - \frac{N_m}{10^{(K_m - \delta_m - 10 \log_{10}(D_m) - \psi_m^q)/10}}\right). \end{aligned} \tag{11}$$

Table 3: Modulation and coding parameters [30].

Level (M)	Modulation	Required SINR (ψ_M)
1	QPSK (1/2)	5 dB
2	QPSK (3/4)	8 dB
3	16-QAM (1/2)	10.5 dB
4	16-QAM (3/4)	14 dB
5	64-QAM (1/2)	16 dB
6	64-QAM (2/3)	18 dB
7	64-QAM (3/4)	20 dB

Theorem 1 gives the upper bound of the FTP if an MU neighbors the FBS. Taking K_{F_1} obtained from (9), the received SINR of U_n, denoted by $\psi_f(d_n)$, can be calculated by

$$\psi_f(d_n) \approx K_{F_1} - \delta_f - 10\eta_f \log_{10}(d_n) - 10\log_{10} \cdot \left(10^{(K_m-\delta_m-10\eta_m \log_{10}(d_m))/10} + N_m\right). \quad (12)$$

Suppose that there are l different AMC levels supported in the PHY layer. The achievable AMC level depends on the received SINR value $\psi_f(d_n)$ of user U_n. Let $M = 1, 2, \ldots, l$ be the index of AMC levels as shown in Table 3, and ψ_M be the minimal required SINR to achieve the modulation level M. Let M_n denote the maximal achievable AMC level for FU U_n, and is given by

$$M_n \triangleq \left\{M \mid \psi_M \le \psi_f(d_n) < \psi_{M+1}\right\}, \quad (13)$$

where $\psi_{M+1} = \infty$ when $M = l$. Let ψ_n^q denote the target SINR of an FU U_n for achieving one of service types with a QoS index q, $q = 1, 2, \ldots, k$. According to (12), the FBS can determine an M_n with K_{F_1} power for QoS level q if it satisfies the condition $\psi_f(d_n) \ge \psi_n^q$. □

3. FTP Adjustment Algorithm

Because FTPA is mainly applied in the FBS, some parameters have to be inputted into FTPA prior for calculating the upper bound of FTP.

3.1. Outage Probability of MU. Based on (1) and (2), Tseng and Huang [25] showed that the MU outage probability in the cellular networks (i.e., $\psi(D_m) < \psi_m^q$) is based on a given distance D_m, a given FBS density λ, and a given target signal-to-interference ratio (SIR). However, in this paper, we consider the SINR as the parameter to obtain the MU outage probability. Hence, the occurring probability of outage events by a given target outage probability $\Pr_O$ (which is the maximal tolerable outage probability of the cellular system) follows the condition

$$\Pr\left[\psi_m(D_m) < \psi_m^q\right] \approx 1 - \exp\left(\frac{-4\pi\lambda \exp\left(3\zeta_u(D_m)\right)}{3}\right) < \Pr_O, \quad (14)$$

where $\zeta_u(D_m)$ is an upper bound value related to ψ_m^q and is given by

$$\zeta_u(D_m) = \frac{\ln(10)}{10\eta_f}\left(\psi_m^q - \left(K_m - K_{F_1}\right) + \left(\delta_m - \delta_f\right) + 10\eta_m \log_{10}(D_m) - 10\log_{10} \cdot \left(1 - \frac{N_m}{10^{(K_m-\delta_m-10\eta_m \log_{10}(D_m)-\psi_m)/10}}\right)\right), \quad (15)$$

where $N_m < 10^{(K_m-\delta_m-10\eta_m \log_{10}(D_m)-\psi_m)/10}$. Because the MU outage probability varies with D_m, the expected MU outage probability in a macrocell with a radius R under a given FBS density λ, MBS transmit power K_m, and the nearest FTP K_{F_1}, can be obtained by

$$P_R = \int_{x=0}^{R} \frac{3x^2}{R^3}\left[1 - \exp\left(\frac{-4\pi\lambda \exp\left(3\zeta_u(x)\right)}{3}\right)\right] dx. \quad (16)$$

3.2. FTPA with Location Information. Assume that all MUs and FBSs equip with the global positioning system (GPS). Each FBS reports the current location via the backhaul connection (i.e., the S1 interface) to the overlaid MBS when each FBS is set by subscribers. Let K_O denote the operating FTP and let the maximal FTP be the initial value of K_O (e.g., 20 dBm). Each MU reports its current location to the MBS by periodic ranging procedures. The MBS can obtain D_m and D_n derived from the location reported from each U_m. The MBS can calculate the received SINR $\psi_m(D_m)$ of each U_m by applying D_m, D_n, and K_F into (8). To ensure that QoS services of MUs can be guaranteed, system operators may set a safe SINR difference value τ (i.e., $\psi_m(D_m) - \psi_m^q \ge \tau$) to increase the reliability of these QoS services. When $\psi_m(D_m) - \psi_m^q < \tau$, the MBS notifies the nearest FBS (i.e., F_1) of the MU with the MU's LI after an observation window T_w (a period of time). After receiving the LI, the interfering FBS reduces its FTP to avoid the interference with the MU according to (9). The value of τ can be set according to the waiting time period or other effects to guarantee the service continuity of nearby MUs.

Let I_k be a set of MUs which are interfered by FBS k, $k = 1, 2, \ldots, N_f$. If an MU U_m is interfered by FBS k (i.e., $\psi_m(D_m) - \psi_m^q < \tau$), the MBS adds the U_m to I_k. Notice that an I_k may contains more than one element (i.e., more than one MU is interfered by FBS k). To reduce LI notification overheads, the MBS waits a period of T_w to observe the situation of MU interference. If the observed MU is still interfered by the FBS k after T_w, the MBS notifies the FBS k of the current LI of MU. Otherwise, the MU is deleted from I_k and the LI message will not be sent to the FBS k.

After T_w elapsing, if $I_k \neq \emptyset$, the MBS sends LI (the location (x, y) of each MU and its corresponding QoS level q) of all neighboring MUs to the FBS via the backhaul networks. Upon receiving the LI, the FBS k executes FTPA to determine the maximal allowable FTP K_F and check whether the use of K_F can satisfy the QoS of its FUs. If no FUs are being served,

the FBS k discards the LI immediately. The details of FTPA are described as follows.

Step 1. Upon receiving the LI, the FBS k selects a tuple $((x, y), q)$ from the LI and determines the maximum allowable FTP K_F according to (9).

Step 2. The FBS k sets $K_O = \min(K_O, K_F)$. Repeat Step 1 until all the tuples of the received LI are treated.

Step 3. The FBS k selects one FU from served FUs to obtain d_n and d_m and applies the obtained K_O (treated as K_{F_1}), d_n, and d_m into (12) to obtain $\psi_f(d_n)$.

Step 4. If $\psi_f(d_n) \geq \psi_n^q$, the FBS k selects the highest achievable AMC level M_n according to (13). Otherwise, the FBS k uses K_O with the lowest AMC level to serve U_n.

Step 5. Repeat Steps 3 and 4 until all served FUs are visited.

When an MU moves close to an FBS (the MBS sends the LI message to the FBS), the FBS executes FTPA to reduce its FTP to avoid interfere with the MU. The FTP reduction of the FBS is temporal. The FBS will return its K_O to the maximal power when the neighboring MUs move away from the FBS's radio coverage. To ensure the FBS can return its maximal FTP, the MBS notifies the FBS of the MU leaving a message once the neighboring MU of the FBS leaves the radio coverage of the FBS (i.e., $\psi_m(D_m) - \psi_m^q \geq \tau$).

4. Simulation Results

This section presents a series of simulations for performance evaluation of FTPA. The scenario of simulation consists of one MBS and multiple FBSs around the MBS located in a sphere area with a radius of 424 meters. Three types of QoS are adopted in the simulation. They are the voice type ($\psi^q = 5.31$ dB), the CBR video type ($\psi^q = 9.32$ dB), and the data type ($\psi^q = 2.94$ dB), for example, email or control messages. To focus on the effect of adjusting FTP on MU's QoS achievement, the antenna gain is not considered in this simulation. The coverage radius varies by different QoS types because different QoS types require different SINR values. Thus, in the simulation, the order of coverage radius of voice, CBR video, and data is data > voice > CBR video.

The number of FBSs is adjusted and controlled by varying the FBS density (λ) which ranges from 10^{-6} to 10^{-8} m^{-3} to observe the impact of femtocell networks on the macrocell network. The communication radius of the femtocell is 20 meters. All the FBSs are OSG FBSs, and the location of FBSs is randomly deployed in the macrocell. The path loss models follow the definitions and descriptions in Section 2.1. All FBSs are in the indoor environment. To evaluate the effect of FTP adjustment on MU interference, three types of scenarios (i.e., different thicknesses of walls are considered $\delta_p = 20, 15, 10$ dB) are adopted in this simulation. The parameter $\delta_p = 20$ dB indicates that FBSs are deployed in the concrete buildings which degrade the signal strength significantly. Contrarily, $\delta_p = 10$ dB represents the wooden houses in which signals can easily pass through the wall. The SINR safe difference value τ is set as 0. The height of each MU is 1.5 m. Other simulation parameters are shown in Table 4.

Table 4: System parameter for simulation.

Parameter	Value
Center frequency	2.5 GHz
Bandwidth	10 MHz
FFT size	1024
Macrocell radius	400 m
FBS radius	20 m
Total BS TX power	46 dBm
Total FBS TX power	20 dBm
Antenna gain	0 dBi
Penetration loss	10, 15, 20 dB
MU distribution	Uniform
Number of MUs	100
Thermal noise	−92.974 dBm

The numerical result (Num) of the MU outage probability obtained from (16), the approximation (Approx) approach based on the dominating interference source only [25], and FTPA are simulated for performance comparison. This simulation does not consider the interfemtocell interference. Each simulation result is obtained by calculating the average for 10,000 random scenarios for all experiments. Four performance criteria are measured to evaluate the performance.

Consider the following.

(i) MU outage probability (P_R): the probability that the received SINR of U_m cannot satisfy the QoS requirement ψ_m^q of U_m.

(ii) Mean MU SINR ($\psi_m(D_m)$): the mean received SINR of MUs.

(iii) System capacity: the capacity achieved by the macrocell and femtocells in the overlaid area.

(iv) MU data rate (Mbps): the mean data rate (mega bits per second) per each MU. This value is obtained by Shannon's formula [31].

Figures 2, 3, and 4 demonstrate the MU outage probability caused by using Approx and FTPA versus the density of FBSs development under three different wall material scenarios. The Num is the upper bound of MU outage probability caused by coexisting FBSs where FBSs do not apply the FTP adjustment scheme. Figure 2 shows the MU outage probability of Num, Approx, and FTPA under three types of wall penetration loss ($\delta_p = 20, 15, 10$ dB) when the QoS type of all FBSs is voice ($\psi_m^q = 5.31$ dB). The MU outage probability of Num and Approx raises quickly as the FBS density increases under different types of penetration loss. The phenomenon becomes more obvious when δ_p is small. Taking $\delta_p = 20$ dB in Figure 2(a); for example, the MU outage probability of Num and Approx is about 0.03 when $\lambda = 10^{-8}$, and raises up to 0.6 when $\lambda = 10^{-6}$. The MU outage probability of Num and Approx raises more significantly when δ_p is smaller as shown in Figures 2(b) and 2(c). Figure 2(b) demonstrates that the MU outage probability

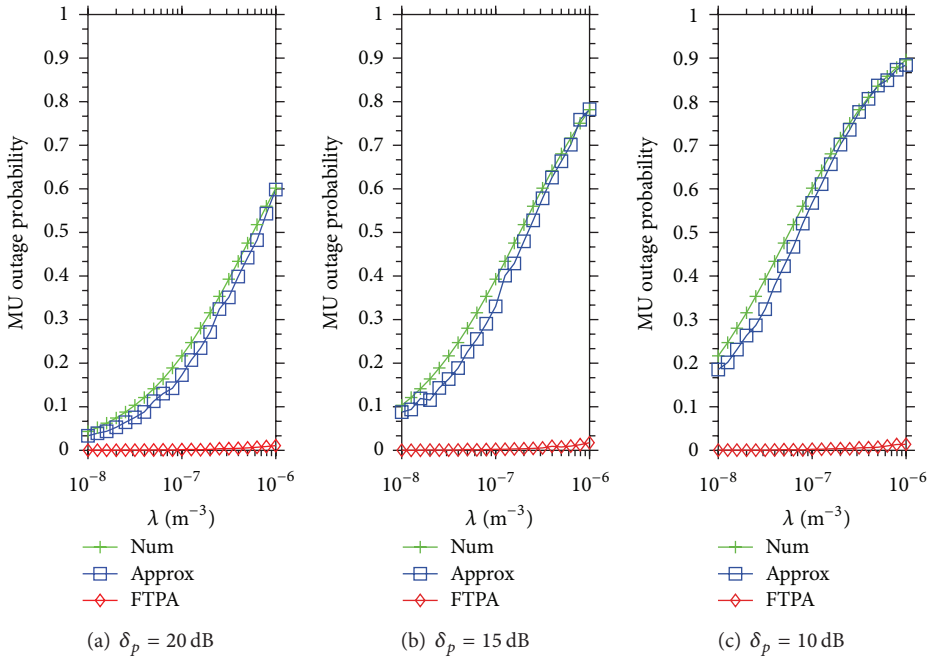


Figure 2: MU outage probability caused by Approx and FTPA versus the density of FBSs development when the QoS type of each MU is voice ($\psi_m^1 = 5.31$ dB).

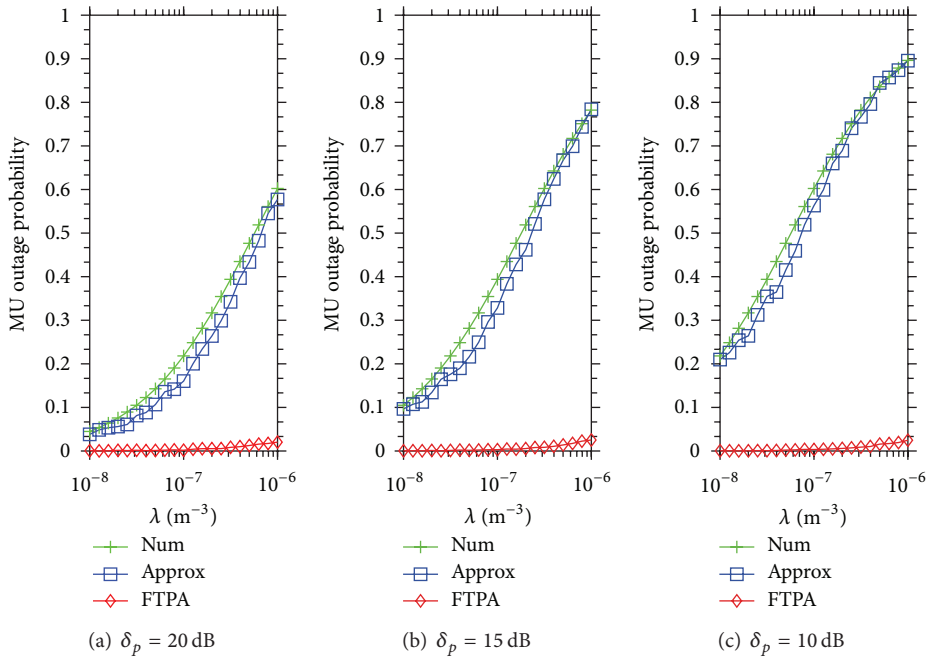


Figure 3: MU outage probability caused by Approx and FTPA versus the density of FBSs development when the QoS type of each MU is CBR video ($\psi_m^3 = 9.32$ dB).

raises from 0.11 to 0.8 as the FBS density raises from 10^{-8} to 10^{-6}. The MU outage probability is up to 0.9 when $\delta_p = 10$ dB and $\lambda = 10^{-6}$, as shown in Figure 2(c). This is because the leaked FTP is higher when δ_p is lower, and thus leads to MU suffering higher unwanted interference power.

Contrarily, the MU outage probability achieved by FTPA is very low ($P_R < 0.02$) even when the FBS density is high ($\lambda = 10^{-6}$) and the penetration loss is low ($\delta_p = 10$ dB). This result shows that FTPA decreases the interfering FTP to satisfy the neighboring MU QoS requirements and then

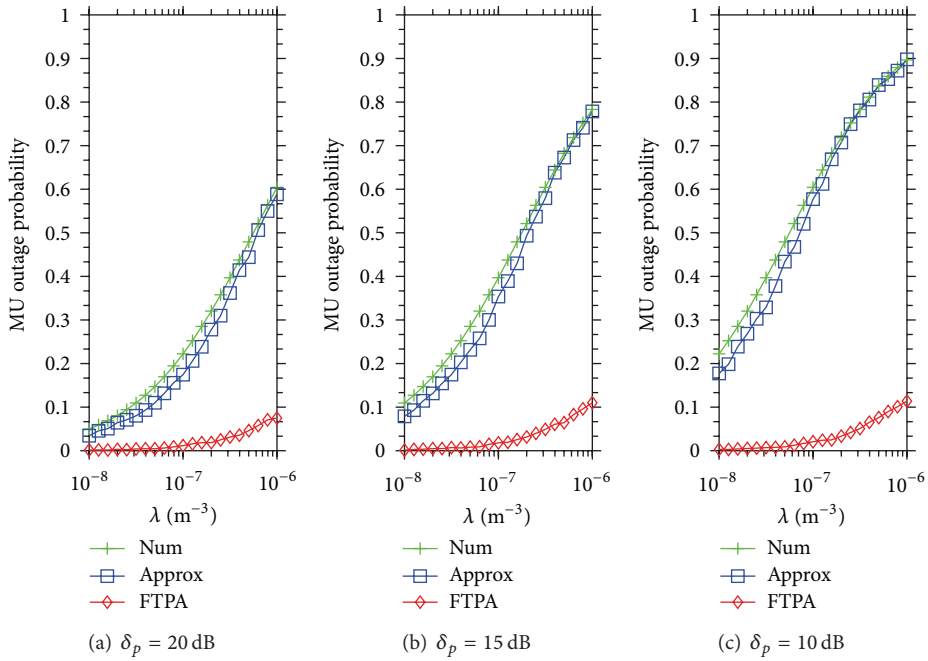


Figure 4: MU outage probability caused by Approx and FTPA versus the density of FBSs development when the QoS type of each MU is data ($\psi_m^5 = 2.94$ dB).

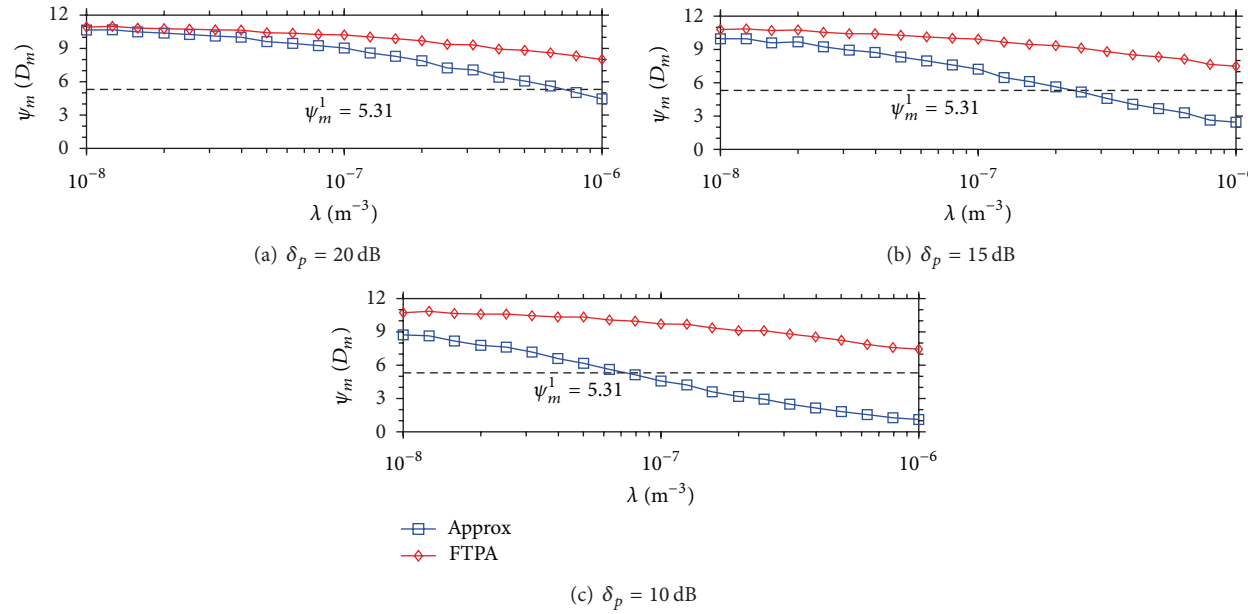
reduces the MU outage probability significantly. For instance, $P_R = 0.0138$ when $\lambda = 10^{-6}$ in Figure 2(c). From the observation of these results, it shows that the outage probability can be controlled easily by only considering the dominating interference because the second interference is not strong enough to interfere with the nearby MUs.

Figure 3 demonstrates the MU outage probability when the QoS service type is CBR video. The simulation result is similar to that shown in Figure 2. The MU outage probability is up to $P_R = 0.6$ when $\delta_p = 20$ dB, $P_R = 0.8$ when $\delta_p = 15$ dB, and $P_R = 0.9$ when $\delta_p = 10$ dB. Meanwhile, FTPA remains in low P_R even in high FBS density $\lambda = 10^{-6}$ ($P_R = 0.025$). Similarly, Figure 4 demonstrates the MU outage probability when the QoS type is data. It shows that the MU outage probability achieved by FTPA reaches 0.1 when $\lambda = 10^{-6}$ and $\delta_p = 10$ dB. The reason that the MU outage probability increases more obviously than the case of ψ_m^1 and ψ_m^3 is as follows. First, because the required SINR of data service $\psi_m^5 = 2.94$ is lower than $\psi_m^1 = 5.31$ and $\psi_m^3 = 9.32$, the FBS uses higher FTP to serve its FUs (according to (9)). As a result, MUs are interfered by FBSs highly. Second, the coverage radius of data service is longer due to the lower ψ_m requirement. Thus, MUs located in the boundary of the transmission range get lower SINR (due to path loss) and then easily interfered by neighboring FBSs. These results show that FTPA can satisfy all types of QoS of MU in all cases because the maximal MU outage probability is lower than 0.1 which is a P_R theorem bound for system operation.

Figure 5 shows the mean MU SINR when the QoS type of all MUs is voice under different types of penetration loss. It shows that the density of FBS deployment has to decrease when δ_p decreases. When $\delta_p = 20$ dB, as shown in Figure 5(a), the feasible FBS density with the QoS requirement of voice ($\psi_m^q = 5.31$ dB) must be lower than $10^{-6.2}$. This situation is much more obvious when δ_p is lower; that is, $\delta_p = 15$ or 10 dB. The feasible FBS density is lower than $10^{-6.7}$ when $\delta_p = 15$ dB (see Figure 5(b)) and is lower than $10^{-7.2}$ when $\delta_p = 10$ dB (see Figure 5(c)). This is because lower δ_p leads to much leaked FTP from nearby FBSs and causes higher interference with MUs. Consequently, the density of FBS development must decrease.

However, as shown in Figure 5, FTPA takes advantage of FTP adjustment to guarantee the QoS requirement of nearby MUs from being interfered by FBS. As we can see from the results, FTPA maintains higher $\psi_m(D_m)$ than the threshold ψ_m^1 in all FBS densities. The mean SINR of FTPA decreases slightly as the FBS density increases, and it is always higher than the threshold ψ_m^1 even when the FBS density is in $\lambda = 10^{-6}$. No matter what the value of δ_p is, the FTPA can still guarantee the mean SINR for MUs about 8 dB when $\lambda = 10^{-6}$.

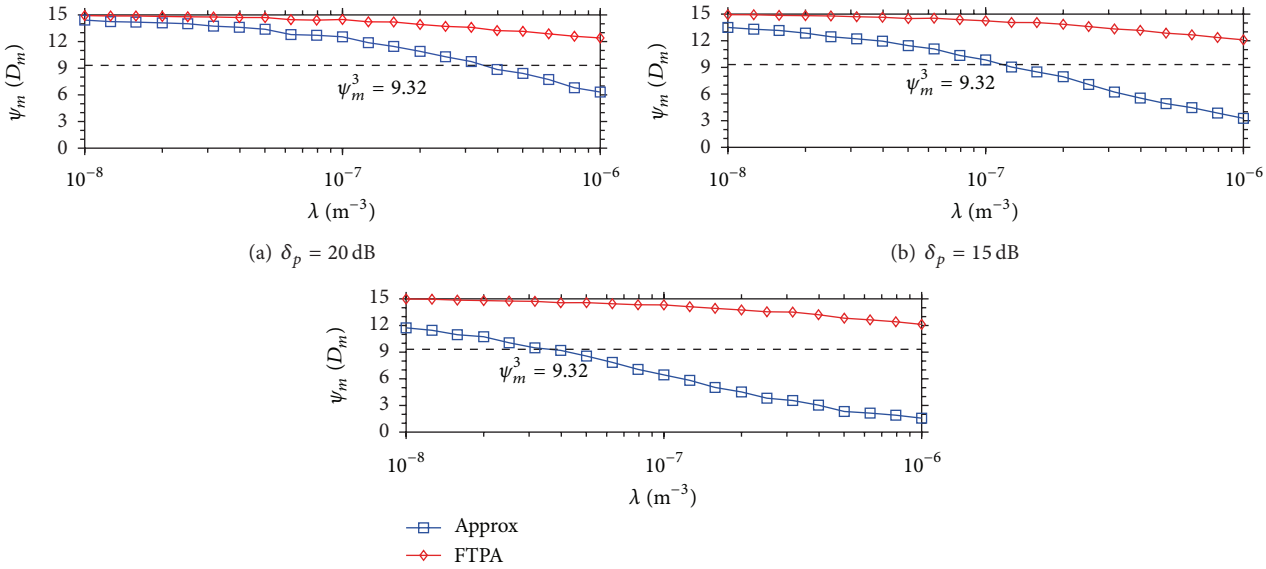
When the QoS type is CBR video, the feasible FBS density with the mean SINR is shown in Figure 6. To meet the higher QoS requirement, the feasible FBS density of Approx is lower than $10^{-6.5}$. As compared with the QoS type of voice, the feasible FBS density (i.e., applying the video service) degrades from $10^{-6.2}$ to $10^{-6.5}$ (see Figure 6(a)). Figure 7 reveals the similar results with that of Figures 5 and 6. Based on these results, a fact can be concluded that FTPA can maintain the mean MU SINR to guarantee the QoS requirement of MU as well as increase the density of FBS development and thus



(a) δ_p = 20 dB (b) δ_p = 15 dB

(c) δ_p = 10 dB

Figure 5: Mean SINR of each MU when the QoS type is voice (ψ_m^1 = 5.31 dB).



(a) δ_p = 20 dB (b) δ_p = 15 dB

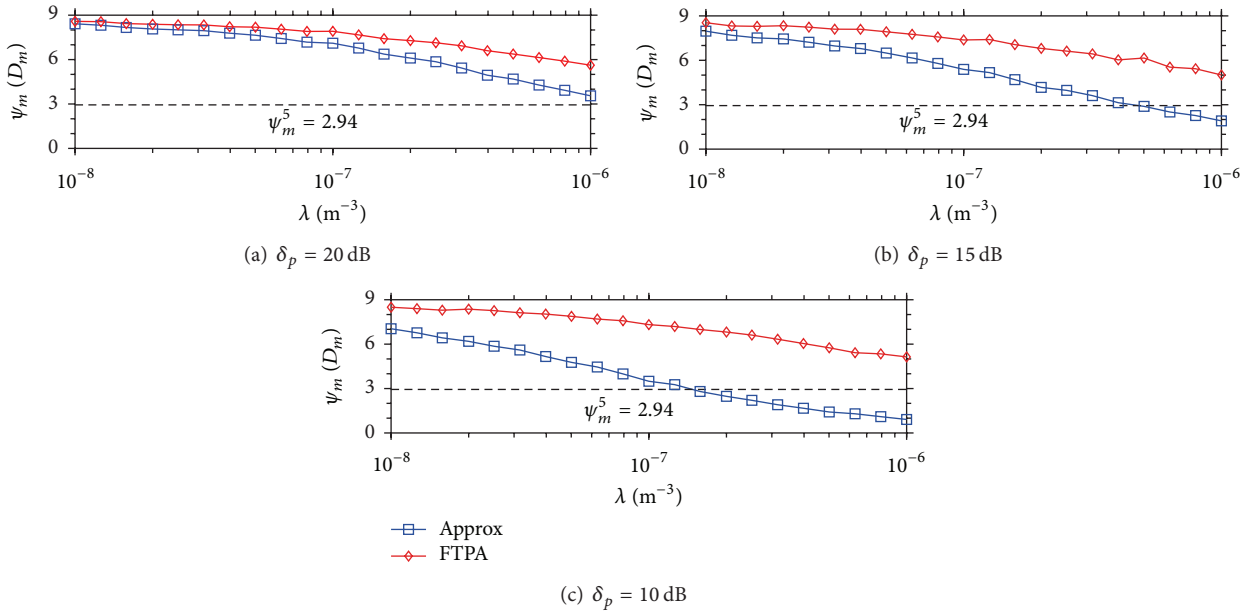
(c) δ_p = 10 dB

Figure 6: Mean SINR of each MU when the QoS type is CBR video (ψ_m^3 = 9.32 dB).

increase the overall system capacity (we discuss it in the following).

To investigate the system capacity achieved by FTPA and Approx approaches, the system capacity of each femtocell is normalized as 1 if the FBS can provide services with a modulation and coding rate of QPSK-1/2 for FUs within 20 m (D_n = 20). The AMC scheme is adopted in the simulation (see Table 3). That is, if the received SINR of an FU located at the point of 20 m from the FBS can support the modulation and coding of 16QAM-1/2, the system capacity of the femtocell becomes 2, and so forth (i.e., 64QAM-1/2 is equal to 3).

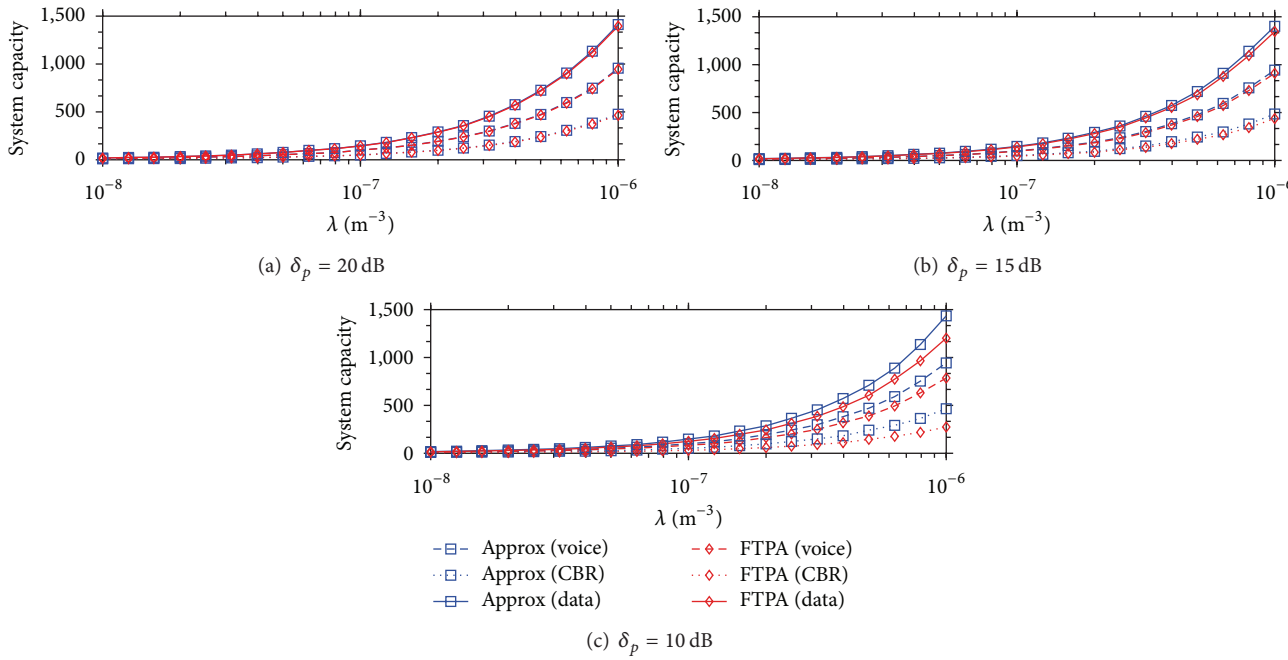
Figure 8 shows the aggregate system capacity under different densities of FBS deployment. The QoS type of voice, CBR video, and data traffic dominate 33% of the total traffic load in the simulation separately. Although increasing the number of FBSs can increase the aggregate system capacity, the MU outage probability will also increase as discussed



(a) δ_p = 20 dB (b) δ_p = 15 dB

(c) δ_p = 10 dB

FIGURE 7: Mean SINR of each MU when the QoS type is data (ψ_m^5 = 2.94 dB).



(a) δ_p = 20 dB (b) δ_p = 15 dB

(c) δ_p = 10 dB

FIGURE 8: Aggregate system capacity achieved by FTPA and Approx versus the FBS deployment density under different δ_p.

above. FTPA can decrease the MU outage probability by adjusting each FTP. However, decreasing each FTP results in decreasing the femtocell capacity. Figure 8(c) shows the consequence of decreasing FTP caused by FTPA that the aggregate system capacity is lower than that of Approx. However, the consequence becomes unapparent when δ_p increases (e.g., concrete walls). These results indicate how FTPA takes advantage of decreasing the femtocell capacity to satisfy the MU QoS requirement.

Figure 9 shows the MU data rate achieved by FTPA and Approx approaches. Although the difference of aggregate system capacity between FTPA and Approx becomes larger when δ_p decreases and λ increases (see Figure 8), the MU data rate decreases very quickly (see Figure 9). This means that the MU QoS is sacrificed to increase the femtocell capacity. However, because FBSs are deployed in the indoor environment, the number of FUs served by one FBS is small (i.e, only 1 or 2 FUs). As a result, FUs may not fully utilize

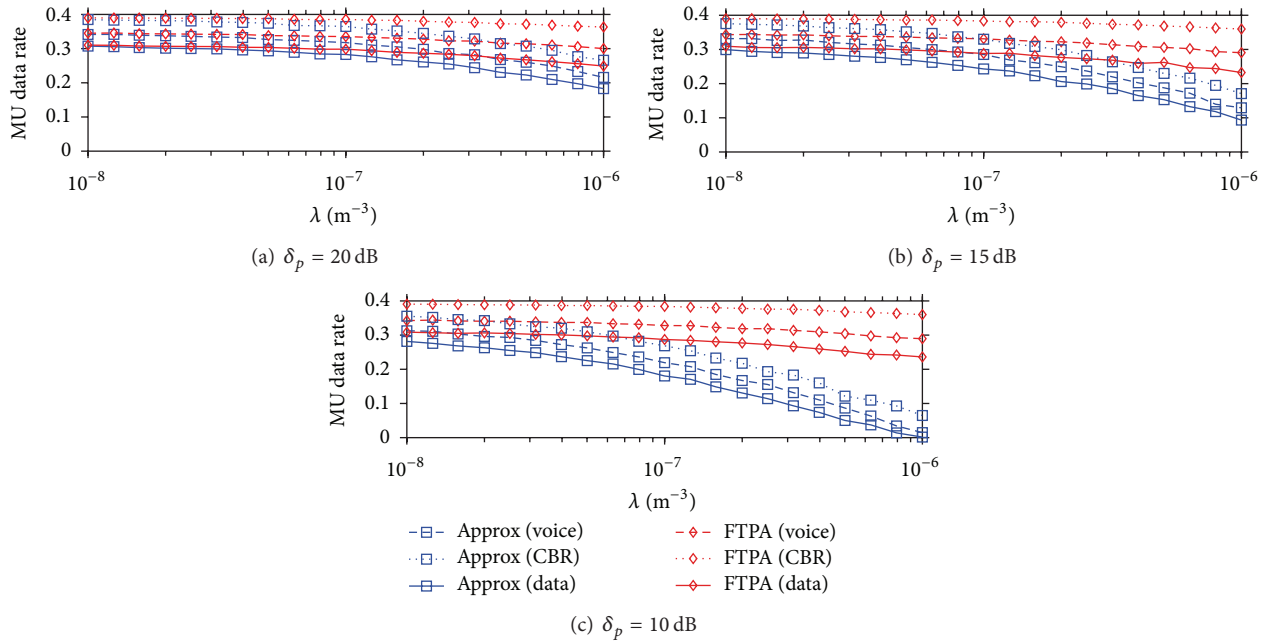


Figure 9: MU data rate achieved by FTPA and Approx versus the FBS deployment density under different δ_p.

the whole femtocell capacity, and the higher MU outage probability is caused (e.g., more than half MUs cannot obtain access service when $\lambda = 10^{-6}$ as shown in Figures 2, 3, and 4). This result provides a strong proof that adjusting FTP to guarantee MU QoS is a feasible solution in the 2-tier heterogeneous femtocell networks.

5. Conclusion

In this study, FTPA is proposed to overcome the interference problem among the MUs and FBSs in the 2-tier heterogeneous femtocell networks. The MU QoS requirement cannot be guaranteed if the interfering FBSs do not reduce the FTP. Considering that the MU locations and penetration loss of different wall materials to assist the FTP adjustment will help the FBS reduce the interfering probability with the nearby MUs as well as provide the indoor network access for FUs. Simulation results give the evidence that FTPA achieves lower MU outage probability, maintains mean MU SINR (i.e., MU data rate), and does not affect the mean FU bandwidth. FTPA is easy to be implemented in the fourth generation networks and meets the trend of NGN in which multimedia applications (voice and video) are the major traffic. In the future work, the femtocell capacity can be improved further if the resource scheduling for MUs is provided to dynamically adjust the FTP. Taking MU handover mechanism to avoid MU interference as well as increase the access bandwidth or offload the loading of macrocell into account is another emerging problem to solve in NGN.

Conflict of Interests

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

Acknowledgments

This work was supported in part by the National Science Council, Taiwan, R.O.C., under Contract NSC102-2221-E-182-032, and the High Speed Intelligent Communication (HSIC) research center, Chang Gung University, Taiwan, R.O.C.

References

[1] 3GPP TS 36.300 V10.10.0, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 10)," July 2013.

[2] J. Hoydis, M. Kobayashi, and M. Debbah, "A cost- and energy-effcient way of meeting the future traffic demands," *IEEE Vehicular Technology Magazine*, vol. 6, no. 1, pp. 37–43, 2011.

[3] T. Zahir, K. Arshad, A. Nakata, and K. Moessner, "Interference management in femtocells," *IEEE Communications Surveys and Tutorials*, vol. 15, no. 1, pp. 293–311, 2012.

[4] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, 2008.

[5] S. Saunders, "The Role of Cooperation in Establishing an Efficient Femto Economy," *Femto Forum*, 2008.

[6] K. Das and S. D. Morgera, "Interference and SIR in integrated voice/data wireless DS-CDMA networks—a simulation study," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 8, pp. 1527–1537, 1997.

[7] I. F. Akyildiz, D. A. Levine, and I. Joe, "A slotted CDMA protocol with BER scheduling for wireless multimedia networks," *IEEE/ACM Transactions on Networking*, vol. 7, no. 2, pp. 146–158, 1999.

[8] X. Wang, "Wide-band TD-CDMA MAC with minimum-power allocation and rate- and BER-scheduling for wireless multimedia networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 1, pp. 103–116, 2004.

[9] M. Luo, G. Villemaud, J.-M. Gorce, and J. Zhang, "Realistic prediction of BER and AMC for indoor wireless transmissions," *IEEE Antennas and Wireless Propagation Letters*, vol. 11, pp. 1084–1087, 2012.

[10] A. Mercado and K. J. R. Liu, "Adaptive QoS for wireless multimedia networks using power control and smart antennas," *IEEE Transactions on Vehicular Technology*, vol. 51, no. 5, pp. 1223–1233, 2002.

[11] G. Guvenc, M.-R. Jeong, F. Watanabe, and H. Inamura, "A hybrid frequency assignment for femtocells and coverage area analysis for co-channel operation," *IEEE Communications Letters*, vol. 12, no. 12, pp. 880–882, 2008.

[12] D.-C. Oh, H.-C. Lee, and Y.-H. Lee, "Power control and beamforming for femtocells in the presence of channel uncertainty," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 6, pp. 2545–2554, 2011.

[13] M. R. Mili and K. A. Hamdi, "On the minimum transmit power in cochannel femtocells," *IEEE Communications Letters*, vol. 16, no. 7, pp. 1026–1029, 2012.

[14] H.-S. Jo, C. Mun, J. Moon, and J.-G. Yook, "Self-optimized coverage coordination in femtocell networks," *IEEE Transactions on Wireless Communications*, vol. 9, no. 10, pp. 2977–2982, 2010.

[15] R. Zhang and L. Hanzo, "Fractional frequency reuse aided twin-layer femtocell networks: analysis, design, and optimization," *IEEE Transactions on Communications*, no. 99, pp. 1–2, 2013.

[16] P. Xu, X. Fang, M. Chenc, and Y. Xu, "A stackelberg game-based spectrum allocation scheme in macro/femtocell hierarchical networks," *Computer Communications*, vol. 36, no. 14, pp. 1552–1558, 2012.

[17] Y. Chen, J. Zhang, and Q. Zhang, "Utility-aware refunding framework for hybrid access femtocell network," *IEEE Transactions on Wireless Communications*, vol. 11, no. 5, pp. 1688–1697, 2012.

[18] O. N. Gharehshiran, A. Attar, and V. Krishnamurthy, "Collaborative sub-channel allocation in cognitive LTE femto-cells: a cooperative game-theoretic approach," *IEEE Transactions on Communications*, vol. 61, no. 1, pp. 325–334, 2013.

[19] S.-M. Cheng, W.-C. Ao, F.-M. Tseng, and K.-C. Chen, "Design and analysis of downlink spectrum sharing in two-tier cognitive femto networks," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 5, pp. 2194–2207, 2012.

[20] W. C. Cheung, T. Q. S. Quek, and M. Kountouris, "Throughput optimization, spectrum allocation, and access control in two-tier femtocell networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 561–574, 2012.

[21] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous LTE-A cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 9, pp. 1479–1489, 2010.

[22] K. Lee, O. Jo, and D.-H. Cho, "Cooperative resource allocation for guaranteeing intercell fairness in femtocell networks," *IEEE Communications Letters*, vol. 15, no. 2, pp. 214–216, 2011.

[23] S. Mukherjee, "Distribution of downlink SINR in heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 575–585, 2012.

[24] H. -S. Jo, Y. J. Sang, P. Xing, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: a comprehensive downlink SINR analysis," *IEEE Transactions on Wireless Communications*, vol. 11, no. 10, pp. 3484–3495, 2012.

[25] Y. L. Tseng and C. Y. Huang, "Analysis of femto base station network deployment," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 2, pp. 748–757, 2012.

[26] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 7, pp. 1029–1046, 2009.

[27] M. Haenggi, "On distances in uniformly random networks," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3584–3586, 2005.

[28] ITU-R Rec M.1225, "Guidelines for evaluation of radio transmission technologies for IMT-2000," *ITU-R Rec M.1225*, February 1997.

[29] V. Mordachev and S. Loyka, "On node density—outage probability tradeoff in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 7, pp. 1120–1131, 2009.

[30] S. H. Ali, K.-D. Lee, and V. C. M. Leung, "Dynamic resource allocation in OFDMA wireless metropolitan area networks," *IEEE Wireless Communications*, vol. 14, no. 1, pp. 6–13, 2007.

[31] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, NY, USA, 2nd edition, 2006.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 306423, 8 pages
http://dx.doi.org/10.1155/2013/306423

Research Article

Multiantenna Relay Beamforming Design for QoS Discrimination in Two-Way Relay Networks

Ke Xiong,[1,2] Yu Zhang,[1] Dandan Li,[1] Chih-Yung Chang,[3] and Zhangdui Zhong[1]

[1] *School of Computer and Information Technology, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China*
[2] *Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*
[3] *Department of Computer Science and Information Engineering, Tamkang University, New Taipei City 25137, Taiwan*

Correspondence should be addressed to Ke Xiong; kxiong@bjtu.edu.cn

Received 13 September 2013; Accepted 27 October 2013

Academic Editors: W. Sun and J. Zhou

This paper investigates the relay beamforming design for quality of service (QoS) discrimination in two-way relay networks. The purpose is to keep legitimate two-way relay users exchange their information via a helping multiantenna relay with QoS guarantee while avoiding the exchanged information overhearing by unauthorized receiver. To this end, we propose a physical layer method, where the relay beamforming is jointly designed with artificial noise (AN) which is used to interfere in the unauthorized user's reception. We formulate the joint beamforming and AN (BFA) design into an optimization problem such that the received signal-to-interference-ratio (SINR) at the two legitimate users is over a predefined QoS threshold while limiting the received SINR at the unauthorized user which is under a certain secure threshold. The objective of the optimization problem is to seek the optimal AN and beamforming vectors to minimize the total power consumed by the relay node. Since the optimization problem is nonconvex, we solve it by using semidefinite program (SDP) relaxation. For comparison, we also study the optimal relay beamforming without using AN (BFO) under the same QoS discrimination constraints. Simulation results show that both the proposed BFA and BFO can achieve the QoS discrimination of the two-way transmission. However, the proposed BFA yields significant power savings and lower infeasible rates compared with the BFO method.

1. Introduction

This paper considers the relay beamforming design to achieve the quality of service (QoS) discrimination between the legitimate users and unauthorized receiver in two-way relay networks, where the system is required to design with such ability; namely, the system is capable of guaranteeing the required communication QoS for the legitimate users meanwhile avoiding the information overhearing by the unauthorized receivers (eavesdroppers). This kind of problems actually can be found in many wireless applications, because of the broadcast nature of wireless medium. For example, the reception performance discrimination is often required between paid and unpaid users in TV broadcast systems, between the members and nonmembers of some organizations in information sharing, and also between authorized users and eavesdropping receivers in secure communications [1].

Traditionally, such QoS discrimination problem was addressed with the employment of application level cryptography and user authentication mechanisms, but recent developments in physical layer secrecy [2] show that this problem can also be effectively handled in the physical layer by exploiting the difference of the fading channels among legitimate and unauthorized users [1].

As for secure communication on QoS discrimination, some works can be found in the literature; see, for example, [1, 3–5], where in [3], the secure communication over broadcast channel was investigated, while it was assumed that the transmitter can broadcast signals with a nonzero coding rate to the legitimate users without any information being eavesdropped by unauthorized users if the mutual information between the transmitter and the legitimate users is higher than that between transmitter and unauthorized users. This physical layer paradigm in information theory

has inspired several recent research efforts; see, for example, [4, 5], where the physical layer secrecy was enhanced via signal processing techniques. It is deserved to be stressed that some of existing works began to consider using the spatial degree of freedom with deploying multiple antennas on the transmitters to achieve and enhance the secrecy for physical layer communications. Specifically, with the use of multiple antennas at the transmitter, beamforming techniques can be adopted to effectively discriminate the reception performances between the legitimate and unauthorized users [4, 5].

Most recently, the authors in [6] proposed an interesting and insightful beamforming scheme where an artificial noise (AN) is purposely added in the transmitted signal for raising the interference level at eavesdroppers. With the channel state information (CSI) known at the transmitter, the AN-aided method may allocate the vector of AN in an orthogonal space of the transmitter-to-legitimate-user's channel in a spatially uniform fashion [7, 8]. By exploiting the CSIs of the unauthorized user, the AN energies may be concentrated on the eavesdroppers' directions to make the QoS discrimination even more effective in practice.

Due to the efficiency of AN-based beamforming design, it has attracted much attention; see, for example, [7–10]. However, most of them considered the joint AN and source beamforming design for one-way transmission wireless systems, where the works in [7, 8] addressed the problem in one-hop broadcast channels and those in [9, 10] addressed the problem for one-way two-hop relay channels.

As the two-way relay transmission is considered as a promisingly applicable transmission model in many wireless systems [11, 12], in this paper, we focus on the joint AN and beamforming design for two-way relay networks.

The contributions of our work are summarized as follows. *Firstly*, we propose a physical layer scheme to achieve QoS discrimination by using AN together with relay beamforming design (BFA) for two-way relay networks. To the best of our knowledge, only two works have investigated the AN-based beamforming design for two-way relay networks, where in both [11, 12] the two-way relay networks with single-antenna relays were considered. Different from existing works, we consider a two-way relay system with a multiantenna relay node. Our goal is to jointly optimally design the AN and beamforming vectors to discriminate the receiving performance between the legitimate and unauthorized users. *Secondly*, we formulate the joint beamforming design into an optimization problem. Our goal is to seek for the jointly optimalized AN and beamforming vectors to minimize the total power consumed by the relay node under the QoS discrimination constraints. Since the optimization problem is nonconvex, we solve it with semidefinite program (SDP) relaxation. *Thirdly*, for comparison, the optimal relay beamforming without using AN method (BFO) is also studied for two-way relay networks. Based on this, extensive simulation results are presented, which show that both the proposed BFA and BFO can realize the QoS discrimination for the two-way relay transmissions. However, the proposed BFA yields significant power savings than BFO.

The rest of this paper is organized as follows. Section 2 introduces the system model. Section 3 describes the proposed BFA, where the joint AN and beamforming vectors design is formulated into an optimization problem and then relaxed into a convex optimization problem by using the SDR method. Section 4 investigates the optimal beamforming design for BFO for compassion. Section 5 presents some simulation results to evaluate the performance of the proposed method, and the paper is summarized with conclusions in Section 6.

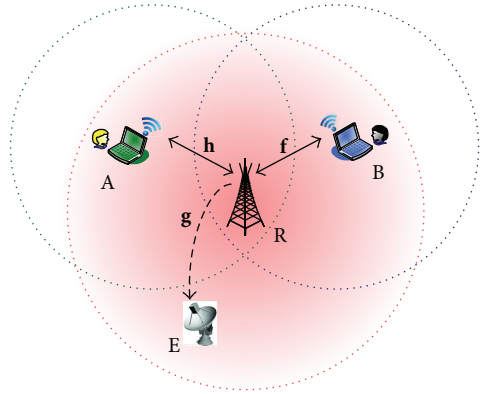


Figure 1: System model and the illustration of secure relay beamforming with artificial noise.

2. System Model

Consider a two-way relay transmission scenario as shown in Figure 1, where two sources, A and B, exchange their information via a relay node R with K antennas, in the presence of an eavesdropper E. We assume that A and B are too far away from each other so that there is no direct link between them. Thus, all information exchange between A and B is helped by the assistant relay R.

It is assumed that the two sources are equipped with single antenna and all channel matrices of the links are known to the transceivers. It is also assumed that the eavesdropper's CSI can be obtained when the eavesdroppers are active in the network [10]. This assumption is applicable in wireless systems, particularly in the multicast and unicast coexisted networks, where terminals play dual roles as legitimate receivers for some signals and eavesdroppers for others.

Half-duplex constraint is considered, so that two phases, that is, the Multiple Access (MA) phase and the Broadcast (BC) phase, are involved to complete a round of information exchange between A and B. In the MA phase, A and B send their signal to the relay node simultaneously. AF protocol is employed, so in the BC phase, relay amplifies the received signals and then broadcasts them through its K antennas to A and B simultaneously. As A and B know their own transmitted signals, the self-interference is able to be canceled and the desired information can be extracted from the received mixed signals. Note that, due to the broadcast nature of wireless links, the eavesdropper E which resides in the system may overhear the information exchange between A and B. Here we assume that E is not within the coverage areas of both A and B but within the service range of relay R. In

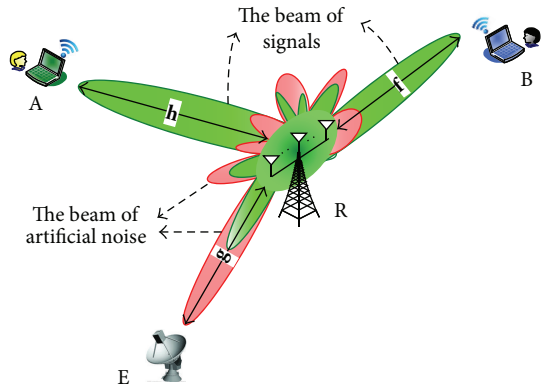


Figure 2: Relay beamforming with artificial noise.

this case, all signals transmitted from R for A and B may be collected by E.

To keep the QoS discrimination for such a two-way relay transmission, the relay node should (1) guarantee the information exchange quality between A and B and, at the same time, (2) prevent the information leakage to the eavesdropper to keep the QoS discrimination. In order to meet these two goals simultaneously, similar to many secure physical layer system design, see, for example, [7], we adopt the received SINR as a measurement. Specifically, the received SINR at each source node should be kept above a predefined QoS threshold to keep the quality of the two-way relay information exchange, while the received SINR at E should be limited below a predefined security threshold, avoiding the information overhearing by the eavesdropper.

Let $\mathbf{h} = [h_1, \ldots, h_K]$, $\mathbf{f} = [f_1, \ldots f_K]$, and $\mathbf{g} = [g_1, \ldots g_K]$ denote the quasistationary flat-fading channel coefficient vectors between A, B, E, and the relay node, respectively. To keep the QoS discrimination for the two-way relay network, we propose a beamforming design method by using artificial noise.

Our basic idea is that (as shown in Figure 2), by proper joint AN and relay beamforming design, the beam pattern of artificial noise can be mainly directed to the eavesdropper, which may effectively interfere with the unauthorized user's reception, while the beam pattern of the signals for the two-way relay users can be mainly directed to the two sources. By doing so, the QoS discrimination for the two-way relay network is able to be enhanced.

Let γ_{A} and γ_{B} be predefined QoS thresholds, for A and B, respectively, and let $\gamma_{\mathrm{E_A}}$ and $\gamma_{\mathrm{E_B}}$ be the predefined secure thresholds at E, for A and B, respectively. Then, an optimization framework of our joint AN and relay beamforming design can be mathematically given by

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\Sigma}} \quad & P_{\mathrm{R}} + P_{\mathrm{AN}} \\ \text{s.t.} \quad & \mathrm{SINR_A} \geq \gamma_{\mathrm{A}}, \\ & \mathrm{SINR_B} \geq \gamma_{\mathrm{B}}, \\ & \mathrm{SINR_{E_A}} \leq \gamma_{\mathrm{E_A}}, \\ & \mathrm{SINR_{E_B}} \leq \gamma_{\mathrm{E_B}}, \end{aligned} \tag{1}$$

where P_{R} is total power of the relay beamforming vectors and P_{AN} is total power of the artificial noise. $\mathrm{SINR_A}$ and $\mathrm{SINR_B}$ are the end-to-end received SNR at A and B, respectively. $\mathrm{SINR_{E_A}}$ and $\mathrm{SINR_{E_B}}$ are the received SNR at E for the signal transmitted from B and A and the signal transmitted from B and A, respectively. $\mathbf{w}$ and $\boldsymbol{\Sigma}$ represent the relay beamforming vector and artificial noise vector, respectively.

3. Optimal Relay Beamforming with Artificial Noise

This section presents the proposed BFA method and then designs the optimal AN and beam vectors for it.

3.1. BFA. In the MA phase of BFA, A and B send their information to the relay node simultaneously. Thus, the received signals at the relays and at E can be, respectively, given by

$$\mathbf{y}_{\mathrm{R}} = \sqrt{P_{\mathrm{A}}}\mathbf{h}_{\mathrm{R}} x_{\mathrm{A}} + \sqrt{P_{\mathrm{B}}}\mathbf{f}_{\mathrm{R}} x_{\mathrm{B}} + \mathbf{n}_{\mathrm{R}}, \tag{2}$$

where $\mathbf{y}_{\mathrm{R}}$ is a $K \times 1$ complex vector of the received signal at the K antennas at R and y_{E} is the received signal at E in the MA phase. P_{A} and P_{B} denote the transmit power of A and B, respectively. $\mathbf{n}_{\mathrm{R}}$ is a $K \times 1$ complex vector of Additive White Gaussian Noise (AWGN) at the relay.

In the BC phase, the relay amplifies the received signal for the ith antenna by a complex beamforming weight w_i. Thus the beam for the amplified signals can be expressed as $\boldsymbol{\Omega}\mathbf{y}_{\mathrm{R}}$. Here, we adopt the artificial noise method [6], where the relay transmits artificial noise (interference) to mask the concurrent transmission of information bearing signal to the eavesdroppers. Let ε be the $K \times 1$ artificial noise vector. Then, the transmit signal vector $\mathbf{x}$ at the relay node can be expressed as

$$\mathbf{x}_{\mathrm{R}}^{(\mathrm{RFA})} = \boldsymbol{\Omega}\mathbf{y}_{\mathrm{R}} + \varepsilon. \tag{3}$$

Note that the design artificial noise follows the zeros-mean complex Gaussian distribution with covariance matrix $\boldsymbol{\Sigma} \succeq 0$. Thus, the signal received at A, B, and E can be expressed as

$$\begin{aligned} y_{\mathrm{A}}^{(\mathrm{BFA})} &= \mathbf{h}_{\mathrm{R}}^{T}\mathbf{x}^{(\mathrm{BFA})} + n_{\mathrm{A}} \\ &= \underbrace{\sqrt{P_{\mathrm{A}}}\mathbf{h}_{\mathrm{R}}^{T}\boldsymbol{\Omega}\mathbf{h}_{\mathrm{R}} x_{\mathrm{A}}}_{\text{self-interference}} + \underbrace{\sqrt{P_{\mathrm{B}}}\mathbf{h}_{\mathrm{R}}^{T}\boldsymbol{\Omega}\mathbf{f}_{\mathrm{R}} x_{\mathrm{B}}}_{\text{desired signal}} \\ &\quad + \underbrace{\mathbf{h}_{\mathrm{R}}^{T}\boldsymbol{\Omega}\mathbf{n}_{\mathrm{R}} + n_{\mathrm{A}}}_{\text{noise}} + \underbrace{\mathbf{h}_{\mathrm{R}}^{T}\varepsilon}_{\text{AN}}, \end{aligned}$$

$$
\begin{aligned}
y_{\mathrm{B}}^{(\mathrm{BFA})} &= \mathbf{f}_{\mathrm{R}}^{T}\mathbf{x}^{(\mathrm{BFA})} + n_{\mathrm{B}} \\
&= \underbrace{\sqrt{P_{\mathrm{A}}}\mathbf{f}_{\mathrm{R}}^{T}\boldsymbol{\Omega}\mathbf{h}_{\mathrm{R}}x_{\mathrm{A}}}_{\text{desired signal}} + \underbrace{\sqrt{P_{\mathrm{B}}}\mathbf{f}_{\mathrm{R}}^{T}\boldsymbol{\Omega}\mathbf{f}_{\mathrm{R}}x_{\mathrm{B}}}_{\text{self-interference}} \\
&\quad + \underbrace{\mathbf{f}_{\mathrm{R}}^{T}\boldsymbol{\Omega}\mathbf{n}_{\mathrm{R}} + n_{\mathrm{B}}}_{\text{noise}} + \underbrace{\mathbf{f}_{\mathrm{R}}^{T}\boldsymbol{\varepsilon}}_{\text{AN}}, \\
y_{\mathrm{E}}^{(\mathrm{BFA})} &= \mathbf{g}_{\mathrm{R}}^{T}\mathbf{x}^{(\mathrm{BFA})} + n_{\mathrm{B}} \\
&= \underbrace{\sqrt{P_{\mathrm{A}}}\mathbf{g}_{\mathrm{R}}^{T}\boldsymbol{\Omega}\mathbf{h}_{\mathrm{R}}x_{\mathrm{A}}}_{\text{desired signal}} + \underbrace{\sqrt{P_{\mathrm{B}}}\mathbf{g}_{\mathrm{R}}^{T}\boldsymbol{\Omega}\mathbf{f}_{\mathrm{R}}x_{\mathrm{B}}}_{\text{self-interference}} \\
&\quad + \underbrace{\mathbf{g}_{\mathrm{R}}^{T}\boldsymbol{\Omega}\mathbf{n}_{\mathrm{R}} + n_{\mathrm{E}}}_{\text{noise}} + \underbrace{\mathbf{g}_{\mathrm{R}}^{T}\boldsymbol{\varepsilon}}_{\text{AN}},
\end{aligned}
\tag{4}
$$

respectively. In terms of (4), the end-to-end received SINRs at A and B can be given by

$$
\begin{aligned}
\mathrm{SINR}_{\mathrm{A}}^{(\mathrm{BFA})} &= \frac{P_{\mathrm{B}}\mathbf{w}^{H}\mathbf{G}_{\mathrm{AB}}\mathbf{w}}{\sigma^{2}\mathbf{w}^{H}\mathbf{D}_{\mathrm{A}}\mathbf{w} + \mathrm{tr}\left(\mathbf{H}_{\mathrm{A}}\boldsymbol{\Sigma}\right) + \sigma^{2}}, \\
\mathrm{SINR}_{\mathrm{B}}^{(\mathrm{BFA})} &= \frac{P_{\mathrm{A}}\mathbf{w}^{H}\mathbf{G}_{\mathrm{AB}}\mathbf{w}}{\sigma^{2}\mathbf{w}^{H}\mathbf{D}_{\mathrm{B}}\mathbf{w} + \mathrm{tr}\left(\mathbf{H}_{\mathrm{B}}\boldsymbol{\Sigma}\right) + \sigma^{2}},
\end{aligned}
\tag{5}
$$

where $\mathbf{H}_{\mathrm{A}} = \mathbf{h}_{\mathrm{R}}\mathbf{h}_{\mathrm{R}}^{H}$, $\mathbf{H}_{\mathrm{B}} = \mathbf{f}_{\mathrm{R}}\mathbf{f}_{\mathrm{R}}^{H}$, and

$$
\begin{aligned}
\mathbf{G}_{\mathrm{AB}} &= \mathrm{diag}\left(\mathbf{h}_{\mathrm{R}}\right)\mathbf{f}_{\mathrm{R}}\left(\mathrm{diag}\left(\mathbf{h}_{\mathrm{R}}\right)\mathbf{f}_{\mathrm{R}}\right)^{H}, \\
\mathbf{D}_{\mathrm{A}} &= \mathrm{diag}\left(\mathbf{h}_{\mathrm{R}}\right)\left(\mathrm{diag}\left(\mathbf{h}_{\mathrm{R}}\right)\right)^{H}, \\
\mathbf{D}_{\mathrm{B}} &= \mathrm{diag}\left(\mathbf{f}_{\mathrm{R}}\right)\left(\mathrm{diag}\left(\mathbf{f}_{\mathrm{R}}\right)\right)^{H}, \\
\mathbf{w} &= \left[w_{1}, w_{2}, \ldots w_{K}\right]^{H}.
\end{aligned}
\tag{6}
$$

$\mathbf{w}$ is the beam vector which is required to design. By using some signal detection methods, E tries to decode x_{A} and x_{B}. The received SINR at E for decoding x_{A} and x_{B} can be, respectively, given by

$$
\begin{aligned}
\mathrm{SINR}_{\mathrm{E_A}}^{(\mathrm{BFA})} &= \frac{P_{\mathrm{A}}\mathbf{w}^{H}\mathbf{G}_{\mathrm{AE}}\mathbf{w}}{P_{\mathrm{B}}\mathbf{w}^{H}\mathbf{G}_{\mathrm{BE}}\mathbf{w} + \sigma^{2}\mathbf{w}^{H}\mathbf{D}_{\mathrm{E}}\mathbf{w} + \mathrm{tr}\left(\mathbf{F}_{\mathrm{E}}\boldsymbol{\Sigma}\right) + \sigma^{2}}, \\
\mathrm{SINR}_{\mathrm{E_B}}^{(\mathrm{BFA})} &= \frac{P_{\mathrm{B}}\mathbf{w}^{H}\mathbf{G}_{\mathrm{BE}}\mathbf{w}}{P_{\mathrm{A}}\mathbf{w}^{H}\mathbf{G}_{\mathrm{AE}}\mathbf{w} + \sigma^{2}\mathbf{w}^{H}\mathbf{D}_{\mathrm{E}}\mathbf{w} + \mathrm{tr}\left(\mathbf{F}_{\mathrm{E}}\boldsymbol{\Sigma}\right) + \sigma^{2}},
\end{aligned}
\tag{7}
$$

where $\mathbf{F}_{\mathrm{E}} = \mathbf{g}_{\mathrm{R}}\mathbf{g}_{\mathrm{R}}^{H}$ and

$$
\begin{aligned}
\mathbf{G}_{\mathrm{AE}} &= \mathrm{diag}\left(\mathbf{h}_{\mathrm{R}}\right)\mathbf{g}_{\mathrm{R}}\left(\mathrm{diag}\left(\mathbf{h}_{\mathrm{R}}\right)\mathbf{g}_{\mathrm{R}}\right)^{H}, \\
\mathbf{G}_{\mathrm{BE}} &= \mathrm{diag}\left(\mathbf{f}_{\mathrm{R}}\right)\mathbf{g}_{\mathrm{R}}\left(\mathrm{diag}\left(\mathbf{f}_{\mathrm{R}}\right)\mathbf{g}_{\mathrm{R}}\right)^{H}, \\
\mathbf{D}_{\mathrm{E}} &= \mathrm{diag}\left(\mathbf{g}_{\mathrm{R}}\right)\left(\mathrm{diag}\left(\mathbf{g}_{\mathrm{R}}\right)\right)^{H}.
\end{aligned}
\tag{8}
$$

Based on the description above, we are going to find the jointly optimized $\mathbf{w}$ and $\boldsymbol{\Sigma}$ for BFA.

3.2. Optimal Beam Vectors Design for BFA. In this section, we design the jointly optimized beamforming vector $\mathbf{w}$ and artificial noise vector $\boldsymbol{\Sigma}$ for BFA to minimize the total transmit powers at the relay R. To meet the security requirement of the system, two constraints are considered.

Since

$$
P_{\mathrm{R}} = E\left\{\mathbf{x}_{\mathrm{R}}^{(\mathrm{PA})}\left(\mathbf{x}_{\mathrm{R}}^{(\mathrm{PA})}\right)^{H}\right\}, \tag{9}
$$

according to (3), we can extend the expression for P_{R} as

$$
\begin{aligned}
P_{\mathrm{R}} &= E\left\{\mathbf{x}_{\mathrm{R}}^{(\mathrm{PA})}\left(\mathbf{x}_{\mathrm{R}}^{(\mathrm{PA})}\right)^{H}\right\} \\
&= P_{\mathrm{A}}\mathbf{w}^{H}\mathbf{D}_{\mathrm{A}}\mathbf{w} + P_{\mathrm{B}}\mathbf{w}^{H}\mathbf{D}_{\mathrm{B}}\mathbf{w} + \sigma^{2}\mathbf{w}^{H}\mathbf{w} + \mathrm{tr}\left(\boldsymbol{\Sigma}\right),
\end{aligned}
\tag{10}
$$

where $\mathrm{tr}(\boldsymbol{\Sigma})$ actually is the power of artificial noise. By substituting (5), (7), (9) and (10) into problem (1), then we have that

$$
\begin{aligned}
\min_{\mathbf{w},\boldsymbol{\Sigma}} \quad & P_{\mathrm{A}}\mathbf{w}^{H}\mathbf{D}_{\mathrm{A}}\mathbf{w} + P_{\mathrm{B}}\mathbf{w}^{H}\mathbf{D}_{\mathrm{B}}\mathbf{w} + \sigma^{2}\mathbf{w}^{H}\mathbf{w} + \mathrm{tr}\left(\boldsymbol{\Sigma}\right) \\
\text{s.t.} \quad & \frac{P_{\mathrm{B}}\mathbf{w}^{H}\mathbf{G}_{\mathrm{AB}}\mathbf{w}}{\sigma^{2}\mathbf{w}^{H}\mathbf{D}_{\mathrm{A}}\mathbf{w} + \mathrm{tr}(\mathbf{H}_{\mathrm{A}}\boldsymbol{\Sigma}) + \sigma^{2}} \geq \gamma_{\mathrm{A}} \\
& \frac{P_{\mathrm{A}}\mathbf{w}^{H}\mathbf{G}_{\mathrm{AB}}\mathbf{w}}{\sigma^{2}\mathbf{w}^{H}\mathbf{D}_{\mathrm{B}}\mathbf{w} + \mathrm{tr}(\mathbf{H}_{\mathrm{B}}\boldsymbol{\Sigma}) + \sigma^{2}} \geq \gamma_{\mathrm{B}} \\
& \frac{P_{\mathrm{A}}\mathbf{w}^{H}\mathbf{G}_{\mathrm{AE}}\mathbf{w}}{P_{\mathrm{B}}\mathbf{w}^{H}\mathbf{G}_{\mathrm{BE}}\mathbf{w} + \sigma^{2}\mathbf{w}^{H}\mathbf{D}_{\mathrm{E}}\mathbf{w} + \mathrm{tr}\left(\mathbf{F}_{\mathrm{E}}\boldsymbol{\Sigma}\right) + \sigma^{2}} \leq \gamma_{\mathrm{E_A}} \\
& \frac{P_{\mathrm{B}}\mathbf{w}^{H}\mathbf{G}_{\mathrm{BE}}\mathbf{w}}{P_{\mathrm{A}}\mathbf{w}^{H}\mathbf{G}_{\mathrm{AE}}\mathbf{w} + \sigma^{2}\mathbf{w}^{H}\mathbf{D}_{\mathrm{E}}\mathbf{w} + \mathrm{tr}\left(\mathbf{F}_{\mathrm{E}}\boldsymbol{\Sigma}\right) + \sigma^{2}} \leq \gamma_{\mathrm{E_B}}.
\end{aligned}
\tag{11}
$$

Since problem (11) is also nonconvex, by using the SDP relaxation method [13], problem (11) also can be relaxed to be convex. Let $\mathbf{W} = \mathbf{w}\mathbf{w}^{H}$; problem (11) then can be relaxed as (12a). Consider

$$
\min_{\mathbf{W},\boldsymbol{\Sigma}} \quad P_{\mathrm{A}}\,\mathrm{tr}\left(\mathbf{D}_{\mathrm{A}}\mathbf{W}\right) + P_{\mathrm{B}}\,\mathrm{tr}\left(\mathbf{D}_{\mathrm{B}}\mathbf{W}\right) + \sigma^{2}\,\mathrm{tr}\left(\mathbf{W}\right) + \mathrm{tr}\left(\boldsymbol{\Sigma}\right) \tag{12a}
$$

$$
\begin{aligned}
\text{s.t.} \quad & \sigma^{2}\gamma_{\mathrm{A}}\,\mathrm{tr}\left(\mathbf{D}_{\mathrm{A}}\mathbf{W}\right) - P_{\mathrm{B}}\,\mathrm{tr}\left(\mathbf{G}_{\mathrm{AB}}\mathbf{W}\right) \\
& \quad + \gamma_{\mathrm{A}}\,\mathrm{tr}\left(\mathbf{H}_{\mathrm{A}}\boldsymbol{\Sigma}\right) + \sigma^{2}\gamma_{\mathrm{A}} \leq 0
\end{aligned} \tag{12b}
$$

$$
\begin{aligned}
& \sigma^{2}\gamma_{\mathrm{B}}\,\mathrm{tr}\left(\mathbf{D}_{\mathrm{B}}\mathbf{W}\right) - P_{\mathrm{A}}\,\mathrm{tr}\left(\mathbf{G}_{\mathrm{AB}}\mathbf{W}\right) \\
& \quad + \gamma_{\mathrm{B}}\,\mathrm{tr}\left(\mathbf{H}_{\mathrm{B}}\boldsymbol{\Sigma}\right) + \sigma^{2}\gamma_{\mathrm{B}} \leq 0
\end{aligned} \tag{12c}
$$

$$
\begin{aligned}
& P_{\mathrm{A}}\,\mathrm{tr}\left(\mathbf{G}_{\mathrm{AE}}\mathbf{W}\right) - \sigma^{2}\gamma_{\mathrm{E_A}}\,\mathrm{tr}\left(\mathbf{D}_{\mathrm{E}}\mathbf{W}\right) \\
& \quad - \left(\sigma^{2} + P_{\mathrm{B}}\,\mathrm{tr}\left(\mathbf{G}_{\mathrm{BE}}\mathbf{W}\right) + \mathrm{tr}\left(\mathbf{F}_{\mathrm{E}}\boldsymbol{\Sigma}\right)\right)\gamma_{\mathrm{E_A}} \leq 0
\end{aligned} \tag{12d}
$$

$$
\begin{aligned}
& P_{\mathrm{B}}\,\mathrm{tr}\left(\mathbf{G}_{\mathrm{BE}}\mathbf{W}\right) - \sigma^{2}\left(\gamma_{\mathrm{E_B}} - \eta_{\mathrm{E_B}}\right)\mathrm{tr}\left(\mathbf{D}_{\mathrm{E}}\mathbf{W}\right) \\
& \quad - \left(\sigma^{2} + P_{\mathrm{A}}\,\mathrm{tr}\left(\mathbf{G}_{\mathrm{AE}}\mathbf{W}\right) + \mathrm{tr}\left(\mathbf{F}_{\mathrm{E}}\boldsymbol{\Sigma}\right)\right) \\
& \quad \times \left(\gamma_{\mathrm{E_B}} - \eta_{\mathrm{E_B}}\right) \leq 0
\end{aligned} \tag{12e}
$$

$$
\mathbf{W} \succeq 0 \tag{12f}
$$

$$
\mathrm{rank}\left(\mathbf{W}\right) = 1. \tag{12g}
$$

Following the SDP relaxation theory, the hard constraint $\text{rank}(\mathbf{W}) = 1$ also can be neglected and then the new relaxed problem is given by

$$\begin{aligned} \min_{\mathbf{W},\boldsymbol{\Sigma}} \quad & P_{\mathrm{A}} \operatorname{tr}\left(\mathbf{D}_{\mathrm{A}}\mathbf{W}\right) + P_{\mathrm{B}} \operatorname{tr}\left(\mathbf{D}_{\mathrm{B}}\mathbf{W}\right) + \sigma^2 \operatorname{tr}\left(\mathbf{W}\right) + \operatorname{tr}\left(\boldsymbol{\Sigma}\right) \\ \text{s.t.} \quad & (12\text{b}), (12\text{c}), (12\text{d}), (12\text{e}), (12\text{f}), \end{aligned} \tag{13}$$

which is a convex SDP, and therefore can be efficiently solved to obtain the global optimum by the available solvers, for example, CVX [14].

It should be noted that, since the rank-one constraint is dropped in (12a), the optimal solution $\mathbf{W}^*$ is not necessarily rank-one. Based on the rank reduction results for general SDPs, namely, Lemma 3.1 in [15], we can derive that $\text{rank}(\mathbf{W}^*) = 1$ or $\text{rank}(\mathbf{W}^*) = 2$. Therefore, if $\text{rank}(\mathbf{W}^*) = 1$, the optimal beamforming vector $\mathbf{w}^*$ can be retrieved from $\mathbf{W}^*$ exactly. If $\text{rank}(\mathbf{W}^*) = 2$, Gaussian randomization method [12] can be applied to obtain an approximated $\mathbf{w}^*$. Interestingly, the optimal solutions $\mathbf{W}^*$ in our simulations are all rank-one, which means that $\mathbf{w}^*$ can be retrieved from $\mathbf{W}^*$ exactly.

4. Relay Beamforming without Artificial Noise (BFO)

4.1. BFA. In this subsection, we describe the BFO scheme. The process in the MA phase of BFA is the same as that of BFO, so we do not repeat the description of it again.

In the BC phase, the ith antenna amplifies the received signal by a complex beamforming weight w_i. Thus, the processed signal vector at the relay node can be written as a $K \times 1$ complex vector as follows:

$$\mathbf{x}_{\mathrm{R}}^{(\mathrm{BFO})} = \boldsymbol{\Omega}\mathrm{y}_{\mathrm{R}}, \tag{14}$$

where $\boldsymbol{\Omega} = \text{diag}([w_1, w_2, \ldots w_{\mathrm{K}}])$.

After this, the relay broadcasts the processed signals to A and B. So, the signals received at A, B, and E can be expressed as

$$\begin{aligned} y_{\mathrm{A}}^{(\mathrm{BFO})} &= \mathbf{h}_{\mathrm{R}}^T\mathbf{x}^{(\mathrm{BFO})} + n_{\mathrm{A}} = \mathbf{h}_{\mathrm{R}}^T\boldsymbol{\Omega}\mathbf{y}_{\mathrm{R}} + n_{\mathrm{A}} \\ &= \underbrace{\sqrt{P_{\mathrm{A}}}\mathbf{h}_{\mathrm{R}}^T\boldsymbol{\Omega}\mathbf{h}_{\mathrm{R}}x_{\mathrm{A}}}_{\text{self-interference}} + \underbrace{\sqrt{P_{\mathrm{B}}}\mathbf{h}_{\mathrm{R}}^T\boldsymbol{\Omega}\mathbf{f}_{\mathrm{R}}x_{\mathrm{B}}}_{\text{desired signal}} + \underbrace{\mathbf{h}_{\mathrm{R}}^T\boldsymbol{\Omega}\mathbf{n}_{\mathrm{R}} + n_{\mathrm{A}}}_{\text{noise}} \\ y_{\mathrm{B}}^{(\mathrm{BFO})} &= \mathbf{f}_{\mathrm{R}}^T\mathbf{x}^{(\mathrm{BFO})} + n_{\mathrm{B}} = \mathbf{f}_{\mathrm{R}}^T\boldsymbol{\Omega}\mathbf{y}_{\mathrm{R}} + n_{\mathrm{B}} \\ &= \underbrace{\sqrt{P_{\mathrm{A}}}\mathbf{f}_{\mathrm{R}}^T\boldsymbol{\Omega}\mathbf{h}_{\mathrm{R}}x_{\mathrm{A}}}_{\text{desired signal}} + \underbrace{\sqrt{P_{\mathrm{B}}}\mathbf{f}_{\mathrm{R}}^T\boldsymbol{\Omega}\mathbf{f}_{\mathrm{R}}x_{\mathrm{B}}}_{\text{self-interference}} + \underbrace{\mathbf{f}_{\mathrm{R}}^T\boldsymbol{\Omega}\mathbf{n}_{\mathrm{R}} + n_{\mathrm{B}}}_{\text{noise}} \\ y_{\mathrm{E}}^{(\mathrm{BFO})} &= \mathbf{g}_{\mathrm{R}}^T\mathbf{x}^{(\mathrm{BFO})} + n_{\mathrm{E_{BC}}} = \mathbf{g}_{\mathrm{R}}^T\boldsymbol{\Omega}\mathbf{y}_{\mathrm{R}} + n_{\mathrm{E_{BC}}} \\ &= \sqrt{P_{\mathrm{A}}}\mathbf{g}_{\mathrm{R}}^T\boldsymbol{\Omega}\mathbf{h}_{\mathrm{R}}x_{\mathrm{A}} + \sqrt{P_{\mathrm{B}}}\mathbf{g}_{\mathrm{R}}^T\boldsymbol{\Omega}\mathbf{f}_{\mathrm{R}}x_{\mathrm{B}} + \underbrace{\mathbf{g}_{\mathrm{R}}^T\boldsymbol{\Omega}\mathbf{n}_{\mathrm{R}} + n_{\mathrm{E_{BC}}}}_{\text{noise}}, \end{aligned} \tag{15}$$

where n_{A}, n_{B}, and $n_{\mathrm{E_{BC}}}$ are the noise received at A, B, and E, respectively. Since A and B know their own transmitted signals, that is, x_{A} and x_{B}, respectively, they can cancel the self-interference. Thus, in terms of (5) and (7), the end-to-end received SINR at A and B can be, respectively, given in

$$\begin{aligned} \mathrm{SINR}_{\mathrm{A}}^{(\mathrm{BFO})} &= \frac{P_{\mathrm{B}}\mathbf{w}^H\mathbf{G}_{\mathrm{AB}}\mathbf{w}}{\sigma^2\mathbf{w}^H\mathbf{D}_{\mathrm{A}}\mathbf{w} + \sigma^2} \\ \mathrm{SINR}_{\mathrm{B}}^{(\mathrm{BFO})} &= \frac{P_{\mathrm{A}}\mathbf{w}^H\mathbf{G}_{\mathrm{AB}}\mathbf{w}}{\sigma^2\mathbf{w}^H\mathbf{D}_{\mathrm{B}}\mathbf{w} + \sigma^2}. \end{aligned} \tag{16}$$

For E, it collects the signals in both phases; we assume that the MRC is used at E to extract the desired signals. Therefore, the received SINR for the signals transmitted from A and B at E can be given, respectively, by

$$\begin{aligned} \mathrm{SINR}_{\mathrm{E_A}}^{(\mathrm{BFO})} &= \frac{P_{\mathrm{A}}\mathbf{w}^H\mathbf{G}_{\mathrm{AE}}\mathbf{w}}{P_{\mathrm{B}}\mathbf{w}^H\mathbf{G}_{\mathrm{BE}}\mathbf{w} + \sigma^2\mathbf{w}^H\mathbf{D}_{\mathrm{E}}\mathbf{w} + \sigma^2} \\ \mathrm{SINR}_{\mathrm{E_B}}^{(\mathrm{BFO})} &= \frac{P_{\mathrm{B}}\mathbf{w}^H\mathbf{G}_{\mathrm{BE}}\mathbf{w}}{P_{\mathrm{A}}\mathbf{w}^H\mathbf{G}_{\mathrm{AE}}\mathbf{w} + \sigma^2\mathbf{w}^H\mathbf{D}_{\mathrm{E}}\mathbf{w} + \sigma^2}. \end{aligned} \tag{17}$$

4.2. Optimal Beam Vector Design for BFO. Compared with the P_{R} of BFA, the P_{R} of BFO has only one term, that is, without $\text{tr}(\boldsymbol{\Sigma})$. Similarly to the analysis for BFA,

$$\begin{aligned} P_{\mathrm{R}} = P_{\mathrm{R}}^{(\mathrm{BFO})} &= E\left\{\mathbf{x}_{\mathrm{R}}^{(\mathrm{BFO})}\left(\mathbf{x}_{\mathrm{R}}^{(\mathrm{BFO})}\right)^H\right\} \\ &= P_{\mathrm{A}}\mathbf{w}^H\mathbf{D}_{\mathrm{A}}\mathbf{w} + P_{\mathrm{B}}\mathbf{w}^H\mathbf{D}_{\mathrm{B}}\mathbf{w} + \sigma^2\mathbf{w}^H\mathbf{w}. \end{aligned} \tag{18}$$

By substituting (18), (16), (17), and $P_{\mathrm{AN}} = 0$ into (7), we have that

$$\begin{aligned} \min_{\mathbf{w}} \quad & P_{\mathrm{A}}\mathbf{w}^H\mathbf{D}_{\mathrm{A}}\mathbf{w} + P_{\mathrm{B}}\mathbf{w}^H\mathbf{D}_{\mathrm{B}}\mathbf{w} + \sigma^2\mathbf{w}^H\mathbf{w} \\ \text{s.t.} \quad & \frac{P_{\mathrm{B}}\mathbf{w}^H\mathbf{G}_{\mathrm{AB}}\mathbf{w}}{\sigma^2\mathbf{w}^H\mathbf{D}_{\mathrm{A}}\mathbf{w} + \sigma^2} \geq \gamma_{\mathrm{A}} \\ & \frac{P_{\mathrm{A}}\mathbf{w}^H\mathbf{G}_{\mathrm{AB}}\mathbf{w}}{\sigma^2\mathbf{w}^H\mathbf{D}_{\mathrm{B}}\mathbf{w} + \sigma^2} \geq \gamma_{\mathrm{B}} \\ & \frac{P_{\mathrm{A}}\mathbf{w}^H\mathbf{G}_{\mathrm{AE}}\mathbf{w}}{P_{\mathrm{B}}\mathbf{w}^H\mathbf{G}_{\mathrm{BE}}\mathbf{w} + \sigma^2\mathbf{w}^H\mathbf{D}_{\mathrm{E}}\mathbf{w} + \sigma^2} \leq \gamma_{\mathrm{E_A}} \\ & \frac{P_{\mathrm{A}}\mathbf{w}^H\mathbf{G}_{\mathrm{AE}}\mathbf{w}}{P_{\mathrm{B}}\mathbf{w}^H\mathbf{G}_{\mathrm{BE}}\mathbf{w} + \sigma^2\mathbf{w}^H\mathbf{D}_{\mathrm{E}}\mathbf{w} + \sigma^2} \leq \gamma_{\mathrm{E_A}}. \end{aligned} \tag{19}$$

Since the problem of (19) is also nonconvex, we handle it using SDP relaxation theory similar to that for BFA. By

introducing a new variable $\mathbf{W} = \mathbf{w}\mathbf{w}^H$, the problem of (19) can be transformed into

$$\min_{\mathbf{W}} \quad P_A \operatorname{tr}(\mathbf{D}_A\mathbf{W}) + P_B \operatorname{tr}(\mathbf{D}_B\mathbf{W}) + \sigma^2 \operatorname{tr}(\mathbf{W}) \tag{20a}$$

$$\text{s.t.} \quad \sigma^2\gamma_A \operatorname{tr}(\mathbf{D}_A\mathbf{W}) - P_B \operatorname{tr}(\mathbf{G}_{AB}\mathbf{W}) + \sigma^2\gamma_A \le 0 \tag{20b}$$

$$P_A \operatorname{tr}(\mathbf{G}_{AE}\mathbf{W}) - \sigma^2\gamma_{E_A} \operatorname{tr}(\mathbf{D}_E\mathbf{W}) - \left(\sigma^2 + P_B \operatorname{tr}(\mathbf{G}_{BE}\mathbf{W})\right)\gamma_{E_A} \le 0 \tag{20c}$$

$$P_B \operatorname{tr}(\mathbf{G}_{BE}\mathbf{W}) - \sigma^2\gamma_{E_B} \operatorname{tr}(\mathbf{D}_E\mathbf{W}) - \left(\sigma^2 + P_A \operatorname{tr}(\mathbf{G}_{AE}\mathbf{W})\right)\gamma_{E_B} \le 0 \tag{20d}$$

$$\mathbf{W} \succeq 0 \tag{20e}$$

$$\operatorname{rank}(\mathbf{W}) = 1. \tag{20f}$$

From problem (20a), it can be observed that the resulting objective function is linear, and all constraints are convex sets except the rank-one constraint. Following the SDP relaxation theory, if we drop the rank-one constraint, we can arrive at

$$\min_{\mathbf{W}} \quad P_A \operatorname{tr}(\mathbf{D}_A\mathbf{W}) + P_B \operatorname{tr}(\mathbf{D}_B\mathbf{W}) + \sigma^2 \operatorname{tr}(\mathbf{W}) \quad \text{s.t.} \quad (20b), (20c), (20d), (20e), \tag{21}$$

which is a convex SDP, and therefore it can be efficiently solved to obtain the global optimum by the available solvers, for example, CVX [13].

It should be noted that the two schemes presented in our paper can easily be extended to the case of multieavesdropper scenario directly. When multiple eavesdroppers are present in the system, the number of the secure constraint will be two times of the eavesdropper's number for the optimization problem and the SDP relaxation method can also be applied to solve the optimization problem.

Moreover, it also can be observed that when $\mathbf{\Sigma} = 0$, problem (11) can be degenerated into problem (19), which implies that the feasible set of problem (11) is a subset of the feasible set of problem (19) and $P_R^{*(\mathrm{BFA})} \le P_R^{*(\mathrm{BFO})}$.

5. Numerical Results

This section will provide some simulation results to validate the effectiveness of our proposed schemes and also compare the performance of the proposed schemes with other existing schemes. In the simulations, the numerical results are obtained by solving the relaxed convex optimization problems (12a) and (20a) by using CVX tools [14].

5.1. Effectiveness Discussion. In this subsection, we select an example to show the effectiveness of the proposed schemes. The number of antennas K is 4. The uniform linear array (ULA) channel model is adopted to keep the space between successive array elements half of the carrier wavelength,

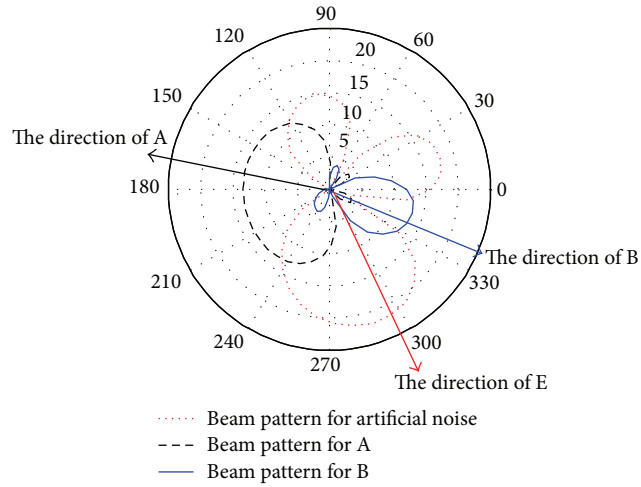


Figure 3: Optimal beam patterns of BFA.

where the channel vectors $\mathbf{h}_R$, $\mathbf{f}_R$, and $\mathbf{g}_R$ are generated in terms of the Vandermonde structure. The vector $V(\varphi) = [1, e^{j\theta}, \ldots, e^{j(K-1)\theta}]^T/\sqrt{K}$, where $\varphi \in [0^\circ, 360^\circ)$ and $\theta = -\pi \sin(\varphi\pi/180)$. The directions of A, B, and E are set to be 170°, 340°, and 299°, respectively. Therefore, $\mathbf{h}_R = V(170^\circ)$, $\mathbf{f}_R = V(340^\circ)$ and $\mathbf{g}_R = V(299^\circ)$.

By solving the problem in (12a) and (20a), the total transmit power consumed by the relay nodes in BFO and BFA is 20.6 dBm and 17 dBm, respectively, which apparently demonstrates that our proposed BFA consumes much less power than BFO.

To show more detail information on the beamforming vectors design, the beam patterns of $\mathbf{w}^* \operatorname{diag}(\mathbf{h}_R)$, $\mathbf{w}^* \operatorname{diag}(\mathbf{f}_R)$ and $\mathbf{\Sigma}^*$ for problem (11) are plotted in Figure 3 and the beam patterns of $\mathbf{w}^* \operatorname{diag}(\mathbf{h}_R)$ and $\mathbf{w}^* \operatorname{diag}(\mathbf{f}_R)$ for problem (19) are shown in Figure 4, respectively, where $\mathbf{w}^* \operatorname{diag}(\mathbf{h}_R)$ and $\mathbf{w}^* \operatorname{diag}(\mathbf{f}_R)$ actually are the beam pattern for A and B, respectively, and $\mathbf{\Sigma}^*$ is the beam pattern of the AN.

From Figure 3 and Figure 4, it can be seen that in the two proposed schemes, the obtained main power (information) of beam patterns $\mathbf{w}^* \operatorname{diag}(\mathbf{h}_R)$ and $\mathbf{w}^* \operatorname{diag}(\mathbf{f}_R)$ focuses towards A and B very well, respectively. From Figure 3 and Figure 4, it also can be observed that both $\mathbf{w}^* \operatorname{diag}(\mathbf{h}_R)$ and $\mathbf{w}^* \operatorname{diag}(\mathbf{f}_R)$ degrade sharply along the direction of E, which implies that, using our proposed schemes, the main power (information) can be focused towards to B and A with less leakage power towards E. So, the simulation results indicate that the major part of power for the signals is transmitted to the authorized users while only a little power of the signal is leaked to the eavesdropper.

Besides, from Figure 3, one can also see that $\mathbf{\Sigma}^*$ focus its main beam power (interference) towards E, and the artificial noise power towards A and B are relatively very low. As $\mathbf{\Sigma}^*$, in fact, represents the power consumption of the designed artificial noise at the relay nodes, it demonstrates that the designed artificial noise can greatly bring down the received SINR at E while only causing very limited impact on the received SINR of the two authorized users, A and B. By doing

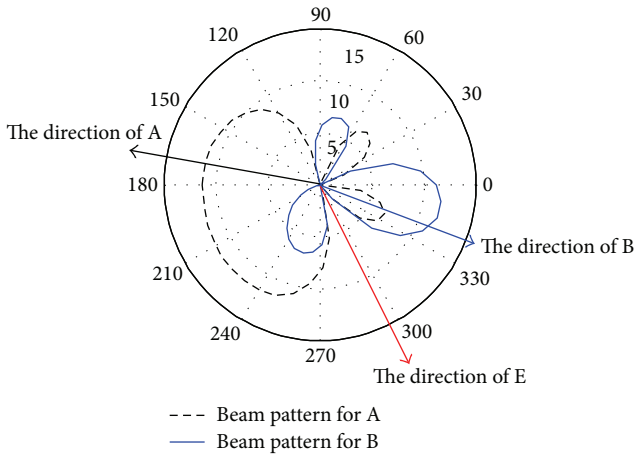


Figure 4: Optimal beam patterns of BFO.

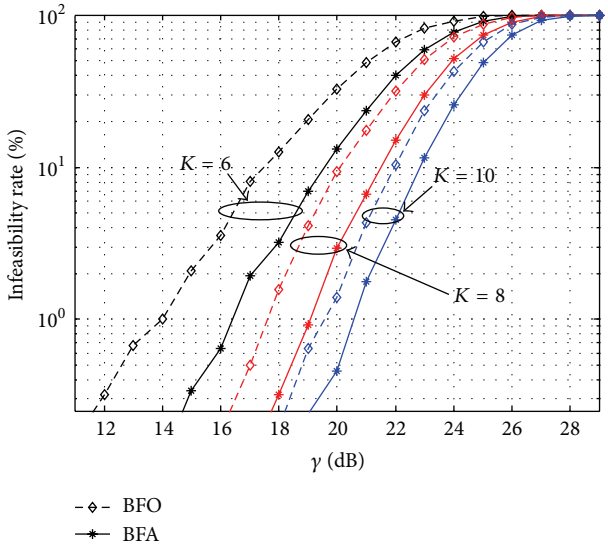


Figure 5: Comparison of the infeasibility rate.

so, secure two-way relay transmission can be achieved by using our proposed scheme.

5.2. Performance Comparison in terms of Infeasible Rate and Power Consumption. In this subsection, we compare our proposed BFA with BFO in terms of infeasibility rate and power consumption, where the infeasibility rate is defined as the percentage of infeasibility (%) of problems (12a) and (20a) out of 1000 simulations, which is used to evaluate the capability of the schemes in problem solving, and the power consumption is the total power consumed at the relay node. In the simulations, the channel vectors are generated as complex zeros-mean Gaussian random vectors.

Figure 5 plots the infeasibility rate versus γ when K are selected to be 6, 8, and 10, respectively. It can be observed that the infeasibility rate of BFA is always lower than BFO and infeasibility rates of both schemes decrease with the increase of the value of K. Figure 6 plots the total consumed

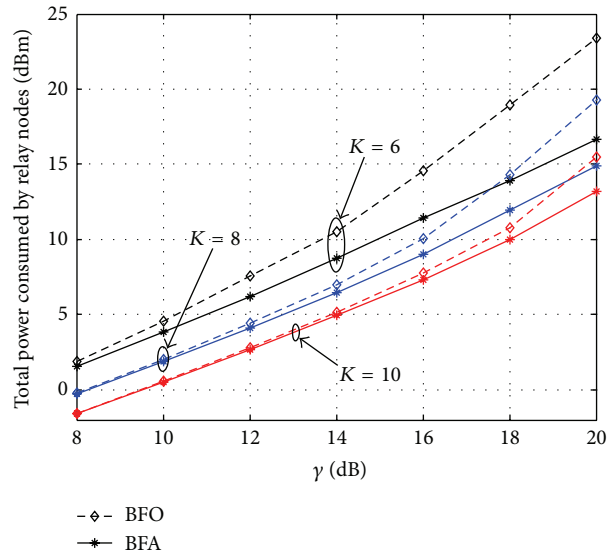


Figure 6: Comparison of the total consumed power.

power P_R by the relay node versus γ when both of the two proposed schemes are feasible. It can be seen that the power consumption of BFA is always less than that of BFO. Moreover, the power consumption gap between the two schemes becomes gradually larger with the increase of γ. It also shows that the total power consumed by the relay node decreases with the growth of the number of antennas. It therefore can be stated that more relay nodes could lead to low power consumption and low infeasibility rate, and by introducing optimally designed artificial noise, secure bemforming performance can be improved.

From the simulations presented above, it can be stated that BFA always has better performance than BFO both in energy saving and feasible rate.

6. Conclusions

This paper studied the relay beamforming design for multiantenna two-way relay networks in the presence of an eavesdropper. We presented two beamforming methods, that is, BFA and BFO. The received SINR at the receiver was used as the QoS measurement. We formulated optimization problems for the two methods to optimally design beamforming vectors and artificial noise vector to minimize the total energy consumption. SDP relaxation theory was used to solve the problems. Simulation results demonstrated the effectiveness of our proposed schemes and showed that BFA outperforms BFO in terms of high power efficiency and low infeasibility rate, which indicated that by jointly design the artificial noise and beamforming vector, the performance secure beamforming for two-way relay systems can be greatly improved.

Acknowledgment

This work was supported by the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University,

under Grant no. RCS2012ZT008, and by the National Natural Science Foundation of China under Grant nos. 61201203 and 61171064.

References

[1] W.-C. Liao, T.-H. Chang, W.-K. Ma, and C.-Y. Chi, "Joint transmit beamforming and artificial noise design for QoS discrimination in wireless downlink," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10)*, pp. 2562–2565, March 2010.

[2] P. K. Gopala, L. Lai, and H. El Gamal, "On the secrecy capacity of fading channels," *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4687–4698, 2008.

[3] A. L. Swindlehurst, "Fixed SINR solutions for the MIMO wiretap channel," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09)*, pp. 2437–2440, April 2009.

[4] A. Khisti and G. W. Wornell, "Secure transmission with multiple antennas I: the MISOME wiretap channel," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3088–3104, 2010.

[5] F. Oggier and B. Hassibi, "The secrecy capacity of the MIMO wiretap channel," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 4961–4972, 2011.

[6] S. Goel and R. Negi, "Guaranteeing secrecy using artificial noise," *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, pp. 2180–2189, 2008.

[7] W.-C. Liao, T.-H. Chang, W.-K. Ma, and C.-Y. Chi, "QoS-based transmit beamforming in the presence of eavesdroppers: an optimized artificial-aoise-aided approach," *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 1202–1216, 2011.

[8] N. Romero-Zurita, M. Ghogho, and D. McLernon, "Physical layer security of MIMO-OFDM systems by beamforming and artificial noise generation," *Physical Communication*, vol. 4, no. 4, pp. 313–321, 2011.

[9] Y. Yang, Q. Liang, W. K. Ma, J. H. Ge, and P. C. Ching, "Cooperative secure beamforming for AF relay networks with multiple eavesdroppers Physical layer security of MIMO-OFDM systems by beamforming and artificial noise generation," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 35–38, 2013.

[10] X. Y. Wang, K. Wang, and X. D. Zhang, "Secure relay beamforming with imperfect channel side information," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 2140–2155, 2013.

[11] H.-M. Wang, Q. Yin, and X.-G. Xia, "Improving the physical-layer security of wireless two-way relaying via analog network coding," in *Proceedings of the 54th Annual IEEE Global Telecommunications Conference (GLOBECOM '11)*, pp. 1–6, December 2011.

[12] D. D. Li, K. Xiong, G. Y. Du, and Z. D. Qiu, "Secure beamforming with artificial noise for two-way relay networks," *KSII Transactions on Internet and Information Systems*, vol. 7, no. 6, pp. 1418–1432, 2013.

[13] Z.-Q. Luo, W.-K. Ma, A. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 20–34, 2010.

[14] M. Grant and S. Boyd, "CVX: MATLAB software for disciplined, Convex Programming," June 2009, http://cvxr.com/.

[15] Y. Huang and D. P. Palomar, "Rank-constrained separable semidefinite programming with applications to optimal beamforming," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, pp. 664–678, 2010.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 380265, 10 pages
http://dx.doi.org/10.1155/2013/380265

Research Article

Peering Strategic Game Models for Interdependent ISPs in Content Centric Internet

Jia Zhao,[1] Jianfeng Guan,[2] Changqiao Xu,[2,3] Wei Su,[1] and Hongke Zhang[1]

[1] *National Engineering Laboratory for Next Generation Internet Interconnection Devices, Beijing Jiaotong University, Beijing 100044, China*
[2] *State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China*
[3] *Institute of Sensing Technology and Business, Beijing University of Posts and Telecommunications, Wuxi, Jiangsu, China*

Correspondence should be addressed to Jia Zhao; 11111004@bjtu.edu.cn

Received 22 September 2013; Accepted 27 October 2013

Academic Editors: W. Sun and J. Zhou

Emergent content-oriented networks prompt Internet service providers (ISPs) to evolve and take major responsibility for content delivery. Numerous content items and varying content popularities motivate interdependence between peering ISPs to elaborate their content caching and sharing strategies. In this paper, we propose the concept of peering for content exchange between interdependent ISPs in content centric Internet to minimize content delivery cost by a proper peering strategy. We model four peering strategic games to formulate four types of peering relationships between ISPs who are characterized by varying degrees of cooperative willingness from egoism to altruism and interconnected as profit-individuals or profit-coalition. Simulation results show the price of anarchy (PoA) and communication cost in the four games to validate that ISPs should decide their peering strategies by balancing intradomain content demand and interdomain peering relations for an optimal cost of content delivery.

1. Introduction

Tremendous volume of traffic from content-oriented services such as media streaming and file download motivates the evolution of Internet architecture for more efficient content delivery. With the emergence of new networking paradigms such as content-centric networking (CCN) [1], the design of future Internet trends towards the way to take content as a central entity. In such content-centric Internet, Internet service providers (ISPs) attach storage to their distributed network nodes (e.g., routers) for in-network caching and delivering content locally [2]. Such extended caching function prompts ISPs in the future Internet to evolve from traffic managers to content managers that will take the major responsibility for content delivery [3, 4].

Caching and delivering content by ISPs give rise to the question as to what kind of interrelationship ISPs can build to fulfill quality of service (QoS) for their intradomain content requesters. Unlike the traditional interconnection (through peering or transit contracts) to maintain global reachability in the current Internet, description of interrelations between ISPs in content centric Internet is more complex. This is so because achieving QoS of content delivery is a tough work that calls for interdependence between ISPs. Although an ISP can cache its intradomain popular content, there are numerous content items whose heavy tail of popularity distribution [5] and varying content popularity across different networks make it difficult for the ISP to satisfy intradomain content requests independently with its limited cache capacity. Accordingly, cooperative caching or inter-ISP content sharing is a more reasonable way to a win-win situation for ISPs [5]. Traditional ISP interconnection through traffic transit or peering should be supplemented with an interdependent relationship (i.e., peering for content) between ISPs so that the communication cost of content delivery is optimized.

A topology at autonomous system (AS) level consists of multiple ISPs with selfish profit utility and variety of complicated bilateral or multilateral relationships. So, it is difficult to let all ISPs converge to a consensus of cooperation.

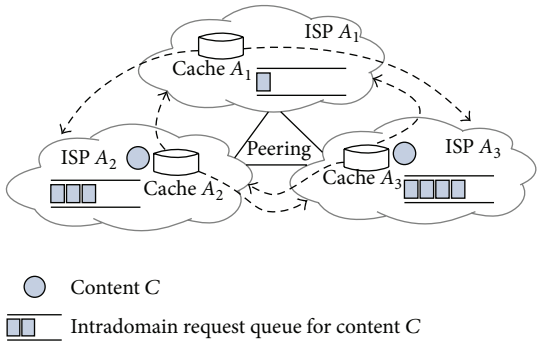


Figure 1: Interdependent ISPs A_1, A_2, and A_3 build peering relationships between each other through cache-to-cache content sharing. Request queues for content C within the three ISPs indicate intradomain popularity of content C. Both A_2 and A_3 cache the content C. A_1 does not cache C but can receive it from A_2 or A_3.

According to intradomain content popularity and profit goal for optimal communication cost (e.g., content delivery distance or latency and cache updating overhead), an ISP should utilize flexible and applicable caching and sharing strategies to build peering relationships with other ISPs. Peering ISPs can make full use of their content caching and sharing flexibility and commit themselves to deliver their intradomain requested content at low cost. For example, in Figure 1, requests for content C in ISP A_2 or A_3 are much more than the requests in A_1. A_2 and A_3 cache the popular content C to lower cross-ISP delivery cost, while A_1 can receive content transit from A_2 or A_3 to serve its less popular content demand without caching the content. If the peering is based on reciprocity, both A_2 and A_3 have the right to decide how they will contribute according to A_1's contribution. If A_1, A_2 and A_3 are three sub-ISPs organized as coalition by a larger ISP, the dominant ISP can coordinate the caches and adjust sharing mechanism for coalitional profit maximization. Variety of peering strategies is due to two aspects: (i) the degree of an ISP's cooperative willingness varying from egoism to altruism, (ii) profit relationships such as profit-individual ISPs or a profit-coalition of ISPs.

In this paper, we model four peering strategic games with interdependent ISPs as participants, each of which decides its caching strategy (i.e., whether to cache a content item) and sharing strategy (i.e., how many interdomain content requests to respond to) to minimize the communication cost of content delivery. In the egoistic game, each ISP can decide its own caching strategy and respond to all its peering ISP's content requests. In the tit-for-tat game, an ISP decides its sharing strategy according to its peer's sharing strategy so as to achieve the reciprocity. In the altruistic game, two peeing ISPs embody a common profit goal in their respective cost function to save cost of content exchange between each other. In the cooperative game, ISPs cooperate as a coalition to gain more profit than noncoalitional ISPs. These four types of interdependent patterns are modeled to describe possible relationships between peering ISPs in content-centric Internet. In subsequent sections, we base our problem analysis and propositions in this paper on the circumstance where ISPs are equipped with the content-centric in-network cache and responsible for content delivery. Some conclusions are also well suited to ISP-operated content delivery networks (CDNs) and Internet economics (e.g., peering or transit) happening between ISPs in the current Internet. We simulate the four games based on an AS-level topology representing the interconnection between some large networks of the Chinaese Internet. Performance evaluation shows that comparatively altruistic games (altruism and cooperation) manifest agreeable properties (low PoA and cost) under high frequency of content requests and that sharing strategy also affects delivery cost in the tit-for-Tat case. This means that ISPs should decide their peering strategies by balancing intradomain content demand and interdomain peering relations for an optimal cost of content delivery.

The main contributions of this paper includes the following: (1) we propose the concept of peering for content exchange between interdependent ISPs in the content centric Internet to minimize their respective content delivery cost by proper peering strategies; (2) we use four game models to formulate and describe the interdependence between peering ISPs; (3) we study the peering ISPs' communication cost of content delivery and each game's divergence from social optimum to validate that ISPs should decide their peering strategies by balancing intradomain content demand and interdomain peering relation.

The remainder of the paper is organized as follow. Section 2 introduces related work. In Section 3, we detail the peering strategy and give the cost utility expression. In Section 4, we model the four peering strategic games. Section 5 makes some game-theoretic analysis. Simulation results are shown in Section 6. Section 7 discusses some involving issues. Section 8 concludes the paper.

2. Related Work

Peering strategy for inter-ISP content sharing is essential to both the current and the future Internet. With the Internet evolution from hierarchical to flat structure, settlement-free peering prevails among ISPs and prompts reciprocal traffic exchange between interconnected ISPs. Widespread video content distribution and its tremendous traffic volume motivate the peering between CDN providers and ISPs. In the paper [6], the authors investigate the content peering between ISPs and large content providers and develop a model to probe into the interaction between different types of ISPs. The authors demonstrate that the situation of asymmetric traffic from and into the CDN providers may not benefit the ISPs and violate the basic peering principle of reciprocity.

Content distributed systems such as CDN and P2P pose significant challenge on the traffic peering or transit relationship between ISPs [7, 8]. ISPs deploy intradomain content cache so as to decrease inter-ISP traffic. In the paper [5], the author formulates two game models to illustrate that ISPs can cooperate to improve selfish interests with cooperative caching strategies. For the efficiency of content delivery such

as P2P streaming, collaborative caching policy of peering ISPs is also proposed to save content-receiving cost [9].

ISP interconnectivity in the current Internet experiences change and adjustment. In [10], the authors build an agent-based network model to study interdomain ecosystem and demonstrate evolutionary Internet transition from a transit hierarchy to a peering mesh. Varying peering or transit strategies are proposed and implemented according to ISP's selfish profit utility and relationship with each other, such as the work in [11].

In the current Internet, ISPs and content distributed system operators share a common profit goal to respond to content requests as locally as possible. Our proposed peering by cooperative caching strategy can satisfy intradomain content demand at optimal content-receiving cost and alleviate the situation of asymmetric profits in [6]. Also differing from the work in [5, 9], due to an ISP's extended duty to deliver content, our proposed peering strategic game models gain equilibrium solution to the optimization of an ISP's content delivery utility (i.e., intradomain contentreceiving cost measured by distance or latency). Additionally, intended to elaborate peering strategy for inter-ISP content sharing, our work also supplements, diversifies and evolves the peering patterns of ISPs.

3. Preliminaries

3.1. From Interconnection to Interdependence. Based on the principle of global reachability in the current Internet, interconnected ISPs commit themselves to agreements or contracts to deal with peering traffic or transit traffic. As the Internet architecture evolves to be content-oriented, ISPs will actively participate in the content delivery more than just traffic management. ISPs will act as CDN operators or cache content in their distributed network nodes so as to satisfy intradomain content demand. Content centric Internet motivates the peering relationship changed from interconnection to interdependence for more efficient content delivery. Peering ISPs can take advantage of the cross-ISP content delivery (if the content requester is geographically far away from its intradomain cache but nearby the content source of another ISP), and the huge global cache capacity can favor every ISP and enhance its respective caching flexibility (i.e., caching the popular by itself and requesting the less popular from other peers). Peering for content calls for solutions to select proper caching and sharing strategy, which is the focus of our work.

3.2. Peering Strategies. Let $\Phi = \{A_1, \ldots, A_m\}$ represent a set of peering ISPs. For all $A_i \in \Phi$, ISP A_i has a total cache capacity $C(A_i)$. Here we consider the distributed cache of an ISP as a whole, because every node (without generality, regarded as a router) with storage for content caching function belongs to the ISP's network whose cache function is independent of the other peering ISPs.

$\Gamma = \{c_0, c_1, \ldots c_n\}$ represents the content item set, $S(c_j)$ is the size of content c_j (also applicable to the case of total size of multiple copies of c_j within an ISP), and $P(c_j)$ denotes the global popularity of content c_j. Content popularity can reflect the request frequency for a content item. Yet, in the real life of applications such as P2P systems, content popularity distribution is of the heavy tail [5]. This means that the less popular content items are too numerous to omit. We conform to this pattern and set $P(c_j)$ to follow a Zipf-like distribution [12] in our simulation. With this content popularity over multiple ISPs, we can assign ISP A_i a part of $P(c_j)$ to denote the intradomain content popularity $p_{A_i}^{(c_j)}$ which is subject to $\sum_{A_i} p_{A_i}^{(c_j)} = P(c_j)$.

Given the previous denotations, for the peering relationship between ISPs, each ISP can independently decide two peering strategies as follows.

(i) Caching strategy: $I_{A_i}^{(c_j)}$ denotes the caching strategic function getting value 1 if ISP A_i caches the content c_j and value 0 otherwise. The decision $I_{A_i}^{(c_j)}$ of ISP A_i is influenced by both the cache capacity constraint of $\sum_{c_j} I_{A_i}^{(c_j)} \cdot S(c_j) \leq C(A_i)$ and the intradomain content popularity $p_{A_i}^{(c_j)}$.

(ii) Sharing strategy: $X_{ik}^{(c_j)} \in [0, 1]$ is a fraction of $p_{A_k}^{(c_j)}$ and denotes how much the ISP A_i is sharing to respond to ISP A_k's requests for content c_j. Because of the contradiction between the limited cache capacity and the content popularity in heavy tail distribution, ISPs receive the shared content from each other for their respective intradomain less popular requests and build an interdependent relationship on this sharing mechanism. How much sharing depends on, however, the patterns of peering is we will discuss in the next section. After all, the sharing is also the basis of our proposed peering for the content centric Internet.

3.3. QoS Metric and Utility Function. We employ game models to study the peering patterns. Every peering ISP, as a participant in a game, has a utility function to optimize its own profit objective. In the content centric Internet, peering ISPs are supposed to deliver content efficiently for their respective intradomain requesters. So, we set the objective to optimize a QoS metric of content delivery.

To define the QoS metric, we have to select a metric of cost to deliver content. Let $D_{A_i}^{(c_j)}$ denote the cost of ISP A_i to deliver the content c_j, and the QoS metric of ISP A_i can be formulated as follows:

$$U_{A_i} = \sum_{c_j} p_{A_i}^{(c_j)} \cdot D_{A_i}^{(c_j)}, \tag{1}$$

where U_{A_i} is the total cost as the utility. $D_{A_i}^{(c_j)}$ depends on the strategies of peering ISPs. An ISP decides its caching and sharing strategies to minimize its QoS metric. In application, $D_{A_i}^{(c_j)}$ has its meaning at distinct grain-levels as follows.

(i) Coarse-grained: $D_{A_i}^{(c_j)}$ denotes the AS-level hop-count distance from the responder ISP to the requester ISP A_i.

(ii) Fine-grained: $D_{A_i}^{(c_j)}$ denotes the average delay or distance that the requester ISP A_i cost to receive content c_j within itself or from the responder ISPs.

Formula (1) with a simple form can calculate the QoS metric with the fine-grained denotation of $D_{A_i}^{(c_j)}$, but it cannot show the ISP's peering strategies $I_{A_i}^{(c_j)}$ and $X_{ik}^{(c_j)}$ obviously. Accordingly, we use a composition of the two definitions of $D_{A_i}^{(c_j)}$. Let $d_{A_i}^{(c_j)}$ denote the content delivery cost (intradomain content-delivering distance or delay) within an ISP A_i, $q_{A_i}^{(c_j)}$ denote the cross-ISP communication cost to receive content from a neighboring ISP (one AS-hop away), and $r_{A_i}^{(c_j)}$ denote the cost to receive content from a remote ISP (multi-AS-hop away) or directly from content providers (CPs). Then the utility of ISP A_i can be expressed as follows:

$$U_{A_i} = \sum_{c_j} p_{A_i}^{(c_j)} \cdot \left(d_{A_i}^{(c_j)} \cdot I_{A_i}^{(c_j)} + \left(1 - I_{A_i}^{(c_j)}\right) \cdot q_{A_i}^{(c_j)} \cdot \sum_{k \neq i} I_{A_k}^{(c_j)} \cdot X_{ki}^{(c_j)} + r_{A_i}^{(c_j)} \cdot \left(1 - I_{A_i}^{(c_j)}\right) \cdot \left(1 - \sum_{k \neq i} I_{A_k}^{(c_j)} \cdot X_{ki}^{(c_j)}\right)\right), \tag{2}$$

with constraint of $\sum_{k \neq i} I_{A_k}^{(c_j)} \cdot X_{ki}^{(c_j)} \leq 1$. The formulating part as $\sum_{k \neq i} I_{A_k}^{(c_j)} \cdot X_{ki}^{(c_j)}$ means the condition that the peering ISPs cache the content c_j and contribute to ISP A_i with their respective sharing strategies. $q_{A_i}^{(c_j)}$ is normally greater than $d_{A_i}^{(c_j)}$, but when the content requester is geographically far away from its intradomain cache but nearby the content source of another ISP, we have $q_{A_i}^{(c_j)} < d_{A_i}^{(c_j)}$. In Section 4, we use the utility of form like (2) to model the peering strategic games. In the simulation, for simplicity without generality, we use the AS-hop-count to evaluate the content delivery cost.

3.4. Number of Game Participants. Peering for traffic in the current Internet is traditionally built between two same tier level ISPs (though there is also the peering between two different levels of ISPs). In the content centric Internet, ISPs have various patterns of interdependent relationship (which will be seen in the next section). Because of the variety of peering for content and an ISP's dependent choice of peering strategy, the complex games may exist among multiple participants that play with each other directly or indirectly. Yet, some of these complex games (e.g., egoism, tit-for-tat, and altruism in the next section) can be generated by a series of noncooperative games between a pair of ISPs. So, we will use the basic two-player game to describe these peering patterns. There are still, however, some peering patterns (e.g., cooperation) involving multiple participants, and we employ cooperative or noncooperative games of multiple players to deal with these cases.

4. Interdependence Patterns

In this section, interdependence patterns of ISPs are formulated as the peering strategic games. An ISP caches popular content for in-network content delivery, while it can also request content from its peer ISP if it does not cache the content, or this cross-ISP content delivery unfolds efficiency (e.g., content source belonging to ISP A_1 is geographically located nearby the requester in ISP A_2).

4.1. Egoism. This peering is built between two profit-individual ISPs in content centric Internet. Either ISP is egoistic to optimize its own content delivery utility. A noncooperative game [13] with ISP A_1 and A_2 as two players can be modeled as follows

A_1 decides $I_{A_1}^{(c_j)}$ to minimize:

$$U_{A_1} = \sum_{c_j} p_{A_1}^{(c_j)} \cdot \left(d_{A_1}^{(c_j)} \cdot I_{A_1}^{(c_j)} + q_{A_1}^{(c_j)} \cdot \left(1 - I_{A_1}^{(c_j)}\right) \cdot I_{A_2}^{(c_j)} + r_{A_1}^{(c_j)} \cdot \left(1 - I_{A_1}^{(c_j)}\right) \cdot \left(1 - I_{A_2}^{(c_j)}\right)\right), \tag{3}$$

subject to $\sum_{c_j} I_{A_1}^{(c_j)} \cdot S(c_j) \leq C(A_1)$.

A_2 decides $I_{A_2}^{(c_j)}$ to minimize

$$U_{A_2} = \sum_{c_j} p_{A_2}^{(c_j)} \cdot \left(d_{A_2}^{(c_j)} \cdot I_{A_2}^{(c_j)} + q_{A_2}^{(c_j)} \cdot \left(1 - I_{A_2}^{(c_j)}\right) \cdot I_{A_1}^{(c_j)} + r_{A_2}^{(c_j)} \cdot \left(1 - I_{A_1}^{(c_j)}\right) \cdot \left(1 - I_{A_2}^{(c_j)}\right)\right), \tag{4}$$

subject to $\sum_{c_j} I_{A_2}^{(c_j)} \cdot S(c_j) \leq C(A_2)$.

Both the ISPs decide their caching strategies selfishly and independently, and their profit goals to minimize their respective cost of content delivery converge them to a strategic equilibrium (i.e., the best caching strategy, whose existence will be proved in next section). To reduce latency or distance of content delivery, egoistic ISPs have to respond to content requests of each other, since there may be $q_{A_i}^{(c_j)} < d_{A_i}^{(c_j)}$. The free exchange of content (with sharing strategy as $X_{12}^{(c_j)} = X_{21}^{(c_j)} = 1$) in this game is similar to the settlement-free agreement between peering ISPs in the current Internet to exchange traffic by transit-free. Hence, the egoistic peering for content can be built between two ISPs at a same tier level because of similar number of clients and cache capacities.

4.2. Tit-for-Tat. In the "tit-for-tat," ISPs A_1 and A_2, A_1 will decide its sharing strategy $X_{12}^{(c_j)}$ according to the A_2's decision

$X_{21}^{(c_j)}$ and vice versa. A noncooperative game with ISPs A_1 and A_2 as two players can be modeled as follows:

A_1 decides $I_{A_1}^{(c_j)}$ and $f_1(X_{21}^{(c_j)})$ to minimize

$$U_{A_1} = \sum_{c_j} p_{A_1}^{(c_j)} \cdot \left(d_{A_1}^{(c_j)} \cdot I_{A_1}^{(c_j)} + q_{A_1}^{(c_j)} \cdot f_2\left(X_{12}^{(c_j)}\right) \cdot \left(1 - I_{A_1}^{(c_j)}\right) \cdot I_{A_2}^{(c_j)} + r_{A_1}^{(c_j)} \cdot \left(1 - I_{A_1}^{(c_j)}\right) \cdot \left(1 - X_{21}^{(c_j)} \cdot I_{A_2}^{(c_j)}\right)\right), \tag{5}$$

subject to $\sum_{c_j} I_{A_1}^{(c_j)} \cdot S(c_j) \le C(A_1)$, $0 \le f_1(X_{21}^{(c_j)}) \le 1$.

A_2 decides $I_{A_2}^{(c_j)}$ and $f_2(X_{12}^{(c_j)})$ to minimize

$$U_{A_2} = \sum_{c_j} p_{A_2}^{(c_j)} \cdot \left(d_{A_2}^{(c_j)} \cdot I_{A_2}^{(c_j)} + q_{A_2}^{(c_j)} \cdot f_1\left(X_{21}^{(c_j)}\right) \cdot \left(1 - I_{A_2}^{(c_j)}\right) \cdot I_{A_1}^{(c_j)} + r_{A_2}^{(c_j)} \cdot \left(1 - I_{A_2}^{(c_j)}\right) \cdot \left(1 - X_{12}^{(c_j)} \cdot I_{A_1}^{(c_j)}\right)\right), \tag{6}$$

subject to $\sum_{c_j} I_{A_2}^{(c_j)} \cdot S(c_j) \le C(A_2)$, $0 \le f_2(X_{12}^{(c_j)}) \le 1$.

We use function $f_1(X_{21}^{(c_j)})$ to express $X_{12}^{(c_j)}$ not only because we want to show the reciprocity in "tit-for-tat," but also we will analyze the type and influence of sharing strategies in the simulations. An ISP in this game can partially respond to content requests from the other and build the peering on the base of reciprocity. In the content sharing, the content requester ISP only optimizes its utility to reduce content delivery delay or distance for quality of service, while the responder ISP has to take the expense of content transit. Hence, differing from free exchange in the egoism, "tit-for-tat" allows an ISP to relate its own sharing strategies with the other ISP's contribution. An example is the equational exchange formulated as $X_{12}^{(c_j)} = f_1(X_{21}^{(c_j)}) = f_2(X_{12}^{(c_j)}) = X_{21}^{(c_j)}$.

4.3. Altruism. Altruistic ISPs can join in a coalition. The utility of a coalition member embodies not only the selfish utility as formula (3) or (4) but also a coalition cost as follows:

$$U_{A_1,A_2} = \sum_{c_j} \left(p_{A_1}^{(c_j)} + p_{A_2}^{(c_j)}\right) \cdot \left(D \cdot \left(I_{A_1}^{(c_j)} \cup I_{A_2}^{(c_j)}\right) + D' \cdot \left(\overline{I_{A_1}^{(c_j)} \cup I_{A_2}^{(c_j)}}\right)\right), \tag{7}$$

where D denotes the content-receiving cost within the coalition and D' denotes the cost to receive content from an ISP out of the coalition. A game between ISP A_1 and A_2 is modeled as follows:

A_1 decides $I_{A_1}^{(c_j)}$ to solve the problem as follows:

$$\text{Min} \quad U_{A_1} + U_{A_1,A_2}, \tag{8}$$

subject to $\sum_{c_j} I_{A_1}^{(c_j)} \cdot S(c_j) \le C(A_1)$.

A_2 decides $I_{A_2}^{(c_j)}$ to solve the problem as follows:

$$\text{Min} \quad U_{A_2} + U_{A_1,A_2}, \tag{9}$$

subject to $\sum_{c_j} I_{A_2}^{(c_j)} \cdot S(c_j) \le C(A_2)$.

The altruistic peering is applicable to the relation between multiple ISPs that agree to be a coalition and share the cost of intracoalitional content delivery. Although the ISPs serve their respective clients, this peering agreement allows the members of the coalition to pursue an optimal coalitional utility without losing their selfish profit. To understand the necessity of this altruism case, we see a similar example where two public peering ISPs buy bandwidth from Internet exchange point (IXP) operators or build private link between each other.

4.4. Cooperation. Altruistic coalition in Section 4.3 is a noncooperative game between two selfish ISPs although they embody a coalition cost utility in their own utility functions. In this subsection, we propose another type of coalition that can be modeled by a cooperative game [14]. A cooperative game differs from a noncooperative game in that participants in the coalition pursue a higher profit than situation where they do not participate in the coalition or cooperate with other ones out of the coalition.

Proposition 1. *Two ISPs A_1 and A_2 with same content demand ($p_{A_1}^{(c_j)} = p_{A_2}^{(c_j)}$) can cooperate to form a stable coalition to save content delivery cost.*

Proof. A_1, A_2, and A_3 are three ISPs. Let $\{c_1, c_2, c_3\}$ denote the set of content items. A_1 caches the content c_1, A_2 caches the content c_2, and A_3 caches the content c_3. The advantage of the coalition lies in that cross-ISP content-receiving cost D_1 equals intra-ISP cost D_0 within the coalition. The three ISPs have same content demand ($p_{A_1}^{(c_j)} = p_{A_2}^{(c_j)} = p_{A_3}^{(c_j)} = 1/3$). If A_1 and A_2 cooperate to be a coalition $\{A_1, A_2\}$, a cooperative game is formed as follows.

According to formula (1), we get cost utility of A_1 in the coalition as

$$\begin{aligned} U_{A_1}^{\{A_1,A_2\}} &= \sum_{c_j} p_{A_1}^{(c_j)} \cdot \left(D_0 \cdot I_{A_1}^{(c_j)} + D_0 \cdot \left(1 - I_{A_1}^{(c_j)}\right) \cdot I_{A_2}^{(c_j)} + D_1 \cdot \left(1 - I_{A_1}^{(c_j)}\right) \cdot \left(1 - I_{A_2}^{(c_j)}\right) \cdot I_{A_3}^{(c_j)}\right) \\ &= \frac{2}{3} \cdot D_0 + \frac{1}{3} \cdot D_1. \end{aligned} \tag{10}$$

If the two ISPs do not cooperate, the utility of A_1 is as follows.

$$\begin{aligned} U_{A_1}^{\{A_1\}} &= \sum_{c_j} p_{A_1}^{(c_j)} \cdot \left(D_0 \cdot I_{A_1}^{(c_j)} + D_1 \cdot \left(1 - I_{A_1}^{(c_j)}\right) \cdot I_{A_2}^{(c_j)} + D_1 \cdot \left(1 - I_{A_1}^{(c_j)}\right) \cdot \left(1 - I_{A_2}^{(c_j)}\right) \cdot I_{A_3}^{(c_j)}\right) \\ &= \frac{1}{3} \cdot D_0 + \frac{2}{3} \cdot D_1. \end{aligned} \tag{11}$$

Because $D_1 > D_0$, we have $U_{A_2}^{\{A_1,A_2\}} = U_{A_1}^{\{A_1,A_2\}} < U_{A_2}^{\{A_2\}} = U_{A_1}^{\{A_1\}}$. Either A_1 or A_2 in other coalitions such as $\{A_1, A_3\}$ or $\{A_2, A_3\}$ will not gain more profit. So, A_1 and A_2 can cooperate to be a stable coalition for cost saving. □

An ISP may be in charge of multiple ASs and arrange content-caching to organize the ASs as a profit-coalition. Let $p_{\mathrm{AS}_i}^{(c_j)}$ denote the content demand of c_j within AS_i and $D_{\mathrm{AS}_i}^{(c_j)}$ denote AS_i's cost to deliver content c_j. The dominant ISP assigns each AS a caching strategy to optimize coalitional content delivery cost by solving the following problem:

$$\text{Min} \quad U_{\text{Coalition}} = \sum_{c_j} \sum_{\mathrm{AS}_i} p_{\mathrm{AS}_i}^{(c_j)} \cdot D_{\mathrm{AS}_i}^{(c_j)}. \tag{12}$$

In this optimization, ASs in the coalition cooperate to achieve a common profit goal.

4.5. Global Optimum. Global optimum is an ideal state in which all the ISPs decide their caching strategy by optimizing a global profit utility together. Let $D_{A_i}(x)$ denote the ISP A_i's cost to deliver content x and $p_{A_i}(x)$ denote the content demand (i.e., content popular density) of content x within ISP A_i. The global popular density of content x is expressed as $p(x) = \sum_{A_i} p_{A_i}(x)$. Because $p(x)$ follows a Zipf distribution, we have $\int_{c_0}^{c_n} p(x) \cdot \mathrm{d}x = 1$. Here, c_o and c_n, respectively, denote the most and the least popular content. The utility of ISP A_i is expressed as $U_{A_i}(x) = p_{A_i}(x) \cdot D_{A_i}(x)$. ISPs decide their caching strategies by solving the problem as follows:

$$\text{Min} \quad U_{\text{Global}} = \int_{c_0}^{c_n} \left(\sum_{A_i} U_{A_i}(x) \right) \cdot \mathrm{d}x, \tag{13}$$

subject to $\int_{c_0}^{c_n} I_{A_i}(x) \cdot S(x) \cdot \mathrm{d}x \le C(A_i)$.

In practice, the global AS-level topology has a multitude of ASs operated by multiple ISPs that pursue their optimal selfish profit. Although the global optimum exists, the caching strategy at the global optimum is not adopted by ISPs. Yet, if an equilibrium solution exists in a peering strategic game, the global optimum can be used to evaluate different game's optimum cost divergences from the social optimum so as to maintain a global efficiency of content delivery.

5. Game-Theoretic Analysis

We have built the game models to study the interdependence between peering ISPs, and some following questions arise: (1) whether the equilibrium solutions exist in these games? (2) How can the peering ISPs as participants converge to the equilibrium in the actual game process? (3) What properties can be used to evaluate the equilibrium. This section will respond to these questions.

5.1. Existence of Equilibria. Existence of the equilibrium point in the game of egoism case is firstly investigated with the following proposition.

Proposition 2. *A Nash equilibrium solution exists in the noncooperative game with two egoistic ISPs as participants.*

Proof. Formulas (3) and (4) are the utility functions of two egoistic ISPs A_1 and A_2. Let two sets $I_1 = \{I_{A_1}^{(c_0)}, \ldots, I_{A_1}^{(c_n)}\}$ and $I_2 = \{I_{A_2}^{(c_0)}, \ldots, I_{A_2}^{(c_n)}\}$ represent their caching strategy sets. The utility functions can be rewritten as $U_{A_1}(I_1, I_2)$ and $U_{A_2}(I_1, I_2)$, and both their values depend on the two strategy set I_1 and I_2. From the formulas (3) and (4) we see that $U_{A_1}(I_1, I_2)$ is convex and continuous on the strategy sets I_1 of ISP A_1 and so is the function $U_{A_2}(I_1, I_2)$ on I_2 of ISP A_2. Since the value of any caching strategy $I_{A_i}^{(c_j)}$ belongs to the 2-element field $\{0, 1\}$, the caching strategic vector spaces I_1 and I_2 are both closed spaces. ISPs A_1 and A_2 independently decide their own strategies I_1 and I_2 to minimize their respective utility functions. Hence, if function $U_{A_1}(I_1, I_2)$ gets the minimal value at a solution I_1^* and $U_{A_2}(I_1, I_2)$ gets the minimal value at I_2^*, we have the inequations $U_{A_1}(I_1^*, I_2^*) \le U_{A_1}(I_1, I_2^*)$ and $U_{A_2}(I_1^*, I_2^*) \le U_{A_2}(I_1^*, I_2)$. When ISP A_1 chooses I_1^* and ISP A_2 chooses I_2^*, they will not deviate from this equilibrium solution, because they cannot gain more profit with other strategies. Hence, (I_1^*, I_2^*) is a Nash equilibrium point. □

In the tit-for-tat case, the sharing strategies $X_{ik}^{(c_j)} \in [0, 1]$ of ISP A_i also generate a closed space. So, following a similar way, the existence of equilibria in tit-for-tat and altruism can be proved. As for the cooperation case, Proposition 1 has proved that a stable coalition exists in the cooperative game.

5.2. Dynamic Convergence. After building the peering, an ISP commits itself to respond to content requests from intradomain or from its peers. During a period of time, the peering ISPs can dynamically converge to equilibrium of their peering strategic game. We can detail the dynamic convergence by an example of the egoistic game between two ISPs. In this noncooperative game, ISPs A_1 and A_2 take several rounds to converge to a strategic equilibrium, that is, the optimal content delivery cost which they will not deviate from by choosing other caching or sharing strategies. In each round, according to the strategic decision of ISP A_2, ISP A_1 will react and adjust with an optimal strategy (of this round) by updating its cache. Conversely, ISP A_2 will react to strategic change of ISP A_1 in the next round and make an optimal adjustment. Although this noncooperative game alternates the two ISPs to make strategic change, it will converge to the equilibrium after several rounds, since the strategies of both arrive at the optimal one and do not need to change again. This dynamic convergence is also the base of algorithm design for simulating the game process.

5.3. Divergence from Social Optimum. Evaluating the equilibrium of a game may go separate ways since we view the optimal solution of the game from different perspectives. Now we consider a peering strategic game with multiple ISPs as participants. From the point of view of every single ISP, the equilibrium solution is optimal because it cannot gain a lower cost of content delivery by choosing other peering strategies. However, from a global point of view, a highly evaluated cost at the equilibrium of a game should not diverge too much from the social optimum, which we have defined in Section 4. To study the game's divergence from the social

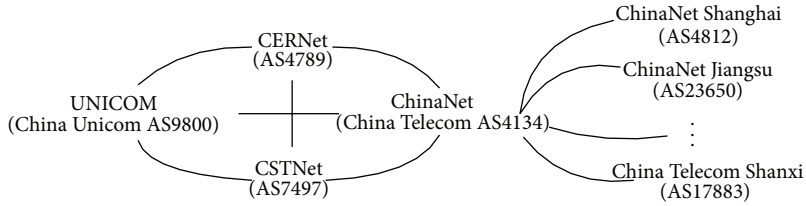


Figure 2: Connection of some large networks in the Chinese Internet.

optimum, we use the price of anarchy (PoA) [15] to measure the inefficiency of the decentralized optimization. Suppose that n ISPs participate in the cost optimization. Let $I_i = \{I_{A_i}^{(c_0)}, \ldots, I_{A_i}^{(c_l)}\}$ represent the caching strategy of ISP A_i, and $\mathrm{X}_i = \{X_{ik}^{(c_0)}, \ldots, X_{ik}^{(c_l)}\}$ represent the sharing strategy. For a noncooperative game, $U_{A_i}(I_1, \ldots, I_n, \mathrm{X}_1, \ldots, \mathrm{X}_n)$ denotes the cost of ISP A_i. For the global optimization, U_{Global} denotes the total cost of n ISPs at the global optimum. PoA is expressed as follows:

$$\mathrm{PoA} = \sup_{i,j} \frac{\sum_{c_j} \sum_i U_{A_i}(I_1, \ldots, I_n, \mathrm{X}_1, \ldots, \mathrm{X}_n)}{U_{\mathrm{Global}}}. \quad (14)$$

In the numerical results, PoA of the four peering games will be given.

6. Numerical Results

To evaluate performance of interdependent ISPs in the proposed peering strategic game models, we simulate the games on a toy system whose AS-level peering topology represents the interconnection between some large networks of the Chinaese Internet [16] as shown in Figure 2. The topology has totally 32 autonomous systems (ASs) including two academic networks, CERNet and CSTNet, two commercial networks, ChinaNet and UNICOM, and 28 local ISPs of provincial and municipal access networks connected to the ChinaNet Backbone. Based on this topology, we study global-optimum-divergence and communication cost changed with the caching and sharing strategies of ISPs in different game models and explore the proper peering strategy to balance intra-ISP content demand and inter-ISP relation for cost saving. According to the scale of networks (ChinaNet has 100 million broadband Internet access customers [17]; UNICOM has 30 million broadband subscribers [18]; CERNet has 20 million end users [19]; CSTNet has 1 million end users [20]), we assign ChinaNet a cache capacity of 1, UNICOM a capacity of 0.3, CERNet a capacity of 0.2, and CSTNet a capacity of 0.01. We also assign each of the 28 access ISPs a capacity of 0.005.

6.1. Price of Anarchy. Figure 3 shows the popularity of content that follows Zipf-like distribution in [12] with a default parameter $\alpha = 0.6$. Let $p_{A_i}^{(c_j)}$ denote the percentage of requests for c_j within ISP A_i. For each ISP A_i, we select $p_{A_i}^{(c_j)}$ from the ranked content items in Figure 3 and let the average content

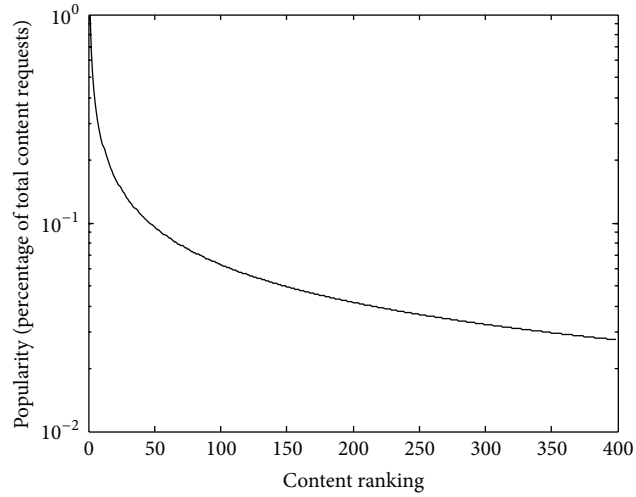


Figure 3: Content popularity in a Zipf-like distribution.

popularity of selected content items vary in the range $0, 1$. Then we study the PoA change with content popularity.

We deploy the four peering strategic games on the topology of 4 ASs (CERNet, CSTNet, ChinaNet, and UNICOM) in Figure 2 and let U^* denote the total content delivery cost of the 4 ASs at the equilibrium of each game. U_{Global} denotes the total cost at the global optimum solution as formula (13). Price of anarchy (PoA) [15] is expressed as $\mathrm{PoA} = U^*/U_{\mathrm{Global}}$. Value of PoA indicates how far a game's equilibrium solution deviates from the social optimum solution. We evaluate PoA because we want to know whether a near-social-optimal peering exists in the four games. Figure 4 shows the four games' PoA values changed with popularity of our selected content sets. The four games' PoA values vary in a range from nearly 1 to 1.28. The tit-for-tat model's PoA has a varying range length up to nearly 0.2, while the varying range lengths in the other three models are all less than 0.1. Both egoism and tit-for-tat have a broadly increasing figure patterns. Yet, tit-for-tat has a much larger PoA than egoism, since tit-for-tat makes a partial rather than total content contribution and limits the sharing relationship between ISPs. When the content popularity is less than 0.5, the egoism case is the near-optimal peering model. When there is too much content demand with popularity more than 0.6, the relatively altruistic peering cases (altruism and cooperation) outperform the relatively selfish case (egoism and Tit-for-tat) with respect to the divergence from global optimum. The cooperation case shows a roughly decreasing figure

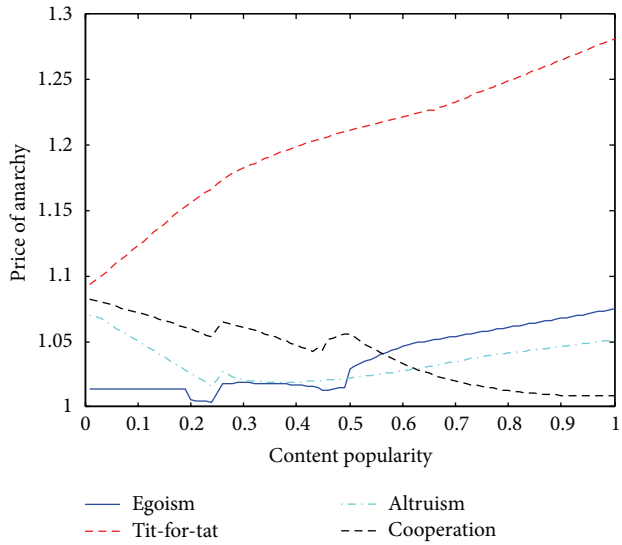


Figure 4: Price of anarchy versus content popularity.

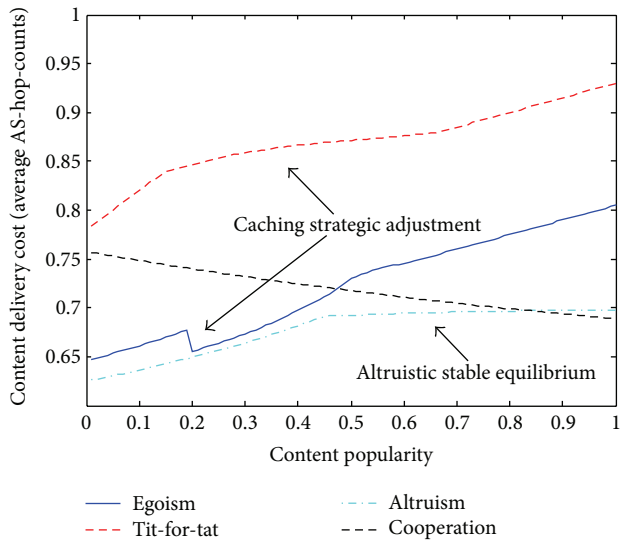


Figure 5: Content delivery cost versus content popularity.

pattern and has near-optimal PoA values in high popularity range.

6.2. Content Delivery Cost. We use average AS-hop-count as the metric of the communication cost to receive content within the AS or from other ASs. As shown in Figure 5, costs of all the four models are less than one AS-hop-count due to proper content-caching strategy decided by a rational ISP to cache more locally popular content and decrease times of cross-AS communication for content delivery cost saving. When content popularity increases up to very high values, caching strategy encounters the bottleneck. It is difficult for an ISP to serve its clients independently. Increasing frequency of cross-AS communication results in increasing cost for

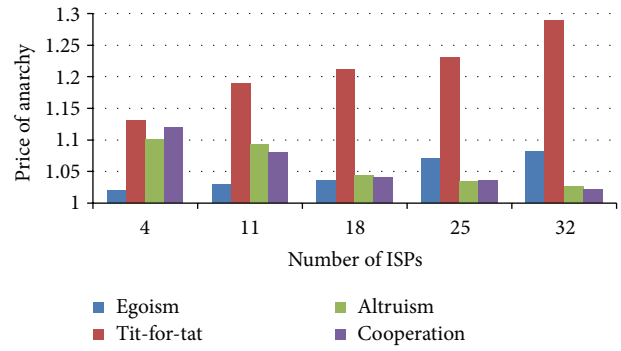


Figure 6: Price of anarchy versus ISP number.

the three peering relationships except cooperation. Although having relative high cost in low popularity range, cooperation shows a decreasing figure pattern and outperforms other models in very high popularity range. Egoism and tit-for-tat show caching strategic adjustment in their figures. Through caching more and more intradomain popular contents, egoistic peering ISPs lower the cost at the popularity range from 0.18 to 0.2, while tit-for-tat ISPs alleviate the sharply increasing trend of cost figure with a smaller slope during popularity range from 0.2 to 0.7. Additionally, at popularity range from 0.45 to 1, altruism model shows a stable equilibrium cost values which nearly do not increase with content popularity. Such good properties are instructive for ISPs to select proper peering strategy in different situations of content popularity.

6.3. Impact of ISPs Number. The number of ISPs also exerts an impact on the performance of the four games. We fix the values of content popularity and let the number of ISPs vary from 4 to all 32 ASs. Figure 6 shows the values of PoA change with ISPs number. The PoA of comparatively egoistic types (egoism and tit-for-tat) increases with the number of ISPs, yet the PoA of comparatively altruistic types (altruism and cooperation) decreases with the number of ISPs. According to formulas (3), (5), (8), and (12) and Proposition 1, the changes in Figure 6 is due to the advantage of coalitional and cooperative games that are well adapted to a large scale of AS-level topology with a large number of peering ISPs. Figure 7 shows the content delivery costs change with ISPs number. The average AS-hop-counts of all the four types increase with the number of ISPs. This is so because when the ISP number is relatively small, an ISP has few choices, caches most content items by itself, and accordingly lowers the times to do the cross-ISP delivery; when the ISP number becomes large, an ISP has increasing possibility to receive content from its adjacent ISPs or other ISPs.

6.4. Sharing Strategy in Tit-for-Tat. Sharing strategies in the peering relationship also have influence on the ISP's content delivery cost. In this subsection, we study two types of sharing function $f(x)$ in formula (5).

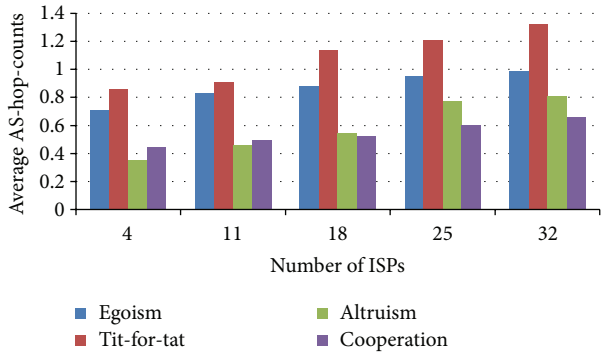


Figure 7: Content delivery cost versus ISP number.

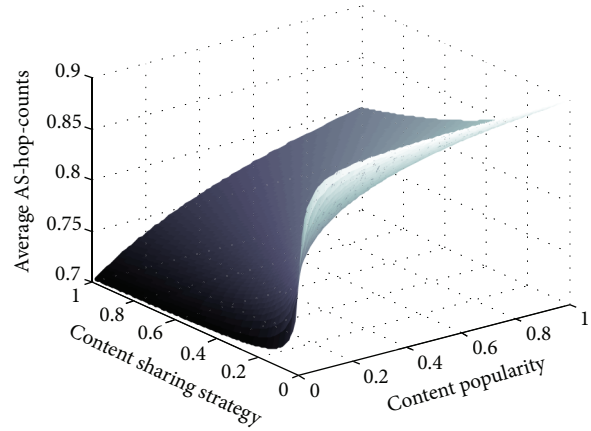


Figure 8: Content delivery cost in tit-for-tat peering.

Considering the peering relationships between n ISPs, we formulate a sharing function as follows:

$$f_j\left(X_{ij}^{(c_l)}\right) = \frac{p_{A_j}^{(c_l)} \cdot X_{ij}^{(c_l)}}{\sum_{k=1}^{n} p_{A_k}^{(c_l)} \cdot X_{ik}^{(c_l)}}. \tag{15}$$

This formula denotes that peering ISP A_j will decide its sharing strategy according to the proportion of the contribution $p_{A_j}^{(c_l)} \cdot X_{ij}^{(c_l)}$ to ISP A_i's total contribution to all peers. The cost utility function of ISP A_i can be expressed as follows:

$$U_{A_i} = \sum_{c_l} p_{A_i}^{(c_l)} \cdot \left(D_0 \cdot I_{A_i}^{(c_l)} + \sum_{A_j \neq A_i} D_1 \cdot f_j\left(X_{ij}^{(c_l)}\right) \cdot \left(1 - I_{A_i}^{(c_l)}\right) \cdot I_{A_j}^{(c_l)} \right). \tag{16}$$

To optimize the utility, an ISP has to balance its intradomain content demand and the peering relationship with other ISPs. Figure 8 shows a peering ISP's content-receiving cost change with both content popularity and content sharing strategies in the tit-for-tat game model. The average AS-hop-counts still broadly increase with content popularity. When content popularity is very low (from 0 to 0.05), sharing strategy varying from 0 to 1 does not affect the cost remarkably. Yet, when content popularity increases up to 1, the cost changes significantly from 0.9 to 0.8 across the sharing strategic varying range from 0 to 1. Hence, the peering ISP should decide proper sharing strategy in different situations. With great demand for content, peering ISPs should contribute more to each other for reciprocal content delivery.

Another type of the peering functions are expressed as $f_i(X_{ji}^{(c_l)}) = X_{ij}^{(c_l)}$ and $f_j(X_{ij}^{(c_l)}) = X_{ji}^{(c_l)}$. Function $f_i(x)$ is the inverse function of $f_j(x)$ and vice versa. Such mutually inverse property can instruct peering ISPs to design their sharing strategy as a function symmetric to the line $y = x$. The advantage of this design lies in that the two peering ISPs will have the same sharing function and monotonicity so that one ISP can precisely predict the sharing strategy of the other one.

7. Discussion

There are other highly involving issues that we ought to pay attention to.

First, the cost evaluation in our numerical analysis uses the average value of equilibrium in games between couples of peering ISPs. The actual interdependence and interaction processes among the involved ISPs are more complex than the simple combination of some two-person games. Yet, our models are the elementary games that are able to generate the comprehensive interrelationships, thus showing similar properties as the actually complex interactions. In addition, since distinct patterns about the cooperative willingness and peer's independent decisions are embodied in the game models, the numerical results will also reflect the cost trends of content delivery in distinct peering relationships. These results are instructive for an ISP's peering strategy choice.

Second, from a practical point of view, ISPs still base their peering relations on the fundament of peering agreements and data exchange protocols between networks. With regard to the current Internet policies, ISPs rely on two basic types of contracts (peering and transit) for exchanging traffic. Compared with traditional interconnection agreements among ISPs, peering for content needs not only the agreements to route traffic, but also a series of distributed networking protocols deployed on the boundary nodes of ISPs' networks to efficiently exchange content between the mutually recognized ISP peers. This should be elaborated in the new networking design.

Third, the Internet is experiencing the transition from a transit hierarchy to a peering mesh and becoming flat [10, 11]. It means that there are more and more peering ISPs at the same tier level. This evolution is good for wide adoption of content-based peering in the future Internet architecture. From our numerical analysis about the values of PoA and average AS-hop-count, we see that the tit-for-tat case, whose interaction between peers is similar to the transit hierarchy,

does not have a good performance in the evaluation. This may prompt ISPs to adapt themselves with more cooperatively peering strategies, which will motivate sharing of content-based ISP peers.

Finally, networks in the future Internet will arrive at a high level of intelligence, which means that an ISP's network will no longer be limited to the "stupid" data carrier and will actively involve content management for fulfilling QoS [21, 22]. Such intelligence is always related to the optimization mechanism with which the network can decide the optimal policy for content delivery. Multiple intelligent networks are supposed to use cooperative way to a "win-win" situation. Our proposed peering strategic game models can be applied for finding the optimal equilibrium solutions.

8. Conclusion

In this paper we propose the concept of peering for content and four game models for peering ISPs to decide content caching and sharing strategies and minimize communication cost of content delivery in content centric Internet. In different games, peering ISPs are characterized by distinct degree of cooperative willingness from egoism to altruism and are interconnected as profit-individuals or profit-coalition. Based on the AS-level topology of part of the Chinese Internet, we simulate the four games to evaluate the performance of price of anarchy value and content delivery cost resulted from ISPs' peering strategy selection. Peering relationships in our proposed games show some good properties (e.g., near-social-optimal PoA, caching strategic adjustment for cost-saving, and sharing strategic balance between intracontent demand and inter-ISP relation), which are instructive for ISPs in both the current and the future content-centric Internet.

Acknowledgments

This work is supported by the National Basic Research Program of China (973 programs) under Grant 2013CB329101 and partially supported by the National High-Tech Research and Development Program of China (863) under Grant no. 2011AA010701, the National Natural Science Foundation of China (NSFC) under Grant nos. 61232017, 61003283, 61001122, 61372112, Jiangsu Natural Science Foundation of China under Grant no. BK2011171, and the DNSLAB Project (China Internet Network Information Center) under Contact no.DNSLAB-2012-N-U-13.

References

[1] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proceedings of the ACM Conference on Emerging Networking Experiments and Technologies (CoNEXT '09)*, pp. 1–12, Rome, Italy, December 2009.

[2] J. Choi, J. Han, E. Cho, T. T. Kwon, and Y. Choi, "A survey on content-oriented networking for efficient content delivery," *IEEE Communications Magazine*, vol. 49, no. 3, pp. 121–127, 2011.

[3] K. Cho, H. Jung, M. Lee, D. Ko, T. Kwon, and Y. Choi, "How can an ISP merge with a CDN?" *IEEE Communications Magazine*, vol. 49, no. 10, pp. 156–162, 2011.

[4] N. Kamiyama, T. Mori, R. Kawahara, S. Harada, and H. Hasegawa, "ISP-operated CDN," in *Proceedings of the IEEE INFOCOM Workshops*, April 2009.

[5] G. Dán, "Cache-to-cache: could ISPs cooperate to decrease peer-to-peer content distribution costs?" *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 9, pp. 1469–1482, 2011.

[6] B. Krogfoss, M. Weldon, and L. Sofman, "Internet architecture evolution and the complex economies of content peering," *Bell Labs Technical Journal*, vol. 17, no. 1, pp. 163–184, 2012.

[7] C. Xu, G.-M. Muntean, E. Fallon, and A. Hanley, "A balanced tree-based strategy for unstructured media distribution in P2P networks," in *Proceedings of the IEEE International Conference on Communications (ICC '08)*, pp. 1797–1801, Beijing, China, May 2008.

[8] C. Xu, G.-M. Muntean, E. Fallon, and A. Hanley, "Distributed storage-assisted data-driven overlay network for P2P VoD services," *IEEE Transactions on Broadcasting*, vol. 55, no. 1, pp. 1–10, 2009.

[9] J. Dai, B. Li, F. Liu, B. Li, and H. Jin, "On the efficiency of collaborative caching in ISP-aware P2P networks," in *Proceedings of the IEEE INFOCOM*, pp. 1224–1232, Shanghai, China, April 2011.

[10] A. Dhamdhere and C. Dovrolis, "The internet is flat: modeling the transition from a transit hierarchy to a peering mesh," in *Proceedings of the 6th International Conference on Emerging Networking Experiments and Technologies (Co-NEXT '10)*, Philadelphia, Pa, USA, December 2010.

[11] V. Valancius, C. Lumezanu, N. Feamster, R. Johari, and V. V. Vazirani, "How many tiers? Pricing in the internet transit market," in *Proceedings of the ACM SIGCOMM Conference (SIGCOMM '11)*, pp. 194–205, Toronto, Canada, August 2011.

[12] S.-B. Lee, S. H. Y. Wong, K.-W. Lee, and S. Lu, "Content management in a mobile ad hoc network: beyond opportunistic strategy," in *Proceedings of the IEEE INFOCOM*, pp. 266–270, Shanghai, China, April 2011.

[13] J. F. Nash, "Non-cooperative games," *The Annals of Mathematics*, vol. 54, no. 2, pp. 286–295, 1951.

[14] R. Davis and D. Maschler, "The kernel of a cooperative game," *Naval Research Logistics Quarterly*, vol. 12, no. 3, pp. 223–259, 1965.

[15] T. Roughgarden, *Selfish Routing and the Price of Anarchy*, MIT Press, 2005.

[16] http://www1.cnnic.cn/IDR/GlobalInternetNews/.

[17] http://en.chinatelecom.com.cn/corp/index.html.

[18] http://eng.chinaunicom.com/about/Eng-gsgl/index.html.

[19] http://www.edu.cn/cernet_1377/index.shtml.

[20] http://www.cstnet.net.cn/english/aboutcstnet/briefintroduction.html.

[21] C. Xu, T. Liu, J. Guan, H. Zhang, and G. -M Muntean, "CMT-QA: quality-aware adaptive concurrent multipath data transfer in heterogeneous wireless networks," *IEEE Transactions on Mobile Computing*, vol. 12, no. 11, pp. 2193–2205, 2013.

[22] C. Xu, F. Zhao, J. Guan, H. Zhang, and G. -M. Muntean, "QoE-driven user-centric VoD services in urban multihomed P2P-based vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 2273–2289, 2013.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 915963, 14 pages
http://dx.doi.org/10.1155/2013/915963

Research Article

A Relation Routing Scheme for Distributed Semantic Media Query

Zhuhua Liao,[1] Guoqiang Zhang,[2] Aiping Yi,[3] Guoqing Zhang,[4] and Wei Liang[4]

[1] *Key Laboratory of Knowledge Processing and Networked Manufacturing, University of Hunan Province, Xiangtan 411201, China*
[2] *School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, China*
[3] *Hunan University of Science and Technology, Xiangtan 411201, China*
[4] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China*

Correspondence should be addressed to Guoqiang Zhang; guoqiang@ict.ac.cn

Received 11 August 2013; Accepted 4 September 2013

Academic Editors: K. Dejhan, J. He, and C. L. Hsu

Performing complex semantic queries over large-scale distributed media contents is a challenging task for rich media applications. The dynamics and openness of data sources make it uneasy to realize a query scheme that simultaneously achieves precision, scalability, and reliability. In this paper, a novel *relation routing scheme (RRS)* is proposed by renovating the routing model of Content Centric Network (CCN) for directly querying large-scale semantic media content. By using proper query model and routing mechanism, semantic queries with complex relation constrains from users can be guided towards potential media sources through *semantic guider nodes*. The scattered and fragmented query results can be integrated on their way back for semantic needs or to avoid duplication. Several new techniques, such as semantic-based naming, incomplete response avoidance, timeout checking, and semantic integration, are developed in this paper to improve the accuracy, efficiency, and practicality of the proposed approach. Both analytical and experimental results show that the proposed scheme is a promising and effective solution for complex semantic queries and integration over large-scale networks.

1. Introduction

Over years, many research efforts have been devoted to large-scale data retrieving in dynamic self-organized network environments, for example, WWW and various forms of peer-to-peer networks. Database and multimedia technologies provide the means to connect and interact with rich elements of static data, which stimulates media artifacts generation in various kinds of networks. However, since the state of nodes may change at any time, data may be moved from one place to another, and links between data may be broken; many new characteristics of the rich media emerge in these networks. For example, many semantically interrelated data are located on different nodes, numerous distributed data fragments are about the same scenario, and many *knowledge islands* are emerging. These issues pose great challenges to realize efficient data query and integration schemes that support complex semantics. In addition to supporting queries with complex semantics, it is also of equal importance to improve the precision, resilience, and scalability of the data query schemes.

In traditional web search engine based solutions [1, 2], distributed media contents are crawled from the Internet to one or a few sites, and then, a central, big index structure is built for centralized querying after data is extracted and sorted. Although centralized searching can support complex semantic query, it is, however, hard to meet the dynamism of the network itself and the speed of network information updating.

A number of schemata have been proposed to timely search the fully decentralized network data. However, these schemata fail to simultaneously meet the requirements of efficient routing and the ability to express complex semantic queries. According to whether index technique is used, these solutions can be classified into three categories [3]: mechanisms without an index, mechanisms with dedicated

index nodes, and mechanisms with indices at each node. According to the semantic capability, these solutions can also be classified into three categories: document-oriented data retrieval, topic searching, and structural or semantic query. Unfortunately, all these solutions that achieve efficient routing only support keywords or topic-based queries; they are not capable of expressing complex structural or semantic queries.

However, in reality, the ability to support distributed semantic query over large-scale networks will enable two promising features: (1) filtering query results by imposing constraints over the content to be searched. In keywords based search, many irrelevant or duplicated results may be returned, whereas with semantic query and data integration, irrelevant results will be pruned at data sources, and duplicated results will be suppressed at intermediate routers; (2) retrieving relevant data by appropriately specifying relation constraints. Relation constraints allow users to retrieve relevant data (e.g., fragmented data) in one query, rather than issuing multiple queries for different data or data parts.

This paper is concerned with the elementary theory and implementation of relation routing for semantic queries over large-scale networks, which can support complex queries over large-scale semantic media that vary with location and semantic schema. The ability of querying and integrating decentralized media over large-scale networks is of great benefit to individual learning, customized media authoring, and so forth. In our schema, applications can complete various semantic queries and sequential integrations over large-scale semantic media without having to build a global semantic view of the media at each node or at a central node in advance. The contributions of this paper are as follows.

(i) We presented the whole framework of relation routing scheme (RRS) that supports complex semantic query makeup, query routing, and result integration.

(ii) We addressed several fundamental issues that are specific to RRS: semantic-based naming, the incomplete response problem arising from the $1:n$ relationship between query and potential results, and the relation integration for semantic or performance considerations.

(iii) We provided a prototype implementation of RRS by extending the CCN protocol. We inherited several building blocks or designing principles from CCN, such as the guided information flooding routing model and soft-state design principle, but meanwhile, we also made several extensions to CCN, such as the introduction of Response Relation Table to tackle the incomplete response problem.

The rest of the paper is organized as follows. Section 2 describes related work. Section 3 details how to extend the forwarding engine of CCN to realize our relation routing scheme. Section 4 discusses how to perform relation integration. Section 5 presents both analytical and experimental performance evaluation of RRS. Finally, Section 6 summarizes our work and concludes the paper.

2. Related Work

Query and integration of relational data in distributed systems can be realized in three paradigms: with central repository, with federation, and with peer-to-peer organization. In the central repository paradigm, structural or semantic data are crawled from data sources and put into a centralized data store. In the federation paradigm [4, 5], data are distributed. A centralized portal is used to receive all data requests, decompose them, and finally distribute the resulting requests to appropriate data stores. Peer-to-peer paradigm itself can be classified into two forms: unstructured and structured. Structured P2P systems build around the theory of distributed hash table which uses flat identifiers to store and locate data [6, 7]. Using flat identifiers, however, is not suitable for complex semantic queries. Unstructured P2P systems, on the other hand, have the potential to employ more complex semantics to query data. Early days' unstructured P2P systems, as well as sensor networks and mobile networks [8–10], rely on flooding or random walk to locate data, which has poor scalability. In response, several methods are proposed to address the scalability issue, for example, range query [11], multidimensions query [12], and the routing indices (RIs) [3]. RIs provide a list of "directions" towards the potential content sources for the query, which is similar to our semantic guider. However, the "directions" information maintained in a node's indices is a list of "coarser" topics and the number of files falling into each topic summarized from its nearby neighbors, which cannot fully support complex semantic queries (e.g., queries with relation constraints).

Support of semantic queries over large-scale networks has attracted much attention in recent years. In the layered semantic overlay networks (SONs) [13], nodes with semantically similar contents are "clustered" together, which makes network topology dependent on the contents and causes substantial maintenance overhead. In the SQPeer middleware [14] approach, each peer has to broadcast its data schema—which includes all RDF classes and properties—to (or requested by) other peer nodes to support semantic queries.

Most recently, several content centric networking approaches [15–19] are proposed to elevate content/information to the first-class citizen in the network. Data-Oriented (and beyond) Network Architecture (DONA) [15] focuses on content-based anycast. It uses flat, self-certifying names to identify content and relies on the Resolution Handlers (RHs) infrastructure to route a content request to the appropriate data source. Publish/Subscribe for Internet Routing Paradigm (PSIRP) [16] proposes a publish/subscribe based networking architecture. In Content Centric Network (CCN) [17], contents are identified by hierarchical names, and these names are announced, aggregated, and propagated into the routing system. It uses the guided information flooding model to route a request towards potential source(s). When a request is being forwarded, it leaves bread crumbs in the Pending Interest Table (PIT), which is later used to forward data packets back to the requester(s)(PIT is also used to suppress requests for the same content). In addition, CCN also supports transparent in-network content caching to reduce data traffic.

These aforementioned content centric networking approaches, however, mainly focus on facilitating users to efficiently retrieve a single named content. They do not tackle the problem of querying the network for relevant data with semantic constraints. The approach of similarity content search [20] moves one step further. Based on CCN, it introduces "search" as a top level namespace and uses flooding to search similar objects in a network, but it does not consider complex queries. Zahariadis et al. [21] have presented an Autonomic Layer-Less Object Architecture (ALLUA) framework which assigns different types of properties to content objects. From these properties, one can know several things such as the creator of the object, its relationship with other objects, and the way it is used.

In short, although there are various distributed index techniques and distributed query models, none of them can effectively deal with the semantic media data query and integration in large-scale networks. However, this capability is becoming increasingly important as network is more and more considered as a large distributed data store with rich semantic media than a simple communication medium.

3. Relation Routing Model

In this section, we introduce our relation routing model. We choose to extend the CCN to meet the goal of effectively dealing with the semantic media data query and integration in large-scale networks. Several reasons motivate the selection of CCN as our starting point. Firstly and most importantly, CCN's several design principles, such as name-based routing, guided information flooding model, soft-state PIT design, and in-network caching, naturally fit the needs of semantic query over large-scale networks. Secondly, CCN can be layered upon the existing Internet infrastructure or directly upon layer 2 networks, which has good potential of deployment. Finally, CCN is an open source which we can modify as we want.

Certainly, the goal of RRS is quite different from CCN. CCN focuses on the retrieval of a single named content, whereas RRS emphasizes on the ability to specify complex semantic queries and retrieve related data satisfying relation constraints. This distinction is the root cause of many differences between CCN and RRS, such as naming, packet format, forwarding procedures, and cached content.

3.1. Naming and Semantic Query Format. In rich media networking, since the semantic relation depends mainly on data itself, it is natural to use the name of data as well as its associated properties to identify and summarize its content. Because hierarchical namespace is very useful when one does not know the exact name of a data but has in mind only a data category [22] and hierarchical names are easy to be aggregated, we use the hierarchical namespace and adopt the URN in each level. However, different from traditional hierarchical names, where each part of the name corresponds to a location or organization related domain, our hierarchical names are solely based on semantic classification of data.

We propose *semantic guider node* (SGN) that is able to give multigranularity directions towards the right sources for a referenced subject. With hierarchical namespace, every SGN only needs to hold the *hierarchical prefix* about the name of data rather than all the semantics of data to direct queries to right sources. The relationship between SGN and other network entities is briefly illustrated in Figure 1. A user's query can be represented by a reference subject (represented as a hierarchical name) plus constraints on the subject. The query will be forwarded by SGNs towards potential sources. When the query finally arrives at right sources, the constraints are used to extract matched data. The results will be returned and may be integrated on their way back.

In order for SGNs on different nodes to communicate and cooperate with each other, it is necessary that all SGNs conform to the same standard; that is, metadata of the name, relation, and property of data on all SGNs should be uniformly defined. Interaction between SGNs and heterogeneous data sources, however, requires transformation so that standard user queries are transformed to source-specific query languages and response results are transformed back to standard format metadata representation. In this way, heterogeneity of data sources is shielded from users.

Usually, a query can have property or relation constraints. We thus define two kinds of basic query formats: (1) query with property constraint; (2) query with relation constraint. The detailed formats of the queries are as follows.

(1) "Prefix + property constraint": the prefix is a hierarchical classification that restricts the subject of the query. The property constraint is a mixed logical and arithmetic function on any property of the given "prefix", which is in the form: $\langle property1\ MOP\ V1\ LOP\ property2\ MOP\ V2 \ldots\rangle$, where MOP is an arithmetic operator, $V1$ or $V2$ are predefined values, and LOP is a logical operator. For example, if a user wants to query the media of the subject (prefix) "*computer/network*", with creation time (T) to be within the last 2 years and media format (F) to be "MPG", then the constraint can be represented as "$p\{T > Year(\text{"}NOW\text{"}) - 2 \wedge F = \text{"MPG"}\}$".

(2) "Prefix (+name) + relation constraint": the prefix is a hierarchical classification or the hierarchical classification plus a "name" as the major reference subject. The relation constraint expresses the relation between the prefix (+name) and the data that the user wants. It is in the form of $\langle relation\ 1\ OP\ relation\ 2\ OP \ldots\rangle$, where the OP is a logical operator. For example, if a user wants to query all the precursors and successors of the known media (or prefix) b constrained by relations R and S, the constraint can be represented as "$r\{R(*, b) \wedge S(b, *)\}$".

The two basic query formats can be used to describe various semantic queries in a unified format "hierarchical prefix + constraints". To compactly express semantic queries, we adopt the relation algebra method. Nevertheless, such expressions can be easily rewritten in existing query languages (e.g., SQL or RQL [23]). In this paper, we focus on the second query format, as the routing processes of the two

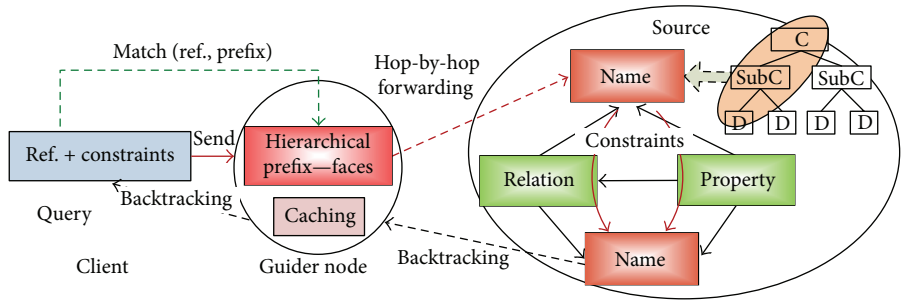


Figure 1: The interactions between "semantic guider" and other network entities.

Table 1: Packet structures.

(a) Relation query packet

Name	Query name in the form of name_r<constraints>
Selector	Order preference, publisher filter
Scope	Response quantity, timeout, and so forth
Nonce	Time stamp or sequence number to avoid routing loops

(b) Relation response packet

Name	Same as in the relation query packet
Publisher ID	Identification of the publisher for this relational data
Validation Key	Information needed to verify the validity of relational data.
Data	A sequence of relational data results

kinds of query formats are the same and the only difference lies in the procedures taken when the query reaches potential data sources.

3.2. Packet Structures. Two types of packets are defined in RRS: relation query packets and relation response packets. They are shown in Table 1.

(1) *Relation Query Packet* (*Table 1*(*a*)). The packet contains four fields: Name, Selector, Scope, and Nonce. In the *Name* field, the user defines a referenced hierarchical name and/or prefix of the subject of interest, plus additional relation and/or property constraints. The *Selector* and *Scope* fields can be used to narrow the query's scope. In order to completely retrieve the relational data for a pending query, the SGN does not discard the pending query as soon as a response packet arrives. It waits some additional time to receive other related results from other sites. The time to wait is determined by the settings in the *Scope* field. While a user is querying for media content, he/she can estimate and set the value of certain query parameters, such as estimated maximum round trip time, publisher, and so forth. These items are defined in the *Selector* field. The *Nonce* field is inherited from CCN's request packet, used to avoid packet forwarding loops.

(2) *Relation Response Packet* (*Table 1*(*b*)). The packet is mainly used for carrying a list of relational data with additional information, which also contains four fields: Name, Publisher ID, Validation Key, and Data. The *Name* field is simply copied from the counterpart in the relation query packet. The *Data* field carries relational data that matches the query. The *Publisher ID* field identifies the publisher of the relational data. And the *Validation Key* field contains information needed to verify the validity of the relational data, for example, public key.

3.3. Forwarding Engine of SGN. In RRS, the main task of SGN is to offer query guidance, data backtracking, and result integration. We extend the forwarding engine of CCN to support the relation routing and result integration, which is shown in Figure 2. Besides the Pending Interest Table (PIT) (renamed as Pending Query Table in RRS model for reasons as below), Forwarding Information Base (FIB), and content store(renamed as metadata store for reasons as below), a new data structure called Response Relation Table (RRT) is added to the engine. The following is a detailed description of these components.

(1) *Pending Query Table* (*PQT*). Similar to the PIT in CCN, PQT is used to temporarily record the semantic queries that have been forwarded but still wait for responses. PQT also follows the soft-state design principle, which means that each entry will be deleted after a reasonable period of time. However, the choice of when to delete the query in PQT is different from CCN. In CCN, a request is for a definitely named content, so the relationship between request and response is 1 : 1. Hence, the query in PIT can be deleted immediately after one copy of the data returns, whereas in RRS, the relationship between

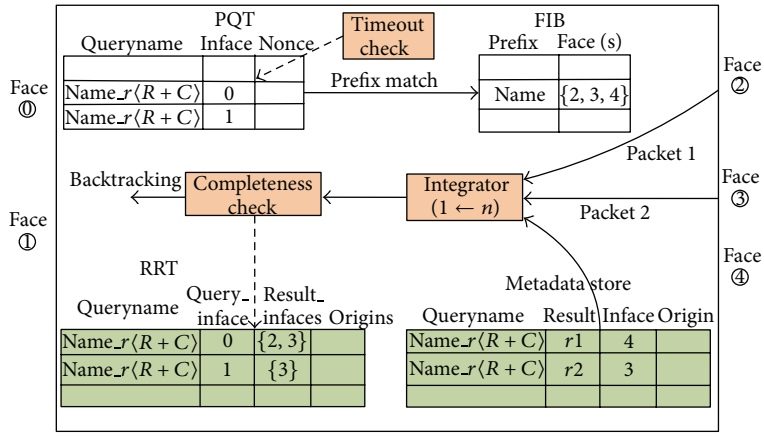


Figure 2: The forwarding and integration engine for relation routing scheme.

semantic query and response is $1 : n$; that is, a semantic query may result in many different results, so it is not appropriate to delete the entry immediately after one response returns.

(2) *Forwarding Information Base* (*FIB*). FIB stores the names and/or the prefixes of all media that are distributed across the whole network and the outgoing faces(Here face refers to the port that is used to exchange information directly with their neighbors for application processes) through which they can be reached them. Except for the semantic distinctions in names, the FIB in RRT is the same as in CCN.

(3) *Response Relation Table* (*RRT*). RRT is a newly introduced data structure, which is mainly used to handle the incomplete response problem introduced in Section 3.4. Briefly speaking, the incomplete response problem also arises from the $1 : n$ relationship between semantic query and response. In case when different users issue the same query, it is possible that some users can only get incomplete responses. RRT is used to record the presently returned results for a query, so that incomplete response to a query can be detected.

(4) *Metadata store*. Unlike CCN's content cache, which caches the data content itself, the metadata store in RRS caches relation response results and related information (such as publisher). This information gives detailed descriptions of the data contents, including the means to fetch the contents. Despite this distinction, the purpose of metadata store is similar to the content store in CCN, that is, to satisfy popular queries from cache, rather than forwarding the query outward each time. Traditional cache replacement policies, such as *Least Recently Used* (LRU), can be used.

The major attributes as well as their descriptions for these components (except for FIB) are given in Table 2.

3.4. The Incomplete Response Problem. Since the data of interest to a user may be distributed over several data sources, it is important to retrieve the response packets from multiple sources. However, when different users issue the same query (i.e., the name and preference are the same), it is possible that some users only get incomplete responses (or no response packets at all) before the due time of the pending query. For example, consider the situation when user A has sent out a query $Q1$ and got *partial (or complete)* response packets, and before the deletion time of its PQT record, some user B sends out the same query $Q1$. This query will be suppressed and combined in the PQT. If the metadata store evicts the cached results of $Q1$ before B's query arrives, then early responses to $Q1$ before the arriving time of B's query will be lost. This is called the *incomplete response* problem. To address this issue, we take two modifications to the basic CCN forwarding engine.

Firstly, when a first response packet is received, the SGN does not remove its query record in PQT immediately(in CCN, the query is removed immediately after the expected content arrives); instead, it sets a due time according to two options. One is the timeout option, which determines how much time the pending query takes to be deleted after it has arrived. Generally speaking, the due time is affected by several factors such as the network transit delays, the duplicate number of the referenced subject, and the complexity of the query. Another option is the "maximum number of response packets" option, which determines how many packets to be received before the pending query can be deleted.

Secondly, we introduce an additional data structure, called *Response Relation Table* (RRT). This table is used to temporarily record which query results have been returned to corresponding pending queries in PQT. The major fields of the RRT are shown in Figure 2. When an SGN receives a query, it will take different actions depending on whether queries with the same query name have been received within a given time interval. If the query is the first one with this query name, then the SGN will search the FIB with the longest prefix match algorithm and forward the query to all matching faces (except the arriving face). Otherwise, if the query is not the first one, then the SGN will first look up

Table 2: Major attribute names and their descriptions for the PQT, RRT, and metadata store.

Attribute name	Description
	Pending Query Table
Queryname	Name of the query represented in the format as discussed in Section 3.1
Inface	The incoming face of the query
Nonce	The nonce of the query packet for loop avoidance
	Metadata store
Queryname	Same as in PQT
Result	Result of the query
Inface	Incoming face of the result
Origin	Source that generates the result
	Response Relation Table
Queryname	Same as in PQT
Query_inface	Incoming face of the query
Result_infaces	Incoming faces of the results that have already been forwarded through query_inface for this query
Origns	Sources that generate these results

the metadata store and return any results that match the query. Because it is possible that early arrived results can be evicted by a cache replacement policy, it is required to validate the cache completeness. The basic solution is to assume that the Response Relation Table for the first query with this query name records all the incoming faces of the results that have so far been returned, so by comparing this incoming faces with the incoming faces of the results for the query in the metadata store, it is possible to deduce whether the cached result is complete. If the cached result is incomplete, that it is necessary to forward the query to *some faces (the difference between the above two incoming faces)* to ensure response completeness. The complete flow chart of the query processing logic of SGN is shown in Figure 3.

Certainly, it is necessary to periodically check whether records in PQT are timeout (note that only one timer is set for all pending queries with the same queryname), which is not shown in the flow chart. For those pending queries, that all response packets have been sent back before timeout, the associated RRT records are also deleted. Otherwise, the query should be reforwarded to appropriate faces.

4. Relation Integration

When multiple response packets are available, it is necessary to integrate these results and filter redundant ones before forwarding the results back so that computation and bandwidth overhead can be reduced. This is what relation integration does. A relation integrator accepts multiple inputs and produces one output by logical integration.

In the following, we first introduce several relation operations and four basic integrations and then show how complex semantic query processing can be facilitated through these basic operations.

4.1. Relation Operations. Relation operations can be executed on SGNs or the data sources where a set of media related to a given query can be integrated. It can also be executed on the client node before the results are presented to the user. Four basic operations are defined.

(i) *Ranking*. This operation is used to sort the elements in multiple relation response packets based on property or relationship. For example, regarding the query "$C1/C2/C3\_r\{\cdots\}$", if an SGN has received 3 result packets, $\{e1\}$ for prefix "$C1/C2/C3$", $\{e2\}$ for prefix "$C1$", and $\{e3\}$ for prefix "$C1/C2$", the results can be sorted as $e1 > e3 > e2$ if ranked by the length of matched prefix.

(ii) *Intersection*. This operation is used to select common data out of multiple relation response packets. For example, $\{b1, b2\} \cup \{b2, b3\} = \{b2\}$.

(iii) *Union*. This operation is used to merge all relational data packets of the same query and remove duplicated ones. For example, two relation response packets are merged: $\{b1, b2\} \cup \{b2, b3\} = \{b1, b2, b3\}$.

(iv) *Complement*. This operation is used to select differential data from multiple relation response packets. For example, $\{b1, b2\} \setminus \{b2, b3\} = \{b1, b3\}$.

4.2. Basic Relation Integration. Four basic relation integrations are defined: forward composition, backward composition, joint composition, and bridging composition. These basic integrations can be combined to produce more complex integrations. In the following, we first introduce the four basic integrations and then show how to use these basic integrations to generate more complex ones.

(1) *Forward Composition*. The composition can be expressed as $R \mid (a, *) = \{(a, b)R(a, b)\}$, and the name of the query can be defined as "$a\_r\langle R(a, ?)\rangle$". It is used to retrieve and integrate the media objects that are the successors of the prefix (+name) a, and the relationship between the data object a and these media data is R. Figure 4 gives an example of forward composition. When the query is executed at two potential sources, source 1 and source 2, the subresults

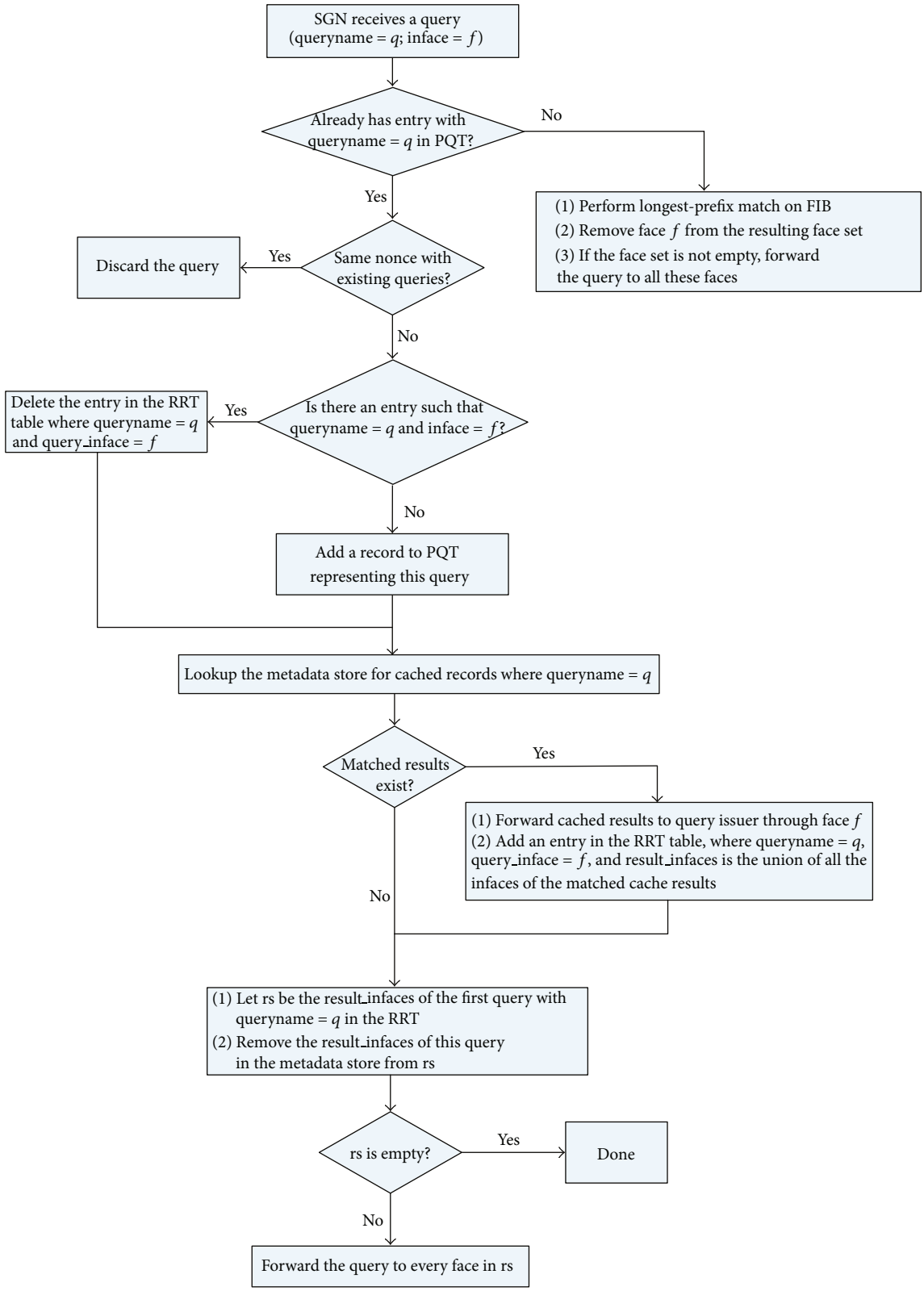


Figure 3: Query process logic.

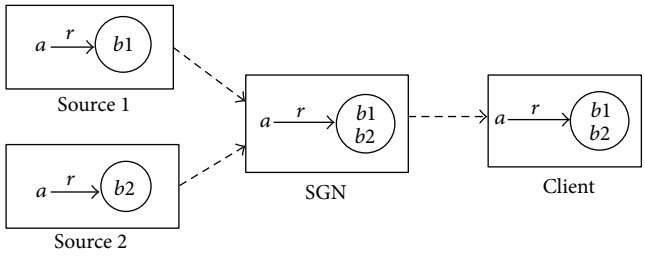


Figure 4: The integration process of the forward composition.

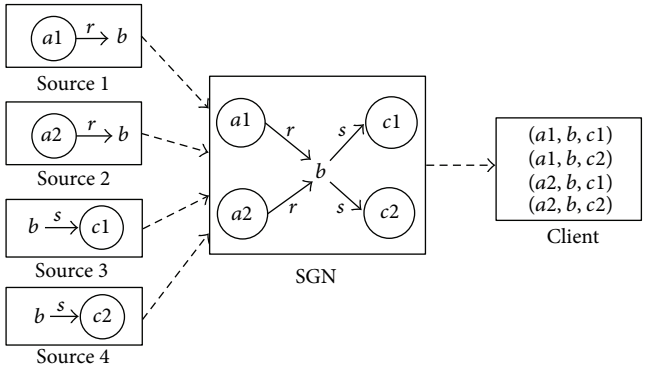


Figure 5: The integration process of the joint composition.

$r(a, b1), r(a, b2)$ will be returned and can be combined at either the SGN or the client.

(2) *Backward Composition*. The backward composition is opposite to the forward composition, which can be expressed as $R \mid (*, b) = \{(a, b) \mid R(a, b)\}$. The name of the query can be represented as "$b\_r\ \langle R(?, b)\rangle$". The integration process of the backward composition is similar to that of the forward composition.

(3) *Joint Composition* (*Jcomp*). This composition can be represented as $R * S \mid (*, b, *) = \{(a, b, c)R(a, b) \wedge S(b, c)\}$, and the name of the query can be represented as "$b\_r\ \langle Jcomp(R(?, b), S(b, ?))\rangle$". It is used to integrate all the precursors and successors of the known media (or prefix) b with relation constraints R and S. Figure 5 visually exemplifies this integration process. When the query is executed at the potential sources, suppose that the subresults are $R(a1, b)$ in source 1, $R(a2, b)$ in source 2, $S(b, c1)$ in source 3, and $S(b, c2)$ in source 4 and can be returned to the guider node. The joint composition $b\_r\ \langle Jcomp(R(?, b), S(b, ?))\rangle$ at the guider node will result in four data sequences, that is, $(a1, b, c1)$, $(a1, b, c2)$, $(a2, b, c1)$, and $(a2, b, c2)$.

(4) *Bridging Composition* (*Bcomp*). This composition can be represented as $R * S \mid (a, *, c) = \{(a, b, c) \mid R(a, b) \wedge S(b, c)\}$. The name of the query can be represented as "$a \circ c\_r\ \langle Bcomp(R(a, ?), S(?, c))\rangle$". It is used to integrate the media data that are associated with a known precursor and a known successor with relation constraints R and S. The integration process of the bridging composition is illustrated in Figure 6. When the query is executed at the potential sources, and the subresults are $R(a, b1)$ at source 1, $R(a, b2)$ at source 2, $S(b1, c)$ at source 3, and $S(b2, c)$ at source 4, the bridging composition

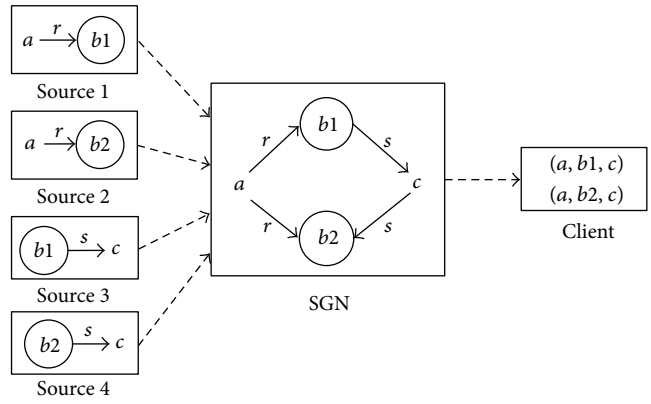


Figure 6: The integration process of the bridging composition.

"$a \circ c\_r\ \langle Bcomp(R(a, ?), S(?, c))\rangle$" at the guider node will result in two data sequences, that is, $(a, b1, c)$ and $(a, b2, c)$.

4.3. More Complex Integrations. The above basic compositions can be used to generate a variety of complex integrations of practical use. In this subsection, we illustrate four complex integrations.

(1) *Iterative Forward Composition* (*IFC*). It iteratively uses the forward composition to generate longer data sequences consecutively constrained by the same relationship R. The integration can be defined as $R \mid (a, **) = \{(a, b, c, \ldots) \mid R(a, b) \wedge R(b, c) \wedge \cdots\}$. The name of its query is expressed as "$a_r\ \langle$while $R(a, ?)$+*constraint*$\rangle$", where *constraint* describes the integration restriction, such as the length limit of generated sequences. For example, if a user wants to generate the sequences that start with object a, consecutively constrained by relationship R, and the length is no greater than 3, the query name can be expressed as "$a\_r\langle$*while* $R(a, ?) +$ "≤ 3"$\rangle$".

(2) *Iterative Backward Composition* (*IBC*). Similar to IFC, IBC iteratively uses backward composition to generate longer data sequences consecutively constrained by the same relationship R. This integration can be defined as $R \mid (**, a) = \{(\ldots, c, b, a) \mid R(b, a) \wedge R(c, b) \wedge \cdots\}$. The name of its query is represented as "$a_r\ \langle$while $R(?, a)$ + *constraint*$\rangle$", where *constraint* is the same as that of IFC.

(3) *Base Hybrid Composition* (*BHC*). BHC is to generate longer data sequences that link specified objects and expected objects constrained by different relationships. In this integration strategy, a user can specify several objects as well as the relationships between the specified objects and the expected objects. For example, given the objects a, c, and e, the relationships $R1$, $R2$, $R3$, $R4$, and $R5$, the subsequence among the objects a and c is the bridging composition with relationships $R1$ and $R2$, the subsequence near the object e is the joint composition with relationships $R4$ and $R5$, and the relationship between the object c and the precursor of the object e is $R3$. So this integration can be expressed as $R1 * R2 * R3 * R4 * R5 \mid (a, \%, c, \%, e, \%) = \{(a, b, c, d, e, f) \mid R1 * R2 \mid (a, *, c), R4 *$

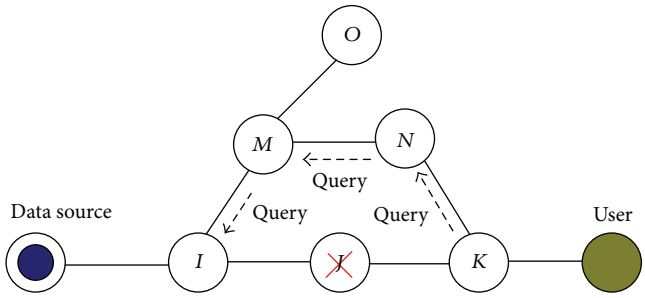


Figure 7: Solution for recovery of the return path.

$R5 \mid (*, e, *), R3(c, d)\}$, and the name of its query can be represented as "$a\_r\ \langle Bcomp(R1(a, ?),\ R2(?, c)), Jcomp(d = R4(?, e), R5(e, ?)), R3(c, d)\rangle$".

(4) *Limited Hybrid Composition* (*LHC*). The integration strategy can generate long sequences with various relationships and length constraints. According to the referenced object from which a sequence is generated, LHC can be categorized into three kinds: (1) limited hybrid composition starting from the front (LHC-F); (2) limited hybrid composition starting from the middle (LHC-M); (3) limited hybrid composition starting from the end (LHC-E). Here, we elaborate on the first kind. For example, consider composing a finite sequence which starts from object a. The subsequence (a, b, c) is constrained by relationship $R1$, subsequence (c, d, e) is constrained by relationship $R2$, and subsequence (e, f) is constrained by relationship $R3$. The integration strategy can be represented as $R1(2)R2(2)R3(1) \mid a = \{(a, b, c, d, e, f) \mid R1(2) \mid (a, **), R2(2) \mid (c, **), R3 \mid (e, *)\}$. The value in parenthesis "( )" behind the relationship describes the length limit of the subsequence. The name of this query can be written as "$a\_r\ \langle (b, c) = \text{while } R1(a, ?) + \text{"} \leq 2\text{"}, (d, e) = \text{while } R2(c, ?) + \text{"} \leq 2\text{"}, f = R3(e, ?)\rangle$" or "$a\_r\ \langle b = R1(a, ?), c = R1(b, ?), d = R2(c, ?), e = R2(d, ?), f = R3(e, ?)\rangle$".

5. Performance Evaluation

We evaluate the system performance of RRS from both analytical and experimental perspectives.

5.1. Analytical Evaluation. In this subsection, we analytically evaluate the performance of RRS from several aspects: index precision and completeness, routing scalability, and reliability.

5.1.1. Index Precision and Completeness. Index precision means the extent to which a user's query can be accurately routed to potential sources. Since the FIB in RRS is built in the same way as in CCN, which holds media names or prefixes advertised by all media sources, the precision and completeness of indices of RRS should be the same as in CCN. In comparison, since RIs only maintain local indices, their precision and completeness should be lower than RRS.

5.1.2. Routing Scalability. The routing protocols used in RRS and CCN are similar to conventional routing protocols in the present-day Internet, such as BGP, whose scalability have already been demonstrated. The main difference lies in naming. The hierarchical naming mechanism can describe different classes of media and be used to perform matching at different granularities, which greatly improves the ability to cope with the diversity of user queries. Prefix aggregation can be used to condense the FIB size (e.g., "computer/network/TCP" and "computer/network/UDP" into "computer/network"), provided that the aggregation will not result in routing uncertainty. Other software or hardware enhancements could also be adopted to improve the forwarding and matching efficiency, such as those being developed in the ongoing NDN (Named Data Networking) project [24].

5.1.3. Reliability. If a node in the network goes offline, the query accessibility (query accessibility means reachability between the query source and potential data sources) should be maintained and the backtracking path for the results should be resumed. Query accessibility is an important factor that affects the results returned. Since the name is independent of location, a mechanism similar to the Opaque LSA [25] in the OSPF protocol can be added to guider nodes to describe the state of its neighbors and update the corresponding faces in FIB and PQT when the neighbor changes. If a guider node can propagate all prefixes to its neighbors in due time, queries can be guided to potential sources. In order to resume a broken backtracking path due to node or link failure, it needs to discover a redundant return path for response packets. One possible solution is illustrated in Figure 7. When a node (say J) fails, its adjacent neighbors (I, M, and K) can sense this failure and then readvertise their routing table to their neighbors (according to the routing protocol). After the routing table stabilizes, the downstream node K of the failed node then resends the query toward the data source. When the query reaches the upstream node I of the failed node, I replaces the invalid face of J with the new face of M, hence, recovers the return path. In addition, reliability is also benefited from the increase in duplication, either by source or by cache.

5.2. Comparative Evaluation. Table 3 summarizes the important features of the RRS and some of the current distributed media query schemes. In Table 3, we compare the following aspects that dominate the query performance of these systems.

(1) *Query Description*. Most systems consider either complex user requirements without taking into account the routing or simply defined requirement for routing. We have considered both requirement description (constraints) and routing information (prefix) included in the query of RRS.

(2) *Query Model*. Before the advent of the P2P networks, most systems should provide a global view or central index structure for the purpose of distributed query. P2P networks do not have this requirement, but efficiently delivering complex user requests to the right sources cannot be handled

Table 3: Comparison of distributed media query schemes.

Scheme	Distributed database	Structural P2P network	Unstructured P2P network	Web search engine	RRS
Query description	Structural query language (e.g. SQL)	Set(s) of flat identifiers (e.g. DHT)	No limited the forms	Keywords plus Boolean operations	Prefix + semantic constraints
Query model	To query global view	To match structural mapping between identifier and location	Flooding, random walk, supernode routing, and so forth	To query the centrical indexes	Prefix-based routing
Adaptable to dynamic sources	Poor	Good	Very good	Poor	Very good
The optimality of transmission path	Poor	Poor	Poor	Poor	Very good
Acquisition directivity	Good	Very good	Poor	Very good	Very good
Search capability	Stronger	Weak	Rely on the matching capability of forwarding nodes	Normal	Stronger

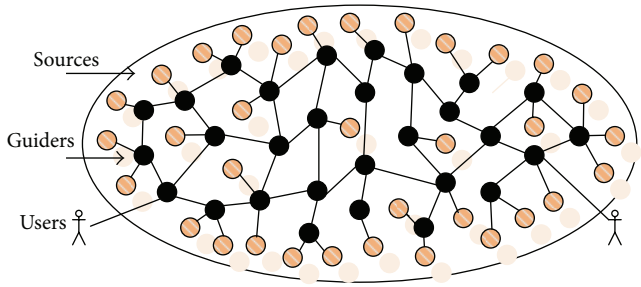


Figure 8: The schematic topology of the experimental network (sources are more than guider nodes).

well. To deal with all these issues, we employ the prefix-based routing in the RRS in view of the prefix provided by the user query.

(3) *Dynamic Adaptability*. It is not easy for traditional systems to retrieve the data distributed across a network when the data moves from one place to another before the system updates the global view or indices. The RRS has the advantage of dynamic adaptability even when the name and position of all distributed data are changing, which is similar to the unstructured P2P network.

(4) *Transmission Optimality*. Most systems build the query system above the communication facilities of network, so they cannot identify the local duplicates of data on the network. The RRS, however, can quickly access the local content and support transparent in-network content caching to cut short the transmission path.

(5) *Acquisition Directivity*. Most systems have considered the query direction provided by global view, central indices, or distributed hash table except for unstructured P2P networks. However, all those schemata should be limited to the global view, central indices, or distributed hash mapping to show the right direction towards certain host(s). To the RRS, the direction is easily established by all sources multicasting all the prefixes of content names.

(6) *Search Capability*. Generally, we consider the query capability from the recall and precision of search. In this way, the distributed database has a stronger capability for searching the data among the federated database systems by structural query language (e.g., SQL). But the search capability of the unstructured P2P network relies on the semantic matching capability of forwarding nodes as it imposes no constraint on the query format and the web search engine cannot provide the high precision of search due to the keywords search pattern. However, the RRS can provide much stronger search capability by using unified names on the network and carrying semantic constraints in the query.

5.3. Experimental Evaluation

5.3.1. Experiment Setup. To evaluate the functionality and efficacy of RRS, we build an overlay network with 2074 nodes on a PC cluster composed of 68 physical computers. We randomly select 1054 nodes as data sources and take the others as guider nodes. Data nodes are randomly connected to one or more SGNs. The CPU of these computers are Intel Pentium 4, and the RAMs range from 512 MB to 1 GB. Each data source stores relational media data about all or part of 30 college courses, and the relational data (more than 3 thousand records at each source) are stored in MySQL databases. The semantic description of local relational media at all sources conforms to a uniform name and relation specification. We set the number of duplicates (duplicate factor) of individual media objects (name or prefix) to be between 1 and 16. These duplicates are distributed over different sources. A conservative threshold of 3 minutes is used for the residence time of each arrived query in the PQT. The schematic topology of the experimental network is shown in Figure 8. Relations of partial media data are shown in Figure 9, which gives the complex relations, such as hierarchical classification, inclusion, and intersection, between course media data.

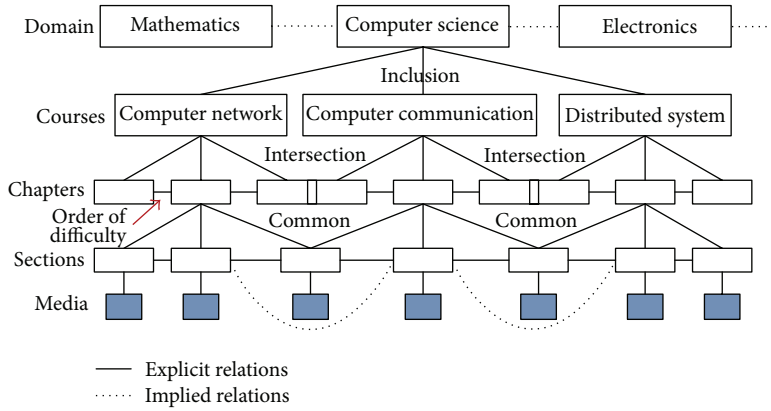


FIGURE 9: The relations of partial media data.

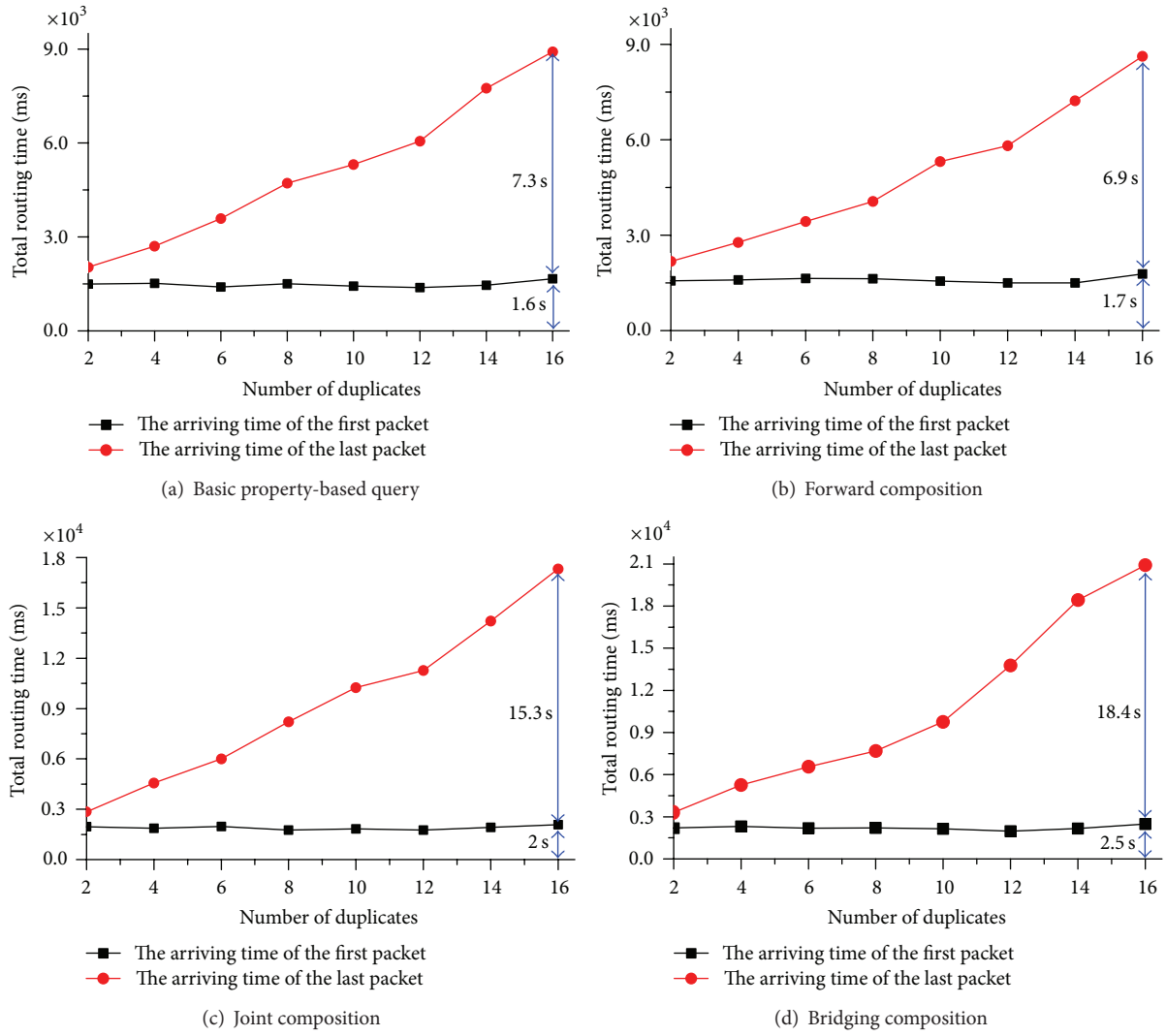


FIGURE 10: Testing results for total routing time with different duplicates (hops are 6).

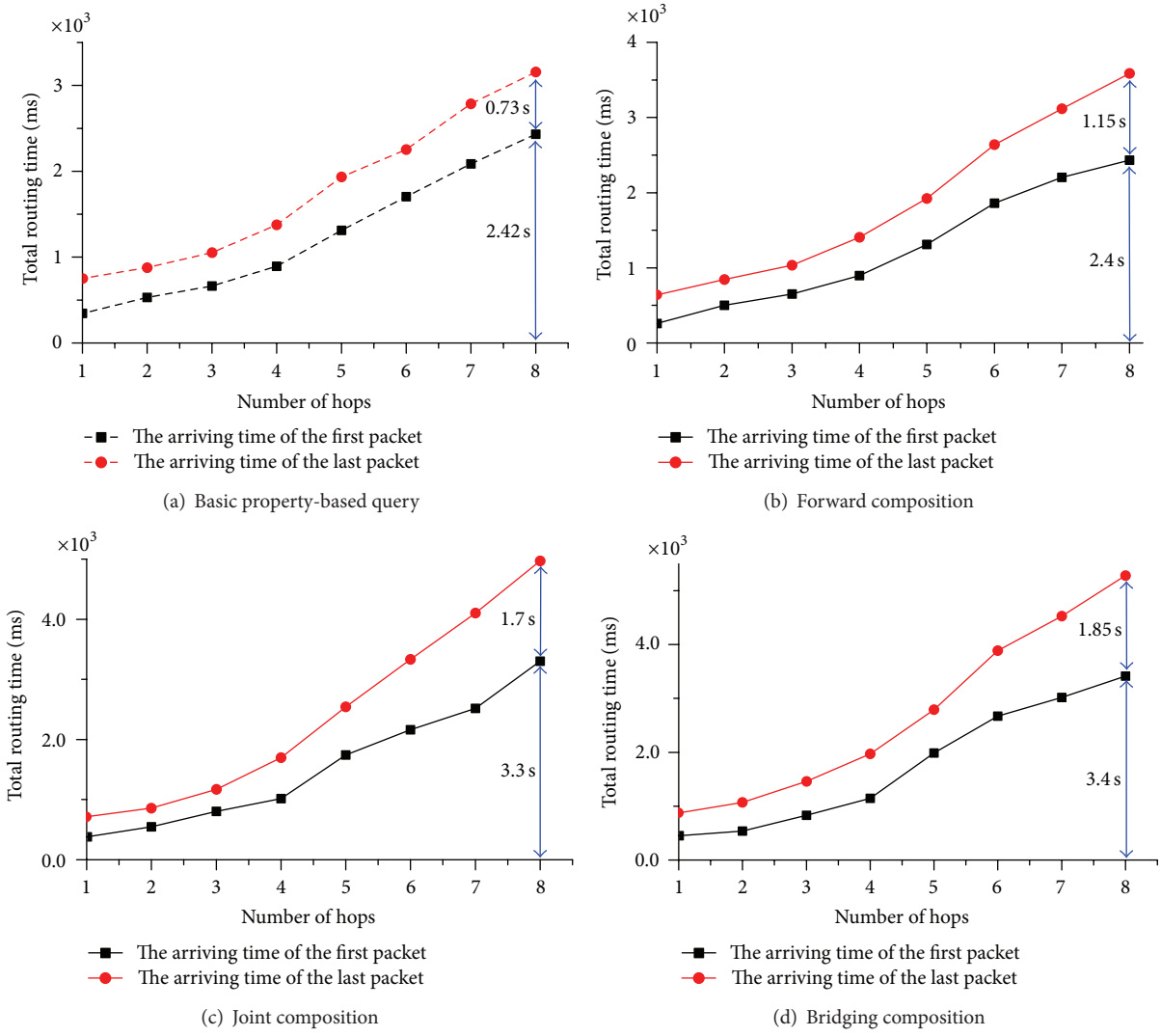


FIGURE 11: Testing results for total routing time with different hops (duplicates are 3).

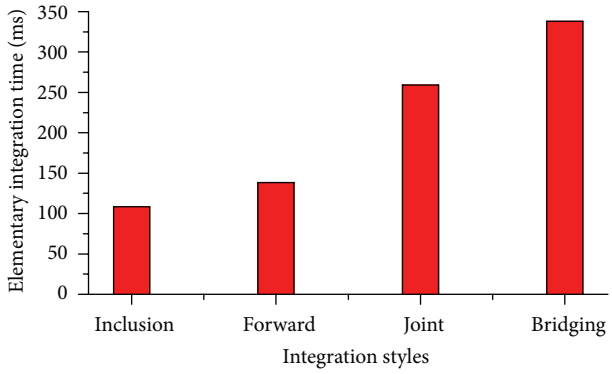


FIGURE 12: Testing results for elementary integration time.

In addition, we assume that users are directly connected to guider nodes rather than to data sources.

In the following, we evaluate the efficiency and cost of our relation routing model with different number of duplicates and different number of hops between the user and the data sources.

5.3.2. Experimental Results. We first look at the routing efficiency of different queries with various duplication factors. Figure 10 shows the total routing time of the queries with property and relation constraints. The query with relation constraints include forward composition, joint composition and bridging composition. The total routing time is the time from sending out an query until the reception of the last response packet.

In the first experiment, we fix the average number of hops from client node to data sources to be 6 and vary the duplication factor from 2 to 16. We repeat each query more than 10 times and average the results. Figures 10(a)–10(d) present the arrival time of the first and last response packets for four queries (basic property based query, forward composition (we omit backward composition here because it is similar to forward composition), joint composition,

and bridging composition), respectively. The arriving time of the first response packet shows that it can retrieve the metadata from closest sources in a very short interval. And the arriving time of the first response keeps being stable as the duplication factor increases. The arriving time of last response packets shows that the total routing time increases slowly as the duplication factor increases. The four kinds of media queries can all be completed in a few seconds even when the duplication factor is 16.

This is not surprising because the sources are distributed, the query and responses are concurrently transmitted over the network, and the number of transmission paths from the user to the duplicates of reference subject is increasing as the duplication increases.

In the second experiment, we fix the duplication factor to be 3 but vary the number of hops from client to data sources from 1 to 8. Figure 11 shows the total routing time for the four kinds of queries mentioned earlier. The results indicate that the arriving times of the first and last response packets both increase slowly as the number of hops increases.

This implies that the complete, decentralized results can be returned to user in a short time.

It should be pointed out that the total query processing time is short for each integration query because the size of relational data, the forwarding time, and the average number of routing hops are all very small in real networks.

Figure 12 presents the elementary integration time of the 4 types of queries. The elementary integration time refers to the average time that is spent for querying data and performing an elementary integration at each data source. The results indicate that the elementary integration time is less than 1 s for these basic integration styles. Finally, the results reveal that the computation complexities are very low and it is larger for the bridging composition than others.

6. Summary

With the development of media technologies, efficient support of complex semantic media queries over large-scale networks will greatly enhance the sharing of large decentralized media and provide a powerful data integration service for users.

For this purpose, we proposed a *relation routing scheme* (RRS) and described its makeups. The semantic-guider is the core component in RRS. By extending the CCN forwarding engine to implement the semantic guider, we inherit many merits from CCN and hope it will have good opportunity to be incrementally deployed. However, RRS and CCN have very different purposes, which makes them differ in several ways. First, RRS adopts a semantic based hierarchy naming convention rather than location or organization based hierarchy naming convention because RRS mainly deals with semantic query. Second, RRS's query format should support complex property and relationship constraints. Thirdly, the 1:*n* relationship between query and responses in RRS results in different timeout strategy of the pending query, the introduction of a new component called RRT to handle incomplete response problem, and the relation integration to condense relevant results and filter redundant ones. Finally, the cache no longer caches the content itself but the relational metadata about the content. Both analytical analysis and experimental studies based on a prototype implementation demonstrated the feasibility and efficiency of this new scheme.

Certainly, it should be acknowledged that there are many possible optimizations that can be made to this scheme, for example, optimization of RRS for high scalability and resilience. Also, the performance evaluation should be carried out to justify the appropriateness of various designing choices and parameter settings, such as the correlation between the query complexity and response time. These are all in our future research agenda.

Acknowledgments

This work was supported in part by the project supported by National Natural Science Foundation of China (Grants nos. 61370227, 61100178, and 61303243), the Hi-Tech Research and Development Program of China under Grant no. 2013AA013503, Hunan Provincial Natural Science Foundation of China (Grant no. 13JJB004, 12JJ3059), and the Beijing Natural Science Foundation (Grant no. 4112057). The authors would like to thank Professor Gregor v. Bochmann of University of Ottawa, Professor Jianxun Liu of Hunan University of Science and Technology, Dr. Di Wu of Sun Yat-Sen University, Dr. Wei Tan of IBM T. J. Watson Research Center, and anonymous reviewers for providing helpful suggestions and comments.

References

[1] S. Agrawal, K. Chakrabarti, S. Chaudhuri, V. Ganti, A. C. Konig, and D. Xin, "Exploiting web search engines to search structured databases," in *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*, pp. 501–510, ACM, New York, NY, USA, 2009.

[2] A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres, and S. Decker, "Searching and browsing Linked Data with SWSE, The Semantic Web Search Engine," *Journal of Web Semantics*, vol. 9, no. 4, pp. 365–401, 2011.

[3] A. Crespo and H. Garcia-Molina, "Routing indices for peer-to-peer systems," in *Proceedings of the 22nd International Conference on Distributed Systems*, pp. 23–32, July 2002.

[4] L. M. Haas, E. T. Lin, and M. A. Roth, "Data integration through database federation," *IBM Systems Journal*, vol. 41, no. 4, pp. 578–596, 2002.

[5] G. Olaf and S. Steffen, "Federated data management and query optimization for linked open data," in *New Directions in Web Data Management*, vol. 331, pp. 109–137, Springer, Berlin, Germany, 2011.

[6] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, "Chord: a scalable peer-to-peer lookup service for internet applications," in *Proceedings of the ACM SIGCOMM Applications, Technologies, Architectures, and Protocols for Computers Communications-*, pp. 149–160, San Diego, Calif, USA, August 2001.

[7] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A scalable content-addressable network," in *Proceedings of ACM SIGCOMM*, San Diego, Calif, USA, 2001.

[8] Gnutella, http://gnutella.wego.com/.

[9] J. N. Al-Karaki and A. E. Kamal, "Routing techniques in wireless sensor networks: a survey," *IEEE Wireless Communications*, vol. 11, no. 6, pp. 6–27, 2004.

[10] M. Meisel, V. Pappas, and L. Zhang, "Ad hoc networking via named data," in *Proceedings of the 5th ACM International Workshop on Mobility in the Evolving Internet Architecture*, pp. 3–8, ACM, New York, NY, USA, 2010.

[11] F. Banaei-Kashani and C. Shahabi, "Swam: a family of access methods for similarity-search in peer-to-peer data networks," in *Proceedings of the 13th ACM Conference on Information and Knowledge Management (CIKM '04)*, pp. 304–313, ACM, New York, NY, USA, November 2004.

[12] X. Sun, "SCAN: a small-world structured p2p overlay for multi-dimensional queries," in *Proceedings of the 16th International World Wide Web Conference (WWW '07)*, pp. 1191–1192, ACM, New York, NY, USA, May 2007.

[13] A. Crespo and H. Garcia-Molina, "Semantic overlay networks for p2p systems," in *Agents and Peer-to-Peer Computing, Lecture Notes in Computer Science*, vol. 3601, pp. 1–13, Springer, Berlin, Germany, 2005.

[14] G. Kokkinidis and V. Christophides, "Semantic query routing and processing in p2p database systems: the ics-forth sqpeer middleware," in *Proceedings of the Current Trends in Database Technology Workshops (EDBT '05)*, vol. 3268 of *Lecture Notes in Computer Science*, pp. 433–436, Springer, Berlin, Germany, 2005.

[15] T. Koponen, M. Chawla, B.-G. Chun et al., "A data-oriented (and beyond) network architecture," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM '07)*, pp. 181–192, ACM, New York, NY, USA, August 2007.

[16] S. T. D. Lagutin and K. Visala, "Publish/subscribe for Internet: psirp perspective," in *Towards the Future Internet CA European Research Perspective*, pp. 75–85, IOS Press, 2010.

[17] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proceedings of the ACM Conference on Emerging Networking Experiments and Technologies (CoNEXT '09)*, pp. 1–12, ACM, New York, NY, USA, December 2009.

[18] A. Anand, F. Dogar, D. Han et al., "XIA: an architecture for an evolvable and trustworthy internet," in *Proceedings of the 10th ACM SIGCOMM Workshop on Hot Topics in Networks (HotNets '10)*, Cambridge, Mass, USA, November 2011.

[19] C. Dannewitz, "NetInf: an information-centric design for the future Internet," in *Proceedings of 3rd GI/ITG KuVS Workshop on the Future Internet*, 2009.

[20] P. Daras, T. Semertzidis, L. Makris, and M. G. Strintzis, "Similarity content search in content centric networks," in *Proceedings of the 18th ACM International Conference on Multimedia ACM Multimedia (MM '10)*, pp. 775–778, ACM, New York, NY, USA, October 2010.

[21] T. Zahariadis, P. Daras, J. Bouwen et al., "Towards a content-centric internet," in *Towards the Future Internet CA European Research Perspective*, pp. 227–236, IOS Press, 2010.

[22] H. Jagadish, L. Lakshmanan, and D. Srivastava, "Hierarchical or relational? a case for a modern hierarchical data model," in *Proceedings of Workshop on Knowledge and Data Engineering Exchange (KDEX '99)*, pp. 3–10, 1999.

[23] G. Karvounarakis, S. Alexaki, V. Christophides, D. Plexousakis, and M. Scholl, "Rql: a declarative query language for rdf," in *Proceedings of the 11th international conference on World Wide Web (WWW '02)*, pp. 592–603, ACM, New York, NY, USA, 2002.

[24] L. Zhang, V. Jacobson, B. Zhang et al., "named data networking (ndn) project," Tech. Rep. parc ndn-0001, 2010.

[25] R. Coltun, "The OSPF Opaque LSA Option," 1998, http://gamay.tools.ietf.org/html/rfc2370.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 656259, 16 pages
http://dx.doi.org/10.1155/2013/656259

Research Article

Design and Analysis of a Dynamic Mobility Management Scheme for Wireless Mesh Network

Abhishek Majumder[1] and Sudipta Roy[2]

[1] *Department of Computer Science & Engineering, Tripura University, Suryamaninagar, 799022 Tripura West, India*
[2] *Department of Information Technology, Assam University, Silchar, 788011 Assam, India*

Correspondence should be addressed to Abhishek Majumder; abhi2012@gmail.com

Received 16 August 2013; Accepted 9 September 2013

Academic Editors: W. Sun and J. Zhou

Seamless mobility management of the mesh clients (MCs) in wireless mesh network (WMN) has drawn a lot of attention from the research community. A number of mobility management schemes such as mesh network with mobility management (MEMO), mesh mobility management (M^3), and wireless mesh mobility management (WMM) have been proposed. The common problem with these schemes is that they impose uniform criteria on all the MCs for sending route update message irrespective of their distinct characteristics. This paper proposes a session-to-mobility ratio (SMR) based dynamic mobility management scheme for handling both internet and intranet traffic. To reduce the total communication cost, this scheme considers each MC's session and mobility characteristics by dynamically determining optimal threshold SMR value for each MC. A numerical analysis of the proposed scheme has been carried out. Comparison with other schemes shows that the proposed scheme outperforms MEMO, M^3, and WMM with respect to total cost.

1. Introduction

Now-a-days wireless mesh network (WMN) [1, 2] has emerged as one of the promising technologies for providing network connectivity to increasing number of mobile users. Moreover, due to advantage of cost effectiveness, robustness, and easy and fast deploy ability; it has become an attractive technology for future network implementation.

WMN consists of three types of nodes: mesh client (MC), mesh router (MR), and gateway (GW). MCs are the mobile users of the WMN. MRs are the wireless routers used for routing of packets from one mesh node to another. An MR having a wired interface to the Internet is called GW. There are two kinds of traffic that flows in the WMN: Internet and Intranet. The Internet packets pass through the GW. The GW receives the downstream Internet packets and sends those to the destination MCs through WMN. In case of upstream Internet traffic, packets are sent from the MC to the GW. On the other hand, Intranet communication takes place between two MCs of same WMN.

One of the major problems in WMN is provisioning of seamless network connectivity for the MCs as it moves from one MR to another. For solving this problem, several mobility management techniques such as MEsh networks with MObility management (MEMO) [3], Mesh Mobility Management (M^3) [4], and Wireless mesh Mobility Management (WMM) [5] have been proposed. MEMO [3] restricts transmission of control message in the WMN to reduce control overhead of the network. But, if the mobility of the MC is high, more numbers of control packets are transmitted by the MCs. So, mobility is an important characteristic of MC. To reduce the control overhead the concept of forward chain has been introduced in M^3. Packets are forwarded through the forward chain. But, in case of a network where session arrival and departure rate to and from the MC are high, large number of packets has to traverse through the forward chain. Thus packet delivery cost increases. So, session arrival and departure rate is also an important issue. WMM uses the concept of forward chain and further takes additional measures to reduce the control overhead and limit the forward chain length. The common problem of the above mentioned schemes is that they are uniform for all the MCs and do not consider the characteristics of an individual MC

while performing its mobility management. So, mobility and session activities of each MC need to be considered for mobility management.

In this paper, a session-to-mobility ratio (SMR) [6] based dynamic mobility management scheme has been proposed. A new SMR calculation scheme is introduced to adapt it in WMN. MC considers both its mobility and the session activity, in the form of SMR, before sending location update to the gateway (GW) and corresponding MRs. Here a threshold SMR value is used, which plays a critical role in the cost of mobility management. Therefore, the optimal threshold SMR value is determined for each individual mesh client dynamically based on the mesh client's specific mobility pattern and session activities.

In this paper an analytical model has been developed to compare the proposed scheme with other baseline schemes such as MEMO, M^3, and WMM. It has been observed that the proposed scheme outperforms the baseline schemes with respect to total signalling cost.

The rest of the paper is organized as follows. Section 2 discusses an overview on relevant mobility management schemes. In Section 3, the SMR based dynamic mobility management scheme has been proposed. The system model and assumptions are discussed in Section 4. The proposed scheme along with some other baseline schemes such as MEMO, M^3 and WMM are numerically analyzed in Section 5. Section 6 presents the performance analysis and comparison among the schemes. Finally, Section 7 presents the conclusion and future work.

2. Related Work

For the purpose of mobility management in WMN, several techniques have been proposed [7, 8]. In this section, some of the existing strategies such as MEMO [3], M^3 [4], and WMM [5] have been discussed.

MEsh networks with MObility management (MEMO) [3] use a modified form of AODV (Ad-hoc on-demand distance vector) protocol, called AODV-MEMO, for integrated routing and mobility management. In this scheme, when the MC moves from one MR to another the new MR proactively sends a route reply to the GW to maintain Internet connectivity. On the other hand for maintaining Intranet connectivity it uses a reactive approach. The old MR of the MC floods route error message in the entire network telling other MRs to delete the outdated MC entry from their routing table. On receiving the route error message the corresponding MRs which still need to communicate with the MC transmit route request message for the MC. After receiving the route request message the new MR of MC sends route reply message to the corresponding MRs. The main drawback of this scheme is its signaling overhead due to flooding of route request and route error messages. This signaling overhead becomes much higher if the MNs of the WMN are highly mobile.

Huang et al. proposed a forward pointer based mobility management scheme named Mesh Mobility Management (M^3) [4]. In this scheme, the GW keeps track of serving MR for each MC. When the MC moves from one MR to another a forward pointer is added from old MR to new one. MC sends location update message to the GW periodically to update its location information in database of the GW. Thus the forward chain is reset. When the GW receives any Internet packet destined to an MC, it searches for the serving MR of the MC in its database. Then it tunnels the packets to the serving MR of the MC. The serving MR forwards the packet to the MR, within whose vicinity the destination MC currently reside (current MC), through the forward pointer. The uplink Internet packets are sent from current MR of the MC to the GW without tunneling. For Intranet communication, the source MC sends the packets to the GW through its current MR and then GW tunnels the packets to the serving MR of destination MC. The serving MR of destination MC handles the Intranet packet the same way as it does with Internet packets. This scheme decreases location update cost but the drawback with this scheme is its periodic location update procedure which makes the entire scheme very much static. In case of high speed MC, the forward chain length will be large and the packet delivery cost will increase drastically if Internet as well as Intranet traffic to the MC is high.

Huang et al. proposed a mobility management scheme called Wireless mesh Mobility Management (WMM) [5]. In this scheme, each mesh node (MN) maintains a routing table and a proxy table. The routing table stores the routing paths between the MNs. The proxy table keeps track of other MCs' location information. No separate message is used by the MCs for location update. Instead of that the IP header of each packet carries the location information of source MC. On receiving the packets, intermediate MNs update their proxy table corresponding to the source MC. Thus WMM scheme does not incur any location update cost. When the MC enters into the vicinity of a new MR, the old MR forwards all the packets, destined to MC, routed to it to the new MR. For routing of packets from source MC to destination MC, MRs use their routing and proxy table. If serving MR of source MC does not know the serving MR of the destination MC, it sends all the packets to the GW. The GW checks whether the MC belongs to the WMN or not. If it does not, the packets are considered as Internet packets and are sent to the wired network. Otherwise, the packets are Intranet packets and after receiving the packets, the GW initiates a query procedure by flooding a query message for the destination MC in the entire network. On receiving response from the destination MC, the GW transmits those packets to the destination. The destination MC updates its proxy table and routing table corresponding to the source MC. Now the destination MC can send packets to the source MC directly (not via GW). The drawback of this scheme is its signaling overhead incurred by the query procedure. Moreover, the characteristics of MCs are not considered to achieve the optimal performance.

The common problem with MEMO, M^3, and WMM is that the schemes do not consider the characteristics of individual MCs for their mobility management rather they use a static approach which is uniform for all MCs.

3. Proposed Scheme

This section presents the proposed mobility management scheme. It uses forward pointer to reduce the number of route update message sent by the MC. To limit the increase in forward chain length, each MC resets the forward chain if its SMR crosses a threshold SMR value. The optimal value for threshold SMR (SMR_{oth}) that minimizes the total communication cost per time unit is dynamically determined for each individual MC. The primary objective of this scheme is to minimize the total cost for mobility management.

In this scheme when the MC joins a WMN, it first gets associated with a nearby MR and sets it as serving MR. Then it sends location update to the GW. The update contains the information about its serving MR. The GW maintains a database recording the serving MRs of all the MCs roaming inside the WMN. On receiving the location update message the GW checks its database whether an entry of the MC is present or not. If there is no entry a new entry of the MC is created. Otherwise, the entry corresponding to the MC is updated. For handling the Intranet traffic each MR maintains a database of the serving MRs of corresponding MCs. There are four major parts in the proposed scheme: calculation of session-to-mobility ratio, calculation of optimal threshold session-to-mobility ratio, mobility management, and routing.

3.1. Session-to-Mobility Ratio (SMR) Calculation. In [6] Pack et al. have defined session-to-mobility ratio (SMR) as the ratio of session arrival rate to mobility rate. A session is a stream of consecutive packet at the IP layer. A timer-based approach is used to identify a session [9]. It is similar to the session management technique used in Universal Mobile Telecommunication System (UMTS) [10]. In this technique, each MC will have an active state timer with length T_A. If time duration between the receiving of two consecutive packets is greater than T_A, the current packet is considered as the first packet of a new session. Otherwise, the packet belongs to the ongoing session. Mobility rate is the MR crossing rate of the MC. In [6] authors have considered only the session arrival rate for computing SMR. But in case of WMN sessions will arrive to as well as depart from the MCs. So, both the factors need to be considered for computing SMR. This technique, considers both session arrival rate to the MC and session departure rate from the MC. The modification in computation of SMR enhances it to capture session and mobility characteristics of the MC more accurately.

3.2. Calculation of Optimal Threshold Session-to-Mobility Ratio. SMR_{oth} can be calculated for each individual user using Bat Algorithm (BA) [11]. This is because BA is superior to many other popular optimization algorithms such as Genetic Algorithm (GA) [12–14] and Particle Swarm Optimization (PSO) [15, 16]. From [4] it can clearly observed that BA requires lesser number of function evaluations for a given tolerance or accuracy than that of GA and PSO. Moreover, for a fixed number of function evaluations the accuracy is higher in case of BA.

When the MC enters into the vicinity of new MR it needs to calculate SMR_{oth}. But, if the value of SMR_{oth} is calculated after every handoff MC has to perform a lot of computation and it is not feasible for MCs with limited battery power. Section 6 presents a detailed discussion over the fitness function used. All the components of the fitness function are either constant or average values of variables that are obtained from continuous measurements by the MC. Instantaneous value of any variable has very limited impact on the fitness function. That is why MC can calculate SMR_{oth} value periodically. This periodic calculation reduces the computational work to be carried out by the MC. Thus, battery power consumption rate of the MC is reduced and at the same time it becomes dynamic.

3.3. Mobility Management. When the MC moves into the vicinity of a new MR it computes its session-to-mobility ratio (SMR_{MC}) and compares it with SMR_{oth}. If SMR_{MC} is less than the SMR_{oth}, the MC notifies the new MR about its handoff from old MR and also sends the addresses of the corresponding mesh nodes. On receiving the notification, new MR informs the old MR about the handoff and enquires about the serving MR of corresponding mesh nodes of the MC. The old MR replies back by sending addresses of serving MRs of corresponding mesh nodes. A forward pointer is also added from old MR to the new one which is also the current MR of the MC. After receiving the reply from old MR, the new MR updates its database. Thus, the forward chain length of the MC increases by 1. On the other hand, if the SMR_{MC} is greater than or equal to SMR_{oth}, same procedure is followed as discussed in the earlier case but no forward chain is added from old MR to new MR, rather the new MR sends location update message to the gateway and the corresponding MRs (if any). When the gateway and corresponding MRs receive the location update message, they search for the entry of the MC in their database and set the current MR as the serving MR of the MC and the forward chain length is reset.

3.4. Routing. In the proposed scheme routing of packets is carried out in two different ways depending on the nature of the packet (Internet or Intranet).

3.4.1. Routing of Internet Packets. Tunneling is used to forward the downstream Internet packets. When the gateway receives a downstream Internet packet, it finds out the serving MR of the destination MC from its database and adds an extra IP header having the address of the serving MR as destination address. This is done because the downstream packets do not have serving MR's address as the destination address and without the serving MR's address intermediate MRs are not able to forward the packet to the serving MR. When the serving MR of the destination MC receives the packet, it decapsulates and if required forwards the packet to current MR through the forward chain. The current MR transmits the packet to the destination MC.

Tunneling is not used in case of upstream packets. Current MR sends the upstream Internet packets received from the MC through the direct route towards the GW.

Figure 1 shows an example scenario representing routing of Internet traffic. Initially the source MC (SMC) was under the vicinity of source serving MR1 (SSMR1). The SMC sent and received its upstream and downstream Internet packets, respectively, through the SSMR1 to the GW. When the SMC moves from SSMR1 to another MR and gets associated with it, the SMC computes its SMR_{MC} value and compares SMR_{oth} with SMR_{MC}. Let its SMR_{MC} value be less than that of SMR_{oth}. The SMC does not send location update to the GW rather a forward pointer is added from SSMR1 to new MR named as source current MR1 (SCMR1). When the GW receives a downstream Internet packets destined at the SMC, it encapsulates the packet with a header having address of SSMR1 as destination address. The GW then sends the packets to SSMR1. On receiving, SSMR1 decapsulates the packet and forwards it to SCMR1 through forward chain. Subsequently the SCMR1 delivers it to the SMC. But the upstream packets are sent directly to the GW without tunneling. In this case, SMC sends the upstream packets through the direct path from SCMR1 to the GW. It does not go through SSMR1. After residing some time in SCMR1 the SMC again changes its point of attachment and gets associated with another MR. The SMC computes its SMR_{MC} and compares SMR_{MC} with SMR_{oth}. Suppose this time its SMR_{MC} is higher than SMR_{oth}. The SMC sends location update message to the GW. This message informs the GW about SMC's association with source serving MR2 (SSMR2). The GW updates SMC's serving MR in its location database. After that all the packets, destined to SMC, are encapsulated by a header having SSMR2's address as destination address and are directly sent to SSMR.

3.4.2. Routing of Intranet Packet. Intranet traffic routing procedure is almost similar to Internet packet routing. But, in this case packets do not go through the GW. MC sends the upstream packets to its current MR. If the packets are part of a continuing session the current MR adds an extra IP header with each packet and the address of corresponding mesh node's serving MR is set as destination address. After that it sends the packet to the corresponding MR directly. On the other hand, if the packets initiate or respond to a new session, before sending the first packet the current MR sends a query message to the GW to know the serving MR of corresponding MC. After getting reply from the GW the current MR updates its database. Now it can follow the normal packet uplink process.

Routing of downstream Intranet packets is straight forward. The serving MR of an MC receives the packets destined to it. Then the MR decapsulates and forwards the packets to the current MR of the MC. The routing of Intranet packets by the intermediate MRs is similar to that of Internet traffic.

Figure 2 shows an example scenario representing routing of Intranet traffic. Let SMC communicate with corresponding MC (CMC). The scenario is similar to that of Internet shown

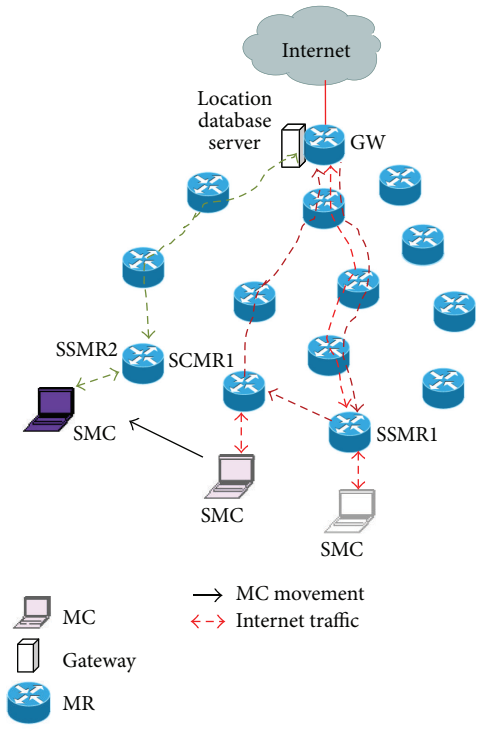


Figure 1: Routing of Internet packet.

in Figure 1. Handling of Intranet traffic is also similar. The only difference is that, here CMC is at the other end of communication instead of GW. When SMC is within the vicinity of SSMR1, it sends upstream packets through SSMR1. SSMR1 adds an IP header to those packets. Address of corresponding serving MR (CSMR) of CMC is set as destination address in the IP header. After receiving the packets the CSMR decapsulates and delivers those to CMC. In case of downlink packets towards SMC, the same process is followed. When the SMC moves to SCMR1 the downlink packets for SMC are received by SSMR1 and forwarded to SCMR1. The uplink packets of SMC are directly sent from SCMR1 to CSMR. When the SMC moves from SCMR1 to SSMR2, SSMR2 handles the uplink and downlink packets in the same way as it was in case of SSMR1.

Figure 3 shows an example representing message communication among the mesh nodes for new session initialization between SMC and CMC. At first SMC sends session request to SCMR. After receiving the session initialization request the SCMR sends a query message to the GW for the address of CSMR of CMR. The GW sends reply to SCMR informing it about CSMR of CMC. Then SCMR sends encapsulated packet to CSMR. CSMR decapsulates the packet and forwards it to corresponding current MR (CCMR). CCMR delivers the packet to CMC. On receiving the request CMC sends response message to CCMR. Again CCMR searches for the SSMR of SMC as SCMR has done for CMR. After that CCMR sends encapsulated packet to SSMR of SMC. SSMR decapsulates the packet and forwards it to SMC through SCMR.

Table 1: Parameters and their interpretations for mathematical modeling and analysis.

Parameter	Interpretation
M	Total number of MRs in the network
α	Average distance (hop count) from an arbitrary MR to the gateway
β	The average distance between two arbitrary MRs
γ	Per hop communication latency
δ_{th}	Threshold value for SMR
λ_p	Average number of packets in a session per time unit
t_{M3}	Time interval between two consecutive location updates in M^3 scheme
N_{active}	Average number of corresponding MCs in the WMN per MC
ω	Rate of reconnection when an MC switches from sleep mode back to active mode
λ_{sc}	Average mobility rate of corresponding MC
λ_{ac}	Average session arrival rate of corresponding MC
λ_{dc}	Average session departure rate of corresponding MC
δ_{thc}	Average threshold value for SMR of corresponding MC
r_{inter}	Percentage of downlink packets per Internet session
r_{intra}	Percentage of downlink packets per Intranet session
I_a	Probability that an arriving session to an MC be an Internet session
I_d	Probability that a departing session from an MC be an Internet session
p_g	Average probability that current MR of MC does not know the location information of destination MC
p_r	Average probability that an MR broadcasts a network control message in its neighborhood
p_q	Average probability that the location query procedure is executed in WMM scheme

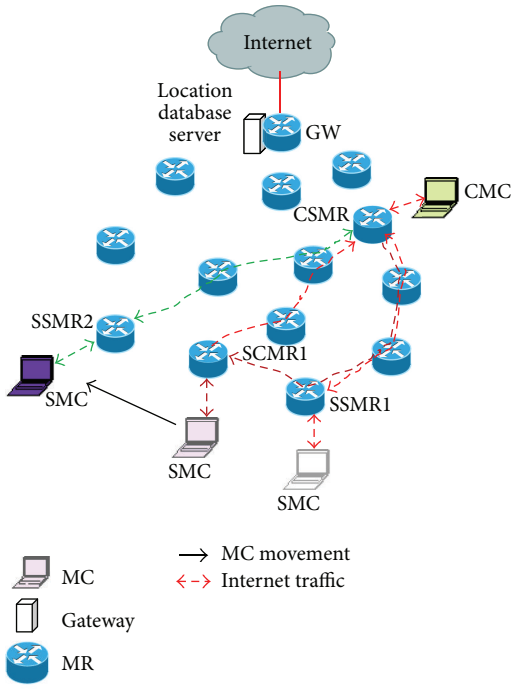


Figure 2: Routing of Intranet packet.

4. System Model and Assumptions

This section describes the system model and assumptions. Without loss of generality the following assumptions can be made.

(i) Incoming and outgoing sessions at an MC occur according to Poisson process with parameter λ_a and λ_d, respectively [17].

(ii) Residence time (t_s) of the MC in the MR follows an exponential distribution with parameter λ_s [18].

The number of occurrences of an event within a time unit follows a Poisson distribution with parameter λ if and only if the time elapsed between two consecutive occurrences of the event has an exponential distribution with parameter λ and it is independent of previous occurrences [19]. It has been assumed that the session arrival and departure rate follow Poisson process with rates λ_a, and λ_d, respectively, and arrival and departure of a session are independent of previous sessions. So, intersession arrival time (t_a) and departure time (t_d) follow exponential distribution with parameters λ_a and λ_d, respectively. Similarly, the number of associations of an MC with MRs in a time unit follows Poisson distribution with rate (mobility rate) λ_s.

In addition to the above mentioned assumptions some parameters are also used for numerical analysis of the proposed scheme and comparison with other baseline schemes. Table 1 shows the parameters used for mathematical modeling and their interpretations.

5. Numerical Analysis

This section presents a numerical analysis of the proposed SMR based scheme, MEMO, M^3, and WMM. Based on the assumptions and system model described in the previous section total communication cost/time unit for each of the schemes has been calculated. Handoff cost, packet delivery

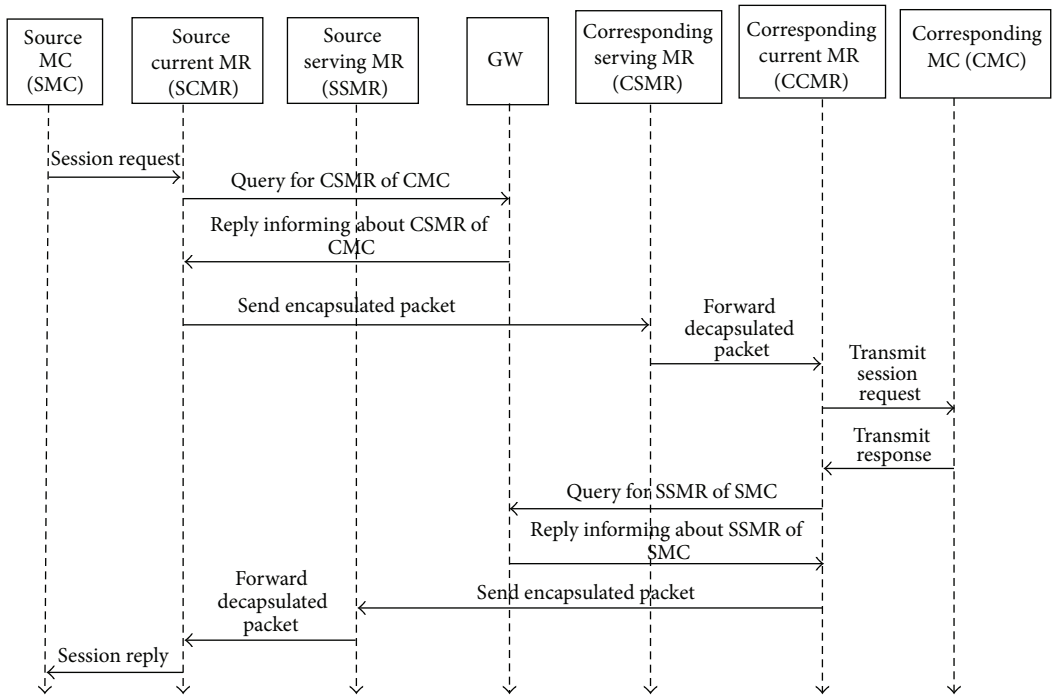


Figure 3: Communication between mesh nodes for session initialization.

cost and query cost per time unit together form the total communication cost/time unit.

Considering both session arrival and session departure $\mathrm{SMR}_{\mathrm{MC}}$ value can be calculated as

$$\mathrm{SMR}_{\mathrm{MC}} = \frac{t_s}{(t_a + t_d)}. \tag{1}$$

Probability that $\mathrm{SMR}_{\mathrm{MC}}$ value be smaller than δ_{th} is calculated as

$$\begin{aligned} P_{\mathrm{th}} &= P\left(\mathrm{SMR}_{\mathrm{MC}} < \delta_{\mathrm{th}}\right) = P\left(\frac{t_s}{(t_a + t_d)} < \delta_{\mathrm{th}}\right) \\ &= P\left(t_s < \delta_{\mathrm{th}}\left(t_a + t_d\right)\right). \end{aligned} \tag{2}$$

Let $t_a = t_1$, $t_d = t_2$, $\lambda_a = \lambda_1$ and $\lambda_d = \lambda_2$. So, $t_a + t_d = t_1 + t_2$. Let $t_1 + t_2 = S_2$. The probability density function of t_1 and t_2 for $t \geq 0$ is given as

$$\begin{aligned} f_{t_a}(t) &= f_{t_1}(t) = \lambda_1 e^{-t\lambda_1}, \\ f_{t_d}(t) &= f_{t_2}(t) = \lambda_2 e^{-t\lambda_2}. \end{aligned} \tag{3}$$

If $\lambda_1 \neq \lambda_2$, from the theorem of convolution of exponential distribution with different parameters, presented in [20], and (3), probability density function of S_2 can be calculated as

$$f_{s_2}(t) = \sum_{i=1}^{2}\left(\frac{\lambda_1\lambda_2}{\prod_{\substack{j=1 \\ i \neq j}}^{2}\left(\lambda_j - \lambda_i\right)} e^{-t\lambda_i}\right). \tag{4}$$

Using (4) on (2) P_{th} can be calculated as

$$\begin{aligned} P_{\mathrm{th}} &= P\left(\mathrm{SMR}_{\mathrm{MC}} < \delta_{\mathrm{th}}\right) \\ &= \int_0^{\infty}\left[P\left(t_s < \delta_{\mathrm{th}}\tau\right) \cdot \sum_{i=1}^{2}\left(\frac{\lambda_1\lambda_2}{\prod_{\substack{j=1 \\ i \neq j}}^{2}\left(\lambda_j - \lambda_i\right)} e^{-\tau\lambda_i}\right)\right] d\tau \\ &= \int_0^{\infty}\left[P\left(t_s < \delta_{\mathrm{th}}\tau\right)\right. \\ &\qquad \left. \cdot \left\{\frac{\lambda_1\lambda_2}{(\lambda_2 - \lambda_1)} e^{-\tau\lambda_1} + \frac{\lambda_1\lambda_2}{(\lambda_1 - \lambda_2)} e^{-\tau\lambda_2}\right\}\right] d\tau \\ &= \frac{\lambda_1\lambda_2}{(\lambda_2 - \lambda_1)}\left[\left\{\int_0^{\infty} P\left(t_s < \delta_{\mathrm{th}}\tau\right) e^{-\tau\lambda_1} d\tau\right\}\right. \\ &\qquad \left. - \left\{\int_0^{\infty} P\left(t_s < \delta_{\mathrm{th}}\tau\right) e^{-\tau\lambda_2} d\tau\right\}\right]. \end{aligned} \tag{5}$$

If $\delta_{\mathrm{th}} \cdot \tau = y$, (5) becomes

$$\begin{aligned} &= \frac{\lambda_1\lambda_2}{\delta_{\mathrm{th}}(\lambda_2 - \lambda_1)}\left[\left\{\int_0^{\infty} P\left(t_s < y\right) e^{-y\lambda_1/\delta_{\mathrm{th}}} dy\right\}\right. \\ &\qquad \left. - \left\{\int_0^{\infty} P\left(t_s < y\right) e^{-y\lambda_2/\delta_{\mathrm{th}}} dy\right\}\right], \end{aligned}$$

$$\int_0^\infty P(t_s < y)\, e^{-y\lambda_1/\delta_{\text{th}}} dy = \int_0^\infty \left(\int_0^y \lambda_s e^{-\lambda_s x} dx \right) e^{-y\lambda_1/\delta_{\text{th}}} dy = \frac{\delta_{\text{th}}^2 \lambda_s}{\lambda_1 (\delta_{\text{th}}\lambda_s + \lambda_1)}. \tag{6}$$

Similarly,

$$\int_0^\infty P(t_s < y)\, e^{-y\lambda_2/\delta_{\text{th}}} dy = \frac{\delta_{\text{th}}^2 \lambda_s}{\lambda_2 (\delta_{\text{th}}\lambda_s + \lambda_2)}. \tag{7}$$

From (5), (6), and (7) the value of P_{th} can be computed as

$$P_{\text{th}} = P(\text{SMR}_{\text{MC}} < \delta_{\text{th}}) = \frac{\lambda_1\lambda_2\lambda_s\delta_{\text{th}}}{\lambda_2 - \lambda_1} \left[\frac{1}{\lambda_1 (\delta_{\text{th}}\lambda_s + \lambda_1)} - \frac{1}{\lambda_2 (\delta_{\text{th}}\lambda_s + \lambda_2)} \right] = \frac{\lambda_a\lambda_d\lambda_s\delta_{\text{th}}}{\lambda_d - \lambda_a} \left[\frac{1}{\lambda_a (\delta_{\text{th}}\lambda_s + \lambda_a)} - \frac{1}{\lambda_d (\delta_{\text{th}}\lambda_s + \lambda_d)} \right]. \tag{8}$$

If $\lambda_1 = \lambda_2 = \lambda$, from the theorem of convolution of exponential distribution with same parameter [20, 21] and (3) it can be written that S_2 follows gamma distribution with parameter $(2, \lambda)$ and has probability distribution function:

$$f_{s_2}(t) = \lambda^2 t e^{-\lambda t}. \tag{9}$$

Using (9) on (2) P_{th} can be calculated as

$$P_{\text{th}} = P(\text{SMR}_{\text{MC}} < \delta_{\text{th}}) = \int_0^\infty \left[P(t_s < \delta_{\text{th}}\tau)\, \lambda^2 e^{-\lambda\tau} \tau \right] d\tau. \tag{10}$$

If $\delta_{\text{th}} \cdot \tau = y$, (10) becomes

$$= \frac{\lambda^2}{\delta_{\text{th}}} \int_0^\infty \left\{ P(t_s < y)\, e^{-(\lambda y/\delta_{\text{th}})} \frac{y}{\delta_{\text{th}}} \right\} dy = \frac{\lambda^2}{\delta_{\text{th}}} \int_0^\infty \left\{ \left(1 - e^{\lambda_s t}\right) e^{-(\lambda y/\delta_{\text{th}})} \frac{y}{\delta_{\text{th}}} \right\} dy = 1 - \frac{\lambda_s^2}{(\lambda_s\delta_{\text{th}} + \lambda)^2}. \tag{11}$$

From (8) and (11) it can be written that

$$P_{\text{th}} = \begin{cases} \dfrac{\lambda_a\lambda_d\lambda_s\delta_{\text{th}}}{\lambda_d - \lambda_a} \times \left[\dfrac{1}{\lambda_a (\delta_{\text{th}}\lambda_s + \lambda_a)} - \dfrac{1}{\lambda_d (\delta_{\text{th}}\lambda_s + \lambda_d)} \right] & \text{if } \lambda_a \neq \lambda_d \\ 1 - \dfrac{\lambda_s^2}{(\lambda_s\delta_{\text{th}} + \lambda)^2} & \text{if } \lambda_a = \lambda_d = \lambda. \end{cases} \tag{12}$$

5.1. Handoff Cost. In MEMO, when the MC moves from one MR to another the old MR broadcasts route error message informing other MRs that the MC has moved out of its vicinity. Since each MR broadcasts the route error message, cost of broadcasting the message is calculated as M. The new MR proactively sends route reply message to the GW. The cost for this route reply is $\alpha \times \gamma$. After Handoff, if the MC wants to continue the communication with any corresponding MC, the new MR broadcast route requests message to all the MRs for searching the corresponding MC. It has the cost of M. The host MR of corresponding MC sends back route reply message costing $\beta \times \gamma$. If the corresponding MC wants to continue, its host MR performs the same procedure and the new MR sends back route reply message. Since there are N_{active} number of corresponding MCs, the cost incurred in this process is $\{N_{\text{active}} \times (M + \beta \times \gamma)\}$. So, for MEMO handoff cost in a time unit is calculated as

$$C_{\text{hMEMO}} = \{M + \alpha \times \gamma + N_{\text{active}} \times (M + \beta \times \gamma)\} \times \lambda_s. \tag{13}$$

In M^3, the MC sends location update message to the GW periodically once in every t_{M^3} time units. Otherwise, a forward pointer is added from old MR to new MR. So, for M^3 the handoff cost in a time unit is calculated as

$$C_{\text{hM}^3} = \left\{ 2 \times \gamma \times \frac{t_{M^3} - 1}{t_{M^3}} + \alpha \times \gamma \times \frac{1}{t_{M^3}} \right\} \times \lambda_s. \tag{14}$$

In WMM the no separate location update message is used but location update is done through the packets sent from source MC. Only a forward pointer is added from old MR to new MR. Handoff cost in a time unit for WMM can be computed as

$$C_{\text{hWMM}} = 2 \times \gamma \times \lambda_s. \tag{15}$$

In SMR based scheme, when the MC moves form vicinity of one MR to another, the MC calculates its SMR_{MC} value, compares it with δ_{th}, and performs the handoff process. If SMR_{MC} is less than δ_{th} a forward pointer is added from old MR to new MR. Otherwise, new MR sends location update message to the GW as well as all corresponding MRs. So, the handoff cost can be calculated as

$$C_{\text{hSMR}} = \begin{cases} 2 \cdot \gamma & \text{if } \text{SMR}_{\text{MC}} < \delta_{\text{th}} \\ (\alpha + N_{\text{active}} \cdot \beta) \cdot \gamma & \text{if } \text{SMR}_{\text{MC}} \geq \delta_{\text{th}}. \end{cases} \tag{16}$$

The handoff cost in a time unit will be

$$C_{\text{hSMR}} = \{2 \times \gamma \times P(\text{SMR}_{\text{MC}} < \delta_{\text{th}}) + (\alpha + N_{\text{active}} \times \beta) \times \gamma \times P(\text{SMR}_{\text{MC}} \geq \delta_{\text{th}})\} \times \lambda_s = \{2 \times \gamma \times P_{\text{th}} + (\alpha + N_{\text{active}} \times \beta) \times \gamma \times P_{\text{th}}\} \times \lambda_s. \tag{17}$$

5.2. Packet Delivery Cost. As the MC roams within the WMN with respect to a starting reference MR, there are two possibilities when it moves from one MR to another. The MC can move to a previously visited MR or a new MR.

If it moves to a previously visited MR, the displacement (hop count) from the reference MR decreases. Otherwise, the displacement increases by 1. The displacement depends on the topology of the WMN and also the mobility pattern of the MC. When the MC moves form one MR to another and gets associated with it, calculation of MN's displacement from the reference MR is very complicated. For simplicity it has been assumed that average displacement (with respect to hop count) of the MC per MR association is c [22]. For example, let a MC move from MR A to MR B. While moving from A to B the MC gets through l number of association with MRs and the distance (hop count) between A and B is k. So, from the assumption it can be written that $k = cl$.

Packet delivery cost incurred by MEMO in a time unit (C_{pMEMO}) consists of downlink Internet packet delivery cost ($C_{\text{pinterdMEMO}}$), uplink Internet packet delivery cost ($C_{\text{pinteruMEMO}}$), and Intranet packet delivery cost ($C_{\text{pintraMEMO}}$):

$$\begin{aligned} C_{\text{pMEMO}} &= C_{\text{pinterdMEMO}} \times \lambda_p \times \lambda_a \times I_a \\ &\quad \times \frac{r_{\text{inter}}}{100} + C_{\text{pinteruMEMO}} \times \lambda_p \times \lambda_a \\ &\quad \times I_a \times \left(1 - \frac{r_{\text{inter}}}{100}\right) + C_{\text{pinterdMEMO}} \times \lambda_p \\ &\quad \times \lambda_d \times I_d \times \frac{r_{\text{inter}}}{100} + C_{\text{pinteruMEMO}} \times \lambda_p \\ &\quad \times \lambda_d \times I_d \times \left(1 - \frac{r_{\text{inter}}}{100}\right) + C_{\text{pintraMEMO}} \\ &\quad \times \lambda_p \times \lambda_a \times (1 - I_a) + C_{\text{pintraMEMO}} \\ &\quad \times \lambda_p \times \lambda_d \times (1 - I_d). \end{aligned} \tag{18}$$

In MEMO the downlink Internet packets of MC are directly sent from the GW to host MR. The host MR then transmits it to MC. On the other hand, in case of uplink packets the reverse procedure is followed. So,

$$C_{\text{pinterdMEMO}} = C_{\text{pinteruMEMO}} = \alpha \cdot \gamma. \tag{19}$$

For Intranet communication the host MR of source MC transmits and receives the packets to and from the host MR of destination MC directly. Thus, Intranet packet delivery cost can be calculated as,

$$C_{\text{pintraMEMO}} = \beta \cdot \gamma. \tag{20}$$

Packet delivery cost/time unit incurred by M^3 in a time unit (C_{pM^3}) consists of downlink Internet packet delivery cost (C_{pinterdM^3}), uplink Internet packet delivery cost (C_{pinteruM^3}), downlink Intranet packet delivery cost (C_{pintradM^3}), and uplink Intranet packet delivery cost (C_{pintrauM^3}):

$$\begin{aligned} C_{\text{pM}^3} &= \lambda_a \times \lambda_p \times I_a \\ &\quad \times \left(C_{\text{pinterdM}^3} \times \frac{r_{\text{inter}}}{100} + C_{\text{pinteruM}^3} \times \left(1 - \frac{r_{\text{inter}}}{100}\right)\right) \\ &\quad + \lambda_d \times \lambda_p \times I_d \\ &\quad \times \left(C_{\text{pinterdM}^3} \times \frac{r_{\text{inter}}}{100} + C_{\text{pinteruM}^3} \times \left(1 - \frac{r_{\text{inter}}}{100}\right)\right) \\ &\quad + \lambda_a \times \lambda_p \times (1 - I_a) \\ &\quad \times \left(C_{\text{pintradM}^3} \times \frac{r_{\text{intra}}}{100} + C_{\text{pintrauM}^3} \times \left(1 - \frac{r_{\text{intra}}}{100}\right)\right) \\ &\quad + \lambda_d \times \lambda_p \times (1 - I_d) \\ &\quad \times \left(C_{\text{pintradM}^3} \times \frac{r_{\text{intra}}}{100} + C_{\text{pintrauM}^3} \times \left(1 - \frac{r_{\text{intra}}}{100}\right)\right). \end{aligned} \tag{21}$$

The cost of transferring the downlink Internet packets from GW to the serving MR of the MC is $\alpha \times \gamma$. The serving MR forwards the packets to the current MR through forward chain. Since average displacement of the MC per MR association is c, the average chain length is $(c \times t_{\text{M}^3} \times \lambda_s)/2$. Thus, the downlink Internet packet delivery cost is

$$C_{\text{pinterdM}^3} = \left(\alpha + \frac{(c \times t_{\text{M}^3} \times \lambda_s)}{2}\right) \times \gamma. \tag{22}$$

The uplink Internet packets are directly sent from current MR to the GW without tunneling. The cost for uplink Internet packet is

$$C_{\text{pinteruM}^3} = \alpha \times \gamma. \tag{23}$$

At first the downlink Intranet packets are sent from the corresponding MC to the GW incurring cost of $\alpha \times \gamma$. The GW then sends the packets to the MC the same way it did with downlink Internet packets which costs $\{(\alpha + (c \times t_{\text{M}^3} \times \lambda_s)/2) \times \gamma\}$. So, downlink packet delivery cost can be calculated as

$$C_{\text{pintradM}^3} = \left(2 \times \alpha + \frac{(c \times t_{\text{M}^3} \times \lambda_s)}{2}\right) \times \gamma. \tag{24}$$

Routing of uplink Intranet packets from current MR of MC to GW has cost of $\alpha \times \gamma$. Then tunneling and followed by forwarding of those packets to the destination MC cost $\{(\alpha + (c \times t_{\text{M}^3} \times \lambda_{sc})/2) \times \gamma\}$. So, uplink packet delivery cost can be calculated as

$$C_{\text{pintrauM}^3} = \left(2 \times \alpha + \frac{(c \times t_{\text{M}^3} \times \lambda_{sc})}{2}\right) \times \gamma. \tag{25}$$

In WMM, packet delivery cost per time unit (C_{pWMM}) consists of downlink Internet packet delivery cost ($C_{\text{pinterdWMM}}$), uplink Internet packet delivery cost

($C_{\text{pinteruWMM}}$), cost for downlink Intranet packet delivery through GW ($C_{\text{pintragdWMM}}$), downstream direct Intranet packet delivery cost ($C_{\text{pintradWMM}}$), cost for upstream Intranet packet delivery through GW ($C_{\text{pintraguWMM}}$), and upstream direct Intranet packet delivery cost ($C_{\text{pintrauWMM}}$):

$$\begin{aligned}
C_{\text{pWMM}} &= C_{\text{pinterdWMM}} \times \lambda_p \times \lambda_a \times I_a \times \frac{r_{\text{inter}}}{100} \\
&+ C_{\text{pinteruWMM}} \times \lambda_p \times \lambda_a \times I_a \times \left(1 - \frac{r_{\text{inter}}}{100}\right) \\
&+ C_{\text{pinterdWMM}} \times \lambda_p \times \lambda_d \times I_d \times \frac{r_{\text{inter}}}{100} \\
&+ C_{\text{pinteruWMM}} \times \lambda_p \times \lambda_d \times I_d \times \left(1 - \frac{r_{\text{inter}}}{100}\right) \\
&+ C_{\text{pintragdWMM}} \times \lambda_p \times \lambda_a \times (1 - I_a) \times p_g \times \frac{r_{\text{intra}}}{100} \\
&+ C_{\text{pintradWMM}} \times \lambda_p \times \lambda_a \times (1 - I_a) \times \left(1 - p_g\right) \\
&\times \frac{r_{\text{intra}}}{100} + C_{\text{pintragdWMM}} \times \lambda_p \times \lambda_d \\
&\times (1 - I_d) \times p_g \times \frac{r_{\text{intra}}}{100} \\
&+ C_{\text{pintradWMM}} \times \lambda_p \times \lambda_d \times (1 - I_d) \times \left(1 - p_g\right) \\
&\times \frac{r_{\text{intra}}}{100} + C_{\text{pintraguWMM}} \times \lambda_p \times \lambda_a \\
&\times (1 - I_a) \times p_g \times \left(1 - \frac{r_{\text{intra}}}{100}\right) \\
&+ C_{\text{pintrauWMM}} \times \lambda_p \times \lambda_a \times (1 - I_a) \times \left(1 - p_g\right) \\
&\times \left(1 - \frac{r_{\text{intra}}}{100}\right) + C_{\text{pintraguWMM}} \times \lambda_p \times \lambda_d \\
&\times (1 - I_d) \times p_g \times \left(1 - \frac{r_{\text{intra}}}{100}\right) \\
&+ C_{\text{pintrauWMM}} \times \lambda_p \times \lambda_d \times (1 - I_d) \\
&\times \left(1 - p_g\right) \times \left(1 - \frac{r_{\text{intra}}}{100}\right).
\end{aligned} \tag{26}$$

In WMM, MC sends its location information with every packet (Internet or Intranet). Between the arrivals of two consecutive packets at an MR, the movement of source MC may result in a number of handoffs and a chain of proxy table entries may be formed. Such chain is similar to forward pointer. The chain associated with source MC gets reset with the arrival of packet from it to the MR. Average time interval between such chain reset operations is the same as the interarrival time of two consecutive packets.

Average rate of Internet packet arrival at the GW from a source MC is

$$\begin{aligned}
\text{AR}_{\text{interg}} &= \lambda_a \times \lambda_p \times I_a \times \left(1 - \left(\frac{r_{\text{inter}}}{100}\right)\right) \\
&+ \lambda_d \times \lambda_p \times I_d \times \left(1 - \left(\frac{r_{\text{inter}}}{100}\right)\right).
\end{aligned} \tag{27}$$

On the other hand, average Intranet packet arrival rate at the GW from the source MC is

$$\begin{aligned}
\text{AR}_{\text{intrag}} &= \lambda_a \times \lambda_p \times (1 - I_a) \times p_g \\
&\times \left(1 - \frac{r_{\text{intra}}}{100}\right) + \lambda_d \times \lambda_p \\
&\times (1 - I_d) \times p_g \times \left(1 - \frac{r_{\text{intra}}}{100}\right).
\end{aligned} \tag{28}$$

The summation of Internet and Intranet packet arrival rate at the GW from the MC will together form the total packet arrival rate. The interarrival time of two consecutive packets is

$$T_{\text{iag}} = \frac{1}{\text{AR}_{\text{interg}} + \text{AR}_{\text{intrag}}}. \tag{29}$$

The average distance an MC can move between two consecutive forward chain reset operations and can be computed as

$$\text{FC}_{\text{gmWMM}} = T_{\text{iag}} \times \lambda_s \times c. \tag{30}$$

So, the downlink Internet packet delivery cost is

$$C_{\text{pinterdWMM}} = \left(\alpha + \text{FC}_{\text{gmWMM}}\right) \times \gamma. \tag{31}$$

Since uplink packets are sent to the GW directly from the current MC of the MR without using any proxy table chain entries, uplink Internet packet delivery cost is

$$C_{\text{pinteruWMM}} = \alpha \times \gamma. \tag{32}$$

The downstream Intranet packets that are sent through the GW will first traverse from the corresponding MC to the GW. This is similar to uplink of Internet packets. Then GW sends those packets to the MC through the forward chain, as it does with downstream Internet packets. So, cost for downlink Intranet packet delivery through GW is

$$C_{\text{pintragdWMM}} = \left(2 \times \alpha + \text{FC}_{\text{gmWMM}}\right). \tag{33}$$

The forward chain length for downstream Internet communication of corresponding MC can be computed by following the same process discussed above. Only λ_s, λ_a, and λ_d will be replaced by λ_{sc}, λ_{ac}, and λ_{dc}, respectively. So the forward chain is

$$\begin{aligned}
\text{FC}^c_{\text{gmWMM}} &= (\lambda_{\text{sc}} \times c) \\
&\times \left(\lambda_{\text{ac}} \times \lambda_p \times I_a \times \left(1 - \frac{r_{\text{inter}}}{100}\right)\right. \\
&+ \lambda_{\text{dc}} \times \lambda_p \times I_d \times \left(1 - \frac{r_{\text{inter}}}{100}\right) \\
&+ \lambda_{\text{ac}} \times \lambda_p \times (1 - I_a) \times p_g
\end{aligned}$$

$$\times \left(1 - \frac{r_{\text{intra}}}{100}\right) + \lambda_{\text{dc}} \times \lambda_p \times (1 - I_d) \times p_g \times \left(1 - \frac{r_{\text{intra}}}{100}\right)\Big)^{-1}. \tag{34}$$

The routing of upstream Intranet packets through GW is the same as downstream Intranet packets through GW. The only difference is that the source MC will directly send the packets to the GW and forward chain will be used at the end of corresponding MC. So, cost for upstream Intranet packet delivery through GW is

$$C_{\text{pintraguWMM}} = \left(2 \times \alpha + \text{FC}^{c}_{\text{gmWMM}}\right). \tag{35}$$

The effective arrival rate of downstream Intranet packets originated from a corresponding MC reaching the MC is

$$\text{AR}_{\text{intrad}} = \frac{\left\{\lambda_p \times \lambda_a \times (1 - I_a) + \lambda_d \times \lambda_p \times (1 - I_d)\right\} \times r_{\text{intra}}}{N_{\text{active}} \times 100}. \tag{36}$$

Average interarrival time between two such consecutive packet is

$$T_{\text{intrad}} = \frac{1}{\text{AR}_{\text{intrad}}}. \tag{37}$$

The average distance that the corresponding MC can move within this time interval is

$$\text{FC}^{c}_{\text{intraWMM}} = T_{\text{intrad}} \times \lambda_{\text{sc}} \times c. \tag{38}$$

Delivery cost of upstream Intranet packets from source MC to corresponding MC through the forward chain is

$$C_{\text{pintrauWMM}} = \left(\beta + \text{FC}^{c}_{\text{intraWMM}}\right) \times \gamma. \tag{39}$$

The effective arrival rate of upstream Intranet packets of source MC reaching a corresponding MC is

$$\text{AR}_{\text{intrau}} = \frac{\left\{\lambda_p \times \lambda_a \times (1 - I_a) + \lambda_d \times \lambda_p \times (1 - I_d)\right\} \times (1 - r_{\text{intra}})}{N_{\text{active}} \times 100}. \tag{40}$$

In this case average interarrival time is

$$T_{\text{intrau}} = \frac{1}{\text{AR}_{\text{intrau}}}. \tag{41}$$

The average distance that the source MC can move within this time interval is

$$\text{FC}_{\text{intraWMM}} = T_{\text{intrau}} \times \lambda_s \times c. \tag{42}$$

Delivery cost of downstream Intranet packets from corresponding MC to source MC through the forward chain is

$$C_{\text{pintradWMM}} = \left(\beta + \text{FC}_{\text{intrawmm}}\right) \times \gamma. \tag{43}$$

In the proposed scheme, when the MC moves to a new MR with respect to the forward chain from serving MR and its $\text{SMR}_{\text{MC}} < \delta_{\text{th}}$, the forward chain length will increase by 1. But if it moves to any MR in the forward chain, the chain length will decrease. Since average displacement of the MC per MR association is c, it can be written that, when the MC moves from one MR to another and its $\text{SMR}_{\text{MC}} < \delta_{\text{th}}$ the forward chain length will increase by c. The upper bound for the forward chain length is $(M - 1)$ and the required number of MR association is $((M - 1)/c)$.

In this scheme packet delivery cost in a time unit (C_{pSMR}) consists of downstream Internet packet delivery cost ($C_{\text{pinterdSMR}}$), upstream Internet packet delivery cost ($C_{\text{pinteruSMR}}$), downstream Intranet packet delivery cost ($C_{\text{pintradSMR}}$), and upstream Intranet packet delivery cost ($C_{\text{pintrauSMR}}$):

$$\begin{aligned} C_{\text{pSMR}} = {} & C_{\text{pinterdSMR}} \times \lambda_p \times \lambda_a \times I_a \times \frac{r_{\text{inter}}}{100} \\ & + C_{\text{pinteruSMR}} \times \lambda_p \times \lambda_a \times I_a \times \left(1 - \frac{r_{\text{inter}}}{100}\right) \\ & + C_{\text{pinterdSMR}} \times \lambda_p \times \lambda_d \times I_d \times \frac{r_{\text{inter}}}{100} \\ & + C_{\text{pinteruSMR}} \times \lambda_p \times \lambda_d \times I_d \times \left(1 - \frac{r_{\text{inter}}}{100}\right) \\ & + C_{\text{pintradSMR}} \times \lambda_p \times \lambda_a \times (1 - I_a) \times \frac{r_{\text{intra}}}{100} \\ & + C_{\text{pintrauSMR}} \times \lambda_p \times \lambda_a \times (1 - I_a) \times \left(1 - \frac{r_{\text{intra}}}{100}\right) \\ & + C_{\text{pintradSMR}} \times \lambda_p \times \lambda_d \times (1 - I_d) \times \frac{r_{\text{intra}}}{100} \\ & + C_{\text{pinteruSMR}} \times \lambda_p \times \lambda_d \times (1 - I_d) \times \left(1 - \frac{r_{\text{intra}}}{100}\right). \end{aligned} \tag{44}$$

The GW will send the downlink Internet packets to the serving MR and the serving MR will forward these packets to the MC through the forward chain. Let $P(i)$ be the probability that forward chain length of the MC is $i \times c$ and $C_{\text{pinterdSMR}}(i)$ be the downstream Internet packet delivery cost with that forward chain length. The cost per downstream packet delivery can be calculated as

$$\begin{aligned} C_{\text{pinterdSMR}} = {} & \sum_{i=0}^{(M-1)/c} C_{\text{pinterdSMR}}(i) \cdot P(i) \\ = {} & (1 - P_{\text{th}}) \cdot \alpha \cdot \gamma + (1 - P_{\text{th}}) \cdot P_{\text{th}} \cdot (\alpha + c) \cdot \gamma \\ & + (1 - P_{\text{th}}) \cdot P_{\text{th}}^2 \cdot (\alpha + 2c) \cdot \gamma + \cdots + (1 - P_{\text{th}}) \\ & \cdot P_{\text{th}}^{((M-1)/c)-1} \cdot (1 - P_{\text{th}}) \end{aligned}$$

$$\cdot \left(\alpha + \left(\frac{M-1}{c} - 1\right) \cdot c\right) \cdot \gamma + P_{\text{th}}^{(M-1)/c} \cdot (\alpha + (M-1)) \cdot \gamma = \alpha \cdot \gamma + \gamma \cdot c \cdot P_{\text{th}} \cdot \left[\frac{1 - P_{\text{th}}^{((M-1)/c)}}{(1 - P_{\text{th}})}\right]. \quad (45)$$

As a check it can be noted that

$$\sum_{i=0}^{(M-1)/c} P(i) = P\left(\bigcup_{i=0}^{((M-1)/c)-1} N = i\right) + P^{(M-1)/c} = \left((1 - P_{\text{th}}) \cdot \sum_{i=0}^{((M-1)/c)-1} P_{\text{th}}^{i}\right) + P_{\text{th}}^{(M-1)/c} = 1. \quad (46)$$

The upstream packets from the MC will not be tunneled by serving MR rather will directly be sent to the GW. So, cost per upstream packet delivery is

$$C_{\text{pinteruSMR}} = \alpha \cdot \gamma. \quad (47)$$

Current MR of corresponding MC will tunnel downstream Intranet packets of MC to its serving MR through β number MRs. Then serving MR forwards the packets to MC through forward chain. So, calculation of downstream Intranet packet delivery cost is similar to that of downstream Internet packets:

$$\begin{aligned} C_{\text{pintradSMR}} &= \sum_{i=0}^{(M-1)/c} C_{\text{pintradSMR}}(i) \cdot P(i) \\ &= (1 - P_{\text{th}}) \cdot \beta \cdot \gamma + (1 - P_{\text{th}}) \cdot P_{\text{th}} \cdot (\beta + c) \cdot \gamma \\ &\quad + (1 - P_{\text{th}}) \cdot P_{\text{th}}^2 \cdot (\beta + 2c) \cdot \gamma \\ &\quad + \cdots + (1 - P_{\text{th}}) \cdot P_{\text{th}}^{((M-1)/c)-1} \cdot (1 - P_{\text{th}}) \\ &\quad \cdot \left(\beta + \left(\frac{M-1}{c} - 1\right) \cdot c\right) \cdot \gamma + P_{\text{th}}^{(M-1)/c} \\ &\quad \cdot (\beta + (M-1)) \cdot \gamma \\ &= \beta \cdot \gamma + \gamma \cdot c \cdot P_{\text{th}} \cdot \left[\frac{1 - p_{\text{th}}^{((M-1)/c)}}{(1 - P_{\text{th}})}\right]. \end{aligned} \quad (48)$$

In case of corresponding MCs, let the probability that its SMR value is smaller than δ_{thc} be P_{thc}. Calculation of P_{thc} is similar to (8)

$$P_{\text{thc}} = \frac{\lambda_{\text{ac}} \lambda_{\text{dc}} \lambda_{\text{sc}} \delta_{\text{thc}}}{\lambda_{\text{dc}} - \lambda_{\text{ac}}} \times \left[\frac{1}{\lambda_{\text{ac}}(\delta_{\text{thc}} \lambda_{\text{sc}} + \lambda_{\text{ac}})} - \frac{1}{\lambda_{\text{dc}}(\delta_{\text{thc}} \lambda_{\text{sc}} + \lambda_{\text{dc}})}\right]. \quad (49)$$

Upstream Intranet packets are tunneled from the current MR of source MC to serving MR of destination MC. Then the serving MR of corresponding MC forwards the packets to it. The cost for upstream Intranet packets is

$$C_{\text{pintrauSMR}} = \beta \cdot \gamma + \gamma \cdot c \cdot P_{\text{thc}} \cdot \left[\frac{1 - p_{\text{thc}}^{((M-1)/c)}}{(1 - P_{\text{thc}})}\right]. \quad (50)$$

5.3. Query Cost. In MEMO, when the source MC sends request to its host MR for initialization of Intranet session and if the host MR does not have a route to the destination MC it broadcasts a route request message which costs $P_r \times M$. On receiving the route request the host MR of destination MC sends back route reply costing $\beta \times \gamma$. So, query cost for MEMO in a time unit is

$$C_{\text{qMEMO}} = (P_r \times M + \beta \times \gamma) \times (1 - I_d) \times \lambda_d \times P_g. \quad (51)$$

In M^3, MC does not send any query message for Internet or Intranet communication because all the packets are sent through the GW.

In WMM, GW will execute location query procedure only when it has packets to send to an MC before the MC initiates the first Internet session, after the MC joins a WMN or wakes up and reconnects to the WMN after staying in sleep mode for some time. Let ω_w and ω_s be switching rate from sleep mode to active mode and the reverse mode switching rate, respectively. As discussed in [5] the reconnection rate of MC is

$$\omega = \frac{1}{1/\omega_w + 1/\omega_s} = \frac{\omega_w \times \omega_s}{\omega_s + \omega_s}. \quad (52)$$

For searching the MC, GW broadcasts route request message to all the MRs of the WMN which costs $P_r \times M$. The serving MR of corresponding MC sends back response message costing $\alpha \times \gamma$. So, query cost in a time unit is

$$C_{\text{qWMM}} = (P_r \times M + \alpha \times \gamma) \times P_q \times \omega. \quad (53)$$

As Figure 3 shows, in the proposed scheme for initialization of every departing session, the current MR of source MC sends query message to the GW for locating destination MR. The GW sends back the information of serving MR of destination MC. From the end of destination MC, the current MR of destination MC will also send such query message to the GW for sending the reply packets to source MC. These four message transfers will cost $4 \times \alpha \times \gamma$. The query cost for this scheme in a time unit is

$$C_{\text{qSMR}} = 4 \times \alpha \times \gamma \times \lambda_d. \quad (54)$$

5.4. Total Communication Cost. Total communication cost in a time unit for MEMO, M^3, WMM, and the proposed SMR based scheme is denoted by TC_{MEMO}, TC_{M^3}, TC_{WMM}, and TC_{SMR}, respectively. So, it can be written that

$$\begin{aligned} \text{TC}_{\text{MEMO}} &= C_{\text{hMEMO}} + C_{\text{pMEMO}} + C_{\text{qmemo}}, \\ \text{TC}_{\text{M}^3} &= C_{\text{hM}^3} + C_{\text{pM}^3}, \\ \text{TC}_{\text{WMM}} &= C_{\text{hWMM}} + C_{\text{pWMM}} + C_{\text{qWMM}}, \\ \text{TC}_{\text{SMR}} &= C_{\text{hSMR}} + C_{\text{pSMR}} + C_{\text{qSMR}}. \end{aligned} \quad (55)$$

6. Fitness Function for Optimization

Each MC will periodically compute optimal threshold SMR value SMR_{oth} using bat algorithm and will be set as δ_{th}. The objective of finding the SMR_{oth} is to minimize total communication cost/time unit (TC_{SMR}). The fitness function for the objective can be obtained by using (17), (44), (45), (47), (48), (50), and (55). So, the fitness function is

$$\begin{aligned} f(\delta_{th}) = {} & \{2 \cdot \gamma \cdot P_{th} + (\alpha + N_{active} \cdot \beta) \cdot \gamma \cdot P_{th}\} \\ & \cdot \lambda_s + \gamma \cdot c \cdot P_{th} \cdot \frac{\lambda_p}{100} \left(\frac{1 - P_{th}^{((M-1)/c)}}{1 - P_{th}} \right) \\ & \cdot [r_{inter} (\lambda_a \cdot I_a + \lambda_d \cdot I_d) \\ & + r_{intra} (\lambda_a \cdot (1 - I_a) + \lambda_d \cdot (1 - I_d))]. \end{aligned} \quad (56)$$

7. Performance Analysis

In this section, the performances of MEMO, M^3, and WMM are compared with the proposed scheme in terms of handoff cost, packet delivery cost, and total communication cost per time unit under different mobility rate, session arrival rate, and session departure rate. To compare the cost incurred by any of the three schemes (C_{other}) with that of proposed SMR based scheme (C_{SMR}) the following metric is used

$$F\% = \left(\frac{|C_{other} - C_{SMR}|}{C_{SMR}} \right) \times 100. \quad (57)$$

If $C_{other} > C_{SMR}$, it can be said that the cost incurred by the other schemes is $F\%$ higher compared to the SMR based scheme. Otherwise, the cost of the other scheme is $F\%$ less compared to the SMR based scheme. The variation in the value of SMR_{oth} with the change in λ_s, λ_a, and λ_d is also analyzed. Table 2 shows the default values of the parameters used in the numerical analysis. Here second is considered as time unit. To find the optimal threshold SMR value (SMR_{oth}) for minimum total communication cost/time unit bat algorithm is used with population size 100, number of generation 50, loudness 100, and pulse rate 0.5.

Figure 4 shows the variation in handoff cost/sec of MEMO, M^3, WMM, and the proposed SMR based scheme with respect to the increase in λ_a. In this case, the values of λ_s and λ_d are assumed to be 0.2 and 0.5, respectively. In MEMO after every handoff the old MR of MC broadcasts route error message in the entire network and new MR sends route reply message to the GW proactively. Thus it incurs high handoff cost/sec. In M^3 location update messages are sent by the MC to the GW after a static time interval. Therefore, the MC has to send location update message less frequently and handoff cost/sec becomes less compared to MEMO. In case of WMM, MC does not send any location update message rather location information is carried by each data packet originated from the MC. So, the handoff cost/sec is the lowest among all other schemes. In the proposed scheme, the MC sends location update message to the GW and the host MRs of corresponding MCs only when its

Table 2: Default parameter values.

Symbol	Value
γ	0.1
λ_p	200
α	50
M	1000
P_r	0.5
β	50
P_q	0.1
P_g	0.5
t_{M3}	1200
C	0.4
N_{active}	30
I_a	0.8
I_d	0.7
r_{inter}	0.8
r_{intra}	0.5
ω_w	1/1200
ω_s	1/600
λ_{sc}	0.3
λ_{ac}	0.05
λ_{dc}	0.5
δ_{thc}	0.36

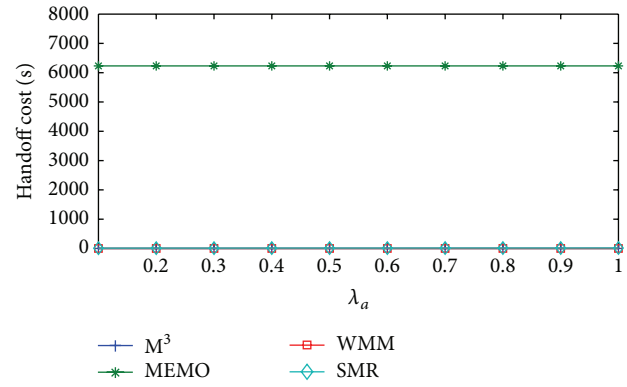


Figure 4: Handoff cost/sec versus λ_a.

SMR is greater than or equal to δ_{th}. This leads to less frequent location updates. Handoff cost/sec of MEMO, M^3, and WMM is constant even if λ_a increases. This is because the frequency of sending location update message of the three schemes remains unchanged with the increase in λ_a. On the other hand, in SMR based scheme, as λ_a increases SMR value of MC increases and frequency of SMR value crossing δ_{th} also increases. Thus handoff cost/sec increases with the increase in λ_a but the rate of increase is very little. Therefore, average handoff cost/sec of MEMO is 45,588.35% higher compared to the proposed SMR based scheme. On the other hand, compared to the SMR based scheme M^3 and WMM incur 99.70% and 99.71% less average handoff cost/sec, respectively.

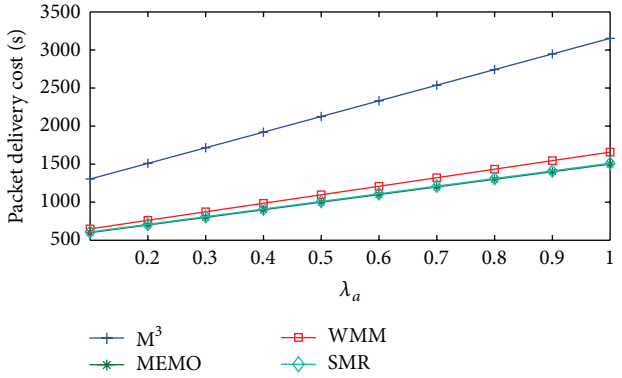


Figure 5: Packet delivery cost/sec versus λ_a.

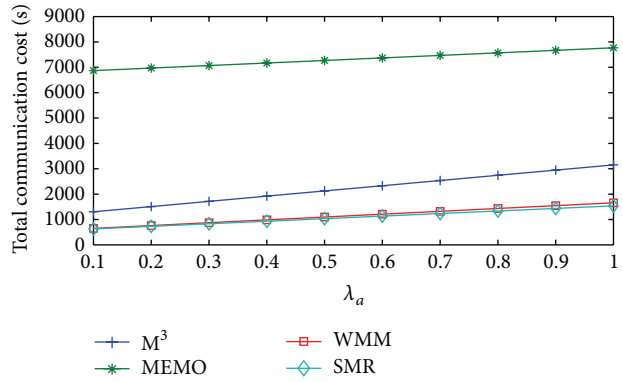


Figure 6: Total communication cost/sec versus λ_a.

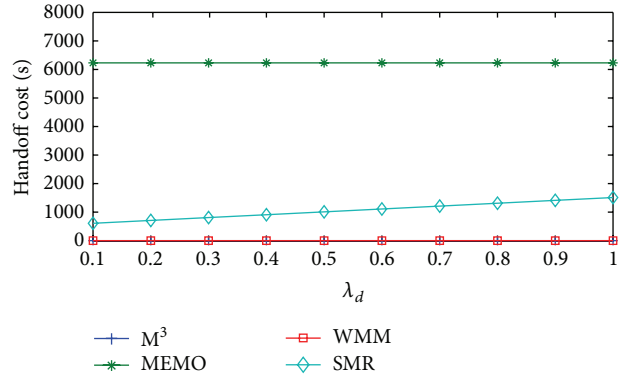


Figure 7: Handoff cost/sec versus λ_d.

Figure 5 shows the change in packet delivery cost/sec of MEMO, M^3, WMM, and SMR based scheme with respect to the increase in λ_a. MEMO has the lowest packet delivery cost/sec since packets are directly delivered to the current MR of MC and no chain for packet forwarding is used. In M^3, the packets are forwarded through long forward chain. The forward chain length keeps growing and gets reset after some fixed time interval. This results in the highest packet delivery cost/sec in M^3. In WMM a chain of proxy table entries is formed due to the mobility of MC within the time interval between two consecutive packet arrivals originated from the MC to the GW or corresponding MC. The packets have to go through the forward chain which is of significant length. In the proposed SMR based scheme because of the selection of optimal threshold SMR value packets have to traverse through the forward chain of optimal length. The packet delivery cost/sec of the MEMO increases with increase in λ_a but it is due to the increased number of packet transfer/sec. As λ_a increases, packet delivery cost/sec of M^3 also increases because more numbers of packets are forwarded through the long forward chain. In WMM because of increase in λ_a though the forward chain resets will be frequent but more number of packets has to go through the forward chain. That is why the packet delivery cost/sec increases as λ_a increases. In SMR based scheme λ_a is incorporated in the calculation of SMR and optimal value for δ_{th} is selected dynamically. As a result the packet delivery cost/sec remains close to that of MEMO with the increase in λ_a. Therefore, compared to SMR based scheme M^3 and WMM have 110.16% and 8.8% higher average packet delivery cost/sec, respectively. On the other hand, MEMO incurs 0.99% less average packet delivery cost/sec compared to the SMR based scheme.

Figure 6 shows the change in total communication cost/sec of MEMO, M^3, WMM, and SMR based scheme with respect to the increase in λ_a. MEMO has the highest total communication cost/sec. This is due to its huge location update cost/sec and significant query cost/sec. Total communication cost/sec of M^3 is lower than MEMO because M^3 incurs much lesser handoff cost/sec than MEMO. Handoff cost/sec of WMM is the minimum but it incurs a significant amount of packet delivery cost/sec. Because of this, its total communication cost/sec is higher than SMR based scheme. In case of the proposed scheme due to the incorporation of λ_a in SMR calculation and dynamic selection of optimal δ_{th}, total communication cost/sec is the least among all the schemes discussed. So, compared to SMR based scheme, MEMO, M^3, and WMM incur 575.11%, 105.59%, and 6.43% higher average total communication cost/sec, respectively.

Change in handoff cost/sec, packet delivery cost/sec, and total communication cost/sec of the four schemes with respect to λ_d are shown in Figures 7, 8, and 9, respectively. Here, the values of λ_s and λ_a are assumed to be 0.2 and 0.5, respectively. With the increase in λ_d handoff cost/sec and packet delivery cost/sec of all the four schemes will show the similar behavior as it was in case of λ_a. The reasons for that are also similar. Average handoff cost/sec of MEMO is 45596.39% higher than SMR based scheme. But compared to SMR based scheme, M^3 and WMM incur 99.70% and 99.71% less average handoff cost/sec, respectively. On the other hand, average packet delivery cost/sec of MEMO is 0.99% less than SMR based scheme. But, in comparison with SMR based scheme, M^3 and WMM has 101.83% and 8.58% higher average packet delivery cost/sec, respectively. Other than, handoff cost/sec and packet delivery cost/sec, query cost/sec is also another factor that will affect total communication cost/sec of MEMO and SMR based scheme with the increase in λ_d. In case of MEMO, if an MC wants to initiate a session and its host

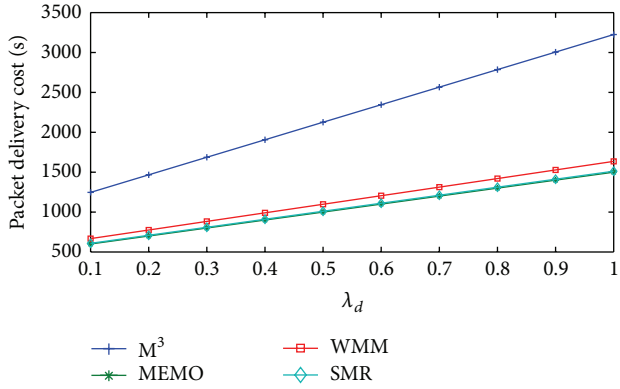


Figure 8: Packet delivery cost/sec versus λ_d.

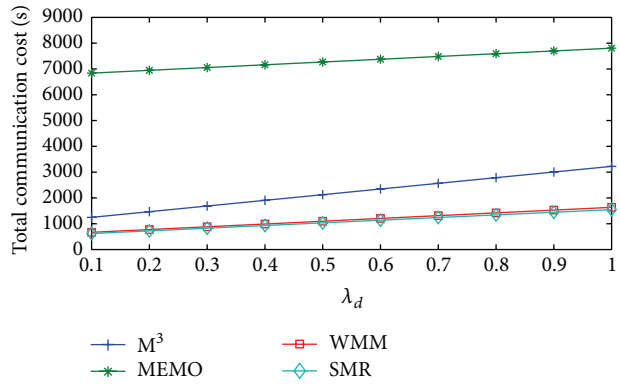


Figure 9: Total Communication cost/sec versus λ_d.

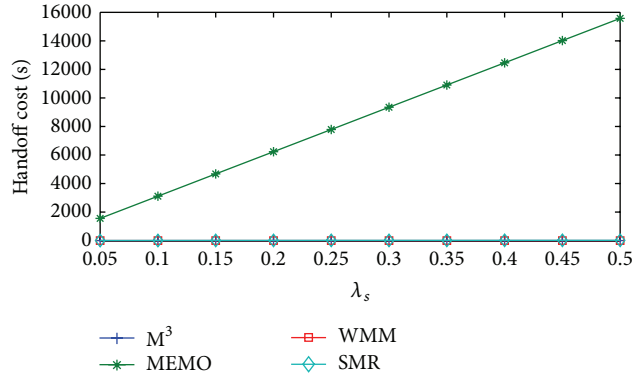


Figure 10: Handoff cost/sec versus λ_s.

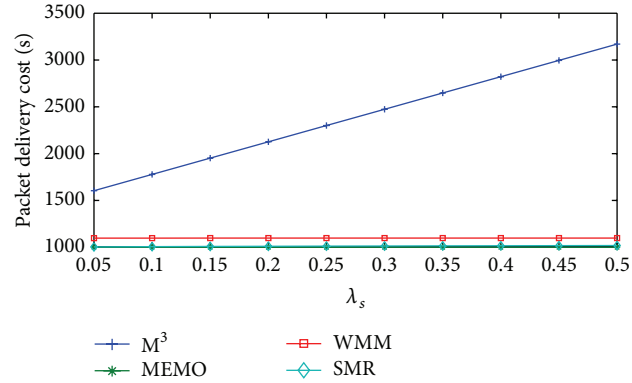


Figure 11: Packet delivery cost/sec versus λ_s.

MR does not have a route to the destination, the host MR of the MC broadcasts query message to all the MRs and GW of WMN. This cost increases as λ_d increases and thus total communication/sec will also increase. In case of the SMR based scheme, before initiating a session the current MR of the MC sends query message to the GW and as λ_d increases, query cost/sec will increase. So, total communication cost/sec will increase as λ_d increases but it will be the least due to dynamic selection of optimal threshold SMR value. Average total communication cost/sec of MEMO, M^3, and WMM is 574.8%, 106.04%, and 6.12% higher than SMR based scheme, respectively.

Figure 10 shows the change in packet delivery cost/sec of MEMO, M^3, WMM, and SMR based scheme with respect to λ_s. Here, both λ_a and λ_d are assumed to be 0.5. In MEMO, if λ_s increases, number handoff/sec will also increase and the host MR has to send and receive more route request and route reply. So, handoff cost/sec will increase at higher rate as λ_s increases. The handoff cost/sec of M^3 and WMM will not increase much as λ_s increases. This is because in M^3 periodic location update messages are sent and in WMM no explicit location update is sent. In case of SMR based scheme with the increase in λ_s there is slight increase handoff cost/sec because of dynamic selection of optimal threshold value δ_{th}. But the rate of increase is very small. Average handoff cost/sec of MEMO is 56604.93% higher than SMR based scheme. On the other hand, compared to SMR based scheme M^3 and WMM incur 99.63% and 99.64% less average handoff cost/sec, respectively.

Figure 11 shows the change in packet delivery cost/sec of MEMO, M^3, WMM, and SMR based scheme with respect to λ_s. MEMO does not use any forward pointer, so with increase in λ_s the packet delivery cost/sec will be constant. In M^3 as λ_s increases, the MC performs more number of handoffs between two consecutive location updates. Thus the packets have to go through longer forward chain and packet delivery cost/sec will increase. Packet delivery cost/sec of WMM increases a little as λ_s increases. This is because here session arrival and departure rate is assumed to be constant. It results in a constant but small time interval between two consecutive location updates and with the increase in λ_s the packets will traverse little larger forward chain. In SMR based scheme the value of δ_{th} is adjusted dynamically. Though λ_s increases, the forward chain does not increase much and as a result there will be very little increase in packet delivery cost/sec. Compared to the SMR based scheme, M^3 and WMM have 135.85% and 8.47% higher average packet delivery cost/sec, respectively. But MEMO has 1.2% less average packet delivery cost/sec than SMR based scheme.

Figure 12 shows the change in packet total communication cost/sec of MEMO, M^3, WMM, and SMR based scheme with respect to λ_s. Total communication cost/sec

incurred by MEMO will increase at the highest rate as λ_s increases. This is because of its highest increase rate of handoff cost/sec. The increase in total communication cost/sec of M^3 is higher than WMM and SMR based scheme as λ_s increases. Highest increase rate of packet delivery cost/sec incurred by M^3 is the main reason behind this. WMM's total communication cost/sec remains higher than SMR based scheme as λ_s increases. This is because it has higher packet delivery cost/sec and significant amount of query cost/sec. In SMR based scheme both handoff cost/sec and packet delivery cost/sec are optimal, so the total communication cost/sec of SMR remains the least as λ_s increases. Compared to SMR based scheme average total communication cost/sec of MEMO, M^3, and WMM is 826.09%, 130.14%, and 5.85% higher, respectively.

Two components of total communication cost/sec: handoff cost/sec and packet delivery cost/sec, are opposite to each other. It means, if for an MC the earlier increases the later will decrease and vice versa. Thus both of the costs need to be optimized so that total communication cost/sec is minimized. The scheme was proposed to reduce total communication cost/sec. Optimal threshold SMR (δ_{oth}) value is determined dynamically and it is set as δ_{th} for minimum total communication cost/sec. Figure 13 shows the effect of change in λ_a and λ_s on δ_{oth}. If the value of λ_a increases SMR value will also increase. Initially optimal threshold SMR value (δ_{oth}) increases with increase in λ_a to avoid drastic decrease in forward chain length causing very high handoff cost/sec. At higher values of λ_a, δ_{oth} decreases as λ_a increases. It helps to keep the packet delivery cost/sec as minimum because it has become more important component of total communication cost/sec with higher λ_a. On the other hand, if λ_s increases more handoffs will take place and to reduce high handoff cost/sec longer forward chain is desirable. But too long forward chain will result in higher packet delivery cost/sec. If the value of λ_s increases SMR value will decrease. For small value of λ_s optimal threshold SMR value (δ_{oth}) increases with increase in λ_s. This reduces frequency of location update operation. For small value of λ_s, less frequency of location update operation will not cause drastic increase in forward chain length. So, both handoff cost/sec and packet delivery cost/sec are balanced such that total communication cost/sec is the minimum. In later section of the figure, when λ_s is higher, δ_{oth} decreases as λ_s increases. This is to avoid very long forward chain which in turn reduces packet delivery cost/sec. Thus minimizing total communication cost/sec. Figure 14 shows effect of change in λ_d and λ_s on δ_{oth}. From the figure it can be observed that the change in δ_{oth} due to variation in λ_d and λ_s follows the same trend as in case of Figure 13 and the reasons for that are also similar.

8. Conclusion and Future Work

In this paper, a SMR based mobility management scheme has been proposed. The proposed scheme is per user based. Each MC periodically calculates optimal value of threshold SMR for minimum total communication cost/sec. The value of δ_{th} is adjusted dynamically based on that optimal threshold SMR value.

An analytical model for performance evaluation of the proposed scheme has been developed. The proposed scheme is also compared with MEMO, M^3, and WMM. Analytical results show that the proposed scheme performs better than the rest of the three schemes. The variation in average forward chain length with respect to increase in session arrival rate, session departure rate, and mobility rate is also analyzed in this paper.

Investigation on the possibility of extending the proposed scheme for WMNs with more than one GWs remains as future work. In addition exploration possibility for use of caching in the proposed scheme to reduce signaling overhead also remains as future work.

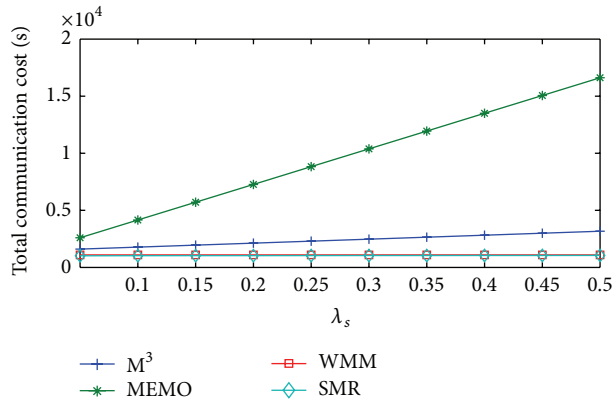


Figure 12: Total communication cost/sec versus λ_s.

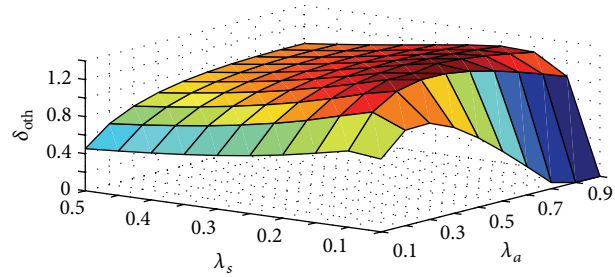


Figure 13: Effect of λ_s and λ_a on δ_{oth}.

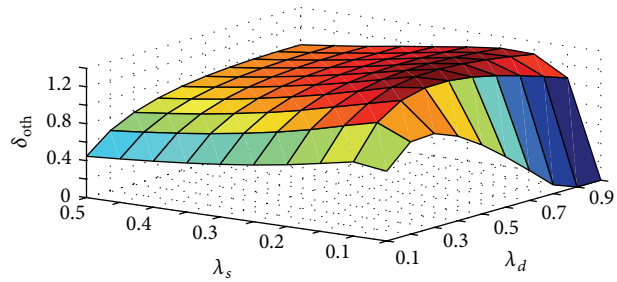


Figure 14: Effect of λ_s and λ_d on δ_{oth}.

References

[1] I. F. Akyildiz and X. Wang, "A survey on wireless mesh networks," *IEEE Communications Magazine*, vol. 43, supplement 9, pp. S23–S30, 2005.

[2] I. F. Akyildiz, X. Wang, and W. Wang, "Wireless mesh networks: a survey," *Computer Networks*, vol. 47, no. 4, pp. 445–487, 2005.

[3] M. Ren, L. Chao, Z. Huizhou, Z. Tong, and Y. Wei, "MEMO: an applied wireless mesh network with client support and mobility management," in *Proceedings of the 50th Annual IEEE Global Telecommunications Conference (GLOBECOM '07)*, pp. 164–170, November 2007.

[4] R. Huang, C. Zhang, and Y. Fang, "A mobility management scheme for wireless mesh networks," in *Proceedings of the 50th Annual IEEE Global Telecommunications Conference (GLOBECOM '07)*, pp. 5092–5096, November 2007.

[5] D. Huang, P. Lin, and C. Gan, "Design and performance study for a mobility management mechanism. (WMM) using location cache for wireless mesh networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 5, pp. 546–556, 2008.

[6] S. Pack, T. Kwon, and Y. Choi, "A performance comparison of mobility anchor point selection schemes in Hierarchical Mobile IPv6 networks," *Computer Networks*, vol. 51, no. 6, pp. 1630–1642, 2007.

[7] K. Mase, "Layer 3 wireless mesh networks: mobility management issues," *IEEE Communications Magazine*, vol. 49, no. 7, pp. 156–163, 2011.

[8] J. Xie and X. Wang, "A survey of mobility management in hybrid wireless mesh networks," *IEEE Network*, vol. 22, no. 6, pp. 34–40, 2008.

[9] X. Zhang, J. G. Castellanos, and A. T. Campbell, "P-MIP: paging extensions for mobile IP," *Mobile Networks and Applications*, vol. 7, no. 2, pp. 127–141, 2002.

[10] Y. Lin, A. Pang, and H. C. Rao, "Impact of mobility on mobile telecommunications networks," *Wireless Communications and Mobile Computing*, vol. 5, no. 8, pp. 713–732, 2005.

[11] X. Yang, "A new metaheuristic Bat-inspired algorithm," *Studies in Computational Intelligence*, vol. 284, pp. 65–74, 2010.

[12] D. Goldberg, *Genetic Algorithms in Search, Optimiztion and Machine Learning*, Addison-Wesley Longman, Reading, Mass, USA, 1989.

[13] J. H. Holland, *Adapation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, Mich, USA, 1975.

[14] M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, Mass, USA, 1998.

[15] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the 1995 IEEE International Conference on Neural Networks*, vol. 4, pp. 1942–1948, December 1995.

[16] J. Kennedy and R. Eberhart, *Swarm Intelligence*, Academic Press, New York, NY, USA, 2001.

[17] S. Kwon, S. Y. Nam, H. Y. Hwang, and D. K. Sung, "Analysis of a mobility management scheme considering battery power conservation in IP-based mobile networks," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 6, pp. 1882–1890, 2004.

[18] S. Pack, H. Jung, T. Kwon, and Y. Choia, "SNC: a selective neighbor catching scheme for fast handoff in 802. 11 wireless networks," *ACM SIGMOBILE Mobile Computing and Communication Review*, vol. 9, no. 4, pp. 39–49, 2005.

[19] M. Taboga, "Poisson distribution," in *Lectures on Probability Theory and Mathematical Statistics*, pp. 350–352, Amazon CreateSpace, 2012.

[20] M. Akkouchi, "On the convolution of exponential distribution," *Journal of the Chungcheong Mathematical Society*, vol. 20, no. 4, pp. 501–510, 2008.

[21] S. M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists*, Elsevier Academic Press, Amsterdam, The Netherlands, 2004.

[22] A. Majumder, S. Roy, and K. K. Dhar, "Design and analysis of an adaptive mobility management scheme for handling traffic in wireless mesh network," in *Proceedings of the International Conference on Microelectronics, Communications and Renewable Energ (ICMiCR '13)*, pp. 1–6, June 2013.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 757182, 8 pages
http://dx.doi.org/10.1155/2013/757182

Research Article

Smart Caching Based on Mobile Agent of Power WebGIS Platform

Xiaohui Wang,[1] Kehe Wu,[2] and Fei Chen[2]

[1] *Postdoctoral Mobile Research Station of Management Science and Engineering, North China Electric Power University, Beijing 102206, China*
[2] *School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China*

Correspondence should be addressed to Xiaohui Wang; wangxiaohui.cn@hotmail.com

Received 22 August 2013; Accepted 12 September 2013

Academic Editors: G. Zhang and J. Zhou

Power information construction is developing towards intensive, platform, distributed direction with the expansion of power grid and improvement of information technology. In order to meet the trend, power WebGIS was designed and developed. In this paper, we first discuss the architecture and functionality of power WebGIS, and then we study caching technology in detail, which contains dynamic display cache model, caching structure based on mobile agent, and cache data model. We have designed experiments of different data capacity to contrast performance between WebGIS with the proposed caching model and traditional WebGIS. The experimental results showed that, with the same hardware environment, the response time of WebGIS with and without caching model increased as data capacity growing, while the larger the data was, the higher the performance of WebGIS with proposed caching model improved.

1. Introduction

As the state grid corporation "three sets of five" (intensive management in human resources; financial resources; and material resources; large-scale movements in programming, construction, operation, overhaul, and production) system construction plan is put forward, the power information construction develops toward intensive, platform, distributed shared direction. As the foundation platform, GIS should be placed in the first place during information construction. And in order to realize spatial information sharing and interoperability, the demand of integrating WebGIS in power information system is increasingly urgent [1].

Power WebGIS platform could integrate all types of equipment belonging to power enterprise, which contains power equipment, substation, transmission and substation network, power users and power load, production and management, and other core business, and form composite management system to meet the requirements (safe, reliable, high-quality, efficient, and economic operation) of power enterprise and its customers. Plan and manage the power grid by using modern technology and management means; enhance power grid equipment asset management, operation management, and regulatory capacity; improve the power supply reliability and power quality; and provide high quality, efficient, and safe service for electricity customers timely [2].

The caching technology is a common mean to solve the WebGIS space data access efficiency. In WebGIS platform, the main impact of spatial data access efficiency consists of two parts: first, database access efficiency and second, the transmission efficiency of the network [3]. Therefore, considering from the two aspects: database and network, we designed dedicated power WebGIS sharing platform cache mode land; it can greatly improve the WebGIS platform transaction processing ability and the response speed of the map. Literature 4 addresses performance issues in a systematic manner from the aspects of architecture, bottlenecks and performance factors, performance improving techniques, and performance solutions [4]. Literature 5 proposed WebGIS framework based on multiagents and made a more extensive and in-depth research on related technical problems, which overcome the defect of the traditional WebGIS in Internet environment [5]. Luo et al. introduces a multilevel componentized WebGIS system, named Geo-Union, and designed

the space cache framework. It is divided into three levels: spatial database caching and network spatial caching and spatial data proxy server [6]. Literature 7 proposed a prefetching algorithm called Retrospective Adaptive prefetching, which takes the former actions of the user into consideration, to reduce the user-perceived response time and to improve user's navigation efficiency [7].

To solve spatial data access, efficiency problem existed in the process of electric power construction of WebGIS; *i* used to put forward spatial data accelerate engine in literature 8. The spatial data accelerate engine consists of spatial data model, spatial index method, and spatial index caching mechanism [8].

This paper is organized as follows: first it provides a brief description of independent research and development of electric power WebGIS platform architecture, and then it presents dynamic display cache model and agent-based dynamic display cache method, finally it verifies the proposed method; finally, conclusion and further research direction are given.

2. Power WebGIS Platform

Power WebGIS platform is designed based on SOA architecture and is aimed at achieving power enterprise labor and material resources intensive management and business centralized operation, using spatial data sharing and business systems integration as the starting point, to construct a service-oriented, comprehensive, and real time power system comprehensive information integration platform.

Besides the general function and the performance of WebGIS platform, Power WebGIS platform also needs to meet the requirement of electric power industry. Such as integrating electric power business data, providing power characteristic service and providing secondary development function to support related business systems integration at the same time; power WebGIS platform architecture is shown in Figure 1.

The data service layer is at the bottom of power WebGIS platform; it contains all kinds of data mentioned in the spatial data model such as spatial index for spatial data and database-side vector data cache and the background map tile server.

The application services layer is the core of the power WebGIS platform. It contains most application logic of WebGIS and various types of data cache Web server for client service agent server spatial information server (SGA server), Various types of functions are provided as services for the client calls. Unlike general WebGIS, we establish three types of caching service at the layer of the application logic, consisting metadata cache, the permissions cache, and data cache (background map cache and vector graphics data cache). The cache uses a unique design and can greatly improve the speed of response of the WebGIS. The cached data can also be shared between agent server, coordinated distributed deployment servers, and clients.

Platform sharing layer interacts directly with the user, in order to enhance the human-computer interaction capacity of WebGIS and improve client's performance. On the basis of general browser, based on Flex RIA technologies, one can realize the map control and caching mechanism. At the same time in order to reduce the network load and customer service network latency and improve WebGIS strength and fault tolerance, designed multiagent structure in Flex RIA achieves synergies in the client and speed up map response.

By using Flex technology and agent technology, power WebGIS platform consummate power business integration application system, while improve the user experience at the same time. In addition, Power WebGIS platform establishs two types of cache: cache separately for user data (vector data cache and metadata cache) and background map tile cache at display interface layer, which completes the maps collaborative function and cache data sharing between clients by moving agent. User interface control layer functions involve map display, map edit, resources display, resource query, resource location, spatial analysis, thematic map display, and so on.

3. Caching Model and Its Replacement Algorithm

3.1. Dynamic Display Caching Model. Dynamic display cache model is a Caching model set up basing on power GIS spatial data model, making database, GIS server, the client as the research object and caching sharing data in the way of memory and file. Its goal is to improve the response speed of the power WebGIS sharing platform and speed up the map rendering efficiency, rich client space operation level.

The formal description of the Dynamic Display Cache Model is as follows.

(1) Cache Directory (CD): CD $= \{M, F, T\}$, where M is a collection of metadata cache, F is a collection of facilities feature cache, and T is the set of topological relations cache, the CD is dynamic display cached content, and the following will descript the specific content of each directory in detail.

(2) The metadata cache set Metadata (M): $M = \{B, \text{CS}, \text{SL}, \text{FM}, \text{RS}, V\}$

(i) Boundary (B): the scope of the map;

(ii) Coordinate Reference System (CS): map reference coordinate system, including all standard coordinate system defined in EPSG, and supports custom extensions;

(iii) Symbol Library (SL): the platform vector symbol library, for punctate facilities rendering;

(iv) Feature Model (FM): facility model FM = {FNO, F type, F style}, where FNO is facilities model number used to identify a type of facility, F type is the type of facility, possible values (point, line, surface, body, multipoint, multiline, multifaceted, and complex body), and F style is the facilities layer rendering style;

(v) Relation Set (RS) is a set of possible relationships between the facilities, RS = {RNO, R type, Rf_1, Rf_2}, where RNO is relationship identification

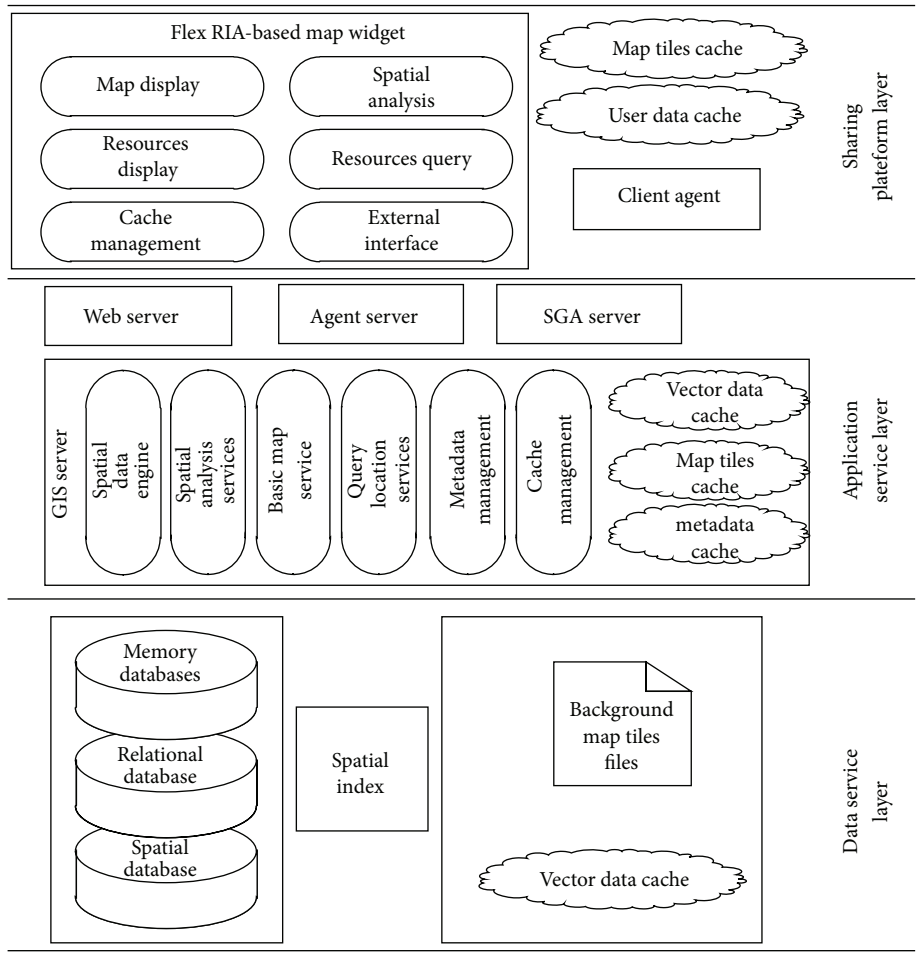


Figure 1: Power WebGIS architecture.

number, R type is the type of relationship, Rf_1 is the first facility number involved in the relationship, and Rf_2 is second facility number involved in the relationship, if it is one-scale relationship, $Rf_2 = 0$;

(vi) Version (V) is the metadata cached version.

(3) Facilities feature cache set Feature (F): $F = \{F_0, F_1, F_2, \ldots, F_n\}$, where $F_i \in F \cap F_i \odot FM$, one of All facilities category, which means instance, the F is an instance of FM. $F_i = \{f_{i0}, f_{i1}, f_{i2}, \ldots, f_{im}\}$, $f_{ij} \in F_i$, is a concrete class facilities.

(4) Topological relations cache set Topology (T): $T = \{t_0, t_1, t_2, \ldots, t_n\}$, $t_i \in T$, stand for a specific topological relations.

(5) Cache Level (CL): CL = $\{cl_1, cl_2, cl_3\}$, where cl_1 stand for level 1 cache, namely, database cache; cl_2 stand for the second level cache, namely, the GIS server cache; cl_3 stand for Level cache, namely, the client cache.

(6) Cache Mode (CM): CM = {cm_1 and cm_2, cubic CM}, where cm_1 stand for memory cache whose cache efficiency is high and is difficult to share and maintain; cm_2 stand for file cache which is difficult to share synchronized copy; cm_3 stand for memory file caching, has advantages of the two cache mentioned above, to maintain cache shared and cache consistency.

DDC model are defined as follows:

$$\mathrm{DDC} = \mathrm{CD} \times \mathrm{CL} \times \mathrm{CM} = \{M, F, R\} \times \mathrm{CL} \times \mathrm{CM}. \quad (1)$$

The creation and updating of cache is closely related to the version number stored in the metadata, and low-level cache relies on high level cache, take level 3 caches for example, to demonstrate the process of cache updating

if ($\mathrm{CD}(cl_3)$ is null), then create (CD_i, cl_3, cm_1);

else if ($V(cl_3) = V(cl_2)$), then updateCache (cl_3),
else dropCache (cl_3),
create (CD_i, cl_3, cm_1),

where $\mathrm{CD}(cl_3)$ represent the client cache directory, create (CD_i, cl_3, cm_1) is to create the client cache directory, cache mode uses the memory cache, $V(cl_3)$ stand for the version number of the client cache, updateCache (cl_3) is to update the client cache, and dropCache (cl_3) is to empty the client cache.

Cache model creation and maintenance should ensure the integrity and consistency of cache. In order to realize the cache efficiency at the same time, some rules need to be followed.

Rule 1. CD must guarantee the integrity, convenient for local client to render graphics, and support simple topology analysis and spatial query.

Rule 2. Cache level CL can support arbitrary combination, which contains eight kinds of state, but often maintain the whole cache state so as to ensure high efficiency of the system.

Rule 3. CM only has single state, can only be one of the following states: memory cache, file cache, or memory file cache. Usually adhere to the following pattern: first carries on the memory cache, when the amount of cache increase to set limit, based on sliding window replacement strategy, transfer a part of the memory cache to memory file cache. When the client is closed, transferred into the file cache.

3.2. Caching Replacement Algorithm. Caching replacement algorithm is established with space limitation and time limitation as theoretical basis. Space limitation behaved as if the most remote distance eliminated first, and time limitation behaved as if the longest time unvisited eliminated first. Comparing with the traditional FIFO, LRU, and LFU, we chose 2Q (two queues algorithm) to improve the efficiency of caching replacement.

2Q algorithm does not eliminate the page least visited from main cache but achieves through swapping with the page most visited. Similar with LRU/2 algorithm, 2Q distinguishes the pages with the time visited the second time, that is to say, 2Q puts the page first visited into a special cache called A_1 queue, which is a FIFO queue, and move it to A_m queue if the page is visited again in the life cycle, which is LRU, likewise, swap it out if it is not visited in the life cycle of A_1 queue.

In 2Q algorithm, A_1 queue exists as a filter, the data could swap to A_m queue when it is visited again after entering the $A_{1\,\mathrm{out}}$ queue. Assuming miss rate is m and 1 out queue space is f, so the data could be swaped to A_m queue if it is visited less than f/m after it entered the $A_{1\,\mathrm{out}}$ queue, we called the probability p_{accept}. Assuming the rate of object i visited is p_i, then the rate that it was swapped to A_m from $A_{1\,\mathrm{out}}$ after the visited k time is $p_i(1-p_i)^{k-1}$, and p_{accept} could be achieved as

$$p_{\mathrm{accept}} = \sum_{k=1}^{f/m} p_i \left(1-p_i\right)^{k-1} = 1-\left(1-p_i\right)^{f/m}. \tag{2}$$

p_{accept} is close to 1 as p_i increased, while close to 0 as p_i decreased, so $A_{1\,\mathrm{out}}$ plays a role of filtering the page least visited. We could define the cutoff hotness of $A_{1\,\mathrm{out}}$ filter as p_{cutoff}, in the case of $p_{\mathrm{accept}} = 1/2$, the derivative of p_{cutoff} is very large, so the object with the visiting rate p_{cutoff} could be swapped into A_m after it was visited the third time not the second time; the visiting rate could be achieved according to (3). If the f and m are given, p_{cutoff} could be expanded as (4) with Taylor series expansion:

$$1-\left(1-p_i\right)^{f/m} = \frac{1}{2}, \tag{3}$$

$$p_{\mathrm{cutoff}} = 1-\left(\frac{1}{2}\right)^{f/m} \approx \frac{m\ln 2}{f}. \tag{4}$$

Formula (4) shows that $A_{1\,\mathrm{out}}$ filter could self-adjusted, p_{accept} will be relatively large when miss rate is large, so the process will swap the object most visited, while the object usual visited could be swapped in when miss rate is low. Assuming p_i and m are given, f_{crit} could be achieved in the condition of p_{accept}, as

$$f_{\mathrm{crit}} = \frac{-m\ln 2}{\ln\left(1-p_i\right)} \approx \frac{m\ln 2}{p_i}. \tag{5}$$

The established condition of (5) is $\ln(1-p_i) \approx -p_i$, and f could be achieved when visited rate limit of A_m queue is given. Assign the value of $f = f_{\mathrm{approx}}$, average visited object could be swapped into A_m queue. If miss rate is m and size of A_m is B, the average visited rate of object in the buffer is $(1-m)/B$, and formula (6) could be achieved when replacing f_{crit} with f:

$$f_{\mathrm{approx}} = \frac{\ln(2)\, mB}{1-m}. \tag{6}$$

The practical significance of (6) is limited as the miss rate need to be estimated, but it illustrates that the ratio of $A_{1\,\mathrm{out}}$ size and buffer size is B. Simulation results showed that the ratio is 1/2 when the value of m is from 10% to 90%, and $A_{1\,\mathrm{out}}$ size is smaller if the m is smaller. Assuming D as the whole data set and $B = rD$, A_m is part of D, shown as (7), then the average visited rate of the object least visited is $m/(1-r)D$. Assuming $X = m/D$, the average visited rate of the object least visited is $(1/(1-r))X$, which is

$$p_{\mathrm{cutoff}} = \frac{2m\ln(2)}{rD}, \tag{7}$$

$$p_{\mathrm{cutoff}} = \frac{2\ln(2)}{r}X. \tag{8}$$

Because r is very small, so $(2\ln(2)/r)X \gg 1/(1-r)$. The object least visited could be filtered as long as r is small enough with the condition $f = B/2$; moreover, $f_{\mathrm{approx}}/B = 1/2$ could result $m = 1/(1+2\ln(2)) \approx 0.419$. As long as m is not large enough, the fluctuation of f_{approx} will not be large, so the algorithm is not sensitive to the setting of f, and $f = B/2$ is always right.

4. Dynamic Display Cache Based on the Agent

Power GIS model has the following characteristics: complex, large amount of data, high real-time requirements, history play back, and real-time tracking operations. So multilevel cache on spatial data can effectively reduce the server pressure and network load, achieve efficient and real-time access to

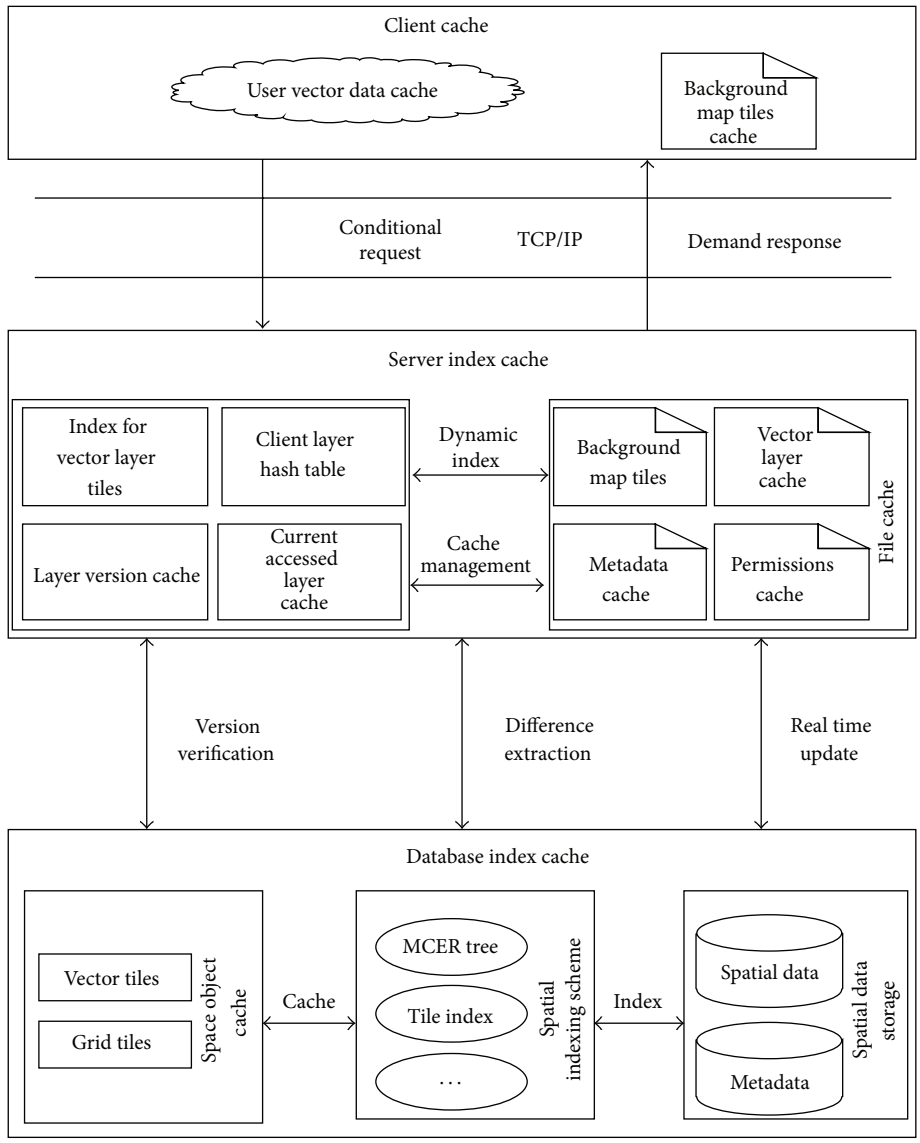


Figure 2: Power WebGIS dynamic display cache.

spatial data. It is also critical to improve the efficiency of the system and reduce the map response time [9].

In order to quickly and effectively deal with huge amounts of spatial data stored in the space database, we implement three-level cache in Power WebGIS platform including the database-side, server-side, and client. The platform dynamic cache model is shown in Figure 2. Create a spatial index and tile cache on database-side and improve query efficiency by querying attribute data instead of the space data. Through the management, the vector cache server, and the map tile server, server-side caches dynamic map layers vector data, user data and layer version data using high-speed dynamic display cache (DDC). Server-side also caches background map in the form of tile data cache. Through a distributed deployment strategy, it reduces network traffic, speeds up the access speed, greatly improves the concurrent user response ability, and realizes high stability of the cache. Through the intelligent agent technology, the client communicates with server-side cache by adopting asynchronous transfer technology, set different cache according to the access level and access frequency, realize sensitive interaction and rich operation experience, improve the efficiency of response.

Server-side dynamic display high-speed cache, and background map tiles cache can satisfy power facilities data real-time change and meet vector graphics rendering requirements. The system can quickly generate maps without query background database so it still has high efficiency while managing huge amounts of data at the same time. The flexible cache can be placed anywhere in the network, considering the system performance, storage space, and network traffic.

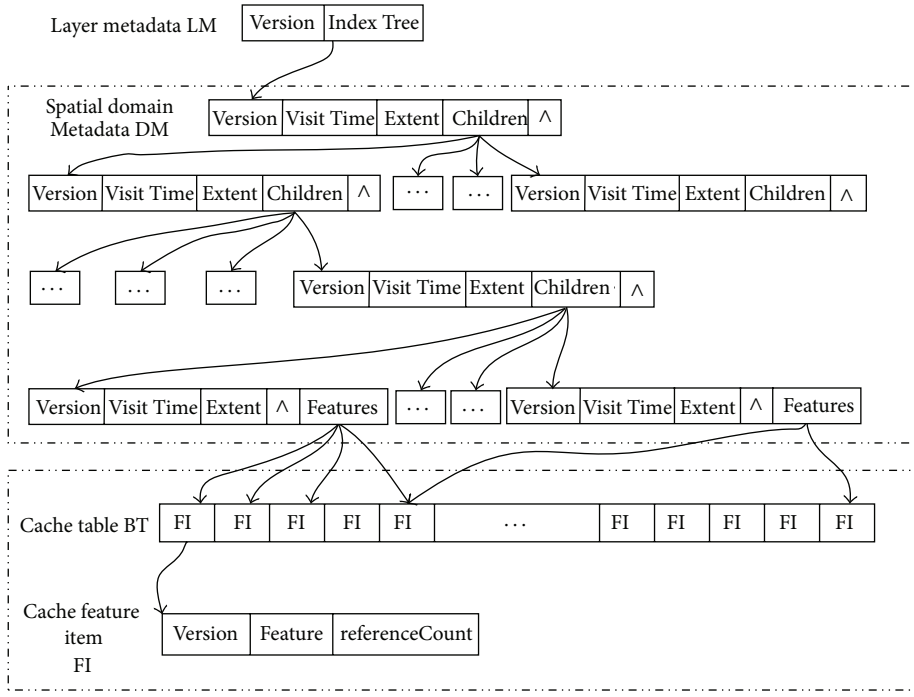


Figure 3: Cache data model.

To meet the requirement of real-time updates, we use DELTA mechanism, update dynamic information through incremental form. Through the combination of DDC and DELTA, it can better support huge amounts of data, accelerate the access speed and ensure data correctness.

The client cache is based on the following fact: in a period of time, user's inspection of the map and the retrieval of power facilities are concentrated in certain layer. According to the access frequency of the layer, the data is cached in the different levels of cache, when a user retrievals data start from the fastest cache. If does not exist, then retrievals in the next cache. Client local caching content includes all or part of the DDC, the background map tiles visited, all kinds of versions of space facilities data model, user permissions metadata, and others. When facilities updated, it realizes real-time requirements by the server-side DELTA mechanism.

By applying multiagent technology, the power WebGIS platform client provides basic maps show function, browse function, space analysis function, resources query localization function, local map cache data, and user data management function. In addition, each client registers a client agent, as the link between the clients completes the function of map synergy and cached data sharing between the clients by mobile agent.

Dynamic cache mechanism based on intelligent agent creates cache spaces on the server-side and the client-side, respectively. Server-side cache is maintained by multiple applications terminal jointing, while the client cache is maintained by each application terminal separately. The caching model adopts version-based caching strategy according to the layer version, the version of the spatial domain feature, and the version of cached data sheets. Due to the quad-split, relationship exists between the spatial domains trellises coded, so when retrieving in the index tree, we can skip certain levels of spatial domain node and query directly to a leaf node. Cache model is shown in Figure 3.

LM represents the layer data block metadata; version is the number of the version for the cache layer elements sets; index Tree is the quad index tree for finding spatial elements, the type of each node of the index tree is domain metadata. A one-to-one relationship exists between each node and a quad-split map (spatial domain). The maximum depth of the search tree equals to the zoom level supported by the platform.

DM represents spatial index tree node, version is the version number of cache space domain element sets; visit time represents the last access time of the spatial domain used in the LRU algorithm; extent represents the scope of the space domain used in priority replacement algorithm based on farthest distance; children represents the pointer pointing to the subspace domain; and features is used to save entry address of feature set in the region.

The feature collections are saved in the cache data sheet Buffer table.

BT represents data sheet, featureColl is cached data collection, the features cached in the data table can be shared by all spatial domain.

FI represents cached feature items model, feature represents the cached elements; version represents the version of the cached feature items; and referenceCount stand for the times a feature is referenced. When the counter is 0, the occupied storage space can be recycled.

Based intelligent agent dynamic caching mechanism, where the power WebGIS platform is able to meet the efficient data access and real-time requirements, significantly improve the hit rate of the data and effectively accelerate the speed of data access and graphics rendering efficiency.

5. Experimental Results

According to the dynamic cache model proposed in this paper, we design the test for vector data cache. That is to say, the elements in the vector layer cache are to be tested. Database cache refers to the unique index created on the feature id field of feature table. Create cluster index so as to index features by spatial domain (tiles) id field. Server-side and client cached data means to establish the mapping between layer and space domain set in layer; the mapping between spatial domain and features set entry address buffered in spatial domain to improve features search efficiency.

Test Case 1. Performance comparison with same concurrent users and different data size.

(i) Test object: low-voltage power lines of a provincial power company.

(ii) Hardware environment: shown in Table 1.

(iii) Test data size: 80000 features on sever-side and 20000 features on client-side.

(iv) Test mode: 10 client access WebGIS server concurrently.

(v) Test goal: comparing the vector data loading time when using dynamic display cache model proposed in this paper with the vector data loading time without caching pattern.

Table 2 shows platform performance comparison in our test. As can be seen from the comparison results, dynamic display caching can shorten the system response time by 2–5 times. Results show that the proposed dynamic display cache technology can significantly improve the response speed of the map and enhance the user experience.

Test Case 2. Response time comparison with same data size and different concurrent users.

(i) Test object and hardware environment was the same as test case 1.

(ii) Test data size: 4000 features.

(iii) Test mode: 10, 20, 30, . . . , 100 concurrent user.

(iv) Test goal: comparing the vector data loading time with dynamic display cache model proposed in this paper with the vector data loading time without caching pattern in different concurrent users.

Figure 4 showed the experimental results, from which we could conclude the following.

(1) The response time of the two ways was low with the smaller concurrent users, but the response time of noncaching mode was 2 times that of caching mode.

Table 1: Hardware environment.

Item	Server	Client
OS	Windows server 2003	Windows 7
CPU	Intel(R) i3	Intel(R) i3
Memory size	4 GB	2 GB
CPU Clock Speed	2.4 GHz	2.4 GHz

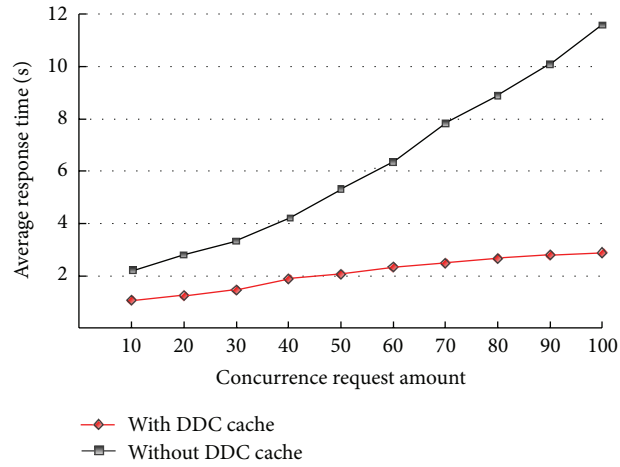


Figure 4: Average response time using different number of concurrency request.

(2) The response time of the two ways both increased as the concurrent users added, but the response time of noncaching mode increased faster, and the response time of noncaching mode was 4.2 times than that of caching mode when the concurrent users reached 100.

(3) The response time could be lower obviously when the dynamic display caching model was used, while the advantage with caching mode was greater and greater as the concurrent users. Dynamic display caching could shorten the system response time by 2–4.2 times.

6. Conclusion

Aiming at the key problem affecting the power performance of WebGIS-caching mechanism, this paper presented a dynamic display cache model, which makes a research on dynamic display cache technology based on intelligent agent, designs the power WebGIS dynamic cache structure, and cached data model. The application of the intelligent caching technology can greatly improve the WebGIS graphics loading speed and the response efficiency when dealing huge amounts of data, improve user concurrent traffic, and balance the network load. But research is still insufficient; the future research direction is described below.

(1) Smart cache can improve the loading speed of vector graphics and meet the needs of two-dimensional GIS, but for Digital Elevation Model (DEM), aerial data, there are no efficient load acceleration solutions.

Table 2: Performance comparison.

Vector data number	Nonindexed cache mode load time (ms)	Indexed cache mode loading time (ms)	Load faster ratio
1,541	942	320	2.94
2,075	1,245	805	1.54
4,111	2,421	1,167	2.10
9,215	5,381	1,675	3.21
17,598	9,472	2,950	3.21
40,393	21,364	5,089	4.19
78,980	45,898	9,374	4.89
141,481	98,507	21,569	4.58

(2) Mobile agent can improve the client cache sharing and improve the response speed of the map, but further study on cache data security and the mobile agent secure communications is needed.

References

[1] Y. Ping, Y. Wei, and Z. Lei, "Design of electric GIS sharing platform based on SOA structure," *Hubei Electric Power*, vol. 29, no. 4, pp. 20–23, 2010.

[2] X. Wang, K. Wu, X. Xiao, and Y. Wang, "A study on power WebGIS platform," *Power System Technology*, vol. 36, no. 4, pp. 174–178, 2012.

[3] Z. Cunmin, *QoS and Key Issues in WebGIS*, China University of Geosciences, 2005.

[4] C. Yang, H. Wu, Q. Huang et al., "WebGIS performance issues and solutions. Advances in Web-based GIS," *Mapping Services and Applications*, pp. 121–138, 2011.

[5] Z. Yuhua, *A Research and Implementation of WebGIS Based Multi Agents*, Information Engineering University of PLA, 2004.

[6] Y. Luo, X. Wang, and Z. Xu, "Spatial cache framework in component-based WebGIS," *Journal of Computer-Aided Design and Computer Graphics*, vol. 17, no. 2, pp. 320–326, 2005.

[7] S. Yesilmurat, *A Prefetching Method for Interactive Web Gis Applications*, Middle East Technical University, 2010.

[8] W. Xiaohui, W. Kehe, and Z. Xiaodong, "A study of WebGIS platform based on acceleration engine," in *Proceedings of the 3rd International Conference on Electronics Computer Technology (ICECT '11)*, pp. 108–112, Kanyakumari, India, April 2011.

[9] Z. Tianhui, D. Jinkang, and G. A. O. Jing, "Application of multi-level cache technique in improving quality of power GIS," *Water Resources and Power*, p. 2, 2009.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 424162, 9 pages
http://dx.doi.org/10.1155/2013/424162

Research Article

Power and Time Slot Allocation in Cognitive Relay Networks Using Particle Swarm Optimization

Pouya Derakhshan-Barjoei,[1] Gholamreza Dadashzadeh,[2] Farbod Razzazi,[1] and S. Mohammad Razavizadeh[3]

[1] *Department of Electrical Engineering, Science and Research Branch, Islamic Azad University, P.O. Box 775/14515, Tehran 1477893855, Iran*
[2] *Electrical and Electronic Engineering Department, Shahed University, Persian Gulf Highway, Tehran, Iran*
[3] *Department of Electrical Engineering, Iran University of Science & Technology, Tehran, Iran*

Correspondence should be addressed to Pouya Derakhshan-Barjoei; derakhshan@m.ieice.org

Received 5 August 2013; Accepted 22 August 2013

Academic Editors: W. Sun, G. Zhang, and J. Zhou

The two main problems in cognitive radio networks are power and time slot allocation problems which require a precise analysis and guarantee the quality of service in both the primary and secondary users. In this paper, these two problems are considered and a method is proposed to solve the resulting optimization problem. Our proposed method provides an improved performance in solving the constrained nonlinear multiobject optimization for the power control and beamforming in order to reach the maximum capacity and proper adaption of time slots, and as a result a new scheme for joint power and time slot allocation in cognitive relay networks is proposed. We adopt space diversity access as the secondary users access scheme and divide the time between multiple secondary users according to their contribution to primary user's transmission. Helping primary users provides more opportunities for secondary users to access the channel since the primary users can release the channel sooner. In contrast, primary network leases portion of channel access time to the secondary users for their transmission using particle swarm optimization (PSO). Numerical studies show good performance of the proposed scheme with a dynamic cost function in a nonstationary environment.

1. Introduction

Cognitive radio (CR) is an intelligent radio communication system which has the potential to make the best of unoccupied licensed spectrum while introducing little interference to licensed or primary user (PU). Cognitive radio is aware of the changing radio environment and is to make real-time adaptation to achieve its targets. Due to the accelerated deployment of broad band communication systems, current fixed frequency allocation schemes spectrum is becoming a major bottleneck. Therefore, there is an increasing interest in this technology among the researchers in academia, industry and spectrum policy makers. Hence, many studies featuring recent advances in theory, design, and analysis of cognitive wireless radio networks have been figured out. In general, cognitive wireless radio network is capable of adapting to the outside existing time varying environment. The cognitive transmitters should use the environmental information to approach the appropriate parameters, such as modulation type, modulation index, coding format, and transmission power level, in order to maximize their data transmission rates under a constrained interference level. One of the main concerns of the networks topology and capacity is the transmission power. To avoid the interference with the primary users, the transmission power of the cognitive wireless radio network should be controlled and limited. Controlling of the transmission power is of vital importance. In [1–3] the power control and spectrum sharing limitations have been studied. According to the descriptions in [4] the power control had an effective impact on the probability of bit error rate. Therefore, in our simulations, minimizing the bit error rate is the optimization criterion. In [5], a relay scheme has been used

for balancing the traffic requests and available spectrum resources in cognitive radio. We extended that idea to consider the secondary network as a relay network in our proposed TDMA scheme for balancing the traffic. Signal-to-interference-plus-noise ratio (SINR) has been enhanced by relays through spatial diversity in [6, 7]. In addition, in [8], directional transmission of relays for exploiting spatial spectrum holes has been studied. According to [9], the relay channels make improvement in the performance through spatial diversity by using additional paths between source and destination. Joint beamforming and power control using weighted least square algorithm have been performed in [10]. Beamforming can be implemented either at the transmitter or at the receiver. Transmitter beamforming concentrates the transmission signal on a certain direction in order to minimize interference with other users. Receiver beamforming is usually useful for signal localization or for taking advantage of spatial diversity. In addition, as mentioned in [11], the distances between base stations and users have an impressive role in the topology of the systems. The spectrum sensing and signal localization have been studied in [12, 13]. A power control method based on genetic algorithm for cognitive radio has been presented in [14]. The problem of transmission power control in cognitive radio networks considering propagation channels has been studied in [15–17]. One of the different features of our proposed method is its utilization in randomly time varying fading channels. In [18, 19], a cooperative communication network for cognitive radios has been investigated. Also, the population adaptation for genetic algorithm based on cognitive radio and bioinspired algorithm for dynamic resource allocation and parameter adaptation have been studied in [15–20]. A precise power control in a randomly time varying environment has been studied by employing an intelligent algorithm in [21, 22]. In addition, dynamic spectrum sensing and spectrum management have been studied in [23–25]. The analysis of detection time for multiple-user cooperative spectrum sensing and best relay selection has been studied in [23]. These researches make it possible for a secondary or cognitive radio network to opportunistically utilize a frequency band initially allocated to a primary network. In [26–29], common spectrum sensing and adaptive power allocation methods have been considered. However, evolutionary power control for cognitive users has not been previously investigated in randomly time varying fading environments considering relay diversity and time slot allocation. In addition, the use of PSO in wireless scenario always faces the problem of computational complexity. The wireless channel varies quickly, and the cognitive radio network requires high number of operations in short time intervals. In [30], by using PSO, the minimum bit error rate has been investigated for multiuser transmission design and computational complexity has been figured out for MIMO channel.

In this paper, we proposed an intelligent method to adapt the channel time slots and allocate transmission power of secondary users (SUs) with relay diversity by using particle swarm optimization. To achieve these goals, we formulate an optimization problem, considering a heuristic strategy for both networks, so we adapt space diversity access for the secondary users access scheme and divide the time between multiple cognitive users regarding their contribution to primary user's transmission. The cognitive user who contributes more to a primary user gets more revenue. Multiple antennas have been assumed to be deployed at the secondary users. Many wireless network standards provision the use of transmit antenna arrays. In our proposed scheme, by using beamforming at the remaining time slot, it is possible to allocate energy in the direction of the intended users, whose channels can often be accurately estimated. Beamforming has been also exploited as a strategy that can serve many users at similar throughput. Due to the variation of radio channel characteristics, as well as the frequency spectrum band availability, cognitive radio networks need to support time varying quality of service requirements. The basic goals of our work are focused on dynamic power and time slot allocation in time division multiplexing access mode, maximizing the transmission rate of both primary and secondary users and minimizing the transmission power of secondary users, with relay diversity in a randomly time varying environment.

2. System Model

We consider a system model where the primary network consists of N primary users (PUs) each having a transceiver. The primary network transmits and communicates with the constant and specific transmission power. The system model of our scenarios is illustrated in Figure 1, and the time slot–division is shown in Figure 2. The secondary network included M secondary users communicating in an ad hoc scenario. The secondary network processing is based on beamforming at both the transmitter (K antennas) and the receiver (K antennas) for each secondary user link. In our scenario, the downlink of the primary network is considered. In the secondary network, secondary users are considered to work in the same frequency band as the primary system. The secondary network has an ad hoc scenario with deployment of K antennas at each secondary transmitter. An efficient transmit beamforming technique is proposed to maximize the total throughput.

The transmit powers of secondary users are limited to a maximum value prescribed by primary users. In Figure 1(a), primary base station transmits signal to primary receivers and secondary users. Figure 1(b) shows a transmission between secondary relays and primary receivers. Secondary users work as relays. In part (c) the transmission between secondary users is shown. The data transmission is divided into frames, whose duration is 1 time unit. t_p denotes the first fraction of the time slot which is dedicated to the transmission of data to both primary receiver and secondary user (SU) as the cooperative relay. Therefore the remaining $1 - t_p$ time unit of slot is separated into two subslots based on parameter γ as shown in Figure 2. In the second part of time slot, $\gamma(1 - t_p)$, relay SUs transmit primary user's data to primary receiver, and finally in the remaining part of time slot, $(1 - \gamma)(1 - t_p)$, multiple secondary users also access channel in space diversity mode. Here the amount of access time for secondary users is related to the contribution they made in relay process.

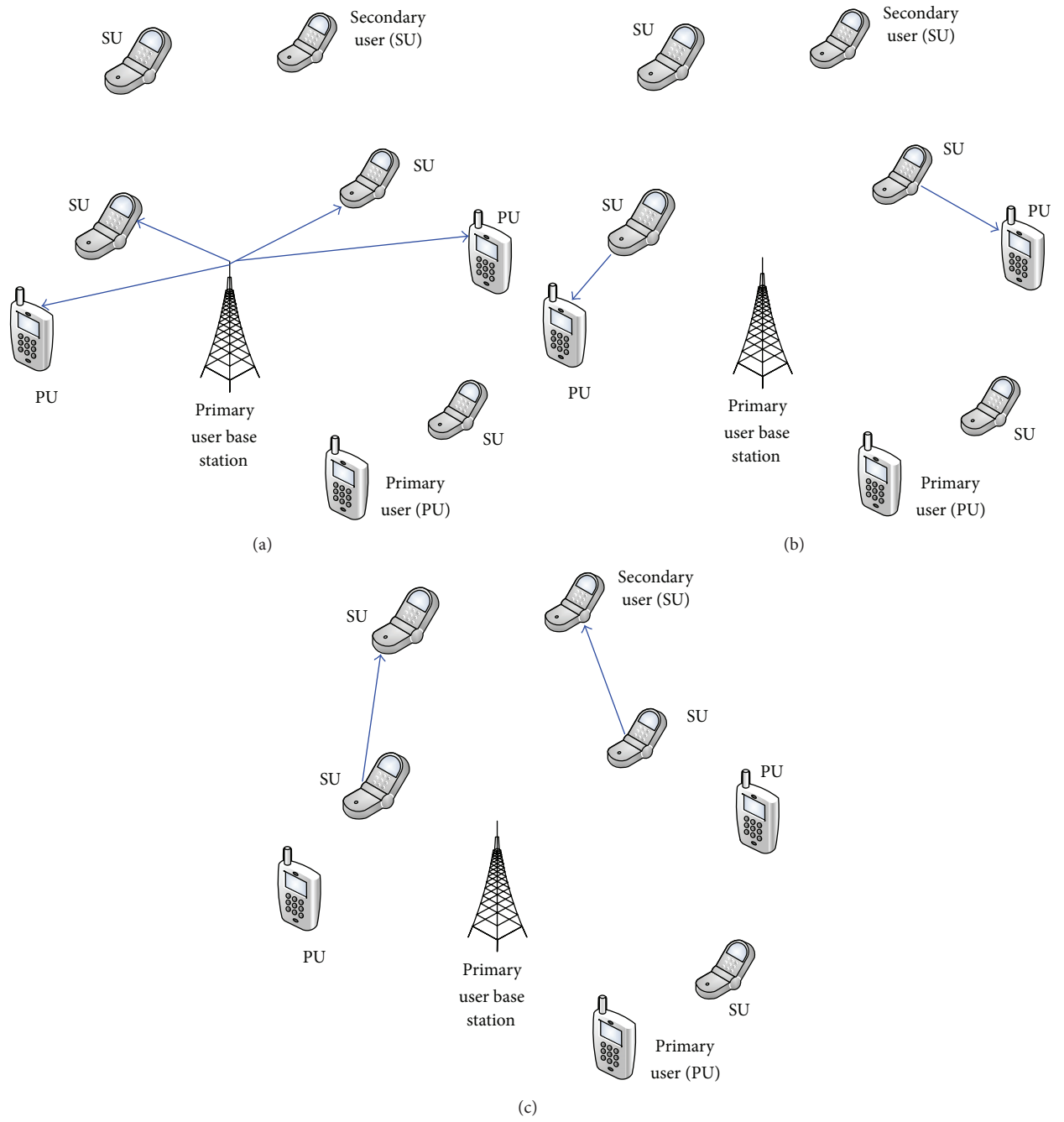


Figure 1: Conceptual diagram of the system model.

This network coexists in the same area with secondary users which are cognitive users.

The transmission scheme is characterized by the power allocation, eigenvectors, and eigenvalues of the transmit covariance matrix.

3. Problem Formulation and Analysis

All secondary users are working intelligently in an ad hoc scenario. A secondary user or cognitive user is talking to

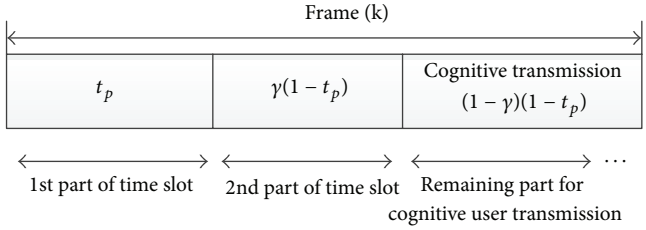


Figure 2: Time slot division for secondary and primary transmission.

a receiver using a frequency band licensed to the primary radio; the objective here is considered as to maximize the transmission capacity of the primary and secondary users and to allocate time slot for an optimal transmission. In addition, beamforming of the transmitted signal from cognitive users is considered as pre- and postbeamforming vectors The received signal at the secondary users functioning as relay is obtained as follows:

$$y_m\left(t_p\right) = \sqrt{p_{pu}} g_{pm} x_{pu}\left(t_p\right) + n_m, \tag{1}$$

where p_{pu} is the transmit power of primary base station, x_{pu} and g_{pm} are the transmitted signals of the primary base station and the channel gain between primary base station and mth secondary user, respectively, and n_m denotes the additive white Gaussian noise. The received signal at the $t' = \gamma(1 - t_p)$ time slot at primary receiver is as follows:

$$y_p\left(t'\right) = \sqrt{p_m} g_{mp} \frac{y_m\left(t_p\right)}{\left|y_m\left(t_p\right)\right|} + n_m, \tag{2}$$

where g_{mp} denotes the channel gain between mth secondary user and primary receiver, p_m denotes the transmit power of secondary user, and $|\cdot|$ represents magnitude operator. Power is constrained by a maximum transmit power limit. Here we present the pre- and postbeamforming vectors, and also we design the transmit and receive beam vectors. In fact, beam vectors associated with each secondary user are determined by optimizing a certain criterion to reach a specific target such as maximizing the throughput or minimizing the interference. To compute the beam vectors, we consider just the secondary users use MIMO system. The reason is that the interference among primary users is nulled in SINR equation. In fact, we propose an algorithm that can minimize the interference between secondary users and maximize the rate. In particular, beam vectors are selected such that they satisfy the interference free condition. Assuming that the secondary users signals are uncorrelated with zero mean, we can express the mth secondary user received signal at the remaining part of time slot as:

$$\mathbf{y}_m^{(1)} = \mathbf{H}_{\mathrm{su}_{mm}} \mathbf{s}_m + \sum_{j=1, j\neq m}^{M} \mathbf{H}_{\mathrm{su}_{jm}} \mathbf{s}_j + \mathbf{n}_m, \tag{3}$$

where $\mathbf{H}_{\mathrm{su}} \in \mathbb{C}^{K\times K}$, a $K \times K$ complex vector of the fading path gains between secondary users. This vector is set with random complex components and obeys the Rayleigh distribution. The additive white Gaussian noise vector $\mathbf{n}_m \in \mathbb{C}^{K\times 1}$ is a Gaussian random process with zero mean and variance N_0 on each vector component [30]. The transmit vector $\mathbf{s}_m$ of size $K \times 1$ is yielded as follows:

$\mathbf{s}_m = x_m \mathbf{b}_m$, where $\mathbf{b}_m \in \mathbb{C}^{K\times 1}$, the prebeamforming vector, and x_m is the transmit sample for m between 1 and M. The mth receiver beam former is

$$y_m = \mathbf{a}_m^H \mathbf{y}_m^{(1)}. \tag{4}$$

So we can express that the mth secondary user received signal as follows:

$$y_m = \mathbf{a}_m^H \mathbf{H}_{\mathrm{su}_{mm}} \mathbf{b}_m x_m + \mathbf{a}_m^H \sum_{j=1, j\neq m}^{M} \mathbf{H}_{\mathrm{su}_{mj}} \mathbf{b}_j x_j + \mathbf{a}_m^H \mathbf{n}_m, \tag{5}$$

where $\mathbf{a}_m \in \mathbb{C}^{K\times 1}$, the postbeamforming vector at the receiving secondary users. The signal to interference noise ratio (SINR) at the mth secondary user is as follows:

$$\begin{aligned}\Gamma_{\mathrm{su}} &= \frac{E\left[\left|\mathbf{a}_m^H \mathbf{H}_{\mathrm{su}_{mm}} \mathbf{b}_m x_m\right|^2\right]}{E\left[\left|\mathbf{a}_m^H \sum_{j=1, j\neq m}^{M} \mathbf{H}_{\mathrm{su}_{mj}} \mathbf{b}_j x_j\right|^2\right] + E\left[\left|\mathbf{a}_m^H \mathbf{n}_m\right|^2\right]} \\ &= \frac{\left|\mathbf{a}_m^H \mathbf{H}_{\mathrm{su}_{mm}} \mathbf{b}_m\right|^2 p_{\mathrm{su}}}{\sum_{j=1, j\neq m}^{M} \left|\mathbf{a}_m^H \mathbf{H}_{\mathrm{su}_{mj}} \mathbf{b}_j\right|^2 p_j + \left\|\mathbf{a}_m^H\right\|^2 N_0},\end{aligned} \tag{6}$$

where $E[|n_{j,m}|^2] = N_0$ and $n_{j,m}$ is jth entry of $\mathbf{n}_m$. Also $E[|x_j|^2] = p_{\mathrm{su}}$ for $j = 1, 2, 3, \ldots, m$, where $E[\cdot]$ denotes statistical expectation, while $(\cdot)^H$ and $\|\cdot\|$ represent the Hermitian and 2-norm operators, respectively. So the per-user total rate is

$$\begin{aligned}R_{\mathrm{su}} &= \sum_{m=1}^{M} \log_2\left(1 + \Gamma_{\mathrm{su}}\right) \\ &= \sum_{m=1}^{M} \log_2\left(1 + \frac{\left|\mathbf{a}_m^H \mathbf{H}_{\mathrm{su}_{mm}} \mathbf{b}_m\right|^2 p_{\mathrm{su}}}{\sum_{j=1, j\neq m}^{M} \left|\mathbf{a}_m^H \mathbf{H}_{\mathrm{su}_{mj}} \mathbf{b}_j\right|^2 p_j + \left\|\mathbf{a}_m^H\right\|^2 N_0}\right).\end{aligned} \tag{7}$$

The total interference plus noise covariance matrix, $\mathbf{\Phi}_{\mathrm{su}}$, at the mth secondary user is defined as follows:

$$\mathbf{\Phi}_{\mathrm{su}} = \sum_{j=1, j\neq m}^{M} \left(\mathbf{H}_{\mathrm{su}_{mj}} \cdot \mathbf{b}_j \cdot \mathbf{b}_j^H \cdot \mathbf{H}_{\mathrm{su}_{mj}}^H\right) + N_0 \mathbf{I}_K, \tag{8}$$

where $\mathbf{I}_K$ is an identity matrix of size $K \times K$. Therefore, the SINR at the mth secondary user can be formulated as follows:

$$\begin{aligned}\Gamma_{\mathrm{su}} &= \frac{\left(\mathbf{a}_m^H \mathbf{H}_{\mathrm{su}_{mm}} \mathbf{b}_m\right)^H \left(\mathbf{a}_m^H \mathbf{H}_{\mathrm{su}_{mm}} \mathbf{b}_m\right)}{\mathbf{a}_m^H \mathbf{\Phi}_{\mathrm{su}} \mathbf{a}_m} \\ &= \left(\mathbf{a}_m^H \mathbf{H}_{\mathrm{su}_{mm}} \mathbf{b}_m\right)^H \left(\mathbf{a}_m^H \mathbf{\Phi}_{\mathrm{su}} \mathbf{a}_m\right)^{-1} \left(\mathbf{a}_m^H \mathbf{H}_{\mathrm{su}_{mm}} \mathbf{b}_m\right) \\ &= \mathbf{b}_m^H \mathbf{H}_{\mathrm{su}_{mm}}^H \mathbf{\Phi}_{\mathrm{su}}^{-1} \mathbf{H}_{\mathrm{su}_{mm}} \mathbf{b}_m.\end{aligned} \tag{9}$$

The postbeamforming vector can be expressed as follows:

$$\mathbf{a}_m = \mathbf{\Phi}_{\mathrm{su}}^{-1} \mathbf{H}_{\mathrm{su}_{mm}} \mathbf{b}_m. \tag{10}$$

This gives us the following maximization of SINR at the mth secondary user:

$$\begin{aligned}\mathbf{b}_m^H \mathbf{H}_{\mathrm{su}_{mm}}^H \mathbf{\Phi}_{\mathrm{su}}^{-1} \mathbf{H}_{\mathrm{su}_{mm}} \mathbf{b}_m &\leq \lambda_{\max}(j)\, \mathbf{b}_m^H \mathbf{b}_m, \\ \lambda_{\max}(j)\, \mathbf{b}_m^H \mathbf{b}_m &= \lambda_{\max}(j) \left\|\mathbf{b}_m\right\|^2.\end{aligned} \tag{11}$$

The power constraint is formulated as follows:

$$\sum_{j=1}^{M} \mathbf{b}_j^H \mathbf{b}_j = \sum_{j=1}^{M} \left\| \mathbf{b}_j \right\|^2 \leq M p_{\max}. \tag{12}$$

The maximum eigenvalue of $\mathbf{H}_{\mathrm{su}_{mm}}^H \boldsymbol{\Phi}_{\mathrm{su}}^{-1} \mathbf{H}_{\mathrm{su}_{mm}}$ must be chosen to maximize the capacity of secondary users, and $\lambda_{\max}(m)$ is the maximum eigenvalue of $\mathbf{H}_{\mathrm{su}_{mm}}^H \boldsymbol{\Phi}_{\mathrm{su}}^{-1} \mathbf{H}_{\mathrm{su}_{mm}}$. $P_{\max}$ denotes the maximum transmit power of each secondary user. The transmission rate between the primary and secondary user is as follows:

$$R_{pm} = \log_2 \left(1 + \frac{p_{pu} g_{pm}^2}{N_0^2} \right). \tag{13}$$

Similarly, the transmission rate between the secondary and primary receiver is defined as

$$R_{mp} = \log_2 \left(1 + \frac{p_m g_{mp}^2}{N_0^2} \right). \tag{14}$$

The data rate of direct transmission is given by

$$R_p = \log_2 \left(1 + \frac{p_{pu} g_{pu}^2}{N_0^2} \right), \tag{15}$$

where g_{pu} denotes the channel gain between primary base station and primary users. Since the transmission links are serially connected at relay, the throughput equals the smaller throughput of the two links. Hence, the overall achievable primary rate of cooperative transmission equals the minimum transmission rate of two links. Then one has:

$$R_{pu_{\mathrm{Coop}}}(t) = \min \left\{ t_p R_{pm}, \gamma \left(1 - t_p \right) R_{mp} \right\}. \tag{16}$$

As it is clear that there is an equilibrium point between $t_p R_{pm}$ and $\gamma(1 - t_p) R_{mp}$; therefore, $R_{pu_{\mathrm{Coop}}}(t)$ is maximum when t_p satisfies $t_p R_{pm} = \gamma(1 - t_p) R_{mp}$. In order to attain maximum relayed and direct SNR, the utility functions for both the primary and secondary users with cooperative transmission rate of two transmission periods are figured out as follows:

$$\begin{aligned} U_{pu} &= t_p \log_2 \left(1 + \frac{p_{pu} g_{pu}^2}{n_m^2} \right) + \frac{\gamma R_{pm} R_{mp}}{R_{pm} + \gamma R_{mp}}, \\ U_{\mathrm{su}} &= \left(1 - \gamma \right) \left(1 - t_p \right) \log_2 \left(1 + \lambda_{\max} \mathbf{b}_m^H \mathbf{b}_m \right), \end{aligned} \tag{17}$$

where p_m denotes contributive transmit power of mth cognitive relay, g_{pu} is channel gain between primary users, and g_{pm}, g_{mp} are channel gains between primary base station and mth secondary user and between the mth secondary user and primary receiver, respectively. The main cost function is defined as follows:

$$\max_{t_p, \gamma, b} U_0 \left(t_p, \gamma, \mathbf{b}_m \right) = \alpha_1 U_{\mathrm{su}} + \alpha_2 U_{pu}. \tag{18}$$

In which the weights are defined as α_1, α_2 and $\alpha_1 + \alpha_2 = 1$. These parameters define the priority and importance of each utility function. So we can choose them as pairs of $\{(0.2, 0.8), (0.5, 0.5), (0.6, 0.4)\}$ arbitrarily. For beamforming, the transmitted power through all the secondary users for the mth secondary user is proportional to $\|\mathbf{b}_m\|^2$. The purpose of this paper is to determine optimal transmit power for all possible fading channel status in nonstationary conditions so as to maximize the channel capacity and time slot allocation with relay diversity based on particle swarm optimization. By considering a cost function, we can convert the constrained optimization process into a multiobjective cost function to meet problem constraints simultaneously. The cost function's behavior is dynamic due to nonstationary environment specifications.

4. Parameters Adjustment Using PSO

4.1. Brief Introduction to PSO. To achieve the optimal performance, weight multiplier α_i is adjusted to satisfy priority. A swarm of particles that represent potential solutions are evolved in the search space. Particle swarm optimization is inspired by observing the bird flocking or fish school. Scientists found that the synchrony of flocking behavior was through maintaining optimal distances between individual members and their neighbors. Furthermore, they simulated the scenario in which birds, fishes, or bees search for food, and observing their social behavior, they perceived that in order to find food, the individual members determined their velocities by two agents, their own best previous experience and the best experience of all other neighbors and members. According to this concept, the so-called PSO for optimization of continuous nonlinear functions has been developed [31, 32]. However, several improvements have been made in this work, we used the original one, and employing the improved versions of PSO for this optimization remained, an open problem for future studies [33]. Here birds, fishes, or bees are called particles, each representing a potential solution, and they all have their position, velocity, and fitness value. To find the optimal solution, each particle adjusts its flying according to its own flying experience and companion's flying experience. A swarm of particles are created. While the initial particles are randomly generated in the search space, each particle keeps its best position in its memory. We should check to ensure that all the particles stay inside the search space, if a particle is outside the search space. It is moved back inside the search space randomly and finally it is forced to stay at the border. In Algorithm 1, the procedure is defined. Based on its searching mechanism, primitive position denotes solution and velocity denotes the mutative direction the solution may take. The pseudocode of the PSO procedure can be written as follows.

In a d-tuple searching space the position and velocity of a particle i are denoted by $X_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$ and $V_i = [v_{i1}, v_{i2}, \ldots, v_{id}]$. Our fitness function is to evaluate every particle to figure out the best solution $P_i = [p_{i1}, p_{i2}, \ldots, p_{id}]$, it may find the best solution P_g for the whole swarm at time t,

```
(1)  Begin
(2)   Randomly initialize particles swarm, [p_m, b_m, γ, t_p]
(3)   while (the stopping criterion is not met)
(4)     Evaluate fitness of particles
(5)     for n = 1 to number of particles
(6)     Find pbest
(7)     Find gbest
(8)       for d = 1 to number of dimension of particle
(9)         update the position of particles by (19), (20)
(10)      next d
(11)    next n
(12)    update the inertia weight value ranges from 0.3 to 0.9
(13)  next generation until stopping criterion
(14)  End
```

Algorithm 1: Pseudocode for PSO.

and the position and velocity are updated with (19) and (20). P_g is the optimum solution. Consider

$$v_{ij}(t+1) = \omega v_{ij}(t) + c_1 r_1 \left[p_{ij} - x_{ij}(t)\right] + c_2 r_2 \left[p_{gj} - x_{ij}(t)\right], \tag{19}$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1), \quad j = 1, \ldots, d, \tag{20}$$

where ω denotes the inertia weight factor, c_1, c_2 denote positive accelerators, and r_1, r_2 are random numbers uniformly distributed in interval $[0, 1]$. The role of inertia weight ω is considered to be crucial for convergence and is to control the impact of the previous history of velocity on the current velocity. Thus it regulates the tradeoff between global and local exploration for the swarm. A large ω makes the searching escape from local minima and facilitates global searching, while small ω facilitates local searching and convergence. When the particles get trapped in local optima, the inertia weight is augmented, and when they are dispersive, the weight is decreased. The velocity interval $[v_{\min}, v_{\max}]$ and position interval $[x_{\min}, x_{\max}]$ are to restrict the searching in the required domain. The above velocity renewal equation (19) comprises three parts. The first part is dominated by current velocity and contributes to the tradeoff between global searching and local searching while the second part embodies cognitive pattern and is to adjust direction based on its recollection to avoid local minima. The third part reflects social effects that shared information contributes to collaboration.

4.2. Allocation Algorithm. According to discussion in Sections 3 and 4, in this simulation for the adaptation of parameters to the PSO, some parameters are selected as particles, so we adapt these parameters to PSO particles and try to update them as iteration goes on. In each iteration of the algorithm, we make the answer after calculating the average amount of particles for specific number of repeating. Our proposed particle is a vector which contains the power of secondary users p_m, the prebeamforming vector $\mathbf{b}_m$, the time slot division factor γ, and the time slot duration t_p. The prebeamforming vector $\mathbf{b}_m$ is set with random complex components by PSO which is one part of the particle. Each component of particle $[p_m, \mathbf{b}_m, \gamma, t_p]$ has a velocity denoted by $v_{ij}(t)$ and position denoted by $x_{ij}(t)$. Each particle remembers its best position visited so far, denoted by p_{ij}, which provides the cognitive information. Every particle also knows the best position visited so far among the entire swarm, denoted by p_g, which provides the social information. p_{ij}, p_g are updated at each iteration. Also each element of particle is checked to ensure that it stays inside the search space. The amount of $p_m \le 5$ mW, $0 \le \gamma \le 1$, and $0 \le t_p \le 1$ and complex component of vector $\mathbf{b}_m$ should be considered for boundary check. For each term of particle the boundary region is defined. As it is clear, selected particle consists of a scalar and vector part. Using (19) and (20), the adaptation parameters to position and velocity interval are set [30, 34, 35]. Because our simulation is run in a nonstationary environment, all channel fading gains are generated randomly with Rayleigh distribution. So during the simulation, the amount of these gains may be changed, and as a result the algorithm will track the cost or fitness function, as all mentioned parameters vary.

5. Numerical Results

In a simulation model, we consider some simulation results to investigate performance of the proposed scenario in which the primary and secondary users are randomly distributed over the considered area. A secondary network coexists and/or shares the radio spectrum with a primary network to which the spectrum is licensed, in an infrastructure scenario. The channels between the transmitters and receivers are assumed to be Rayleigh faded with mean of one. The channel gains are independent across subchannels. The power spectral density of additive Gaussian noise is 10^{-8} W/Hz. Empirically, our results suggest that the velocity limit can be set to $v_m = 0.09$ and the acceleration coefficients can be set to $c_1 = 2$ and $c_2 = 2$. The primitive inertia weight ω is set to 0.9, ranges from 0.3 to 0.9, and varies as the iteration goes on. Path noises are independent zero-mean complex Gaussian random variables with variance $N_0 = 1$. The maximum transmit power $P_{\max}$ for each secondary user is assumed to be 5 mW, and the transmit

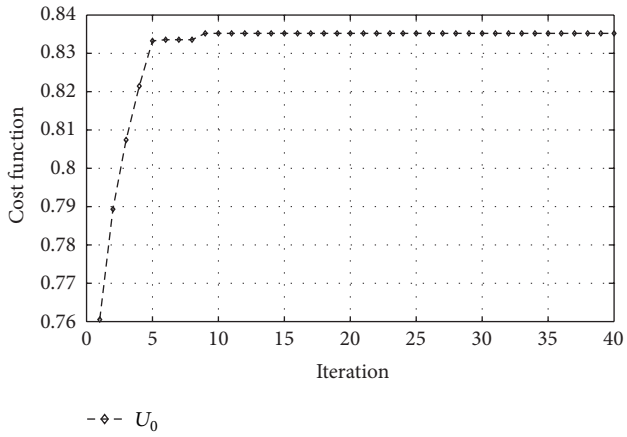


Figure 3: Behavior of the main cost function.

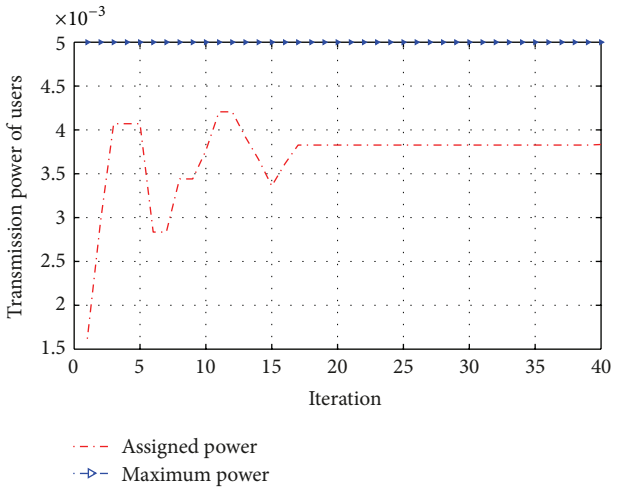


Figure 4: Convergence of transmission power for secondary users.

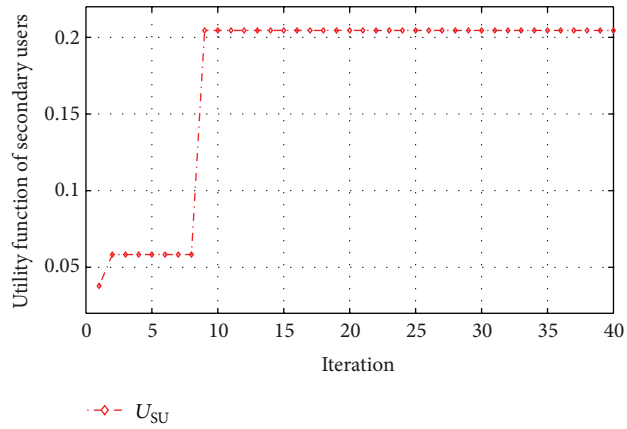


Figure 5: Utility function of the secondary users.

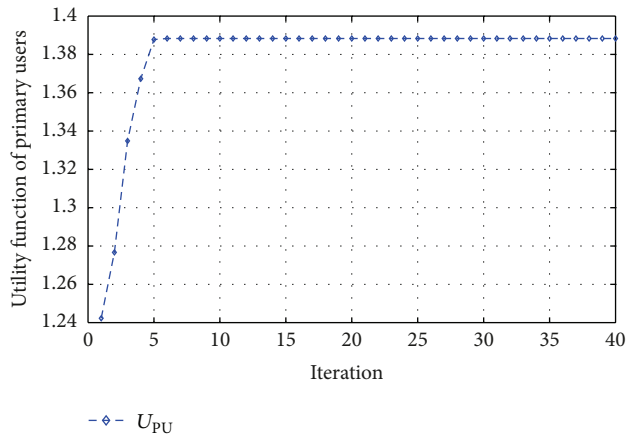


Figure 6: Utility function of the primary users with cooperative relays.

power of primary base station is set to 0.2 W for a 10^5 Hz bandwidth. Interference from primary users to base station is ignored. The amounts of time slot division parameters, γ and t_p, will be set during the iterations by related components in selected particle. It was found that the more repetition of the algorithm in each iteration has the much accuracy. In our algorithm, we repeated each iteration to reach the best accuracy. In Figure 3 the behavior of the cost function is shown.

By making average global information of the best particles, the accuracy of the best particles has been raised as shown in Figure 4. We can see that the behavior of the cost function and its convergence are the attributes of PSO, and it is clear that the all constraints are fulfilled.

Figure 4 shows that there is a difference between maximum transmission power of users and assigned transmission power. This convergence shows the control of the transmission power for secondary users by the proposed algorithm. From Figure 5, it can be seen that the utility function of secondary users arises by increasing the amount of transmission power; however, the power will be limited by our simulations constraints.

From Figure 6, it can be seen that the utility function of primary users arises along with cooperative relays. Furthermore, the fitness functions steer the evolution of the PSO in the correct direction to optimize the given multiobjective cost functions for the secondary and primary users with the defined constraints in a nonstationary environment. Figure 7 shows the performance of cooperative relay. It can be seen that the utility function of primary user arises when cooperative relay is used. However the channel gains are Rayleigh faded in a nonstationary environment. In wireless communication systems, the wireless channel varies quickly, so cognitive radio network requires exhausting operation every few seconds, Also there is a tradeoff between computational complexity and power consumed in this wireless communication scenario. Therefore it might be considered as the next interest.

In our simulation for short time slots, we consider slow varying channels and the algorithm can set the transmission parameters.

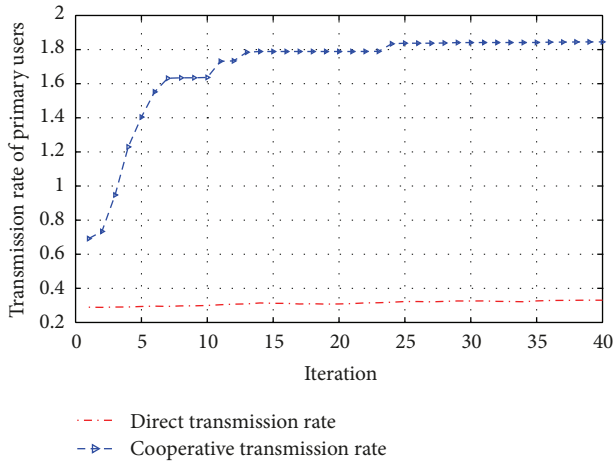


Figure 7: Transmission rate of primary users.

6. Conclusions

We have proposed a PSO-assisted distributed minimum transmission power and time slot allocation with relay diversity in cognitive or secondary radio network in a randomly time varying environment. The scenario is formulated in the downlink mode of the primary user network to maximize the transmission capacity of secondary and primary users. However, the minimum transmission power of each cognitive or secondary user is considered. We have developed joint time slot and transmission power control with relay diversity in which PSO adjusts the parameters, while maintaining a quality of service for the primary user. The PSO-aided algorithm provides improved performance by using appropriate pre- and postbeamforming. The proposed scheme shows the performance of a heuristic improvement in cognitive radio performance in a dynamic environment. During an extended searching space, the fitness function determines the values, and the number of iteration determines the speed. Hence, the idea of considering the practicality of the proposed method in terms of computational complexity is one of the next interests to us.

Acknowledgment

The authors would like to thank the anonymous reviewers, whose insightful comments made this a better paper.

References

[1] S. A. Grandhi and J. Zander, "Constrained power control in cellular radio systems," in *Proceedings of the 44th IEEE Vehicular Technology Conference*, pp. 824–828, Stockholm, Sweden, June 1994.

[2] S. A. Grandhi, J. Zander, and R. Yates, "Constrained power control," *Wireless Personal Communications*, vol. 1, no. 4, pp. 257–270, 1995.

[3] A. Ghasemi and E. S. Sousa, "Fundamental limits of spectrum-sharing in fading environments," *IEEE Transactions on Wireless Communications*, vol. 6, no. 2, pp. 649–658, 2007.

[4] C. S. Park and K. B. Lee, "Transmit power allocation for BER performance improvement in multicarrier systems," *IEEE Transactions on Communications*, vol. 52, no. 10, pp. 1658–1663, 2004.

[5] J. Jia, J. Zhang, and Q. Zhang, "Cooperative relay for cognitive radio networks," in *Proceedings of the 28th Conference on Computer Communications (IEEE INFOCOM '09)*, pp. 2304–2312, April 2009.

[6] X. Gong, W. Yuan, W. Liu, W. Cheng, and S. Wang, "A cooperative relay scheme for secondary communication in cognitive radio networks," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '08)*, pp. 3106–3111, December 2008.

[7] J. Mietzner, L. Lampe, and R. Schober, "Performance analysis for a fully decentralized transmit power allocation scheme for relay-assisted cognitive-radio systems," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '08)*, pp. 3772–3777, December 2008.

[8] G. Zhao, J. Ma, G. Y. Li et al., "Spatial spectrum holes for cognitive radio with relay-assisted directional transmission," *IEEE Transactions on Wireless Communications*, vol. 8, no. 10, pp. 5270–5279, 2009.

[9] Y. Liang, V. V. Veeravalli, and H. V. Poor, "Resource allocation for wireless fading relay channels: max-min solution," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3432–3453, 2007.

[10] M. H. Islam, Y.-C. Liang, and A. T. Hoang, "Joint power control and beamforming for cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 7, pp. 2415–2419, 2008.

[11] W. Ren, Q. Zhao, and A. Swami, "Power control in cognitive radio networks: how to cross a multi-lane highway," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 7, pp. 1283–1296, 2009.

[12] P. Derakhshan-Barjoei, G. Dadashzadeh, F. Razzazi, and S. M. Razavizadeh, "Comparison of radiometry and modified periodogram spectrum detection in wireless radio networks," in *Proceedings of the International Conference on Computer and Information Application (ICCIA '10)*, pp. 386–389, November 2010.

[13] J.-L. Wang, X. Zhang, and Q. Wu, "State transition probability based sensing duration optimization algorithm in cognitive radio," *IEICE Transactions on Communications*, vol. 93, no. 12, pp. 3258–3265, 2010.

[14] T. R. Newman, B. A. Barker, A. M. Wyglinski, A. Agah, J. B. Evans, and G. J. Minden, "Cognitive engine implementation for wireless multicarrier transceivers," *Wireless Communications and Mobile Computing*, vol. 7, no. 9, pp. 1129–1142, 2007.

[15] T. Renk, C. Kloeck, D. Burgkhardt et al., "Bio-inspired algorithms for dynamic resource allocation in cognitive wireless networks," *Mobile Networks and Applications*, vol. 13, no. 5, pp. 431–441, 2008.

[16] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, 2005.

[17] T. R. Newman, R. Rajbanshi, A. M. Wyglinski, J. B. Evans, and G. J. Minden, "Population adaptation for genetic algorithm-based cognitive radios," *Mobile Networks and Applications*, vol. 13, no. 5, pp. 442–451, 2008.

[18] A. F. Molisch, L. J. Greenstein, and M. Shafi, "Propagation issues for cognitive radio," *Proceedings of the IEEE*, vol. 97, no. 5, pp. 787–804, 2009.

[19] K. B. Letaief and W. Zhang, "Cooperative communications for cognitive radio networks," *Proceedings of the IEEE*, vol. 97, no. 5, pp. 878–893, 2009.

[20] 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.

[21] P. Derakhshan-Barjoei, G. Dadashzadeh, F. Razzazi, and S. Mohammad Razavizadeh, "Minimum power transmission design for cognitive radio networks in non-stationary environment," *IEICE Electronics Express*, vol. 8, no. 3, pp. 136–142, 2011.

[22] P. Derakhshan-Barjoei, G. Dadashzadeh, F. Razzazi, and S. Mohammad Razavizadeh, "Bio-inspired distributed beamforming for cognitive radio networks in non-stationary environment," *IEICE Electronics Express*, vol. 8, no. 6, pp. 332–339, 2011.

[23] G. Zhao, C. Yang, G. Y. Li, D. Li, and A. C. K. Soong, "Power and channel allocation for cooperative relay in cognitive radio networks," *IEEE Journal on Selected Topics in Signal Processing*, vol. 5, no. 1, pp. 151–159, 2011.

[24] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey," *Computer Networks*, vol. 50, no. 13, pp. 2127–2159, 2006.

[25] A. Eghbali, H. Johansson, P. Löwenborg, and H. G. Göckler, "Dynamic frequency-band reallocation and allocation: from satellite-based communication systems to cognitive radios," *Journal of Signal Processing Systems*, vol. 62, no. 2, pp. 187–203, 2011.

[26] I. Budiarjo, M. K. Lakshmanan, and H. Nikookar, "Cognitive radio dynamic access techniques," *Wireless Personal Communications*, vol. 45, no. 3, pp. 293–324, 2008.

[27] J. Zhu, B. Y. Zheng, and Y. L. Zou, "Detection time analysis for the multiple-user cooperative spectrum sensing scheme in cognitive radio networks," *Science in China F*, vol. 52, no. 10, pp. 1915–1925, 2009.

[28] J. Ma, G. Y. Li, and B. H. Juang, "Signal processing in cognitive radio," *Proceedings of the IEEE*, vol. 97, no. 5, pp. 805–823, 2009.

[29] I. F. Akyildiz, W.-Y. Lee, and K. R. Chowdhury, "CRAHNs: cognitive radio ad hoc networks," *Ad Hoc Networks*, vol. 7, no. 5, pp. 810–836, 2009.

[30] W. Yao, S. Chen, S. Tan, and L. Hanzo, "Minimum bit error rate multiuser transmission designs using particle swarm optimisation," *IEEE Transactions on Wireless Communications*, vol. 8, no. 10, pp. 5012–5017, 2009.

[31] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, pp. 1942–1948, December 1995.

[32] J. Kennedy, R. C. Eberhart, and Y. Shi, *Swarm Intelligence*, Morgan Kaufmann, San Francisco, Calif, USA, 2001.

[33] R. Mendes, J. Kennedy, and J. Neves, "The fully informed particle swarm: simpler, maybe better," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 3, pp. 204–210, 2004.

[34] X. Shiquan, Z. Qinyu, and L. Wei, "PSO-based OFDM adaptive power and bit allocation for multiuser cognitive radio system," in *Proceedings of the 5th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '09)*, September 2009.

[35] C.-J. Liao, C.-T. Tseng, and P. Luarn, "A discrete version of particle swarm optimization for flowshop scheduling problems," *Computers and Operations Research*, vol. 34, no. 10, pp. 3099–3111, 2007.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 561984, 9 pages
http://dx.doi.org/10.1155/2013/561984

Research Article

A Hybrid OFDM-TDM Architecture with Decentralized Dynamic Bandwidth Allocation for PONs

Taner Cevik

Computer Engineering Department, Fatih University, 34500 Istanbul, Turkey

Correspondence should be addressed to Taner Cevik; tcevik@fatih.edu.tr

Received 4 August 2013; Accepted 27 August 2013

Academic Editors: W. Sun, G. Zhang, and J. Zhou

One of the major challenges of passive optical networks is to achieve a fair arbitration mechanism that will prevent possible collisions from occurring at the upstream channel when multiple users attempt to access the common fiber at the same time. Therefore, in this study we mainly focus on fair bandwidth allocation among users, and present a hybrid Orthogonal Frequency Division Multiplexed/Time Division Multiplexed architecture with a dynamic bandwidth allocation scheme that provides satisfying service qualities to the users depending on their varying bandwidth requirements. Unnecessary delays in centralized schemes occurring during bandwidth assignment stage are eliminated by utilizing a decentralized approach. Instead of sending bandwidth demands to the optical line terminal (OLT) which is the only competent authority, each optical network unit (ONU) runs the same bandwidth demand determination algorithm. ONUs inform each other via signaling channel about the status of their queues. This information is fed to the bandwidth determination algorithm which is run by each ONU in a distributed manner. Furthermore, Light Load Penalty, which is a phenomenon in optical communications, is mitigated by limiting the amount of bandwidth that an ONU can demand.

1. Introduction

In parallel with the improvements occurring at the backbone side, more sophisticated and bandwidth harvesting user demands have emerged lately. In spite of this mutual capacity and demand increase, the portion residing between the user and the backbone which is called the last (first) mile remains as the bottleneck. The recent technology employed in the last mile is the DSL technology. However, this solution will soon not be able to fulfill the bandwidth requirements of the users running multimedia and real time services such as video conferencing [1, 2]. 6 Mbits/s bandwidth per user was satisfying for all users previously. However, by the rapid progress of multimedia applications offered via Internet such as High Definition TV, existing resources have begun not to meet the needs of some users. These users running bandwidth harvesting applications needed broadband access capacity in order not to suffer delays and experience low-quality performance. Furthermore, due to the improvements in optical signal processing and devices, deployment of optical architectures in the last mile is not as costly as before [3].

In order to remedy that bandwidth bottleneck, optical fiber has been deployed instead of traditional copper cable in the last mile. This architecture is called Fiber-to-the-x (FTTx). This appellation differs according to the distance between the end user and the optical fiber [4].

Besides serving the broadband bandwidth to the users, achieving it in lower costs is another challenge to overcome. PONs are considered as the premising solution for that aforementioned challenge by means of their low operational costs, longevity, and huge capacity [5]. In addition to conventional Internet, CATV, and multimedia data, it also supports wireless data such as 3G and 4G [6].

All the elements assigned on the network between source and destination, are passive which means no electrical supply is needed [7]. By means of a splitter, fewer number of fibers are deployed, thereby cost and complexity are reduced. On the contrary, in Active Optical Networks (AON), an active curb switch is employed instead of a passive splitter/combiner. That requires electrical supply, physical preservation, and administration of the switch. However, passive splitter just replicates the incoming signal without performing

any operations on it. In another type of optical network called point-to-point Optical Networks, individual fibers are deployed for each subscriber. That is, each user is connected to the central office through the fiber assigned specifically for it which is a very cumbersome and expensive solution.

PONs roughly comprise three elements. The main component of a PON is the Optical Line Terminal (OLT) device that is placed at the local exchange part (Central Office). OLT is responsible for signal conversion between optical and electrical plane. Furthermore, it coordinates multiple access of ONUs to the common fiber towards upstream direction [8]. Another component is the Optical Network Unit (ONU) which is located at the user side and serves as an interface between the OLT and the user domain. ONUs are usually assigned for every user premise such as a curb or a building. ONUs perform the conversion of optical signals into electrical domain and vice versa. Moreover, they schedule users, thereby provide efficient service quality, fair bandwidth usage and better channel utilization. The last important element of a PON is the Optical Splitter/Combiner that is charged to split the optical signal incoming from OLT into several subsignals and distribute them via the fibers connected to the ONUs placed at the premises side.

ATM is defined as a standard layer 2 protocol of PONs by the Full Service Access Network (FSAN) in the hope that ATM networks will dominate the market. However, Ethernet has won a great victory against ATM and gained great acceptance especially at the user side of PONs. In ATM Passive Optical Networks (APONs), whole IP datagram is segmented into fixed size ATM cells. If one of the cells is dropped or disrupted somehow, all remaining cells comprising the whole IP datagram have to be resent again, even though they arrive to the destination without any corruption. Moreover, the IP datagram cannot be fragmented. That is to say, when an IP datagram is segmented into ATM cells and the remaining part does not contain enough data to fill out completely an ATM cell, empty bits of the ATM cell will be padded which brings an extra overhead. Also, failing to be an inexpensive technology is another important factor for ATM not being considered at the user side of PONs. Due to above-mentioned reasons, APONs have not gained the targeted interest [9].

Ethernet PON (EPON) is an inexpensive, ubiquitous technology that is interoperable with the legacy equipment while supporting different types of data requiring various service qualities. That makes EPONs much more popular and acceptable in the client side. Depending on the fact that 90% of data traffic emerges from or ends at an Ethernet LAN, it is reasonable of Ethernet to be the standard layer-2 protocol for PONs.

This paper presents a hybrid OFDM/TDM architecture that provisions enhanced service quality, fair, dynamic bandwidth allocation, and utilization for EPONS. Nearly all dynamic bandwidth allocation algorithms suggested for Time Division Multiplexing (TDM) architectures, offer very close solutions and performances. Furthermore, most of the TDM architectures are centralized. The most attractive solution breaking this uniformity is the method that applies echo mechanism at the splitter, thereby all the ONUs become aware of each other's transmission. A latter solution offered for multiple access is the Wavelength Division Multiplexing (WDM) in which different wavelengths $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ are assigned to each ONU. Though it is a simple solution to implement, since it requires tunable receiver arrays at the OLT, it is not cost effective. Orthogonal Frequency Division Multiplexing (OFDM) which is a popular wireless broadband access technology has been considered lately to be employed in PONs. By utilizing the penetrating performance of OFDM mechanism, we present a decentralizing dynamic bandwidth allocation algorithm by means of using echo mechanism in the splitter.

The remainder of this paper is presented as follows. Section 2 summarizes the multiple access schemes and dynamic bandwidth allocation algorithms utilized in EPONs; Section 3 describes the architecture we propose; in Section 4, simulation results are depicted and the last section concludes the paper.

2. Scheduling and Dynamic Bandwidth Allocation

In EPON, communication channel is divided into two subchannels, upstream and downstream. In the downstream direction, PON acts as a point-to-multipoint network. Data sent by the OLT is broadcasted to all ONUs by means of the splitter. ONUs getting data look at the address of the packets whether the designated subscriber belongs to their domain or not. If it is not, packets are discarded. Encryption is applied in order to prevent eavesdropping. Since broadcast mechanism is employed, problems like multiple accessing in the upstream do not occur. On the other hand, some challenges arise in the upstream direction. All ONUs share the same fiber and they inherently attempt to access to the common communication channel at the same time. Since it was nearly impossible for ONUs to detect collisions occurring at the OLT, traditional collision detection mechanisms such as Ethernet's CSMA/CD mechanism was previously difficult to be employed in EPONs [10]. However, development and usage of echo splitters have let ONUs detect other ONUs' signals, thereby requiring CSMA/CD to be applied. To this end, different solutions have been suggested and some of these are described generally in this section.

Dynamic Bandwidth Allocation Solutions depend on the multiplexing mechanism employed. Therefore, they are grouped roughly under four headings, TDM, CSMA/CD, WDM, and OFDM solutions.

2.1. TDM. First and the most effective solution offered is the application of TDM in the upstream direction. Time domain is sliced into windows and each ONU is assigned a fix-sized window permanently. This mechanism is very simple and easy to implement. However, it is not an ideal solution due to the bursty nature of the data traffic. Though an ONU does not have any data, the transmission window previously assigned for that ONU cannot serve any other ONU, and the fiber will redundantly remain idle which results in low bandwidth utilization. With the aim of achieving high bandwidth

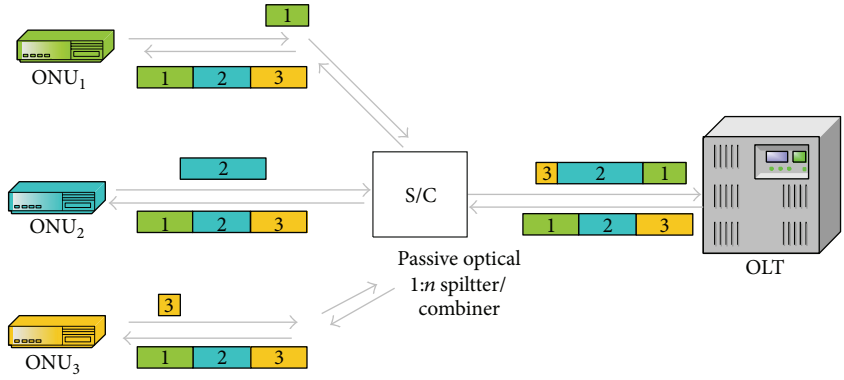


Figure 1: Traffic flow with TDM.

utilization, transmission time windows should be assigned dynamically with varying sizes depending on the burstiness of data [11] as shown in Figure 1.

In order to provide a better channel utilization, resources should be assigned to demanding ONUs temporarily which is called Dynamic Bandwidth Allocation [12]. Moreover, since time domain is being multiplexed, transmissions of the ONUs must be scheduled depending on different criteria. Some of the proposed solutions employ round-robin mechanism and some of them use adaptive scheduling schemes [13–15].

Scheduling is performed in two domains. First one is the Inter-ONU Scheduling, which is the job of organizing the transmissions of multiple ONUs in the system. ONUs inform the OLT about their bandwidth demands. OLT performs required planning operations and replies back to ONUs with their transmission time and length schedules. Another type of scheduling is the Intra-ONU scheduling which is performed by the ONUs in the network. After each ONU gets its GRANT message, obtained bandwidth is apportioned among the users existing in the ONU's domain, depending on their service level agreements or bandwidth demands [16].

All mechanisms suggested for dynamic bandwidth allocation under TDM group are based on multipoint Control Protocol (MPCP) [17]. MPCP is a signaling protocol used by both OLT and ONUs during bandwidth allocation process. MPCP mainly relies on two messages: REPORT and GATE. GATE message is used by the OLT for allocating bandwidth to the ONUs. ONUs send their data together with the REPORT message that carries information about the present situation about the queues of the related ONU.

Various DBA mechanisms are suggested for the purpose of remedying dynamic bandwidth allocation. As is known, the most popular and fundamental study is IPACT, in which an interleaved polling scheme with an adaptive cycle time is presented. The next ONU is polled, before the data of the previous ONU arrives. In IPACT, for in-band signaling a single wavelength is used for both GRANT message and downstream data transmissions. Besides, they suggest that, in order to prevent some ONUs with larger demands to dominate the whole bandwidth, a maximum window size is defined. If an ONU demands a window size larger than the maximum threshold, the maximum bandwidth it can get is the predefined maximum window size. Demands smaller than the maximum window size will be fulfilled by the OLT.

Another similar work presented is an adaptive MAC polling protocol suggested by Zheng and Mouftah [11]. This protocol mainly depends on IPACT and the amendment they suggest is defining the ONU that should send its data first by considering their RTTs.

The problem that emerges during light network loads is the light load penalty identified by Kramer et al. [18]. At the end of every cycle, each ONU puts information about its queues' status into the REPORT message that is attached to the data portion. Obviously, this information denotes the recent bandwidth demand of that ONU. If demand of the ONU is smaller than or equal to the maximum transmission window size, the requested bandwidth is supplied to the ONU by giving the details in the GRANT message. However, until the GRANT message arrives at the ONU, new packets with different priorities inherently may arrive at the ONU from the users located in its domain. The bandwidth granted by the OLT covers the packets arrived previously and the newly arriving high priority packets will suppress the previously arrived lower priority packets. These lower priority packets will have to wait for the next cycle. However, same events may repeat in the next cycle, and the same lower priority packets may suffer unlimited delays and as a result will be dropped. Kramer et al. suggest a two-stage buffer mechanism in which packets arriving at the ONU are placed into different queues depending on their priorities. These queues are called stage 1 queues. Depending on a pulling algorithm, packets are pulled from these differentiated queues and passed to the queue that holds the packets to be transmitted in the next cycle. So, the REPORT message holds information about the status of this stage 2 queue. Newly arriving packets are directly pushed into the stage 1 queues. Thus, they cannot dominate the lower priority packets that have gained transmission grant for the present cycle.

Studies described above are based on the idea of centralized bandwidth allocation and scheduling which is performed by the OLT. ONUs send their bandwidth demands in a REPORT message attached after the data portion. Depending on these requests, OLT allocates bandwidth and organizes

the schedule. Despite that, in decentralized solutions, OLT does not have any mission anything about bandwidth allocation and organization process. On the contrary, ONUs decide everything when and how long they will access to the common fiber. This is achieved by means of an echo featured splitter that provides each signal sent by an ONU to be received by the remaining ONUs. In the paper proposed by Sherif et al. [19], time is portioned into frames each of which represents a cycle. A small slice at the beginning of the cycle is divided into small slots that are each assigned to an ONU permanently to announce its bandwidth demand and make reservation. Each ONU broadcasts its demand in its dedicated time slot. After this notification process, all ONUs in the system become aware of the demands of other ONUs. The values obtained from other ONUs are processed by a DBA algorithm. Since every ONU runs the same algorithm, the results produced at each ONU are all consistent. Hence, the schedule becomes definite and every ONU starts its transmission at the time the algorithm identifies.

2.2. CSMA/CD. As described at the beginning of this section, CSMA/CD couldn't have been applied in passive optical networks previously, because, it was impossible for an ONU to be aware of another ONUs' transmission. There were two major factors causing ONUs not to be able to realize the signals of other ONUs. First one is that splitters lacked echo (repeater) mechanism. Other reason was that it was very difficult for ONUs to sense a collision occuring at the OLT. By the development of repeater type passive splitters, a signal transmitted by an ONU is echoed at the splitter to all other ONUs. Thus, each ONU realizes an ongoing transmission and does not attempt to send its data in order not to cause a collision which is an obvious application of CSMA/CD.

RCMA [20] and its extended version FULL-RCMA [21] propose a MAC protocol similar to CSMA/CD. They employ the echo mechanism by means of a passive optical splitter/combiner similar to the 10BASE-FP Ethernet. According to their solution, an ONU, before sending its data, should broadcast a request message. This request message is repeated at the echo splitter and arrives at all other ONUs together with the originating one. When the echoed request message arrives back at the originating ONU, it is discarded, and not evaluated as a collision by means of a special transceiver designed for Ethernet. The request message is used for making reservation in time domain. At that time, all other ONUs become aware of how long the common medium will be busy. Each ONU applies this process and makes its own reservation. Before an ONU broadcasts its requests, it first senses the medium. If it is free to send, then it puts the request message on the way towards the splitter.

2.3. WDM. TDM PONs with statistical multiplexing offers low cost network maintenance and administration. However, the common fiber can be used by only a single ONU at a time. Considering bandwidth demands of the clients increase very rapidly nowadays, a promising solution for fulfilling these requirements is the WDM technology. Capacity increase can be achieved in the backbone easily by using WDM. However, employing WDM in the access side has been a much more complicated and expensive challenge. In terms of cost of the devices and maintenance, WDM networks are considered as the next-generation follower of TDM PONs. Thus, hybrid TDM-WDM techniques are being proposed as a smooth adaptation bridge for the recent stage [22].

The most important factor that affects WDMs to be an expensive technology to be applied in the access side is the high cost of the WDM light source. Distributed feedback laser diode used in the backbone is not convenient for the access side [23]. However, promising, low cost WDM light sources are proposed such as Amplified Spontaneous Emission (ASE)-injected Fabry-Perot Laser Diode (FP-LD) [24].

In nondynamic WDM networks, ONUs have a single light source (FP-LD) fixed to a single wavelength. Though this approach seems to be a cost effective solution to implement under light loads; the assigned wavelength will not be fully utilized [25]. Using multiple FP-LDs each fixed to a single wavelength at each ONU is another alternative solution. However, the best case is that each ONU owns a tunable laser that can use any of the upstream channel wavelengths. Though full tunability scheme provides the best performance in terms of bandwidth utilization, it requires a centralized tunable DeMux at OLT which is an expensive solution [26].

2.4. OFDM. Another promising technology that is not as costly as WDM for PONs is the Orthogonal Frequency Division Multiplexing. Application of OFDM on PONs has drawn great interest in recent years. By dividing the baseband carrier to the multiple orthogonal subcarriers, significant bandwidth utilization, and enlargements can be achieved [27].

Several research studies [28–30] have been performed about OFDMA PONs. An OFDMA-based 10 Gb/s PON architecture is demonstrated by Qian et al. [31]. In another work [32], a dynamic, asynchronous OFDMA scheme is proposed. Uplink channel is divided into orthogonal subcarriers again. Besides, each ONU has allocated a subchannel permanently dedicated for control messages transmission. Each ONU sends its recent demand to the OLT via that dedicated signaling channel. Depending on these demands arriving from ONUs, OLT allocates the necessary resources to ONUs by means of GRANT messages. In this way, a centralized statistical bandwidth allocation mechanism is achieved.

In the architecture, we propose a statistical subcarrier allocation to each ONU depending on their recent demands. However, instead of charging the OLT for giving this bandwidth allocation decisions, a decentralized approach is employed in which each ONU runs an algorithm that produces the same output at every ONU. Thus, delay suffered because of the REPORT and GRANT signaling messages is alleviated.

3. The Proposed Architecture

In traditional TDM PONs, just a single ONU can access to the common channel at a single time. However, by allocating a portion of the capacity to each of the ONUs in the system,

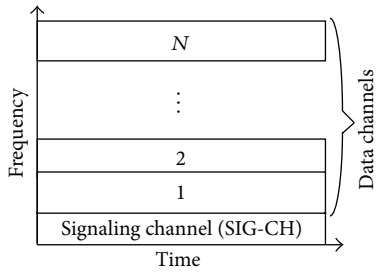


Figure 2: Uplink channel partition in frequency domain.

multiple of them can access to the common fiber simultaneously. This simultaneous multiple access is achieved by applying OFDMA scheme in our architecture. Uplink channel is apportioned to multiple orthogonal subchannels as shown in Figure 2.

Subchannel allocation is not performed centrally by the OLT. As in the previous works, unnecessary delays occur during signaling process. In contrast, a decentralized mechanism is employed. Each ONU in the system runs the same Dynamic Subchannel Allocation algorithm that produces the same output at each ONU. In order to produce the same output, the algorithm has to be applied to the same inputs. These inputs are the bandwidth demands of each ONU in the system. So, these demands should be notified to every ONU. One of the subchannels is assigned for this notification process. Access to the signaling subchannel is achieved in the TDM manner. Hence, signaling subchannel is partitioned into mini-slots in the time domain as depicted in Figure 3, each of which is assigned to an ONU permanently.

3.1. System Architecture. Instead of a traditional splitter/combiner, an echo-splitter/combiner [33] is utilized in the system. Thus, each ONU becomes aware of each other's transmissions via a common signaling channel as shown in Figure 4.

Architecture of an ONU in the system is sketched in Figure 5. Numbers and arrows in the figure denote the order and direction of operations, respectively. In order to prevent light-load penalty, two stage queue mechanism is employed in the ONUs. Packets incoming from users (1) are conducted to the appropriate queues by the classifier depending on their types (2). At the beginning of each cycle, Load Calculation Unit (LCU) calculates the total load (3) and passes this information to the Communication Unit (CU) and Dynamic Subchannel Allocation (DSCA) Unit (4). CU broadcasts the load information in its permanently allocated mini-slot period (5) to inform other ONUs about its load. Also, all other ONUs' notification broadcast messages arrive during this load notification stage (6). All incoming load information is processed by the DSCA Unit, and the result which represents the amount of bandwidth that the corresponding ONU gained is conveyed to the Scheduler (8). Scheduler pulls packets from the queues according to the priority management policy without exceeding the bandwidth defined by the DSCA (9). Packets leaving the Stage1 Queue Unit enter to the stage 2 Queue Unit to be sent in the present cycle (10). Lastly, packets

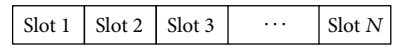


Figure 3: Signaling channel partition in time domain.

ordered in the second queue are forwarded to the CU to be transmitted (11).

3.2. Intra-ONU Scheduling and Decentralized Dynamic Bandwidth Allocation. Network traffic is classified as in Diffserv, [34] and packets are pushed into different queues depending on their types. As shown in Figure 5, each ONU maintains three types of queues that each holds different prioritized packets. ConstantBitRate (CBR) stands for Diffserv's Expedited Forwarding (EF) representing services such as voice services that need low delay jitter. VariableBitRate (VBR) symbolizes Diffserv's Assured Forwarding services, such as video that can tolerate some delay and loss. Lastly, BestEffort (BE) is the lowest prioritized services in the hierarchy [35].

Length of each queue that holds different priority packets are passed to the LCU. Bandwidth demand is calculated according to the load present in the queues. The length of high priority queues are the major decisive factor while calculating the load of each ONU. Load of an ONU is calculated by (1):

$$\mathrm{BWD}_{\mathrm{ONU}(i)} = (\alpha * \mathrm{LqCBR}) + (\beta * \mathrm{LqVBR}) + (\lambda * \mathrm{LqBE}). \quad (1)$$

In the formula given previously, $\mathrm{BWD}_{\mathrm{ONU}(i)}$ denotes the total load of an ONU, in other words, total bandwidth demand of the ONU. LqCBR, LqVBR, and LqBE represent the lengths of CBR, VBR, and BE queues of ONU(i), respectively.

Several simulations are performed for various (α, β, γ) triples. As it will be clarified in the following sections, in order to provide better service qualities for high prioritized packets, larger ratio of $(\alpha/\beta/\gamma)$ must be applied. By this way, ONUs with larger CBR queue lengths will possess more bandwidth.

Total bandwidth demand in the system, which is denoted by TotBwD, is calculated by adding demands of all ONUs in the network:

$$\mathrm{TotBwD} = \sum_{i=1}^{N} \mathrm{BWD}_{\mathrm{ONU}(i)}. \quad (2)$$

In order to provide fairness, there is a maximum bandwidth value (MaxBw) that an ONU can demand. That is because, if all bandwidth demands of ONUs are satisfied carelessly, ONUs with lower demands will never get a chance to reach to the common fiber. Thus, in order to prevent Light Load Penalty, the amount of bandwidth that an ONU can demand is limited in this way. Calculation of the bandwidth allocated to the ONUs is performed in two stages. In the first stage, demands smaller than MaxBw is satisfied completely. If the value that the ONU demands exceeds MaxBw in the first place, a minimum guaranteed bandwidth (MinGrBw) is allocated to that ONU. In this way, ONUs with larger demands

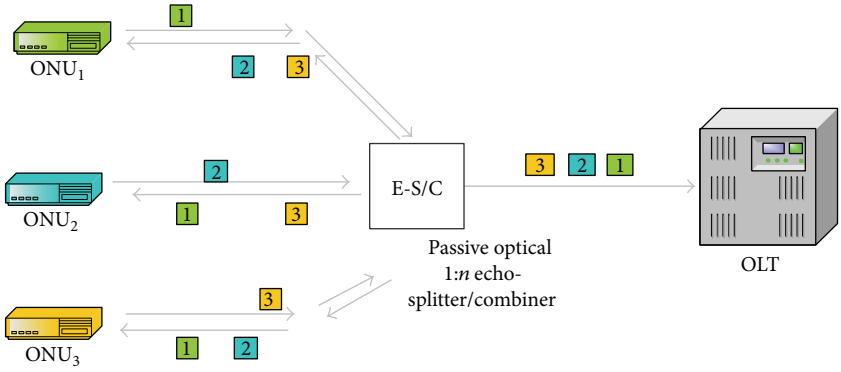


Figure 4: Load notification of ONUs via signaling channel by means of the echo splitter.

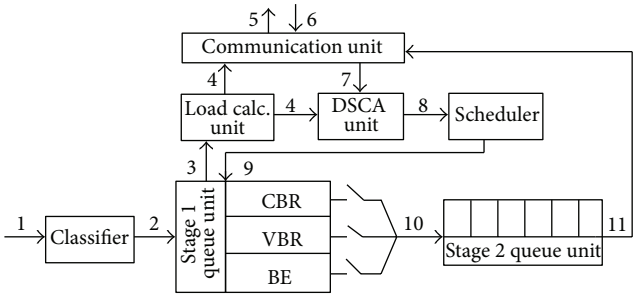


Figure 5: ONU architecture.

are prevented to dominate the common media and also the ONUs with lower demands get their share fairly. Bandwidth allocation process in the first stage is formulated as follows:

$$\mathrm{AlBw}_{\mathrm{ONU}(i)} = \begin{cases} \mathrm{BWD}_{\mathrm{ONU}(i)}, & \mathrm{BWD}_{\mathrm{ONU}(i)} \leq \mathrm{MaxBw}, \\ \mathrm{MinGrBw}, & \mathrm{BWD}_{\mathrm{ONU}(i)} > \mathrm{MaxBw}. \end{cases} \quad (3)$$

After the bandwidth allocation at the first stage, the remaining bandwidth demand of each ONU is shown by

$$\mathrm{RemBwD}_{\mathrm{ONU}(i)} = \mathrm{BWD}_{\mathrm{ONU}(i)} - \mathrm{AlBw}_{\mathrm{ONU}(i)}. \quad (4)$$

Equation (5) represents the total remaining bandwidth demand after the first stage:

$$\mathrm{TotRemBwD} = \sum_{i=1}^{N} \mathrm{RemBwD}_{\mathrm{ONU}(i)}. \quad (5)$$

The total allocated and remaining available bandwidths after the first stage are calculated, respectively, as follows:

$$\mathrm{TotAlBw} = \sum_{i=1}^{N} \mathrm{AlBw}_{\mathrm{ONU}(i)}, \quad (6)$$

$$\mathrm{RemAvBw} = \mathrm{TotBw} - \mathrm{TotAlBw}. \quad (7)$$

In (7), TotBw and RemAvBw denote the total bandwidth capacity of the network and the remaining available bandwidth after the first stage, respectively. In the second stage, residual bandwidth allocation to the ONUs is performed proportional with their remaining demands as identified by

$$\mathrm{AlBw}_{\mathrm{ONU}(i)+} = \frac{\mathrm{RemBwD}_{\mathrm{ONU}(i)}}{\mathrm{TotRemBwD}} * \mathrm{RemAvBw}. \quad (8)$$

Channels are assigned to the ONUs as shown in Figure 6 according to the results calculated by the decentralized dynamic bandwidth allocation algorithm. As represented in Figure 6, time is partitioned into cycles. Each cycle comprises two phases: Load Notification Phase (LDP) and Data Transfer Phase (DTP). During LDP which is negligible compared to the DTP, an ONU broadcasts its load information towards the splitter via the signaling channel in its permanent mini-timeslot. When the broadcasted notification message arrives at the echo-splitter/combiner, it is reflected back towards the other ONUs together with the originating one. By looking at the source id of the message, the originator discards it. Other ONUs check the destination address of the message and after realizing the broadcast address and the type of the message which identifies a notification, their CUs take and pass the message to their DSCA units.

4. Performance Evaluation

In this section, we present the results of the simulations performed with different parametric combinations. Simulations are performed on an event-based testbed. We use a PON with 10 ONUs connected in a stare-shape topology. The uplink channel data rate between each ONU and the splitter is assumed as 1 Gb/s with a total capacity of 10 Gb/s. Cycle length in time domain is set to 20 ms. Total bandwidth is divided into 2048 subchannels each with the capacity of 5 Mb/s. Number of users connected to each ONU is assumed to be the same. However, not all of the users connected to an ONU generate packets. Each user generates packets randomly in each cycle. As described previously, three different types of packets are present in the network: CBR, VBR, and BE. CBR packets are generated with the rate of 4.48 Mb/s and 70 Bytes length. Sizes of VBR and BE packets are assumed to be 100 Bytes.

We first identify the impact of the coefficients α, β, and λ on the maximum delay suffered in the system. As described in (1), load of an ONU which actually symbolizes the bandwidth

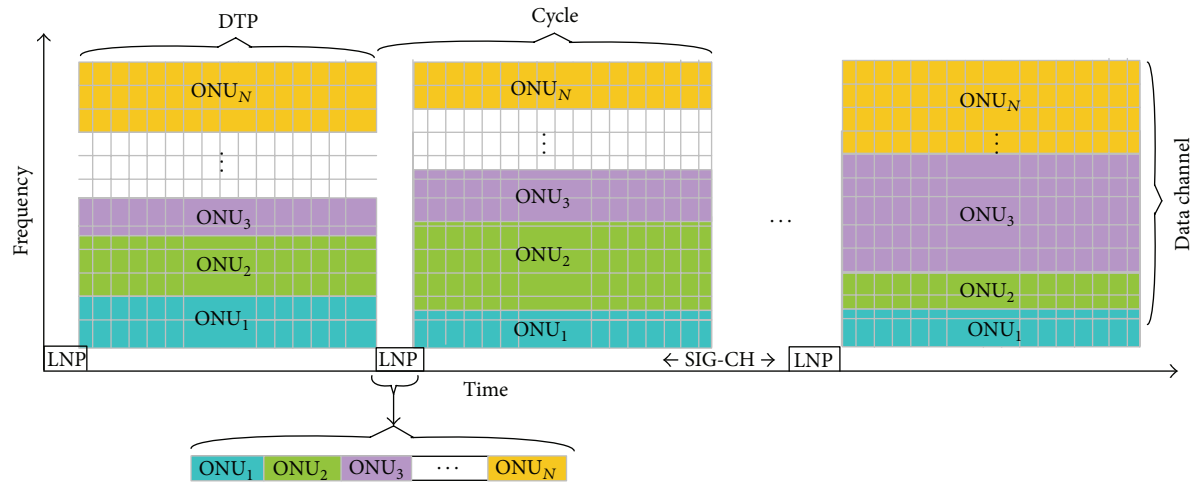


Figure 6: Channel assignment.

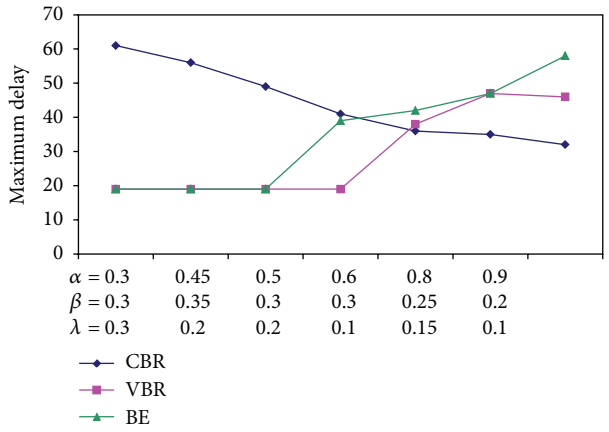


Figure 7: Maximum delay suffered due to different combinations of (α, β, λ) triple.

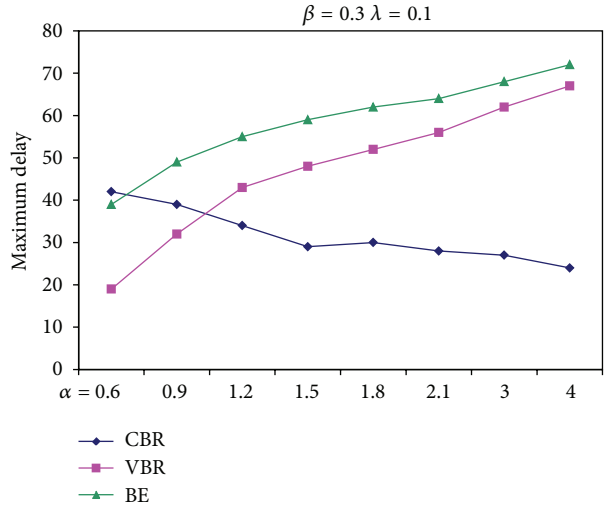


Figure 8: Maximum delay suffered for different values of α when β and λ are kept constant.

demand of an ONU is calculated according to this triplet (α, β, λ) value.

Figure 7 shows the maximum delay suffered depending on different (α, β, λ) combinations. As the ratio of $\alpha/(\beta, \lambda)$ increases, end-to-end delay of a CBR packet decreases. The reason of the decrease in delay is that $\alpha/(\beta, \lambda)$ ratio describes the bandwidth reserved for CBR packets. Figure 8 also encourages the notion asserted above. It is depicted the change in the maximum delay CBR, VBR, and BE packets suffered due to varying values of α when β and λ are kept constant.

Since the parameter β describes the weight of VBR packets during bandwidth reservation, as shown in Figure 9, the increase in the value of β when α and λ are kept constant, the delay of VBR packets decreases.

The same goes for the BE packets. As presented in Figure 10, when it is intended to prevent the CBR and VBR packets to dominate the BE packets in the system, value of λ is increased.

Another important factor affecting the delay is the network traffic. By the time the number of active users in the system increases, the network load also increases. The change in delay due to the increase in the number of users attached to each ONU is shown in Figure 11.

As described in the previous section, during bandwidth partition for each ONU, if it is not prevented, an ONU with a higher load can dominate the total capacity. In this case, users of ONUs with lower loads may suffer to significant delays. In order to avoid this situation, all of the ONUs are given some amount of guaranteed bandwidth which is denoted by MinGrBw in (3). The remaining bandwidth is shared rationally between the ONUs of which bandwidth demands are not fulfilled. As clarified in Figure 12, if the amount of minimum guaranteed bandwidth is increased, the delay that the CBR packets suffer also increases.

5. Conclusion

This paper proposes a hybrid OFDM-TDM architecture with a statistical bandwidth allocation scheme. Traditional

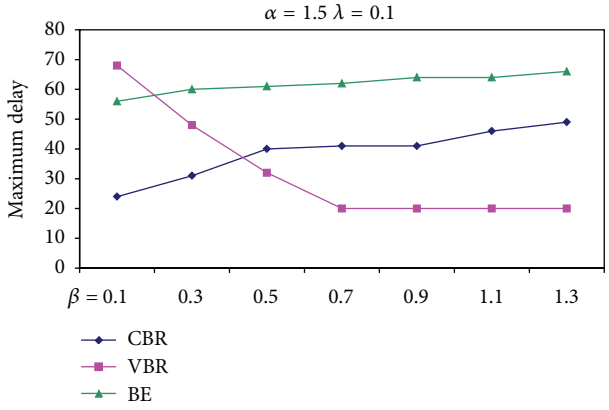


Figure 9: Maximum delay suffered for different values of β when α and λ are kept constant.

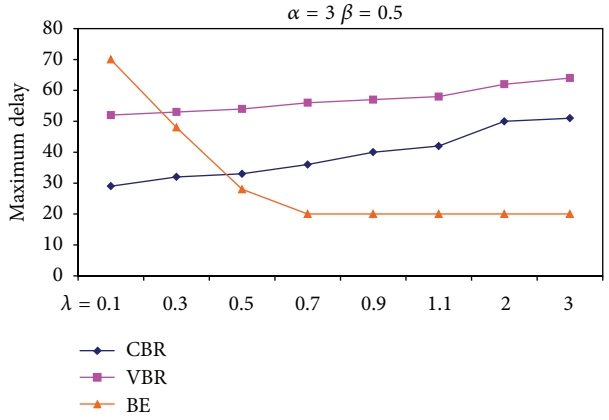


Figure 10: Maximum delay suffered for different values of λ when α and β are kept constant.

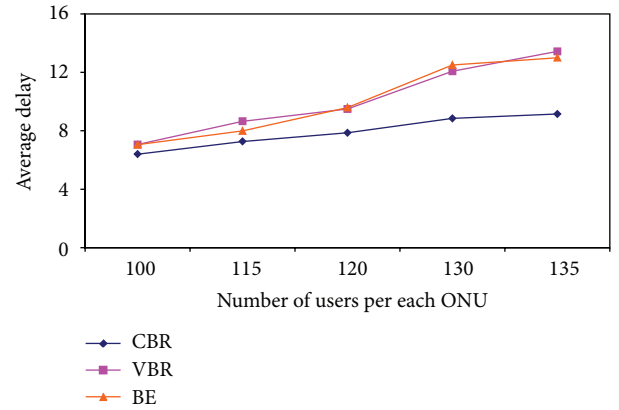


Figure 11: Change in delay due to increase in the number of users.

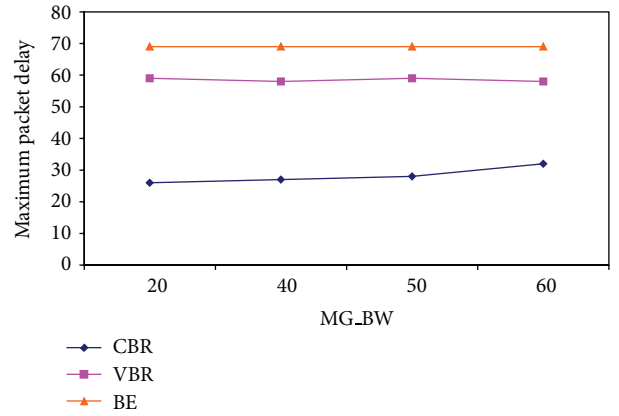


Figure 12: Change in delay due to increase in the amount of minimum guaranteed bandwidth.

centralized structures mostly apply TDM mechanism as the multiple access scheme. With TDM, just a single ONU owns the common shared medium at a time. However, apportioning the available resources among the beneficiaries yields a much more effective and fair utilization. WDM is an alternative and promising solution for this purpose. Though it is an applicable technology in the backbone, it is not cost effective in the access side because it requires a centralized tunable DeMux in OLT which is an expensive solution. One of the latest prominent solutions for PONs is the application of OFDM in the access side. OFDM provides better shared resource access with lower costs. In this paper, TDM is employed as the accompanying multiple access scheme for coordinating the bandwidth apportioned among the users that is assigned by OFDM. Besides, by applying a decentralized architecture, redundant delays suffered because of the signaling between the ONUs and the OLT are prevented. After a short notification stage, each ONU becomes aware of the loads of other ONUs in the system. Load calculation is performed according to the weights of CBR, VBR, and BE packets in the total load. Hence, high priority packets and ONUs dominating the resources are prevented. Allocating resources dynamically in each cycle according to the demands of ONUs provides better resource utilization.

For future work, a performance comparison by employing multiple OLTs in the system will be added.

References

[1] G. Kramer, *Ethernet Passive Optical Networks*, McGraw-Hill Professional, New York, NY, USA, 2005.

[2] G. Kramer and G. Pesavento, "Ethernet passive optical network (EPON): building a next-generation optical access network," *IEEE Communications Magazine*, vol. 40, no. 2, pp. 66–73, 2002.

[3] D. Nowak and J. Murphy, "FTTH: the overview of existing technologies," in *Opto-Ireland: Optoelectronics, Photonic Devices, and Optical Networks*, vol. 5825 of *Proceedings of the SPIE*, pp. 500–509, Dublin, Ireland, April 2005.

[4] G. Keiser, *FTTX Concepts and Applications*, John Wiley & Sons, 2006.

[5] M. P. McGarry, M. Reisslein, and M. Maier, "WDM ethernet passive optical networks," *IEEE Communications Magazine*, vol. 44, no. 2, pp. 15–22, 2006.

[6] F. Rawshan and Y. Park, "Architecture of multi-OLT PON systems and its bandwidth allocation algorithms," *Photonic Network Communications*, vol. 25, no. 2, pp. 95–104, 2013.

[7] G. Kramer, B. Mukherjee, and A. Maislos, "Ethernet passive optical network (EPON)," in *Multiprotocol Over DWDM: Build-Ing the Next Generation Optical Internet*, John Wiley & Sons, 2003.

[8] J. Zhang, T. Wang, and N. Ansari, "Designing energy-efficient optical line terminal for TDM passive optical networks," in *Proceedings of the 34th IEEE Sarnoff Symposium (SARNOFF '11)*, pp. 1–5, Princeton, NJ, USA, May 2011.

[9] M. P. McGarry, M. Maier, and M. Reisslein, "Ethernet PONs: a survey of dynamic bandwidth allocation (DBA) algorithms," *IEEE Communications Magazine*, vol. 42, no. 8, pp. S8–S15, 2004.

[10] C. M. Assi, M. Maier, and A. Shami, "Toward quality of service protection in ethernet passive optical networks: challenges and solutions," *IEEE Network*, vol. 21, no. 5, pp. 12–19, 2007.

[11] J. Zheng and H. T. Mouftah, "An adaptive MAC polling protocol for ethernet passive optical networks," in *Proceedings of the IEEE International Conference on Communications (ICC '05)*, vol. 3, pp. 1874–1878, May 2005.

[12] C. C. Sue and T. C. Lee, "Elastic ONU-class-based idle time elimination algorithm for Ethernet passive optical networks," *Photonic Network Communications*, vol. 25, pp. 105–119, 2013.

[13] M. Ma, Y. Zhu, and T. H. Cheng, "A bandwidth guaranteed polling MAC protocol for ethernet passive optical networks," in *Proceedings of the 22nd Annual Joint Conference on the IEEE Computer and Communications Societies*, vol. 1, pp. 22–31, April 2003.

[14] G. Kramer, B. Mukherjee, and G. Pesavento, "IPACT: a dynamic protocol for an Ethernet PON (EPON)," *IEEE Communications Magazine*, vol. 40, no. 2, pp. 74–80, 2002.

[15] C. M. Assi, Y. Ye, S. Dixit, and M. A. Ali, "Dynamic bandwidth allocation for quality-of-service over Ethernet PONs," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 9, pp. 1467–1477, 2003.

[16] J. Zheng and H. T. Mouftah, "Media access control for ethernet passive optical networks: an overview," *IEEE Communications Magazine*, vol. 43, no. 2, pp. 145–150, 2005.

[17] IEEE 802. 3ah Ethernet in the First Mile Task Force, http://www.ieee802.org/3/efm/public/nov01/kramer_1_1101.pdf.

[18] G. Kramer, B. Mukherjee, S. Dixit, Y. Ye, and R. Hirth, "Supporting differentiated classes of service in Ethernet passive optical networks," *Journal of Optical Networking*, vol. 1, no. 8, pp. 280–298, 2002.

[19] S. R. Sherif, A. Hadjiantonis, G. Ellinas, C. Assi, and M. A. Ali, "A novel decentralized ethernet-based PON access architecture for provisioning differentiated QoS," *Journal of Lightwave Technology*, vol. 22, no. 11, pp. 2483–2497, 2004.

[20] C. H. Foh and M. Zukerman, "A novel and simple MAC protocol for high speed passive optical LANs," in *Proceedings of the 2nd International IFIP-TC6 Networking Conference On Networking Technologies, Services, and Protocols, Performance of Computer and Communication Networks, and Mobile and Wireless Communications*, pp. 467–478, 2002.

[21] C. H. Foh, L. Andrew, E. Wong, and M. Zukerman, "FULL-RCMA: a high utilization EPON," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 8, pp. 1514–1524, 2004.

[22] D. J. Shin, D. K. Jung, H. S. Shin et al., "Hybrid WDM/TDM-PON with wavelength-selection-free transmitters," *Journal of Lightwave Technology*, vol. 23, no. 1, pp. 187–195, 2005.

[23] S.-J. Park, C.-H. Lee, K.-T. Jeong, H.-J. Park, J.-G. Ahn, and K.-H. Song, "Fiber-to-the-home services based on wavelength-division-multiplexing passive optical network," *Journal of Lightwave Technology*, vol. 22, no. 11, pp. 2582–2591, 2004.

[24] H. D. Kim, S.-G. Kang, and C.-H. Lee, "A Low-cost WDM source with an ASE injected Fabry-Perot semiconductor laser," *IEEE Photonics Technology Letters*, vol. 12, no. 8, pp. 1067–1069, 2000.

[25] A. Buttaboni, M. De Andrade, and M. Tornatore, "A multi-threaded dynamic bandwidth and wavelength allocation scheme with void filling for long reach WDM/TDM PONs," *Journal of Lightwave Technology*, vol. 31, no. 8, pp. 1149–1157, 2013.

[26] Y.-L. Hsueh, M. S. Rogge, S. Yamamoto, and L. G. Kazovsky, "A highly flexible and efficient passive optical network employing dynamic wavelength allocation," *Journal of Lightwave Technology*, vol. 23, no. 1, pp. 277–286, 2005.

[27] B. Yan, W. Guo, Y. Jin, and W. Hu, "A novel dynamic wavelength bandwidth allocation scheme over OFDMA PONs," in *Network Architectures, Management, and Applications IX*, vol. 8310 of *Proceedings of the SPIE*, pp. 1–5, Shanghai, China, November 2011.

[28] K. Kanonakis, E. Giacoumidis, and I. Tomkos, "Physical-layer-aware MAC schemes for dynamic subcarrier assignment in OFDMA-PON networks," *Journal of Lightwave Technology*, vol. 30, no. 12, Article ID 6148245, pp. 1915–1923, 2012.

[29] W. Lim, P. Kourtessis, M. Milosavljevic, and J. M. Senior, "Dynamic subcarrier allocation for 100 Gbps, 40 km OFDMA-PONs with SLA and CoS," *Journal of Lightwave Technology*, vol. 31, no. 7, pp. 1055–1062, 2013.

[30] K. Kanonakis, N. Cvijetic, I. Tomkos, and T. Wang, "Dynamic software-defined resource optimization in next-generation optical access enabled by OFDMA-based meta-MAC provisioning," *Journal of Lightwave Technology*, vol. 31, no. 14, pp. 2296–2306, 2013.

[31] D. Qian, H. Junqiang, and Y. Jianjun, "Experimental demonstration of a novel OFDM-A based 10Gb/s PON architecture," in *Proceedings of the 33rd European Conference on optical Communications (ECOC '07)*, pp. 1–2, Berlin, Germany, 2007.

[32] J. Zhang, T. Wang, and N. Ansari, "An efficient MAC protocol for asynchronous ONUs in OFDMA PONs," in *Proceedings of the Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC '11)*, pp. 1–3, Los Angeles, Calif, USA, March 2011.

[33] C.-J. Chae, E. Wong, and R. S. Tucker, "Optical CSMA/CD media access scheme for ethernet over passive optical network," *IEEE Photonics Technology Letters*, vol. 14, no. 5, pp. 711–713, 2002.

[34] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," Tech. Rep. IETF, RFC, 2475, 1998.

[35] Q. Liu, H. Li, Y.-F. Ji, and Y.-J. Qiao, "Resources allocation in an intserv/diffserv integrated EPON system," *Journal of China Universities of Posts and Telecommunications*, vol. 16, no. 3, pp. 108–121, 2009.