Journal of Applied Mathematics

Mathematical and Numerical Modeling of Information Dissemination in Mobile Networks

Guest Editors: Pin-Han Ho, Chih-Hao Lin, and Anyi Chen

Mathematical and Numerical Modeling of Information Dissemination in Mobile Networks

Journal of Applied Mathematics

Mathematical and Numerical Modeling of Information Dissemination in Mobile Networks

Guest Editors: Pin-Han Ho, Chih-Hao Lin, and Anyi Chen

Editorial Board

Contents

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 590872, 2 pages
http://dx.doi.org/10.1155/2013/590872

Editorial

Mathematical and Numerical Modeling of Information Dissemination in Mobile Networks

Pin-Han Ho,[1] Chih-Hao Lin,[2] and Anyi Chen[3]

[1] *Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada N2L 3G1*
[2] *Department of Information Management, Chung Yuan Christian University, Chungli 32023, Taiwan*
[3] *Center of Wireless Broadband Technology, Tatung University, Taipei 104, Taiwan*

Correspondence should be addressed to Pin-Han Ho; p4ho@uwaterloo.ca

Received 24 November 2013; Accepted 24 November 2013

This special issue covers a number of developing topics in mathematical and numerical modeling of information dissemination in mobile networks. The 15 research articles included in this special issue present original research outcomes and future evolutions of mathematics in mobility and networking. From communication mechanisms to mobile applications, the topics of this special issue are classified into three categories, namely, complex models, techniques, and applications.

The first group of papers addresses issues in the area of information dissemination via mathematical modelling approaches. In the paper of "*Efficient periodic broadcasting for mobile networks at small client receiving bandwidth and buffering space*," H.-F. Yu et al. introduced a new Fibonacci-broadcasting scheme (called FiB+) for video broadcasting and achieved smaller client buffering space than that of FiB under two-channel receiving bandwidth. In the paper entitled "*Single-channel data broadcasting under small waiting latency*," H.-F. Yu proposes a single-channel broadcasting scheme for video-on-demand services. By partitioning a video into equal-sized segments, these classified segments have been transferred over a single channel according to a predefined arrangement to yield short waiting time of data broadcasting. In the paper "*A mutual-evaluation genetic algorithm for numerical and routing optimization*," C.-H. Lin and J.-D. He present a mutual-evaluation genetic algorithm (MEGA) to find optimal flow-allocation strategies for multipath-routing problems. In the paper entitled "*Minimum-cost QoS-constrained deployment and routing policies for wireless relay networks*," F.-Y.-S. Lin et al. use Lagrangian relaxation (LR) method to minimize the development cost of wireless relay networks. In the paper entitled "*A rough penalty genetic algorithm for multicast routing in mobile ad hoc networks*," C.-H. Lin and C.-C. Chuang formulated the multicast routing problem in mobile ad hoc networks, where the objective function is to minimize the total cost of the multicast tree subject to QoS constraints. The aforementioned constrained optimization problems are solved by a proposed rough penalty genetic algorithm and achieve near-optimal solutions for a variety of multicast routing problems.

The second group concerns networking techniques and system design, such as optimal methods for the resource management and localization estimation. In the paper entitled "*Interference control for cognitive network with high mobility*," Y. Li et al. aim at maximizing the capacity of the secondary system with the interference constraints via a water-filling style method. The paper entitled "*Modeling and performance analysis of route-over and mesh-under routing schemes in 6LoWPAN under error-prone channel condition*" by T.-H. Lee et al. develops a Markov chain model to analyze the performance of two routing schemes in 6LoWPAN. In "*A multistage control mechanism for group-based machine-type communications in an LTE system*," W.-C. Hung et al. used a Markov chain with $M/G/k/k$ to analyze machine-type communications in an LTE network and proposed a multistage-control (MSC) mechanism to allocate LTE bandwidth effectively. In the paper entitled "*Calculation of weighted geometric dilution of precision*," C.-S. Chen et al. intelligently select measurement units for improving location accuracy in the proposed wireless positioning systems. In the paper entitled

"*Modeling of location estimation for object tracking in WSN*," H.-C. Chu et al. use a range-free-positioning technology as well as centralized data processing technology with data aggregation to reduce the data processing of location estimation for object tracking in large-scale WSNs.

The third group of the included papers concerns some potential applications for mobile networks. The paper entitled "*Characterizing pairwise social relationships quantitatively: interest-oriented mobility modeling for human contacts in delay tolerant networks*" presents an interest-oriented human contacts mobility model (IHC) to reproduce social relationships on a pairwise granularity for wireless mobile networks. The paper entitled "*Recovery and resource allocation strategies to maximize mobile network survivability by using game theories and optimization techniques*" exercises game theory to find the optimal resource allocation for both cyber attacker and mobile network defender. The paper entitled "*Effective proactive and reactive defense strategies against malicious attacks in a virtualized honeynet*" formulates the attack-defense scenario as a mathematical model. In the paper entitled "*Multiagent consensus control under network-induced constraints*," W. I. Kim et al. consider a mean consensus problem for multi-agent systems by using a cone-complementarity-linearization algorithm to exchange information effectively. In the paper entitled "*Application scheduling in mobile cloud computing with load balancing*," X. Wei et al. presented an appropriate architecture of mobile cloud computing and proposed a dedicated scheduling algorithm to improve the quality of service by efficiently exploiting the mobile devices' idle computing, storage, and sensing capacity.

Acknowledgments

The guest editors would like to thank all the authors for their exciting contributions and all the reviewers whose comments have significantly increased the quality of this special issue. In addition, we sincerely appreciate the editorial board members of this journal for their warm support throughout the preparation of this special issue.

Pin-Han Ho
Chih-Hao Lin
Anyi Chen

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 597981, 15 pages
http://dx.doi.org/10.1155/2013/597981

Research Article

Characterizing Pairwise Social Relationships Quantitatively: Interest-Oriented Mobility Modeling for Human Contacts in Delay Tolerant Networks

Jiaxu Chen, Yazhe Tang, Chengchen Hu, and Guijuan Wang

The Department of Computer Science & Technology, Xi'an Jiaotong University, No. 28 Xian Ning Road West, Xi'an, Shaanxi 710049, China

Correspondence should be addressed to Yazhe Tang; yztang@mail.xjtu.edu.cn

Received 19 April 2013; Accepted 20 September 2013

Academic Editor: Chih-Hao Lin

Human mobility modeling has increasingly drawn the attention of researchers working on wireless mobile networks such as delay tolerant networks (DTNs) in the last few years. So far, a number of human mobility models have been proposed to reproduce people's social relationships, which strongly affect people's daily life movement behaviors. However, most of them are based on the granularity of community. This paper presents interest-oriented human contacts (IHC) mobility model, which can reproduce social relationships on a pairwise granularity. As well, IHC provides two methods to generate input parameters (interest vectors) based on the social interaction matrix of target scenarios. By comparing synthetic data generated by IHC with three different real traces, we validate our model as a good approximation for human mobility. Exhaustive experiments are also conducted to show that IHC can predict well the performance of routing protocols.

1. Introduction

Wireless portable devices (e.g., laptop, PDA, and cell phone) are often held by humans in a delay tolerant networking (DTN) scenario. Understanding human mobility and accordingly designing better routing protocols have drawn the attention of researchers working on DTNs. Using real trace to evaluate routing protocol performance does not allow enough flexibility to change the mobility settings in order to perform the analysis for a slightly different scenario [1]. Human mobility models, however, can depict real-life human mobility characteristics and can be used to obtain meaningful routing protocol performance results in simulations [2]. So far, a number of real-life experiments have been conducted to observe and summarize human mobility characteristics, including individual (e.g., spatio/temporal preferences), encounter-based (e.g., inter-contact time and contact duration), and social (e.g., group, community) metrics [3]. Although current human mobility models are good at reproducing individual and encounter-based metrics, their strategies of generating social metrics still need further exploiting. For example, most existing models are able to reproduce the inherent social interactions but on a rather coarse granularity of community. Thus, the intercommunity nodes' social relationships and the social relationships of nodes that do not belong to any community are ignored. This ignorance leads to a considerable deviation of social relationships between the synthetic scenario and the real-life scenario. In addition, it will be unable to utilize the intercommunity nodes' social relationships to assist data dissemination when those mobility models are used to evaluate routing protocols.

This paper will focus on the granularity issue and propose a new mobility model that takes social relationships between each pair of nodes into account. We call this model interest-oriented human contacts (IHC) mobility model. The design principle is based on two intuitionistic and empirical observations. First, people always visit several spots periodically, and visit the more interesting spots at a higher probability. Second, the available contacts between a pair of people are always detected when the two people are relatively static. By implementing these two rules rationally, IHC is able to

reproduce social behavior among people, and characterize pairwise social relationships quantitatively. Our model is verified with three real traces on the metrics of inter-contact time, contact duration and meaningful pair-wise social relationships. Evaluation results show that the synthetic data IHC generates match the three real traces very well. In brief, we make the following contributions.

(i) We propose IHC, which is the first human mobility model dedicated to node interest to the best of our knowledge. We not only take node interest vectors as the exclusive social relationship input but also make a similarity analysis of node interest.

(ii) We design two methods to generate input parameters (interest vectors) based on the social interaction matrix of target scenarios. By using any of the two methods, IHC is able to reproduce social relationships on a pair-wise granularity. That is, IHC can characterize pair-wise social relationships quantitatively. As far as we know, IHC is the first such model.

(iii) We show that IHC can generate synthetic data matching very well the three real traces on the statistical properties of inter-contact time complementary cumulative distribution function (CCDF), contact duration CCDF, and meaningful pairwise social relationships, which undoubtedly leads to a well-matched social community structure.

(iv) We build cases to use the model by comparing the performance of three forwarding protocols working on real traces and synthetic traces generated by IHC. Simulation results indicate that IHC can predict well the performance of the forwarding protocols.

The rest of the paper is organized as follows. Section 2 introduces current research on the field of human mobility modeling. In Section 3, we describe IHC in detail, including the human mobility patterns which inspire our work, the model itself, a similarity analysis, and two methods to generate input parameters (interest vectors) based on the social interaction matrix of target scenarios. We validate our model by comparing synthetic data traces with three different real traces in Section 4. The comparison shows a good matching between IHC-generated and real traces. Section 5 shows that IHC can predict well the performance of routing protocols by comparing the performance of three DTN protocols with IHC-generated and real traces. Section 6 summarizes our conclusions and describes future work.

2. Related Work

Up to now a number of research articles on human mobility modeling are published. Some of them are dedicated to explore exhaustive human mobility itself such as [1, 4]; others are designed for assisting accurate performance evaluation of DTN forwarding protocols in simulations, such as [2, 5]. Our work belongs to the latter. To provide a good understanding of mobility model framework as well as the relationships among human mobility characteristics, mobility models, and routing protocols in a DTN, we draw Figure 1.

In Figure 1, each entity represents a human mobility characteristic: a rectangular entity represents an individual metric; an elliptical entity represents an encounter-based metric, and a circular entity represents a social metric. A bold line transversely divides Figure 1 into two portions. The entities above the bold line are taken as input parameters of mobility models, and the entities below the bold line are considered as output metrics. On the other hand, two erect dashed lines divided Figure 1 into three portions, namely individual metrics module, encounter-based metrics module and social metrics module, respectively (on order of left-to-right). These three modules are demarcated based on the attributes of the metrics, no matter where they are, above or below the bold line. Based the on extensive study of the existing work, we say that the most significant metrics that lead to human-like "inter-contact time" and "contact duration" distributions are the "pause time" and "location preference and periodic re-appearance" (two individual metrics). Therefore, individual metrics that do not affect the two encounter-based metrics are classified into individual metrics module and denoted by "..." in Figure 1. As a work for a model designed for assisting accurate performance evaluation of forwarding protocols in simulations, this paper is oblivious to such individual metrics.

Human mobility is driven by social relationships. Figure 1 indicates this intuition by exhibiting the social metrics module. The "inherent social interaction" in the input portion works together with several individual metrics, such as "pause time" and "location preference and periodic reappearance" to generate "social metrics" in the output portion. However, early human mobility models (such as [6, 7]) only have the individual metrics module and encounter-based metrics module. As far as we know, CMM [8] is the first model that takes social relationships into consideration. Models after CMM always include the social metrics module, with different social based representations such as interaction matrix [8, 9], communities [5, 8, 9], home-points distribution [5], overlapping communities [10], and centrality [10].

Due to the lack of end-to-end paths in DTNs, routing protocols have to utilize nodes' each chance of contact to forward packets (see Figure 1, "encounter-based protocol-dependent metrics"). The nodes' chances of contact are unstable and hard to hold. However, in DTNs, nodes' social relationships are steady and reliable. Hence, social-aware routing protocols utilize some social metrics, such as node community [11], node centrality [12], and node interest [13] to help with routing decisions. Among these metrics, the community is an important conception in social network theory. Therefore, most existing models reproduce the "inherent social interactions" on the granularity of (and are limited to) the community. Our model breaks this limitation and reproduces social relationships on a pair-wise granularity.

We argue that a good human mobility model should, first, be simple, second, have the ability of reproducing crucial human mobility metrics (both encounter-based and social), and third, predict well the performance of forwarding protocols on real DTNs.

TLW [7] generates movement traces using a model which is similar to Levy Walks, except that the flight lengths and

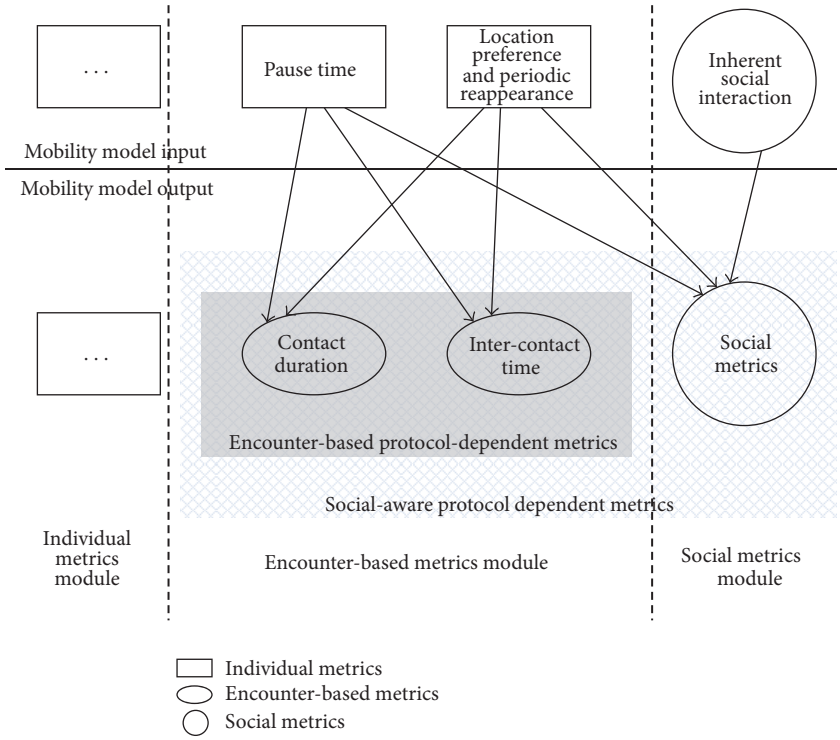


Figure 1: Mobility model framework.

the pause times have power-law distributions. TLW generates the inter-contact time similar to real traces. However, as each node moves independently, TLW does not capture the nodes' social behaviors. As a subsequent work to TLW, SLAW [14] correlates the walks of different nodes based on TLW, and takes heterogeneously bounded mobility areas and fractal waypoints into consideration. SLAW generates the inter-contact time and fight lengths distribution similar to real movements. However, there is no clear explanation about how to characterize nodes' social relationships. SMOOTH [15] is a mobility model with similar principle and performance to SLAW.

As the first mobility model that utilizes social network theory, CMM [8] is not concerned about individual human mobility characteristics. Besides, CMM has been proved defective: specifically, in the majority of configurations, all users collapse into a single location, this practically overthrows the initial setting of the system [9]. To erase this defect, Boldrini and Passarella [9] improve CMM by adding the following two individual human mobility characteristics: people tend to visit just a few locations, where they spend the majority of their time (the same meaning as the "location preference and periodic reappearance" in Figure 1), and people prefer shorter paths to longer ones. They validate their synthetic data with real traces and show a good matching of inter-contact time and flight lengths.

HHW [10] concerns heterogeneous human popularity. The model's input characteristics are "location preferential" (individual) and "overlapping communities" (social). Its output metrics "inter-contact time" (encounter based) and "centrality" (social) also have a good matching with real traces.

The intuition that inspires SWIM [5] is as follows: people go more often to places not very far from their homes and where they can meet a lot of other people. This feature is actually the same as "location preference and periodic reappearance" in Figure 1, with a clear location preference weight decided by the location's distance and popularity. SWIM uses a bounded power-law-distributed "pause time." The authors validate their synthetic data with real traces and show that SWIM has a good matching of inter-contact time, contact duration, number of contacts and community structure.

Maiti et al. [1] collect numbers of human mobility patterns and explore dependencies between them. A framework is also proposed to reproduce these human mobility patterns in the model. Thakur and Helmy [3] propose a framework for mobility model analyzing. The model generated by the framework is validated with real traces on both encounter-based metrics and social metrics. STEPS [16] is designed based on a Markov chain modeling method. The human mobility pattern it depends on is also "location preference and periodic reappearance." A good matching of both encounter-based and social metrics with real traces is shown in the literature. SAGA [17] is dedicated to the geographic diversity of the region of interest, which is different from all of the above-mentioned models. As a result, SAGA is validated with real traces on different metrics as well.

IHC also relies on "location preference and periodic reappearance" but with a different modeling method from existing models. IHC is very simple to implement and able to reproduce statistical human mobility properties, such as inter-contact time, contact duration, and accurate meaningful social relationships on the granularity of pairwise. It also can predict well the protocol performance on real DTNs. Based on the analysis of human mobility patterns and the statistical data of real traces, IHC chooses to ignore the spatial human mobility characteristics which are non-correlative to human contacts. The inherent social interaction in the IHC input parameters is node interest, such that node interest analyzing will be more facile in an IHC simulation environment.

3. Interest-Oriented Human Contacts Mobility Model

In this section, we propose interest-oriented human contacts mobility model. We introduce the human mobility characteristics that inspire our work and describe the model in detail, followed with a similarity analysis of node interest based on IHC and two methods to generate input parameters based on the social interaction matrix of target scenarios.

3.1. Design Principle. We observe that people prefer to visit a few locations and spend plenty of time staying at such locations. In other words, a few locations bring more attractions to people. In this paper, such locations are called hotspots. In real-life scenarios, hotspots are the locations related to people's interests, for example, the locations where we work, study, have meals or do sports. Usually, people spend a lot of time at some of the hotspots (such as work or study) and less time at others (such as having lunch or doing sports).

As people spend the majority of their time at hotspots, they spend little time on the journey. Besides, the vehicle people choose to accomplish the journey is strongly dependent on the distance of the next destination. For example, people prefer to walk to an adjacent destination, ride to a farther one and drive to a much farther one.

Further, we observe that an available contact (an available contact is considered as a contact when one person meets the other, their wireless devices can detect each other and accordingly forward messages) between a pair of people usually happens when the two people are relatively static. Suppose that, in a college, two students with active wireless devices are walking past each other. Due to the transmission range of wireless devices and relative speed of the two students, the actual contact duration is too short for wireless devices to detect each other. Only when the two students appear at the same spot and stay for a while, such as having dinner in the cafeteria or reading in the library, their wireless devices have enough time to detect each other and then forward messages.

Now that an available contact occurs only if the two people are relatively static, the geographical position where the contact happens is not significant anymore if we do not take the case of the synchronized motion (e.g., the two people walk along with each other) of the two people into consideration (because synchronized motion involves very specific social relationships, e.g., a colleague relationship and a particular schedule, which is hard to hold for a mobility model). Since temporary passing-bys (non-available contacts) from one spot to another spot are negligible, the specific geographical position of these two spots and distance between them make no sense. Under these circumstances, the factors that impact the contact metrics of the two people (e.g., A and B) in a period of time are as follows: when A arrives at this spot; how long A will stay for (when A will leave this spot); when B arrives at this spot, how long B will stay for. If we extend this period of time to the overall runtime of the social network that A and B belong to, then the encounter based metrics of A and B (inter-contact time and contact duration) depend on the probability that A and B visit this spot and how long will A and B stay at this spot.

Consequently, theoretically, by rationally setting numbers of hotspots, the probability of visiting each hotspot, and the pause time after each arrival, it is quite promising to generate good matching statistical characteristics and social relationships with real traces.

3.2. Detailed Model. As an interest-oriented mobility model, IHC builds an environment where node interests are manifested as hotspots. In general, one interest stands for one hotspot and vice versa.

Like the most existing mobility models, in IHC, the mobility of a specific node is composed of a set of movement epochs throughout the simulation time. At the beginning of a movement epoch, the node chooses a destination and moves towards it at a speed. After arrival, the node stays at the destination spot for a time period which is known as pause time. Till the end of the pause time, the node begins to choose a new destination and start the next movement epoch. Contacts occur when one node is within the transmission range of another node. However, in IHC, only available contacts are under consideration; that is, a contact is recorded if and only if two nodes are staying at the same hotspot simultaneously. In addition, nodes spend the majority of their time at hotspots. That is, the destination of each epoch can only be chosen from hotspots. Therefore, the contact metrics of a pair of nodes in IHC are determined by their probability to visit the same hotspot and the pause time they stay there. In such a case, the network area, the node transmission range and the positions of hotspots make little sense to contact metrics in IHC, and we choose to omit them. However, in order to ensure that the node interests are mutually independent, the distance between any two hotspots (although IHC does not care about their specific geographical positions) needs to be larger than the node transmission range so that nodes visiting different hotspots will not meet each other.

Since a node has different preferences to different interests, it visits the corresponding hotspots at different probabilities. Suppose that there are n hotspots corresponding to n interests in the network area. These n node interests compose an n-dimensional interest space. Each node has an

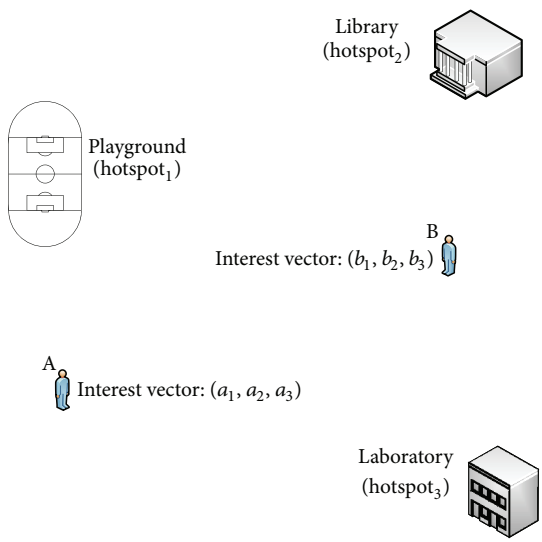


Figure 2: An IHC scenario containing three hotspots and two nodes.

n-dimensional vector, corresponding to a point in the interest space. The interest vector of *nodex* $(x_1, x_2, \ldots, x_n)$ means *nodex* visits hotspot i at probability x_i $(x_i \geq 0)$, $i \in [1, n]$; then $\sum_{i=1}^{n} x_i = 1$.

Figure 2 maps a network area containing three hotspots (i.e., playground, library, and laboratory) and two nodes (A and B) who determine each movement epoch following IHC. Initially, each node, for example, node A, can be anywhere in the network area. Node A's first epoch soon begins: it chooses its destination from all hotspots according to its interest vector. In Figure 2, node A chooses to visit playground (hotspot$_1$), library (hotspot$_2$), and laboratory (hotspot$_3$) at probabilities a_1, a_2, and a_3, respectively, such that $a_1 + a_2 + a_3 = 1$. Once the destination has been chosen, for example, the library (hotspot$_2$), node A starts moving straightly towards it with a constant time ft. A constant flight time actually indicates that the speed is proportional to the flight length (distance between the starting point and the destination) in IHC. This proportional relation is based on the observation that in real-life scenarios people spend little time on the journey by choosing different vehicles for different distances of destinations. After reaching the destination, the pause time will be determined by a variable whose probability density function (PDF) $pt()$ obeys a bounded power-law distribution as in [5]. Now node A is reading in the library and will stay for a time duration (pause time). Note that node B has also been accomplishing its movement epochs. If it is staying at the library coincidently, both node A and node B will be able to detect this contact until one of them runs out of the pause time and begins the next epoch. Later, the other node will leave the library as well for the next epoch. Both A and B keep this kind of movements till the end of the simulation time.

It can be seen that IHC gives a clear expression on node interest and ignores specific geographical information. Such treatments lead to a much more convenient tuning up of node interest parameters. Other models do not provide a direct node interest tuning. For example, in SWIM, the probability that a node visits a spot depends on not only the spot's popularity but also the distance between the spot and the node's home. Although such settings make the preference of each node to each interest (corresponding to a spot) self-controlled, it is not easy to change the probability that a node visits a spot at will. Therefore, by ignoring specific geographical information, IHC replaces the probability of visiting a spot influenced by popularity and distance in SWIM with a single interest value. In this way, IHC gets rid of the inconvenience of altering a spot's visiting probability influenced by a home's position. In addition, IHC keeps all temporal metrics on contact and ignores the information of "the geographical position of contact," which not only needs complicated settings but also lacks corresponding information in real traces and does not affect the performance of forwarding protocols as well. In IHC, the specific geographical position of each hotspot has no influence on either contact metrics or forwarding protocols' performance as long as the distance between any two hotspots exceeds the node transmission range. As a conclusion, Table 1 summarizes all parameters and their meanings in IHC.

3.3. Similarity Analysis. Thakur et al. [18] demonstrate that people with similar behavioral principle tie together, which means that user-location coupling can be used to identify similarity patterns in mobile users. They make similarity analysis for several mobility models and show that many mobility models do not explicitly capture similarity and result in homogeneous users that are all similar to each other. Their similarity analysis is based on spatiotemporal preferences, preferential attachment to locations, and the frequency and duration of visiting these locations, which are actually the first-hand design principle of IHC. Therefore, IHC is suitable for similarity analysis inherently.

Mei et al. [13] try to utilize the cosine similarity of node interest profile to assist data forwarding in social-aware routing protocols because they believe that similar node interest profiles lead to close social interactions. However, as we have mentioned above, the mobility model they use, SWIM, cannot be used to measure node interest either accurately or conveniently. Additionally, whether cosine similarity of node interest profile can represent people's social interactions accurately is still unclear, while for IHC, node interest is taken as input parameters, thus making similarity analysis so natural that we may hopefully get meaningful conclusions.

It is generally believed that a large contact duration represents a close relationship between nodes, so social relationships are always denoted simply by contact durations [8, 12]. We also use this denotation in this paper.

Intuitively, in IHC, the social relationship between two nodes ought to be related to the interests shared by the two nodes, as only their common interests result in the two nodes' meeting at the corresponding hotspots. Based on this intuition, we conduct extensive simulations to observe what the relation between the common interests and contact durations of the two nodes is.

Table 1: IHC parameters.

Parameter	Meaning
num	The amount of nodes
st	Simulation time, measured in seconds
ft	Flight time, measured in seconds
pt()	The PDF of pause time which is measured in seconds
n	The amount of node interests
$H_i(x, y)$	The coordinate of hotspot H_i corresponding to interest $i, i \in [1, n]$
$X_j = (x_1, x_2, \ldots, x_n)$	The interest vector of node j, $j \in [1, num]$, $\sum_{i=1}^{n} x_i = 1$

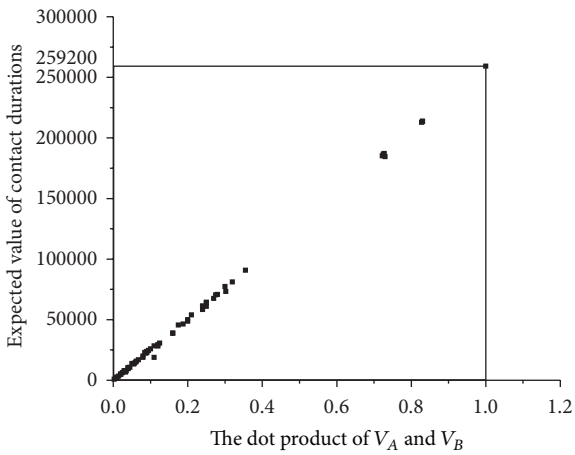


Figure 3: Proportional relation between a node interest metric and a social relationship metric.

In our simulations, there are only two nodes, namely, A and B, in the network area. Without specific input social interaction, pair-wise contacts are mutually independent. Thus, multiple nodes do not bring new insights. Each node has and only has 4 interests, such that node A's interest vector $V_A = (A_1, A_2, A_3, A_4)$, $A_1 + A_2 + A_3 + A_4 = 1$ and node B's interest vector $V_B = (B_1, B_2, B_3, B_4)$, $B_1 + B_2 + B_3 + B_4 = 1$. Let c_i be the amount of A and B's common interests, with value of 1, 2, 3,4, respectively, and set different values of V_A and V_B, such as (0.25, 0.25, 0.25, 0.25), (0.3, 0.3, 0.3, 0.1), (0.1, 0.2, 0.3, 0.4), or (0.05, 0.05, 0.05, 0.85). Note that even for two specific vectors, different common interests should be assigned in each simulation. The simulation time is set as three days, that is, 259200 seconds. To get the expected value, for each scenario, we average the results of contact duration over 10000 runs using different seeds.

Excitedly, we find that there is specific relation between "the dot product of V_A and V_B" and "the expected value of A and B's contact durations," and we show the results in Figure 3. Each black point in Figure 3 represents a simulation scenario. The x-axis shows the dot product of V_A and V_B ($V_A \cdot V_B$) and the y-axis indicates the expected value of contact duration averaged over 10000 runs. The maximum of $V_A \cdot V_B$ is 1 when both nodes have only one, and the same interest. Corresponding to this specific scenario, the two nodes stay forever at the same hotspot. Thus, their contact duration is the simulation time, 259200 s.

Figure 3 shows a proportional relation between "the dot product of V_A and V_B" (x-axis) and "the expected value of A and B's contact durations" (y-axis). The dot product of V_A and V_B is a metric derivative from A and B's interests and the expected value of A and B's contact durations is a metric, which can represent the social relationship between A and B. Figure 3 reveals a promising feature of IHC. That is, IHC may have the ability of accurately reproducing a specific contact duration matrix, which is always regarded as a social interaction map. IHC can generate a specific expected value of contact duration accurately by setting appropriate values to node interest vectors. Note that the y-axis in Figure 3 only shows the expected value of contact durations. As a complement, the distributions of the contact durations for different expected values are shown in Figure 4.

Figure 4 is graphed to assist understanding what distribution the contact durations obey for one expected value in Figure 3. Figure 4 is composed of eight subfigures. The expected value of contact duration in each subfigure is denoted by Exp. We choose Exp for eight scales to show in Figure 4, namely, 2500, 5000, 7500, 10000, 15000, 20000, 50000, and finally a very large one, 180000. As we can see in Figures 4(a)–4(d), for a small Exp (no larger than 10000), the distributions are far from the Gaussian Distribution, such that the expected value shows a considerable deviation from a randomly chosen value. In such a case, maybe IHC cannot reproduce a small contact duration accurately by tuning node interest vectors. Fortunately, a small contact duration makes nearly no sense in social network analysis. When Exp becomes larger, for example, in Figures 4(e)–4(h), the distributions look like a Gaussian Distribution, such that it will be more accurate by representing contact durations with the expected value.

3.4. Interest Vectors Generator. As shown in Table 1, we take node interest vectors as the exclusive social relationship input of IHC. The corresponding parameters, namely, n, $H_i(x, y)$, and X_j, can be derived based on the conclusion in Section 3.3. That is, the expected value of A and B's contact durations is proportional to $V_A \cdot V_B$. Consider an extreme case: when two nodes both have only one interest and their interest is the same as follows their dot production of interest vector is 1, and their contact duration is exactly the simulation time *st*

(see Figure 2) such that the coefficient of proportionality is $1/st$.

Suppose the *num* nodes are *node1*, *node2*, ..., *nodenum*, and their interest vectors are $X_1, X_2, \ldots, X_{num}$, respectively. Denote the sum of elements of vector X by $sum(X)$. Denote the contact duration matrix of the scenario which we want to reproduce by D, such that D is a $num \times num$ matrix. Then the contact duration between *nodei* and *nodej* is D_{ij}. The following equation set holds:

$$\begin{aligned} &sum\left(X_i\right) = 1, \quad i \in [1, num]; \\ &X_i \cdot X_j = \frac{D_{ij}}{st}, \quad i, j \in [1, num], \quad i \neq j. \end{aligned} \tag{1}$$

The equation set has $n \times num$ variables, ($num \times (num - 1)/2 + num$) equations. Obviously a properly selected n can make this equation set have solutions. Approximate solutions can be derived with the Levenberg Marquardt algorithm. The solutions include the parameter settings of n and X_j. $H_i(x, y)$ can be anywhere in the network area as long as the distance between any two hotspots is larger than r.

However, the above method of choosing amount of interests and nodes' interest vectors, named as method-1, may not generate accurately small contact durations when the simulation time is not long enough. The reason can be deduced in Figure 3. For a small contact duration expected value, the smaller the sample size (amount of contacts between a pair of nodes) is, the harder the control of contact duration value (because the distribution is far from the Gaussian Distribution) is. To tackle this problem, we provide another method of choosing amount of interests and nodes' interest vectors, namely, method-2.

The detailed method-2 is as follows.

(1) Let $n = num$. To simulate the small contact durations in scenarios with a short simulation time, we assume that the amount of node interests equals the number of nodes. In other words, it can be regarded that each node has a home spot which the node visits at a high probability. Under this circumstance, a node's interest vector (x_1,x_2,...,x_n) means that this node visit *node1*'s home at probability x_1, *node2*'s home at probability x_2, and *noden*'s home at probability x_n. Therefore, if two nodes have a large contact duration value, one node will certainly visit the other node's home at high probability, and the value of the probability is determined by the contact duration of these two nodes in the corresponding scenario.

(2) Designate a public spot. We use a spot that is visited by all nodes at a specific probability to generate all small contact durations in the contact duration matrix of the scenario. Now, the value of n is actually $num + 1$. A node's interest vector becomes $(p, x_1, x_2, \ldots, x_n)$, where p presents the probability at which the node visits the public spot. Generally, p holds the same value for all nodes.

(3) Assign an initial interest vector value to each node. Initially, we assign *nodei*'s interest value as (p, 0,..., x_i=1 − p,..., 0), $1 \leq i \leq n$. That is, initially, each node only visits two hotspots: the public spot and its own home.

(4) Set a threshold to the target scenario's contact duration matrix. For the contact duration matrix, we set a threshold Tth and select all the values that are no less than Tth to reproduce in IHC. The value of Tth is chosen intuitively and empirically, assuring that contact duration larger than Tth is considerable and meaningful to represent a close social relationship.

(5) Tune up all nodes' interest vector value based on the values exceeding Tth in the target scenario's contact duration matrix. Since all contact duration values no larger than Tth are generated by the visiting of the public spot, the remaining contact duration values (exceeding Tth) can be generated by tuning up the probability of home spots in nodes' interest vector values. Generally, we deal with the nodes one by one on the order of node ID from 1 to *num*. That is, for *node1*, the interest value is (p, $x_1 = 1 - p$, 0, ..., 0). Then, we search the *node1*'s list in contact duration matrix; if *nodea* and *node1*'s contact duration CD_{1a} exceeds Tth, *nodea*'s interest vector will be updated as ($p, x_1 = CD_{1a}/(st(1-p))$, 0, ..., x_a=1 − p − x_1, ..., 0) and so on. Different treatments can also be conducted as long as the contact duration values exceeding Tth are all held and for each node's interest vector (p, x_1, x_2, ..., x_n), $p + \sum_{i=1}^{n} x_i = 1$ is assured.

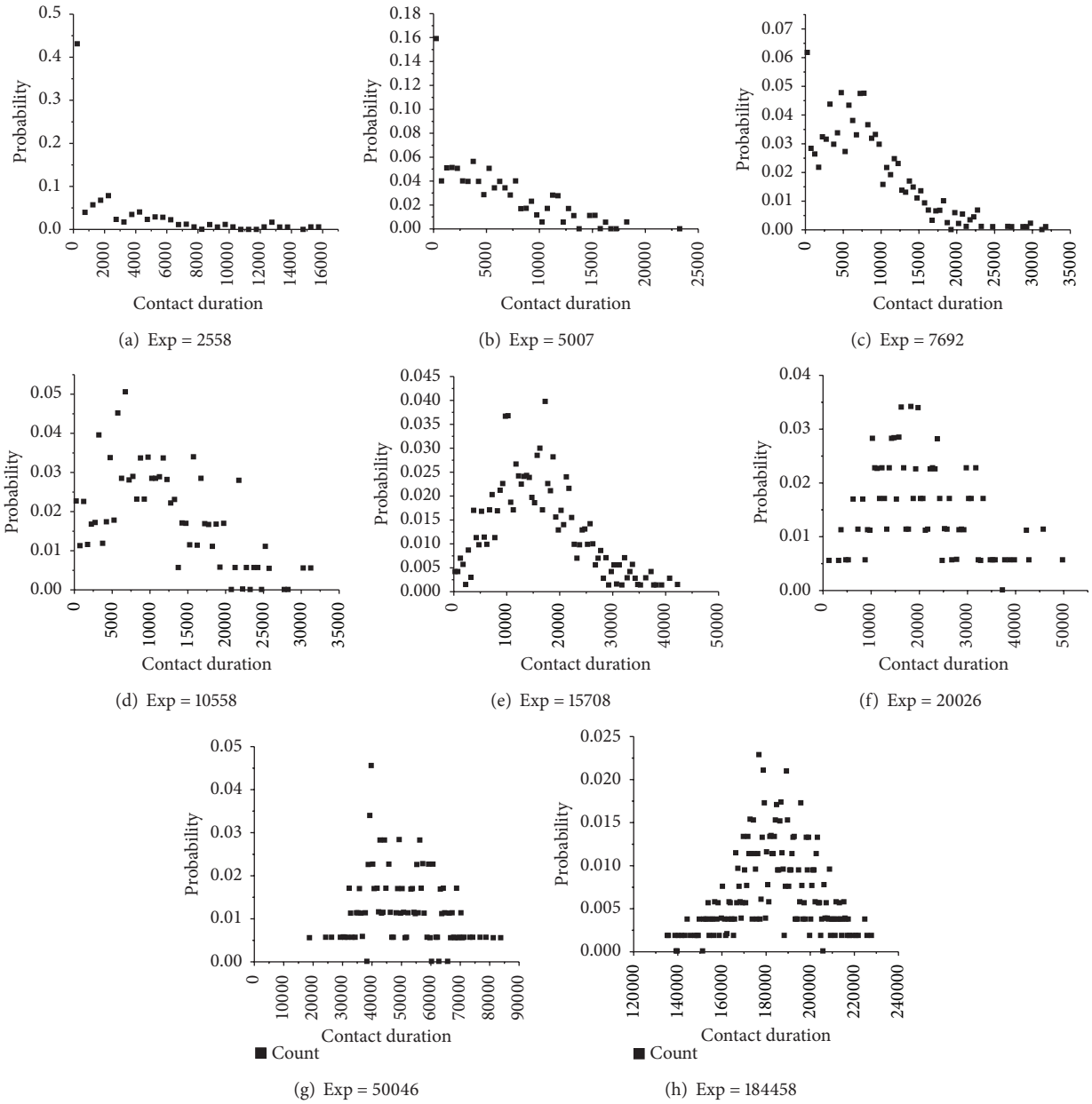
We can derive the interest vectors by using method-1 or method-2 if we want to reproduce a real-life scenario using IHC. However, in general, there are totally two cases when we need to determine the interest vectors of IHC. In the other case, if we want to set up just a simulation scenario, the interest vectors can be set as we need, for example, random values.

Table 2: The three experimental data sets.

Dataset name	Infocom 06 Trace-1	Infocom 05 Trace-2	Cambridge Trace-3
Device	iMote	iMote	iMote
Network type	Bluetooth	Bluetooth	Bluetooth
Duration (days)	3	3	11
Granularity (sec)	120	120	600
Devices number	98 (78 mobile)	41	54 (36 mobile)

4. Model Verification

In order to show the accuracy of IHC in simulating real-life scenarios, we compare IHC with three real traces whose data is gathered from experiments done with wireless devices carried by people. These three traces are known as Infocom 06 trace (trace-1) [19], Infocom 05 trace (trace-2) [20], and Cambridge trace (trace-3) [21]. More details of the three real traces are shown in Table 2.



(a) Exp = 2558 (b) Exp = 5007 (c) Exp = 7692

(d) Exp = 10558 (e) Exp = 15708 (f) Exp = 20026

(g) Exp = 50046 (h) Exp = 184458

Figure 4: Distributions of contact durations under different expected values.

4.1. Trace Data. We illuminate how we use the trace data as follows.

First, we only care for the data generated by mobile and homogeneous nodes. For example, in trace-1, there are totally 98 iMotes in the experiment, but 20 of them are long range (around 100 meters) and static (deployed throughout the area or placed in lift of the hotel). The remaining 78 iMotes are carried by participants of the Infocom student workshop, with transmission range around 30 meters. Hence, these 78 iMotes are our research objects in this experiment. In order to find neighbor iMotes, each iMote performs periodic desynchronized scanning. The scanning takes approximately 5 to 10 seconds with time granularity between two consecutive scanning 120 seconds. An iMote cannot respond to any request when it is active such that the synchronization needs to be avoided. In this experiment, a contact is defined as a period of time where all successive scanning by one iMote receive a positive answer by another. That is, a contact can only be confirmed after at least two scanning. Given that the scanning granularity is 120 s, the speed of a pedestrian is around 1 m/s, and the transmission range is around 30 m, it indicates that a moving iMote can hardly detect a contact such that the rationality of the feature "ignoring the contacts of moving nodes" in IHC is supported.

Second, symmetrize the contact duration matrix. In the three experiments, due to the interference and other limitations, non-mutual sightings are always created. As a result, the inter-contact time and contact duration are not

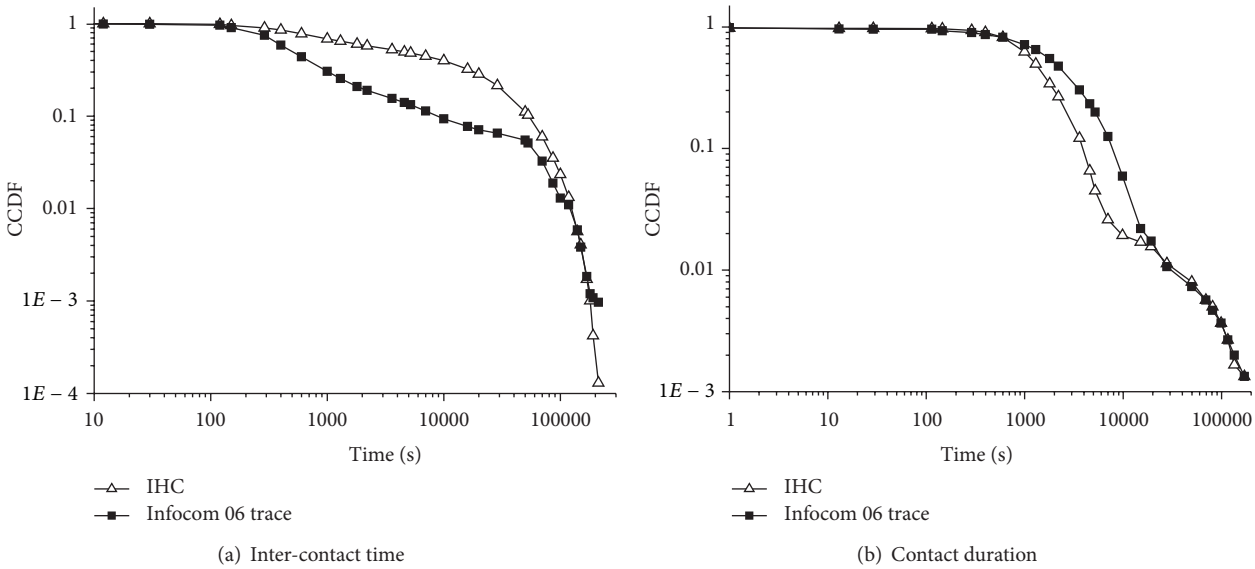


Figure 5: Reproduce Infocom 06 trace using IHC.

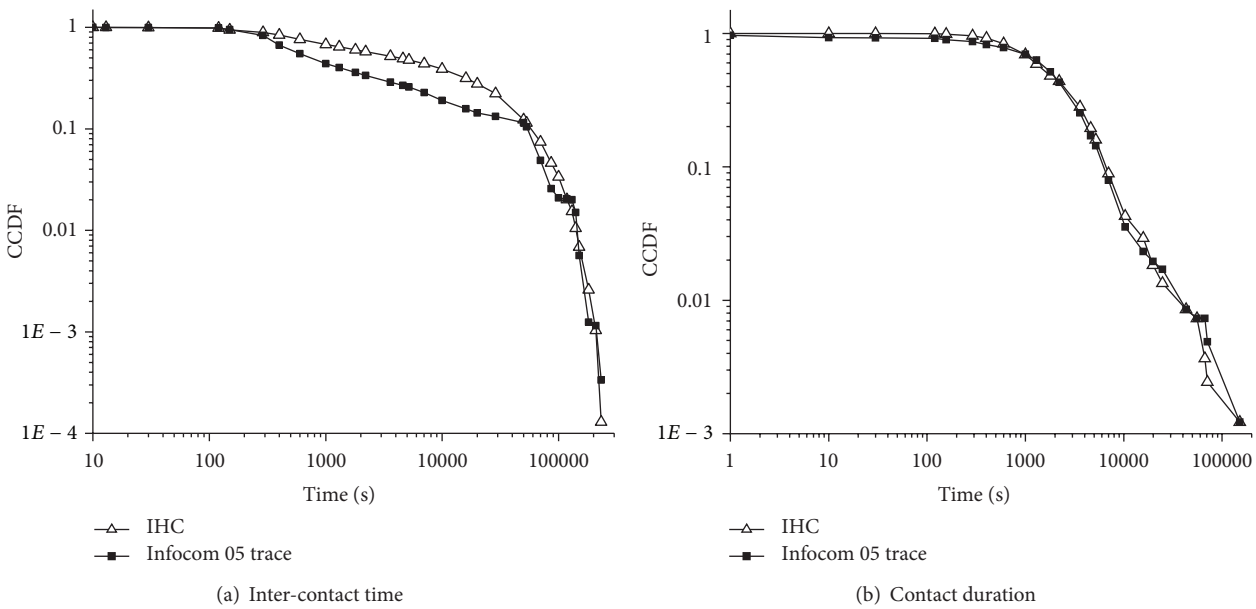


Figure 6: Reproduce Infocom 05 trace using IHC.

symmetrical. We keep all inter-contact times detected by the mobile iMotes. However, for the contact duration between a pair of iMotes, we take the maximum of their detected results as the value. That is, if iMote A has detected that its contact duration with B is CD_{A-B}, while iMote B detected the value as CD_{B-A}, we will take $\max(CD_{A-B}, CD_{B-A})$ as the contact duration value between iMotes A and B.

Note that these two treatments are conducted for all the three real traces.

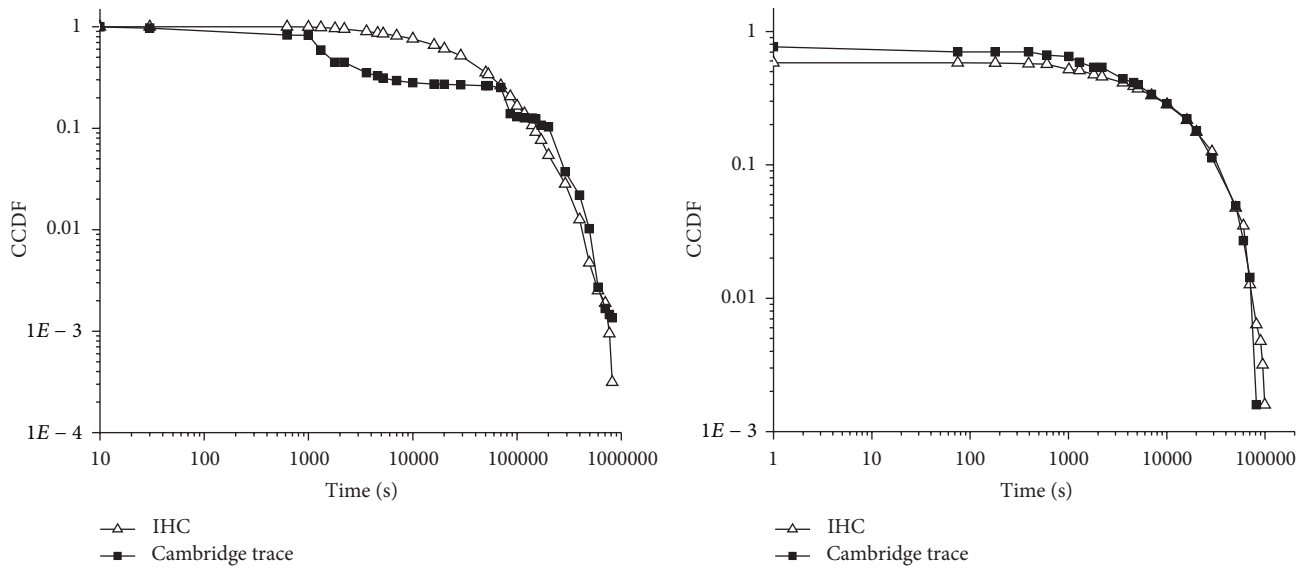
4.2. Simulation Environment and Parameter Settings. IHC takes parameters listed in Table 1 as input. To compare IHC with real traces, we make the output text files containing records on contact metrics and social relationships, including:

(i) inter-contact time.txt: recording all inter-contact times between any two nodes;

(ii) contact duration.txt: recording all contact durations between any two nodes;

(iii) interaction matrix.txt: recording all contact durations between any two nodes in a matrix.

As we build a discrete even simulator of IHC with VC++6.0, we are able to change the output of the simulator to observe each event, such as a node starting moving or finish



(a) Inter-contact time

(b) Contact duration

Figure 7: Reproduce Cambridge trace using IHC.

moving and two nodes meeting each other or departing from each other.

Parameters of IHC are chosen and tuned up based on the scenarios which we want to simulate, for example, Infocom 06 trace (trace-1). For the parameters of *num* and *st*, we assign them exactly the same values as trace-1, that is, 78 and 3 days (259200 seconds). ft is set to be 10 seconds based on the intuition that people spend few time on the journey. The pause time, which makes the best output (e.g., inter-contact time) matching with the real traces, is a bounded power law over the range of [120s, 4800s] with *slope* 6, denoted by (*slope, lower_bound, upper_bound*) in Table 3. Among them, the *lower_bound* affects the head of the inter-contact time CCDF, the *upper_bound* affects the tail of the inter-contact time CCDF, and the *slope* weakly affects the *slope* of inter-contact time CCDF in a very small range. The *lower_bound* is determined by scanning granularity because it is the scanning granularity that strongly affects the head of inter-contact time CCDF of the real traces. The values of *slope* and the *upper_bound* are determined by matching between real trace inter-contact time CCDF and simulation results.

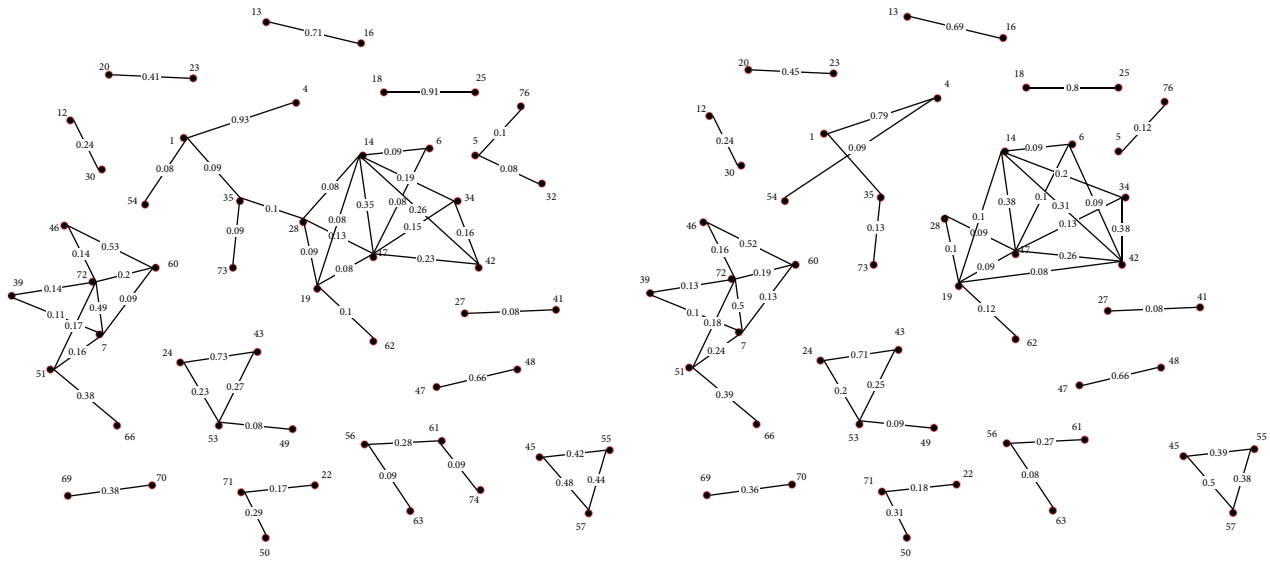
Table 3 summarizes all parameter settings of the three scenarios. $H_i(x, y)$ and X_j are too expatiatory to show in Table 3 and thus omitted, since we have indicated the specific method to get them in detail in Section 3.4.

4.3. Simulation Results. We show the simulation results of inter-contact time and contact duration of Infocom 06 trace, Infocom 05 trace, and Cambridge trace in Figure 5, Figure 6, and Figure 7, respectively. Figures 5–7 validate that IHC can generate statistical metrics that approximate real traces. For a quantitative comparison, we calculate the Jensen-Shannon divergence between the distributions of the real traces and IHC traces in Table 4, as well as the corresponding results of SWIM traces whose data can be found in [5], since SWIM is a very outstanding work on human mobility modeling. The results shown in Table 4 indicate that our model outperforms SWIM in the accuracy of reproducing inter-contact time and contact duration.

We draw meaningful social relationships in real and IHC traces in Figures 8, 9, and 10, which are weighted undirected graphs. A vertex in the graph (Figures 8, 9, and 10) represents the node with the same ID in the networks. The edge between two vertices indicates that the social relationship (contact duration) between these two nodes exceeds a certain threshold. For Infocom 06 trace, Tth is assigned as 20000 seconds because only the top 1.665% (50 out of 3003) largest contact durations are larger than 20000. For Infocom 05 trace, Tth is set to be 10000 seconds because only the top 3.9% (32 out of 820) largest contact durations are larger than 10000. The threshold for Cambridge trace is chosen with similar principle with the value of 50000. The weights of the edges are calculated as the ratio of contact duration between the two nodes to the network simulation time, retaining two decimal places.

Figures 8, 9, and 10 visually show the social relationship similarity between real traces and corresponding IHC-generated ones. For a quantitative view, we conduct Mantel Test on the real and IHC-generated social interaction matrices where the raw data Figures 8, 9, and 10 comes from. Mantel Test measures the correlativity between two matrices. Since Figures 5(b)–7(b) have shown a very similar scale of social interaction matrices between real traces and corresponding IHC-generated ones, a high correlativity can complementarily prove that the IHC-generated social interaction matrices are very similar to the real ones. The Mantel Test results are shown in Table 5.

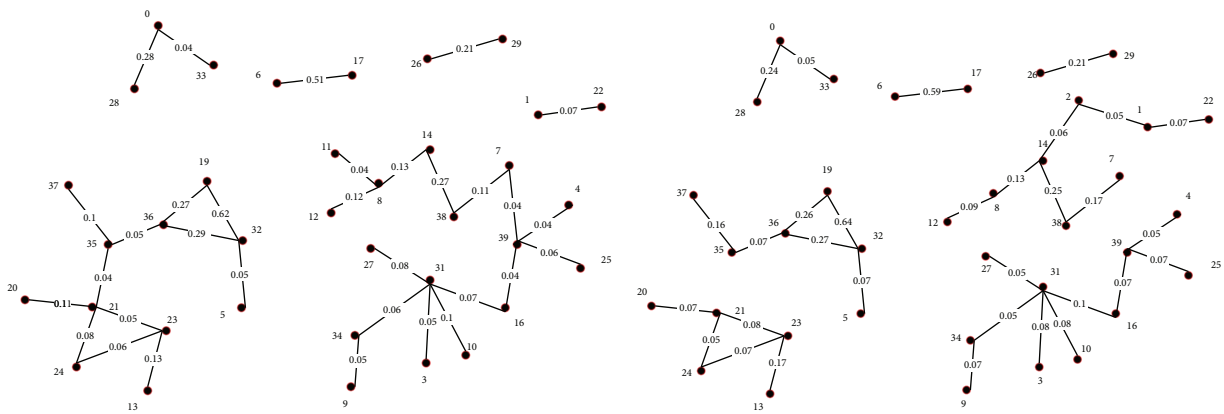
Figures 5(b)–7(b) and Table 5 prove that IHC can accurately reproduce the overall social relationships in real-life scenarios. Further, Figures 8, 9, and 10 indicate that



(a) Meaningful social relationships in Infocom 06 trace (Tth > 20000)

(b) Meaningful social relationships in IHC's synthetic trace (Tth > 20000)

Figure 8: Comparisons between Infocom 06 trace and IHC: meaningful social relationships.



(a) Meaningful social relationships in Infocom 05 trace (Tth > 10000)

(b) Meaningful social relationships in IHC's synthetic trace (Tth > 10000)

Figure 9: Comparisons between Infocom 05 trace and IHC: meaningful social relationships.

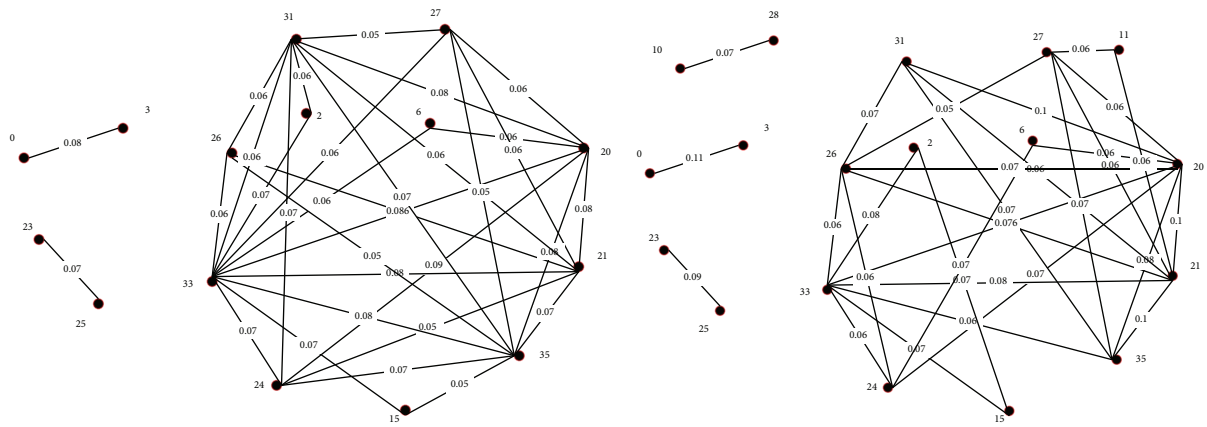
IHC is also able to characterize pairwise social relationships quantitatively. For example, there exist some close social relationships that cannot be detected by a community detection algorithm (e.g., k-clique [22], $k > 2$), such as the relationship between 13 and 16 and 18 and 25 in Figure 8(a). This kind of relationship is defined as "friendship" in [12]. IHC has the ability of reproducing the "friendship" in the target scenario. Further, the inter/intra-community social relationships and pairwise social relationships belonging to no communities that IHC generates all match real traces very well. As far as we know, no model has such a feature. Note here that we only compare and show pair-wise social relationships of real traces and IHC traces, since other mobility models cannot reproduce the social relationships on the basis of pair-wise.

5. Building Cases to Use IHC

In this section, we build cases to use our model. We compare the performances of forwarding protocols running with real traces and our simulated scenarios to validate that IHC can be used to predict protocols' performance. We use the three real traces (Infocom 06 trace, Infocom 05 traces and Cambridge trace) and the three corresponding synthetic traces generated by IHC as the network environments. Our goal is to validate that IHC is able to predict the performance of forwarding protocols rather than evaluating which forwarding protocol performs better. Therefore, the protocols we choose, that is, Epidemic Forwarding [23] and Spray and Wait [24], which are very mature and get extensively utilized in DTNs and

Table 3: Parameter settings.

Scenario	Infocom 06	Infocom 05	Cambridge
num	78	41	36
st	259200	259200	950400
ft	10	10	10
pt: (slope, lower_bound, upper_bound)	(6, 120, 4800)	(6, 120, 7200)	(2.45, 600, 14400)
Interest vectors generator	Method-2	Method-2	Method-1
n	79	42	36



(a) Meaningful social relationships in Cambridge trace (contact duration >50000)

(b) Meaningful social relationships in IHC's synthetic trace (contact duration >50000)

Figure 10: Comparisons between Cambridge trace and IHC: meaningful social relationships.

Table 4: Jensen-Shannon divergence between distributions of the real and IHC traces, comparing with the corresponding results of SWIM traces.

Trace	Infocom 06	Infocom 05	Cambridge
Intercontact time (IHC)	0.048	0.037	0.043
Intercontact time (SWIM)	0.049	0.062	0.058
Contact duration (IHC)	0.022	0.004	0.014
Contact duration (SWIM)	0.18	0.21	0.15

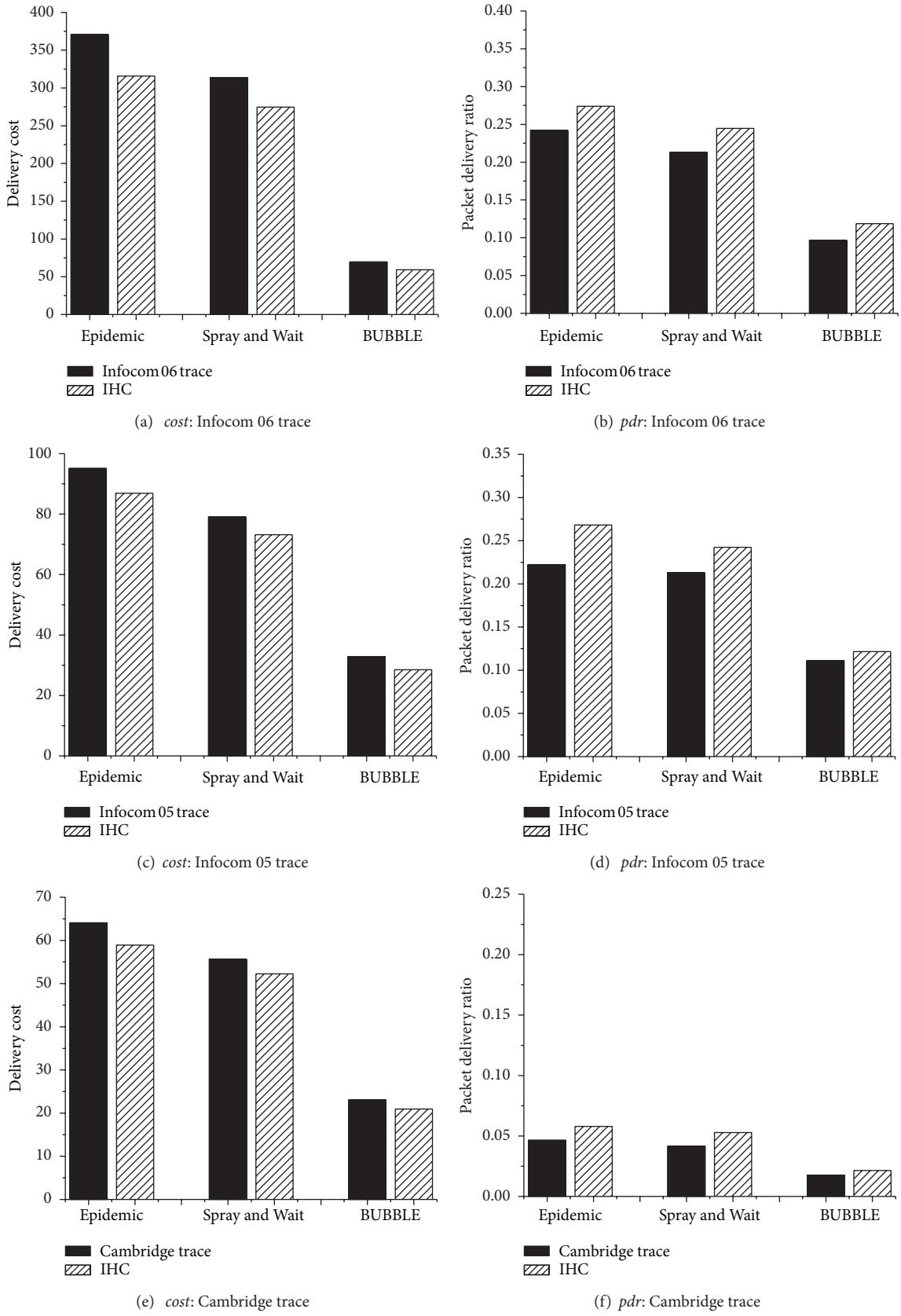
Table 5: Mantel Test results on the real and IHC-generated social interaction matrices.

Scenario	Infocom 06	Infocom 05	Cambridge
Correlativity	0.9496	0.9407	0.9256

BUBBLE [12], which is a sophisticated social-aware protocol are appropriate for our goal.

As in [5, 25], we choose two metrics to evaluate the performance of forwarding protocols. They are delivery cost (*cost*) and packet delivery ratio (*pdr*). The former indicates the price of forwarding a data packet successfully and accounts for the efficiency of the protocol. The delivery cost is calculated by the ratio of "the amount of received control packets plus the amount of data packets' replicas" to "the amount of received data packets". The packet delivery ratio, instead, is actually the successful rate of forwarding data packets and accounts for the effectiveness of the protocol. Packet delivery ratio is calculated as the ratio of "amount of received data packets" to "amount of generated data packets".

The following settings are validated for each scenario: a set of messages is generated with sources and destinations chosen uniformly at random with interval of 20 minutes, as we simulated the overall periods, that is, 3 days or 11 days, which is significantly different from that of [5] where each simulation runs only for 3 hours (choosing 3 hours out of 3 days or 11 days incurs too many uncertainties). However, the interest vector setting in Section 4.2 is based on the overall experiments duration (i.e., 3 days and 11 days). In IHC, all movement epochs are consecutive, but the actual movement epochs differ greatly in daytime and nighttime. Therefore, for a more meaningful simulation and for making statistical analysis in separate scenario, we divide the experiment duration into fragments equally. Concretely, the scale of the fragment is chosen as 12 hours to capture daytime and nighttime motions, respectively. That is, Infocom 06 and Infocom 05 scenarios are divided into 6 fragments and the Cambridge scenario is divided into 22 fragments. More importantly, interest vectors are dynamic and the values are derived using "Interest Vectors Generator" based on the current experiment fragment. To avoid end-effects, no messages are generated in the last hour; the time-to-live of messages is set as 1 hour. The accumulated forwarding protocols' results are shown in Figure 11. That is, in each simulation in Figure 11, the statistical metrics (*cost* and *pdr*) are continuously calculated except that interest vectors change with the alternate fragments. Table 6 shows



(a) *cost*: Infocom 06 trace

(b) *pdr*: Infocom 06 trace

(c) *cost*: Infocom 05 trace

(d) *pdr*: Infocom 05 trace

(e) *cost*: Cambridge trace

(f) *pdr*: Cambridge trace

Figure 11: Performance of forwarding protocols (interest vectors change with the alternate fragments).

Table 6: The average error percentage of all fragments in each separate scenario.

Scenario	Infocom 06	Infocom 05	Cambridge
cost (Epidemic)	−0.14	−0.08	−0.08
cost (Spray & Wait)	−0.11	−0.07	−0.06
cost (BUBBLE)	−0.14	−0.12	−0.13
pdr (Epidemic)	0.10	0.21	0.24
pdr (Spray & Wait)	0.15	0.17	0.25
pdr (BUBBLE)	0.23	0.09	0.22

the average error percentage of all fragments in each separate scenario. In our simulations, IHC-generated traces always lead to lower *cost* and higher *pdr* than the corresponding real ones. Thus in Table 6 the average error percentage of *cost* are all negative and those of *pdr* are all positive. However, the error percentage results in Table 6 have small absolute value, showing that each of the three forwarding protocols has similar performance in both real and synthetic traces generated by IHC.

Figure 11 is composed of six subfigures, namely Figures 11(a)–11(f). Figures 11(a)-11(b), 11(c)-11(d), and 11(e)-11(f) depict the performance of forwarding protocols in Infocom 06 trace, Infocom 05 trace, and Cambridge trace, respectively. In each subfigure, we draw six pillars representing consecutively the performance (corresponding to the subfigure, such as *cost* or *pdr*) of Epidemic Routing in the real trace and synthetic trace, Spray and Wait in the real trace and synthetic trace and BUBBLE. Figure 11 shows that the trend of the protocols in the real traces is the same as that of the corresponding synthetic ones. That is, the ones that perform better in the real world do the same things in the IHC-generated one. Figure 11 and Table 6 both indicate that IHC can predict well the performance of all the three protocols. As a result, IHC is a good model for protocol validation; the performance of protocols in the real life scenarios can be accurately predicted by running the protocols on the synthetic traces generated by IHC.

6. Conclusions

In this paper, we propose a mobility model, IHC. IHC merges a few human mobility characteristics and is very simple to implement. IHC takes node interest as input to reproduce nodes' social relationships. Correspondingly, we explore 2 methods to generate node interest vectors based on a contact duration matrix. Through the comparisons with real-life human mobility metrics of inter-contact time and contact duration, we validate that IHC can generate synthetic traces that approximate real traces. Being different from any existing mobility models, IHC has the ability of characterizing pairwise social relationships quantitatively. Further simulations have been conducted to show that IHC can predict the performance of forwarding protocols well.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (NNSFC) under Grant no. 61170245.

References

[1] R. R. Maiti, A. Mallya, and N. Ganguly, "Characterizing Mobility Models for Human Movement".

[2] C. Zhao and M. L. Sichitiu, "N-body: social based mobility model for wireless ad hoc network research," in *Proceedings of the 7th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON '10)*, June 2010.

[3] G. S. Thakur and A. Helmy, "COBRA: A Framework for the Analysis of Realistic Mobility Models," 2012.

[4] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11)*, pp. 1082–1090, August 2011.

[5] S. Kosta, A. Mei, and J. Stefa, "Large-scale synthetic social mobile networks with SWIM," *IEEE Transactions on Mobile Computing*, 2012.

[6] W.-J. Hsu, T. Spyropoulost, K. Psounis, and A. Helmy, "Modeling time-variant user mobility in wireless mobile networks," in *Proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM '07)*, pp. 758–766, May 2007.

[7] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM Transactions on Networking*, vol. 19, no. 3, pp. 630–643, 2011.

[8] M. Musolesi and C. Mascolo, "Designing mobility models based on social network theory," *ACM SIGMOBILE Mobile Computing and Communication Review*, vol. 11, no. 3, pp. 59–70, 2007.

[9] C. Boldrini and A. Passarella, "HCMM: modelling spatial and temporal properties of human mobility driven by users' social relationships," *Computer Communications*, vol. 33, no. 9, pp. 1056–1074, 2010.

[10] S. Yang, X. Yang, C. Zhang, and E. Spyrou, "Using social network theory for modeling human mobility," *IEEE Network*, vol. 24, no. 5, pp. 6–13, 2010.

[11] T. Abdelkader, K. Naik, A. Nayak, N. Goel, and V. Srivastava, "SGBR: a routing protocol for delay tolerant networks using social grouping," *IEEE Transactions on Parallel and Distributed Systems*, 2012.

[12] P. Hui, J. Crowcroft, and E. Yoneki, "BUBBLE Rap: social-based forwarding in delay-tolerant networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 11, pp. 1576–1589, 2011.

[13] A. Mei, G. Morabito, P. Santi, and J. Stefa, "Social-aware stateless forwarding in pocket switched networks," in *Proceedings of the 30th IEEE International Conference on Computer Communications (INFOCOM '11)*, pp. 251–255, April 2011.

[14] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "SLAW: a mobility model for human walks," in *Proceedings of the 28th Conference on Computer Communications (INFOCOM '09)*, pp. 855–863, April 2009.

[15] A. Munjal, T. Camp, and W. C. . Navidi, "SMOOTH: a simple way to model human walks," in *ACM SIGMOBILE Mobile Computing and Communications Review*, pp. 34–36, 2010.

[16] A. D. Nguyen, P. Sénac, V. Ramiro, and M. Diaz, "STEPS-an approach for human mobility modeling," in *Proceedings of the 10th International IFIP TC 6 Networking Conference*, vol. 6640, pp. 254–265, Springer, Valencia, Spain, 2011.

[17] B. Astuto, A. Nunes, K. Obraczka, and A. Rodrigues, "SAGA: socially-and geography-aware mobility modeling framework," in *Proceedings of the 15th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pp. 367–376, ACM, 2012.

[18] G. S. Thakur, A. Helmy, and W.-J. Hsu, "Similarity analysis and modeling in mobile societies: The missing link," in *Proceedings of the 5th ACM Workshop on Challenged Networks (CHANTS '10)*, pp. 13–20, September 2010.

[19] http://crawdad.cs.dartmouth.edu/cambridge/haggle/imote/infocom2006.

[20] http://crawdad.cs.dartmouth.edu/cambridge/haggle/imote/infocom.

[21] http://crawdad.cs.dartmouth.edu/upmc/content/imote/cambridge.

[22] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[23] A. Vahdat and D. Becker, "Epidemic routing for partially connected ad hoc networks," Tech. Rep. CS-200006, Duke University, 2000.

[24] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and wait: An efficient routing scheme for intermittently connected mobile networks," in *Proceedings of the ACM SIGCOMM Conference on Computer Communications*, pp. 252–259, August 2005.

[25] P. Costa, C. Mascolo, M. Musolesi, and G. P. Picco, "Socially-aware routing for publish-subscribe in delay-tolerant mobile ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 5, pp. 748–760, 2008.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 409539, 13 pages
http://dx.doi.org/10.1155/2013/409539

Research Article

Application Scheduling in Mobile Cloud Computing with Load Balancing

Xianglin Wei,[1] Jianhua Fan,[1] Ziyi Lu,[1] and Ke Ding[2]

[1] *Nanjing Telecommunication Technology Research Institute, Nanjing 21007, China*
[2] *College of Command Information Systems, PLA University of Science and Technology, Nanjing 21007, China*

Correspondence should be addressed to Xianglin Wei; wei_xianglin@163.com

Received 19 April 2013; Revised 15 September 2013; Accepted 27 September 2013

Academic Editor: Chih-Hao Lin

Mobile cloud computing (MCC) enables the mobile devices to offload their applications to the cloud and thus greatly enriches the types of applications on mobile devices and enhances the quality of service of the applications. Under various circumstances, researchers have put forward several MCC architectures. However, how to reduce the response latency while efficiently utilizing the idle service capacities of the mobile devices still remains a challenge. In this paper, we firstly give a definition of MCC and divide the recently proposed architectures into four categories. Secondly, we present a Hybrid Local Mobile Cloud Model (HLMCM) by extending the Cloudlet architecture. Then, after formulating the application scheduling problems in HLMCM and bringing forward the Hybrid Ant Colony algorithm based Application Scheduling (HACAS) algorithm, we finally validate the efficiency of the HACAS algorithm by simulation experiments.

1. Introduction

Recent years have witnessed the rapid development of mobile devices, such as PDAs and smartphones. The tremendous improvement of hardware and software enables them to make calls and send short messages and emails, but it also gives them the ability to sense the environment and make social contacts, health care, and mobile learning. Moreover, the inherent mobility of the mobile devices enables the users to interact with the devices, environment, and social community without time and space restriction. Thus the mobile devices are able to integrate the capabilities of communication, work, medical treatment, and mobile learning, and they become important components of people's daily life. According to the International Data Corporation (IDC) Worldwide Quarterly Mobile Phone Tracker, it is estimated that 982 million smartphones will be shipped worldwide in 2015 [1]. However, mobile devices also have some inherent defects, such as their limited battery energy, low CPU speed, insufficient storage space, and inadequate sensing capacities [2]. These limitations have brought for mobile applications many challenges in mobility management, quality of service (QoS) insurance, energy management, and security issues. On the other hand, the lack of resources also motivates the researchers in mobile computing area to search for the infrastructure which can provide the needed resources for the mobile devices [3]. Consequently, cloud computing was introduced to fulfill this gap since it can theoretically provide nearly inexhaustible resources for mobile computing. The combination of cloud computing and mobile computing has stimulated the emergence of mobile cloud computing (MCC).

MCC brings rich resources of the cloud computing for mobile devices and applications, as well as inheriting the cloud's advantages, such as low cost, high scalability, and robustness. Therefore, it greatly improves the potential of mobile computing. In MCC environment, mobile devices can offload full or part of their mobile applications to the data centers of the cloud in order to relieve their own burden in CPU load and energy consumption. This enables them to support more sophisticated and richer applications and services, such as mobile game [4], mobile locating [5], voice, key words and picture searching [6–8], and mobile sensing [9]. In order to support these applications, researchers have proposed various architectures for MCC, such as MobiCloud, MAUI, CloneCloud, Cloudlet, and Hyrax. These proposals

attempt to utilize the cloud infrastructure as well as the mobile devices' idle CPUs or sensing capacities to achieve high QoS. However, offloading to remote cloud infrastructure will introduce long response latency. Hence, to reduce the response latency, the offloaded applications should be handled by local cloud infrastructure, which usually has fewer resources than those of the remote ones. Therefore, how to efficiently schedule the limited resources while fulfilling the requirements of a large number of offloaded applications is a critical problem for MCC to address.

In this paper, we first give the definition of MCC and divide the recently proposed architectures into four categories. Secondly, Hybrid Local Mobile Cloud Model (HLMCM) is put forward after combining the advantages of current proposals, such as Cloudlet and Hyrax. Thirdly, we formulate the application scheduling problems in HLMCM and bring forward the Hybrid Ant Colony algorithm based Application Scheduling (HACAS) algorithm. Finally, the effectiveness of HACAS algorithm is validated by simulation experiments.

The rest of the paper is organized as follows. Section 2 summarizes related work. Section 3 gives the definition of MCC and introduces the HLMCM. Section 4 formulates the application scheduling problem and proposes HACAS algorithm. Simulation experiments and results are shown in Section 5. Finally, we conclude our main work and present further research directions in Section 6.

2. Related Work

Application scheduling in cloud infrastructure has attracted much attention in recent years. In fact, the application scheduling and load balancing in MCC are burgeoning areas. Liu et al. have established a macroscopic scheduling model with cognition and decision components for the cloud computing, which considers both the requirements of different jobs and the circumstances of computing infrastructure. They have also put forward a job scheduling algorithm based on Multiobjective Genetic Algorithm (MO-GA), taking into account the energy consumption and the profits of the service providers [10]. In order to reduce the operator cost and to increase the reliability of the cloud service provider, Feller et al. have modeled the workload placement problem as an instance of the multidimensional bin-packing (MDBP) problem and have designed a novel, nature-inspired algorithm based on the Ant Colony Optimization (ACO) metaheuristics to compute the placement dynamically, according to the current load [11]. Goudarzi and Pedram have considered a multitier cloud computing environment, in which the clients have Service Level Agreements (SLAs) and the total profit in the system depends on how the system can meet these SLAs. They have proposed an algorithm based on force-directed search to allocate the resources, such as processing, memory requirement, and communication resources [12].

In the shared data center environment, Nagendram et al. have depicted the resource scheduling problem to a bounded multidimensional knapsack problem, taking into account the requirement dependency among multidimensional resources including memory, storage, CPU, and network bandwidth. Then, they have presented a Multidimensional resource integrated scheduling (MRIS), an inquisitive algorithm to obtain the approximate optimal solution [13]. To schedule the tasks and achieve load balance, Tayal has put forward an optimized algorithm based on the Fuzzy-GA optimization which makes a scheduling decision by evaluating the entire group of tasks in the job queue [14]. In the private cloud environment constructed for e-learning, Morariu et al. have presented a workload scheduling algorithm based on genetic algorithm [15]. For the load balancing problem of the VM scheduling in the cloud computing, Gu et al. have proposed a scheduling strategy on load balancing of VM resources based on genetic algorithm [16]. Yamauchi et al. proposed a distributed parallel scheduling methodology for MCC and developed a simulator to analyze the bottleneck of MCC [17].

Ant Colony Optimization (ACO) has also been used to balance the load in cloud environment. In [18], Mishra et al. have proposed a method to utilize the ACO for load balancing in cloud environment. The routing packets in this environment are treated as the ants in the network. Moreover, they replaced the routing tables in the network nodes by tables of probabilities. These tables are also called "pheromone tables" since the pheromone strengths used in ACO are represented by these probabilities. Every node has a pheromone table for every possible destination in the network, and each table has an entry for every neighbor. The entries in the tables are the probabilities which influence the ants' selection of the next node on the way to their destination node. Consequently, at each node, the ant should choose the next node toward its final destination node according to the probabilities. After arriving at a node, the ants update the probabilities of that node's pheromone table entries corresponding to their source node; that is, ants lay the kind of pheromone associated with the node they were launched from. They alter the table to increase the probability pointing to their previous node. Besides, in order to distribute the load among many paths from the source node to the destination node, the ants from one colony will consult the routing tables of other colonies so as to avoid routing packets to those paths that are highly preferred by the other groups. Through this method, the loads are separated among many possible paths in the network.

Zhu et al. considered the task scheduling in cloud environment from the perspectives of QoS fulfillment and shortest path [19]. Nishant et al. have proposed an algorithm for load distribution of workloads among nodes of a cloud by the use of ACO. Moreover, they have presented another load balancing algorithm which ensures that there is no conflict of interests based on relocating the tasks among nodes [20]. In these works, they do not take each task's profit into consideration and cannot maximize the profit of the system, which is an import target of the scheduling algorithm for the commercial mobile cloud environment. In grid computing, an ACO algorithm is proposed by Suryadevera et al. for load balancing which will determine the best resource to be allocated to the jobs, based on resource capacity, and at the same time balance the load of entire resources on grid. The main objective of this algorithm is to achieve high throughput and thus increases the performance in grid environment

[21]. A review on the load balancing studies for the cloud environment is presented in [22].

Different from the scheduling algorithm in cloud environment, the scheduling algorithms for MCC should take the energy consumption into consideration. To make the system last longer, the scheduling algorithms should balance the load of the mobile devices to avoid some heavy-loaded nodes leaving the system too early. Therefore, these scheduling algorithms for cloud computing cannot be applied to the MCC environment directly. In order to bridge this gap, we propose an architecture for MCC and then present a scheduling algorithm for MCC which can maximize the profit and balance the load of the mobile devices.

3. Definition and Architecture

3.1. The Definition of MCC and Current Proposed Architectures. The Mobile Cloud Computing Forum defines MCC as follows [23]: "Mobile cloud computing, at its simplest, refers to an infrastructure where both the data storage and the data processing happen outside of the mobile device. Mobile cloud applications move the computing power and data storage away from mobile phones and into the cloud, bringing applications and mobile computing to not just smartphone users but a much broader range of mobile subscribers." Aepona describes MCC as a new paradigm for mobile applications whereby the data processing and storage are moved from the mobile device to powerful and centralized computing platforms located in clouds [24]. These centralized applications are then accessed over the wireless connection based on a thin native client or web browser on the mobile devices. In [25, 26], MCC is described as a combination of mobile web and cloud computing, which is the most popular tool for mobile users to access applications and services on the Internet. Dinh et al. defined MCC as an entity that provides mobile users with the data processing and storage services in clouds [27].

These definitions are mostly descriptive. This paper gives the following definition: MCC is a mobile application-oriented computing paradigm in mobile and dynamic environment, which makes use of the resources provided by clouds, mobile devices, and network facilities to fulfill users' requirements on QoS, quality of experience (QoE), security and privacy, with some particular cost, energy, and programming model and context information.

In order to efficiently utilize the available resources, researchers have brought forward several MCC architectures. In this paper, we divide them into four categories. In the proposals of the first category, mobile devices first offload applications to the remote large data centers of the clouds, from which the results will be returned; the typical proposals include MAUI [28] and CloneCloud [29]. In the second category, Satyanarayanan et al. introduced the Cloudlet entity to the system, which is a local service infrastructure logically implemented at the access point of mobile devices, and the mobile devices only have to offload applications to the Cloudlet rather than remote data centers [27, 30]. It should be noted that Cloudlet can reduce the server response latency since the offloading happens locally most of the time. In the proposals of the third category, mobile devices collaborate with each other to run applications without the need to rely on any cloud infrastructure. This sounds like the mobile Peer-to-Peer system and mobile grid computing, but different from these computing paradigms, the typical schemes (such as Hyrax [31] Misco [32], and the virtual cloud [33]) of this category adopt the unique characteristics of MCC such as fault tolerant and application partition. The typical schemes of the fourth category move the cloud infrastructure close to the users to improve the timeliness of the service, such as MobiCloud [34–36].

Figure 1 illustrates the traditional architecture which uses remote cloud infrastructure via the backbone network. As shown in Figure 1, the latency of this architecture consists of the time spent on the access network, the backbone network, and the time spent inside the cloud infrastructure. In the Cloudlet architecture, as shown in Figure 2, most of the time, the application will not deliver to the remote cloud infrastructure, and thus the latency is composed of the one-hop time spent on the access network, which is much lower than those using architecture of Figure 1 [30, 37]. Under some special conditions, Cloudlet cannot handle the offloaded applications locally, and it needs the remote cloud infrastructure's help for processing them. The latency under these conditions approximates the traditional architecture which directly uses the remote cloud infrastructure.

3.2. Hybrid Local Mobile Cloud Model. In order to provide high QoS for mobile applications, the MCC architecture should have low response latency. Moreover, the mobile devices' participation for providing their idle computing and sensing capabilities is also very critical for the promotion of the users' QoE (quality of experience). This is due to the fact that one single mobile device's sensing result can be easily influenced by its local environment and hence is error-prone, while aggregating a few mobile devices' sensing results about the area can provide more correct context information. Therefore, we have modified the Cloudlet architecture to make the mobile devices contribute their computing and sensing capabilities like they do in Hyrax [31] and Misco models [32]. This new mobile cloud computing model is called the Hybrid Local Mobile Cloud Model (HLMCM) and is illustrated in Figure 3.

From Figure 3, we can see that HLMCM consists of a Cloudlet and a set of mobile devices. The mobile devices are connected to the Cloudlet via wireless links, such as WiFi and WiMAX. Similar to the original Cloudlet architecture, the Cloudlet in HLMCM is logically attached to the access point of the mobile devices to achieve low response latency, and the mobile devices can offload their mobile applications to the Cloudlet. However, different from Satyanarayanan's scheme [30], in HLMCM, the mobile devices collaborate with the Cloudlet to provide service. This designation is based on the following considerations.

(1) Mobile devices' computing, storage, and sensing capabilities are increasingly becoming powerful. However, the utilization ratio of these resources is low

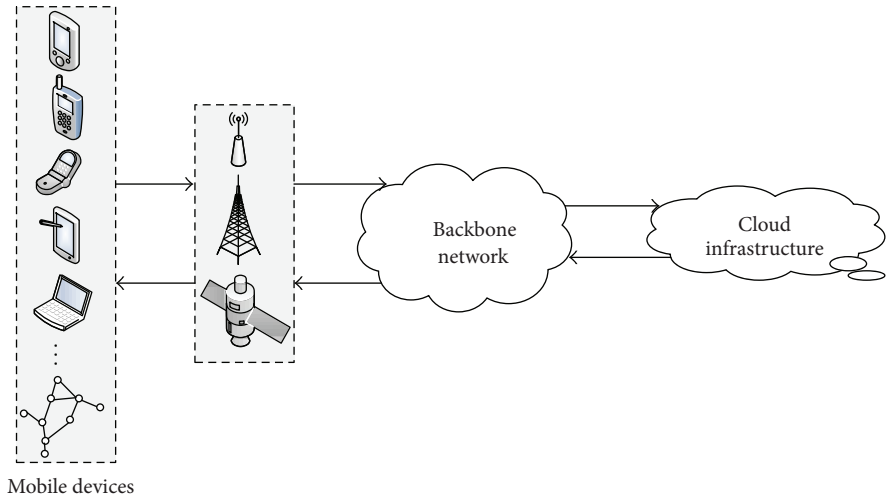


Figure 1: The architecture that uses the remote cloud infrastructure.

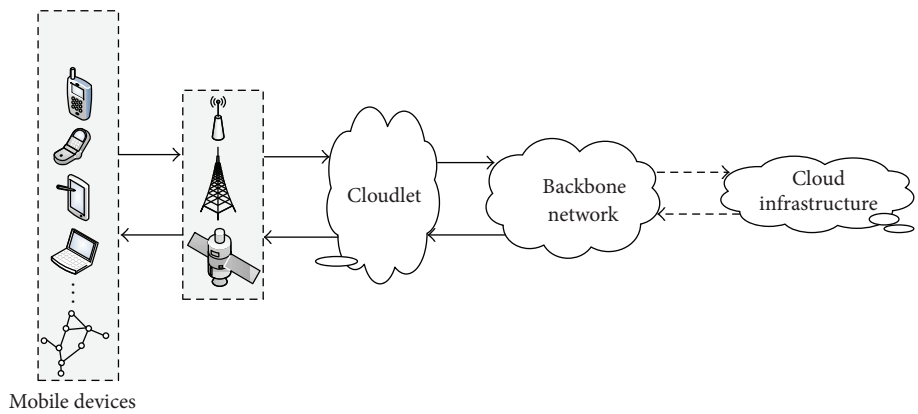


Figure 2: The Cloudlet architecture.

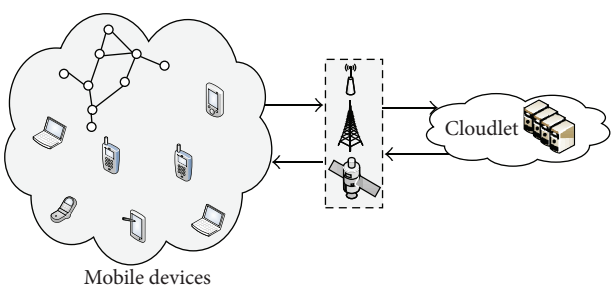


Figure 3: The architecture of the hybrid local mobile cloud model.

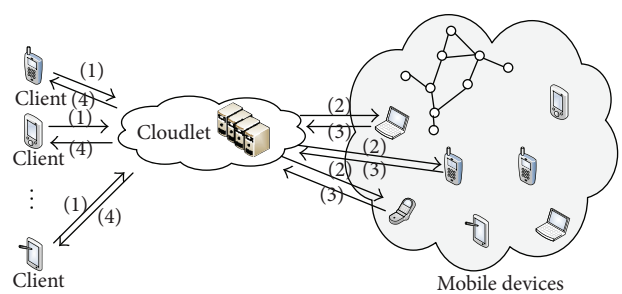


Figure 4: The general working process of HLMCM.

at most times, which means that the mobile devices usually have idle resources for sharing.

(2) The Cloudlet usually only contains a few servers and is much less powerful than the data center of the typical large-scale cloud. Therefore, the mobile devices' participation can promote the scalability of HLMCM.

(3) The involvement of the mobile devices can help improve the QoS provided by HLMCM, especially the sensing capabilities.

The general working process of HLMCM is shown in Figure 4, and it mainly contains the following four steps.

(1) The clients offload part or full of their applications to the Cloudlet. Note that the clients are mobile devices

as well, and they can provide service for other devices' mobile applications.

(2) The Cloudlet executes the application scheduling algorithm to offload the applications to a few mobile devices that are willing to provide resources.

(3) The mobile devices handle the applications received and send their results to the Cloudlet.

(4) The Cloudlet sends the results of the applications from the mobile devices to the clients.

Note that there may be a large number of mobile devices requesting for offloading applications to the HLMCM whose sensibility, computing, and storage capabilities are usually much less than those of the large cloud infrastructure. Therefore, efficient application scheduling algorithm is critical for HLMCM to provide high QoS. Moreover, the application scheduling algorithm should consider the profits of the HLMCM as well as balancing the load of the mobile devices to make the whole system last longer.

4. Model and Algorithm

4.1. Model and Problem Statement. In HLMCM environment, the scheduling algorithm is in charge of allocating the offloaded applications from the mobile devices to the service providers, including the Cloudlet and $m-1$ mobile devices in the system. For ease of description, the service provider will be referred to as provider in the following analysis.

Assume the dimension of the resources is d and each provider's resources can be expressed as a vector $\vec{c}_i = (c_i^1, \ldots, c_i^d)$, in which c_i^k is the kth dimensional resource that the provider i has. Assume that the set of applications that arrives at some particular time slot is $I = \{1, 2, \ldots, n\}$, and the value of the application j is p_j, $j = 1, 2, \ldots, n$.

The resources consumed by application j when executed on provider i are a vector $\vec{r}_{ij} = (r_{ij}^1, \ldots r_{ij}^d)$, $i = 1, 2, \ldots, m$. Assume that each application can only be executed on one provider and cannot be further partitioned. Once an application is executed successfully on some provider, HLMCM will receive the value of this application as its profit. Here, the scheduling target is to maximize the total profits of HLMCM with the constraint of resource capacity of each service provider. Therefore, the scheduling problem can be formulated as follows.

$$\begin{aligned} &\text{Maximize} \quad \sum_{j=1}^{n} p_j \sum_{i=1}^{m} x_{ij} \\ &\text{subject to} \quad \sum_{j=1}^{n} \vec{r}_{ij} x_{ij} \le \vec{c}_i, \quad i = 1, 2, \ldots, m \\ &\qquad \sum_{i=1}^{m} x_{ij} \le 1, \quad j = 1, 2, \ldots, n \\ &\qquad x_{ij} \in \{0, 1\}, \quad j = 1, 2, \ldots, n. \end{aligned} \tag{1}$$

From (1), we can see that this can be seen as a multidimensional 0-1 knapsack problem and is NP-hard.

Besides, in order to make HLMCM last longer, the applications should be uniformly executed on the mobile devices. This will make the devices consume their energy evenly to avoid the phenomenon that some mobile devices with heavy load consume their energy too early and have to leave the system.

For some particular provider i, the load of its kth dimensional resource is defined as.

$$L_{ik} = \frac{\sum_{j=1}^{n} r_{ij}^k x_{ij}}{c_i^k}. \tag{2}$$

i's load L_i is defined as the mean value of all its d-dimensional resources' loads; that is,

$$L_i = \frac{\sum_{k=1}^{d} L_{ik}}{d}. \tag{3}$$

4.2. HACAS Algorithm. In order to solve this problem, this paper proposes a scheduling algorithm based on the hybrid ant colony algorithm which has been widely used to solve complex combinatorial optimization problems [38]. The following part of this section presents HACAS algorithm, which contains the pheromone value and its update model, local heuristic value, application scheduling probability, tabu list and the bulletin board, and provider selection scheme.

4.2.1. Pheromone Value and Its Update Model. The application scheduling problem in this paper belongs to the subset problem [39]; that is, given a set S which contains n applications for scheduling and the evaluation function $f(\,)$, the target is to select the best subset of S to maximize or minimize $f(\,)$. There may be more than one evaluation functions while this section focuses on the case where there is only one evaluation function. In this situation, the order of the selected applications is not important, and the pheromone value is placed on the application rather than the connection among the applications, which means that applications with a higher pheromone value can better satisfy the requirements of the evaluation function. When the specific condition is met, such as a partial solution with some particular length is obtained, the pheromone value needs to be updated. The update process includes two parts. Firstly, the pheromone value of each application is reduced by a certain percentage to emulate the real-life behavior of evaporation of pheromone count over time; Secondly, the pheromone value increment laid by the new partial solutions of the ants will be added. Assume that the pheromone value on application i at time t is $\tau_i(t)$; then at the next update time t', the value is updated to $\tau_i(t')$:

$$\tau_i\left(t'\right) = \left(1 - \rho\right) \tau_i\left(t\right) + \Delta\tau_i\left(t, t'\right), \tag{4}$$

where $0 < \rho \le 1$ is a coefficient which represents pheromone evaporation, $\Delta\tau_i(t, t')$ is the pheromone value increment obtained from all the ants' partial solutions; that is,

$$\Delta\tau_i\left(t, t'\right) = \sum_{j=1}^{q} \Delta\tau_i^j\left(t, t'\right), \tag{5}$$

where q is the number of the ants and $\Delta\tau_i^j(t,t')$ is the pheromone value laid on application i by ant j's partial solution at time t' and is defined as

$$\Delta\tau_i^j\left(t,t'\right) = \begin{cases} G\left(f\left(\widetilde{S}_j\left(t'\right)\right)\right), & \text{if } j\text{th ant incoporates application } i \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where $\widetilde{S}_j(t')$ is the partial solution of ant j at time t' and $f(\widetilde{S}_j(t'))$ is the value of the evaluation function of this solution. To maximize the profit, the evaluation is defined as

$$f\left(\widetilde{S}_j\left(t'\right)\right) = \sum_{k\in\widetilde{S}_j(t')} p_k; \tag{7}$$

that is, the total value of the applications belongs to $\widetilde{S}_j(t')$. The function G in (4) depends on the problem; in this paper, it is defined as $G(f(\widetilde{S}_j(t'))) = Qf(\widetilde{S}_j(t'))$, in which Q is a parameter of the method.

4.2.2. Local Heuristic Value. The positive feedback of the ant colony algorithm is usually combined with some local heuristic schemes to accelerate the search process. In HLMCM, the local heuristic scheme needs to consider the profits of the applications as well as the resources they consume.

Let $\vec{\mu}_k(j,t) = \sum_{l\in\widetilde{S}_j(t)} \vec{r}_{kl}$ be the resources consumed on provider k by the partial solution $\widetilde{S}_j(t)$ constructed by ant j. Then, the remaining resources on provider k are $\vec{\gamma}_k(j,t) = \vec{c}_k - \vec{\mu}_k(j,t) = (\gamma_k^1(j,t),\ldots,\gamma_k^d(j,t))$, in which $\gamma_k^i(j,t)$ is the remaining amount of the ith dimensional resource. The tightness of application h on provider k on the ith dimensional resource is defined as

$$\left|\frac{r_{kh}^i}{\gamma_k^i\left(j,t\right)}\right| \tag{8}$$

that is the ratio between r_{kh}^i, the amount of provider k's resource consumed by application h, and $\gamma_k^i(j,t)$. Moreover, the tightness of application h on provider k is defined as

$$\delta_{kh}\left(j,t\right) = \left|\frac{r_{kh}^1}{\gamma_k^1\left(j,t\right)}\right| + \cdots + \left|\frac{r_{kh}^d}{\gamma_k^d\left(j,t\right)}\right|. \tag{9}$$

In (9), the tightness of the d-dimensional resources is converted into a single value which comprehensively considers all dimensions of the resources.

The average tightness on all providers in case of application h being chosen to be included in $\widetilde{S}_j(t)$ is

$$\overline{\delta}_h\left(j,t\right) = \frac{\sum_{k=1}^m \delta_{kh}\left(j,t\right)}{m}. \tag{10}$$

In order to consider the application h's profit as well as its resource requirement, the local heuristic value $\eta_h(\widetilde{S}_j(t))$ is defined as

$$\eta_h\left(\widetilde{S}_j\left(t\right)\right) = \frac{p_h}{\overline{\delta}_h\left(j,t\right)}. \tag{11}$$

4.2.3. Application Scheduling Probability. After obtaining the pheromone value and local heuristic value on each application, the probability that h has to be selected as the next scheduling application of $\widetilde{S}_j(t)$ is

$$P_h^j\left(\mathrm{t}\right) = \begin{cases} \dfrac{[\tau_h\left(t\right)]^\alpha[\eta_h(\widetilde{S}_j\left(t\right))]^\beta}{\sum_{k\in\text{allowed}_j(t)}[\tau_k\left(t\right)]^\alpha[\eta_k(\widetilde{S}_j\left(t\right))]^\beta}, & \text{if } h\in\text{allowed}_j\left(t\right) \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

where $\text{allowed}_j(t) \subseteq S - \widetilde{S}_j(t)$, is the set of the remaining schedulable applications. From (12), we can see that the more pheromone value and local heuristic value an application has, the higher the probability will be scheduled.

4.2.4. Tabu List and the Bulletin Board. A data structure, called a *tabu list*, is associated to each ant in order to avoid that ant from scheduling an application more than once. This list $\text{tabu}_j(t)$ maintains a set of scheduled applications up to time t by ant j. Let the applications that can be executed on at least one of the providers be F; then we have $\text{allowed}_j(t) = F \cap \{h \mid h \notin \text{tabu}_j(t)\}$.

In addition, we set a bulletin board to record the best solution up to time t, with which each ant can compare its own solution. If its solution is better than the best one, it will update the best one with its solution.

4.2.5. Provider Selection Scheme. After deciding the scheduled application, there is usually more than one provider who have sufficient resources to execute the application. Note that traditional hybrid algorithm only focuses on the application scheduling probability. In order to balance the load of the providers, we take the provider selection scheme into consideration in this section.

For some particular feasible provider i, the load of its kth dimensional resource is L_{ik} as defined in (2). After adding application h, the expected load of its kth dimensional resource is

$$L'_{ik} = L_{ik} + \frac{r_{ih}^k}{c_i^k}. \tag{13}$$

The expected load of provider i is defined as

$$L'_i = \frac{\sum_{k=1}^d L'_{ik}}{d}. \tag{14}$$

This means that if is h executed on provider i, i's expected load will be L'_i.

In order to balance the load of all the providers, the provider with the lowest expected load will be selected to execute application h.

4.2.6. Algorithm. The HACAS algorithm is illustrated in Algorithm 1. The parameters needed to be settled include C,

```
(1) Initialize r⃗_ij, c⃗_i, p_j, 1 ≤ i ≤ m, 1 ≤ j ≤ n, q = n, number
    of cycles C, τ_j (0) = 1/q, 1 ≤ j ≤ q, tabu_list = [],
    best_solution = 0, application_provider, α = β = 1, ρ = 0.3, Q = 1
(2) for (t = 1: t <= C; t++)
(3)   for (j = 1; j < q; j++)
(4)     random_first = 0
(5)     while allowed_j (t) ≠ ∅
(6)       if random_first == 0
(7)         Select the first scheduled application randomly
(8)         random_first = 1
(9)       else
(10)        Select the scheduled application h according to (12)
(11)      end if
(12)        Calculate the expected loads of all feasible providers according to (14)
(13)        Rank the feasible providers according to their expected loads in an increasing order
(14)        The provider with the lowest expected load is selected for h
(15)        Add h to S̃_j (t) and the tabu_list
(16)    end while
(17)        Calculate f(S̃_j (t)), which is the object function of the generated solution of ant j
(18)    if f(S̃_j (t)) > best_solution
(19)        best_solution = f(S̃_j (t))
(20)        Save ant j's solution in application_provider
(21)    end if
(22) end for
(23) Calculate the incremental pheromone on each application according to (5)
(24) Clear the tabu_list for each ant
(25) end for
(26) print best_solution
(27) print application_provider
```

Algorithm 1: The HACAS algorithm.

α, β, ρ, q, Q, $\vec{r}_{ij}$, $\vec{c}_i$, and p_j. The initial pheromone trail value on each application is set to be $1/q$.

Step 1 initiates the parameters. Step 4 introduces a variable *random_first* to enable each ant to randomly select its first scheduling application. Steps 5 to 16 present the solution-searching process of an ant. Firstly, Steps 6 to 11 select the next scheduled application, that is, randomly selecting the first one and then scheduling other applications according to the probabilities calculated in (12). Secondly, Step 12 calculates the expected loads of all feasible providers according to (14), Step 13 ranks the feasible providers according to their expected loads in an increasing order, Step 14 selects the provider with the lowest expected load for h, and finally, Step 15 adds the newly scheduled application to the partial solution as well as the *tabu list*. At the end of Step 16, each ant has found its solution $\widetilde{S}_j(t)$; Step 17 calculates $f(\widetilde{S}_j(t))$. If $f(\widetilde{S}_j(t))$ is larger than the *best_solution* in the bulletin board, then Step 19 will assign $f(\widetilde{S}_j(t))$ to *best_solution* and save the scheduling results into *application_provider*. After Step 22, all the ants have finished searching for the solutions, and one cycle is finished. Step 23 calculates the pheromone value increment according to (5). Step 24 clears the *tabu lists* of the ants. Steps 26 to 27 print the best solution found at the Cth cycle and the scheduling results.

5. Simulation, Results, and Discussion

5.1. Experimental Settings

5.1.1. Experimental Environment. To evaluate the performance of the algorithms proposed in this paper, we have conducted many simulation experiments, whose parameters are listed in Table 1.

In this simulation, the dimension of the resources is 2. Both the first and the second dimensional resources possessed by each provider obey uniform distribution in the interval $[a_1, a_2]$. There are m providers and n applications in the system. At time t, these applications arrive simultaneously. The applications' resource consumption of the first and the second dimensional resources on some provider obeys uniform distribution in the intervals $[a_3, a_4]$ and $[a_5, a_6]$, respectively. Moreover, the applications' profits obey uniform distribution in the interval $[a_7, a_8]$. The number of cycles (i.e., C) is set to be 10. The default parameters are listed in Table 1.

Based on these parameters, a series of simulation experiments has been conducted. The experiments contain 10 cycles. In each cycle, each ant searches for its own scheduling result based on the method presented in the scheduling algorithm in the above section. At the end of each cycle,

Table 1: Simulation parameters.

Parameter	Default Value
n	100
a_1	31
a_2	100
a_3	10
a_4	30
a_5	10
C	10
β	1
Q	1
m	20
a_6	30
a_7	5
a_8	30
α	1
ρ	0.3
θ	1
λ	1

the pheromone value on the applications will be updated and the bulletin board is used to record the best scheduling result.

5.1.2. Comparison Benchmark and Metrics. We evaluate HACAS algorithm from two different angles. Firstly, in order to validate the effectiveness of HACAS algorithm, we compare it with the First-Come-First-Served (FCFS) algorithm. In FCFS, the applications are scheduled according to their arrival order, and the providers are selected randomly from those who can execute the application. The profit of the scheduling algorithm, which is defined as the total profits of all the applications scheduled by the algorithm, is chosen as the metrics to compare them.

Secondly, in order to evaluate the provider selection scheme of HACAS algorithm, a scheduling algorithm with random provider selection is adopted as the comparison benchmark, in which steps 12–14 in Algorithm 1 are replaced with the following step.

(12) *Randomly select the service provider for h from those feasible providers.*

The scheduling algorithm with this modification is called the Hybrid Ant Colony algorithm based Application Scheduling with Random Provider Selection (HACASRPS) algorithm.

For the solution of ant j at the kth cycle, let provider i's load be L_i^{jk}. The mean value (μ_k^j) and the standard deviation (σ_k^j) of all the m providers' loads at the kth cycle of ant j's solution are defined as

$$\mu_k^j = \frac{\sum_{i=1}^{m} L_i^{jk}}{m}, \qquad \sigma_k^j = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(L_i^{jk} - \mu_k^j\right)^2}. \tag{15}$$

σ_k^j reflects the deviation of all the providers' loads of ant j's solution at the kth cycle. Then, at the end of the kth cycle, the mean value of the standard deviation of all the providers' loads of all the q ants' solution is defined as

$$\mu_\sigma^k = \frac{\sum_{j=1}^{q} \sigma_k^j}{q}. \tag{16}$$

In order to simplify the expression, μ_σ^k will be referred to as the load variation of the scheduling algorithm at the kth cycle. Then, the average load variation of the scheduling algorithm of all the simulation cycles can be defined as

$$\mu_\sigma = \frac{\sum_{k=1}^{C} \mu_\sigma^k}{C}. \tag{17}$$

We define the average load of a scheduling algorithm in kth cycle by

$$\mu_\mu^k = \frac{\sum_{j=1}^{q} \mu_k^j}{q}. \tag{18}$$

Then the average load of a scheduling algorithm is defined by

$$\mu_\mu = \frac{\sum_{k=1}^{C} \mu_\mu^k}{C}. \tag{19}$$

σ_k^j, μ_μ, and μ_σ^k are selected as the metrics to evaluate the effectiveness of the provider selection scheme of the HACAS algorithm.

5.2. Experimental Results

5.2.1. The Profits. With the parameters in Table 1, the profit of FCFS algorithm is 1324. The profits of HACASRPS and HACAS algorithms are shown in Figure 5.

From Figure 5, we can see that as cycle increases, the profits of both HACASRPS and HACAS algorithms increase. This is due to the fact that as cycle increases, both the HACASRPS and HACAS algorithms will find more profitable scheduling results which will bring more profits. At the end of the 10th cycle, the profits of HACASRPS and HACAS algorithms are 1747 and 1794, respectively, which are more than 30% higher than that of FCFS algorithm. This phenomenon can be attributed to the local heuristic value adopted by both algorithms, which enables them to schedule those applications which consume less resources while bringing more profits with preference.

5.2.2. Load Balancing. Balancing the load of all the providers to make the system last longer is an important target of HACAS algorithm's provider selection scheme. This section investigates the effectiveness of this selection scheme and compares it with random provider selection method adopted in HACASRPS algorithm.

With the parameters in Table 1, the average loads of HACAS and HACASRPS algorithms are 0.800 and 0.779,

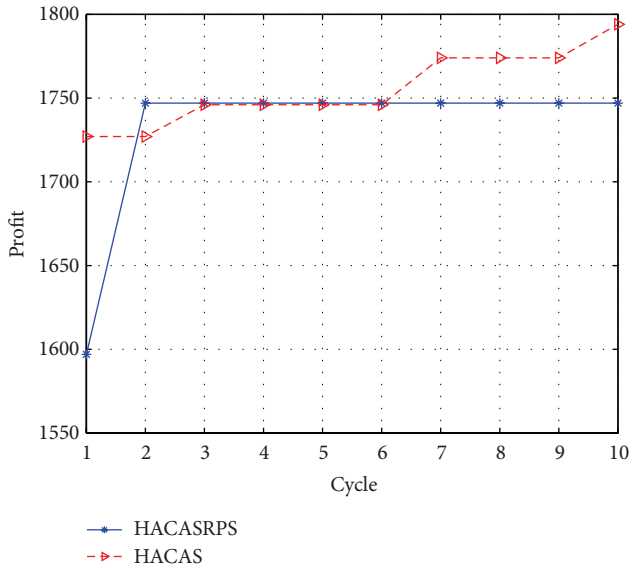


Figure 5: The profits of HACASRPS and HACAS algorithms. The horizontal axis represents the simulation cycle; the longitudinal axis represents the profit of the algorithm.

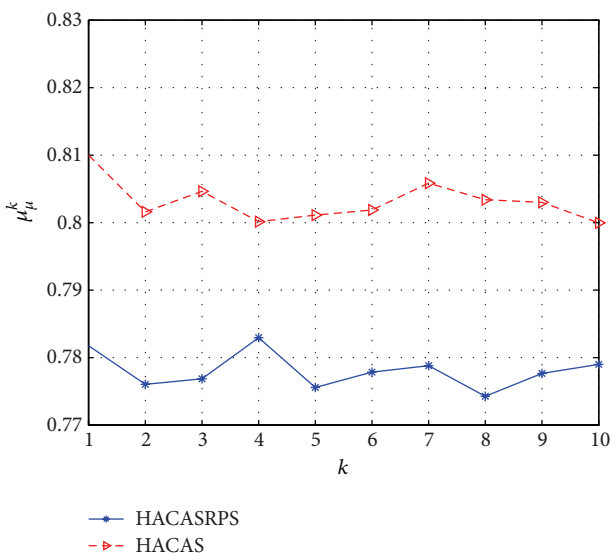


Figure 6: The average load of HACAS and HACASRPS algorithms in different simulation cycles; k is the simulation cycle, and μ_μ^k is the average load of the algorithm in the kth cycle.

respectively. The former is slightly higher than the latter, which is due to the fact that HACAS schedules more applications than HACASRPS algorithm as shown in Figure 5. The average load of HACAS and HACASRPS algorithms in different simulation cycles (as defined in (18)) are shown in Figure 6. From Figure 6, we can see that the average load of both algorithms stays stable as cycle increases.

Figure 6 tells us that the average loads of HACAS and HACASRPS algorithms are 0.8 and 0.78, respectively. We further investigate the load variations of both algorithms

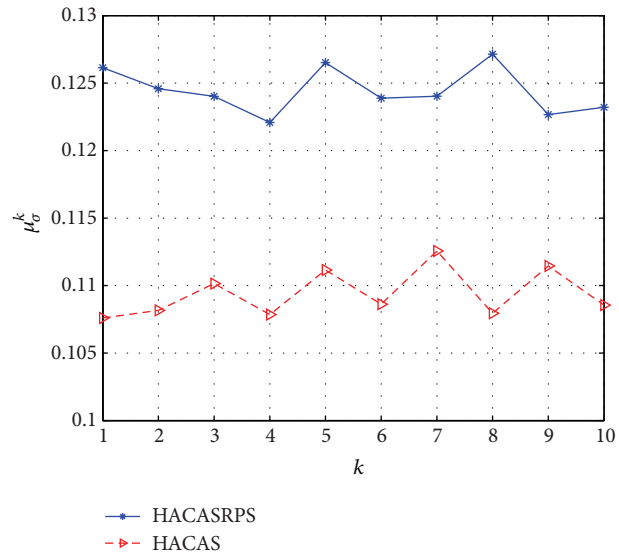


Figure 7: The load variations of HACAS and HACASRPS algorithms in different simulation cycles; k is the simulation cycle, and μ_σ^k is the load variation of the algorithm in the kth cycle.

in different simulation cycles, and the results are shown in Figure 7.

From Figure 7, we can see that the load variations of these two algorithms maintain stable as simulation cycle increases. Moreover, the load variation of HACAS algorithm is much lower than that of HACASRPS algorithm. In some particular cycle, HACAS algorithm's load variation is about 13% lower than that of the HACASRPS algorithm. This is because the provider selection scheme adopted in HACAS algorithm takes the providers' load into account when choosing the provider for the scheduled applications. This means that the provider selection scheme in HACAS algorithm can effectively balance the load of the providers more effectively.

5.2.3. Parameters' Influence. In the above experiments, the number of the applications for scheduling is large and the load of the providers is high. This section shows the results when the number of the applications for scheduling is relatively small. Concretely speaking, we set $m = 20$ with $n = 30$ in this experiment.

Firstly, we investigate the load of all the providers of the 10th ant's solution at the 5th cycle, and the results are shown in Figure 8.

From Figure 8, we can see that the load of the providers in HACASRPS algorithm fluctuates in a much wider range than that of HACAS algorithm. In HACASRPS algorithm, the load of the 7th provider is almost 0.8, while the loads of the 16th and the 18th provider are 0. In contrast, the load of the providers in HACAS algorithm is mostly between 0.3 and 0.6. The notable differences of the resource consumption among different applications (from 10 to 30) have led to the differences among the loads of the providers.

Then we show the standard deviation of all the providers' loads of the ant's solution in the 5th cycle in Figure 10. Note that there are 30 ants in the algorithm since $q = n$.

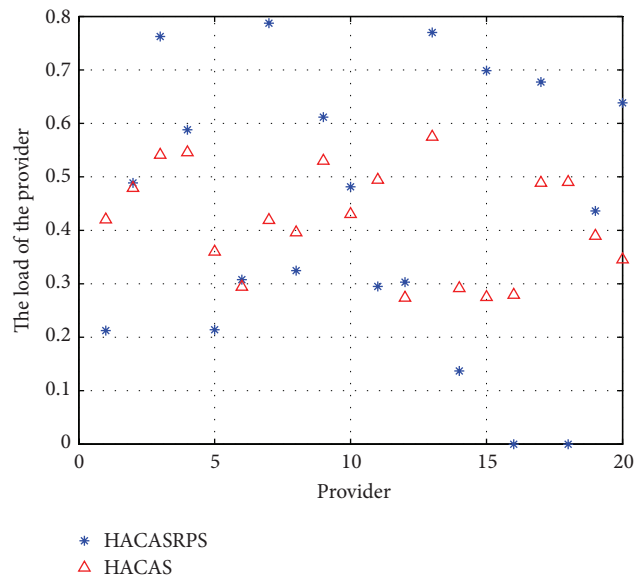


Figure 8: The load of all the providers of the 10th ant's solution at the 5th cycle.

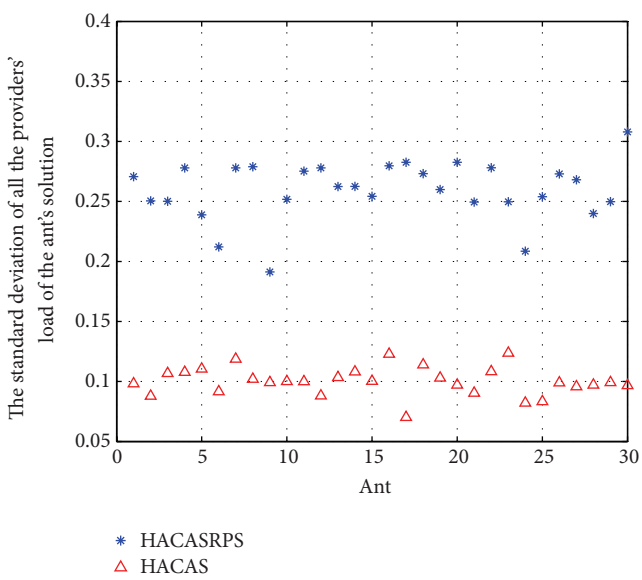


Figure 9: The standard deviation of all the providers' loads of the ant's solution in the 5th cycle.

Figure 9 reveals that the standard deviations of all the providers' loads of the ant's solution of HACAS algorithm are much lower than those of the HACASRPS algorithm.

Similar to Figure 7, Figure 10 further reveals the load variations of HACAS and HACASRPS algorithms in different simulation cycles when $n = 30$, from which we can derive similar observations with those drawn from Figure 7. Moreover, in some particular cycle in Figure 10, HACAS algorithm's load variation is about 60% lower than that of the HACASRPS algorithm. Therefore, after combining Figures 7 and 10, we know that the effectiveness of the provider

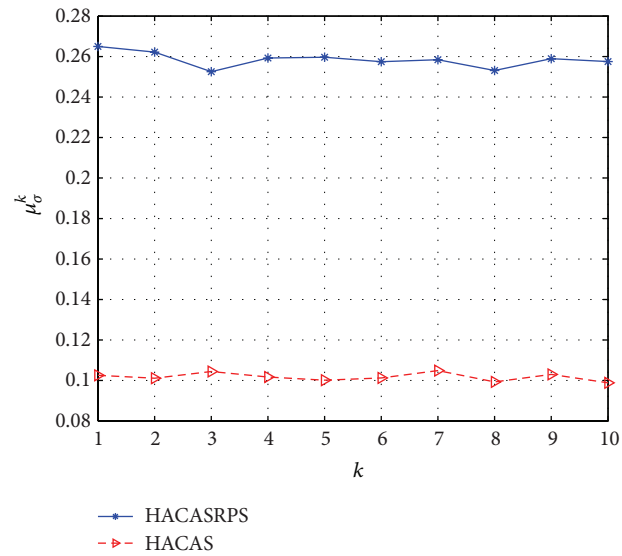


Figure 10: The load variations of HACAS and HACASRPS algorithms in different simulation cycles; k is the simulation cycle, and μ_σ^k is the load variation of the algorithm in the kth cycle.

selection scheme of the HACAS algorithm becomes more prominent when the load of the system is low.

5.3. Discussion and Application Scenario

5.3.1. Discussion. As shown in Section 5.2, the performance of HACAS algorithm is better than that of FCFS and HACASRPS algorithms. This phenomenon can be attributed to HACAS algorithm's pheromone value and its update model, application scheduling model, and provider selection scheme. Concretely speaking, pheromone value and its update model make HACAS learn from its historical decision to raise the profit of the system. Moreover, application scheduling model takes the pheromone value as well as application's resource consumption into consideration and can help HACAS algorithm choose those applications with the highest profit and the lowest resource consumption. Last but not the least, the provider selection scheme can balance the load of the mobile devices, which is very important to make the system last longer.

In addition, the following section shows the reason why we propose HLMCM and choose simulation parameters.

5.3.2. The Rationale behind Proposing HLMCM. We put forward HLMCM since we cannot fulfill users' QoE requirement through simply extending the cloud infrastructure or the Cloudlet entity's computing or storage capacity. For instance, in a cooperation sensing environment, the accurate sensing result can only be drawn through jointly using many mobile devices' diverse sensing results (such as location, orientation, and temperature) [3]. Besides, in some circumstances, communication using backbone links may not be always present in some isolated areas, during rescue missions, uprisings, and disaster scenarios [37, 40]. Under these circumstance,

the mobile devices can only search for the local infrastructure such as the Cloudlet entity which is easy to implement.

5.3.3. Rationale for Choosing the Simulation Parameters Presented in Table 1. In the simulation part, a mobile device is assumed to have at least enough resources to run an offloaded application. This is also the basic assumption in the knapsack problem that the maximum volume of the objects (in this paper, 30) is smaller than the minimum capacity of the knapsacks (in this paper, 31). Moreover, we notice that in Table 1, the resources possessed by each provider obey uniform distribution in the interval $[a_1, a_2]$, while $a_1 = 31$ and $a_2 = 100$. This setting is attributed to following observation. The resources in mobile devices mainly contain CPU, storage, and sensors, and so forth. The process frequency of mainstream smartphones' CPU ranges from 800 MHz to 2 GHz. Moreover, the storage capacity of the mainstream smartphones ranges from 16 GB to 64 GB. The number of sensors (including proximity sensor, Global Positioning System, accelerometer, compass, and gyros) on each smartphones ranges from 2 to 6. If we treat these devices as general resources, then we know that the volume difference of the resource possessed by the devices is about 3 times. Therefore, in Table 1, the volume difference of the resource processed by the devices is also set to be around 3. Based on this consideration, the maximum and the minimum resources possessed by each device in Table 1 are set to be in the interval (31, 100). The profit and the energy consumption difference of the applications are based on the observations and current studies [41] on the mainstream applications (such as game, web browser, etc.) in the app store (such as the iPhone App store). The other parameters, such as α, β, ρ, and Q, are decided by the default parameters used by the hybrid ant colony algorithm.

In this paper, the parameters used in Table 1 can validate that HACAS algorithm is effective under heavy load environment (i.e., the applications' total resource requirements exceed the resources possessed by the system). Moreover, in the section "Parameters' influence," n is set to be 20 to evaluate the effectiveness of HACAS under light load environment. Notice that these parameters can be adjusted as the simulation needs and the proposed HACAS algorithm can adapt to various circumstances.

5.3.4. Application Scenario. The case for mobile cloud computing can be argued by considering the unique advantages of empowered mobile computing, and a wide range of potential mobile cloud applications have been recognized in the literature. These applications fall into different areas such as image processing, natural language processing, sharing GPS, sharing Internet access, sensor data applications, querying, crowd computing, and multimedia search. A survey of the possible applications can be referred to [42].

Here, we show an application scenario that applies MCC for disaster rescue. In a disaster-stricken environment (such as hurricane, tsunamim, and earthquake), the communication infrastructure can be seriously damaged, if not completely destroyed. Moreover, many roads could also get blocked. These damages make it difficult for the rescuers to find the location of wounded people or even to get a global view of the disaster area. Under such circumstances, the rescuers can deploy some emergency communication facilities (such as communication vehicles) with Cloudlet entities. Then, the mobile devices (especially the smartphones) near the vehicles can communicate with each other, report the location of the wounded people, and upload the pictures or videos around themselves for processing to help the rescue process. Moreover, the devices also need the Cloudlet to provide them with the needed information (such as the latest map in the area) and to process their images captured. Among these requests, searching for wounded or missing persons is one of the most critical yet excruciating tasks (applications). Therefore, the HLMCM, which is composed by the Cloudlet and the mobile devices, can run HACAS algorithm to effectively schedule these applications.

6. Conclusion and Future Work

Efficiently exploiting the mobile devices' idle computing, storage, and sensing capacity can greatly improve the quality of service provided by mobile cloud computing (MCC). To achieve this goal, an appropriate architecture of MCC and a dedicated scheduling algorithm are considered important. To address these issues, this paper contributes in several ways by providing suitable definitions of critical aspects and proposing efficient algorithms and approaches.

Our simulation results have revealed that when the load of the system is heavy, HACAS algorithm can select those applications with maximum profit and minimum energy consumption. With the parameters setting in the simulation, the profit of HACAS algorithm is about 30% higher than that of FCFS algorithm. Besides, when the load of the system is light, the provider selection scheme adopted in HACAS can effectively balance the load of the devices in the system. Concretely speaking, HACAS algorithm's load variation is about 60% better than that of the random provider selection scheme. Moreover, in the discussion part of the paper, we have presented the rationale for devising HLMCM and for selecting those simulation parameters. In simple terms, HLMCM can effectively use mobile devices' diverse sensing results which cannot be realized by extending the cloud infrastructure or the Cloudlet entity's service capability. Moreover, the simulation parameters are chosen based on the observation of mainstream smartphones. The discussion section also gives an application scenario where HACAS algorithm is used for disaster rescue.

In the future, we will further extend the scheduling algorithm by considering the dynamic resource requirement of the applications.

Acknowledgments

This research was supported in part by the Major State Basic Research Development Program of China (973 Program) no. 2012CB315806, National Natural Science Foundation of China under Grant no. 61070173, National Natural Science Foundation of China under Grant no. 61201216, Jiangsu

Province Natural Science Foundation of China under Grant no. BK2010133, Jiangsu Province Natural Science Foundation of China under Grant no. BK2009058, and China Postdoctoral Science Foundation funded project under Grant no. 201150M1512.

References

[1] IDC, Worldwide smartphone market expected to grow 55 in 2011 and approach shipments of one billion in 2015, according to IDC, http://www.idc.com/getdoc.jsp?containerId=prUS22871611.

[2] M. Conti, S. Chong, S. Fdida et al., "Research challenges towards the Future Internet," *Computer Communications*, vol. 34, no. 18, pp. 2115–2134, 2011.

[3] D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking (Mobicom '12)*, pp. 173–184, ACM, 2012.

[4] L. Duan, T. Kubo, K. Sugiyama, J. Huang, T. Hasegawa, and J. Walrand, "Incentive mechanisms for smartphone collaboration in data acquisition and distributed computing," in *Proceedings of the Annual IEEE International Conference on Computer Communications (INFOCOM '12)*, pp. 1701–1709, Orlando, Fla, USA, March 2012.

[5] Y.-K. Kwok, K. Hwang, and S. Song, "Selfish grids: game-theoretic modeling and NAS/PSA benchmark evaluation," *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no. 5, pp. 621–636, 2007.

[6] R. Subrata, A. Y. Zomaya, and B. Landfeldt, "A cooperative game framework for QoS guided job allocation schemes in grids," *IEEE Transactions on Computers*, vol. 57, no. 10, pp. 1413–1422, 2008.

[7] P. Ghosh, K. Basu, and S. K. Das, "A game theory-based pricing strategy to support single/multiclass job allocation schemes for bandwidth-constrained distributed computing systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no. 3, pp. 289–306, 2007.

[8] G. Danezis, S. Lewis, and R. Anderson, "How much is location privacy worth?" in *Proceedings of Workshop on the Economics of Information Security Series (WEIS '05)*, 2005.

[9] J.-S. Lee and B. Hoh, "Sell your experiences: a market mechanism based incentive for participatory sensing," in *Proceedings of the 8th IEEE International Conference on Pervasive Computing and Communications (PerCom '10)*, pp. 60–68, April 2010.

[10] J. Liu, X. Luo, X. Zhang, F. Zhang, and B. Li, "Job scheduling model for cloud computing based on multi-objective genetic algorithm," *IJCSI International Journal of Computer Science Issues*, vol. 10, no. 1, 2013.

[11] E. Feller, L. Rilling, and C. Morin, "Energy-aware ant colony based workload placement in clouds," in *Proceedings of the 12th IEEE/ACM International Conference on Grid Computing (Grid '11)*, pp. 26–33, September 2011.

[12] H. Goudarzi and M. Pedram, "Multi-dimensional SLA-based resource allocation for multi-tier cloud computing systems," in *Proceedings of the IEEE 4th International Conference on Cloud Computing (CLOUD '11)*, pp. 324–331, July 2011.

[13] S. Nagendram, J. Vijaya Lakshmi, and D. Venkata Narasimha Rao, "Efficient resource scheduling in data centers using MRIS," *Indian Journal of Computer Science and Engineering*, vol. 2, no. 5, pp. 764–769, 2011.

[14] S. Tayal, "Task scheduling optimization for the cloud computing systems," *International Journal of Adcanced Engineering Sciences and Technologies*, vol. 5, no. 2, pp. 111–115, 2011.

[15] O. Morariu, C. Morariu, and T. Borangiu, "A genetic algorithm for workload scheduling in cloud based e-Learning," in *Proceedings of the 2nd International Workshop on Cloud Computing Platforms (CloudCP '12)*, ACM, April 2012.

[16] J. Gu, J. Hu, T. Zhao, and G. Sun, "A new resource scheduling strategy based on genetic algorithm in cloud computing environment," *Journal of Computers*, vol. 7, no. 1, pp. 42–52, 2012.

[17] H. Yamauchi, K. Kurihara, T. Otomo, Y. Teranishi, T. Suzuki, and K. Yamashita, "Effective distributed parallel scheduling methodology for mobile cloud computing," in *Proceedings of the 17th Workshop on Synthesis and System Integration of Mixed Information Technologies (SASIMI '12)*, pp. 516–521, 2012.

[18] R. Mishra and A. Jaiswa, "Ant colony optimization: a solution of load balancing in cloud," *International Journal of Web & Semantic Technology*, vol. 3, no. 2, 2012.

[19] L. Zhu, Q. Li, and L. He, "Study on cloud computing resource scheduling strategy based on the ant colony optimization algorithm," *IJCSI International Journal of Computer Science*, vol. 9, no. 5, 2012.

[20] K. Nishant, P. Sharma, V. Krishna et al., "Load balancing of nodes in cloud using ant colony optimization," in *Proceedings of the 14th International Conference on Computer Modelling and Simulation (UKSim '12)*, pp. 3–8, March 2012.

[21] S. Suryadevera, J. Chourasia, S. Rathore, and A. Jhummarwala, "Load balancing in computational grids using ant colony optimization algorithm," *International Journal of Computer & Communication Technology*, vol. 3, no. 3, 2012.

[22] N. J. Kansal and I. Chana, "Cloud load balancing techniques: a step towards green computing," *IJCSI International Journal of Computer Science*, vol. 9, no. 1, 2012.

[23] http://www.mobilecloudcomputingforum.com/.

[24] White Paper, Mobile Cloud Computing Solution Brief, AEPONA, November 2010.

[25] J. H. Christensen, "Using RESTful web-services and cloud computing to create next generation mobile applications," in *Proceedings of the 24th Annual ACM Conference on Object-Oriented Programming, Systems, Languages and Applications (OOPSLA '09)*, pp. 627–633, October 2009.

[26] L. Liu, R. Moulic, and D. Shea, "Cloud service portal for mobile device management," in *IEEE International Conference on E-Business Engineering (ICEBE '10)*, pp. 474–478, January 2011.

[27] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," *Wireless Communications and Mobile Computing*, 2012.

[28] E. Cuervoy, A. Balasubramanian, D.-K. Cho et al., "MAUI: making smartphones last longer with code offload," in *Proceedings of the 8th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '10)*, pp. 49–62, June 2010.

[29] B.-G. Chun and P. Maniatis, "Augmented smartphone applications through clone cloud execution," in *Proceedings of the 12th Workshop on Hot Topics in Operating Systems (HotOS XII '09)*, Monte Verita, Switzerland, 2009.

[30] M. Satyanarayanan, P. Bahl, R. Cáceres, and N. Davies, "The case for VM-based cloudlets in mobile computing," *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14–23, 2009.

[31] E. E. Marinelli, *Hyrax: cloud computing on mobile devices using MapReduce [M.S. thesis]*, Carnegie Mellon University, 2009.

[32] A. Dou, V. Kalogeraki, D. Gunopulos, T. Mielikainen, and V. H. Tuulos, "Misco: a MapReduce framework for mobile systems," in *Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '10)*, ACM, June 2010.

[33] G. Huerta-Canepa and D. Lee, "A virtual cloud computing provider for mobile devices," in *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond (MCS '10)*, New York, NY, USA, June 2010.

[34] D. Huang, X. Zhang, M. Kang, and J. Luo, "MobiCloud: building secure cloud framework for mobile computing and communication," in *Proceedings of the 5th IEEE International Symposium on Service-Oriented System Engineering (SOSE '10)*, pp. 27–34, June 2010.

[35] T. Xing, D. Huang, S. Ata, and D. Medhi, "MobiCloud: a geo-distributed mobile cloud computing platform," in *Proceedings of the 8th International Conference on Network and Service Management (CNSM '12)*, Las Vegas, Nev, USA, October 2012.

[36] Q. Liu, X. Jian, J. Hu, H. Zhao, and S. Zhang, "An optimized solution for mobile environment using mobile cloud computing," in *Proceedings of the 5th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '09)*, pp. 1–5, September 2009.

[37] D. Fesehaye, Y. Gao, K. Nahrstedt, and G. Wang, "Impact of cloudlets on interactive mobile cloud applications," in *IEEE 16th Internationa Enterprise Distributed Object Computing Conference (EDOC '12)*, pp. 123–132, 2012.

[38] M. Dorigo, G. Di Caro, and L. M. Gambardella, "Ant algorithms for discrete optimization," *Artificial Life*, vol. 5, no. 2, pp. 137–172, 1999.

[39] G. Leguizamon and Z. Michalewicz, "A new version of ant system for subset problems," in *Proceedings of the Congress on Evolutionary Computation (CEC '99)*, vol. 2, p. 1464, 1999.

[40] H. Mehendale, A. Paranjpe, and S. Vempala, "Lifenet: a flexible ad hoc networking solution for transient environments," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4, pp. 446–447, 2011.

[41] Y. Cui, X. Ma, H. Wang, I. Stojmenovic, and J. Liu, "A survey of energy efficientWireless transmission and modeling in mobile cloud computing," *Mobile Networks and Applications*, vol. 18, no. 1, pp. 148–155, 2012.

[42] N. Fernando, S. W. Loke, and W. Rahayu, "Mobile cloud computing: a survey," *Future Generation Computer Systems*, vol. 29, pp. 84–106, 2013.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 517846, 19 pages
http://dx.doi.org/10.1155/2013/517846

Research Article

Minimum-Cost QoS-Constrained Deployment and Routing Policies for Wireless Relay Networks

Frank Yeong-Sung Lin,[1] Chiu-Han Hsiao,[1] Kuo-Chung Chu,[2] and Yi-Heng Liu[2]

[1] *Department of Information Management, National Taiwan University, No. 1 Section 4, Roosevelt Road, Taipei 106, Taiwan*
[2] *Department of Information Management, National Taipei University of Nursing & Health Sciences, No. 365, Ming Te Road, Taipei 112, Taiwan*

Correspondence should be addressed to Chiu-Han Hsiao; chiuhanhsiao@gmail.com

Received 13 April 2013; Accepted 7 July 2013

Academic Editor: Chih-Hao Lin

With the continued evolution of wireless communication technology, relaying is one of the features proposed for the 4G LTE Advanced (LTE-A) system. The aim of relaying is to enhance both coverage and capacity. The idea of relays is not new, but relaying is being considered to ensure that the optimum performance is achieved to enable the expectations or good quality of service (QoS) of the users to be met while still keeping capital expenditure (CAPEX) within the budgeted bounds of operators. In this paper, we try to stand for an operator to propose a solution that determines where and how many relays should be deployed in the planning stages to minimize the development cost. In the planning stages, we not only derive a multicast tree routing algorithm to both determine and fulfill the QoS requirements to enhance throughput, but we also utilize the Lagrangian relaxation (LR) method in conjunction with optimization-based heuristics and conduct computational experiments to evaluate the performance. Our contribution is utilizing the LR method to propose an optimal solution to minimize the CAPEX of operators to build up a relay network with more efficiency and effectiveness and the QoS can be guaranteed by service level agreement.

1. Introduction

Providing a guaranteed service and good performance with budgets constraint is always an optimization problem of operators and vendors. During the last decade, this problem has however become much more difficult, because the traffic has grown significantly and demand for broadband data services is expected to increase tremendously [1]. The business challenges of operators would be that increasing revenues will have to come from nonvoice services which means they have to increase total communication market shares with extending service coverage and offering good service by capacity expansion as well as increased bandwidth and improved quality of services (QoSs) [2]. But building out of the macronetworks significantly will require huge investments, especially where access would otherwise be limited or unavailable without the need for expensive cellular towers. So, the operators may plan to compensate with new revenues and cost reduction at the same time. The aim of this paper is related to the relays should be deployed strategies to investigate how the relays are suitable for providing good services by minimize operator capital expenditure (CAPEX) significantly [3].

In the ongoing standardization technology development by third-Generation Partnership Project (3GPP), relaying is one of the features proposed for the LTE Advanced (LTE-A) system [4–8]. The aim of LTE-A relaying is to enhance both coverage and capacity. However, this idea of relays is not new, but the LTE-A relaying is being considered to ensure that the optimum performance is achieved to enable the expectations of the users to be met while still keeping CAPEX within the budgeted bounds. As cell edge performance is becoming more critical, with some of the technologies being pushed towards their limits, it is necessary to look at solutions that

will enhance performance at the cell edge for a comparatively low cost. One solution that is being investigated and proposed is that of the use of relays [9–11].

In order for the cellular telecommunications technology to be able to keep pace with technologies that may compete, it is necessary to ensure that new cellular technologies are being formulated and developed. But there are many realistic conditions influencing operators including the tough economic environment, declining budgets, limited resources, time pressures, and high user expectations.

This paper proposes a solution approach for relay network planning of where to build relays, and how to configure each relay, how the routing algorithm of relays and mobile stations is worked properly. This research can be divided into two parts. First, we constructed the relay network architecture with multicast tree routing concepts. Secondly, we proposed a precise mathematical expression to model this network and developed algorithms based on Lagrangian Relaxation Method to solve this problem. These model approaches might nevertheless be regarded as useful engineering guidelines for operators to build up a good network to extend services and reduce CAPEX efficiently and effectively.

This paper is organized as follows. Section 2 is Literature Survey, and then we introduce a mathematical formulation for the wireless relay networks design problem in Section 3. In Section 4, we present the solution approach by using Lagrangian Relaxation, in which heuristics for calculating a good primal feasible solution are developed, and conduct computational experiments. In Section 5, we conclude and discuss the direction of future research.

2. Literature Survey

Multihop Wireless Networking has been widely studied and implemented throughout ad hoc networks and mesh networks to exploit the user diversity concept and improve overall performance. The original concept of general relaying problems was defined in [12, 13] and was inspired by the development of the ALOHA system at the University of Hawaii. Based on this concept, a relay network can be designed as a tree-based topology with one end of the path being the base station (BS) relaying multiple connections to provide services and improve the coverage.

Relay stations (RSs) have some characteristics or cost efficiency for the following reasons.

(1) The transmission range is much less than a BS, meaning that the transmit power is also less than that of a BS. Relays are generally cheaper than BS, meaning reduced costs without site survey and easy to construct relays in the place which is not suitable to build a base station tower.

(2) Relays do not have a wired connection to the backhaul. Instead, they receive signals from the BS and retransmit to destination users wirelessly and vice versa. The leases of wired broadband backhauls can be saved.

(3) Relaying techniques have the dual advantages of performance improvement and coverage extension at the cell edge. These could feasibly be a deployment solution for the high-frequency band in which propagation is significantly more vulnerable to nonline-of-sight (NLOS) conditions to overcome shadowing [14].

Coordinated multipoint (CoMP) is a relatively new class of spatial diversity techniques that are enabled by relaying [15, 16] and cooperative communications which is shown in Figure 1 [17, 18]. This concept has been the focus of many studies by 3GPP for LTE-A as well as the IEEE for the WiMAX, 802.16 standards. But still no conclusion about CoMP has been reached regarding the full implementation, because CoMP has not been included in Rel.10 of the 3GPP standards. As the work is ongoing, CoMP is likely to reach a greater level of consensus; when this occurs, it will be included in future releases of the standards.

CoMP is a complex set of techniques which are distributed radios that jointly transmit information in wireless environments. The main purpose may be improved for the reliability of communications in terms of coverage extension, reduced outage probability, symbol-error, or bit-error probability for a given transmission rate [19–21]. It brings many advantages to the user as well as the network operator as follows.

(1) It makes better utilization of network: by providing connections to several BSs or RSs at once, using CoMP, data can be passed through least loaded BS or RS for better resource utilization.

(2) It provides enhanced reception performance: using several sources cooperative BSs or RSs for each connection means that overall reception will be improved and the number of dropped calls should be reduced.

(3) Multiple site reception increases received power: the joint reception from multiple BSs or RSs using CoMP techniques enables the overall received power at the handset to be increased.

When building a relay wireless network in a metropolis, various factors influence the design such as QoS requirements, throughput requirements, and total cost. The objective of our research is "to minimize the total building costs subject to QoS and throughput requirements." Nonetheless, this objective is obviously a tradeoff because total building costs will increase if the QoS and throughput requirements increase. Based on this conventional tradeoff, we take multipath routing algorithms into consideration to solve the critical problem [22, 23].

The purpose of this research is different from that of conventional network design problems. The assumptions are that multiple source nodes jointly transmit one single source of information if the signal strength is not robust enough in the link between one source node to the destination. The routing policy is no longer a single path but a more complex and interesting multipath algorithms.

3. Problem Formulation

3.1. Problem Description. A sequence of the wireless relay network design may be described as fellows. First, the location

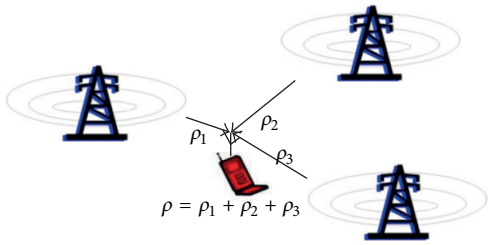


Figure 1: Coordinated multipoint.

of each BS could be determined by site survey and how many BSs can cover the service area. Second, the set of BSs roughly divide the entire network into several subnetworks, each of which is rooted by one BS connected to the core network which is shown in Figure 2. Meanwhile, there are many candidate locations suitable for the deployment of relays. The decision of deploying a relay at a specific location depends on the users surrounding the location; once a location is selected, the relay must associate with one of the BSs mentioned previously. So that total costs will depend on where relays are developed and how many relays should be developed of the network. But there is another important factor that should be considered by how to provide a good QoS of users at the stage of the network design. This problem may be solved by designing a mechanism or routing algorithm through which users can reach the suitable relays or BSs which can serve users well. In our work, we introduce the multicast tree-based routing algorithm (MTBR) to apply the multipath concept, as represented later.

Figure 2 illustrates the entire network design: the triangles represent the BSs; the cell phones represent the mobile stations (MSs); the circles with solid line represent the relay stations being built on the selected locations; and the circles with dotted line represent the locations not selected to build RSs. The whole area is divided into several subnetworks and rooted at associated BSs. If a subnetwork is concerned that each OD pair, like the BS-to-MS, can be expressed in Figure 3, transmits through the routing multicast tree to the associated BS. In DL transmission, data is multicasted from BS to the RSs selected by the MS, and cooperatively relayed to the destination MS to achieve the spatial diversity gains through CoMP techniques. The same routing multicast tree in UL is represented in Figure 4, in which the aggregation of traffic from the MS can overcome the weak signal strength when the MS is far from the BS or RSs. Because the channels, bandwidth, and even transmission power are different between DL and UL, the DL tree and UL tree of an MS may be different. In this paper, we derive a near optimal RSs development policy to minimize total development costs; we also maintain both DL and UL spanning trees and use the MTBR to ensure that BER and data rate requirements of each MS are satisfied.

3.2. Mathematical Modeling

Assumption

(i) The relaying protocol in this model is *Decode-and-Forward*.

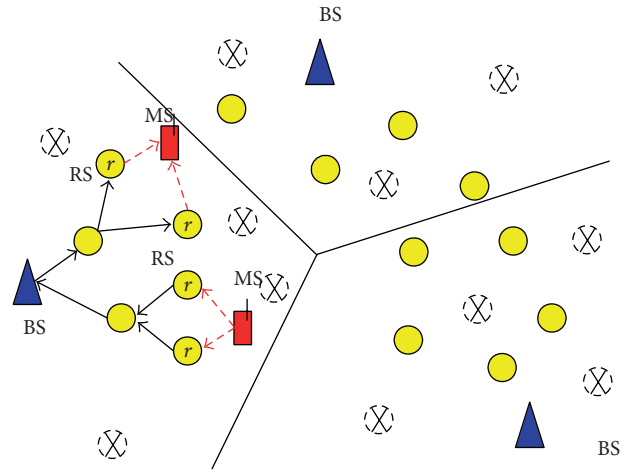


Figure 2: Network separations with several BSs.

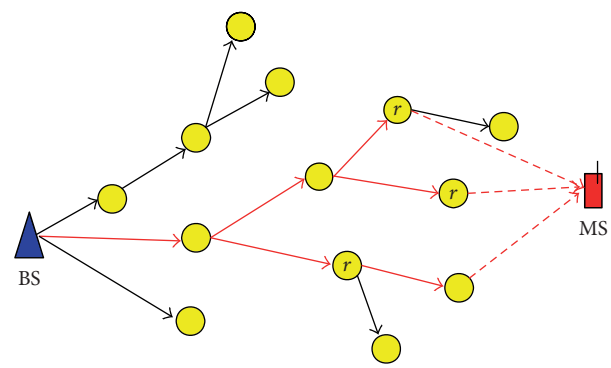


Figure 3: One OD pair routing multicast tree in DL transmission.

(ii) Once a location is selected to build an RS, it must home to one BS.

(iii) Each MS must home to either a BS or RS(s).

(iv) The RSs selected by an MS must associate with the same BS.

(v) The routing path of each OD pair in DL (UL) is a multicast tree.

(vi) The capacity of a link uv is decided by adaptive modulation with respect to the signal-to-noise ratio (SNR) received at node v.

(vii) The spatial diversity gains are represented by the aggregate SNR with CoMP techniques.

(viii) The bit error rate (BER) of a transmission is measured by the receiving SNR value

(ix) The aggregate BER of the destination is the summation of BER of each node on the routing multicast tree.

(x) The numbers of links of each path adopted by each MS are assumed to be equal to ensure that the CoMP can be achieved within limited delay.

(xi) Error corrections and retransmissions are not considered in this problem.

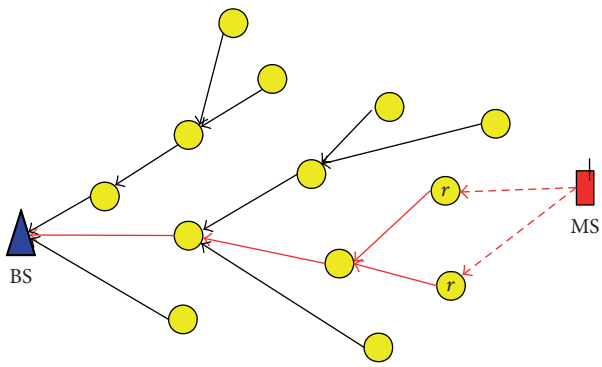


Figure 4: One OD pair routing multicast tree in UL transmission.

Given

(i) The set of BSs, candidate locations and configurations of RSs, MSs,

(ii) required data rate of an MS in DL and UL,

(iii) fixed and configured cost of an RS,

(iv) the set of all spanning trees, paths,

(v) distance of each link,

(vi) attenuation factor,

(vii) thermal noise function,

(viii) transmit power of BS, RS, and MS,

(ix) sNR function,

(x) the minimum SNR requirement for an MS in DL and UL to home to a BS or an RS,

(xi) the maximum BER threshold of a OD pair transmission in DL and UL,

(xii) nodal and link capacity functions,

(xiii) the maximum spatial diversity of an MS in DL and UL.

Objective. To minimize the total cost of wireless relay network deployment.

Subject to

(i) RS selection constraints,

(ii) nodal capacity constraints,

(iii) cooperative relaying constraints in DL and UL,

(iv) routing constraints in DL and UL,

(v) link capacity constraints in DL and UL.

To Determine

(i) Whether or not a location should be selected to build an RS,

(ii) the cooperative RSs of each MS,

(iii) the routing paths of an OD pair (a BS to an MS or contrary), which form a multicast tree from the BS to the cooperative RSs selected by each MS.

Objective Function

$$\min \sum_{r\in R}\sum_{k\in K}\left(\psi_r + \Phi_r(k)\right)\eta_{rk} \tag{IP 1}$$

Subject to

Relay Selection Constraints

$$\sum_{k\in K}\eta_{rk} \le 1 \quad \forall r \in R, \tag{1}$$

$$\sum_{b\in B} h_{rb}^{\mathrm{dir}} \le 1 \quad \forall r \in R,\ \mathrm{dir} \in \mathrm{DIR}, \tag{2}$$

$$\sum_{b\in B} h_{rb}^{\mathrm{dir}} \le \sum_{k\in K}\eta_{rk} \quad \forall r \in R,\ \mathrm{dir} \in \mathrm{DIR}. \tag{3}$$

Nodal Capacity Constraints

$$\begin{aligned}\sum_{u\in\{R\cup B\}}\sum_{n\in N} y_{nur}^{1}\theta_n^1 + \sum_{w\cup\{R\cup B\}}\sum_{n\in N} y_{nrw}^{2}\theta_n^2 \\ \le \sum_{k\in K}\eta_{rk}\overline{C_r(k)} \quad \forall r \in R,\end{aligned} \tag{4}$$

$$\begin{aligned}\sum_{r\in R}\sum_{n\in N} y_{nbr}^{1}\theta_n^1 + \sum_{i\in R}\sum_{n\in N} y_{nib}^{2}\theta_n^2 \\ + \sum_{n\in N}\sum_{\mathrm{dir}\in\mathrm{DIR}} \kappa_{nb}^{\mathrm{dir}}\theta_n^{\mathrm{dir}} \le \overline{C_b} \quad \forall b \in B.\end{aligned} \tag{5}$$

Cooperative Relay Constraints

$$\begin{aligned}\kappa_{nr}^{\mathrm{dir}} \le \sum_{k\in K}\eta_{rk} \quad \forall n \in N, \\ r \in R, \quad \mathrm{dir} \in \mathrm{DIR},\end{aligned} \tag{6}$$

$$\begin{aligned}\sum_{b\in B}\kappa_{nb}^{\mathrm{dir}} + \kappa_{nr}^{\mathrm{dir}} \le 1 \quad \forall n \in N, \\ r \in R, \quad \mathrm{dir} \in \mathrm{DIR},\end{aligned} \tag{7}$$

$$1 \le \sum_{s\in\{R\cup B\}}\kappa_{ns}^{\mathrm{dir}} \quad \forall n \in N,\ \mathrm{dir} \in \mathrm{DIR}, \tag{8}$$

$$\sum_{r\in R}\kappa_{nr}^{\mathrm{dir}} \le \mathrm{SD}^{\mathrm{dir}} \quad \forall n \in N,\ \mathrm{dir} \in \mathrm{DIR}, \tag{9}$$

$$\kappa_{ns}^{1}P_{\min}^{N} \le \kappa_{ns}^{1}\pi_{sn}^{1} \quad \forall n \in N,\ s \in \{R\cup B\}, \tag{10}$$

$$\kappa_{ns}^{2}P_{\min}^{R} \le \kappa_{ns}^{2}\pi_{ns}^{2} \quad \forall n \in N,\ s \in \{R\cup B\}, \tag{11}$$

$$\omega_n \le \sum_{s\in\{R\cup B\}}\kappa_{ns}^{1}\pi_{sn}^{1} \quad \forall n \in N, \tag{12}$$

$$\begin{aligned}0 \le \pi_{sn}^{\mathrm{dir}} \le \overline{\pi_n} \quad \forall s \in \{R\cup B\}, \\ n \in N, \quad \mathrm{dir} \in \mathrm{DIR},\end{aligned} \tag{13}$$

$$\kappa_{ns}^{1}\pi_{sn}^{1} \le \sum_{k\in K}\eta_{sk}\left[P\left(\rho_{s}^{1}(k), D_{sn}, \tau\right) - P_{N}(n)\right] \quad \forall s \in \{R\cup B\}, \quad n \in N, \tag{14}$$

$$\kappa_{ns}^{2}\pi_{ns}^{2} \le \sum_{k\in K}\eta_{sk}\left[P\left(\rho_{n}^{N}, D_{ns}, \tau\right) - P_{N}(s)\right] \quad \forall s \in \{R\cup B\}, \quad n \in N, \tag{15}$$

$$y_{nuv}^{\mathrm{dir}}P_{\min}^{R} \le y_{nuv}^{\mathrm{dir}}\phi_{uv}^{\mathrm{dir}} \quad \forall n \in N, \quad u, v \in \{R\cup B\}, \quad \mathrm{dir} \in \mathrm{DIR}, \tag{16}$$

$$\varepsilon_{nv} \le \sum_{u\in\{R\cup B\}} y_{nuv}^{2}\phi_{uv}^{2} \quad \forall n \in N, \ v \in \{R\cup B\}, \tag{17}$$

$$0 \le \phi_{uv}^{\mathrm{dir}} \le \overline{\phi_{v}} \quad \forall u, v \in \{R\cup B\}, \ \mathrm{dir} \in \mathrm{DIR}, \tag{18}$$

$$y_{nuv}^{\mathrm{dir}}\phi_{uv}^{\mathrm{dir}} \le \sum_{k\in K}\eta_{uk}\left[P\left(\rho_{u}^{\mathrm{dir}}(k), D_{uv}, \tau\right) - P_{N}(v)\right] \quad \forall n \in N, \quad u, v \in \{R\cup B\}, \quad \mathrm{dir} \in \mathrm{DIR}, \tag{19}$$

$$\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}} y_{nuv}^{1}\mathrm{BER}\left(\phi_{uv}^{1}\right) + \mathrm{BER}\left(\omega_{n}\right) \le \mathrm{BER}^{1} \quad \forall n \in N, \tag{20}$$

$$\sum_{v\in\{R\cup B\}}\mathrm{BER}\left(\varepsilon_{nv}\right) + \sum_{s\in\{R\cup B\}}\mathrm{BER}\left(\kappa_{ns}^{2}\pi_{ns}^{2}\right) \le \mathrm{BER}^{2} \quad \forall n \in N. \tag{21}$$

Routing Constraints

$$\kappa_{nr}^{\mathrm{dir}} \le h_{rb}^{\mathrm{dir}} \quad \forall n \in N, \quad r \in R, \quad b \in B, \quad \mathrm{dir} \in \mathrm{DIR}, \tag{22}$$

$$h_{rb}^{\mathrm{dir}} \le \sum_{p\in P_{br}} x_{nrp}^{\mathrm{dir}} \quad \forall n \in N, \quad b \in B, \quad \mathrm{dir} \in \mathrm{DIR}, \tag{23}$$

$$\sum_{b\in B}\sum_{p\in P_{br}} x_{nrp}^{\mathrm{dir}} \le 1 \quad \forall n \in N, \quad r \in R, \quad \mathrm{dir} \in \mathrm{DIR}, \tag{24}$$

$$\sum_{p\in P_{bi}}\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}} x_{nip}^{\mathrm{dir}}\delta_{puv} \le \sum_{p\in P_{bj}}\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}} x_{njp}^{\mathrm{dir}}\delta_{puv} + \left(1-\kappa_{nj}^{\mathrm{dir}}\right)\overline{M} \quad \forall n \in N, \quad i, j \in R, \quad b \in B, \quad \mathrm{dir} \in \mathrm{DIR}, \tag{25}$$

$$y_{nuv}^{1} \le \sum_{k\in K}\eta_{vk} \quad \forall n \in N, \quad u \in \{R\cup B\}, \quad v \in R, \tag{26}$$

$$y_{nuv}^{2} \le \sum_{k\in K}\eta_{uk} \quad \forall n \in N, \quad u \in R, \quad v \in \{R\cup B\}, \tag{27}$$

$$\sum_{b\in B}\sum_{p\in P_{br}} x_{nrp}^{\mathrm{dir}}\delta_{puv} \le y_{nuv}^{\mathrm{dir}} \quad \forall n \in N, \quad r \in R, \quad u, v \in \{R\cup B\}, \quad \mathrm{dir} \in \mathrm{DIR}. \tag{28}$$

Link Capacity Constraints

$$\sum_{n\in N} y_{nuv}^{\mathrm{dir}}\theta_{n}^{\mathrm{dir}} \le C_{uv}\left(\phi_{uv}^{\mathrm{dir}}\right), \quad \forall u, v \in \{R\cup B\}, \quad \mathrm{dir} \in \mathrm{DIR} \tag{29}$$

Integer Constraints

$$\begin{aligned} &\eta_{rk} = 0 \text{ or } 1 \quad \forall r \in R, \ k \in K, \\ &h_{rb}^{\mathrm{dir}} = 0 \text{ or } 1 \quad \forall r \in R, \quad b \in B, \quad \mathrm{dir} \in \mathrm{DIR}, \\ &\kappa_{ns}^{\mathrm{dir}} = 0 \text{ or } 1 \quad \forall n \in N, \quad s \in \{R\cup B\}, \quad \mathrm{dir} \in \mathrm{DIR}, \\ &x_{nrp}^{\mathrm{dir}} = 0 \text{ or } 1 \quad \forall n \in N, \ r \in R, \quad p \in P_{br}, \quad b \in B, \quad \mathrm{dir} \in \mathrm{DIR}, \\ &y_{nuv}^{\mathrm{dir}} = 0 \text{ or } 1 \quad \forall n \in N, \quad u, v \in \{R\cup B\}, \quad \mathrm{dir} \in \mathrm{DIR}. \end{aligned} \tag{30}$$

Explanation of Objective Function. The objective function (IP) is to minimize the total cost of RSs deployment: (1) Fix costs of RS such as land acquisition and hardware purchases; (2) The configured costs of each RS. The detail parameters in the formulation are noted and shown in Tables 1 and 2.

Explanation of Constraint

(1) Relay Assignment Constraints. Constraint (1) requires that each location is selected to install an RS with exactly only one configuration or none.

Constraint (2) requires that each RS can associate with one BS or none in direction dir.

Constraint (3) indicates that once an RS r associates with a BS, r must be built.

(2) Nodal Capacity Constraints. Constraint (4) requires that each RS's total amount of traffic in DL and UL cannot be greater than its nodal capacity.

Constraint (5) requires that each BS's total amount of traffic in DL and UL cannot be greater than its nodal capacity.

(3) Cooperative Relay Constraints. Each MC will select an RS r in direction dir only if r is installed in (6).

An MC must select either one BS or RS(s) in direction dir in (7).

Table 1: Notations of given parameters.

Given parameters	
Notation	Description
General	
DIR	The set of transmission direction, where dir $\in$ DIR, DIR = {1 (downlink), 2 (uplink)}
B	The set of BSs, where $b \in B$
R	The set of RS candidate locations, where $r \in R$
K	The set of RS configurations, where $k \in K$
N	The set of MCs, where $n \in N$
θ_n^{dir}	The data rate required to be transmitted of MC n in direction dir in (packets/sec)
ψ_r	The fix cost of building an RS on location r
$\Phi_r(k)$	The configured cost of building RS r, which is a function of configuration k
$\overline{M}$	An arbitrarily large number
Routing	
P_{br}	The set of paths from BS b to RS r, where $p \in P_{br}$
δ_{puv}	The indicator function which is 1 if link uv is on path p and 0 otherwise
SNR and attenuation	
D_{uv}	The distance of link uv
τ	Attenuation factor
$\rho_r^{\text{dir}}(k)$	Transmit power of RS r in direction dir, which is a function of configuration k
$P_N(s)$	Thermal noise strength function in dBm/Hz, where $s \in \{N, R, B\}$ represents receiving node type.
ρ_b^B	Transmit power of BS b
ρ_n^N	Transmit power of MC n
$P\left(\rho_u^{\text{dir}}(k), D_{uv}, \tau\right)$	Signal strength received by node v in dBm, which is a function of $\rho_r^{\text{dir}}(k)$, D_{uv} and τ
$P_{\min}^N$	The minimum SNR requirement for a MC to receive from a RS in DL
$P_{\min}^R$	The minimum SNR requirement for a RS to receive from a MC in UL
$\overline{\phi}_v$	The maximum SNR can be received by node v in link uv, where $u, v \in \{R \cup B\}$
$\overline{\pi}_v$	The maximum SNR can be received by node v in link uv, where $u \in \{R \cup B\}, v \in N$ in DL; and $u \in N, v \in \{R \cup B\}$ in UL
BER	
BER^{dir}	The BER requirement for the transmission received by a destination in direction dir where the destination in DL is MC and in UL is BS
$\text{BER}(\text{SNR}_s)$	The BER value of each node s, which is a function of the receiving SNR, where $s \in \{R \cup B \cup N\}$

Table 1: Continued.

Given parameters	
Notation	Description
Capacity	
$\overline{C_b}$	The nodal capacity of BS b in (packets/sec)
$\overline{C_r(k)}$	The nodal capacity of RS r in (packets/sec), which is a function of configuration k
$C_{uv}(\text{SNR})$	The capacity of link uv in (packets/sec), which is a function of the receiving SNR of node v, where $u, v \in \{R \cup B\}$
Relaying	
SD^{dir}	The maximum spatial diversity of a MC in direction dir

Table 2: Notations of decision variables.

Notation	Description
Decision variables	
η_{rk}	1 if candidate location r is selected to build a RS with configuration k and 0 otherwise
h_{rb}^{dir}	1 if RS r associates with BS b in direction dir and 0 otherwise
κ_{ns}^{dir}	1 if node s is selected to cooperatively relay the transmission of MC n in direction dir and 0 otherwise, where $s \in \{R \cup B\}$
y_{nuv}^{dir}	1 if link uv is on the multicast tree adopted by MC n in direction dir and 0 otherwise
x_{nrp}^{dir}	1 if path p is selected for MC n to cooperative RS r in direction dir and 0 otherwise, where $p \in P_{br}$
Auxiliary variables	
ϕ_{uv}^{dir}	The SNR received by node v in link uv, where $u, v \in \{R \cup B\}$
π_{uv}^{dir}	The SNR received by node v in link uv, where $u \in \{R \cup B\}, v \in N$ in DL; and $u \in N, v \in \{R \cup B\}$ in UL
ω_n	The summation of SNR received by MC n in DL
ε_{ns}	The summation of SNR received by node s in UL oriented by MC n, where $s \in \{R \cup B\}$

Constraints (8) and (9) represent the boundaries of the number of cooperative RSs an MC can select.

The minimum SNR constraints for an MC to receive from a BS or an RS in DL, and for an MC to transmit to a BS or an RS in UL, are expressed in (10) and (11), respectively.

Constraint (12) requires that the SNR value received by an MC n in DL cannot exceed the summation of the SNR values n receives from the cooperative RSs selected by n.

Constraint (13) represents the boundaries of decision variable π_{uv}^{dir}.

Once MC n selects RS (or BS) s to be its cooperative RS, the SNR value on link ns cannot exceed the SNR transmitted from source node to destination node in DL and UL in constraints (14) and (15).

The minimum SNR constraint for a link uv selected by MC n is expressed in (16), while $u, v \in \{R \cup B\}$.

Constraint (17) requires that the SNR value received by an RS (or BS) v in UL cannot exceed the summation of the SNR values on the link uv selected by MC n.

Constraint (18) represents the boundaries of decision variable ϕ_{uv}^{dir}.

Once MC n selects a link uv in direction dir, the SNR value on uv cannot exceed the SNR transmitted from u to v in (19).

The aggregative BERs constraints for DL in MC and UL in BS are expressed in (20) and (21), respectively.

(4) Routing Constraints. Constraint (22) requires that once RS r is selected by MC n, r must associate with one BS in direction dir.

Constraint (23) requires that once RS r associates with BS b in direction dir, the paths from b to r must be selected by one or more than one MC.

Constraint (24) requires that there is exactly one path to be selected by an MC from the associated BS to RS r only if the MC selects RS r in direction dir.

There are two constructions in (24): first, every two RSs selected by an MC must associate with the same BS; second, the numbers of links of every two paths selected by an MC must be the same.

For each MC, every receiving RS v on a link uv in DL is installed in (26) and every transmitting RS u on a link uv in UL is installed in (27).

Constraint (28) requires that, if link uv is on the path p adopted by the MC n to reach RS r in direction dir, then y_{nuv}^{dir} must be 1.

(5) Link Capacity Constraints. The aggregate flow of link uv in direction dir is restricted in (29).

(6) Integer Constraint. Constraints (30) are integer properties of the decision variables.

4. Solution Approach and Computational Experiments

4.1. Lagrangian Relaxation Techniques. By applying the Lagrangian Relaxation (LR) Method and the Subgradient Method to solve the complex problem, based on the problem formulation mentioned previously, the first step would be that the constraints of the primal problem are relaxed by using the LR Method [26]. In this step, we can not only determine a theoretical lower bound of the primal problem, but, can also glean some hints of feasible solutions captured by the primal problems. After iterations, the end result of the Lagrangian Relaxation Problem is guaranteed to a feasible solution by a feasible step which is satisfied with all constraints of the primal problem, if not, we have to make some modifications.

4.2. Getting Primal Feasible Heuristics. To obtain the primal feasible solutions for (IP 1), the first step is considered the solutions to the Lagrangian Relaxation. Two major decision variables, κ_{ns}^{dir} and y_{nuv}^{dir} are taken into consideration. According to κ_{ns}^{dir}, the RS(s) (or BS) can be obtained to serve MS n selected in dir direction, and y_{nuv}^{dir} represents the link uv which n selected on the routing multicast tree in dir direction. In addition to κ_{ns}^{dir} and y_{nuv}^{dir}, for the complexity of this problem including five 0-1 integer decision variables, we still need other clues to help solving this problem in good quality. Thus, the coefficient $\mu_{ns1}^{4} + \sum_{b\in B}\mu_{nsb1}^{15} + P_{\min}^{N}\mu_{ns}^{5} + \overline{M}\sum_{i\in R}\sum_{b\in B}\mu_{nisb1}^{16} - \pi_{sn}^{1}(\mu_{ns}^{5} + \mu_{n}^{7} - \mu_{ns}^{8})$ of κ_{ns}^{1} in DL, namely, C_{κ}^{1}; $\mu_{nr2}^{4} + \sum_{b\in B}\mu_{nrb2}^{15} + P_{\min}^{R}\mu_{nr}^{6} + \overline{M}\sum_{i\in R}\sum_{b\in B}\mu_{nirb2}^{16} - \pi_{nr}^{2}(\mu_{nr}^{6} - \mu_{nr}^{9}) + \mu_{n}^{14}\text{BEP}(\pi_{nr}^{2})$ of κ_{ns}^{2} in UL, namely, C_{κ}^{2} is introduced in our solution to sort κ_{ns}^{dir} for further calculations.

The main purpose of determining the primal feasible heuristic is, in both DL and UL directions, and for each MS sorted by the distance to BS, to fully utilize the RSs built already to meet the BER requirement, and if not, to at least minimize the number of RSs necessary to reach the previous goal. The detailed procedure that decomposites the Lagrangian Relaxation Problem into several subproblems is described in the appendix.

4.3. Experiments Environment. In this session, we conduct several computational experiments to justify the proposed algorithms. Due to limitation of available experiment scenarios and parameters, we focus on IEEE 802.16j instead of LTE-A; it is easier to build the network based on realistic and operable environment parameters. In order to effectively analyze the physical operations of an 802.16j network, Table 3 lists all system parameters utilized in this research with reference to "Mobile WiMAX" published by WiMAX forum. Adaptive Modulation and Coding (AMS) applied in 802.16j is illustrated specifically in Table 4 with the same reference to "Mobile WiMAX."

In the meantime, and for the purpose of evaluating our solution of quality, two simple algorithms, minimum BER algorithm (MBA) and density-based algorithm (DBA), are implemented for comparison. The purpose of each MBA is, for each MS n, always to find the best paths that can generate the smallest BER value n receives in DL and BS b receives in UL. This algorithm will provide every transmission the minimum BER. The other one is DBA, the main concept would be the building of an RS with the first priority of the highest density area which is not served at the edges of coverage.

Path Loss Function [27]

$$\begin{aligned}\overline{PL(d)}\,(\text{dB}) &= 32.45 + 10\times n\log f_c\ (\text{MHz})\\ &\quad + 10\times n\log d\ (\text{km}),\end{aligned}\tag{31}$$

where n is attenuation factor, f_c is operation frequency, d is distance.

Thermal Noise Function

$$\begin{aligned}N &= KT_0BF, \quad \text{transfer into (dB):}\\ N &= -174\ (\text{dBm}) + 10\log_{10}B + F\ (\text{dB}),\end{aligned}\tag{32}$$

where B is channel bandwidth, F is noise figure.

Table 3: System parameters [24, 25].

Parameters	Value
Operation frequency	2500 MHz
Channel bandwidth	10 MHz
BS antenna gain	15 dBi
RS basic antenna gain	5 dBi
MS antenna gain	−1 dBi
BS noise figure	4 dB
RS noise figure	5 dB
MS noise figure	7 dB
BS transmit power	43 dBm
RS basic transmit power	33 dBm
MS transmit power	23 dBm
RS config, set	3
Attenuation factor	3.2
Thermal noise figure	−174 dB
Min. RS to RS SNR	7.9515 dB
Min. SNR received by MS	2.6505 dB
BER threshold	0.0001
Max. spatial diversity	3
Traffic required by MS (DL)	1 Mbps
Traffic required by MS (UL)	0.5 Mbps
BS capacity	100 Mbps
RS basic capacity	15 Mbps
RS fix cost	1 M dollars
RS config. cost	0.2 M dollars

In this research, the SNR function we apply is listed as follows:

$$\text{SNR (dBm)} = P_t + G_t + G_r - \overline{PL(d)} - N, \tag{33}$$

where P_t is transmit power, G_t is transmit gain, G_r is receive gain, $\overline{PL(d)}$ is path loss function, N is thermal noise function.

The BER evaluation functions we apply have been moderated with various modulation schemes [28, 29] are demonstrated the theoretical and simulated results of BER value in four different modulation schemes.

4.4. Experiment Scenarios. For the unique characteristics of this network deployment problem, the given circumstances are BS and MS locations, but RSs would be candidate locations. There is no RS built at the beginning. The word "topology" introduced in the following refers to the geographic distribution (the position) of locations where an RS could be built. Two types of topologies, grid and random, are proposed with different numbers of RS and MS in one BS environment to analyze the impact on deployment cost. We then apply different numbers of RS and MS with two BSs in a random topology to analyze the deployment in multiple BSs environment. Table 5 lists the experiment scenarios; Figures 5 and 6 show the graphic examples of grid and random networks. For each scenario, all MSs are guaranteed to have transmission paths and each scenario can be solved in our experiments. In these scenarios, BSs are at the center of the network, RSs are in Grid/Random topology, and MSs are in random topology.

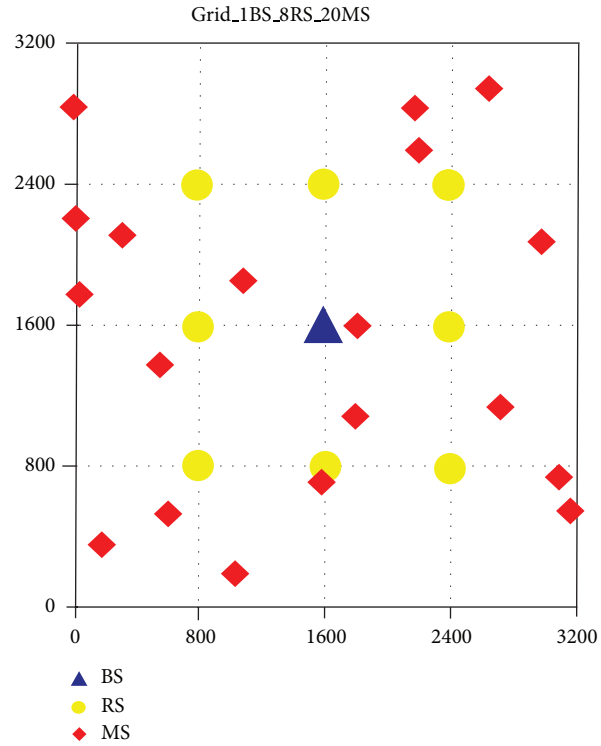


Figure 5: Grid topology example.

4.5. Experiment Results. In Lagrangian relaxation approach, an upper bound (UB) of the problem, is the best primal feasible solution, while the solution to the Lagrangian dual problem guarantees the lower bound (LB) of the problem. By solving the Lagrangian dual problem iteratively and getting a primal feasible solution, we derive the LB and the UB, respectively. Thus, the gap between the UB and LB, computed by (UB − LB)/LB × 100%, illustrates the quality (optimality) of the problem solution.

Figure 7 and Table 6 show the total deployment cost calculated by different algorithms within 1 BS and grid RS topology configuration with different numbers of RS and MS are deployed, respectively. It is obvious that LR-based algorithm results in superior solution in comparison with MBA and DBA, especially when the RS number is large. Additionally, DBA has lower costs than MBA. This illustrates that the LR algorithm has a trend of choosing RS with large MS density instead of RS, which results in minimum BER.

In RS grid topology, for a given network scale, the distance of RS is the farthest locations from BS to receive signals under BER threshold should be included mandatorily. This phenomenon can be observed in Figures 8 and 9. The RS locations in grid topologies of RS = 24 exclude the RS locations in the same topologies of RS = 8 where the about farthest locations BS can reach an RS. The costs in the scenarios of RS = 24 are all higher than in the scenarios of RS = 8 except MBA. We infer that this is because some MSs

Table 4: Modulation and code rate [24, 25].

Modulation	Code rate	SNR	DL rate (Mbps)	UL rate (Mbps)
QPSK	1/2 CTC	SNR ≤ 9.4	6.34	4.70
QPSK	3/4 CTC	9.4 < SNR ≤ 11.2	9.50	7.06
16 QAM	1/2 CTC	11.2 < SNR ≤ 16.4	12.67	9.41
16 QAM	3/4 CTC	16.4 < SNR ≤ 18.2	19.01	14.11
64 QAM	2/3 CTC	18.2 < SNR ≤ 22.7	25.34	18.82
64 QAM	3/4 CTC	22.7 < SNR	28.51	21.17

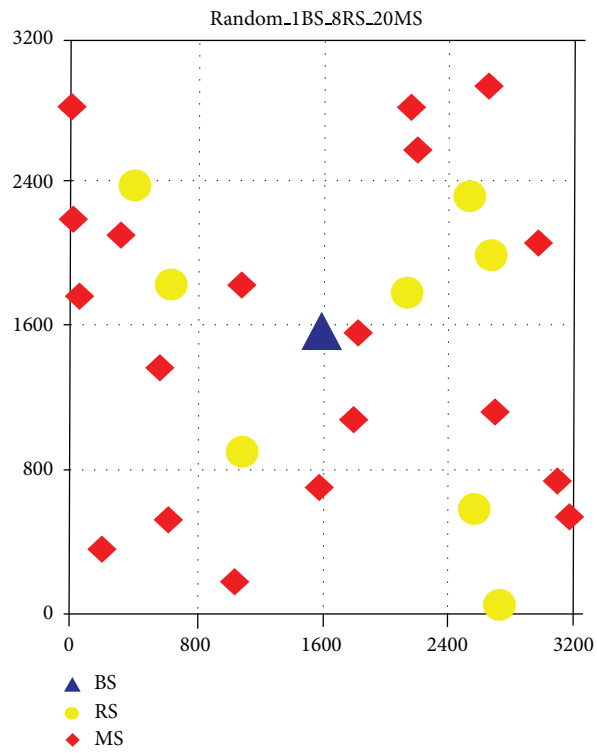


Figure 6: Random topology example.

Table 5: Experiment scenarios.

Topology	Network scale	No. of BS	No. of RS	No. of MS
Grid	3.2 km	1	8, 24, 48	20, 30, 40, 50
Grid	6.4 km	1	24, 48	20, 30, 40, 50
Grid	9.6 km	1	80	20
Random	3.2 km	1	8, 24, 48	20, 30, 40, 50
Random	6.4 km	2	16, 48	40, 60, 80

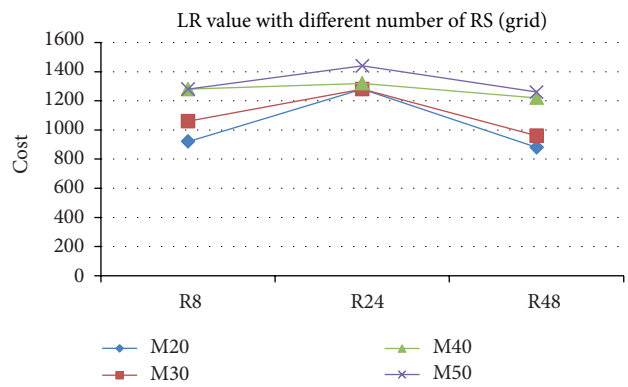


Figure 8: Deployment cost with different number of RS (1 BS, grid, 3.2 km).

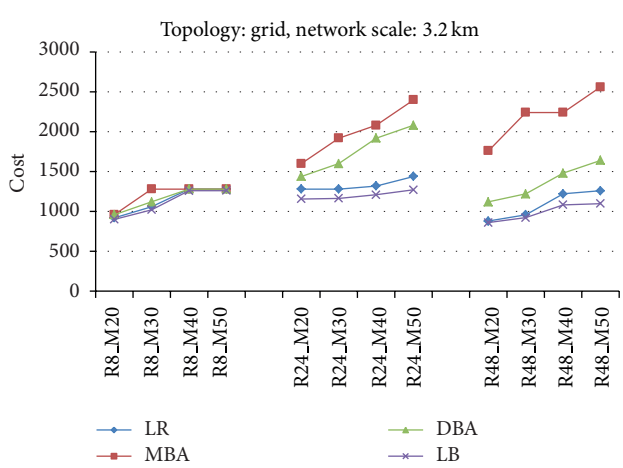


Figure 7: Deployment cost with different number of RS and MS (1 BS, grid, 3.2 km).

in RS = 24 need more hops than RS = 8 to reach the BS, thus inducing costs.

From Figures 10 and 11, we can come to the conclusion that with a fixed number of MSs, total deployment costs are reduced with an increasing number of RSs. Meanwhile, with a fixed number of RSs, total deployment costs are reduced with an increase in the number of MSs.

Figure 10 and Table 7 show total deployment costs as calculated by different algorithms through 1 BS with different numbers of RS and MS in a random topology. Again, it is obvious that the Lagrangian Relaxation-based algorithm receives better solution of quality in comparison with MBA and DBA (bold font), particularly so with a large number of RSs.

Figures 11 and 12 indicate the same conclusion in grid topology. With a fixed number of MSs, total deployment costs are reduced with an increase in the number of RSs. At the same time, with a fixed number of RSs, total deployment costs are reduced with an increase in the number of MSs.

Table 6: Algorithm comparison (1 BS, grid, 3.2 km).

No. of RS	No. of MC	LB	UB	GAP (%)	MBA	Imp. ratio of MBA (%)	DBA	Imp. ratio of DBA (%)
8	20	901.7678	920	1.98176	960	**4.347826**	960	**4.347826**
8	30	1020.242	1060	3.750792	1280	**20.75472**	1120	**5.660377**
8	40	1258.947	1280	1.644797	1280	**0**	1280	**0**
8	50	1260.484	1280	1.524727	1280	**0**	1280	**0**
24	20	1156.286	1280	9.665148	1600	**25**	1440	**12.5**
24	30	1164.774	1280	9.002031	1920	**50**	1600	**25**
24	40	1208.743	1320	8.428545	2080	**57.57576**	1920	**45.45455**
24	50	1269.846	1440	11.81623	2400	**66.66667**	2080	**44.44444**
48	20	860.6118	880	2.203205	1760	**100**	1120	**27.27273**
48	30	921.4716	960	4.013375	2240	**133.3333**	1220	**27.08333**
48	40	1082.548	1220	11.2666	2240	**83.60656**	1480	**21.31148**
48	50	1098.812	1260	12.79272	2560	**103.1746**	1640	**30.15873**

Table 7: Algorithm comparison (1 BS, random, 3.2 km).

No. of RS	No. of MC	LB	UB	GAP (%)	MBA	Imp. ratio of MBA (%)	DBA	Imp. ratio of DBA (%)
8	20	867.3459	900	3.628233	960	**6.666667**	960	**6.666667**
8	30	850.3321	900	5.518652	960	**6.666667**	960	**6.666667**
8	40	846.2536	900	5.971822	1120	**24.44444**	960	**6.666667**
8	50	909.7847	980	7.164823	1280	**30.61224**	1020	**4.081633**
24	20	811.1707	860	5.67783	1600	**86.04651**	960	**11.62791**
24	30	798.1251	860	7.19476	1920	**123.2558**	1020	**18.60465**
24	40	805.3412	900	10.51765	2080	**131.1111**	1020	**13.33333**
24	50	860.9539	980	12.14756	2400	**144.898**	1340	**36.73469**
48	20	734.8455	820	10.3847	1760	**114.6341**	1020	**24.39024**
48	30	766.4685	860	10.87563	1820	**111.6279**	1080	**25.58142**
48	40	744.6947	880	15.3756	2260	**156.8182**	1140	**29.54545**
48	50	768.6480	920	16.4513	2420	**163.0435**	1260	**36.95652**

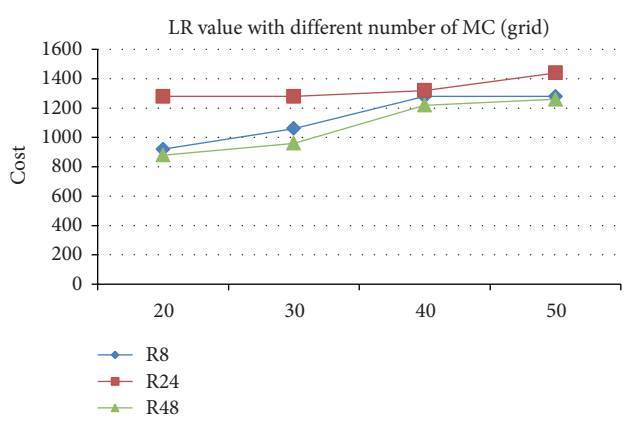


Figure 9: Deployment cost with different number of MS (1 BS, grid, 3.2 km).

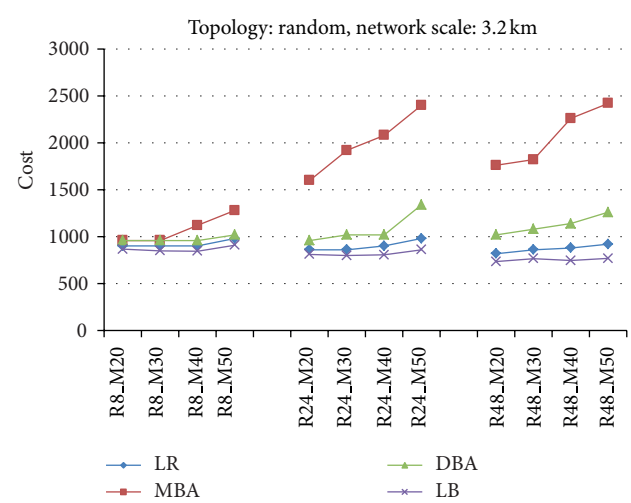


Figure 10: Deployment cost with different number of RS and MS (1 BS, random, 3.2 km).

In random topology, it is difficult to generate a network capable of satisfying every MS's transmission when a few RSs (ex. RS = 8) are deployed. In general, RSs are not distributed uniformly enough to fully cover all MSs. Figure 13 and Table 8 show total deployment costs calculated by different algorithms under 2 BS and random topology, with different number of RS and MS. We come to the same conclusion: the Lagrangian Relaxation-based algorithm still gets better

Table 8: Algorithm comparison (2 BSs, random, 6.4 km).

No. of RS	No. of MC	LB	UB	GAP (%)	MBA	Imp. ratio of MBA (%)	DBA	Imp. ratio of DBA (%)
16	40	1542.505	1620	4.78362	1760	**8.641975**	1680	**3.703704**
16	60	1726.22	1840	6.18371	2240	**21.73913**	1840	**0**
16	80	1737.373	1920	9.5118	2400	**25**	2020	**5.208333**
48	40	1409.38	1540	8.48179	3040	**97.4026**	1760	**14.28571**
48	60	1533.7687	1720	10.8274	3360	**95.34884**	1940	**12.7907**
48	80	1550.1775	1820	14.82541	3840	**110.989**	2280	**25.27473**

Table 9: Experiment results (1 BS, grid, 6.4 km).

No. of RS	No. of MC	LB	UB	GAP (%)	MBA	Imp. ratio of MBA (%)	DBA	Imp. ratio of DBA (%)
8	20	N/A	N/A	N/A	N/A	**N/A**	N/A	**N/A**
8	30	N/A	N/A	N/A	N/A	**N/A**	N/A	**N/A**
8	40	N/A	N/A	N/A	N/A	**N/A**	N/A	**N/A**
8	50	N/A	N/A	N/A	N/A	**N/A**	N/A	**N/A**
24	20	2155.106	2340	7.901457	2720	**16.23932**	2420	**3.418803**
24	30	2212.838	2480	10.77267	3520	**41.93548**	2840	**14.51613**
24	40	2469.994	2820	12.41156	4800	**70.21277**	3360	**19.14894**
24	50	2699.765	3440	21.51847	5440	**58.13953**	4480	**30.23256**
48	20	1537.11	1700	9.581763	2880	**69.41176**	1960	**15.29412**
48	30	2059.57	2320	11.22543	4480	**93.10345**	2640	**13.7931**
48	40	2309.617	2720	15.08761	5600	**105.8824**	3480	**27.94118**
48	50	2445.661	3280	25.43715	6240	**90.2439**	4880	**48.78049**

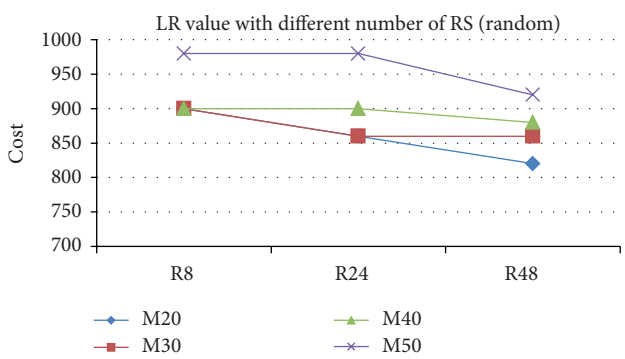


Figure 11: Deployment cost with different number of RS (1 BS, random, 3.2 km).

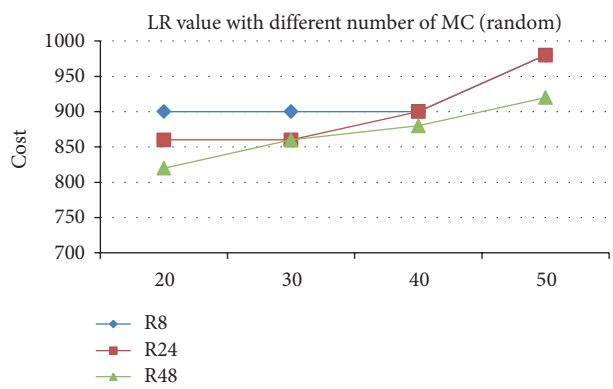


Figure 12: Deployment cost with different number of MS (1 BS, random, 3.2 km).

solution of quality in comparison with MBA and DBA, more so with a large number of RSs.

If these scenarios experimented previously are double the size of those in which BS = 1, how the result is. With random topology in BS = 1, it is also difficult to get a feasible network when RS number is small (ex. RS = 16 here). Figure 13 illustrates total deployment costs in random topologies with BS = 1 and BS = 2. Since the RS locations are different in both conditions, it would be fruitless to compare their costs. However, it is still obvious that the gaps are all larger in every scenarios in BS = 2 than in BS = 1 for network complexity.

Figure 14 and Table 9 show the scenario of 1 BS, with different number of RS and MS in a grid topology with a 6.4 km network scale. Since the network (6.4 km) is larger than that of previous experiments (3.2 km), 8 RSs is no longer sufficient to fulfill all transmissions. Therefore, 24 RSs (two layers from the BS) becomes the smallest size of this network scale.

Figure 15 demonstrates the deployment costs of 20 MSs with various numbers of RS among three kinds of network scale, 3.2 km with 8 RSs (1 layer from the BS), 6.4 km with 24 RSs, and 9.6 km with 80 RSs (4 layers from the BS). As explained previously, in the 6.4 km network scale the smallest grid size is 24 RSs (2 layers from the BS). One can see the same situation in 9.6 km, with the smallest grid size being 80 RSs (3 layers from the BS).

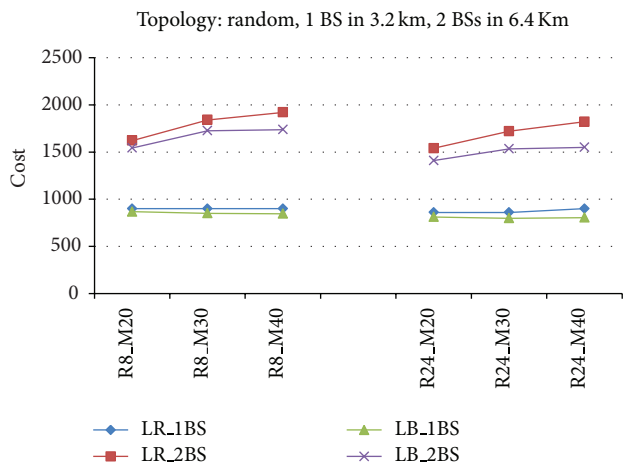


Figure 13: Deployment cost with different number of BS in random topology.

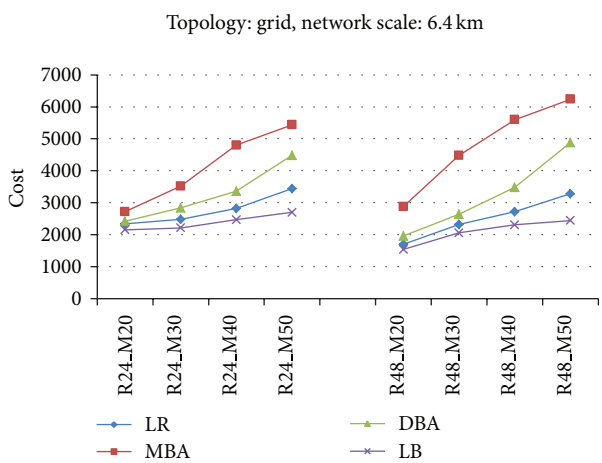


Figure 14: Deployment cost with different number of RS and MS (1 BS, gird, 6.4 km).

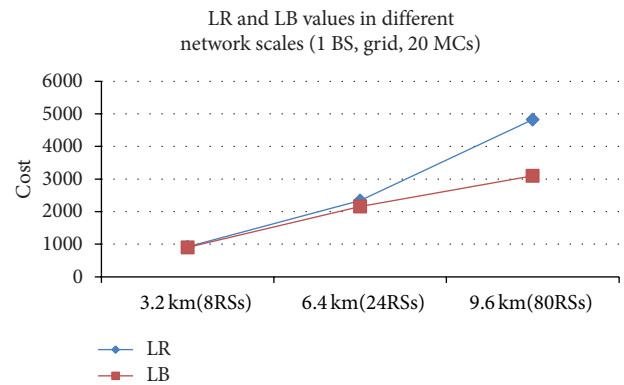


Figure 15: Deployment cost and lower bound (LB) values in different network scales (1 BS, grid, 20 MSs).

5. Conclusions and Future Work

With 3G technology established, it was obvious that the traffic is increased significantly, but the average revenue per user (ARPU) is decreased very fast. The business challenges of operators would be that increasing revenues by finding other solutions more efficiency and effectiveness. But the network development of a new 4G system started to be investigated and made huge investments of macro base station deployments. In one early investigation which took relays would be able to speed up extend services and expanded market share at this stage economically. So, the operators can make new revenues and cost reduction balance.

Although our experiments do not cover large network scales with large number of RSs and MSs for the restrictions of computational capabilities, these model approaches can nevertheless be regarded as useful engineering guidelines for future LTE-A relay network development.

In this paper, we stand for an operator to propose a solution that determines where and how many relays should be deployed in the planning stages to minimize the development cost. In the planning stages, we not only derive a Multicast Tree routing algorithm to both determine and fulfill the QoS requirements and also enhance throughput on both down-link and up-link communications, but we also utilize the Lagrangian Relaxation Method in conjunction with optimization-based heuristics and conduct computational experiments to evaluate the performance of the proposed algorithms.

Our contributions in this research can be divided into three parts. First, we have constructed the network architecture with multicast tree routing concepts. Secondly, we proposed a precise mathematical expression to model the network architecture problem. This is not an intuitive mathematical model for considering the solvability of this problem. We have designed the entire model not only to be solvable but also to not violate the physical meanings. Finally, we provide the lagrangian relaxation and optimization-Based algorithms to solve this problem; we prove it to have superior quality after verification with other simple algorithms and lower bound value. This optimal solution is a good strategic method to minimize the CAPEX of operators to build up a relay network with more efficiency and effectiveness and the QoS can be guaranteed.

Appendix

Solution Approach. The wireless relay deployment problem is emulated as a mixed integer and linear programming (MILP) problem. To solve this problem, the optimal development cost for network planning is minimized to relay selection constraints, nodal, and link capacity constraints, cooperatively relaying constraints, and routing constraints for both UL and DL transmissions. The Lagrangian Relaxation method is proposed in conjunction with the optimization-based heuristics to solve the problem. The primal problem (IP 1) is transformed into the following Lagrangian Relaxation Problem, where constraints (3), (4), (5), (6), (10), (11), (12), (14), (15), (16), (17), (19), (20), (21), (22), (23), (25), (26), (27), (28), and (29) are relaxed by introducing Lagrangian multiplier vector $\mu_1 \sim \mu_{22}$.

Optimal Problem. One has

$$
\begin{aligned}
&Z_D\left(\mu_1,\mu_2,\mu_3,\mu_4,\mu_5,\mu_6,\mu_7,\mu_8,\mu_9,\mu_{10},\mu_{11},\mu_{12},\mu_{13},\mu_{14},\mu_{15},\mu_{16},\mu_{17},\mu_{18},\mu_{19},\mu_{20},\mu_{21},\mu_{22}\right)\\
&\quad=\min\sum_{r\in R}\sum_{k\in K}\left(\psi_r+\Phi_r(k)\right)\eta_{rk}+\sum_{r\in R}\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu^1_{r\mathrm{dir}}\left[\sum_{b\in B}h^{\mathrm{dir}}_{rb}-\sum_{k\in K}\eta_{rk}\right]\\
&\qquad+\sum_{r\in R}\mu^2_r\left[\sum_{u\in\{R\cup B\}}\sum_{n\in N}y^1_{nur}\theta^1_n+\sum_{w\cup\{R\cup B\}}\sum_{n\in N}y^2_{nrw}\theta^2_n-\sum_{k\in K}\eta_{rk}\overline{C_r(k)}\right]\\
&\qquad+\sum_{b\in B}\mu^3_b\left[\sum_{r\in R}\sum_{n\in N}y^1_{nbr}\theta^1_n+\sum_{i\in R}\sum_{n\in N}y^2_{nib}\theta^2_n+\sum_{n\in N}\sum_{\mathrm{dir}\in\mathrm{DIR}}\kappa^{\mathrm{dir}}_{nb}\theta^{\mathrm{dir}}_n-\overline{C_b}\right]\\
&\qquad+\sum_{n\in N}\sum_{r\in R}\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu^4_{nr\mathrm{dir}}\left[\kappa^{\mathrm{dir}}_{nr}-\sum_{k\in K}\eta_{rk}\right]+\sum_{n\in N}\sum_{s\in\{R\cup B\}}\mu^5_{ns}\left[\kappa^1_{ns}P^N_{\min}-\kappa^1_{ns}\pi^1_{sn}\right]\\
&\qquad+\sum_{n\in N}\sum_{s\in\{R\cup B\}}\mu^6_{ns}\left[\kappa^2_{ns}P^R_{\min}-\kappa^2_{ns}\pi^2_{ns}\right]+\sum_{n\in N}\mu^7_n\left[\omega_n-\sum_{s\in\{R\cup B\}}\kappa^1_{ns}\pi^1_{sn}\right]\\
&\qquad+\sum_{n\in N}\sum_{s\in\{R\cup B\}}\mu^8_{ns}\left[\kappa^1_{ns}\pi^1_{sn}-\sum_{k\in K}\eta_{sk}\left(P\left(\rho^1_s(k),D_{sn},\tau\right)-P_N(n)\right)\right]\\
&\qquad+\sum_{n\in N}\sum_{s\in\{R\cup B\}}\mu^9_{ns}\left[\kappa^2_{ns}\pi^2_{ns}-\sum_{k\in K}\eta_{sk}\left(P\left(\rho^N_n,D_{ns},\tau\right)-P_N(s)\right)\right]\\
&\qquad+\sum_{n\in N}\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu^{10}_{nuv\mathrm{dir}}\left[y^{\mathrm{dir}}_{nuv}P^R_{\min}-y^{\mathrm{dir}}_{nuv}\phi^{\mathrm{dir}}_{uv}\right]+\sum_{n\in N}\sum_{v\in\{R\cup B\}}\mu^{11}_{nv}\left[\varepsilon_{nv}-\sum_{u\in\{R\cup B\}}y^2_{nuv}\phi^2_{uv}\right]\\
&\qquad+\sum_{n\in N}\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu^{12}_{nuv\mathrm{dir}}\left[y^{\mathrm{dir}}_{nuv}\phi^{\mathrm{dir}}_{uv}-\sum_{k\in K}\eta_{uk}\left(P\left(\rho^{\mathrm{dir}}_u(k),D_{uv},\tau\right)-P_N(v)\right)\right]\\
&\qquad+\sum_{n\in N}\mu^{13}_n\left[\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}y^1_{nuv}\mathrm{BEP}_v\left(\phi^1_{uv}\right)+\mathrm{BEP}_n(\omega_n)-\mathrm{BEP}^1\right]\\
&\qquad+\sum_{n\in N}\mu^{14}_n\left[\sum_{v\in\{R\cup B\}}\mathrm{BEP}_v(\varepsilon_{nv})+\sum_{s\in\{R\cup B\}}\mathrm{BEP}_s\left(\kappa^2_{ns}\pi^2_{ns}\right)-\mathrm{BEP}^2\right]+\sum_{n\in N}\sum_{r\in R}\sum_{b\in B}\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu^{15}_{nrb\mathrm{dir}}\left[\kappa^{\mathrm{dir}}_{nr}-h^{\mathrm{dir}}_{rb}\right]\\
&\qquad+\sum_{n\in N}\sum_{i\in R}\sum_{j\in R}\sum_{b\in B}\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu^{16}_{nijb\mathrm{dir}}\left[\sum_{p\in P_{bi}}\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}x^{\mathrm{dir}}_{nip}\delta_{puv}-\sum_{p\in P_{bj}}\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}x^{\mathrm{dir}}_{njp}\delta_{puv}-\left(1-\kappa^{\mathrm{dir}}_{nj}\right)\overline{M}\right]\\
&\qquad+\sum_{n\in N}\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\mu^{17}_{nuv}\left[y^1_{nuv}-\sum_{k\in K}\eta_{vk}\right]+\sum_{n\in N}\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\mu^{18}_{nuv}\left[y^2_{nuv}-\sum_{k\in K}\eta_{uk}\right]\\
&\qquad+\sum_{n\in N}\sum_{r\in R}\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu^{19}_{nruv\mathrm{dir}}\left[\sum_{b\in B}\sum_{p\in P_{br}}x^{\mathrm{dir}}_{nrp}\delta_{puv}-y^{\mathrm{dir}}_{nuv}\right]\\
&\qquad+\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu^{21}_{uv\mathrm{dir}}\left[\sum_{n\in N}y^{\mathrm{dir}}_{nuv}\theta^{\mathrm{dir}}_n-C_{uv}\left(\phi^{\mathrm{dir}}_{uv}\right)\right]+\sum_{n\in N}\sum_{r\in R}\sum_{b\in B}\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu^{22}_{nrb\mathrm{dir}}\left[h^{\mathrm{dir}}_{rb}-\sum_{p\in P_{br}}x^{\mathrm{dir}}_{nrp}\right],
\end{aligned}
\tag{A.1}
$$

subject to

$$\begin{gathered}
\sum_{k\in K}\eta_{rk}\le 1 \quad \forall r\in R,\\
\sum_{b\in B}h_{rb}^{\mathrm{dir}}\le 1 \quad \forall r\in R,\ \mathrm{dir}\in \mathrm{DIR},\\
\sum_{b\in B}\kappa_{nb}^{\mathrm{dir}}+\kappa_{nr}^{\mathrm{dir}}\le 1 \quad \forall n\in N,\\
r\in R,\quad \mathrm{dir}\in \mathrm{DIR},\\
1\le \sum_{s\in\{R\cup B\}}\kappa_{ns}^{\mathrm{dir}} \quad \forall n\in N,\\
s\in\{R\cup B\},\quad \mathrm{dir}\in \mathrm{DIR},\\
\sum_{r\in R}\kappa_{nr}^{\mathrm{dir}}\le \mathrm{SD}^{\mathrm{dir}} \quad \forall n\in N,\ \mathrm{dir}\in \mathrm{DIR},\\
0\le \pi_{sn}^{\mathrm{dir}}\le \overline{\pi} \quad \forall s\in\{R\cup B\},\\
n\in N,\quad \mathrm{dir}\in \mathrm{DIR},\\
0\le \phi_{uv}^{\mathrm{dir}}\le \overline{\phi} \quad \forall u,v\in\{R\cup B\},\ \mathrm{dir}\in \mathrm{DIR},\\
\sum_{b\in B}\sum_{p\in P_{br}}x_{nrp}^{\mathrm{dir}}\le 1 \quad \forall n\in N,\\
r\in R,\quad \mathrm{dir}\in \mathrm{DIR},\\
\eta_{rk}=0 \text{ or } 1 \quad \forall r\in R,\ k\in K,\\
h_{rb}^{\mathrm{dir}}=0 \text{ or } 1 \quad \forall r\in R,\\
b\in B,\quad \mathrm{dir}\in \mathrm{DIR},\\
\kappa_{ns}^{\mathrm{dir}}=0 \text{ or } 1 \quad \forall n\in N,\\
s\in\{R\cup B\},\quad \mathrm{dir}\in \mathrm{DIR},\\
x_{nrp}^{\mathrm{dir}}=0 \text{ or } 1 \quad \forall p\in P_{br},\ b\in B,\\
r\in R,\quad n\in N,\\
y_{nuv}^{\mathrm{dir}}=0 \text{ or } 1 \quad \forall n\in N,\\
u,v\in\{R\cup B\},\quad \mathrm{dir}\in \mathrm{DIR}.
\end{gathered} \tag{A.2}$$

Subproblem 1 (related to decision variable η_{rk}). One has

$$\begin{aligned}
&Z_{\mathrm{sub3.1}}\left(\mu_1,\mu_2,\mu_4,\mu_8,\mu_9,\mu_{12},\mu_{17},\mu_{18}\right)\\
&\quad=\min\Bigg\{\sum_{r\in R}\sum_{k\in K}\Bigg[\psi_r+\Phi_r(k)-\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu_{r\mathrm{dir}}^{1}-\mu_r^{2}\overline{C_r(k)}\\
&\qquad+\sum_{n\in N}\Bigg[-\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu_{nr\mathrm{dir}}^{4}\\
&\qquad\quad-\mu_{nr}^{8}\left(P\left(\rho_r^{1}(k),D_{rn},\tau\right)-P_N(n)\right)\\
&\qquad\quad-\mu_{nr}^{9}\left(P\left(\rho_n^{N},D_{nr},\tau\right)-P_N(r)\right)\\
&\qquad\quad-\sum_{u\in\{R\cup B\}}\left(\mu_{nur}^{17}+\mu_{nru}^{18}\right)\\
&\qquad\quad-\sum_{v\in\{R\cup B\}}\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu_{nrv\mathrm{dir}}^{12}\\
&\qquad\qquad\times\left(P\left(\rho_r^{\mathrm{dir}}(k),D_{rv},\tau\right)\right.\\
&\qquad\qquad\left.-P_N(v)\right)\Bigg]\Bigg]\eta_{rk}\\
&\qquad-\sum_{r\in B}\sum_{k\in K}\Bigg[\sum_{n\in N}\left[\mu_{nr}^{8}\left(P\left(\rho_r^{1}(k),D_{rn},\tau\right)-P_N(n)\right)\right.\\
&\qquad\qquad\left.+\mu_{nr}^{9}\left(P\left(\rho_n^{N},D_{nr},\tau\right)-P_N(r)\right)\right]\\
&\qquad+\sum_{v\in\{R\cup B\}}\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu_{nrv\mathrm{dir}}^{12}\\
&\qquad\qquad\times\left(P\left(\rho_r^{\mathrm{dir}}(k),D_{rv},\tau\right)\right.\\
&\qquad\qquad\left.-P_N(v)\right)\Bigg]\eta_{rk}\Bigg\},
\end{aligned} \tag{Sub 3.1}$$

subject to

$$\sum_{k\in K}\eta_{rk}\le 1 \quad \forall r\in R, \tag{A.3a}$$

$$\eta_{rk}=0 \text{ or } 1 \quad \forall r\in R,\ k\in K. \tag{A.3b}$$

Because the configuration of BS is constant, (Sub 3.1) can be further decomposed into $|R|$ independent subproblems. For each candidate location r:

$$\begin{aligned}
&\min\Bigg\{\sum_{k\in K}\Bigg[\psi_r+\Phi_r(k)-\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu_{r\mathrm{dir}}^{1}-\mu_r^{2}\overline{C_r(k)}\\
&\qquad+\sum_{n\in N}\Bigg[-\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu_{nr\mathrm{dir}}^{4}\\
&\qquad\quad-\mu_{nr}^{8}\left(P\left(\rho_r^{1}(k),D_{rn},\tau\right)-P_N(n)\right)\\
&\qquad\quad-\mu_{nr}^{9}\left(P\left(\rho_n^{N},D_{nr},\tau\right)-P_N(r)\right)\\
&\qquad\quad-\sum_{u\in\{R\cup B\}}\left(\mu_{nur}^{17}+\mu_{nru}^{18}\right)\\
&\qquad\quad-\sum_{v\in\{R\cup B\}}\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu_{nrv\mathrm{dir}}^{12}\\
&\qquad\qquad\times\left(P\left(\rho_r^{\mathrm{dir}}(k),D_{rv},\tau\right)\right.\\
&\qquad\qquad\left.-P_N(v)\right)\Bigg]\Bigg]\eta_{rk}\Bigg\}
\end{aligned} \tag{Sub 3.1.1}$$

subject to (A.3a) and (A.3b).

For each (Sub 3.1.1), find the configuration k corresponding to the smallest coefficient value of η_{rk}. If the coefficient is negative, then set η_{rk} to be 1 and 0 otherwise.

Subproblem 2 (related to decision variables h_{rb}^{dir}). One has

$$Z_{\text{sub3.2}}(\mu_1, \mu_{15}, \mu_{22}) = \min\left\{\sum_{r\in R}\sum_{b\in B}\sum_{\text{dir}\in\text{DIR}} h_{rb}^{\text{dir}}\left[\mu_{r\text{dir}}^{1} + \sum_{n\in N}\left(\mu_{nrb\text{dir}}^{22} - \mu_{nrb\text{dir}}^{15}\right)\right]\right\}, \quad \text{(Sub 3.2)}$$

subject to

$$\sum_{b\in B} h_{rb}^{\text{dir}} \le 1 \quad \forall r \in R,\ \text{dir} \in \text{DIR}, \quad \text{(A.4a)}$$

$$h_{rb}^{\text{dir}} = 0 \text{ or } 1 \quad \forall r \in R,\ b \in B,\ \text{dir} \in \text{DIR}. \quad \text{(A.4b)}$$

Equation (Sub 3.2) can be further decomposed into $|R| \times |\text{DIR}|$ subproblems. For each RS r and direction dir,

$$\min\left\{\sum_{b\in B} h_{rb}^{\text{dir}}\left[\mu_{r\text{dir}}^{1} + \sum_{n\in N}\left(\mu_{nrb\text{dir}}^{22} - \mu_{nrb\text{dir}}^{15}\right)\right]\right\} \quad \text{(Sub 3.2.1)}$$

Subject to (A.4a) and (A.4b).

For each (Sub 3.2.1), find the BS b which can result in the smallest coefficient $\mu_{r\text{dir}}^{1} + \sum_{n\in N}(\mu_{nrb\text{dir}}^{22} - \mu_{nrb\text{dir}}^{15})$ of h_{rb}^{dir}; if the coefficient is negative, then set h_{rb}^{dir} to be 1 and 0 otherwise.

Subproblem 3 (related to decision variables κ_{ns}^{dir}, π_{sn}^{dir}). One has

$$\begin{aligned} Z_{\text{sub3.2}}&(\mu_3, \mu_4, \mu_5, \mu_6, \mu_7, \mu_8, \mu_9, \mu_{14}, \mu_{15}, \mu_{16}) \\ &= \min \sum_{n\in N}\left\{\sum_{s\in B}\left[\sum_{\text{dir}\in\text{DIR}} \mu_s^3 \kappa_{ns}^{\text{dir}} \theta_n^{\text{dir}} + \mu_{ns}^5\left(\kappa_{ns}^1\left(P_{\min}^N - \pi_{sn}^1\right)\right) + \mu_{ns}^6\left(\kappa_{ns}^2\left(P_{\min}^R - \pi_{ns}^2\right)\right) + \left(\mu_{ns}^8\pi_{sn}^1 + \mu_{ns}^9\pi_{ns}^2\right)\right]\right. \\ &\quad + \sum_{s\in R}\left[\sum_{\text{dir}\in\text{DIR}}\left(\mu_{ns\text{dir}}^4\kappa_{ns}^{\text{dir}} - \mu_{ns\text{dir}}^{15}\kappa_{ns}^{\text{dir}} + \sum_{i\in R}\sum_{b\in B}\mu_{nisb\text{dir}}^{16}\kappa_{ns}^{\text{dir}}\overline{M}\right) + \mu_{ns}^5\left(\kappa_{ns}^1\left(\overline{M} - \pi_{sn}^1\right)\right) + \mu_{ns}^6\left(\kappa_{ns}^2\left(\overline{M} - \pi_{ns}^2\right)\right) + \left(\kappa_{ns}^1\mu_{ns}^8\pi_{sn}^1 + \kappa_{ns}^2\mu_{ns}^9\pi_{ns}^2\right)\right] \\ &\quad \left. + \sum_{s\in\{R\cup B\}}\left[\mu_n^7\kappa_{ns}^1\left(\pi_{sn}^1\right) + \mu_n^{14}\kappa_{ns}^2\text{BEP}\left(\pi_{ns}^2\right)\right]\right\}, \end{aligned} \quad \text{(Sub 3.3)}$$

subject to

$$\sum_{b\in B}\kappa_{nb}^{\text{dir}} + \kappa_{nr}^{\text{dir}} \le 1 \quad \forall n \in N,\ r \in R,\ \text{dir} \in \text{DIR}, \quad \text{(A.5a)}$$

$$1 \le \sum_{s\in\{R\cup B\}}\kappa_{ns}^{\text{dir}} \quad \forall n \in N,\ \text{dir} \in \text{DIR}, \quad \text{(A.5b)}$$

$$\sum_{r\in R}\kappa_{nr}^{\text{dir}} \le \text{SD}^{\text{dir}} \quad \forall n \in N,\ \text{dir} \in \text{DIR}, \quad \text{(A.5c)}$$

$$0 \le \pi_{sn}^{\text{dir}} \le \overline{\pi} \quad \forall n \in N,\ s \in \{R\cup B\},\ \text{dir} \in \text{DIR}, \quad \text{(A.5d)}$$

$$\kappa_{ns}^{\text{dir}} = 0 \text{ or } 1 \quad \forall n \in N,\ s \in \{R\cup B\},\ \text{dir} \in \text{DIR}. \quad \text{(A.5e)}$$

Equation (Sub 3.3) can be further decomposed into $|N|$ independent subproblems. For each MC n,

$$\begin{aligned} \min&\left\{\sum_{s\in B}\left[\sum_{\text{dir}\in\text{DIR}} \mu_s^3\kappa_{ns}^{\text{dir}}\theta_n^{\text{dir}} + \mu_{ns}^5\left(\kappa_{ns}^1\left(P_{\min}^N - \pi_{sn}^1\right)\right) + \mu_{ns}^6\left(\kappa_{ns}^2\left(P_{\min}^R - \pi_{ns}^2\right)\right) + \left(\mu_{ns}^8\pi_{sn}^1 + \mu_{ns}^9\pi_{ns}^2\right)\right]\right. \\ &+ \sum_{s\in R}\left[\sum_{\text{dir}\in\text{DIR}}\left(\mu_{ns\text{dir}}^4\kappa_{ns}^{\text{dir}} - \mu_{ns\text{dir}}^{15}\kappa_{ns}^{\text{dir}} + \sum_{i\in R}\sum_{b\in B}\mu_{nisb\text{dir}}^{16}\kappa_{ns}^{\text{dir}}\overline{M}\right) + \mu_{ns}^5\left(\kappa_{ns}^1\left(\overline{M} - \pi_{sn}^1\right)\right) + \mu_{ns}^6\left(\kappa_{ns}^2\left(\overline{M} - \pi_{ns}^2\right)\right) + \left(\kappa_{ns}^1\mu_{ns}^8\pi_{sn}^1 + \kappa_{ns}^2\mu_{ns}^9\pi_{ns}^2\right)\right] \end{aligned}$$

$$+\sum_{n\in N}\sum_{s\in\{R\cup B\}}\left[\mu_n^7\kappa_{ns}^1\left(\pi_{sn}^1\right)+\mu_n^{14}\kappa_{ns}^2\mathrm{BEP}\left(\pi_{ns}^2\right)\right]\Bigg\} \tag{Sub 3.3.1}$$

subject to (A.5a), (A.5b), (A.5c), (A.5d), and (A.5e).

The two directions of DL and UL are independent; (Sub 3.3.1) can be decomposed into DL and UL subproblems. Constraint (A.5a) illustrates that once an MC homes to exactly a BS, it cannot home to RS anymore, and vice versa, so (Sub 3.3.1) can be rewritten into following forms. In both directions, while MC n homes to a BS b, the decision variable $\kappa_{nb}^{\mathrm{dir}}=1$. For each MC n

$$\left(\min\left\{\sum_{b\in B}\left[\mu_b^3\theta_n^1+P_{\min}^N\mu_{nb}^5-\pi_{bn}^1\left(\mu_{nb}^5+\mu_n^7-\mu_{ns}^8\right)\right]\kappa_{nb}^1,\ \sum_{r\in R}\left[\mu_{nr1}^4+\sum_{b\in B}\mu_{nrb1}^{15}+P_{\min}^N\mu_{nr}^5+\overline{M}\sum_{i\in R}\sum_{b\in B}\mu_{nirb1}^{16}-\pi_{rn}^1\left(\mu_{nr}^5+\mu_n^7-\mu_{nr}^8\right)\right]\times\kappa_{nr}^1\right\}\right)\quad\text{(for DL)}$$

$$+\left(\min\left\{\sum_{b\in B}\left[\mu_b^3\theta_n^2+P_{\min}^R\mu_{nb}^6-\pi_{nb}^2\left(\mu_{nb}^6-\mu_{ns}^9\right)+\mu_n^{14}\mathrm{BEP}\left(\pi_{nb}^2\right)\right]\kappa_{nb}^2,\ \sum_{r\in R}\left[\mu_{nr2}^4+\sum_{b\in B}\mu_{nrb2}^{15}+P_{\min}^R\mu_{nr}^6+\overline{M}\sum_{i\in R}\sum_{b\in B}\mu_{nirb2}^{16}-\pi_{nr}^2\left(\mu_{nr}^6-\mu_{nr}^9\right)+\mu_n^{14}\mathrm{BEP}\left(\pi_{nr}^2\right)\right]\kappa_{nr}^2\right\}\right)\quad\text{(for UL)} \tag{Sub 3.3.2}$$

subject to (A.5a), (A.5b), (A.5c), (A.5d), and (A.5e).

The algorithm to optimally solve (Sub 3.3.2) is illustrated in the following.

For DL

Step 1. Use SNR function to calculate the SNR value π_{bn}^1 from every BS to MC n.

Step 2. Find the BS b which can result in the smallest coefficient $\mu_b^3\theta_n^1+P_{\min}^N\mu_{nb}^5-\pi_{bn}^1(\mu_{nb}^5+\mu_n^7-\mu_{ns}^8)$ of κ_{nb}^1.

Step 3. Examining all sets of configuration for each RS in the SNR function to determine the SNR value π_{rn}^1; meanwhile, to find the RSs which can result in the first SD^1 smallest coefficient $\mu_{ns1}^4+\sum_{b\in B}\mu_{nsb1}^{15}+P_{\min}^N\mu_{ns}^5+\overline{M}\sum_{i\in R}\sum_{b\in B}\mu_{nisb1}^{16}-\pi_{sn}^1(\mu_{ns}^5+\mu_n^7-\mu_{ns}^8)$ of κ_{nr}^1. The summation of these SD^1 amount of coefficients can be taken into consideration excluding positive ones bigger than the smallest coefficient for the further calculations. That is, we at least had the smallest coefficient for further steps, whether it is negative or not.

Step 4. If the coefficient of κ_{nb}^1 in Step 2 is smaller than the summation of coefficient of SD^1 smallest coefficient of κ_{nr}^1 in Step 3, then set κ_{nb}^1 to be 1; otherwise, set these SD^1 of κ_{nr}^1 to be 1.

For UL. Repeat Step 1 to Step 4 to determine κ_{ns}^2 and π_{sn}^2 with spatial diversity number SD^2.

Subproblem 4 (related to decision variable x_{nrp}^{dir}). One has

$$Z_{\mathrm{sub}}\left(\mu_{16},\mu_{19},\mu_{22}\right)=\min\left\{\sum_{n\in N}\sum_{r\in R}\sum_{\mathrm{dir}\in\mathrm{DIR}}\sum_{b\in B}\sum_{p\in P_{br}}\left[\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\delta_{puv}\left(\mu_{nruv\mathrm{dir}}^{19}+\sum_{j\in R}\mu_{nrjb\mathrm{dir}}^{16}-\sum_{i\in R}\mu_{nirb\mathrm{dir}}^{16}\right)-\mu_{nrb\mathrm{dir}}^{22}\right]x_{nrp}^{\mathrm{dir}}\right\} \tag{Sub 3.4}$$

subject to

$$\sum_{b\in B}\sum_{p\in P_{br}}x_{nrp}^{\mathrm{dir}}\le 1\quad\forall n\in N,\quad r\in R,\quad \mathrm{dir}\in\mathrm{DIR}, \tag{A.6a}$$

$$x_{nrp}^{\mathrm{dir}}=0\text{ or }1\quad\forall n\in N,\ r\in R,\quad p\in P_{br},\quad b\in B,\quad \mathrm{dir}\in\mathrm{DIR}. \tag{A.6b}$$

Equation (Sub 3.4) can be further decomposed into $|N|\times|R|\times|\mathrm{DIR}|$ independent shortest path problems with arc weight $\mu_{nruv\mathrm{dir}}^{19}+\sum_{j\in R}\mu_{nrjb\mathrm{dir}}^{16}-\sum_{i\in R}\mu_{nirb\mathrm{dir}}^{16}$. For each shortest

path problem, it can be effectively solved by Bellman Ford's minimum cost shortest path algorithm. For each RS n, RS r, DIR dir, if the total cost of the shortest path is smaller than $\mu^{22}_{nrb\mathrm{dir}}$, then set x^{dir}_{nrp} to be 1 and 0 otherwise.

Subproblem 5 (related to decision variables y^{dir}_{nuv}, ϕ^{dir}_{uv}). One has

$$\begin{aligned}
&Z_{\mathrm{sub}}\left(\mu_2,\mu_3,\mu_{10},\mu_{11},\mu_{12},\mu_{13},\mu_{17},\mu_{18},\mu_{19},\mu_{21}\right)\\
&\quad=\min\Bigg\{\sum_{n\in N}\left[\sum_{u\in\{R\cup B\}}\sum_{v\in R}\mu^2_v+\sum_{u\in B}\sum_{v\in R}\mu^3_u\right]y^1_{nuv}\theta^1_n\\
&\qquad+\sum_{n\in N}\left[\sum_{u\in R}\sum_{v\in\{R\cup B\}}\mu^2_u+\sum_{u\in R}\sum_{v\in B}\mu^3_v\right]y^2_{nuv}\theta^2_n\\
&\qquad-\sum_{n\in N}\sum_{v\in\{R\cup B\}}\sum_{u\in\{R\cup B\}}\mu^{11}_{nv}y^2_{nuv}\phi^2_{uv}\\
&\qquad+\sum_{n\in N}\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\mu^{13}_n y^1_{nuv}\mathrm{BEP}\left(\phi^1_{uv}\right)\\
&\qquad+\sum_{n\in N}\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\left[\mu^{17}_{nuv}y^1_{nuv}+\mu^{18}_{nuv}y^2_{nuv}\right]\\
&\qquad+\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\sum_{\mathrm{dir}\in\mathrm{DIR}}\Bigg[\sum_{n\in N}\Bigg(\mu^{10}_{nuv\mathrm{dir}}\left(\overline{M}-\phi^{\mathrm{dir}}_{uv}\right)\\
&\qquad\qquad+\mu^{21}_{uv\mathrm{dir}}\theta^{\mathrm{dir}}_n\\
&\qquad\qquad-\sum_{r\in R}\mu^{19}_{nruv\mathrm{dir}}\Bigg)y^{\mathrm{dir}}_{nuv}\Bigg]\\
&\qquad+\sum_{n\in N}\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\sum_{\mathrm{dir}\in\mathrm{DIR}}\left[\mu^{12}_{nuv\mathrm{dir}}\phi^{\mathrm{dir}}_{uv}+\overline{M}\right]y^{\mathrm{dir}}_{nuv}\\
&\qquad-\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\sum_{\mathrm{dir}\in\mathrm{DIR}}\mu^{21}_{uv\mathrm{dir}}C_{uv}\left(\phi^{\mathrm{dir}}_{uv}\right)\Bigg\},
\end{aligned}\tag{Sub 3.5}$$

subject to

$$0\le\phi^{\mathrm{dir}}_{uv}\le\overline{\phi}\quad\forall u,v\in\{R\cup B\},\ \mathrm{dir}\in\mathrm{DIR},\tag{A.7a}$$

$$\begin{aligned}&y^{\mathrm{dir}}_{nuv}=0\text{ or }1\quad\forall n\in N,\\&u,v\in\{R\cup B\},\quad\mathrm{dir}\in\mathrm{DIR}.\end{aligned}\tag{A.7b}$$

Similar to (Sub 3.3), Since DL and UL are two independent transmission directions, we can rewrite (Sub 3.5) into DL and UL subproblems:

$$\begin{aligned}
&\min\Bigg\{\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\Bigg[\sum_{n\in N}\Big[\mu^{10}_{nuv1}P^R_{\min}\\
&\qquad+\left(\mu^{12}_{uv1}-\mu^{10}_{nuv1}\right)\phi^1_{uv}\theta^1_n\\
&\qquad+\mu^{17}_{nuv}-\sum_{r\in R}\mu^{19}_{nruv1}\\
&\qquad+\left(\mu^2_v+\mu^3_u+\mu^{21}_{uv1}\right)\\
&\qquad+\mu^{13}_n\mathrm{BEP}\left(\phi^1_{uv}\right)\Big]y^1_{nuv}\\
&\qquad-\mu^{21}_{uv1}C_{uv}\left(\phi^1_{uv}\right)\Bigg]\Bigg\}\quad(\text{for DL})\\
&+\min\Bigg\{\sum_{u\in\{R\cup B\}}\sum_{v\in\{R\cup B\}}\Bigg[\sum_{n\in N}\Big[\mu^{10}_{nuv2}P^R_{\min}\\
&\qquad+\left(\mu^{12}_{uv2}-\mu^{10}_{nuv2}\right)\phi^2_{uv}\\
&\qquad+\mu^{18}_{nuv}-\sum_{r\in R}\mu^{19}_{nruv2}\\
&\qquad+\left(\mu^2_u+\mu^3_v+\mu^{21}_{uv2}\right)\theta^2_n\\
&\qquad-\mu^{11}_n\phi^2_{uv}\Big]y^2_{nuv}\\
&\qquad-\mu^{21}_{uv2}C_{uv}\left(\phi^2_{uv}\right)\Bigg]\Bigg\}\quad(\text{for DL}),
\end{aligned}\tag{Sub 3.5.1}$$

subject to (A.7a) and (A.7b).

Equation (Sub 3.5.1) can then be decomposed into $|R\cup B|\times|R\cup B|$ independent subproblems. For each link uv,

$$\begin{aligned}
&\min\Bigg\{\sum_{n\in N}\Big[\mu^{10}_{nuv1}P^R_{\min}+\left(\mu^{12}_{uv1}-\mu^{10}_{nuv1}\right)\phi^1_{uv}\\
&\qquad+\mu^{17}_{nuv}-\sum_{r\in R}\mu^{19}_{nruv1}\\
&\qquad+\left(\mu^2_v+\mu^3_u+\mu^{21}_{uv1}\right)\theta^1_n\\
&\qquad+\mu^{13}_n\mathrm{BEP}\left(\phi^1_{uv}\right)\Big]y^1_{nuv}\\
&\qquad-\mu^{21}_{uv1}C_{uv}\left(\phi^1_{uv}\right)\Bigg\}\quad(\text{for DL})\\
&+\min\Bigg\{\sum_{n\in N}\Big[\mu^{10}_{nuv2}P^R_{\min}\\
&\qquad+\left(\mu^{12}_{uv2}-\mu^{10}_{nuv2}\right)\phi^2_{uv}\\
&\qquad+\mu^{18}_{nuv}-\sum_{r\in R}\mu^{19}_{nruv2}\\
&\qquad+\left(\mu^2_u+\mu^3_v+\mu^{21}_{uv2}\right)\theta^2_n\\
&\qquad-\mu^{11}_n\phi^2_{uv}\Big]y^2_{nuv}\\
&\qquad-\mu^{21}_{uv2}C_{uv}\left(\phi^2_{uv}\right)\Bigg\}\quad(\text{for DL}),
\end{aligned}\tag{Sub 3.5.2}$$

subject to (A.7a) and (A.7b).

The algorithm to optimally solve (Sub 3.5.2) is illustrated in the following.

For DL

Step 1. Examining all sets of configuration for each source node u in the SNR function to determine the SNR value ϕ_{uv}^1 which can result in the smallest summation of coefficient $\sum_{n\in N}[\mu_{nuv1}^{10}P_{\min}^R+(\mu_{uv1}^{12}-\mu_{nuv1}^{10})\phi_{uv}^1+\mu_{nuv}^{17}-\sum_{r\in R}\mu_{nruv1}^{19}+(\mu_v^2+\mu_u^3+\mu_{uv1}^{21})\theta_n^1+\mu_n^{13}\mathrm{BEP}(\phi_{uv}^1)]$ of y_{nuv}^1. If all individual coefficients are positive, then set ϕ_{uv}^1 and all y_{nuv}^1 to be 0.

Step 2. For each MC n, if the coefficient $\mu_{nuv1}^{10}P_{\min}^R+(\mu_{uv1}^{12}-\mu_{nuv1}^{10})\phi_{uv}^1+\mu_{nuv}^{17}-\sum_{r\in R}\mu_{nruv1}^{19}+(\mu_v^2+\mu_u^3+\mu_{uv1}^{21})\theta_n^1+\mu_n^{13}\mathrm{BEP}(\phi_{uv}^1)$ is negative, then set y_{nuv}^1 to be 1 and 0 otherwise.

For UL. Repeat Steps 1 and 2 to determine ϕ_{uv}^2 and y_{nuv}^2.

Subproblem 6 (related to decision variable ω_n). One has

$$Z_{\mathrm{sub}}(\mu_7,\mu_{13}) = \min\left\{\sum_{n\in N}\left[\mu_n^{13}\mathrm{BER}(\omega_n)+\mu_n^7\omega_n\right]\right\}, \quad \text{(Sub 3.6)}$$

subject to

$$\omega_n^{\min}\le\omega_n\le\omega_n^{\max}\quad\forall n\in N, \quad \text{(A.8a)}$$

$$\omega_n\in\Omega_n=\{0,\Delta,2\Delta,3\Delta,4\Delta,\ldots\}\quad\forall n\in N. \quad \text{(A.8b)}$$

For (Sub 3.6) can be solvable, we introduced constraint (A.8b) into this subproblem to transform ω_n from continuous to discrete. Then, (Sub 3.6) can be further decomposed into $|N|$ independent subproblems. For each MC n,

$$\min\left\{\mu_n^{13}\mathrm{BER}(\omega_n)+\mu_n^7\omega_n\right\}, \quad \text{(Sub 3.6.1)}$$

subject to (A.8a) and (A.8b).

We can calculate the value of (Sub 3.6.1) by examining every ω_n exhaustively. Set ω_n while it can result in the smallest value of (Sub 3.6.1). Here, we applied the interval Δ to be 0.01.

Subproblem 7 (related to decision variable ε_{nv}). One has

$$Z_{\mathrm{sub}}(\mu_{11},\mu_{14}) = \min\left\{\sum_{n\in N}\sum_{v\in\{R\cup B\}}\left[\mu_n^{14}\mathrm{BER}(\varepsilon_{nv})+\mu_{nv}^{11}\varepsilon_{nv}\right]\right\}, \quad \text{(Sub 3.7)}$$

subject to

$$\varepsilon_{nv}^{\min}\le\varepsilon_{nv}\le\varepsilon_{nv}^{\max}\quad\forall n\in N,\ v\in\{R\cup B\}, \quad \text{(A.9a)}$$

$$\varepsilon_{nv}\in E_{nv}=\{0,\Delta,2\Delta,3\Delta,4\Delta,\ldots\}\quad\forall n\in N,\ v\in\{R\cup B\}. \quad \text{(A.9b)}$$

In the same situation like Subproblem 7, we introduced constraint (A.9b) into subproblem 7 to transform ε_{nv} from continuous to discrete. Equation (Sub 3.7) can be further decomposed into $|N|\times|R\cup B|$ independent subproblems. For each MC n, RS (or BS) v,

$$\min\left\{\mu_n^{14}\mathrm{BER}(\varepsilon_{nv})+\mu_{nv}^{11}\varepsilon_{nv}\right\}, \quad \text{(Sub 3.7.1)}$$

subject to (A.9a), and (A.9b).

Similar to (Sub 3.6.1), (Sub 3.7.1) can be solved by exhaustively examining ε_{nv} to find out the smallest value of this problem; then set ε_{nv}. Here, we applied the interval Δ to be 0.01.

The Dual Problem and the Subgradient Method. According to the algorithms proposed previously, the Lagrangian relaxation problem can be solved effectively and optimally. Based on the weak Lagrangian duality theorem, the objective value of $Z_D(\mu_1,\mu_2,\mu_3,\mu_4,\mu_5,\mu_6,\mu_7,\mu_8,\mu_9,\mu_{10},\mu_{11},\mu_{12},\mu_{13},\mu_{14},\mu_{15},\mu_{16},\mu_{17},\mu_{18},\mu_{19},\mu_{20},\mu_{21},\mu_{22})$ is a lower bound of Z_{IP}. The following dual problem is constructed to calculate the tightest lower bound and solved the dual problem by using the subgradient method.

Dual Problem (D). One has

$$\begin{aligned}Z_D=\max\ Z_D(&\mu_1,\mu_2,\mu_3,\mu_4,\mu_5,\mu_6,\\&\mu_7,\mu_8,\mu_9,\mu_{10},\mu_{11},\\&\mu_{12},\mu_{13},\mu_{14},\mu_{15},\mu_{16},\\&\mu_{17},\mu_{18},\mu_{19},\mu_{20},\mu_{21},\mu_{22}),\end{aligned} \quad \text{(A.10)}$$

subject to

$$\begin{aligned}&\mu_1,\mu_2,\mu_3,\mu_4,\mu_5,\mu_6,\mu_7,\mu_8,\mu_9,\\&\mu_{10},\mu_{11},\mu_{12},\mu_{13},\mu_{14},\mu_{15},\\&\mu_{16},\mu_{17},\mu_{18},\mu_{19},\mu_{20},\mu_{21},\mu_{22}\ge 0.\end{aligned} \quad \text{(A.11)}$$

Let the vector S be a subgradient of $Z_D=\max Z_D(\mu_1,\mu_2,\mu_3,\mu_4,\mu_5,\mu_6,\mu_7,\mu_8,\mu_9,\mu_{10},\mu_{11},\mu_{12},\mu_{13},\mu_{14},\mu_{15},\mu_{16},\mu_{17},\mu_{18},\mu_{19},\mu_{20},\mu_{21},\mu_{22})$. Then, in iteration k of the subgradient procedure, the multiplier vector $m^k=(\mu_1^k,\mu_2^k,\mu_3^k,\mu_4^k,\mu_5^k,\mu_6^k,\mu_7^k,\mu_8^k,\mu_9^k,\mu_{10}^k,\mu_{11}^k,\mu_{12}^k,\mu_{13}^k,\mu_{14}^k,\mu_{15}^k,\mu_{16}^k,\mu_{17}^k,\mu_{18}^k,\mu_{19}^k,\mu_{20}^k,\mu_{21}^k,\mu_{22}^k)$ is updated by $m^{k+1}=m^k+t^kS^k$. The step size t^k is determined by $t^k=\lambda((Z_{IP}^*-Z_D(m^k))/\|S^k\|^2)$. Z_{IP}^* is the best primal objective function value found by iteration k. λ is a constant where $0\le\lambda\le 2$.

Conflict of Interests

It is hereby asserted that all the coauthors of this paper do not have any personal or financial interest with any model or system used in this paper.

References

[1] E. Fox, "North American ARPU growth outpaces the world: a look at wireless forecast drivers," Yankee Group, March 2006.

[2] D. Wang, J. Li, K. Xing, S. Jin, and K. Liu, "Real-time, real-location, and real-identity service information (R3SI) based application enabled mobile service architecture in cellular networks," in *Proceedings of the 32nd International Conference on Distributed Computing Systems Workshops (ICDCSW '12)*, pp. 315–323, Macau.

[3] A. Engels, M. Reyer, and R. Mathar, "Profit-oriented combination of multiple objectives for planning and configuration of 4G multi-hop relay networks," in *Proceedings of the 7th International Symposium on Wireless Communication Systems (ISWCS '10)*, pp. 330–334, September 2010.

[4] 3GPP, http://www.3gpp.org/.

[5] "3GPP TR 36. 814," Evolved Universal Terrestrial Radio Access (E-UTRA), Further Advancements for E-UTRA Physical Layer Aspects.

[6] "3GPP TR 36. 806," Evolved Universal Terrestrial Radio Access (E-UTRA), Relay Architectures for E-UTRA (LTE-Advanced).

[7] "3GPP R3-093149," Comprehensive Comparison among Type-I Relay Alternatives, 2009.

[8] "3GPP R1-101825," Backhaul UL Subframe Allocation in TDD LTE-A Relay, 2010.

[9] Y. Yang, H. Hu, J. Xu, and G. Mao, "Relay technologies for WiMAX and LTE-Advanced mobile systems," *IEEE Communications Magazine*, vol. 47, no. 10, pp. 100–105, 2009.

[10] S. W. Peters, A. Y. Panah, K. T. Truong, and R. W. Heath, "Relay architectures for 3GPP LTE-advanced," *Eurasip Journal on Wireless Communications and Networking*, vol. 2009, Article ID 618787, 2009.

[11] X. Wang, S.-J. Horng, R.-G. Cheng, and P. Fan, "Call dropping performance analysis of the eNB-first channel access policy in LTE-Advanced relay networks," in *Proceedings of the IEEE 7th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob '11)*, pp. 43–50, October 2011.

[12] C. Huang, M. Zeng, and S. Cui, "Achievable rates of two-hop interference networks with conferencing relays," in *Proceedings of the 54th Annual IEEE Global Telecommunications Conference: "Energizing Global Communications" (GLOBECOM '11)*, pp. 1–6, Houston, Tex, USA, December 2011.

[13] V. D. Meulen and C. Edward, *Transmission of Information in a T-Terminal Discrete Memoryless Channel*, Department of Statistics, University of California, Berkeley, Calif, USA, 1968.

[14] R. Pabst, B. H. Walke, D. C. Schultz et al., "Relay-based deployment concepts for wireless and mobile broadband radio," *IEEE Communications Magazine*, vol. 42, no. 9, pp. 80–89, 2004.

[15] H. Li, L. Liu, G. Li, Y. Kim, and J. Zhang, "Multicell cooperation and MIMO technologies for broadcasting and broadband communications," *International Journal of Digital Multimedia Broadcasting*, vol. 2010, Article ID 848527, 2 pages, 2010.

[16] K. Schober, R. Wichman, and T. Roman, "Layer arrangement for single-user coordinated multi-point transmission," in *Proceedings of the 46th Annual Conference on Information Sciences and Systems (CISS '12)*, pp. 1–5.

[17] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity, Part I: system description," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927–1938, 2003.

[18] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity, Part II: implementation aspects and performance analysis," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1939–1948, 2003.

[19] J. Yang, Z. Zhang, Z. Jiang, X. Jiang, and M. He, "Study on joint-coding based on pre-coding and STBC in coordinated multi-point system," in *Proceedings of the Spring Congress on Engineering and Technology (S-CET '12)*, pp. 1–4.

[20] M. Rahman and H. Yanikomeroglu, "Interference avoidance through dynamic downlink OFDMA subchannel allocation using intercell coordination," in *Proceedings of the IEEE 67th Vehicular Technology Conference-Spring (VTC '08)*, pp. 1630–1635, May 2008.

[21] M. Rahman, H. Yanikomeroglu, and W. Wong, "Interference avoidance with dynamic inter-cell coordination for downlink LTE system," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '09)*, pp. 1238–1243, April 2009.

[22] G. Amarasuriya, C. Tellambura, and M. Ardakani, "Performance analysis framework for transmit antenna selection strategies of cooperative MIMO AF relay networks," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 7, pp. 3030–3044, 2011.

[23] A. B. Saleh, O. Bulakci, J. Hämäläinen, S. Redana, and B. Raaf, "Analysis of the impact of site planning on the performance of relay deployments," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 7, pp. 3139–3150, 2012.

[24] N. Kong and L. B. Milstein, "Average SNR of a generalized diversity selection combining scheme," *IEEE Communications Letters*, vol. 3, no. 3, pp. 57–79, 1999.

[25] R. K. Mallik, M. Z. Win, J. W. Shao, M.-S. Alouini, and A. J. Goldsmith, "Channel capacity of adaptive transmission with maximal ratio combining in correlated Rayleigh fading," *IEEE Transactions on Wireless Communications*, vol. 3, no. 4, pp. 1124–1133, 2004.

[26] M. L. Fisher, "The Lagrangian relaxation method for solving integer programming problems," *Management Science*, vol. 27, no. 1, pp. 1–18, 1981.

[27] T. N. Lin, *Content Delivery over Wireless Network- Radio Propagation: Issues & Models*, Department of Electrical Engineering, National Taiwan University, Taiwan.

[28] C. Langton, "Intuitive guide to principles of communications: all about modulation: Part II," 2002, http://complextoreal.com/.

[29] "Wireless Communications Laboratory," National Chung Cheng University, Taiwan.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 242483, 9 pages
http://dx.doi.org/10.1155/2013/242483

Research Article

Modeling and Performance Analysis of Route-Over and Mesh-Under Routing Schemes in 6LoWPAN under Error-Prone Channel Condition

Tsung-Han Lee,[1] Hung-Chi Chu,[2] Lin-Huang Chang,[1] Hung-Shiou Chiang,[1] and Yen-Wen Lin[1]

[1] *Department of Computer Science, National Taichung University of Education, No. 140, Minsheng Road, West District, Taichung 40306, Taiwan*
[2] *Department of Information and Communication Engineering, Chaoyang University of Technology, No. 168, Ji-Fong East Road, Wu-Fong District, Taichung 41349, Taiwan*

Correspondence should be addressed to Tsung-Han Lee; thlee@mail.ntcu.edu.tw

Received 21 June 2013; Accepted 20 August 2013

Academic Editor: Chih-Hao Lin

6LoWPAN technology has attracted extensive attention recently. It is because 6LoWPAN is one of Internet of Things standard and it adapts to IPv6 protocol stack over low-rate wireless personal area network, such as IEEE 802.15.4. One view is that IP architecture is not suitable for low-rate wireless personal area network. It is a challenge to implement the IPv6 protocol stack into IEEE 802.15.4 devices due to that the size of IPv6 packet is much larger than the maximum packet size of IEEE 802.15.4 in data link layer. In order to solve this problem, 6LoWPAN provides header compression to reduce the transmission overhead for IP packets. In addition, two selected routing schemes, mesh-under and route-over routing schemes, are also proposed in 6LoWPAN to forward IP fragmentations under IEEE 802.15.4 radio link. The distinction is based on which layer of the 6LoWPAN protocol stack is in charge of routing decisions. In route-over routing scheme, the routing distinction is taken at the network layer and, in mesh-under, is taken by the adaptation layer. Thus, the goal of this research is to understand the performance of two routing schemes in 6LoWPAN under error-prone channel condition.

1. Introduction

6LoWPAN [1, 2] is an IETF standardized IPv6 adaptation layer that allows IP over low-power, lossy networks. Extending IP to LoWPANs (low-power, wireless personal area networks) faces different challenges than traditional network. The microcontrollers typically embedded with LoWPAN radios have limited memory and compute power. Thus, the technique for transmitting IPv6 packets over Low-power Wireless Personal Area Networks is called 6LoWPAN. However, 6LoWPAN is difficult to implement because the size of IPv6 packet is much larger than the packet size of IEEE 802.15.4 data link layer. In order to make it possible, the IETF 6LoWPAN working group introduces the adaptation layer between network and data link layers. It provides header compression to reduce transmission overhead, fragmentation, and reassembly of IPv6 packet. It can also be involved in routing decisions, and the routing scheme in 6LoWPAN can be divided into two categories. In the mesh-under method, the routing decision is taken in adaptation layer. On the other hand, the route-over method makes the routing decision in network layer.

Mesh-under and route-over routing schemes can be considered as end-to-end and hop-by-hop transmission, respectively. Hop-by-hop fragmentation and reassembly generate more delay but achieve better fragment arrival ratio. Whereas end-to-end scheme has less latency, but fragment loss has high probability. Therefore, the goal of this research is to understand the performance of two routing schemes in 6LoWPAN under error-prone channel condition.

In 6LoWPAN protocol stack, the last two layers are based on IEEE 802.15.4 physical and data-link layers. A two-dimensional discrete-time Markov chain for unslotted CSMA/CA mechanism in 6LoWPAN is used to analyze the performance in unsaturated 6LoWPAN for rout-over and mesh-under routing schemes. Based on this Markov chain, the packet successful transmission probability for route-over and mesh-under routing schemes is evaluated under different numbers of competing nodes and wireless channel condition. To the best of our knowledge, this study is not yet reported in the literature. Finally, we attempt to compare route-over against mesh-under routing schemes in 6LowPAN in terms of goodput for IP end-to-end communication.

The rest of the paper is organized as follows. Next section describes related research works. Section 3 introduces the 6LoWPAN unslotted CSMA/CA mechanism. Section 4 describes the mesh-under and the route-over routing schemes in 6LoWPAN. Section 5 describes the proposed two-dimensional discrete-time Markov chain to model 6LowPAN unslotted CSMA/CA mechanism and analyzes the goodput for 6LowPAN routing schemes under error-prone channel condition. Section 6 shows the analysis results for probability of IP packet successful transmission in both route-over and mesh-under routing schemes. Finally, Section 7 presents our conclusions.

2. Related Work

In the case of the WLAN, Bianchi [3] presented an analytical model to compute the saturation throughput performance evaluation of both RTC/CTS access mechanisms.

In WSN, [4] presented a similar analytical model to predict energy consumption as well as the throughput of saturated and unsaturated 802.15.4 networks, based on which some design guidelines can be derived. In order to address system goodput and energy efficiency enhancement, [5] studies packet size optimization for IEEE 802.15.4 networks. Taking into account the CSMA-CA contention, protocol overhead, and channel condition, analytical models are proposed to calculate the goodput and the energy consumption. In [6], the authors try to analyze the complete CSMA/CA in IEEE 802.15.4. First, to analyse the performance of the slotted CSMA/CA of IEEE 802.15.4 by integrating the discrete-time Markov chain models of the node states and the channel states and then extend the models by adopting a modification to the CAP. The extended models could be used to analyze the performance of the unslotted and slotted CSMA/CA strategy. In nonbeacon-enabled mode [7], build a process chain to model unslotted CSMA/CA mechanism. However, the backoff procedure is not only a Markov chain but also the backoff time counter is an accumulation which value depends on how many times the node has tried to access the channel without success. According to the proposed process chain and mathematical model, the distribution of traffic changes has been estimated when different loads are offered to the network. Moreover, the proposed model can evaluate the proper packet size to improve the success probability.

In [8], an analytical comparison between route-over and mesh-under schemes based on 6LoWPAN experimental research which tries to analyze the fragment arrival probability, the total number of transmissions and the total delay from source and destination. However, the authors analyze the arrival probability for single fragment in multihop environment. Furthermore, they assume that the fragment arrival probability for one hop distance is fixed, which means that the number of competing nodes in each hop has not been concerned. In our research, the probability of successful transmission for each fragment is different depending on the contention window in MAC layer for IEEE 802.15.4. And we analyze the complete IP packet arrival probability including all IP fragments for route-over and mesh-under.

3. Overview of CSMA/CA Mechanism in 6LowPAN

6LoWPAN protocol stack adopts IEEE 802.15.4 standard PHY and MAC layers which are specified in [1, 2]. Figure 1 [9] illustrates the steps of unslotted CSMA/CA algorithm of 6LoWPAN. *NB* is the number of times the CSMA/CA algorithm was required to back off while attempting the current transmission (the value will be initialized to 0). In the first-time backoff, the range of random contention window (CW) is from $[0, 2^{\text{macMinBE}} - 1]$. Here macMinBE is the minimum value of the backoff exponent (BE) that has the default value which is 3. When backoff counter reaches 0, the node performs channel sensing immediately.

The random backoff mechanism is used to decrease the probability of collisionsand ensure that the channel is clear for a node to access it. The channel clear assessment in unslotted CSMA/CA is one backoff period (in slotted CSMA/CA, which performs two channel clear assessments before transmission). If the channel is detected to be busy, BE is increased by 1, and the new backoff stage begins before channel sensing. This process is repeated until BE equals upperbounded macMaxBE (maximum value of BE, the default is 5), and then the BE is frozen at macMaxBE. When the number of backoff stage is equal to macMaxCSMABackoffs (the default value is 4), the node access channel is failure.

4. Routing Scheme in 6LoWPAN

To enable the transmission of large IPv6 packets over size constrained link layer payload size (102 bytes of payload) in IEEE 802.15.4, the 6LoWPAN adaptation layer provides IP packet fragmentation mechanism [1]. All fragments are transmitted into multiple link-layer frames for reassembling them at the other end under the mesh-under or route-over routing scheme in 6LoWAN.

As mentioned in the previous section, 6LoWPAN divides routing schemes into mesh-under and route-over [2, 8] schemes. The distinction is based on which layer of the 6LoWPAN protocol stack is in charge of routing decisions; in route over they are taken at the network layer, and in mesh under at the adaptation layer. Figure 2 shows routing decision

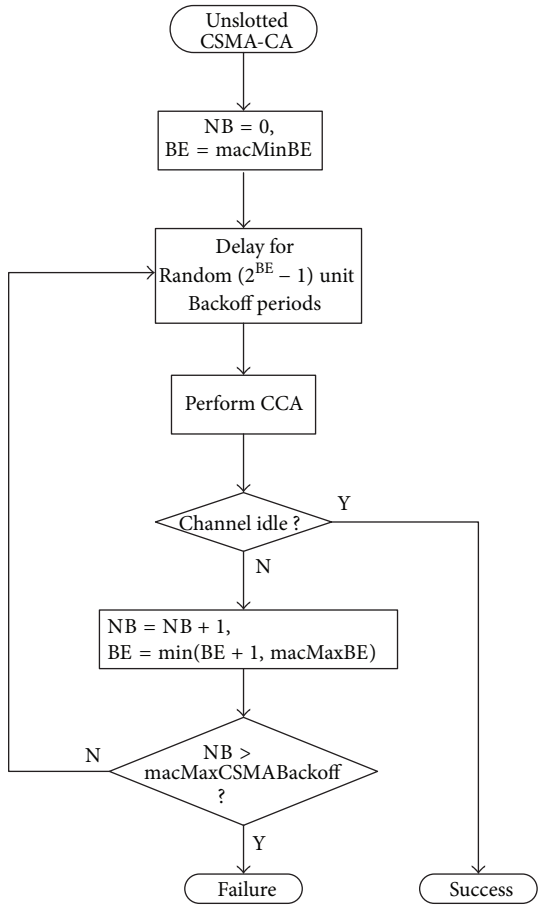


Figure 1: 6LoWPAN unslotted CSMA/CA algorithm [9].

layer for both mesh-under and route-over routing schemes [1].

4.1. Mesh-Under Routing Scheme. In the mesh-under routing scheme, the routing functions are placed at the link layer based on IEEE 802.15.4 frame structure and the 6LoWPAN header [2, 8]. All fragments will be sent to the next hop by mesh routing and finally reach to the destination. Different fragments of one IP packet might reach the destination via different route-paths. If all fragments are received at the destination successfully, the destination's adaptation layer reassembles all fragments into an IP packet. The adaptation layer of destination node starts reconstruction process. However, any fragment is missing in forwarding process; all fragments of this IP packet are retransmitted from the source to the destination.

4.2. Route-Over Routing Scheme. In route-over scheme, each sensor node inside the route path acts as an IP router. The IP packet is forwarded hop by hop from the source node to the destination node [2, 8]. The IP packet's payload is encapsulated with IPv6 header. After that, IP packet is fragmented by the adaptation layer and all IP fragments will be sent to the next hop based on routing table. The next hop has to reassemble them in order to reconstruct the original IP packet in adaptation layer when all fragments are received successfully. The reconstruction process starts only when the last fragment arrives. Once reconstructed, the IP packet will be sent to the network layer. Finally, the IP packet will be fragmented again and these fragments will be delivered to the nexthop. However, the retransmission executes only in one-hop distance if there is any fragment lost in this forwarding process.

5. Numerical Analytical Model for 6LoWPAN Routing Schemes in Error-Prone Channel Condition

In this section, we propose mathematical models to analyze the IP packet successful transmission probability for route-over and mesh-under, respectively. In addition, we present the goodput analysis to compare the performance of these two routing schemes under error-prone channel condition.

5.1. Markov Chain Model for Unslotted CSMA/CA Mechanism in Error-Prone Channel Condition. In Figure 3, a two-dimensional discrete-time Markov chain model has been used to analyze the unslotted CSMA/CA mechanism in 6LoWPAN under error-prone channel condition. Define the

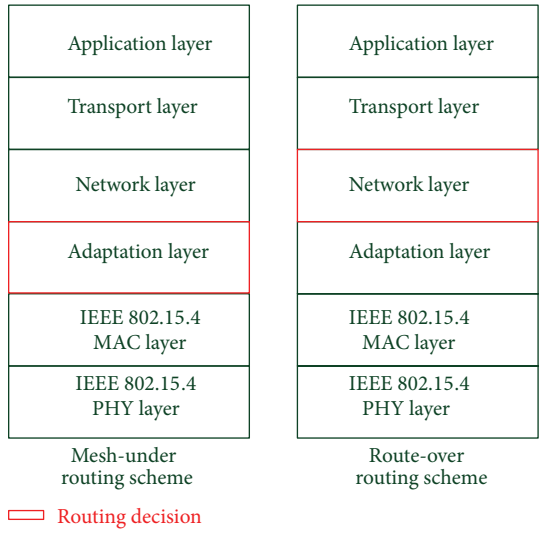


Figure 2: Routing decision layer for both mesh-under and route-over routing schemes in 6LoWPAN [1].

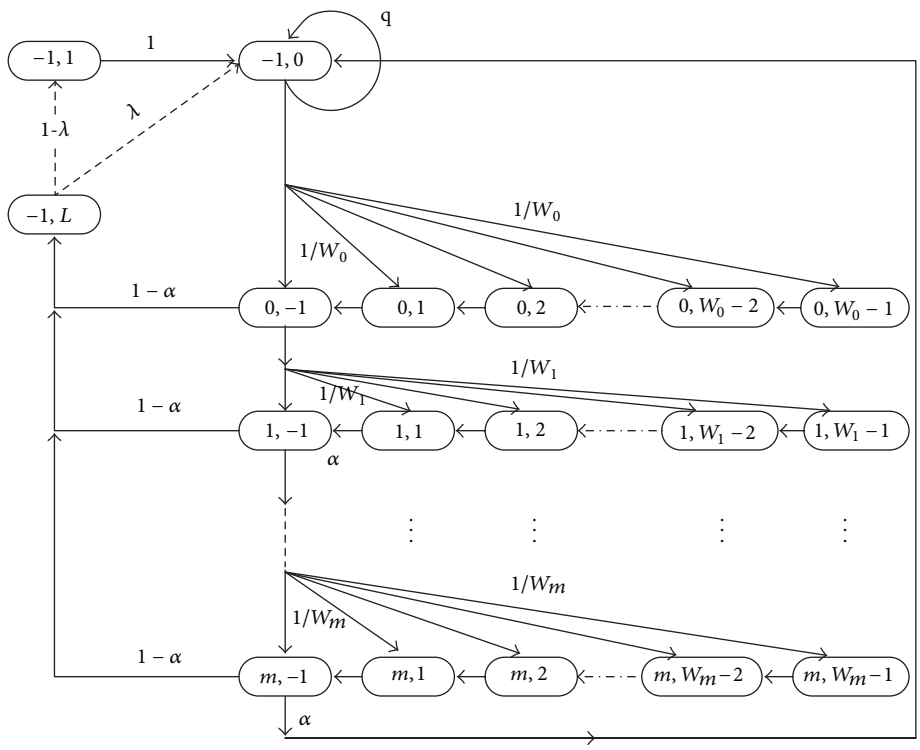


Figure 3: The Discrete-time Markov chain model for unslotted CSMA/CA mechanism in 6LoWPAN under error-prone channel condition.

state as $\{nb(t), bc(t)\}$; $nb(t), bc(t)$ as the stochastic process representing the number of backoff times and the backoff counter at time slot t, respectively. When $nb(t) \in [0, m]$, m is the maximum number of backoff stage which is equal to macMaxCSMABackoffs (the default value is 4); $bc(t) \in [0\ ,\ W_i - 1]$. Note that when $bc(t) = 0$, the node enters CCA state immediately. Thus, $bc(t) = 0$ is replaced by $bc(t) = -1$ which represents the successful CCA attempt. In addition, $\{0, -1\} \sim \{m, -1\}$ represents the successful CCA state for backoff stage from 0 to the maximum backoff number m. According to the protocol, the duration of the backoff counter is

$$W_i = 2^{\text{macMinBE}} 2^{\min\{\text{macMaxBE}-\text{macMinBE}, i\}}, \quad i \in [0, m] \tag{1}$$

α is the probability that the channel is busy at the CCA detection. When $(t) = -1$, it represents the successful transmission attempt. The range of $nb(t)$ is from $\{-1, L\}$ to $\{-1, 0\}$, where L is the number of time slots when a packet is under the transmission duration. Hence, the length of a

transmission period must equal the length of packet. We propose an interference model in the transmission process based on this Markov chain, it occurs when the transmission process. We let λ be defined as frame error rate (FER). Bit error rate (BER) is used to indicate the wireless channel condition. Thus, the given certain BER value, λ, can be calculated by

$$\lambda = 1 - (1 - \mathrm{BER})^{l} \tag{2}$$

$$l = L \cdot T_{\mathrm{Slot}} \cdot R, \tag{3}$$

where T_{Slot} is a aUnit Backoff Period and R is a physical layer bit rate. Equation (2) is shown that the different channel conditions and the length of packet transmission will impact the FER. While the value of λ is too high, the packet transmission error, a node, will back to the idle state $\{-1, 0\}$. And q is defined as the probability that the user is still in the idle state in the next time slot.

Let the stationary probabilities of the Markov chain be $b_{n,b} = P\{(nb(t), bc(t)) = (n, b)\}$. Note that backoff counter reaches 0 and the node enters CCA state immediately. Hence, $b_{n,0}$ has a same value as $b_{n,-1}$. We obtain that

$$\begin{gathered} b_{n,-1} = \alpha^{n} b_{0,-1}, \quad n \in [1, m], \\ b_{-1,L} = \left(1 - \alpha^{m+1}\right) b_{0,-1}, \\ b_{-1,0} = \frac{b_{0,-1}}{1 - q}. \end{gathered} \tag{4}$$

The sum of probabilities of all the states should be equal to 1, and we have

$$\begin{aligned} & b_{-1,0} + \sum_{l=1}^{L} b_{-1,l} + \sum_{n=1}^{m} b_{n,-1} \\ & \quad + b_{0,-1} + \sum_{n=0}^{m} \sum_{b=1}^{W_n - 1} b_{n,b} = 1, \end{aligned} \tag{5}$$

where τ is the probability that a node attempts carrier sensing; we get

$$\tau = \sum_{n=0}^{m} b_{n,-1}. \tag{6}$$

Assume that the system has N nodes. From [4], the probability that the channel is busy in CCA is

$$\begin{aligned} \alpha &= \left[1 - \frac{1}{1\left(1 + 1/\left(1 - (1-\tau)^{N}\right)\right)}\right]\left(1 - (1-\tau)^{N-1}\right) \\ &= \frac{1 - (1-\tau)^{N-1}}{2 - (1-\tau)^{N}}. \end{aligned} \tag{7}$$

According to the proposed Markov chain model, we can get the probability to enter transmission stage which is

$$P_{\mathrm{tr}} = \tau \cdot (1 - \alpha). \tag{8}$$

While a node access channel is successful, it will transmit data, and the transmission task is completed in data link layer. But this transmission cannot ensure that the packet arrival to receiver is correctly. It is possible occurring interference in air propagation. Hence, a successful transmission will not have any FER from sender to receiver. We can get

$$P_{\mathrm{suc}} = P_{\mathrm{tr}} \cdot (1 - \lambda). \tag{9}$$

5.2. IP Packet Successful Transmission Probability for Route-Over and Mesh-Under. Consider

$$P_{\mathrm{suc}}^{\mathrm{mu}} = \left\{\left[\sum_{i=0}^{k} \left(P_{\mathrm{suc}}\right)\left(1 - P_{\mathrm{suc}}\right)^{i}\right]^{f}\right\}^{h}. \tag{10}$$

Let k represent macMaxFrameRetries which has default value which is 3, and f is the number of fragments with hop counts h. In mesh-under scheme, all fragments must aggregate at destination. Each fragment is sent from source to the destination in h hops. Thus, the IP packet successful transmission probability decreases gradually after f number of fragments through h hops transmission route-path:

$$P_{\mathrm{suc}}^{\mathrm{ro}} = \left[\sum_{i=0}^{k} \left(P_{\mathrm{suc}}\right)\left(1 - P_{\mathrm{suc}}\right)^{i}\right]^{f}. \tag{11}$$

Equation (11) is the IP packet successful probability of 6LoWPAN route-over routing scheme. The major feature of route-over is hop-by hop fragmentation and reassembly. In each hop, all fragments will recover to a completed IPv6 packet. Moreover, in the modeling assumption, each hop contents ideal channel condition. Consequently, we can consider route-over scheme as hop-by-hop forwarding from source to destination. Although this scheme consumes more energy and delay time, but it brings the robust packet transmission rate.

5.3. Goodput Analysis for 6LoWPAN Routing Schemes. To evaluate the goodput, we consider that a cycle of transmission includes idle, contention, and transmission states. These states define as the duration, and each one is normalized which contains probability. The equations are shown as follows.

$$E[\mathrm{idle}] = \left(1 - P_{\mathrm{tr}}\right)^{N} \cdot \sigma, \tag{12}$$

where σ is defined as the duration of an empty slot time and N is number of competing nodes. Equation (12) is the expectation of node's idle duration. If any competing node generates frame to transmit, it will transit to contention state. The expectation of a node in contention state is

$$\begin{aligned} E[\mathrm{wait}] = {} & P_{\mathrm{tr}} \cdot \left(7 \times T_{\mathrm{Slot}} + T_{\mathrm{CCA}}\right) \\ & + P_{\mathrm{tr}} \cdot \left(1 - P_{\mathrm{tr}}\right) \cdot \left(23 \times T_{\mathrm{Slot}} + T_{\mathrm{CCA}}\right) \\ & + P_{\mathrm{tr}} \cdot \left(1 - P_{\mathrm{tr}}\right)^{2} \cdot \left(55 \times T_{\mathrm{Slot}} + T_{\mathrm{CCA}}\right) \\ & + P_{\mathrm{tr}} \cdot \left(1 - P_{\mathrm{tr}}\right)^{3} \cdot \left(87 \times T_{\mathrm{Slot}} + T_{\mathrm{CCA}}\right) \\ & + P_{\mathrm{tr}} \cdot \left(1 - P_{\mathrm{tr}}\right)^{4} \cdot \left(119 \times T_{\mathrm{Slot}} + T_{\mathrm{CCA}}\right). \end{aligned} \tag{13}$$

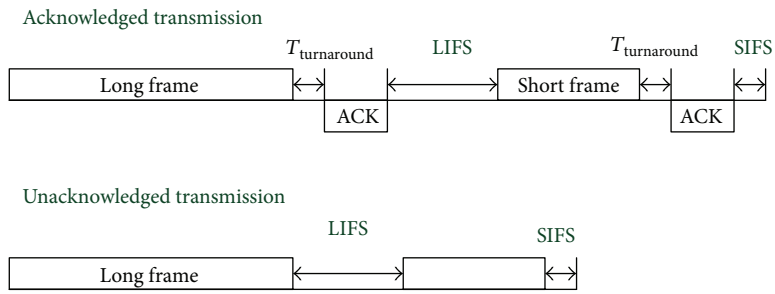


FIGURE 4: The procedure of data transmission in 6LoWPAN using the acknowledged and unacknowledged transmission.

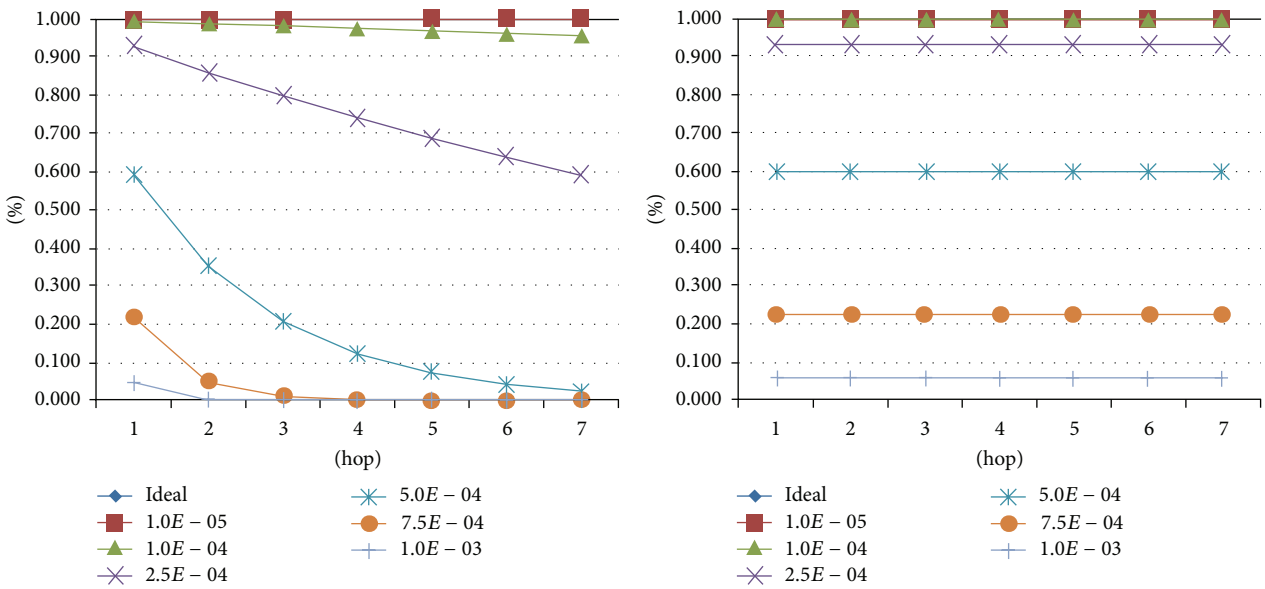


FIGURE 5: The successful IP packet transmission probability with 3 competing nodes for mesh-under and route-over routing schemes in 6LowPAN under error-prone channel condition.

In (13), a timeslot and CCA durations are 320 μs and 128 μs, respectively. In the first backoff stage, the maximum number of slots for backoff counter is 8 (from 0 to 7). If access channel is failure in first stage, it will enter to next stage and backoff again. Since the main difference between route-over and mesh-under schemes is hop-by-hop fragmentation and reassembly, it means that the transmission delay in route-over scheme is higher than mesh-under scheme. Thus, we define δ is the delay time for the IP packet assembly and reassembly for route-over scheme. The expectation of transmission state for these two schemes is different. Moreover, we also consider the transmission can be with and without retransmission, respectively. From [9], the duration of transmission state for two types is shown in Figure 4.

Figure 4 illustrates the procedure of data transmission in 6LoWPAN using the acknowledged and unacknowlwdged transmission, respectively. The length of the IFS depends on the size of the frame that has been transmitted. The packet length greater than 18 bytes will be followed by a long IFS (LIFS is 40 symbols) and short frames by a short IFS (SIFS is 12 symbols). Thus, a mathematical analysis on transmission performance for the 6LoWPAN can be expressed in Acknowledged and Unacknowledged types.

5.3.1. Goodput Analysis for Unacknowledged Transmission. The goodput analysis model for unacknowledged transmission in both route-over and mesh-under routing schemes is presented in this section. We first obtain the expected transmission time for route-over and mesh-under routing schemes.

Consider

$$E_{\text{noack}}^{\text{ro}}[\text{tx}] = P_{\text{suc}} \cdot (T_{\text{data}} + T_{\text{LIFS}} + \delta),$$
$$E_{\text{noack}}^{\text{mu}}[\text{tx}] = P_{\text{suc}} \cdot (T_{\text{data}} + T_{\text{LIFS}}). \quad (14)$$

Equations (14) present the expected transmission time for route-over and mesh-under routing schemes, respectively. T_{data} represents the transmission time for one IP fragment. T_{LIFS} is LIFS period which is 640 μs. SIFS is used when the MPDU is smaller than or equal to 18 bytes. In addition, δ is the process delay for IP packet fragmentation and reassembly IP packet. We assume that the δ is around 2 ms.

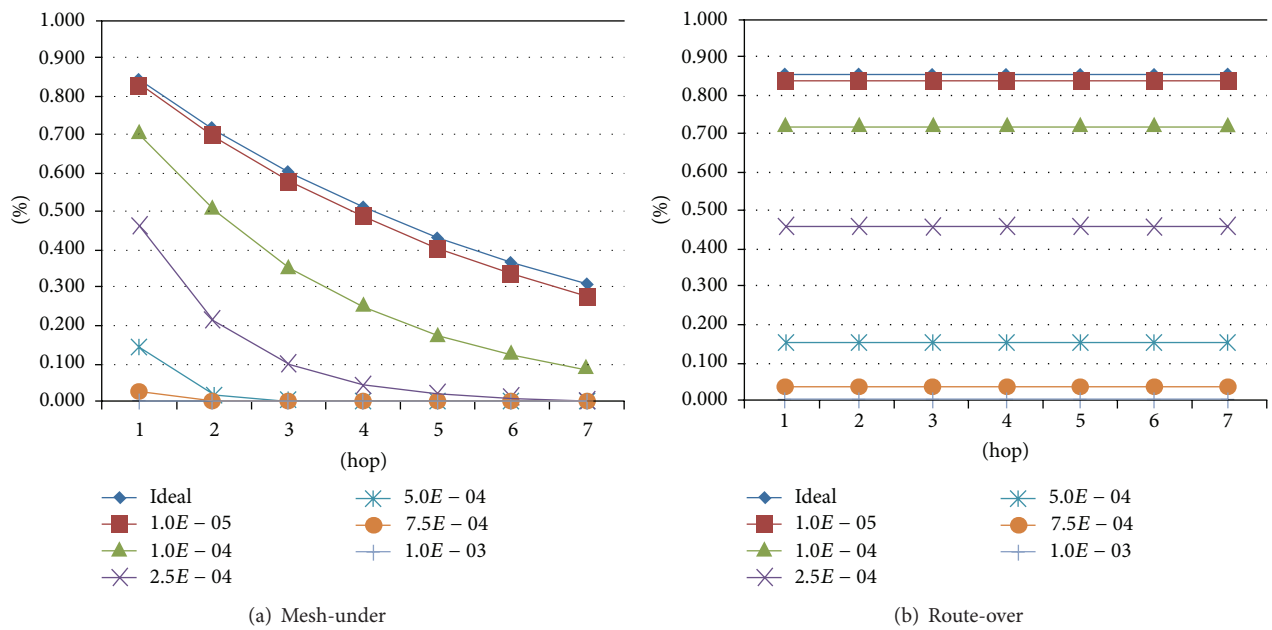


Figure 6: The successful IP packet transmission probability with 5 competing nodes for mesh-under and route-over routing schemes in 6LowPAN under error-prone channel condition.

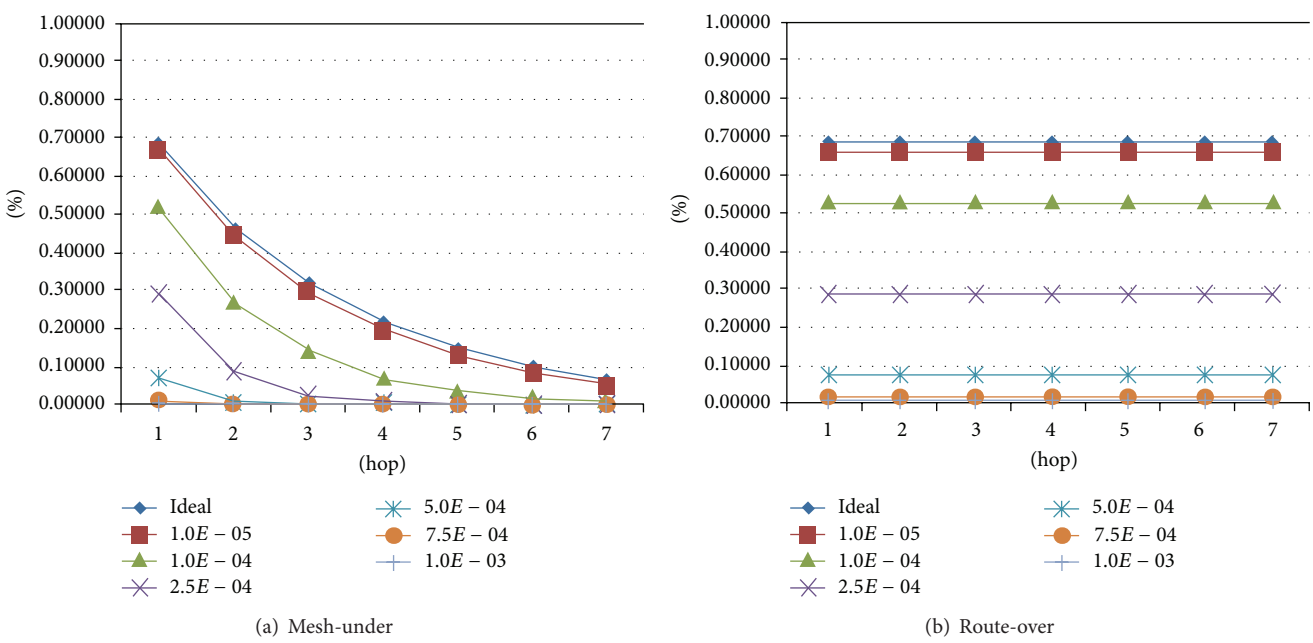


Figure 7: The successful IP packet transmission probability with 7 competing nodes for mesh-under and route-over routing schemes in 6LowPAN under error-prone channel condition.

Thus, the transmission goodput of route-over and mesh-under routing schemes can be obtained from (15):

$$S_{\text{noack}}^{\text{ro}} = \frac{P_{\text{suc}}^{\text{ro}} \cdot \text{IP}_{\text{payload}}}{\left(E\left[\text{idle}\right] + E\left[\text{wait}\right] + E_{\text{noack}}^{\text{ro}}\left[\text{tx}\right]\right) \cdot f \cdot h},$$
$$S_{\text{noack}}^{\text{mu}} = \frac{P_{\text{suc}}^{\text{mu}} \cdot \text{IP}_{\text{payload}}}{\left(E\left[\text{idle}\right] + E\left[\text{wait}\right] + E_{\text{noack}}^{\text{mu}}\left[\text{tx}\right]\right) \cdot f \cdot h}, \quad (15)$$

where $P_{\text{suc}}^{\text{ro}}$ and $P_{\text{suc}}^{\text{mu}}$ are the successful transmission probability for route-over and mesh-under routing schemes. f and h are the number of fragmentations and hop count from source to destination, respectively. $E[\text{idle}]$, $E[\text{wait}]$, and $E[\text{tx}]$ represent the expected idle time, competing time, and transmission time, respectively.

5.3.2. Goodput Analysis for Acknowlwdged Transmission. In this section, we first obtain the expected transmission time

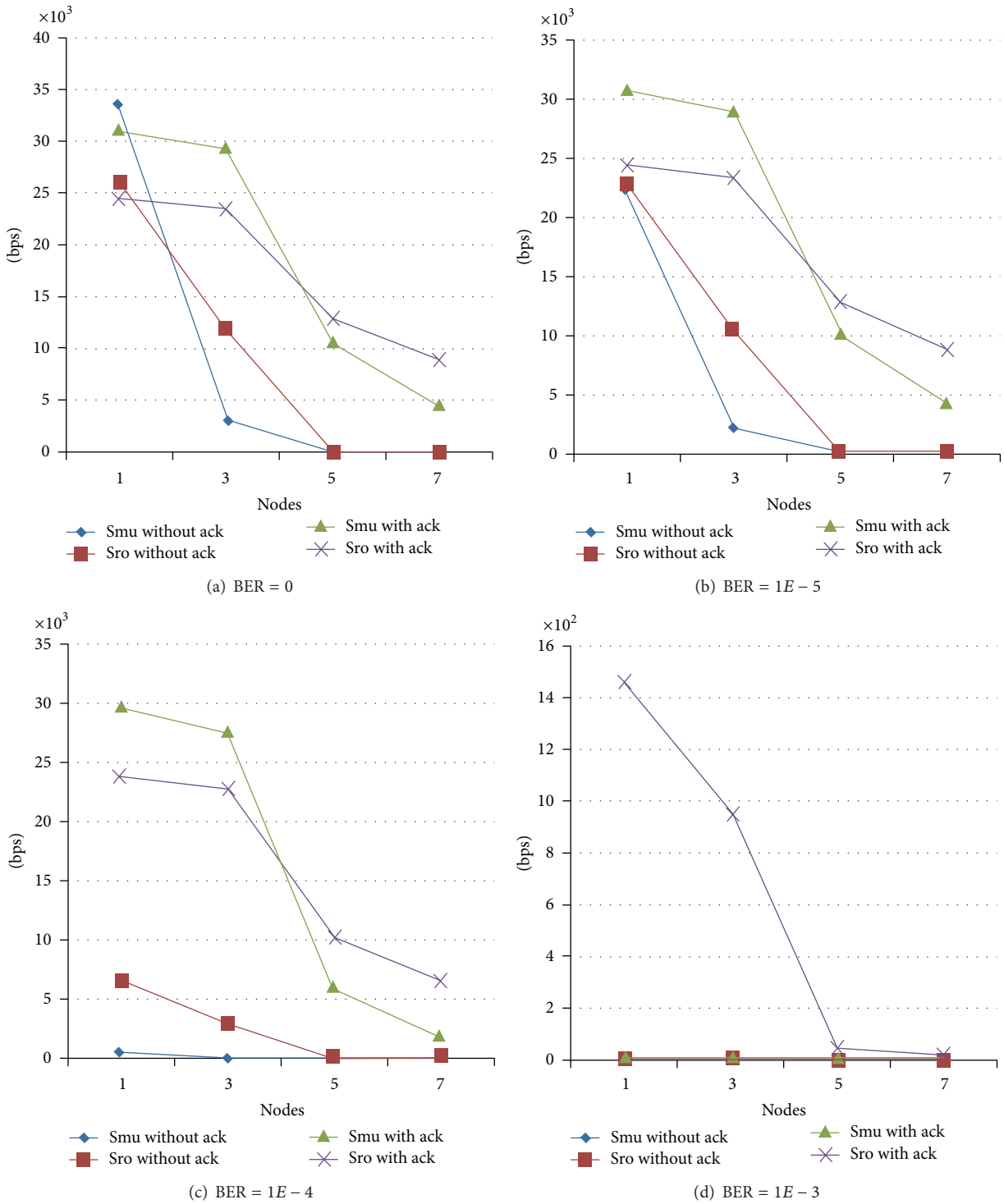


FIGURE 8: Goodput for mesh-under and route-over routing schemes in 6LowPAN under error-prone channel condition.

for route-over and mesh-under routing schemes for acknowledged transmission:

$$E_{\text{ack}}^{\text{ro}}[\text{tx}] = P_{\text{suc}} \cdot \left(T_{\text{data}} + T_{\text{turn around}} + T_{\text{ack}} + T_{\text{LIFS}} + \delta\right),$$
$$E_{\text{ack}}^{\text{mu}}[\text{tx}] = P_{\text{suc}} \cdot \left(T_{\text{data}} + T_{\text{turn around}} + T_{\text{ack}} + T_{\text{LIFS}}\right). \quad (16)$$

T_{ack} and $T_{\text{turn around}}$ are the acknowledgement transmission time (352 μs), and turnaround time (192 μs), respectively. If there is no acknowledgement, then turnaround time $T_{\text{turn around}}$ and T_{ack} is equal to zero. Finally, the goodput of acknowledged transmission for route-over and mesh-under routing schemes in error-prone wireless environment can be

Table 1: Numerical evaluation parameters.

IPv6 packet size	1280 bytes
Number of fragments	14
Number of competing nodes	3, 5, and 7
Hop counts	From 1 to 7
BER	From 0 to 1E-3

computed as follows:

$$S_{\text{ack}}^{\text{ro}} = P_{\text{suc}}^{\text{ro}} \cdot \text{IP}_{\text{payload}} \times \left(\left[\sum_{i=0}^{k} (1 - P_{\text{suc}})^{i} \cdot (E[\text{idle}] + E[\text{wait}] + E^{\text{ro}}[\text{tx}]) \right] \cdot f \cdot h \right)^{-1},$$
$$S_{\text{ack}}^{\text{mu}} = P_{\text{suc}}^{\text{mu}} \cdot \text{IP}_{\text{payload}} \times \left(\left[\sum_{i=0}^{k} (1 - P_{\text{suc}})^{i} \cdot (E[\text{idle}] + E[\text{wait}] + E^{\text{mu}}[\text{tx}]) \right] \cdot f \cdot h \right)^{-1}, \quad (17)$$

where macMaxFrameRetriesis 3, f, and h are the number of fragmentations and hop count from source to destination, respectively.

6. Performance Emulation Results for 6LoWPAN Routing Schemes in Error-Prone Channel Condition

In this section, we present the numerical analysis results for 6LoWPAN routing schemes in error-prone channel condition. Our probabilistic model was emulated by PRISM [10]. Let the IPv6 packet size be minimum MTU (1280 bytes), and it will be fragmented into 14 maximum IEEE 802.15.4 frames (133 bytes), so called fragments. Assume f fragments were sent from source to destination through h hop counts. The evaluation parameters are shown in Table 1.

From results in figures 5, 6, and 7, we observed that the number of competing nodes increases, the probability of successful transmission for two routing schemes are decreases due to the busy channel condition. The result shows that the rout-over routing is beneficial compared to the mesh-under routing scheme since it reduces the probability of collisions from competing nodes and hop count. It is because that the transmission probability would resume to 1 due to that the route-over scheme reassembled all fragments for each hop. Therefore, the IP packet transmission successful probability would not decrease after multihops routing path. However, route-over consumes more delay time from hop-by-hop fragments assembly and reassembly.

In Figure 8, as we can observe from both route-over and mesh-under routing schemes, the critical influence on the goodput in both schemes is the channel condition.

7. Conclusions

In this paper, we investigate the 6LoWPAN transmission performance by using the proposed mathematical model in 6LowPAN under varying number of competing nodes and error-prone channel condition. Analysis results show that route-over scheme has higher transmission probability than mesh-under.

Acknowledgments

This research was partially supported by the National Science Council of Republic of China, Taiwan under contracts, NSC102-2221-E-142-005 and NSC 101-2119-M-142-001 as well as National Taichung University regarding the MoE project (No. 1020035480A).

References

[1] Z. Shelby and C. Bormann, *6LoWPAN: The Wireless Embedded Internet*, vol. 43 of *Wiley Series on Communications Networking & Distributed Systems*, John Wiley & Sons, New York, NY, USA, 2009.

[2] A. Ludovici, A. Calveras, and J. Casademont, "Forwarding techniques for IP fragmented packets in a real 6LoWPAN network," *Sensors*, vol. 11, no. 1, pp. 992–1008, 2011.

[3] 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.

[4] S. Pollin, M. Ergen, S. C. Ergen et al., "Performance analysis of slotted carrier sense IEEE 802.15.4 medium access layer," *IEEE Transactions on Wireless Communications*, vol. 7, no. 9, pp. 3359–3371, 2008.

[5] Y. Zhang and F. Shu, "Packet size optimization for goodput and energy efficiency enhancement in slotted IEEE 802.15.4 networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '09)*, pp. 1–6, Budapest, Hungary, April 2009.

[6] F. Wang, D. Li, and Y. Zhao, "Analysis of CSMA/CA in IEEE 802.15.4," *IET Communications*, vol. 5, no. 15, pp. 2187–2195, 2011.

[7] C. Buratti and R. Verdone, "Performance analysis of IEEE 802.15.4 non beacon-enabled mode," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 7, pp. 3480–3493, 2009.

[8] A. H. Chowdhury, M. Ikram, H. S. Cha et al., "Route-over vs mesh-under routing in 6LoWPAN," in *Proceedings of the ACM International Wireless Communications and Mobile Computing Conference (IWCMC '09)*, pp. 1208–1212, June 2009.

[9] IEEE 802.15 Work Group, "Part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (LR-WPANs)," ANSI/IEEE Std 802.15.4, 2006.

[10] PRISM Website, http://www.prismmodelchecker.org/.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 953048, 10 pages
http://dx.doi.org/10.1155/2013/953048

Research Article

Calculation of Weighted Geometric Dilution of Precision

Chien-Sheng Chen,[1] Yi-Jen Chiu,[2] Chin-Tan Lee,[3] and Jium-Ming Lin[4]

[1] *Department of Information Management, Tainan University of Technology, Tainan, Taiwan*
[2] *Department of Digital Entertainment and Game Design, Taiwan Shoufu University, Tainan, Taiwan*
[3] *Department of Electronic Engineering, National Quemoy University, Kinmen, Taiwan*
[4] *Department of Communication Engineering, Chung-Hua University, Hsinchu, Taiwan*

Correspondence should be addressed to Chien-Sheng Chen; t00243@mail.tut.edu.tw

Received 20 April 2013; Revised 28 August 2013; Accepted 3 September 2013

Academic Editor: Anyi Chen

To achieve high accuracy in wireless positioning systems, both accurate measurements and good geometric relationship between the mobile device and the measurement units are required. Geometric dilution of precision (GDOP) is widely used as a criterion for selecting measurement units, since it represents the geometric effect on the relationship between measurement error and positioning determination error. In the calculation of GDOP value, the maximum volume method does not necessarily guarantee the selection of the optimal four measurement units with minimum GDOP. The conventional matrix inversion method for GDOP calculation demands a large amount of operation and causes high power consumption. To select the subset of the most appropriate location measurement units which give the minimum positioning error, we need to consider not only the GDOP effect but also the error statistics property. In this paper, we employ the weighted GDOP (WGDOP), instead of GDOP, to select measurement units so as to improve the accuracy of location. The handheld global positioning system (GPS) devices and mobile phones with GPS chips can merely provide limited calculation ability and power capacity. Therefore, it is very imperative to obtain WGDOP accurately and efficiently. This paper proposed two formations of WGDOP with less computation when four measurements are available for location purposes. The proposed formulae can reduce the computational complexity required for computing the matrix inversion. The simpler WGDOP formulae for both the 2D and the 3D location estimation, without inverting a matrix, can be applied not only to GPS but also to wireless sensor networks (WSN) and cellular communication systems. Furthermore, the proposed formulae are able to provide precise solution of WGDOP calculation without incurring any approximation error.

1. Introduction

In positioning the location estimates are determined through the received signals transmitted by the mobile devices at a set of base stations (BSs), satellites, or other sensors. First, the length or direction of the radio path is determined through signal measurements. Secondly, the MS position is derived from radio location algorithms and known geometric relationships. Mobile positioning systems have received significant attention, and various location technologies have been proposed in the past few years. Among the techniques for mobile positioning there are two major categories—handset-based and network-based schemes. Both approaches have their advantages and limitations. Global positioning system (GPS) is a positioning system that can provide position, velocity, and time information to a user. Handset-based solutions generally require a handset modification to calculate its own position when they are fully or partially equipped with a GPS receiver. The advantages of using handset-based methods are that they have global coverage and usually provide much more accurate location measurements. The drawbacks of the handset-based methods include cost, redundant hardware, and economical integrated technology. The reliability of GPS measurements is greatly compromised in a building or shadowed environments, where direct line-of-sight (LOS) propagation is not available. Without the aid of satellite systems, network-based positioning schemes use time and angle measurements to determine the MS location

or to assist the process of MS location determination. Instead of using all seven BSs, four BSs with better geometry are good enough to provide sufficient measurements for positioning in cellular communication networks. The network-based location schemes are relatively less complex on hardware when compared with the handset-based methods. They can be employed in many situations where GPS signal cannot, for example, indoor environment and urban canyon areas, or when GPS-embedded handsets are not available. For many applications in wireless sensor networks (WSN), like environmental sensing and activities measuring, it is crucial to know the locations of the sensor nodes in network-based positioning; this is known as a "localization problem" [1]. An ideal location technology should be able to provide a robust estimate of location in all environments.

This paper considers both the network-based method and the handset-based method, employing the concept of geometric dilution of precision (GDOP), which was initially developed as a criterion for selecting the optimal 3D geometric configuration of satellites in GPS. The general object of the GPS satellite selection algorithm is to minimize the GDOP to improve the position accuracy. The smaller value of GDOP is calculated, the better geometric configuration we will have. The redundant measurements will bring large amount of computation and may not provide significantly improved location accuracy. When enough measurements are available, the optimal measurements selected with the minimum GDOP can prevent the poor geometry effects and have the potential of obtaining greater location accuracy.

There have been extensive researches trying to obtain an approximate GDOP value without executing matrix inversion in the past few years. Simon and El-Sherief [2, 3] proposed the employment of back-propagation neural network (BPNN) [4] to obtain an approximation for the GDOP function. The BPNN is employed to learn the relationship between the entries of a measurement matrix and the eigenvalues of its inverse. Three other input-output relationships were proposed in [5]. We present the resilient back-propagation (Rprop) architectures to obtain the approximate GDOP [6]. The matrix inversion method for GDOP calculation is born with significant computational burden. GDOP is approximately inversely proportional to the volume of the tetrahedron formed by the tips of four unit vectors directed to the selected satellites in GPS [7]. The four satellites evenly distribute with the maximum volume which brings the more accurate location estimation. The maximum volume method requires low computing time in selecting a subset with the largest tetrahedron as the optimum [8]. However, it is not suitable to use this method because it may not select the desired satellites with the minimum GDOP. The main disadvantage of these methods is to incur approximation errors. To avoid these disadvantages, a simple closed-form formula for GDOP calculation is proposed in [9].

Traditionally, the GDOP computation assumes that the pseudorange errors are independent and identical [10]. Several methods based on GDOP have been proposed to improve the GPS positioning accuracy [7, 9, 11]. In fact, measurements usually have different error variances [12]. Ranging error of GPS is caused by many sources, such as the effect of ionosphere delay, tropospheric delay, carrier-to-noise ratio, and multipath. GDOP and the effect of these errors can be considered simultaneously; the extension of GDOP criteria is used for satellite selection in [13]. The satellite signal is also approximated by combining the user range accuracy value, carrier-to-noise ratio, elevation angle, and the date of ephemeris. The weighted GDOP (WGDOP) which takes these errors into account was proposed in [14]. The elevation of each satellite and signal-to-noise-ratio (SNR) are introduced as fuzzy subset to weight GDOP and provide the positioning solution [15]. When baro-altitude measurements or *a priori* terrain elevation information is used, the conventional GDOP formula cannot be applied and must be modified [16] in order to reduce the influence of satellites with a large error and evaluates the influence of each satellite on the arrangement of satellites. The GDOP was focused as a factor to determine the weight matrix and improve precision in GPS measurements [17]. The combinations of the GPS and Galileo satellite constellations will provide more visible satellites with better geometric distribution, and the availability of satellites will be significantly improved. A novel algorithm, namely, the WGDOP minimum algorithm, was proposed in [18] for the combined GPS-Galileo navigation receiver. In addition to the aforementioned, several papers which focus on WGDOP concepts have been proposed to improve the GPS positioning accuracy [19–21]. Taking the different variances of the satellites into account, researchers have proposed various WGDOP measures [13–21]. Much of the research literature needs matrix inversion to calculate WGDOP. Though they can guarantee to achieve the optimal subset, the computational complexity is usually too expensive to be practical.

High accuracy in wireless positioning system requires both the accurate measurement and a good effect of GDOP. When the measurements have different error variances or come from integrated positioning systems, WGDOP minimum criterion is appropriate to select the appropriate measurement units to diminish the positioning error. The optimal measurements selected with the minimum WGDOP can help reduce the adverse geometry effects. Increasing the number of satellites will always reduce the WGDOP value, since the best WGDOP can be obtained by computing all satellites in view. If the number of visible satellites is not large, the all-in-view method is a good choice to provide high accuracy positioning [15]. In order to further improve the positioning accuracy, the combined use of multiconstellation can be employed. There will be 70~90 navigation satellites operating at the same time when Glonass and Galileo reach full operation capability [22]. In any moment, there are more than 30 satellites in view in the multiconstellation navigation systems. To employ all-in-view method for positioning is very difficult for us in the future. Due to limited resources associated with many mobile devices and because the number of visible satellites is very large [18], measurement unit selection techniques can be used. If we select 4 out of 30 satellites, the number of possible subset is 27405. The calculation of WGDOP is a time and power consuming process, and fast calculation of WGDOP is most anticipated. WGDOP is computed for all subsets, and

the subset which gives the smallest WGDOP is selected for location estimation.

The growth of GPS embedded into current mobile phones continues to grow rapidly, as many mobile phones now are already equipped with GPS inside. Despite their performance increases, these devices still possess limited resources, such as the number of channels, battery capacities, and processing capability. Satellite selection can reduce the number of satellites used to position and as a result reducing the amount of calculation greatly. The number of measurements can be restricted and the resulting saving in load on the processor can be used to offer more spare processing time which can be used for other user specific requirements. On the other hand, reducing the signal-processing time of the receiver dedicated to satellite selection implies both increasing the processing capabilities available for other purposes and saving battery. The conventional method for calculating WGDOP is to use matrix inversion, which requires enormous amount of computation. This can present challenges to real-time practical applications. Therefore, it is very critical to select a subset with the most appropriate measurement units rapidly and reasonably before positioning.

To calculate WGDOP in the form of 2D and 3D formulations effectively, the closed-form solutions for two WGDOP formations are proposed for the case of each measurement with a unique variance and one of the measurements with higher location precision. The computation load of the proposed formulae is greatly less than that of the matrix inversion method. When exactly four measurements are used, the proposed formulae provide the best computational efficiency. The proposed formulae can also provide the exact solution to the WGDOP calculation and do not incur any approximation errors. The relatively simple closed-form WGDOP formulae can be implemented in the aforementioned papers [13–21]. The calculations of WGDOP for fast evaluation can be applied in GPS, WSN, and cellular communication systems. In practice, the measurement units of GPS, WSN, and cellular communication systems are satellites, sensors, and BSs, respectively.

The author of this paper proposed two novel architectures and presented four original architectures based on Rprop neural network to approximate WGDOP [23]. The disadvantage of Rprop-based WGDOP algorithm is the need of a training phase with several input-output patterns. We collect the elements of related matrix and the desired WGDOP value to train the neural network prior to the practicaluse. After the training, the elements of geometry matrix and weighted matrix as input data can not only pass through the trained Rprop but also predict the better appropriate WGDOP. From simulation results, the proposed WGDOP formulae always provide much better accuracy than Rprop-based WGDOP approximation [23]. But the proposed efficient formulae for WGDOP have been developed when there are exactly four measurement units used.

The remainder of this paper is organized as follows: Section 2 describes the concepts of GDOP and WGDOP. Section 3 reviews an efficient solution for the calculation of GDOP. The closed-form formulae for WGDOP calculations in the case of four measurements with unequal variances are proposed in Section 4. In Section 5, we examine the performance of the proposed formulae through simulation experiments. Conclusion is given in Section 6.

2. GDOP and WGDOP

GDOP is a task of choosing the appropriate measurement units, which results in a better geometric configuration and a more accurate position estimate. In order to achieve better positioning accuracy, it is desirable to select the combination of measurements with GDOP as small as possible. Using a 3D Cartesian coordinate system, the distances between satellite i and the user can be expressed as

$$r_i = \sqrt{(x - X_i)^2 + (y - Y_i)^2 + (z - Z_i)^2} + C \cdot t_b + v_{ri}, \quad (1)$$

where (x, y, z) and (X_i, Y_i, Z_i) are the locations of the user and satellite i, respectively; C is the speed of light, t_b denotes the time offset, and v_{ri} is pseudorange measurements noise. Equation (1) is linearized through the use of a Taylor series expansion around the approximate user position $(\hat{x}, \hat{y}, \hat{z})$ and the first two terms are retained. Defining $\hat{r}_i$ as r_i at $(\hat{x}, \hat{y}, \hat{z})$, we can obtain

$$\Delta r = r_i - \hat{r}_i \cong e_{i1}\delta_x + e_{i2}\delta_y + e_{i3}\delta_z + C \cdot t_b + v_{ri}, \quad (2)$$

where δ_x, δ_y, and δ_z are, respectively, coordinate offsets of x, y, and z,

$$e_{i1} = \frac{\hat{x} - X_i}{\hat{r}_i}, \qquad e_{i2} = \frac{\hat{y} - Y_i}{\hat{r}_i}, \qquad e_{i3} = \frac{\hat{z} - Z_i}{\hat{r}_i},$$
$$\hat{r}_i = \sqrt{(\hat{x} - X_i)^2 + (\hat{y} - Y_i)^2 + (\hat{z} - Z_i)^2}. \quad (3)$$

(e_{i1}, e_{i2}, e_{i3}), $i = 1, 2, \ldots, n$, denote the line-of-sight (LOS) vector from the satellites to the user.

The linearized pseudorange measurement equations take the form

$$z = H\delta + v, \quad (4)$$

where $z = \begin{bmatrix} r_1 - \hat{r}_1 \\ r_2 - \hat{r}_2 \\ \vdots \\ r_n - \hat{r}_n \end{bmatrix}$, $\delta = \begin{bmatrix} \delta_x \\ \delta_y \\ \delta_z \\ c \cdot t_b \end{bmatrix}$, $v = \begin{bmatrix} v_{r1} \\ v_{r2} \\ \vdots \\ v_{rn} \end{bmatrix}$, and $H = \begin{bmatrix} e_{11} & e_{12} & e_{13} & 1 \\ e_{21} & e_{22} & e_{23} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & e_{n3} & 1 \end{bmatrix}$ is the geometry matrix.

According to the least square algorithm (LS), the solution to (4) is given by

$$\hat{\delta} = \left(H^T H\right)^{-1} H z. \quad (5)$$

Assume that the pseudorange errors are uncorrelated with equal variances σ^2, the error covariance matrix can be expressed as

$$E\left[\left(\hat{\delta} - \delta\right)\left(\hat{\delta} - \delta\right)^T\right] = \sigma^2 \cdot \left(H^T H\right)^{-1}. \quad (6)$$

The variances are functions of the diagonal elements of $(H^T H)^{-1}$. The GDOP is a measure of accuracy for positioning systems and depends solely on the geometry matrix H

$$\mathrm{GDOP} = \sqrt{\mathrm{tr}\left(H^T H\right)^{-1}}. \tag{7}$$

In fact, each measurement error does not have the same variance, especially for the combination of different systems. The covariance matrix represents the uncertainty in the pseudorange measurements and has the following form:

$$E\left(vv^T\right) = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \sigma_n^2 \end{bmatrix}. \tag{8}$$

W is a diagonal matrix and defined as a weighted matrix

$$W = \begin{bmatrix} 1/\sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & 1/\sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & 1/\sigma_3^2 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1/\sigma_n^2 \end{bmatrix} = \begin{bmatrix} k_1 & 0 & 0 & 0 & 0 \\ 0 & k_2 & 0 & 0 & 0 \\ 0 & 0 & k_3 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & k_n \end{bmatrix}, \tag{9}$$

where $\sigma_i^2 = 1/k_i$, $i = 1, 2, \ldots, n$ are the variances of the measurement errors.

With the weighting matrix defined above, the solution to the weighted least square (WLS) can be expressed as

$$\hat{\delta} = \left(H^T W H\right)^{-1} H^T W z. \tag{10}$$

Taking into account that all measurement units contain different variances, the positioning algorithm using WLS estimation provides higher location accuracy than the LS estimation. Having considered both the geometric configuration and the *priori* knowledge of error models simultaneously, we choose WGDOP, instead of GDOP, for measurements selection to achieve effective performance improvement. The optimal subset is the one with the minimum WGDOP, which is given by the trace of the inverse of the $H^T W H$ matrix

$$\mathrm{WGDOP} = \sqrt{\mathrm{tr}\left(H^T W H\right)^{-1}}. \tag{11}$$

We can compute the WGDOP value of each subset, and then the subsets with minimum WGDOP are the selected measurement units. The conventional method for calculating WGDOP is to use matrix inversion for all subsets. The method can guarantee the optimal subset; however, it presents a heavy computational burden.

3. Calculation of GDOP for Four Measurements

In the time of arrival (TOA) positioning methods, which is applied to GPS, the TOA circle becomes the sphere in space and the fourth measurement is required to solve the receiver-clock bias for a 3D solution. The bias is the clock synchronization error between the receiver and the satellite. In practice, the time of user is significantly more inaccurate than that of an atomic clock on the satellite. In order to correct the clock bias errors present at the receiver of the users end, the measurement from the fourth satellite is employed. Getting information from the fourth measurement makes it possible to solve for this fourth unknown. Even though there are more than four satellites in view, a subset with four satellites is sufficient providing the sufficient measurements for navigation solution even though more than four satellites are in view, which is called the optimum four GPS satellites positioning [15]. The selection of four visible satellites to provide the suitable GPS positioning accuracy is presented in several papers [13–17]. Thus, we propose to take only four BSs with better geometry out of seven to estimate the MS location in cellular communication networks. For practical real-time applications, the number of selected measurement units should not be large. The efficient closed-form solution with simpler calculation for a four-satellite case is proposed in [9].

By using of the following properties:

$$\begin{aligned} (UV)^{-1} &= V^{-1}U^{-1}, \\ \mathrm{tr}\,(UV) &= \mathrm{tr}\,(VU), \end{aligned} \tag{12}$$

we have $\mathrm{tr}\,(UV)^{-1} = \mathrm{tr}\,(VU)^{-1}$.

From (7), the GDOP can be written as

$$\mathrm{GDOP} = \sqrt{\mathrm{tr}\left(H^T H\right)^{-1}} = \sqrt{\mathrm{tr}\left(HH^T\right)^{-1}}. \tag{13}$$

By defining the variable

$$B_{ij} = e_{i1}e_{j1} + e_{i2}e_{j2} + e_{i3}e_{j3} + 1, \quad 1 \le i < j \le 4, \tag{14}$$

and using the following relation that

$$e_{i1}^2 + e_{i2}^2 + e_{i3}^2 = 1, \quad i = 1, 2, 3, 4, \tag{15}$$

we have

$$HH^T = \begin{bmatrix} 2 & B_{12} & B_{13} & B_{14} \\ B_{12} & 2 & B_{23} & B_{24} \\ B_{13} & B_{23} & 2 & B_{34} \\ B_{14} & B_{24} & B_{34} & 2 \end{bmatrix}. \tag{16}$$

Defining the following variables:

$$a = \left(B_{12}B_{34} + B_{13}B_{24} - B_{14}B_{23}\right)^2 - 4B_{12}B_{34}B_{13}B_{24}, \tag{17a}$$

$$b = 16 - 4\left(B_{12}^2 + B_{13}^2 + B_{14}^2 + B_{23}^2 + B_{24}^2 + B_{34}^2\right), \tag{17b}$$

$$c = 2\left[B_{12}\left(B_{13}B_{23} + B_{14}B_{24}\right) + B_{34}\left(B_{13}B_{14} + B_{23}B_{24}\right)\right], \tag{17c}$$

the GDOP can be written as [9]

$$\mathrm{GDOP} = \sqrt{\frac{16 + b + c}{a + b + 2c}}. \tag{18}$$

Note that both $B_{12}B_{34}$ and $B_{13}B_{24}$ appear twice in the expression of a, and two multiplications can be eliminated. The closed-form equation needs only 39 multiplications, 34 additions, 1 division, and 1 square root.

4. Calculation of WGDOP for Four Measurements

To further reduce the computational overhead and improve the location performance, the selection of the optimal measurement units is necessary. Since the statistics of different location measurement units are, in general, not equal to each other, WGDOP is appropriate to an index for the precision of location in different networks, such as GPS, WSN, and cellular communication systems. The steps for positioning are listed as follows.

(1) We will first select four measurements among n measurement units to generate the subsets; thus, the n measurement units are classified into $C(n, 4)$ possible subsets.

(2) WGDOP is computed for all possible subsets of four measurement units.

(3) The subset which gives the smallest WGDOP is selected as the optimal subset.

(4) Finally, the four measurements of this subset can be used to find out the location solution.

The calculation of WGDOP takes considerable computing time; it is very imperative to accelerate the computation of WGDOP in real-time application. In this paper, we propose the efficient closed-form solution of two WGDOP formations, which includes the effect of GDOP and error statistics properties simultaneously. These solutions, with the simplified form for WGDOP calculation, can apply to all possible subsets in 3D and 2D scenarios and require much less computation compared to the conventional matrix inversion method.

4.1. Type 1: Four Measurements Have Different Error Variances

4.1.1. 3D Case. From (11) and by using the properties of the basic algebra theory, WGDOP can be alternatively recognized as

$$\text{WGDOP} = \sqrt{\operatorname{tr}\left(H^T W H\right)^{-1}} = \sqrt{\operatorname{tr}\left(HH^T W\right)^{-1}}. \tag{19}$$

By using (14) and (15), we have

$$HH^T W = \begin{bmatrix} e_{11} & e_{12} & e_{13} & 1 \\ e_{21} & e_{22} & e_{23} & 1 \\ e_{31} & e_{32} & e_{33} & 1 \\ e_{41} & e_{42} & e_{43} & 1 \end{bmatrix} \begin{bmatrix} e_{11} & e_{21} & e_{31} & e_{41} \\ e_{12} & e_{22} & e_{32} & e_{42} \\ e_{13} & e_{23} & e_{33} & e_{43} \\ 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} k_1 & 0 & 0 & 0 \\ 0 & k_2 & 0 & 0 \\ 0 & 0 & k_3 & 0 \\ 0 & 0 & 0 & k_4 \end{bmatrix}, \tag{20}$$

thus

$$HH^T W = \begin{bmatrix} 2k_1 & k_2 B_{12} & k_3 B_{13} & k_4 B_{14} \\ k_1 B_{12} & 2k_2 & k_3 B_{23} & k_4 B_{24} \\ k_1 B_{13} & k_2 B_{23} & 2k_3 & k_4 B_{34} \\ k_1 B_{14} & k_2 B_{24} & k_3 B_{34} & 2k_4 \end{bmatrix}. \tag{21}$$

The WGDOP parameter is the square root of the sum of diagonal terms of the matrix $(HH^T W)^{-1}$

$$\begin{aligned} &\text{WGDOP} \\ &= \sqrt{\operatorname{tr}\left(HH^T W\right)^{-1}} \\ &= \sqrt{\left(HH^T W\right)^{-1}_{1,1} + \left(HH^T W\right)^{-1}_{2,2} + \left(HH^T W\right)^{-1}_{3,3} + \left(HH^T W\right)^{-1}_{4,4}}. \end{aligned} \tag{22}$$

$(HH^T W)^{-1}_{i,i}$ is defined as the ith element on the diagonal of matrix $(HH^T W)^{-1}$

$$\begin{aligned} \operatorname{tr}\left(HH^T W\right)^{-1} &= \sum_{i=1}^{4} \left(HH^T W\right)^{-1}_{i,i} \\ &= \frac{\operatorname{tr}\left[\operatorname{adj}\left(HH^T W\right)\right]}{\det\left(HH^T W\right)} = \sum_{i=1}^{4} \frac{\operatorname{cof}_{i,i}\left(HH^T W\right)}{\det\left(HH^T W\right)}. \end{aligned} \tag{23}$$

The term $\operatorname{adj}(HH^T W)$ is the adjoint of $HH^T W$ and the cofactor, and $\operatorname{cof}_{i,i}(HH^T W)$ is the determinant of the submatrix of $HH^T W$ by deleting the ith row and the ith column. The cofactors can be obtained as

$$\begin{aligned} &\operatorname{cof}_{1,1}\left(HH^T W\right) \\ &\quad = k_2 k_3 k_4 \left[8 + 2\left(B_{23}B_{24}B_{34} - \left(B_{23}^2 + B_{24}^2 + B_{34}^2\right)\right)\right], \end{aligned} \tag{24a}$$

$$\begin{aligned} &\operatorname{cof}_{2,2}\left(HH^T W\right) \\ &\quad = k_1 k_3 k_4 \left[8 + 2\left(B_{13}B_{14}B_{34} - \left(B_{13}^2 + B_{14}^2 + B_{34}^2\right)\right)\right], \end{aligned} \tag{24b}$$

$$\begin{aligned} &\operatorname{cof}_{3,3}\left(HH^T W\right) \\ &\quad = k_1 k_2 k_4 \left[8 + 2\left(B_{12}B_{14}B_{24} - \left(B_{12}^2 + B_{14}^2 + B_{24}^2\right)\right)\right], \end{aligned} \tag{24c}$$

$$\begin{aligned} &\operatorname{cof}_{4,4}\left(HH^T W\right) \\ &\quad = k_1 k_2 k_3 \left[8 + 2\left(B_{12}B_{13}B_{23} - \left(B_{12}^2 + B_{13}^2 + B_{23}^2\right)\right)\right]. \end{aligned} \tag{24d}$$

After some algebraic manipulation, the determinant of matrix $HH^T W$ can be written as

$$\begin{aligned} \det\left(HH^T W\right) &= k_1 k_2 k_3 k_4 \\ &\quad \times \left\{16 + 2\left[B_{23}B_{24}B_{34} - \left(B_{23}^2 + B_{24}^2 + B_{34}^2\right)\right]\right. \\ &\quad + 2\left[B_{13}B_{14}B_{34} - \left(B_{13}^2 + B_{14}^2 + B_{34}^2\right)\right] \end{aligned}$$

$$
\begin{aligned}
&+2\left[B_{12}B_{14}B_{24}-\left(B_{12}^2+B_{14}^2+B_{24}^2\right)\right]\\
&+2\left[B_{12}B_{13}B_{23}-\left(B_{12}^2+B_{13}^2+B_{23}^2\right)\right]\\
&+\left(B_{12}B_{34}+B_{13}B_{24}-B_{14}B_{23}\right)^2\\
&-4B_{12}B_{34}B_{13}B_{24}\\
&+2\left[B_{12}\left(B_{13}B_{23}+B_{14}B_{24}\right)\right.\\
&\qquad\left.+B_{34}\left(B_{13}B_{14}+B_{23}B_{24}\right)\right]\Big\}. \qquad (25)
\end{aligned}
$$

Defining the following variables:

$$p=\left[B_{23}B_{24}B_{34}-\left(B_{23}^2+B_{24}^2+B_{34}^2\right)\right], \qquad (26a)$$

$$q=\left[B_{13}B_{14}B_{34}-\left(B_{13}^2+B_{14}^2+B_{34}^2\right)\right], \qquad (26b)$$

$$m=\left[B_{12}B_{14}B_{24}-\left(B_{12}^2+B_{14}^2+B_{24}^2\right)\right], \qquad (26c)$$

$$n=\left[B_{12}B_{13}B_{23}-\left(B_{12}^2+B_{13}^2+B_{23}^2\right)\right], \qquad (26d)$$

then we have

$$\mathrm{WGDOP}=\sqrt{\frac{2\cdot\left[(1/k_1)\cdot(4+p)+(1/k_2)\cdot(4+q)+(1/k_3)\cdot(4+m)+(1/k_4)\cdot(4+n)\right]}{a+c-16+2\cdot\left[(4+p)+(4+q)+(4+m)+(4+n)\right]}}. \qquad (27)$$

When four measurements have different error variances, the closed-form solution for WGDOP is given by

$$\mathrm{WGDOP}=\sqrt{\frac{2\cdot\left((1/k_1)\cdot P+(1/k_2)\cdot Q+(1/k_3)\cdot M+(1/k_4)\cdot N\right)}{a+c-16+2\cdot(P+Q+M+N)}}, \qquad (28)$$

where $P=4+p$, $Q=4+q$, $M=4+m$, $N=4+n$.

Note that $B_{12}B_{34}$, $B_{13}B_{24}$, $B_{12}B_{13}B_{23}$, $B_{12}B_{14}B_{24}$, $B_{13}B_{14}B_{34}$, $B_{23}B_{24}B_{34}$, B_{12}^2, B_{13}^2, B_{14}^2, B_{23}^2, B_{24}^2, B_{34}^2, $(4+p)$, $(4+q)$, $(4+m)$, and $(4+n)$ all appear twice in the express for WGDOP; thus sixteen multiplications and four additions can be eliminated. The values of $1/k_i, i=1,2,3,4$, are assumed to be already known before the calculation of (28); thus they can be treated as constants. From Table 1, the closed-form equation needs only 42 multiplications (including the constant multiplications by 4, 2, 2, and 2), 48 additions, 1 division, and 1 square root. Due to many parameters in the numerator and the denominator of (27) simultaneously, the computational complexity of the proposed criteria can be reduced.

4.1.2. 2D Case. From algebra theory, we know that solving the four unknowns requires at least four independent equations. When three measurements are utilized to determine the user location, a 2D position solution is obtained. This means that at least three measurements are required to determine the 2D position of the users. The complexity of computing the inverse of a 3×3 square matrix is very low. When four measurements are available for the 2D scenarios, we propose the simple closed-form formulae of the WGDOP calculations. The geometry matrix which is composed of four location measurement units in 2D environments is

$$H=\begin{bmatrix} e_{11} & e_{12} & 1\\ e_{21} & e_{22} & 1\\ e_{31} & e_{32} & 1\\ e_{41} & e_{42} & 1\end{bmatrix}, \qquad (29)$$

where $e_{i1}=(\hat{x}-X_i)/\hat{r}_i$, $e_{i2}=(\hat{y}-Y_i)/\hat{r}_i$, and $\hat{r}_i=\sqrt{(\hat{x}-X_i)^2+(\hat{y}-Y_i)^2}$, $i=1,2,3,4$.

Denoting

$$B_{ij}=e_{i1}e_{j1}+e_{i2}e_{j2}+1, \quad 1\le i<j\le 4, \qquad (30)$$

and using the fact that

$$e_{i1}^2+e_{i2}^2=1, \qquad (31)$$

WGDOP in the 2D case can be expressed as (28). The difference between the 2D and 3D scenarios of WGDOP calculation is in the calculation of $B_{ij}, 1\le i<j\le 4$. The computational complexity in the 2D case is 6 multiplications and 6 additions fewer than that in the 3D case. Therefore, the closed-form equation needs only 36 multiplications (including the constant multiplications by 4, 2, 2, and 2), 42 additions, 1 division, and 1 square root.

4.2. Type 2: Four Measurements Have Two Types of Error Variances

4.2.1. 3D Case. In the case of one measurement gives better accuracy than the others, the closed-form solution for WGDOP formulation is proposed here. The situation may occur in some positioning systems. For example, the BS serving a particular MS is called the serving BS, which provides the more accurate measurements in cellular communication systems [24]. Assume that the measurement variances of the serving BS and nonserving BSs are σ_1^2 and σ_2^2, respectively. Regarding the two types of the error variances, the weight matrix should be modified as follows:

$$W=\begin{bmatrix} 1/\sigma_1^2 & 0 & 0 & 0\\ 0 & 1/\sigma_2^2 & 0 & 0\\ 0 & 0 & 1/\sigma_2^2 & 0\\ 0 & 0 & 0 & 1/\sigma_2^2\end{bmatrix}=\begin{bmatrix} \omega & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}, \qquad (32)$$

where ω is the ratio of the nonserving BS error variance to the serving BS error variance:

$$\omega=\frac{\sigma_2^2}{\sigma_1^2}. \qquad (33)$$

Table 1: The complexity of WGDOP calculation when the four measurements have different error variances.

	Multiplications	Additions	Division	Square root
B_{12}	3	3	0	0
B_{13}	3	3	0	0
B_{14}	3	3	0	0
B_{23}	3	3	0	0
B_{24}	3	3	0	0
B_{34}	3	3	0	0
a	6	3	0	0
c	7	3	0	0
p	3	3	0	0
q	2	3	0	0
m	1	3	0	0
n	0	3	0	0
WGDOP (numerator)	5	7	0	0
WGDOP (denominator)	0	5	0	0
WGDOP	0	0	1	1
Total	42	48	1	1

Table 2: The complexity of WGDOP calculation when the four measurements have two types of error variances.

	Multiplications	Additions	Division	Square root
B_{12}	3	3	0	0
B_{13}	3	3	0	0
B_{14}	3	3	0	0
B_{23}	3	3	0	0
B_{24}	3	3	0	0
B_{34}	3	3	0	0
a	6	3	0	0
c	7	3	0	0
p	3	3	0	0
q	2	3	0	0
m	1	3	0	0
n	0	3	0	0
WGDOP (numerator)	2	5	0	0
WGDOP (denominator)	0	3	0	0
WGDOP	0	0	1	1
Total	39	44	1	1

In this case, we have

$$HH^TW = \begin{bmatrix} 2\omega & B_{12} & B_{13} & B_{14} \\ \omega B_{12} & 2 & B_{23} & B_{24} \\ \omega B_{13} & B_{23} & 2 & B_{34} \\ \omega B_{14} & B_{24} & B_{34} & 2 \end{bmatrix}, \tag{34}$$

and the cofactors can be obtained to be

$$\begin{aligned} &\mathrm{cof}_{1,1}\left(HH^TW\right) \\ &\quad = \left[8 + 2\left(B_{23}B_{24}B_{34} - \left(B_{23}^2 + B_{24}^2 + B_{34}^2\right)\right)\right], \end{aligned} \tag{35a}$$

$$\begin{aligned} &\mathrm{cof}_{2,2}\left(HH^TW\right) \\ &\quad = \omega\left[8 + 2\left(B_{13}B_{14}B_{34} - \left(B_{13}^2 + B_{14}^2 + B_{34}^2\right)\right)\right], \end{aligned} \tag{35b}$$

$$\begin{aligned} &\mathrm{cof}_{3,3}\left(HH^TW\right) \\ &\quad = \omega\left[8 + 2\left(B_{12}B_{14}B_{24} - \left(B_{12}^2 + B_{14}^2 + B_{24}^2\right)\right)\right], \end{aligned} \tag{35c}$$

$$\begin{aligned} &\mathrm{cof}_{4,4}\left(HH^TW\right) \\ &\quad = \omega\left[8 + 2\left(B_{12}B_{13}B_{23} - \left(B_{12}^2 + B_{13}^2 + B_{23}^2\right)\right)\right]. \end{aligned} \tag{35d}$$

The determinants of matrix HH^TW are found to be

$$\begin{aligned} \det\left(HH^TW\right) = \omega\{&16 + 2\left[B_{23}B_{24}B_{34} - \left(B_{23}^2 + B_{24}^2 + B_{34}^2\right)\right] \\ &+ 2\left[B_{13}B_{14}B_{34} - \left(B_{13}^2 + B_{14}^2 + B_{34}^2\right)\right] \\ &+ 2\left[B_{12}B_{14}B_{24} - \left(B_{12}^2 + B_{14}^2 + B_{24}^2\right)\right] \\ &+ 2\left[B_{12}B_{13}B_{23} - \left(B_{12}^2 + B_{13}^2 + B_{23}^2\right)\right] \\ &+ \left(B_{12}B_{34} + B_{13}B_{24} - B_{14}B_{23}\right)^2 \\ &- 4B_{12}B_{34}B_{13}B_{24} \\ &+ 2\left[B_{12}\left(B_{13}B_{23} + B_{14}B_{24}\right)\right. \\ &\quad \left. + B_{34}\left(B_{13}B_{14} + B_{23}B_{24}\right)\right]\}, \end{aligned} \tag{36}$$

and we have

$$\begin{aligned} &\mathrm{WGDOP} \\ &= \sqrt{\frac{2\cdot\left[(1/\omega)\cdot(4+p) + (4+q) + (4+m) + (4+n)\right]}{a + c - 16 + 2\cdot\left[(4+p) + (4+q) + (4+m) + (4+n)\right]}}. \end{aligned} \tag{37}$$

The closed-form WGDOP for the case of exactly four measurements can be expressed as

$$\begin{aligned} \mathrm{WGDOP} &= \sqrt{\frac{2\cdot\left[(1/\omega)\cdot(4+p) + (12+q+m+n)\right]}{a + c - 16 + 2\cdot\left[(4+p) + (12+q+m+n)\right]}} \\ &= \sqrt{\frac{2\cdot((1/\omega)\cdot P + G)}{(a + c - 16 + 2\cdot(P+G))}}, \end{aligned} \tag{38}$$

where $G = Q + M + N = 12 + q + m + n$.

Notice that $B_{12}B_{34}$, $B_{13}B_{24}$, $B_{12}B_{13}B_{23}$, $B_{12}B_{14}B_{24}$, $B_{13}B_{14}B_{34}$, $B_{23}B_{24}B_{34}$, B_{12}^2, B_{13}^2, B_{14}^2, B_{23}^2, B_{24}^2, B_{34}^2, $(4+p)$, and $(12+q+m+n)$ all appear twice in the WGDOP

Table 3: Comparison of average WGDOP residual for the proposed formulae and Rprop-based algorithm.

	Proposed WGDOP formulae	Rprop-based algorithm
Average WGDOP residual for Type 1	$3.7101 * 10^{-11}$	0.2385
Average WGDOP residual for Type 2	$3.7062 * 10^{-11}$	0.2311

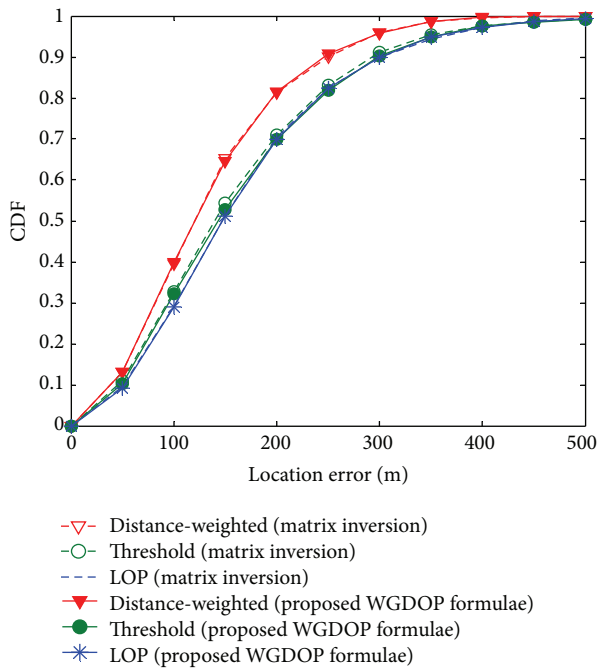


Figure 1: CDFs of the location error for various methods when four measurements have different error variances (Type 1).

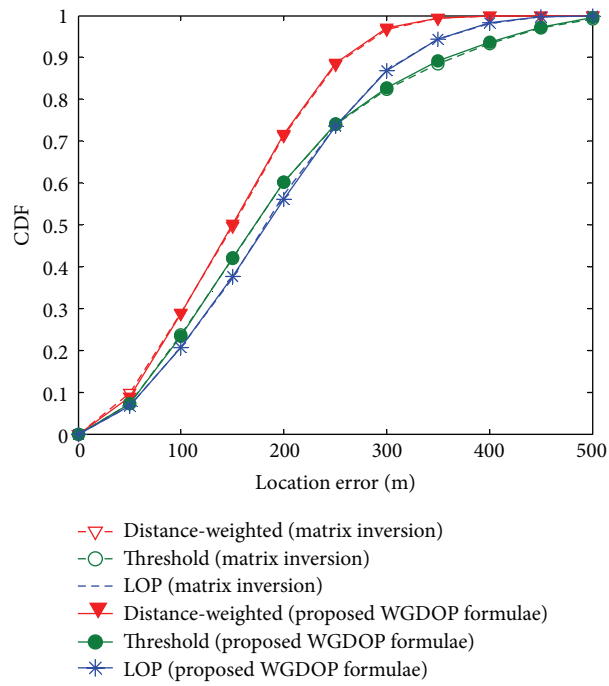


Figure 2: Comparison of CDFs of location error for various methods when four measurements have two types of error variances (Type 2).

express; thus sixteen multiplications and four additions can be eliminated. The value ω is also treated as a constant in the WGDOP calculation. From Table 2, this closed-form solution only needs 39 multiplications (including the constant multiplication by 4, 2, 2, and 2), 44 additions, 1 division, and 1 square root.

4.2.2. 2D Case. The WGDOP in the 2D case is expressed as (38). The WGDOP calculation in the 2D case requires 6 multiplications and 6 additions fewer than that in the 3D case. The closed-form equation needs only 33 multiplications (including the constant multiplications by 4, 2, 2, and 2), 38 additions, 1 division, and 1 square root. An alternative closed-form solution of the WGDOP calculation has been presented in this paper, in which one measurement provides superior location precision over the others.

5. Simulation Results

Time of arrival (TOA) is major time based method and usually used in calculating the mobile station (MS) location in cellular communication systems. It is consisting of seven base stations (BSs) in cellular communication system. The serving BS and its six neighboring BSs are separated by 5 km, and the MS is randomly placed among the BSs [25]. The non-line-of-sight (NLOS) propagation model is based on the uniformly distributed noise model [24], in which the TOA NLOS errors from all the BSs are different and assumed to be uniformly distributed over $(0, U_i)$, for $i = 1, 2, \ldots, 7$ where U_i is the upper bound. For Type 1, the variables are chosen as follows: $U_1 = 200$ m, $U_2 = 400$ m, $U_3 = 350$ m, $U_4 = 700$ m, $U_5 = 300$ m, $U_6 = 500$ m, and $U_7 = 350$ m. For Type 2, the variables are given as follows: $U_1 = 200$ m and $U_i = 500$, for $i = 2, 3, \ldots, 7$. The diagonal elements of the weighted matrix W are utilized with the reciprocal of the square root of an upper bound of the NLOS errors.

In order to verify the superior properties of the proposed formulae, we compare the results of WGDOP calculation accuracy for the proposed formulae and matrix inversion method. The WGDOP residual is defined as the difference between the proposed formulae and matrix inversion method. Table 3 shows average WGDOP residual for the proposed formulae and Rprop-based algorithm. For Type 1 and 2, the proposed formulae always provide much better WGDOP residual than Rprop-based algorithm [23].

We can evaluate the positioning accuracy with minimum WGDOP algorithm; MS location can be estimated by the

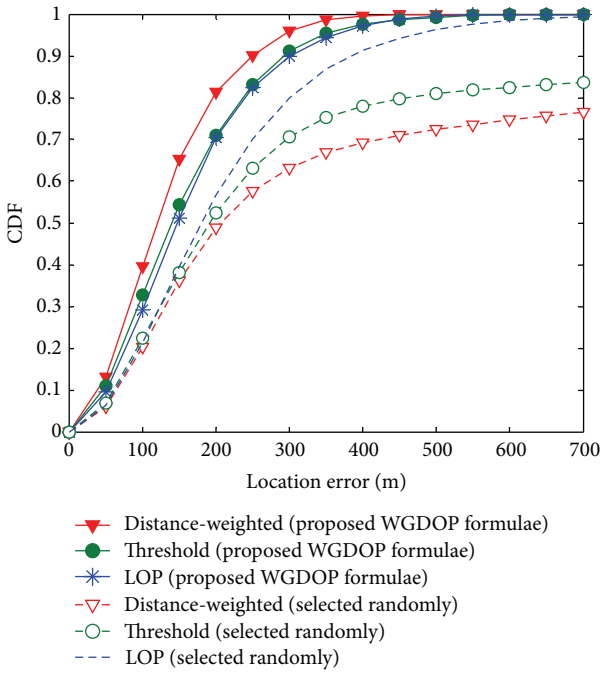


Figure 3: Comparison of location error CDFs using the subset with proposed minimum WGDOP approximation and the subset selected four BSs randomly (Type 1).

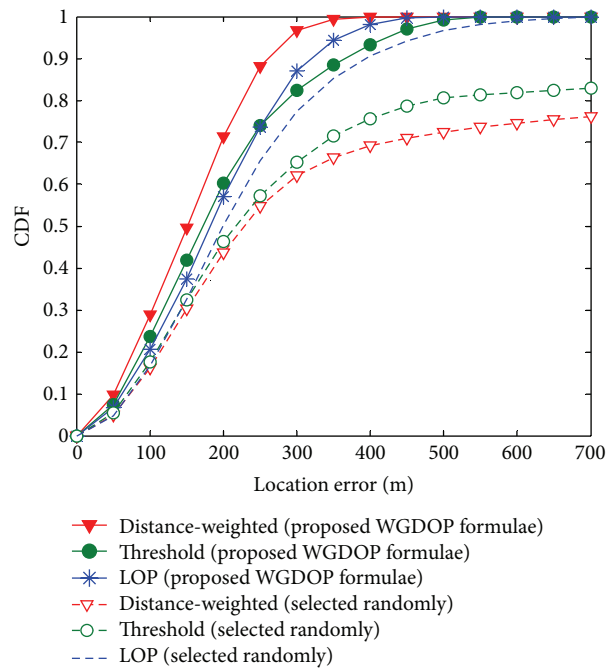


Figure 4: CDFs of location error of the subset with proposed minimum WGDOP formulae, and the subset selected four BSs randomly (Type 2).

linear lines of position algorithm (LOP) [26], distance-weighted method, and threshold method which we have proposed in [27]. When four measurements have different error variances (Type 1) or four measurements have two types of error variances (Type 2), the proposed WGDOP formulae and matrix inversion method provide the nearly identical MS location estimation, as shown in Figures 1 and 2.

For Type 1, Figure 3 shows the CDFs of the average location error of these methods with different subset. Four randomly selected BSs with poor geometry perform extremely worse location estimation, and the location accuracy can be strongly affected by the relative geometry between BSs and MS. The proposed Type 2 of efficient WGDOP formulae can give better location estimation than the subsets with four BSs taken from seven BSs randomly regardless of the different methods, as shown in Figure 4. The positioning accuracy would be seriously affected by the geometric configuration of BSs and MS. In order to eliminate the poor geometry influence and improve the positioning accuracy, the selection of BSs with minimum WGDOP approximation can be used and optimal geometric configuration with four BSs is obtained.

6. Conclusion

To reduce the computational overhead and improve location performance, the selection of optimal measurement units is necessary. The concept of GDOP is commonly used to determine the geometric effect of GPS satellite configurations. The conventional matrix inversion method is rather time consuming and requires a great deal of computational effort. The four measurement units selected from the maximum volume method may not be the optimal selection. Taking into account that the variance of each measurement variance is not equal, we choose the WGDOP instead of GDOP as the criteria to select the optimal measurement units. Due to the limited power and computation capability of many mobile devices and the great number of visible satellites, to obtain WGDOP efficiently from range measurements is very critical. To further reduce the complexity, novel closed-form solutions are proposed in this paper to compute WGDOP when there are exactly four measurements available for location estimation. The efficient closed-form formulae of two formations WGDOP calculations with less effort have been proposed, in which the *priori* error information of each measurement is not the same. If exactly four measurements are used, the proposed formulae can provide the best computational efficiency. The proposed formulae are applicable not only to GPS but also for the WSN and cellular communication systems. The WGDOP calculations for fast evaluation are able to reduce the computational load and eliminate the poor geometry influence. The proposed efficient formulae can provide very precise solution of WGDOP calculation without incurring any approximation error.

References

[1] G. Sun, J. Chen, W. Guo, and K. J. R. Liu, "Signal processing techniques in network-aided positioning: a survey of state-of-the-art positioning designs," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 12–23, 2005.

[2] D. Simon and H. El-Sherief, "Navigation satellite selection using neural networks," *Neurocomputing*, vol. 7, no. 3, pp. 247–258, 1995.

[3] D. Simon and H. El-Sherief, "Fault-tolerant training for optimal interpolative nets," *IEEE Transactions on Neural Networks*, vol. 6, no. 6, pp. 1531–1535, 1995.

[4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[5] D.-J. Jwo and K.-P. Chin, "Applying back-propagation neural networks to GDOP approximation," *Journal of Navigation*, vol. 55, no. 1, pp. 97–108, 2002.

[6] C.-S. Chen and S.-L. Su, "Resilient back-propagation neural network for approximation 2-D GDOP," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 2, pp. 900–904, March 2010.

[7] D. Y. Hsu, "Relations between dilutions of precision and volume of the tetrahedron formed by four satellites," in *Proceedings of the IEEE Position Location and Navigation Symposium*, pp. 669–676, April 1994.

[8] M. Kihara and T. Okada, "A satellite selection method and accuracy for the global positioning system," *Navigation*, vol. 31, no. 1, pp. 8–20, 1984.

[9] J. Zhu, "Calculation of geometric dilution of precision," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 28, no. 3, pp. 893–894, 1992.

[10] E. D. Kaplan and C. J. Hegarty, *Understanding GPS: Principles and Applications*, Artech House Press, Boston, Mass, USA, 2006.

[11] M. Kihara, "Study of a GPS satellite selection policy to improve positioning accuracy," in *Proceedings of the IEEE Position Location and Navigation Symposium*, pp. 267–273, April 1994.

[12] G. M. Siouris, *Aerospace Avionics Systems—A Modern Synthesis*, Academic Press, San Diego, Calif, USA, 1993.

[13] C. Park, I. Kim, J. G. Lee, and G.-I. Jee, "A satellite selection criterion incorporating the effect of elevation angle in GPS positioning," *Control Engineering Practice*, vol. 4, no. 12, pp. 1741–1746, 1996.

[14] H. Sairo, D. Akopian, and J. Takala, "Weighted dilution of precision as quality measure in satellite positioning," *IEE Proceedings*, vol. 150, no. 6, pp. 430–436, 2003.

[15] Y. Yong and M. Lingjuan, "GDOP results in all-in-view positioning and in four optimum satellites positioning with GPS PRN codes ranging," in *Proceedings of the Position Location and Navigation Symposium*, pp. 723–727, April 2004.

[16] M. Pachter, J. Amt, and J. Raquet, "Accurate positioning using a planar pseudolite array," in *Proceedings of the IEEE/ION Position, Location and Navigation Symposium*, pp. 433–440, May 2008.

[17] K. Kawamura and T. Tanaka, "Study on the improvement of measurement accuracy in GPS," in *Proceedings of the SICE-ICASE International Joint Conference*, pp. 1372–1375, October 2006.

[18] B. Xu and S. Bingjun, "Satellite selection algorithm for combined gpsgalileo navigation receiver," in *Proceedings of the 4th International Conference on Autonomous Robots and Agents*, pp. 149–154, February 2009.

[19] C. Hongwei and S. Zhongkang, "A nonlinear optimized location algorithm for bistatic radar system," in *Proceedings of the IEEE National Aerospace and Electronics Conference*, vol. 1, pp. 201–205, May 1995.

[20] N. Levanon, "Lowest GDOP in 2-D scenarios," *IEE Proceedings*, vol. 147, no. 3, pp. 149–155, 2000.

[21] H. Lu and X. Liu, "Compass augmented regional constellation optimization by a multi-objective algorithm based on decomposition and PSO," *Chinese Journal of Electronics*, vol. 21, no. 2, pp. 374–378, 2012.

[22] M. Zhang, J. Zhang, and Y. Qin, "Satellite selection for multi-constellation," in *Proceedings of the IEEE/ION Position, Location and Navigation Symposium*, pp. 1053–1059, May 2008.

[23] C.-S. Chen, J.-M. Lin, and C.-T. Lee, "Neural network for WGDOP approximation and mobile location," *Mathematical Problems in Engineering*, vol. 2013, Article ID 369694, 11 pages, 2013.

[24] S. Venkatraman, J. Caffery Jr., and H.-R. You, "A novel ToA location algorithm using LoS range estimation for NLoS environments," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 5, pp. 1515–1524, 2004.

[25] J. Caffery Jr. and G. Stuber, "Subscriber location in CDMA cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 47, pp. 406–416, 1998.

[26] J. J. Caffery Jr., "A new approach to the geometry of TOA location," in *Proceedings of the 52nd Vehicular Technology Conference*, pp. 1943–1949, September 2000.

[27] C.-S. Chen, S.-L. Su, and Y.-F. Huang, "Hybrid TOA/AOA geometrical positioning schemes for mobile location," *IEICE Transactions on Communications*, vol. E92-B, no. 2, pp. 396–402, 2009.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 207141, 9 pages
http://dx.doi.org/10.1155/2013/207141

Research Article

Recovery and Resource Allocation Strategies to Maximize Mobile Network Survivability by Using Game Theories and Optimization Techniques

Pei-Yu Chen[1,2] and Frank Yeong-Sung Lin[1]

[1] *Department of Information Management, National Taiwan University, Taipei 106, Taiwan*
[2] *CyberTrust Technology Institute, Institute for Information Industry, Taipei 106, Taiwan*

Correspondence should be addressed to Pei-Yu Chen; d96006@im.ntu.edu.tw

Received 19 April 2013; Revised 2 August 2013; Accepted 6 August 2013

Academic Editor: Anyi Chen

With more and more mobile device users, an increasingly important and critical issue is how to efficiently evaluate mobile network survivability. In this paper, a novel metric called Average Degree of Disconnectivity (Average DOD) is proposed, in which the concept of probability is calculated by the contest success function. The DOD metric is used to evaluate the damage degree of the network, where the larger the value of the Average DOD, the more the damage degree of the network. A multiround network attack-defense scenario as a mathematical model is used to support network operators to predict all the strategies both cyber attacker and network defender would likely take. In addition, the Average DOD would be used to evaluate the damage degree of the network. In each round, the attacker could use the attack resources to launch attacks on the nodes of the target network. Meanwhile, the network defender could reallocate its existing resources to recover compromised nodes and allocate defense resources to protect the survival nodes of the network. In the approach to solving this problem, the "gradient method" and "game theory" are adopted to find the optimal resource allocation strategies for both the cyber attacker and mobile network defender.

1. Introduction

Network security problems are often challenging given that the growing complexity and interconnected nature of IT systems lead to a limited capability of observation and control. This is especially the case for mobile networks, in which the cycle time of decision making is reduced from enterprise having access to real-time data. As the enterprise systems are widely relayed on mobile networks, the services are disrupted whenever the network suffers a disruption, such as from physical damage or malicious attacks. Compared to wired network system, mobile network systems are much more vulnerable to security problems [1]. For example, insofar as there is not a precisely defined physical boundary of the mobile network, as soon as an adversary comes in the radio range of a node, he can communicate with that node and thus launch a malicious attack on it [2]; these attacks include eavesdropping, phishing, war driving, and denial of service (DoS) attack [3]. As a result, there is a pressing need to design countermeasures for network attacks. Moreover, it is critical for an enterprise to evaluate and allocate its resources to protect it assets, as well as to be able to continuously provide service.

In the past, the security state of systems or infrastructures was classified in terms of two states: safe or compromised [4]. However, networks often face many situations, such as natural disasters, malicious attacks, and random error conditions, which can lead to different outcomes. Network security professionals must also ensure the availability and continuity of services. For these reasons, the binary concept of safe/compromised is insufficient to describe a system's state, with an increasing number of researchers focusing on the issue of network survivability.

There are many quantitative analyses of network survivability, such as connectivity. In [5], the definition of network connectivity is the minimum number of links or

Table 1: The summary of different DOD metrics.

No.	Name	Concept	Original
1	Degree of disconnectivity (DOD)	The DOD value can be explained as measuring the average numbers of the broken nodes in any O-D pair of the network.	[7]
2	Longest damaged path (LDP-DOD)	The LDP-DOD used to measure the damage degree of the network finds the most damaged O-D pair among all the O-D pairs of the network.	[7]
3	Minimal recovery node (MRN-DOD)	The MRN-DOD discovers the minimal numbers of broken nodes that are needed to repair and reconnect all the O-D pairs of the network.	[7]
4	Partial DOD (P-DOD)	Since the important degree of the different network areas is usually unequal, the network defender could assign different DOD requirements to different areas of the network, that is, the P-DOD.	[8]
5	Weight DOD (W-DOD)	Since the significance degree of each O-D pair can be diversified, the network defender can assign different weights to each O-D pair, that is, the W-DOD.	[8]

nodes that must be recovered from a given O-D (original-destination) pair. In general, the greater the number of links or nodes to be recovered to disconnect an O-D pair, the higher the survivability of the network. Thus, there are many studies adopting the concept of network connectivity to do quantitative analyses of network survivability. In [6], the researchers proposed using the network connectivity to measure the network survivability under intentional attacks and random disasters. Furthermore, the authors in [7] employing network connectivity for a quantitative analysis of network survivability proposed a survivability metric called the degree of disconnectivity (DOD) to estimate the residual network survivability after a malicious attack or any network crash incident.

To date, there have been several proposed degree of disconnectivity (DOD) metrics to evaluate network survivability. In [7], two other metrics called longest damaged path (LDP-DOD) and minimal recovery node (MRN-DOD) were proposed. Unlike the DOD metric, the LDP-DOD is used to measure the damage degree of the network by finding the most damaged O-D pairs among all the O-D pairs of the network. Therefore, the larger value of the LDP-DOD could be used to represent the most damage that a network could endure. On the other hand, the MRN-DOD discovers the minimal number of broken nodes that is necessary to be recovered in order to reconnect all the O-D pairs of the network.

In [8], the partial DOD (P-DOD) and weight DOD (W-DOD) metrics were adopted to evaluate network survivability. Because the important degree of the different network areas is usually unequal, the network defender could assign different DOD requirements according to its area, which is defined as the P-DOD. The network defender could then use the P-DOD value to determine the order to recover compromised nodes. Moreover, the significant degree of each O-D pair could be determined by diversity, where the network defender could assign different weights to each O-D pair, that is, the W-DOD. If the more significant O-D pair is cut to increase the degree of damage to the network, the W-DOD will clearly increase. The above DOD metrics are summarized in Table 1.

The DOD metric proposed in [7] assumed that the cyber attacker would launch the attack either successfully or unsuccessfully. However, this assumption is limited since the attack might not be perfectly successful or even completely unsuccessful. Motivated by previous works, the Average Degree of Disconnectivity (Average DOD) is developed to carry out a quantitative analysis of network survivability, combining the concept of probability as calculated by the contest success function [9] with the DOD metric, thus becoming the Average DOD. When the number of the Average DOD value is large, the damage to the network will be greater.

According to the allocated resources on each node from both cyber attacker and network defender, the contest success function is adopted to calculate the attack success probability of each node. The attack success probability of each node is calculated based on the concept of contest success function, where S_i represents the attack success probability of node i:

$$S_i\left(T_i, t_i\right) = \frac{T_i}{T_i + t_i} = \frac{1}{1 + t_i/T_i}. \tag{1}$$

In [7], the DOD metric is used to measure the damage degree of the network, such that the larger the DOD value, the more the damage degree of the network. The definition of the DOD value (D) is as function (2). In this metric, W is the index set of all given critical O-D pairs, while t_{wi} is the shortest path of O-D pairs w, where $w \in W$; $|W|$ is the O-D pair number of W. The total shortest path cost of each O-D is calculated first. Here, c_i represents the transmission cost of a node i, where a large number M represents the link disconnection:

$$D = \frac{\sum_{w\in W}\sum_{i\in V} t_{wi}c_i}{|W|\,M}. \tag{2}$$

The calculated DOD value could be explained as measuring the average numbers of broken nodes in any O-D pair of the network.

Theoretical models at the system level play an increasingly important role in network security and provide a scientific basis for high-level security-related decision making. To enhance or reduce network survivability, both network defender and cyber attacker usually need to invest a limited number of resources in the network. In these models, the decision makers in network security problems play the role

Table 2: Given parameters.

Given parameter	
Notation	Description
$S_i(T_i, t_i)$	The attack successful probability on node i
T_i	The attack resource allocated on node i
t_i	The defensive resource allocated on node i

of either the attacker or the defender. They often have conflicting goals, in that a cyber attacker attempts to breach the security of the system to disrupt or cause damage to network services, whereas a defender takes appropriate measures or strategies to enhance the system security design or response. Traditionally, although the attack-defense resource allocation problem is usually discussed for only one round [7, 10–12], the interaction frequency between cyber attacker and network defender is usually more than one time in real world. For this reason, several researchers are beginning to discuss multiround attack-defense resource allocation issues [8, 13, 14]. However, most of the existing solutions to multiround attack-defense resource allocation are still not suitable to the field of the network security, because they almost solely focus on the attack-defense problem of the parallel systems [13, 14] and serial systems [15]. In reality, the topology of the network is usually more complicated than the topology of the parallel, serial, or even serial-parallel systems. Thus, a new multiround attack-defense model to solve the resource allocation problem for both cyber attackers and network defenders is needed and developed in this study.

2. Problem Formulation

2.1. The Average DOD. The DOD metric proposed in [12] assumed that the cyber attacker launches the attack either successfully or unsuccessfully, but this binary assumption is limited in its inability to describe attack results that are neither perfectly successful nor unsuccessful. Therefore, the concept of the probability calculated by contest success function combined with the DOD metric was forwarded as a new survivability metric called the Average DOD. According to the allocated resources on each node of both cyber attacker and network defender, the contest success function would be adopted to calculate the attack success probability of each node. The attack success probability of each node is demonstrated, where S_i represents the attack success probability of node i. After each attack-defense interaction, there are 2^v configurations of a given network, where V means the total number of network nodes, and j is the configuration index. For example, in Table 2, the total number of possible configurations of a network is 2^9, and the configuration index j is $1, 2, \ldots, 512$.

In addition, each possible network configuration has a probability P_j, which is related to the safe or compromised state of the configuration. This probability is determined by the attack success probability S_i of each node. For example, if a 9-node network is completely compromised by the attacker, the probability of this network configuration would be $\prod_{i=1}^{9} S_i$ (where S_i means the attack success probability of the node i). However, if all the nodes of the network are still functional, the probability of this network configuration would be $\prod_{i=1}^{9}(1 - S_i)$.

Furthermore, each kind of network configuration would lead to a different damage degree of the network. The degree of disconnectivity (DOD) having been introduced in the preceding part can be adopted to measure the damage degree of network. For example, if all the nodes of the network are still functional, the DOD value would be 0. The probability and DOD value of each kind of network configuration are calculated with the concept of expectation value. The predicted mean value of the result of a statistical experiment would be adopted to evaluate the damage degree of the whole network. The calculated expectation value is defined as the Average DOD $\overline{D}$ here, which is shown in (3):

$$\overline{D} = \sum_{j=1}^{j \in J} D_j P_j. \tag{3}$$

The Average DOD value is influenced by the attack success probability calculated by the resource allocation of both the cyber attacker and network defender. Therefore, the Average DOD value could be induced from the damage degree of the network. The calculation of an Average DOD 9-node-network example is demonstrated in Table 3. In this example, probability P_1 of configuration 1 is $\prod_{i=1}^{9}(1 - S_i)$, since all nodes of this configuration are functional. In (2), the DOD value is the recovered nodes in any given compromised O-D pair; there is no compromised node in configuration 1. Therefore, the DOD value D_1 here is 0.

2.2. Problem Description. In this attack-defense problem, both cyber attacker and network defender employ certain strategies to attain their goals. From the perspective of the network defender, the defender usually aims to minimize the damage degree of the target mobile network. On the other hand, the cyber attacker hopes to maximize the damage degree of the network. However, given that both cyber attacker and network defender are always limited by the invested resources, how to make the decision to efficiently allocate resources to each node is an extremely significant issue for both cyber attacker and network defender. Meanwhile, in the real world, it is impossible that there will only be a one-time interaction between the cyber attacker and network defender, and as such, a multiround attack-defense problem in this mathematical model needs to be considered. A mathematical model to support both cyber attacker and network defender in making the optimal decision is thus developed to solve this problem.

In this model, the damage degree of the mobile network can be evaluated by the Average DOD value. The cyber attacker needs to determine how to allocate resources to attack the targeted network, since the strategies of both cyber attackers and network defenders are usually constrained by the allocated resources in each round. On the other hand, the network defender can choose to reallocate the existing resources in the mobile network, but the problem regarding

Table 3: Calculation of an example of the Average DOD value.

Configuration j	Network configuration*	Probability P_j of configuration j	DOD value D_j on configuration j	Probability $P_j \times$ DOD value D_j
1	1, 2, 3, 4, 5, 6, 7, 8, 9	$\prod_{i=1}^{n}(1-S_i)$	D_1	0
2	$\underline{1}$, 2, 3, 4, 5, 6, 7, 8, 9	$S_1\prod_{i=2}^{n}(1-S_i)$	D_2	$D_2S_1\prod_{i=2}^{v}(1-S_i)$
3	1, $\underline{2}$, 3, 4, 5, 6, 7, 8, 9	$(1-S_1)\,S_2\prod_{i=3}^{n}(1-S_i)$	D_3	$D_3\,(1-S_1)\,S_2\prod_{i=3}^{v}(1-S_i)$
⋮	⋮	⋮	⋮	⋮
512	$\underline{1}$, $\underline{2}$, $\underline{3}$, $\underline{4}$, $\underline{5}$, $\underline{6}$, $\underline{7}$, $\underline{8}$, $\underline{9}$	$\prod_{i=1}^{n}S_i$	D_4	$D_4\prod_{i=1}^{n}S_i$

*$\underline{i}$ means the node i is compromised in configuration j.

the discount factor of those reallocated resources also needs to be considered here. As a result, the total number of resources that the defender could use would be the newly allocated and reallocated resources in each round, and those resources could be used to recover the compromised nodes and to protect the mobile network survival nodes.

In the following, the notations of given parameter and decision variable in this model are listed in Tables 4 and 5.

Using the above notations of the given parameter and decision variable, the problem is formulated as follows:

Objective Function

$$\min_{\vec{b}_r}\max_{\vec{a}_r} \quad \sum_{r\in R} w_r \overline{D}\left(\vec{a}_r, \vec{b}_r\right) \tag{IP 1}$$

$$\text{subject to} \quad \sum_{i\in V} b_{ri} + \sum_{i\in V} e_{ri}z_{ri} \le B_r + \sum_{i\in V}\theta_i d_{ri} \quad r\in R, \tag{IP 1.1}$$

$$\sum_{i\in V} a_{ri} \le A_r \quad r\in R, \tag{IP 1.2}$$

$$\sum_{r\in R} A_r \le \widehat{A}, \tag{IP 1.3}$$

$$\sum_{r\in R} B_r \le \widehat{B}. \tag{IP 1.4}$$

Explanation of the Objective Function

(IP 1) The purpose of the objective function is to minimize both the maximum sum of the product of the Average DOD and the different weight in each round.

Explanation of the Constraint Function

(IP 1.1) The sum of the allocated defense budgets in each node and repaired cost of the compromised nodes should not exceed the sum of the new allocated and reallocated budgets in that round.

(IP 1.2) The sum of the allocated attack budgets in each node should not exceed the attack budgets in that round.

(IP 1.3) The sum of the allocated defense budgets in each round should not exceed the total budget of the defender.

Table 4: Given parameter.

Notation	Description
V	Index set of nodes
R	Index set of rounds in the attack and defense actions
$\widehat{A}$	Total budget of attacker
$\widehat{B}$	Total budget of defender
w_r	The weight of the Average DOD in round r, where $r\in R$
θ_i	Existing defense resources allocated on node i, where $i\in V$
e_{ri}	Repair cost of defender when node i is dysfunctional in round r, where $i\in V$ and $r\in R$
d_{ri}	The discount rate of defender reallocate resources on node i in round r, where $i\in V$ and $r\in R$

Table 5: Decision variable.

Notation	Description
$\vec{a}_r$	Attacker's budget allocation, which is a vector of attack cost a_{r1}, a_{r2} to a_{ri} in round r, where $i\in V$ and $r\in R$
$\vec{b}_r$	Defender's budget allocation, which is a vector of defense cost b_{r1}, b_{r2} to b_{ri} in round r, where $i\in V$ and $r\in R$
a_{ri}	Attacker's budget allocation on node i in round r, where $i\in V$ and $r\in R$
b_{ri}	Defender's budget allocation on node i in round r, where $i\in V$ and $r\in R$
$\vec{z}_{ri}$	Defender's node recovery status, which is a vector of repaired status z_{r1}, z_{r2} to z_{ri} in round r, where $i\in V$ and $r\in R$
z_{ri}	1 if node i is repaired by defender in round r, 0 otherwise, where $i\in V$ and $r\in R$
A_r	Attacker's attack budget in round r, where $r\in R$
B_r	Defender's defense budget in round r, where $r\in R$
$\overline{D}\left(\vec{a}_r, \vec{b}_r\right)$	The Average DOD among r rounds, considering that it is under the attacker's and defender's budget allocations, is $\vec{a}_r$ and $\vec{b}_r$ in round r, where $r\in R$

(IP 1.4) The sum of the allocated attack budgets in each round should not exceed the total budget of the attacker.

3. Solution Approach

Combining game theory with the gradient method is our proposal to solve the optimal resource allocation strategy for both cyber attackers and network defenders. The gradient method is used to calculate the Average DOD value and to find the optimal resource allocation strategy in each node for both cyber attacker and network defender. Game theory is adopted to find the optimal percentage resource allocation in each round for both cyber attacker and network defender. Further details are presented in the following sections.

3.1. Game Theory. Game theory provides the mathematical tools and models for investigating multi-player strategic decision making, where the rational players compete for restricted resources [9]. This demonstrates the modeling situations of conflict and predicts the behavior of the different players. Security games and their solutions are used not only as a basis for formal decision making and algorithm development but also for predicting attacker and defense behavior [16]. The weakness of traditional network security solutions is that they lack a quantitative decision framework [17]. As a result, researchers are starting to advocate the utilization of game theory approaches. According to the surveys in [18, 19], several game theory approaches have in recent years been proposed to address network security issues. In these frameworks, a network administrator and an attacker can be viewed as two competing players participating in a game, with the added benefit that game theory has the capability of examining hundreds of thousands of possible scenarios before taking the best action.

The primary components of the game theory are player, strategy, payoff, and information. In this model, there are the two players: cyber attacker and network defender; strategy means the possible moves that the players would take; the payoff value means the positive or negative reward to the player from a specific strategy; finally, the information can be categorized into two types, one is complete information, and the other one is perfect information, with the former meaning that every player knows both the strategies and payoff values of all players in the game, and the latter meaning that each player is aware of the moves of all players that have already taken place. The nominal definitions of game theory are summarized in Table 6.

Table 6: The nominal definition of the game theory.

Noun	Definition
Player	A basic entity in a game with making choices for actions
Strategy	The possible motion that the players take
Payoff	The positive or negative reward to the player on the specific strategy
Complete information	Every player knows both the strategies and payoffs of all players in the game
Perfect information	Each player is aware of the moves/strategies of all other players that have already taken place

According to the move order, the game can be categorized into simultaneous games (i.e., static games) and sequential games (i.e., dynamic games). If the all the players move simultaneously, this game is called a simultaneous game, in contrast to a sequential game in which players move in a sequence. And depending on whether the game repeats or not, it will be categorized as either a one-shot or repeat game: the former is a game played only one time, whereas the latter is a game that repeats. The game can be further categorized into zero-sum or nonzero sum game, based on whether the gain or loss of one equals the gain or loss of the other. Finally, according to the definition of the complete and perfect information, game theory is categorized into four types: complete and perfect information games, incomplete and perfect information games, complete and imperfect information games, and incomplete and imperfect information games.

In this paper, since both cyber attacker and network defender need to determine how to efficiently allocate resources simultaneously in each node in each round before the attack-defense game, this problem can be viewed as a simultaneous or imperfect information game. Moreover, insofar as both cyber attacker and network defender have complete information about the strategies and payoff values (the Average DOD value) of each other, this problem is regarded as a complete information game. Therefore, a two-player (cyber attacker and network defender), zero-sum, complete, and imperfect information game is used to solve this problem.

3.2. Gradient Method. The gradient method is a general framework used to resolve the optimization problems of how to maximize or minimize functions of continuous parameters. The proposed model in this paper is a min-max formulation, and both cyber attacker and network defender are assumed to be able to allocate continuous resources to each node. Here, the gradient method is adopted to solve this problem. The gradient method can usually be categorized into two types: one is gradient descent and the other one is gradient ascent [14]. The gradient descent method can be used to solve the optimal minimization problem. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient (or of the approximate gradient) of the function at the current point. On the other hand, if instead one takes steps proportional to the positive of the gradient, one approaches a local maximum of that function; the procedure is then known as gradient ascent. The concepts of gradient descent and gradient ascent are extremely similar.

3.3. The Proposed Heuristic. We here describe the detailed process of combining game theory with the gradient method [20] is adopted to find the optimal resource allocation strategy in each node in each round for both cyber attacker and network defender. The gradient method is used to calculate the Average DOD value and to find the optimal resource allocation strategy in each node. Given that how to allocate resources in each round is another issue, game theory is adopted to determine the optimal percentage resource

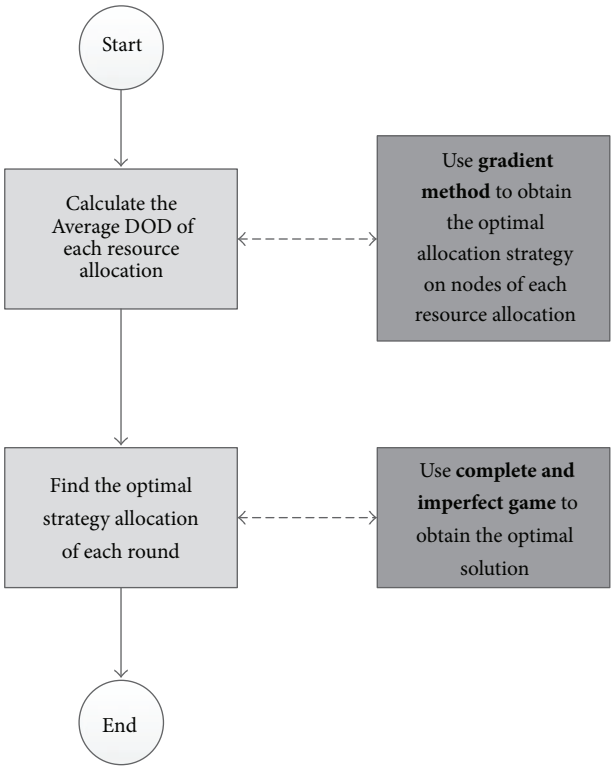


Figure 1: The proposed heuristic.

Step 1. an initial point
Step 2. Determine a positive or negative direction
Step 3. Determine a step size
Step 4. Do {
Find the most impact of all dimensions
Move a step of the most of all dimensions
Update an initial point
} While (a Given Stop Criterion)

Algorithm 1: The algorithm of the gradient method.

allocation in each round. The proposed heuristic, with its two major steps, is illustrated in Figure 1.

First, the gradient method is adopted to find an optimal strategy for each node in the given configuration. Initially, it is assumed that the cyber attacker and network defender would evenly allocate their limited resources on each survival node. The cyber attacker has limited resources in each round, and as a result, the cyber attacker would choose the gradient ascent method to maximize damage degree of the network. At the same time, the defense resources are also limited in each round, leading the network defender to use the gradient descent method to find the minimization solution. The detailed process flow of the gradient method is described in Algorithm 1. The selection criterion of the start point is critical, because it influences the quality of the computational efficiency. Moreover, a positive or negative direction results from the maximization or minimization problem. If the maximization problem is to be solved, the positive direction must be chosen. The gradient method adopts a step-by-step method to find the optimization result.

Table 7: The game matrix.

Strategy		Player 1				
		S_{11}	S_{12}	S_{13}	S_{14}	S_{15}
Player 2	S_{21}	U_{11}	U_{12}	U_{13}	U_{14}	U_{15}
	S_{22}	U_{21}	U_{22}	U_{23}	U_{24}	U_{25}
	S_{23}	U_{31}	U_{32}	U_{33}	U_{34}	U_{35}
	S_{24}	U_{41}	U_{42}	U_{43}	U_{44}	U_{45}
	S_{25}	U_{51}	U_{52}	U_{53}	U_{54}	U_{55}

Here, the derivative method is adopted in Step 4 in Algorithm 1, which is designed to find the most important node in the given configuration. The derivative of the Average DOD value is $\widehat{D_i}$, shown in (4), which represents the importance of the node i; r_i represents the resources on node i. The player would move more resources from the less important to the most important nodes. The procedure is stopped when the resource movement is not significant to the Average DOD. After this, the optimal resource allocation strategy for both cyber attacker and network defender in each node is obtained:

$$\widehat{D_i} = \lim_{h \to 0} \frac{\overline{D}(r_i + h) - \overline{D}(r_i)}{h}. \tag{4}$$

The second part of the proposed heuristic involves game theory, which is adopted to efficiently allocate resources in each round for both cyber attacker and network defender. For two players, the strategy of one is represented in a column, whereas the strategy of the other is represented in a row of a matrix. For example, in Table 7, both players have five different strategies (S_{11} to S_{15} and S_{21} to S_{25}), with the combination of the two players' different strategies resulting in 25 (U_{11} to U_{55}) values (the Average DOD values).

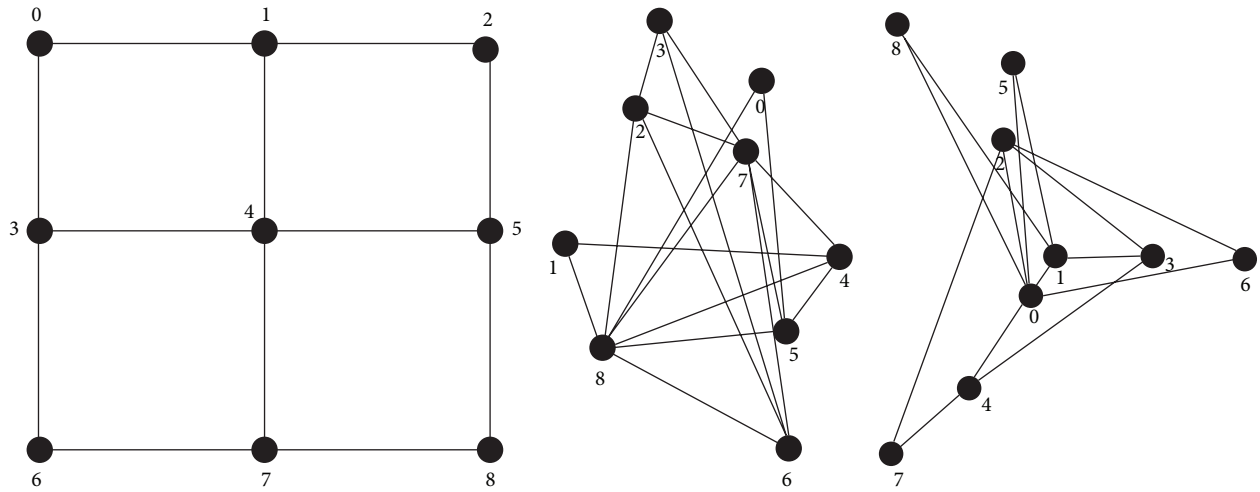
In this paper, the cyber attacker and network defender strategies involve different percentages of resource allocation in each round and can be formulated in a matrix. The payoff of all the resource allocation strategies of each participant is calculated by the Average DOD. The analysis of the complete and imperfect information game is conducted via heuristics. The solution procedure of the complete and imperfect information game [18] is shown in the following steps.

Step 1. Dominant strategy elimination, which means that no matter what kind of strategy the opponent takes, it is better than the other strategies.

Step 2. If only one strategy is left for each participant, it is the optimal strategy. Otherwise, go to Step 3.

Step 3. Use the min-max strategy to find the optimal strategy of each participant. If the min-max strategy still cannot find the optimal strategy, go to Step 4.

Step 4. Use the mixed strategy (linear programming) to find the optimal strategy for each participant.



(a) Grid network (b) Random network (c) Scale-free network

Figure 2: Mobile network topology.

Table 8: Experiment parameters settings.

Parameters	Value
Network topology	(1) Grid, in Figure 2(a) (2) Random, in Figure 2(b) (3) Scale-free, in Figure 2(c)
The number of rounds	2
The number of nodes	9
The number of links	24~36
The total resources of both players	20

4. Computational Experiments

The proposed solution approach is implemented on a PC with AMD Athlon X3 440 CPU 3.00 GHz, 2 GB RAM, and on the OS of MS Windows 7.

The parameters used in the experiments are shown in Table 8.

Because of the complexity of this problem, the number of mobile network nodes considered in the experiments is only 9, and the number of attacker-defender interactions covers only two rounds. Considering the variety of the distributions of mobile nodes, three types of mobile network topologies have been selected to act as attack-defense nodes: the grid network (GD), the scale-free network (SF), and the random network (RD). These three topologies are shown in Figure 2.

Both cyber attacker and network defender would attach a different level of importance to each round, so the different weight of each round would be considered. In this model, given that the weight in the two rounds is (a, b), the first round weight is a, while the second round weight is b. In this paper, we maintain that the importance of these two rounds is equally important, from which we induce the weight to be 0.5.

In this model, three kinds of node recovery policies are proposed. First, in NR1, the defender would choose to recover all the compromised nodes when the resources are sufficient. If the resources are insufficient, they would be used to protect the survival nodes. The second recovery policy is the defender choosing not to recover any compromised node (NR2). Finally, because the defense resources are limited, the third policy determines the order to recover compromised nodes by τ_i in (5) (NR3). Given that e_{ri} is the repair cost of the defender when node i is dysfunctional in round r, where $i \in V$ and $r \in R$, $|W_i|$ is the number of node i on O-D pair w, where $w \in W$:

$$\tau_i = \frac{|W_i|}{e_{ri}} \tag{5}$$

(once the unit cost recovers a larger number of the O-D pairs, this means that this node is more important. For this reason, the above formulation could be used to determine the order to recover compromised nodes).

4.1. The Experiments. There are several different kinds of strategies that the attacker and defender could implement, which result in various possible attack-defense situations. However, insofar as the defense resources are usually limited with resources usually being used to not only protect survival nodes but also recover compromised nodes, three kinds of different node recovery policies, that is, NR1, NR2, and NR3, are proposed in this paper and will be the subject of the following section.

4.2. Experiment Results. The purpose of this experiment is to compare the results from different kinds of node recovery policies (NR). To compare the three different kinds of node recovery policies, it is assumed that in the resource reallocation policy of the defender, the defense resources of each round would not be accumulated (RR1). Further, the weight of two rounds would be $(0.5, 0.5)$. The total resources of players, that is, the attacker and defender, are held to

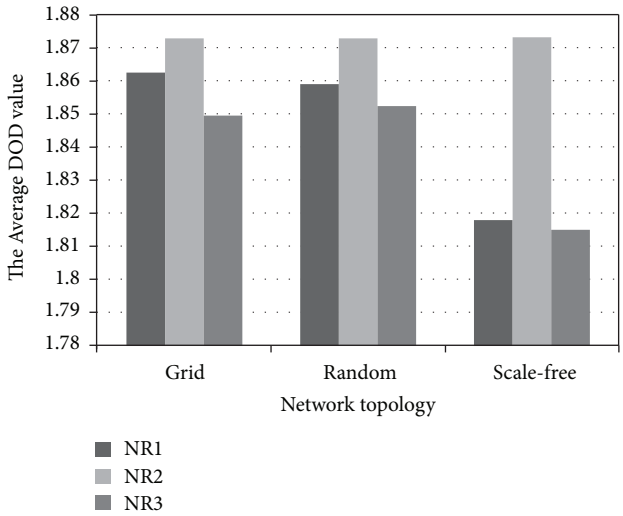


Figure 3: The different node recovery policies in the different network topologies.

Table 9: The experiment results in different kinds of node recovery policy.

Network topology	NR1	NR2	NR3
Grid	1.8626	1.8729	1.8496
Random	1.8592	1.873	1.8525
Scale-free	1.8180	1.8733	1.8151

be equal. The experiment results are listed in Table 9. The different results of the different node recovery policy for the three kinds of network topology are also compared in Figure 3.

4.3. Discussion of Results. The experiment results of the different node recovery policies of the defender have been described. In the following, the results are further discussed.

(i) The recovery policy is advantageous insofar as it improves the Average DOD of the defender. The experiment shows that when the defender has the ability to recover compromised nodes (NR1 and NR3), the Average DOD value is less than when the defender cannot recover any compromised nodes (NR2). Once the defender implements node recovery policies to recover compromised nodes, this decreases the value of the Average DOD. Therefore, when the defender takes node recovery policies to recover certain compromised nodes (NR1 and NR3), the Average DOD value is less than when the defender cannot recover any compromised nodes (NR2).

(ii) Among the three node recovery policies, NR3 is better than the other policies for the grid, random, and scale-free network topologies. NR3 is a strategy for recovering nodes according to their importance. In many experimental cases, the resources are limited and insufficient, thus making it impossible to recover the entire set of compromised nodes. If the resources are restricted, the defender under the NRI policy would use resources to protect survival nodes instead of recovering nodes. However, the node recovery policy is better than the node protection one in improving the network survivability. Hence, the node recovery policy of the NR3 would be better than the NR1 from the view of the defender.

5. Conclusion and Future Works

In this paper, two issues are considered. First, in order to evaluate mobile network survivability, a new survivability metric called Average DOD (degree of disconnectivity) was proposed. In addition, the problem of how to efficiently allocate resources in each node in each round for both cyber attacker and network defender is solved.

This work offers two main contributions. The first was the introduction of the Average DOD metric, which combines the concept of the probability calculated by the contest success function with the DOD metric and which can be a new evaluation tool to demonstrate network survivability. Secondly, a new min-max mathematical formulation was proposed to describe the conflict behavior of a network scenario. Both cyber attacker and network defender could adopt several different policies. The resource reallocation and node recovery problem is considered for the mobile network defender in this paper. As game theory deals with problems in which multiple players with contradictory objectives compete with each other, we developed a combined approach using the gradient method and game theory to resolve the optimal resource allocation for both cyber attacker and network defender in each node in each round. The gradient method can be used to find the optimal resource allocation in each node. Meanwhile, game theory is employed to find the optimal percentage resource allocation in each round. The proposed model provides a mathematical framework for analysing and modeling the posed mobile network security problems.

Although this paper has discussed a two round attack-defense game, it is still difficult to solve the multiround attack-defense scenario because of the complexity of mathematical problem. A possible solution involves the introduction of a threshold for computing or an advanced technology, such as parallel processing systems, in order to improve the efficiency of this model. Furthermore, from the experiment results, compared with the node protection strategy, the node recovery policy is better for defenders to ensure better network survivability. On the other hand, in the multiround attack-defense scenario, the attacker usually gains experience from his previous attack, and as such, the accumulated experience of the attacker should be taken into account in this model. Another consideration is that the resources might have multiple purposes, such as network defenders possibly deploying counterattack strategies to attack the attacker and the cyber attacker possibly using defense strategies to protect his critical information. As a result, since the purpose of resources may not be limited to only one usage for both cyber attacker and network defender, the concept of the

multipurpose resources will be further investigated in future research.

Acknowledgment

This research was supported by the National Science Council of Taiwan, Republic of China, under Grant NSC-102-2221-E-002-104.

References

[1] J. M. Kizza, "Security threats to computer networks," in *Guide to Computer Network Security*, pp. 63–88, Springer, London, UK, 2013.

[2] A. K. Rai, R. R. Tewari, and S. K. Upadhyay, "Different types of attacks on integrated MANET-Internet communication," *International Journal of Computer Science and Security*, vol. 4, no. 3, pp. 265–274, 2010.

[3] D. Ferro and A. Salden, "Self-organizing mobile surveillance security networks," in *Proceedings of the 2nd International Conference on Bio-Inspired Models of Network, Information, and Computing Systems (BIONETICS '07)*, pp. 217–227, December 2007.

[4] R. J. Ellison, D. A. Fisher, R. C. Linger, H. F. Lipson, and T. Longstaff, Survivable network systems: An emerging discipline (No. CMU/SEI-97-TR-013). CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST, 1997.

[5] O. M. Al-Kofahi and A. E. Kamal, "Survivability strategies in multihop wireless networks," *IEEE Wireless Communications*, vol. 17, no. 5, pp. 71–80, 2010.

[6] S. Neumayer and E. Modiano, "Network reliability with geographically correlated failures," in *Proceedings of the IEEE International Conference on Computer Communications (IEEE INFOCOM '10)*, March 2010.

[7] F. Y.-S. Lin, H.-H. Yen, P.-Y. Chen, and Y.-F. Wen, "Evaluation of network survivability considering degree of disconnectivity," in *Hybrid Artificial Intelligent Systems*, pp. 51–58, Springer, Berlin, Germany, 2011.

[8] F. Y.-S. Lin, P.-Y. Chen, Y.-S. Wang, and Y.-Y. Chang, "Network recovery strategies for maximization of network survivability," in *Proceedings of the 7th International Wireless Communications and Mobile Computing Conference (IWCMC '11)*, pp. 1130–1134, July 2011.

[9] S. Skaperdas, "Contest success functions," *Economic Theory*, vol. 7, no. 2, pp. 283–290, 1996.

[10] W. Jiang, Z.-H. Tian, H.-L. Zhang, and X.-F. Song, "A game theoretic method for decision and analysis of the optimal active defense strategy," in *Proceedings of the International Conference on Computational Intelligence and Security (CIS '07)*, pp. 819–823, December 2007.

[11] W. Jiang, B.-X. Fang, H.-L. Zhang, Z.-H. Tian, and X.-F. Song, "Optimal network security strengthening using attack-defense game model," in *Proceedings of the 6th International Conference on Information Technology: New Generations (ITNG '09)*, pp. 475–480, April 2009.

[12] Y.-S. Lin, P.-H. Tsang, C.-H. Chen, C.-L. Tseng, and Y.-L. Lin, "Evaluation of network robustness for given defense resource allocation strategies," in *Proceedings of the 1st International Conference on Availability, Reliability and Security (ARES '06)*, pp. 182–189, April 2006.

[13] G. Levitin and K. Hausken, "Parallel systems under two sequential attacks," *Reliability Engineering and System Safety*, vol. 94, no. 3, pp. 763–772, 2009.

[14] G. Levitin and K. Hausken, "Resource distribution in multiple attacks against a single target," *Risk Analysis*, vol. 30, no. 8, pp. 1231–1239, 2010.

[15] K. Hausken and G. Levitin, "Protection vs. false targets in series systems," *Reliability Engineering and System Safety*, vol. 94, no. 5, pp. 973–981, 2009.

[16] X. Liang and Y. Xiao, "Game theory for network security," *IEEE Communications Surveys & Tutorials*, vol. 15, pp. 472–486, 2013.

[17] T. Alpcan and T. Basar, "An intrusion detection game with limited observations," in *Proceedings of the 12th International Symposium on Dynamic Games and Applications*, Sophia Antipolis, France, July 2006.

[18] R. Machado and S. Tekinay, "A survey of game-theoretic approaches in wireless sensor networks," *Computer Networks*, vol. 52, no. 16, pp. 3047–3061, 2008.

[19] M. H. Hassoun, *Fundamentals of Artificial Neural Networks*, MIT Press, 1995.

[20] S. Roy, C. Ellis, S. Shiva, D. Dasgupta, V. Shandilya, and Q. Wu, "A survey of game theory as applied to network security," in *Proceedings of the 43rd Annual Hawaii International Conference on System Sciences, HICSS-43*, pp. 51–58, January 2010.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 601652, 4 pages
http://dx.doi.org/10.1155/2013/601652

Research Article

Multiagent Consensus Control under Network-Induced Constraints

Won Il Kim,[1,2] Rong Xiong,[1] Qiuguo Zhu,[1] and Jun Wu[1]

[1] *State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China*
[2] *Kimchaek Industry University, Pyongyang 999093, Republic of Korea*

Correspondence should be addressed to Rong Xiong; rxiong@iipc.zju.edu.cn

Received 29 March 2013; Accepted 19 August 2013

Academic Editor: Anyi Chen

Mean consensus problem is studied using a class of discrete time multiagent systems in which information exchange is subjected to some network-induced constraints. These constraints include package dropout, time delay, and package disorder. Using Markov jump system method, the necessary and sufficient condition of mean square consensus is obtained and a design procedure is presented such that multiagent systems reach mean square consensus.

1. Introduction

Cooperative control of networked multiagent systems by information exchange has received extensive attention presently, because of their extensive applications in flocking, swarming, distributed sensor fusion, attitude alignment, and so forth (see [1, 2] for surveys). An important problem for cooperative control is to design an appropriate control law such that a multiagent system reaches consensus in the presence of insecure information exchange. Distributed cooperative control of networked multiagent systems has been investigated in various perspectives [3–7]. In [3], the leaderless consensus problem is studied. The problem of consensus with leader node was researched in [4–7]. For networked multiagent systems of linear dynamics, consensus using state feedback or output feedback was analysed in [8, 9].

Unmodelled time delay during the design phase is an important factor that may affect the performance of dynamical systems. It can even, in some situation, cause instability of a system. In these years, consensus in networked multiagent systems with time delay was discussed using linear matrix inequality (LMI) method [10–12]. In [10], the average-consensus problem for continuous-time multiagents with switching topology and time delay was studied. The work of [11] investigated the average consensus problem in undirected networks with fixed and switching topologies under time-varying communication delays. The consensus problem was solved in [12] on directed graphs of the multiagent system with model uncertainty and time delay.

In the information exchange of network, there are not only time-delay but also other network-induced constraints. The other network-induced constraints, which include package dropout and package disorder, also affect the consensus of networked multiagent systems. However, not many works have studied multiagent systems with these network-induced constraints. Based on Markov jump system method [13–15], this paper considers mean square consensus of multiagent systems of first-order integrator under network-induced constraints such as package dropout, time delay, and package disorder. By system transformation, the necessary and sufficient condition of mean square consensus problem is provided and a corresponding design algorithm is given.

The remainder of the paper is organized as follows. Section 2 contains the formulation of the problem and terminology. The main results are presented in Section 3. Section 4 provides the numerical simulation and Section 5 draws conclusions.

2. Problem Formulation and Preliminaries

In this paper, Z^+ is used to denote the set of all nonnegative integers. The $n\times n$ identity matrix is denoted by I_n. The ith row of I_n is denoted by $e(i,n)$. If a matrix P is positive (negative) definite, it is denoted by $P > 0 (< 0)$. The notation # within a matrix represents the symmetric term of the matrix. The expected value is represented by $E[\bullet]$.

Here a discrete-time system is considered that it consists of 2 agents. Each agent is a first-order integrator, which,

$$\begin{aligned} x_1(k+1) &= x_1(k) + bu_1(k), \\ x_2(k+1) &= x_2(k) + bu_2(k) \quad k \in Z^+, \end{aligned} \tag{1}$$

where $x_1(k) \in R$ and $x_2(k) \in R$ are the state at time step k,$u_1(k) \in R$ and $u_2(k) \in R$ are the control at time step k, and $b \in R$ is constant. The two agents exchange their state through two communication channels: channel no. 1 and channel no. 2. At each k, agent 1 transmits $x_1(k)$ to agent 2 through channel no. 1. Agent 2 utilizes $z(k)$ as the information obtained from channel no. 1 at k. Due to random package dropout, time delay, and package disorder in communication, the receiving scenarios in the side of agent 2 at k are various. Agent 2 may receive one package $x_1(k-t)$ from channel no. 1 at k. The package $x_1(k-t)$ is sent by agent 1 at $k-t$ no later than k. After received, $x_1(k-t)$ is examined to see whether it is of disorder (i.e., whether agent 2 has received any packages sent later than $k-t$). If $x_1(k-t)$ is not of disorder, $z(k) \leftarrow x_1(k-t)$. If $x_1(k-t)$ is of disorder, $x_1(k-t)$ is discarded and $z(k) \leftarrow z(k-1)$. Agent 2 may receive severe package $x_1(k-t_1), x_1(k-t_2), \ldots, x_1(k-t_d)$ from channel no. 1 at k. Except the newest $x_1(k-t^*)$ with $t^* = \min(t_1, t_2, \ldots, t_d)$, these packages are discarded. If $x_1(k-t^*)$ is not of disorder, $z(k) \leftarrow x_1(k-t^*)$. If $x_1(k-t^*)$ is of disorder, it is also discarded and $z(k) \leftarrow z(k-1)$. Agent 2 may receive no package from channel no. 1 at k. In this case, $z(k) \leftarrow z(k-1)$.

From the above mechanism, it is seen that $z(k) = x_1(k-\alpha_k)$ with some random $\alpha_k \in Z^+$ constrained by

$$\alpha_{k+1} \le \alpha_k + 1 \quad \forall k \in Z^+. \tag{2}$$

On α_k, we adopt an assumption which is made by some researchers in networked control [16]; that is, α_k is assumed to be a Markov chain taking values in a finite set $\{0, 1, \ldots, m\}$ with transition probabilities:

$$\phi_{s,l} = \Pr\left(\alpha_{k+1} = l \mid \alpha_k = s\right) \quad \forall s, l \in \{0, 1, \ldots, m\}, \tag{3}$$

where m is a given nonnegative integer. The transition probability matrix

$$\Phi = \begin{bmatrix} \phi_{0,0} & \phi_{0,1} & 0 & \cdots & 0 \\ \phi_{1,0} & \phi_{1,1} & \phi_{1,2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi_{m-1,0} & \phi_{m-1,1} & \phi_{m-1,2} & \cdots & \phi_{m-1,m} \\ \phi_{m,0} & \phi_{m,1} & \phi_{m,2} & \cdots & \phi_{m,m} \end{bmatrix} \in R^{(m+1)\times(m+1)} \tag{4}$$

is also known. The expression (4) of Φ displays that, for the reason of constraint (2), $\phi_{s,l} = 0$ when $l > s+1$. Thus, we have described the communication in channel no. 1 using Markov chain α_k. The same method is applied to describe the communication in channel no. 2. Agent 1 obtains $x_2(k-\beta_k)$ from channel no. 2 at k, where β_k is a Markov chain taking values in a known set $\{0, 1, \ldots, n\}$ with a known transition probability matrix

$$\Psi = \begin{bmatrix} \varphi_{0,0} & \varphi_{0,1} & 0 & \cdots & 0 \\ \varphi_{1,0} & \varphi_{1,1} & \varphi_{1,2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \varphi_{n-1,0} & \varphi_{n-1,1} & \varphi_{n-1,2} & \cdots & \varphi_{n-1,n} \\ \varphi_{n,0} & \varphi_{n,1} & \varphi_{n,2} & \cdots & \varphi_{n,n} \end{bmatrix} \in R^{(n+1)\times(n+1)}. \tag{5}$$

The goal of agents 1 and 2 is a prescribed state $x^* \in R$. In this paper, agent 1 is aware of x^* while agent 2 is not. Consequently, agent 1 employs control law

$$u_1(k) = -h\left(\left(x_2(k-\beta_k) - x_1(k)\right) - \left(x_1(k) - x^*\right)\right) \tag{6}$$

but agent 2 employs control law

$$u_2(k) = -h\left(x_1(k-\alpha_k) - x_2(k)\right), \tag{7}$$

where $h \in R$ is the control parameter.

The above multiagent system is said to be mean square consensus if $\forall x_1(0) \in R$, $\forall x_2(0) \in R$, $\forall \alpha_0 \in \{0, 1, \ldots, m\}$, $\forall \beta_0 \in \{0, 1, \ldots, m\}$,

$$\begin{aligned} \lim_{k\to\infty} E\left[\left(x_1(k) - x^*\right)^2\right] &= 0 \\ \lim_{k\to\infty} E\left[\left(x_2(k) - x^*\right)^2\right] &= 0. \end{aligned} \tag{8}$$

Our objective is to design h such that the two agents reach mean square consensus.

3. Main Result

Define

$$\begin{aligned} y_1(k) &= x_1(k) - x^* \\ y_2(k) &= x_2(k) - x^* \quad \forall k \in Z^+. \end{aligned} \tag{9}$$

Then from (1), (6), (7), and (9), we have

$$\begin{aligned} y_1(k+1) &= (1+2bh)\, y_1(k) - bhy_2(k-\beta_k) \\ y_2(k+1) &= -bhy_1(k-\alpha_k) + (1+bh)\, y_2(k). \end{aligned} \tag{10}$$

Further, denote

$$Y(k) = \begin{bmatrix} y_1(k) \\ \vdots \\ y_1(k-m) \\ y_2(k) \\ \vdots \\ y_2(k-n) \end{bmatrix}^T \in R^{m+n+2}. \tag{11}$$

Obviously, mean square consensus (8) is equivalent to $\lim_{k\to\infty}E[Y^{\mathrm{T}}(k)Y(k)] = 0$. Using (11), system (10) is transformed into

$$Y(k+1) = G_h(\alpha_k, \beta_k)\, Y(k), \tag{12}$$

where

$$\begin{aligned}
G_h(\alpha_k, \beta_k) &= \begin{bmatrix} G_{h11} & G_{h12}(\beta_k) \\ G_{h21}(\alpha_k) & G_{h22} \end{bmatrix} \in R^{(m+n+2)\times(m+n+2)}, \\
G_{h11} &= \begin{bmatrix} 1+2bh & 0 & \cdots & 0 & 0 \\ 1 & & & & 0 \\ & 1 & & & 0 \\ & & \ddots & & \vdots \\ & & & 1 & 0 \end{bmatrix} \in R^{(m+1)\times(m+1)}, \\
G_{h12}(\beta_k) &= bh \begin{bmatrix} -e(\beta_k, n+1) \\ 0 \end{bmatrix} \in R^{(m+1)\times(n+1)}, \\
G_{h21}(\beta_k) &= bh \begin{bmatrix} -e(\alpha_k, m+1) \\ 0 \end{bmatrix} \in R^{(n+1)\times(m+1)}, \\
G_{h22} &= \begin{bmatrix} 1+bh & 0 & \cdots & 0 & 0 \\ 1 & & & & 0 \\ & 1 & & & 0 \\ & & \ddots & & \vdots \\ & & & 1 & 0 \end{bmatrix} \in R^{(n+1)\times(n+1)}.
\end{aligned} \tag{13}$$

On system (12), [16] presented the following.

Lemma 1. *Suppose that Markov chains α_k and β_k are independent. System* (12) *achieves* $\lim_{k\to\infty}E[Y^T(k)Y(k)] = 0$ *if and only if there exist positive definite matrices $P(\alpha,\beta) \in R^{(m+n+2)\times(m+n+2)}$, $\alpha \in \{0,1,\ldots,m\}$, $\beta \in \{0,1,\ldots,n\}$ such that*

$$P(\alpha,\beta) - G_h^T(\alpha,\beta)\left(\sum_{i=0}^{m}\sum_{j=0}^{n}\phi_{\alpha,i}\varphi_{\beta,j}P(i,j)\right)G_h(\alpha,\beta) > 0. \tag{14}$$

Actually, the condition in Lemma 1 can be converted using Schur complement.

Lemma 2 (see [17]). *Let $S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$ be given partitioned matrix. Then $S > 0$ if and only if $S_{22} > 0$ and $S_{11} - S_{12}S_{22}^{-1}S_{21} > 0$.*

Through the above converting, the necessary and sufficient condition is obtained from mean square consensus of the multiagent system.

Theorem 3. *Suppose that Markov chains α_k and β_k are independent. The multiagent system in Section 2 achieves mean square consensus if and only if there exist $h \in R$ and positive definite matrices $P(\alpha,\beta) \in R^{(m+n+2)\times(m+n+2)}$, $Q(\alpha,\beta) \in R^{(m+n+2)\times(m+n+2)}$, $\alpha \in \{0,1,\ldots,m\}$, $\beta \in \{0,1,\ldots,n\}$ such that*

$$\begin{bmatrix} P(\alpha,\beta) & & & \# \\ \sqrt{\phi_{\alpha,0}\varphi_{\beta,0}}G_h(\alpha,\beta) & Q(0,0) & & \\ \vdots & & \ddots & \\ \sqrt{\phi_{\alpha,m}\varphi_{\beta,n}}G_h(\alpha,\beta) & & & Q(m,n) \end{bmatrix} > 0 \tag{15}$$

$$P(\alpha,\beta) - Q^{-1}(\alpha,\beta) = 0.$$

In order to deal with the condition in Theorem 3 using Cone Complementarity Linearisation algorithm [18], for $r \in Z^+$, we construct LMI

$$\begin{bmatrix} P_r(\alpha,\beta) & & & \# \\ \sqrt{\phi_{\alpha,0}\varphi_{\beta,0}}G_h(\alpha,\beta) & Q(0,0) & & \\ \vdots & & \ddots & \\ \sqrt{\phi_{\alpha,m}\varphi_{\beta,n}}G_h(\alpha,\beta) & & & Q(m,n) \end{bmatrix} > 0 \tag{16}$$

$$\begin{gathered}
\begin{bmatrix} P_r(\alpha,\beta) & I \\ I & Q_r(\alpha,\beta) \end{bmatrix} > 0 \\
P_r(\alpha,\beta) \in R^{(m+n+2)\times(m+n+2)}, \\
Q_r(\alpha,\beta) \in R^{(m+n+2)\times(m+n+2)}, \\
\alpha \in \{0,1,\ldots,m\}, \ \beta \in \{0,1,\ldots,n\}
\end{gathered} \tag{17}$$

which is denoted by $L(P_r(\alpha,\beta), Q_r(\alpha,\beta), h) > 0$. The following is our design steps:

Step 1. Specify an enough small real number $\varepsilon > 0$ and an enough large integer T. Set $r = 0$. Find feasible $P_0(\alpha,\beta)$, $Q_0(\alpha,\beta)$, and h_0, $\forall\alpha \in \{0,1,\ldots,m\}$ and $\forall\beta \in \{0,1,\ldots,n\}$ satisfy $L(P_0(\alpha,\beta), Q_0(\alpha,\beta), h_0) > 0$. If there is none, exit.

Step 2. Solve the LMI problem

$$\begin{aligned}
\mu_{r+1} = \min \ \mathrm{trace} \sum_{\alpha=0}^{m}\sum_{\beta=0}^{n} & P_{r+1}(\alpha,\beta)\, Q_r(\alpha,\beta) \\
& + P_r(\alpha,\beta)\, Q_{r+1}(\alpha,\beta) \\
\text{s.t.} \quad & L(P_{r+1}(\alpha,\beta), Q_{r+1}(\alpha,\beta), h_{r+1}) > 0,
\end{aligned} \tag{18}$$

and obtain $P_{r+1}(\alpha,\beta)$, $Q_{r+1}(\alpha,\beta)$ and h_{r+1}.

Step 3. If $|\mu_{r+1} - 2(m+1)(n+1)(m+n+2)| < \varepsilon$, let $h = h_{r+1}$ and terminate. Otherwise, $r \leftarrow r+1$ and go to Step 4.

Step 4. If $r > T$, exist. Otherwise, go to Step 2

It should be pointed out that the above method is easy to be extended to q agents when $q > 2$. Among q agents, there are $q(q-1)$ communication channels. We utilize $q(q-1)$ independent Markov chains to describe communication in these channels and can arrive at a similar result as Theorem 3.

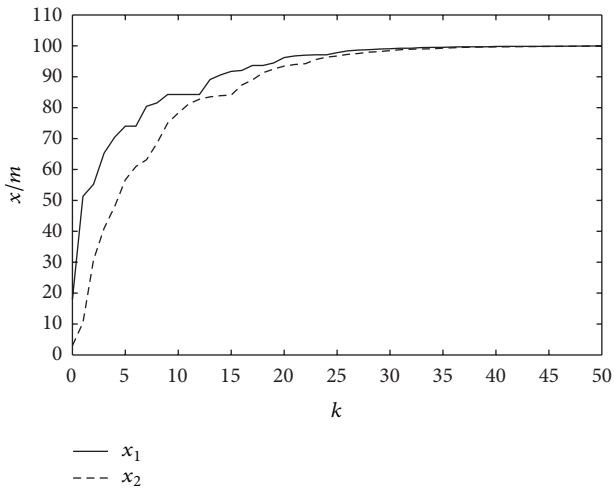


Figure 1: State response of two agents.

4. Numerical Example

In the numerical example, we give $m = 3$, $n = 3$, $b = 0.8$, and transition probability of α_k and β_k is given as

$$\Phi = \begin{bmatrix} 0.6 & 0.4 & 0 & 0 \\ 0.25 & 0.3 & 0.45 & 0 \\ 0.2 & 0.2 & 0.1 & 0.5 \\ 0.1 & 0.55 & 0.1 & 0.25 \end{bmatrix},$$
$$\Psi = \begin{bmatrix} 0.7 & 0.3 & 0 & 0 \\ 0.55 & 0.35 & 0.1 & 0 \\ 0.2 & 0.4 & 0.25 & 0.15 \\ 0.3 & 0.35 & 0.15 & 0.2 \end{bmatrix}. \tag{19}$$

Using the design steps in Section 3, we get $h = -0.6215$. Figure 1 shows the state response of two agents with

$$x_1(0) = 18, \qquad x_2(0) = 3,$$
$$x^* = 100, \quad \alpha_0 = 0, \quad \beta_0 = 0. \tag{20}$$

It can be seen that x_1 and x_2 converge at x^*.

5. Conclusion

The consensus control problem of multiagent systems of first-order integrator is studied under network-induced constraints. A new model is presented to describe the network communication with package dropout, time delay, and package disorder. For the new model, the definition of mean square consensus is given multiagent systems. Further, the necessary and sufficient condition of mean square consensus is proposed in the form of matrix inequalities. Based on this condition and Cone Complementarity Linearisation algorithm, a consensus control law can be designed to make systems reach mean square consensus.

References

[1] R. O. Saber, J. Fax, and R. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[2] W. Ren, R. Beard, and E. Atkins, "A survey of consensus problems in multi-agent coordination," in *Proceedings of the American Control Conference (ACC '05)*, pp. 1859–1864, Portland, Ore, USA, June 2005.

[3] W. Ren, R. Beard, and E. Atkins, "Information consensus in multivehicle cooperative control," *IEEE Control Systems Magazine*, vol. 27, no. 2, pp. 71–82, 2007.

[4] Y. Hong, J. Hu, and L. Gao, "Tracking control for multi-agent consensus with an active leader and variable topology," *Automatica*, vol. 42, no. 7, pp. 1177–1182, 2006.

[5] X. Li, X. Wang, and G. Chen, "Pinning a complex dynamical network to its equilibrium," *IEEE Transactions on Circuits and Systems*, vol. 51, no. 10, pp. 2074–2087, 2004.

[6] X. F. Wang and G. Chen, "Pinning control of scale-free dynamical networks," *Physica A*, vol. 310, no. 3-4, pp. 521–531, 2002.

[7] W. Ren, K. Moore, and Y. Chen, "High-order and model reference consensus algorithms in cooperative control of multi-vehicle systems," *Journal of Dynamic Systems, Measurement, and Control*, vol. 129, no. 5, pp. 678–688, 2007.

[8] H. Zhang, F. L. Lewis, and A. Das, "Optimal design for synchronization of cooperative systems: state feedback, observer and output feedback," *IEEE Transactions on Automatic Control*, vol. 56, no. 8, pp. 1948–1952, 2011.

[9] C.-Q. Ma and J.-F. Zhang, "Necessary and sufficient conditions for consensusability of linear multi-agent systems," *IEEE Transactions on Automatic Control*, vol. 55, no. 5, pp. 1263–1268, 2010.

[10] P. Lin and Y. Jia, "Average consensus in networks of multi-agents with both switching topology and coupling time-delay," *Physica A*, vol. 387, no. 1, pp. 303–313, 2008.

[11] Y. G. Sun, L. Wang, and G. Xie, "Average consensus in networks of dynamic agents with switching topologies and multiple time-varying delays," *Systems & Control Letters*, vol. 57, no. 2, pp. 175–183, 2008.

[12] P. Lin, Y. Jia, and L. Li, "Distributed robust H_∞ consensus control in directed networks of agents with time-delay," *Systems & Control Letters*, vol. 57, no. 8, pp. 643–653, 2008.

[13] A. H. Tahoun and H.-J. Fang, "Adaptive stabilisation of networked control systems tolerant to unknown actuator failures," *International Journal of Systems Science*, vol. 42, no. 7, pp. 1155–1164, 2011.

[14] J. Xiong and J. Lam, "Stabilization of linear systems over networks with bounded packet loss," *Automatica*, vol. 43, no. 1, pp. 80–87, 2007.

[15] J. Xiong and J. Lam, "Robust H_2 control of Markovian jump systems with uncertain switching probabilities," *International Journal of Systems Science*, vol. 40, no. 3, pp. 255–265, 2009.

[16] Y. Xia, G. P. Liu, M. Fu, and D. Rees, "Predictive control of networked systems with random delay and data dropout," *IET Control Theory Application*, vol. 3, pp. 1476–1486, 2009.

[17] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix inequalities in System and Control Theory*, vol. 15 of *SIAM Studies in Applied Mathematics*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa, USA, 1994.

[18] L. El Ghaoui and L. Oustry, "A cone complementarity linearization algorithm for static output-feedback and related problems," *IEEE Transactions on Automatic Control*, vol. 42, no. 8, pp. 1171–1176, 1997.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 541240, 10 pages
http://dx.doi.org/10.1155/2013/541240

Research Article

Modeling of Location Estimation for Object Tracking in WSN

Hung-Chi Chu,[1] Tsung-Han Lee,[2] Lin-huang Chang,[2] and Chung-Jie Li[1]

[1] *Department of Information and Communication Engineering, Chaoyang University of Technology, No. 168, Ji-Fong East Road, Wu-Fong District, Taichung 41349, Taiwan*
[2] *Department of Computer Science, National Taichung University of Education, No. 140, Min-Sheng Road, Taichung 403, Taiwan*

Correspondence should be addressed to Lin-huang Chang; albertchang04@gmail.com

Received 21 June 2013; Accepted 16 August 2013

Academic Editor: Chih-Hao Lin

Location estimation for object tracking is one of the important topics in the research of wireless sensor networks (WSNs). Recently, many location estimation or position schemes in WSN have been proposed. In this paper, we will propose the procedure and modeling of location estimation for object tracking in WSN. The designed modeling is a simple scheme without complex processing. We will use Matlab to conduct the simulation and numerical analyses to find the optimal modeling variables. The analyses with different variables will include object moving model, sensing radius, model weighting value α, and power-level increasing ratio k of neighboring sensor nodes. For practical consideration, we will also carry out the shadowing model for analysis.

1. Introduction

Recently, with the characteristics of low cost and low power consumption, wireless sensor net-work (WSN) communication has attracted lots of attentions on different applications and services, such as monitoring and object tracking. Location estimation for object tracking is one of the important topics in the research of wireless sensor networks (WSNs) [1, 2]. The wireless sensors not only are able to induce and detect the environmental target and the change of environment but also deal with collected data and send the disposed information to the sink or base station via wireless communication.

For WSN applications, tracking moving objects is one of the important issues. Tracking moving objects is more difficult than sensing objects in fixed area because the objects may move from time to time, and sufficient computing power and storage space are required for disposing information of objects and sending the information to users. In the application of wireless sensor networks, users hope to use sensors to collect needed information, such as temperature, gas concentration, position of wildlife, and so forth. For users, the position of object is an important piece of information. For example, in an intelligent building, when sensors detect a fire event, the firemen wish to know the fire site and moving direction through the sensor nodes so that the action can be carried out promptly. Therefore, to acquire the accurate position of the object is one of issues concerning object tracking.

Another important research issue is information processing. Since the object may move everywhere at any moment, the information of object should be updated and sent to the sink timely so that we can maintain the object location accurately. If the monitoring area is very large, the sensing coverage and data aggregation technology needs to be taken into consideration to save energy in terms of sensing and transmission. Therefore, in the paper we will propose the procedure and modeling of location estimation for object tracking with energy issues in mind while maintaining the relatively low estimation error.

The rest of this paper is organized as follows. Section 2 describes the research related to the positioning and location estimation for object tracking. The detailed design and modeling of the proposed scheme is illustrated in Section 3. Section 4 provides the analyses and evaluations as well as discussions based on the simulation and numerical results. We summarize this paper and address future work in Section 5.

2. Related Works

Location estimation of object tracking is to define the object movement path and position [3–7]. In recent years, global positioning system (GPS) [8], implemented in real system, is one of the most popular positioning systems for outdoor environment. The average error of GPS is within 3 m. However, the GPS positioning technology is not suitable for wireless sensor networks due to size, cost, and power consumption constraints.

The location estimation for radio communication can be divided into 2 categories: range-based positioning [9–11] and range-free positioning [12, 13]. The range-based positioning technologies include angle of arrival (AOA) [9], time of arrival (TOA) [9], time difference of arrival (TDOA) [9], and received signal strength indicator (RSSI) [11]. The range-based technology in general can obtain more accurate location estimation by applying the distance and angle information received from specific or expensive network equipments; however, it might not be suitable for large-scale deployment of WSNs.

The range-free positioning technology, on the other hand, estimates the location without the assist from any expensive network equipment. The centroid positioning mechanism [13] employed the centroid of selected reference points to estimate its own position. The general scheme for object tracking is taking the object as the center of a circle and estimating the signal radius with the consideration of moving speed. Besides this general scheme, some other researches [14, 15] have been studied. The research in [14] proposed a method using a mobile agent to track the object moving path and three fixed sensor nodes to calculate the object position. These three sensor nodes broadcast a message to their neighboring nodes for object detection. In order to obtain more accurate object position, some researches [3, 4] defined some sensor nodes within certain range from the object to help tracking the object. Other sensor nodes beyond this range maintain in dormant states. The research in [15] proposed the dynamic adjustment of the signal coverage based on the object behavior to reduce the tracking area in wireless sensor networks. The research can track the object with minimum number of sensor nodes so that the network lifetime can be prolonged.

The deployment of WSNs usually consists of a great number of sensors. To manage the sensor node resources and collected data, some schemes need to be taken into account to provide efficient data processing or message transfer, especially in large-scale WSNs. In general, there are two types of information processing in WSNs, named distributed processing [12, 13, 16, 17] and centralized processing [14, 18–21]. For the distributed processing, when the sensor nodes sense and collect data, they will calculate or process the data followed by sending the data to the sink. GPS system is one typical position system using distributed processing [8]. GPS estimates more accurate location; however, it takes a longer time to first fix (TTFF) and incurs the additional cost of setting up a GPS receiver for each sensor node. Some other researches using related positions of the nodes [16] or area-based position schemes [13] for distributed data processing were also proposed. The research in [12] narrows down the possible region, formed by selecting three anchors among all sensing nodes, in which a particular node may reside. The location of an object can be determined by the center of gravity of the intersection of triangles. To reduce the number of anchors, the research in [17] employed a few mobile anchors equipped with the GPS capability to broadcast their current positions periodically for location estimation.

On the other hand in centralized processing, upon receiving the sensed data, the sensor nodes send it to the sink via some routing protocols for data processing. The data aggregation technology [18–21] might be employed before the data reach the sink so as to reduce the amount of data transfer and consequently provide the energy saving. The research in [14] set a data volume threshold for the detected object data in WSNs. When the data volume detected by an agent node is smaller than the threshold, the agent node will send the data to the sink directly. If the collected data volume is larger than the threshold, the agent node will carry out the data fusion for all collected data and then send them to the sink.

From the related researches discussed above, not all sensor nodes can afford the GPS capability due to the limitations of sensor nodes in size, cost, and power consumption. Also due to the limited computational power of sensor nodes, simpler positioning and data processing mechanisms will be the major consideration in this paper. Therefore, this paper aims at the scenario wherein static sensing nodes are deployed in fixed location. The proposed modeling of location estimation for object tracking developed using range-free positioning technology based on centroid scheme [13] as well as centralized data processing technology to reduce the data processing and traffic loads is suitable for large-scale WSNs. We will also use data aggregation idea to reduce the data volume transfer between the sensing nodes and sink. From the result of related researches above, the range-free positioning technology combined with data aggregation used in the proposed mechanism is suitable for practical use in large-scale WSNs which are constrained in terms of energy consumption, computation power, and device cost. Furthermore, the proposed mechanism using dynamic sensing procedure with different sensing radii and power-level of the transmission signal will improve the positioning accuracy as compared to other related schemes.

3. Proposed Location Estimation Scheme

3.1. Dynamic Sensing Procedure. In this paper, we will make the following assumptions for our location estimation model. First, we assume that the WSN nodes deployment is in a 2-dimensional plane. Secondly, there is no interference between any two sensor nodes. Thirdly, all sensor nodes in WSN have the knowledge of their own IDs and corresponding GPS positions as well as sink GPS value. Lastly, we assume that every node knows its neighboring nodes' ID. All the above mentioned data can be set up or acquired during the initial deployment of the WSN. The sink node maintains the record of IDs and GPS positions of all sensor nodes. Within the sensing range, R, each sensor node can detect the movement

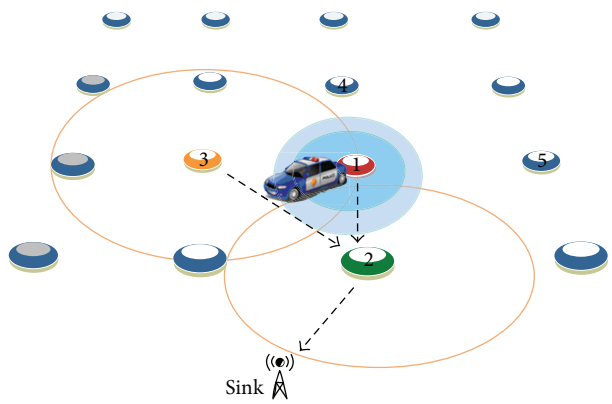


Figure 1: The dynamic power-level sensing topology.

of the object. In this research, we deploy a square network nodes topology to provide fully coverage of sensing network. The proposed location estimation for object tracking in WSN includes three steps.

Step 1. Dynamic power-level sensing: once detecting the appearance of one object, the sensor node informs the one-hop neighboring node to increase the power level for sensing.

Step 2. Cluster head selection: designate a node from the sensing nodes which is the closest node to the sink. The selected node, named cluster head, will collect all data from sensing nodes and then send the data to the sink.

Step 3. Modeling of the location estimation: upon receiving the data from the cluster head, the sink calculates the object location using the location estimation model.

The detailed modeling and mechanism of the proposed location estimation for object tracking is described below.

While estimating the position of moving object in WSN, the sensing node calculates the position according to models such as [13]. In general, using more sensing nodes to detect the object, the estimated location will be more accurate. To increase the accuracy of location estimation for object tracking, the first step in our proposed mechanism is to dynamically adjust the sensing power-level of neighboring nodes which is one hop away from the initial sensing node. The one-hop neighboring nodes increase their power by k time to extend their corresponding signal coverage. The dynamic power-level sensing topology and scheme is illustrated in Figure 1, where sensor node 1 is the initial sensing node to detect an object. Once detecting the appearance of the object, the sensing node, node 1, issues a message, with TTL setting as 1, to its one-hop neighboring nodes, nodes 2, 3, 4, and 5 as shown in Figure 1. Upon receiving the message from the initial sensing node, the one-hop neighboring nodes extend their corresponding signal coverage by increasing the sensing power-level to k time. The moving object location can be calculated using location estimation model, described later, from all sensible nodes corresponding to nodes 1, 2, and 3 as shown in Figure 1 for example.

The next step is to send sensed information back to the sink node by each sensible node. In this research, we designate a node from the sensible nodes being the closest node to the sink as the cluster head. The cluster head will collect all data from sensible nodes and then send the data to the sink. This will reduce the unnecessary messages sending from other sensible nodes, except the cluster head, to the sink.

Once the cluster head is determined, all other sensible nodes will send the sensing information to the cluster head. After the data fusing, the cluster head sends the sensing object information to the sink. Upon receiving the data from the cluster head, the sink calculates the object location using the location estimation model. The modeling of location estimation for object tracking is described in more details in Section 3.2.

3.2. Modeling of Location Estimation for Object Tracking. Because the dynamic power-level sensing is applied to the neighboring sensor nodes which would be different from the initial sensor node, we define the initial sensor node as major node, node 1 in Figure 1, and the other sensible neighboring nodes as minor nodes, for example, nodes 2 and 3 in Figure 1. The algorithm with distance formula for our location estimation is illustrated in Algorithm 1.

In our proposed algorithm, there are two types of sensing data between sensor nodes to the cluster head. For those nodes without changing the sensing range, they will send an *Mposition (cluster, ID)* packet to the cluster head, where *cluster* is the cluster head sensor ID, and *ID* is the sensor node ID. If a sensor node, increasing its power by k time, detects the object, it will send an *Nposition (cluster, ID)* packet to the cluster head. After receiving the information from all sensible nodes, the cluster head processes and compresses the data and then sends it to the sink. Upon receiving the data from the cluster, the sink calculates the object location according to the following equation:

$$\left(O_x, O_y\right) = \left(\frac{\alpha \sum_{i=1}^{z} M_{x_i} + \sum_{i=1}^{j} N_{x_i}}{\alpha z + j}, \frac{\alpha \sum_{i=1}^{z} M_{y_i} + \sum_{i=1}^{j} N_{y_i}}{\alpha z + j}\right), \quad (1)$$

where the (O_x, O_y), (M_x, M_y), and (N_x, N_y) are the coordinate of object, major node, and minor node, respectively. Also, the z and j are the numbers of the major nodes and minor nodes, respectively. In (1), we add a weighting value α to the estimated formula because sensing ranges for major nodes and minor nodes are different. The α value could be related to the increase of the power level k for minor nodes. When we conduct the performance analysis, we compare the numerical result for different k and α values. The algorithm for our proposed location estimation model is shown in Algorithm 2.

Since the practical wireless communications will suffer from all kinds of interference and signal fading issue, the idea model with circular sensing coverage may not be feasible. Therefore, we further consider the shadowing model in our

(1) // Given a graph $G = (V, E)$
(2) // V presents the set of sensor nodes
(3) // E presents the set of communication link between sensor node and its neighbors
(4) // MPosition (M_x, M_y) is the coordinator of the major node
(5) // NPosition (N_x, N_y) is the coordinator of the minor node
(6) // R presents the sensing range of sensor node
(7) $S = \{s_a\}$, for $a = 1, 2, \ldots, i$, and $S \in V$ // S presents the set of sensor nodes detecting object O
(8) $N = \{n_b\}$, for $b = 1, 2, \ldots, j$, and $N \in V$ // N presents the set of node's neighboring nodes with enlarged power, which can detect object O
(9) $dist(O, s_a) = \sqrt{\left(O_x - s_{a_x}\right)^2 + \left(O_y - s_{a_y}\right)^2}$
(10) if $dist(O, s_a) < R$ then // The proposed method will be triggered when a sensor node S_a detects an object O.
(11) $dist(O, n_b) = \sqrt{\left(O_x - n_{b_x}\right)^2 + \left(O_y - n_{b_y}\right)^2}$ // In order to determine object's location, the sensor node notifies its one-hop neighboring nodes to increase their sensing range.
(12) end if

Algorithm 1: Algorithm with distance formula for the proposed location estimation.

(1) *//When Object O into the Sensing Filed*
(2) *//The set of Sensor nodes* $S = \{s_1, s_2 \ldots, s_i\}$ *detect Object O*
(3) *//The set of neighbor nodes* $N = \{n_1, n_2 \ldots, n_j\}$ *detect Object O*
(4) for $(a = 1; a \leq i; a{+}{+})$
(5) if $dist(O, s_a) < R$
(6) s_a Send *enlarge(ID)* to one-hop neighbors;
(7) end if
(8) Broadcast *notice(ID)*;
(9) Broadcast *contend(ID, dist(sink,* s_a*))*;
(10) if *dist(sink,* $S + N - s_a$*) < dist(sink,* s_a*)* then
(11) s_a.State = give up;
(12) else
(13) s_a.State = cluster;
(14) end if
(15) if s_a.State = cluster then
(16) Send *cluster(ID)* to $S + N - s_a$;
(17) else if *dist* $(O, s_a) < R$
(18) Send Mposition*(Cluster, ID)* to Cluster;
(19) else
(20) Send Nposition*(Cluster, ID)* to Cluster;
(21) end if
(22) if s_a.State = cluster then
(23) Send all information to sink;
(24) end if
(25) end for

Algorithm 2: Algorithm of the proposed location estimation model.

design and analysis. We model the shadowing issue with random process. Basically, when the object is close to the sensor node, the sensed probability would be higher than that of the object being remote. The shadowing model applied in our study is shown in the following:

$$\left[\frac{P_r(d)}{P_r(d_0)}\right]_{db} = -10\beta \log\left(\frac{d}{d_0}\right) + X_{dB}, \tag{2}$$

where X_{db} is the random variable with Gaussian distribution, β is the path loss exponent, d is the distance between the object and the sensing node, and d_0 is the sensing radius.

4. Numerical Analyses

The performance analysis is conducted by using Matlab software and the experimental parameters; setup and mobility

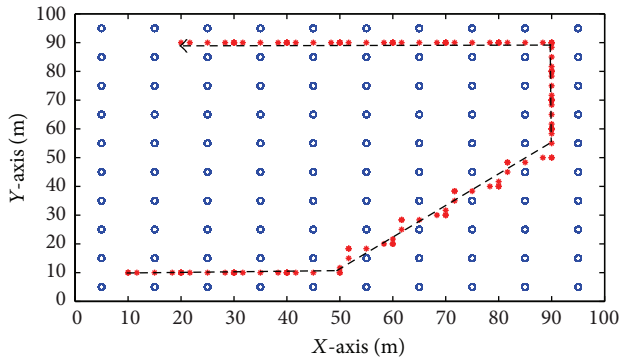


Figure 2: Tracking of moving object with line movement path 1 (black broken lines are the real movements, and red dots stand for predicted object locations).

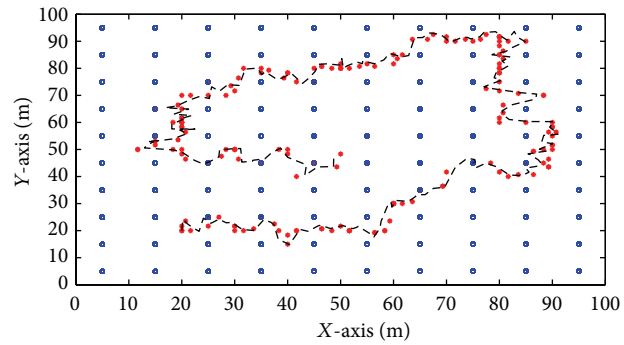


Figure 3: Tracking of moving object with random movement path 2 (black broken lines are the real movements, and red dots stand for predicted object locations).

models are defined according to the survey in [22]. We deployed 100 sensor nodes in a 100 m × 100 m environment. The sensor nodes are equally distributed with square shape in the experimental region. The distance between two neighboring nodes is 10 meters. Two object moving paths are considered in the experiment. Path 1 is line movement with 1 m/s moving speed and path 2 corresponds to the random movement with 1–5 m/s moving speed. The initial sensing radius R is set as 8 m, which will be varied for different simulations. The experiment time for the simulation is 180 sec.

The experiments include object moving model, sensing radius, modeling weight value α and power level increasing ratio k of neighboring sensor nodes. The detailed result for each experiment is discussed in Section 4.1.

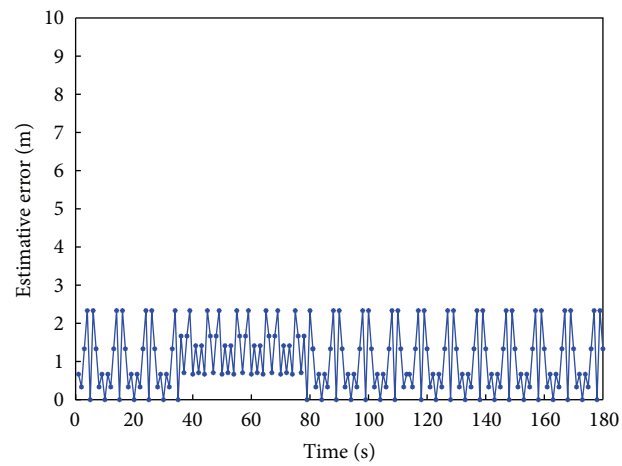


Figure 4: The estimation error for path 1 movement.

4.1. Different Object Moving Models. Figures 2 and 3 illustrate the tracking of moving object with line movement path 1 and random movement path 2, respectively. The location estimation from our model for both path 1 and path 2 are quite close to the real object moving position. The estimations error for path 1 movement is shown in Figure 4 where the x-axis represents the simulation time which corresponds to the object position after movement at each instance. The y-axis in Figure 4 on the other hand stands for the estimation error between the real object position and calculated position. The average estimated error for path 1 movement is 1.17 m with 2.83 m maximum error and standard deviation 0.83. The occurrence of maximum error appears during 35–80 sec simulation time which proceeds with oblique movement instead of horizontal or vertical movements. For rectangular deployment of sensor nodes, the oblique movement of object is expected to come out with large estimation error.

On the other hand, the estimation error for path 2 movement is illustrated in Figure 5. The average estimated error for path 2 random movement is 1.34 m with 3.48 m maximum error and standard deviation 0.62. The random direction characteristics and various moving speeds is the major reason which causes larger estimated error.

4.2. Different Sensing Radii. In general, the difference in sensing radius, resulting in different sensing coverage and consequently different signal overlay area, will affect the accuracy of the location estimation. The change of sensing radius will also affect the numbers of major and minor sensor nodes as well as the estimated position. Figures 6 and 7 show the average estimated error with different sensing radii for path 1 and path 2 movement, respectively. In this research with 10 m separation between each neighboring sensor node, the minimum sensing range would be 7.2 m in order to fully cover the whole experimental environment. However, for the practical deployment while considering the shadowing effects, it is better to provide sensing radius larger than 8 m. On the other hand, larger sensing radius with larger sensing power will introduce the power consumption issue for WSN applications. From the power energy viewpoint, the sensing radius in this design could be between 8 m to 10 m.

As shown from the simulation results in Figure 6, when the sensing radius is between 11.2–11.3 m, the location estimation error reaches the minimum. However, by taking account the energy consumption issue, the sensing radius around 9.7 m with relatively low location estimation error, not larger than 0.1 m as compared to the sensing radius 11.4 m with

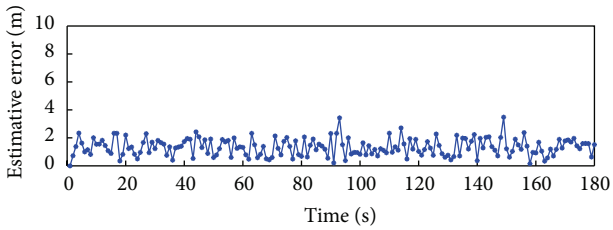


Figure 5: The estimation error for path 2 random movement.

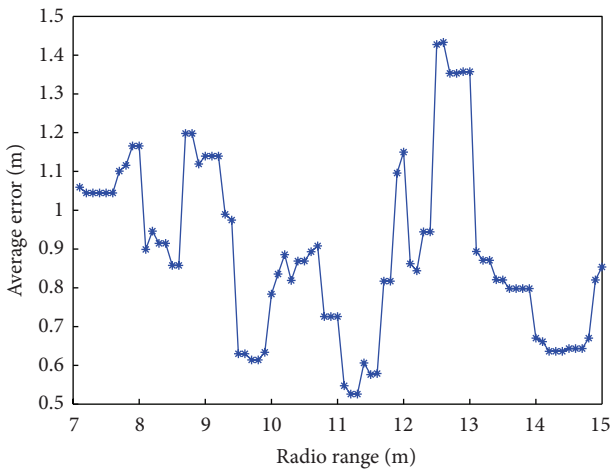


Figure 6: The average estimated error with different sensing radii for path 1 movement.

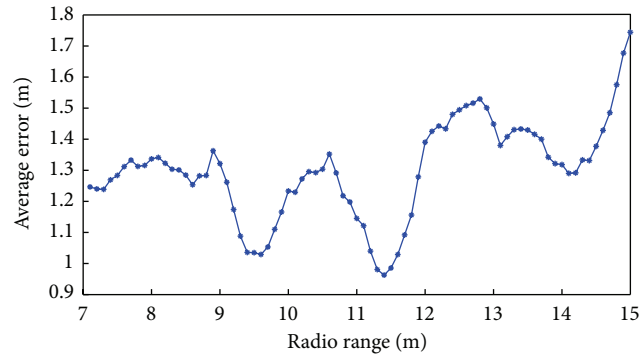


Figure 7: The average estimated error with different sensing radii for path 2 movement.

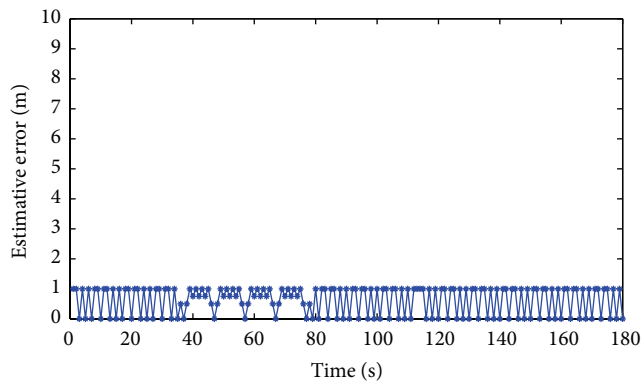


Figure 8: The estimation error result with optimal sensing radius for path 1 movement.

minimum estimation error, could be the optimal candidate in the experimental scenario for path 1 simulation. On the other hand, when the sensing radius is around 11.4 m, the location estimation error reaches the minimum for path 2 case shown in Figure 7. Similarly, we will set the sensing radius 9.6 m as the optimal candidate for energy consumption consideration in path 2 simulation.

With the optimal sensing radius in mind, more accuracy location estimation could be obtained. Figures 8 and 9 show the estimation error result with optimal sensing radius for path 1 and path 2 movement, respectively. As shown in Figure 8 for path 1 case, the average estimation error is about 0.61 m with 1.04 maximum error and standard deviation 0.46 which are much smaller than the results with 8 m sensing radius. Similarly, for the Path 2 movement case in Figure 9 we obtain the average estimation error about 1.03 m with 3.29 m maximum error and standard deviation 0.66. The average estimation error with optimal sensing radius is less than the result with 8 m sensing radius.

4.3. Different Modeling Weight Values. The weight value α of major sensor node will be analyzed in this subsection. Figures 10 and 11 show the dependence of weighting value on the average location estimation error for path 1 movement with sensing radius 9.7 m and path 2 movement with sensing radius 9.6 m, respectively. As shown in Figure 10 for path 1 case, for weighting value between 2 and 3 we will obtain the lowest estimation error. Similarly, we will obtain the relatively low estimation error for weighting value between 2 and 3 in Figure 11 path 2 case. In general, large weighting value (larger than 3) would push the calculation of location estimation too close to major sensor node. On the other hand, small weighting value (smaller than 2) would push the calculation of location estimation too close to minor sensor node. Both situations would increase the estimation error.

4.4. Different Power-Level Increasing Ratios. Figure 12 shows the simulation results of average estimation error with different sensing radii and power-level increasing ratios k for path 1 movement. As shown in Figure 12, when k value of power-level increasing ratio for minor sensor node is larger than 2, the estimation error becomes relatively large. This is because the remote sensor nodes with extended sensing coverage may detect the object and consequently result in the increase of the estimation error. The detailed estimation errors with corresponding variables for some instances of path 1 movement are listed in Table 1. Although the minimum estimation error is obtained with 9.9–10 m sensing radius and $k = 1.7$, the optimal selection would be the case with 9.7 m sensing radius and $k = 1.5$ by considering the energy consumption issue.

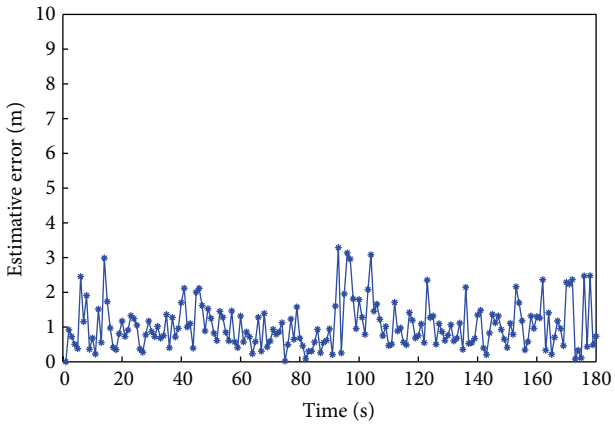


Figure 9: The estimation error result with optimal sensing radius for path 2 movement.

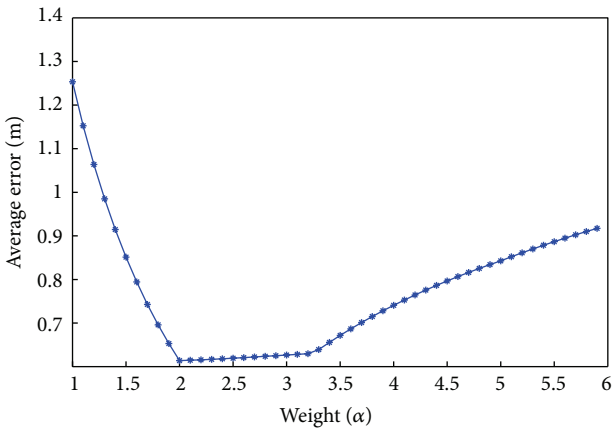


Figure 10: The dependence of weighting value on the average location estimation error for path 1.

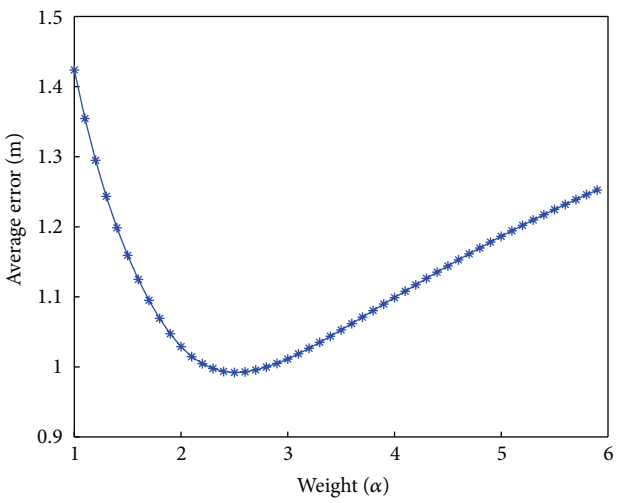


Figure 11: The dependence of weighting value on the average location estimation error for path 2.

Table 1: The estimation errors with corresponding variables for Path 1 movement.

Sensing radius	Power-level (k)	Estimation error
7.1	1.4	0.7001
7.2	1.8	0.7801
7.3	1.3	0.7001
7.4	1.3	0.7001
7.5	1.3	0.7001
7.6	1.3	0.7001
7.7	1.2	0.8337
7.8	1.2	0.8337
7.9	1.2	0.7800
8.0	1.2	0.7435
8.1	1.2	0.7435
8.2	1.2	0.7435
8.3	1.2	0.8352
8.4	1.1	0.9099
8.5	1.5	0.8576
8.6	1.1	0.8062
8.7	2.0	0.9888
8.8	1.9	0.9548
8.9	1.9	0.9334
9.0	1.9	0.8808
9.1	1.9	0.8836
9.2	1.8	0.8497
9.3	1.8	0.8497
9.4	1.8	0.8283
9.5	1.5	0.6297
9.6	1.5	0.6297
9.7	1.5	0.6140
9.8	1.7	0.6139
9.9	1.7	0.5925
10	1.7	0.5925

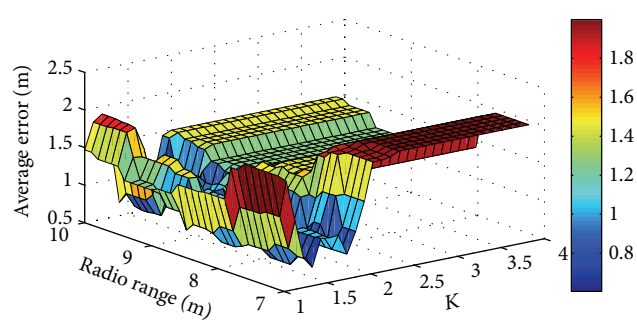


Figure 12: Average estimation error with different sensing radii and power-level ratios k for path 1.

Similar results with different sensing radii and power-level increasing ratios k for path 2 case is shown in Figure 13, and the detailed estimation errors with corresponding variables for path 2 case are listed in Table 2. Again, we may take the optimal sensing radius as 9.6 m and k value as 1.5 for path 2 case from this result.

Table 2: The estimation errors with corresponding variables for Path 2 movement.

Sensing radius	Power-level (k)	Estimation error
7.1	1.3	1.0816
7.2	1.3	1.0725
7.3	1.3	1.0607
7.4	1.3	1.0618
7.5	1.3	1.1070
7.6	1.3	1.1575
7.7	1.2	1.1804
7.8	1.2	1.1950
7.9	1.2	1.2324
8.0	1.6	1.2183
8.1	1.6	1.1980
8.2	1.6	1.2118
8.3	1.6	1.2684
8.4	1.6	1.2453
8.5	1.7	1.2698
8.6	1.5	1.2534
8.7	1.6	1.2341
8.8	1.6	1.1755
8.9	1.6	1.1321
9.0	1.6	1.1219
9.1	1.6	1.1225
9.2	1.6	1.1347
9.3	1.5	1.0880
9.4	1.5	1.0358
9.5	1.5	1.0351
9.6	1.5	1.0290
9.7	1.5	1.0532
9.8	1.5	1.1100
9.9	1.4	1.1194
10	1.4	1.1180

4.5. Result of Shadowing Model. In this subsection, the shadowing model is considered, and the simulation is conducted for 100 times to analyze the result. Figure 14 shows the location estimation error with shadowing model for path 1 movement. The sensing radius is set to 9.7 m. From the result in Figure 14, the estimation error is 1 m with 4.47 m maximum error and standard deviation 0.72. On the other hand, the estimation error with shadowing model for path 2 case with 9.6 m sensing radius is shown in Figure 15. From the result in Figure 15, the estimation error is 1.36 m with 5.43 m maximum error and standard deviation 0.77. The results for path 1 and path 2 for shadowing model are within acceptable range.

4.6. Comparison with Other Schemes. As mentioned earlier, the proposed scheme is based on the low power consumption and computation for moving objecting location tracking. In this sub-section, we therefore, further compare our estimation accuracy with the results of centroid scheme [13]. Figure 16 shows the comparison of location estimation error

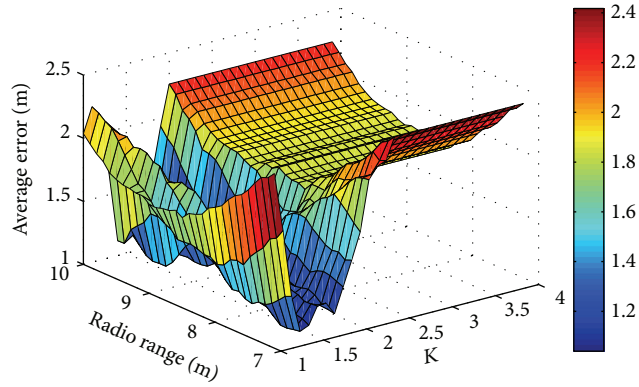


Figure 13: Average estimation error with different sensing radii and power-level ratios k for path 2.

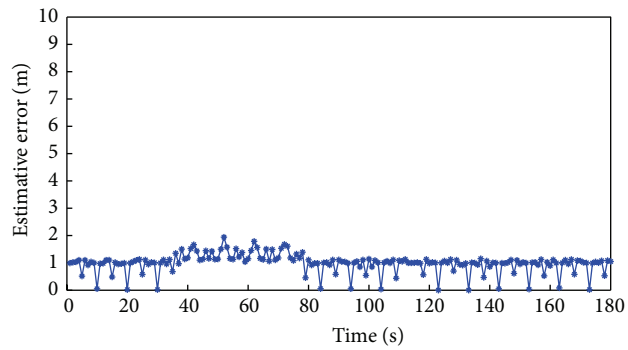


Figure 14: The location estimation error with shadowing model for path 1 movement.

between centroid and our proposed schemes for path 1 movement. As we can see from the results of Figure 16, our proposed scheme performs better estimation accuracy than centroid scheme for most radio sensing radii. The proposed scheme comes out with a little bit higher estimation error around sensing radii from 12.5 m to 13 m. Similarly, the comparison of location estimation error between Centroid and our proposed schemes for path 2 movement is illustrated in Figure 17. Again most results of our proposed scheme show better estimation accuracy than centroid scheme except for the results for radii from 12.6 m to 13.8 m. However, for 100 m $\times$ 100 m experimental setup, the optimal sensing radius would be around 9.6–9.7 m from the previous results. Therefore, from practical deployment viewpoint, our proposed scheme does provide much better accuracy as compared with centroid scheme. In general, the estimation error of our proposed mechanism is less than half of that of centroid scheme.

5. Conclusion

In this paper, we have developed and proposed the mechanism and procedure to model the location estimation for object tracking in large-scale WSNs. The designed modeling is a simple scheme without complex processing which uses range-free-positioning technology as well as centralized data

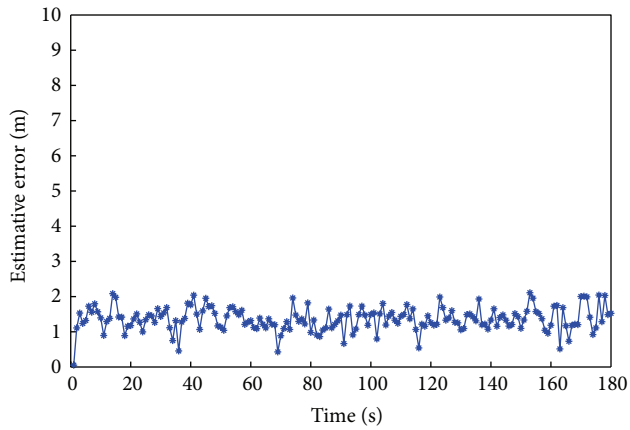


Figure 15: The location estimation error with shadowing model for path 2 movement.

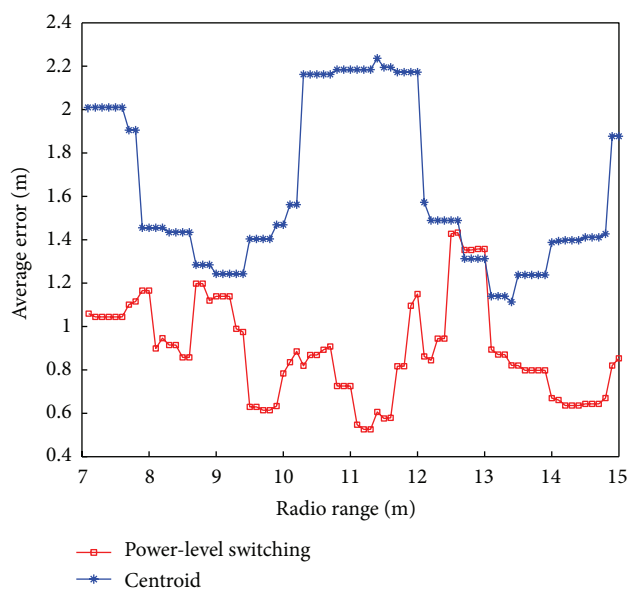


Figure 16: The comparison of location estimation error with centroid model for path 1 movement.

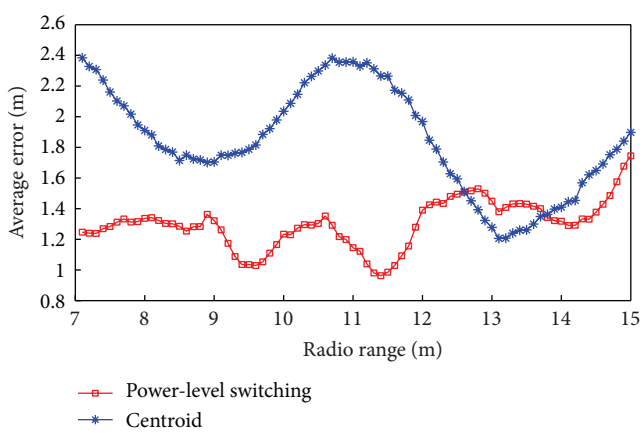


Figure 17: The comparison of location estimation error with centroid model for path 2 movement.

processing technology with data aggregation idea to reduce the data processing and traffic loads. The proposed positioning model and mechanism are suitable for practical use in large-scale WSNs which are constrained in terms of energy consumption, computation power, and device cost.

We have conducted the simulation and numerical analyses on different variables, such as object moving model, sensing radius, model weighting value α, and power-level increasing ratio k of neighboring sensor nodes. The shadowing model was also introduced and analyzed to map the designed scheme to the practical situation. The experimental results showed that the average estimation errors are 0.61 m and 1.03 m with optimized sensing radius around 9.7 m and 9.6 m for path 1 line movement and path 2 random movement, respectively. We have further compared our proposed model and mechanism with centroid scheme. From practical deployment viewpoint, our proposed scheme does provide much better accuracy as compared with centroid scheme. In general, the estimation error of our proposed mechanism is less than half of that of centroid scheme.

In the future, we will investigate the design of mobile sensing nodes to further reduce the number of deployed nodes. We will also conduct experiments on irregular deployment of sensing nodes to simulate some special scenarios or environments.

Acknowledgments

This research was partially supported by the National Science Council of Republic of China, Taiwan, under contracts NSC 102-2221-E-324-023, NSC 101-2221-E-142-003, and NSC 99-2632-E-324-001-MY3.

References

[1] P. Gao, W. Shi, W. Zhou, H. Li, and X. Wang, "A location predicting method for indoor mobile target localization in wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 2013, Article ID 949285, 11 pages, 2013.

[2] O. G. Adewumi, K. Djouani, and A. M. Kurien, "RSSI based indoor and outdoor distance estimation for localization in WSN," in *Proceedings of the IEEE International Conference on Industrial Technology (ICIT '13)*, pp. 1534–1539, February 2013.

[3] T.-S. Chen, W.-H. Liao, M.-D. Huang, and H.-W. Tsai, "Dynamic object tracking in wireless sensor networks," in *Proceedings of the 7th IEEE Malaysia International Conference on Communications*, vol. 1, pp. 475–480, November 2005.

[4] C.-Y. Lin and Y.-C. Tseng, "Structures for in-network moving object tracking in wireless sensor networks," in *Proceedings of the 1st International Conference on Broadband Networks (BroadNets '04)*, pp. 718–727, October 2004.

[5] J. Tan and X. Shan, "Spatiotemporal sensor network and mobile robot coordination in constrained environments," in *Proceedings of the 6th World Congress on Intelligent Control and Automation (WCICA '06)*, vol. 1, pp. 391–396, June 2006.

[6] C.-Y. Lin, W.-C. Peng, and Y.-C. Tseng, "Efficient in-network moving object tracking in wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 5, no. 8, pp. 1044–1056, 2006.

[7] R. R. Brooks, P. Ramanathan, and A. M. Sayeed, "Distributed target classification and tracking in sensor networks," *Proceedings of the IEEE*, vol. 91, no. 8, pp. 1163–1171, 2003.

[8] B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins, *Global Positioning System: Theory and Practice*, Springer, New York, NY, USA, 4th edition, 1997.

[9] M. Vossiek, L. Wiebking, P. Gulden, J. Wieghardt, C. Hoffmann, and P. Heide, "Wireless local positioning," *IEEE Microwave Magazine*, vol. 4, no. 4, pp. 77–86, 2004.

[10] D. Niculescu and B. Nath, "Ad hoc positioning system (APS) using AOA," in *Proceedings of the 22nd Annual Joint Conference on the IEEE Computer and Communications Societies*, vol. 3, pp. 1734–1743, April 2003.

[11] A. Awad, T. Frunzke, and F. Dressler, "Adaptive distance estimation and localization in WSN using RSSI measures," in *Proceedings of the 10th Euromicro Conference on Digital System Design Architectures, Methods and Tools (DSD '07)*, pp. 471–478, August 2007.

[12] T. He, C. Huang, B. Lum, J. Stankovic, and T. Adelzaher, "Range-free localization schemes for large scale sensor networks," in *Proceedings of the ACM Annual International Conference on Mobile Computing and Networking (MobiCom '03)*, September 2003.

[13] N. Bulusu, J. Heidemann, and D. Estrin, "GPS-less low-cost outdoor localization for very small devices," *IEEE Personal Communications*, vol. 7, no. 5, pp. 28–34, 2000.

[14] Y.-C. Tseng, S.-P. Kuo, H.-W. Lee, and C.-F. Huang, "Location tracking in a wireless sensor network by mobile agents and its data fusion strategies," *Computer Journal*, vol. 47, no. 4, pp. 448–460, 2004.

[15] J. Jeong, T. Hwang, T. He, and D. Du, "MCTA: target tracking algorithm based on minimal contour in wireless sensor networks," in *Proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM '07)*, pp. 2371–2375, May 2007.

[16] S. Capkun, M. Hamdi, and J. Hubaux, "GPS-free positioning in mobile ad-hoc networks," in *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, pp. 3481–3490, January 2001.

[17] K.-F. Ssu, C.-H. Ou, and H. C. Jiau, "Localization with mobile anchor points in wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 3, pp. 1187–1197, 2005.

[18] B. Krishnamachari, D. Estrin, and S. Wicker, "The impact of data aggregation in wireless sensor networks," in *Proceedings of the 22nd International Conference on Distributed Computing Systems Workshops*, pp. 575–578, July 2002.

[19] H.-C. Chu and R.-H. Jan, "A GPS-less, outdoor, self-positioning method for wireless sensor networks," *Ad Hoc Networks*, vol. 5, no. 5, pp. 547–557, 2007.

[20] H. C. Chu, L. H. Chang, H. W. Yu, J. J. Liaw, and Y. H. Lai, "Target tracking in wireless sensor networks with guard nodes," *Journal of Internet Technology*, vol. 11, no. 7, 2010.

[21] W. Kim, J. Park, J. Yoo, H. J. Kim, and C. G. Park, "Target localization using ensemble support vector regression in wireless sensor networks," *IEEE Transactions on Cybernetics*, vol. 43, no. 4, pp. 1189–1198, 2013.

[22] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless Communications and Mobile Computing*, vol. 2, no. 5, pp. 483–502, 2002.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 548564, 12 pages
http://dx.doi.org/10.1155/2013/548564

Research Article

A Multistage Control Mechanism for Group-Based Machine-Type Communications in an LTE System

Wen-Chien Hung,[1] Sun-Jen Huang,[1] Feng-Ming Yang,[2] and Chun-Yen Hsu[2]

[1] *Department of Information Management, National Taiwan University of Science and Technology, Taipei, Taiwan*
[2] *Smart Network System Institute, Institute for Information Industry, Taipei, Taiwan*

Correspondence should be addressed to Wen-Chien Hung; d10116002@mail.ntust.edu.tw
and Feng-Ming Yang; fengmingyang@iii.org.tw

Received 18 March 2013; Accepted 17 July 2013

Academic Editor: Anyi Chen

When machine-type communication (MTC) devices perform the long-term evolution (LTE) attach procedure without bit rate limitations, they may produce congestion in the core network. To prevent this congestion, the LTE standard suggests using group-based policing to regulate the maximum bit rate of all traffic generated by a group of MTC devices. However, previous studies on the access point name-aggregate maximum bit rate based on group-based policing are relatively limited. This study proposes a multistage control (MSC) mechanism to process the operations of maximum bit rate allocation based on resource-use information. For performance evaluation, this study uses a Markov chain with $M/G/k/k$ to analyze MTC application in a 3GPP network. Traffic flow simulations in an LTE system indicate that the MSC mechanism is an effective bandwidth allocation method in an LTE system with MTC devices. Experimental results show that the MSC mechanism achieves a throughput 22.5% higher than that of the LTE standard model using the group-based policing, and it achieves a lower delay time and greater long-term fairness as well.

1. Introduction

Machine-type communication (MTC) applications are gradually becoming available for a wide range of potential applications because of the tremendous interest in mobile network operators. This emerging technology is used in machine-to-machine (M2M) communication to provide wireless broadband communications for various types of applications [1, 2]. This dynamic network provides MTC devices with serving network capabilities for various applications, including metering, road security, and consumer electronic devices [3, 4]. The 3GPP TS 22.368 specification initially defined the service requirement for MTC services and MTC devices [5, 6]. The SA2 committee extended architecture requirements to support the aggregate maximum bit rate (AMBR) in quality-of-service (QoS) parameters, creating the TR 23.888 specification for group service application requirements [7]. This extension created new challenges and opportunities in the resource allocation problem. The efficiency of the dynamic bandwidth allocation mechanism plays a vital role in system performance because MTC devices can cause network congestion when performing the attach procedure without bit rate limitations [8]. Figure 1 shows a typical MTC network architecture. In a long-term evolution (LTE) system, the mobility management entity (MME) regulates all communication between the enhanced node B (eNB) and MTC devices in the radio coverage cell of the eNB. The policy and charging rules function (PCRF) is a software node for determining policy rules in a 3GPP network. To meet QoS requirements, the MME can send an attachment report to the Home Subscriber Server (HSS), and the HSS attempts to inform the MME by group-based policing to compute the QoS parameter value. To allow all MTC devices to send the attach request separately, the aggregate maximum bit rate (AMBR) must regulate the bit rate of all traffic generated through a group of nonguaranteed bit rate bearers. In this situation, users are unable to receive proper bandwidth, which causes the attach reject for the user equipment-aggregate maximum bit rate (UE-AMBR) and therefore degrades the QoS [9].

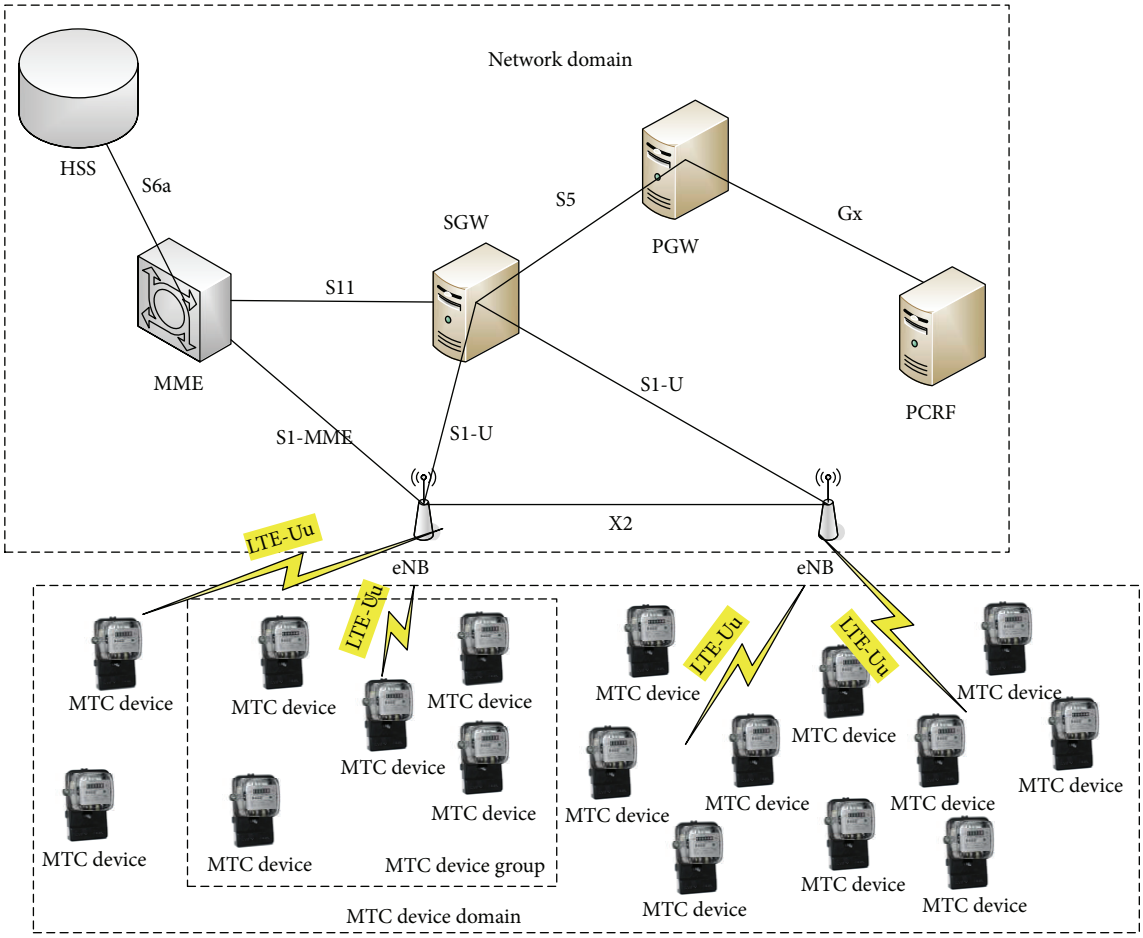


Figure 1: MTC architecture in an LTE system.

A crucial bandwidth group-based policing issue in the design of 3GPP networks is the QoS support required for the tremendous global growth in data and traffic of mobile MTC subscribers [10]. A service flow is based on the report information with a particular set of QoS requirement parameters (e.g., packet delay tolerance, acceptable packet loss rates, required minimum bit rates, and AMBR). The LTE system defines several end-to-end QoS requirements and the QoS information for each evolved packet system (EPS) bearer based on the concept of data flows, including the QoS class identifier (QCI), allocation and retention policy (ARP), guaranteed bit rate (GBR), and maximum bit rate (MBR). The QCI is a scalar that indicates a specific priority, maximum delay, and packet error rate. This index also refers to a set of packet-forwarding treatments (e.g., scheduling weights, admission thresholds, queue management thresholds, and link layer protocol configuration). The ARP is involved in prioritization and preemption decisions regarding bearers, and its primary purpose is to decide whether to accept a bearer establishment request when resources are limited. The GBR denotes the minimum bit rate to be provided to a GBR bearer. For a rate-shaping function, the MBR limits the GBR value to discard excess traffic. The MBR and AMBR regulate the bit rate of traffic generated through one GBR bearer and a group of non-GBR bearers, respectively. The LTE system also supports QoS for EPS bearer aggregates at both the UE-AMBR and access point name-aggregate maximum bit rate (APN-AMBR). The UE-AMBR combines the maximum bit rate across all non-GBR bearers of a UE and enforces bandwidth allocation by the eNB for both uplink and downlink. The APN-AMBR aggregates the maximum bit rate across all non-GBR bearers and across all packet data network (PDN) connections of the same APN and enforces bandwidth allocation by the PDN gateway (PGW) for a downlink [11, 12].

This study considers MTC devices with an Event-Trigger MTC feature for transmission data only at certain predefined times, which include a grant time interval and a forbidden time interval. The network permits an MTC device to transmit data during the grant time interval and prevents it from transmitting data outside the grant time interval; for example, the grant time interval does not overlap with the forbidden time interval. The network does not permit an MTC device to transmit data during the forbidden time interval for various reasons, such as maintenance of the MTC server. MTC device is to communicate with the MTC server. The communication windows of the devices shall be distributed over the predefined time for a group of MTC devices to avoid network overload. This study focuses on a centralized

point-to-multipoint (PMP) architecture that the eNB uses to distribute bandwidth resources to multiple MTC devices. This architecture mode provides superior QoS performance to that of the distributed bearer mode. After receiving an attach request from an MTC device, the centralized MME reports an attach report opportunity in time slots to the HSS from all authorized MTC devices currently using the available bandwidth resource.

The remainder of the paper is organized as follows. Section 2 presents a brief literature review of the attach procedure and MTC application over the LTE system. Section 3 provides a description of the system model and the proposed MSC mechanism. Section 4 introduces the simulation environment. Section 5 presents the simulation results of the proposed scheme. Finally, Section 6 offers the conclusion.

2. Attach Procedures in LTE

The attach procedure involves one or multiple dedicated bearer establishment procedures to establish a dedicated EPS bearer for that UE. Each bearer is associated with a set of QoS parameters that describes the properties of the transport channel. This leads to a flexible bandwidth allocation algorithm that enables differential treatment for traffic with varying QoS requirements. Network operators allow MTC devices to attach to the network during the predefined period.

2.1. Attach Procedure. In the attach procedure, one UE transmits an attach request message to the eNB. The eNB delivers the initial UE message to the MME to enable session management. The MME provides the serving gateway (SGW) with the UE of the out-of-synchronization (OoS) parameter in the create session request message. In the initial context setup request message, the MME determines the bearer ID and OoS parameter based on the create session information. Session creation management is only defined for the uplink data path, and not for the received downlink data path. The eNB generates the initial context setup response message in response to the initial context setup request message from MME to report the eNB address. The 3GPP standard proposes that the modifying bearer management reports the eNB address to the SGW, enabling the PGW to determine the downlink data path. The UE obtains the downlink data according to the modifying bearer management [13].

The Policy Control Enforcement Function (PCEF) sends an Internet Protocol Connectivity Access Network (IP-CAN) session modification message if the IP-CAN QoS exceeds the authorized QoS provided by the PCRF. The dedicated bearer establishment contains the create bearer and bearer setup messages without data payload. The PGW sends a create bearer request message (while maintaining the OoS parameter) to the MME by triggering a dedicated bearer establishment. When the MME is ready to switch the bearer to dedicated bearer activation mode, it sends a bearer setup request message to the eNB. After the eNB receives the bearer setup request message, it replies with a bearer setup response message. To report the bearer setup response message, the eNB sends its address to the MME without any QoS parameters. The reported effective bearer setup feedback is consistent with the RRC connection reconfiguration. The PCEF can then be obtained by the optimal QoS report to meet the specified target error rate before the dedicated bearer establishment.

Figure 2 shows a sequence diagram of the attach procedure. For the enabled service, a UE must register with the network by network attachment. The eNB derives the MME from the initial UE message carrying the attach request and the PDN connectivity. The MME sends a create session request (PGW address, APN, default EPS bearer QoS, APN-AMBR, etc.) message to the selected SGW. The SGW sends a create session request message to the PGW using the PGW address received in the previous message. The PGW considers the received message and returns a create session response (PDN type, PDN address, EPS bearer ID, EPS bearer QoS, APN-AMBR, etc.) message to the SGW. The eNB sends the RRC connection reconfiguration message including the EPS bearer ID and attach accept message to the UE, and the UE returns the RRC connection reconfiguration complete message to the eNB. The MME sends a modify bearer request (EPS bearer ID, eNB address, etc.) message to the SGW, and the SGW acknowledges by sending a modify bearer response (EPS bearer ID) message to the MME.

2.2. Event-Trigger MTC Applications. Although some existing MTC applications use short-range radio technologies, MTC solutions based on mobile access technologies are easier to install. Mobile access-based MTC solutions are also better suited to supporting MTC applications, which require reliable delivery of data to distant MTC devices [14]. Figure 3 shows a MTC application that consists of a modified time and reported device ID at the S6a interface. The MTC device performs initial attach and authentication procedures. When the MME is unaware of the context information of the MTC device, it sends a time period report request, which includes a device ID, to the HSS. After receiving the time period report request from the MME, the HSS determines the period of the MTC device. The HSS returns a time period report, which includes the device ID and time period information of the MTC device, to the MME. To avoid network signaling overhead, the MME calculates the modified grant time interval and determines whether the attach request is from the MTC device. When the attach request is received outside the grant time interval, the MME sends an attach reject message to the MTC device [7]. After an attach request message is accepted, the MME sends a time period report request to the HSS. The time period report request includes a device ID and QoS parameters corresponding to the MTC device making the initial attach request. When the received attach report corresponds to an initial attach report, the HSS determines an APN-AMBR by dividing the group-APN-AMBR assigned to the MTC group by using the group-based policing method. After applying this group-based policing method, the HSS sends an APN-AMBR report, which includes the determined APN-AMBR, to the MME. Finally, the MTC device establishes a PDN connection [9].

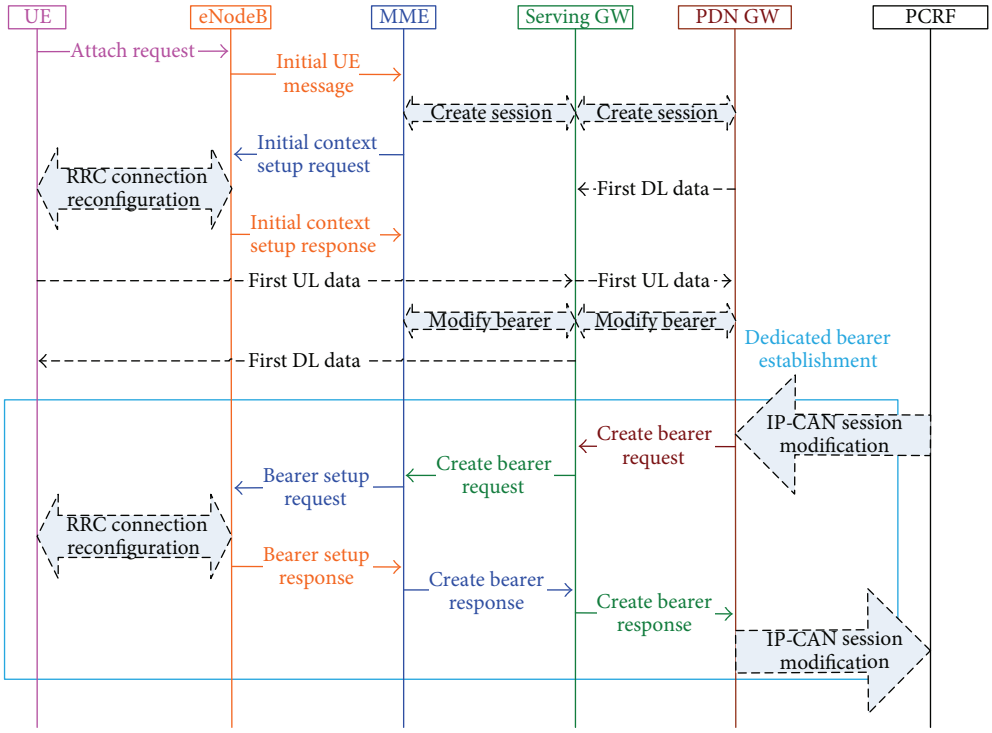


Figure 2: Sequence diagram of LTE attach procedure.

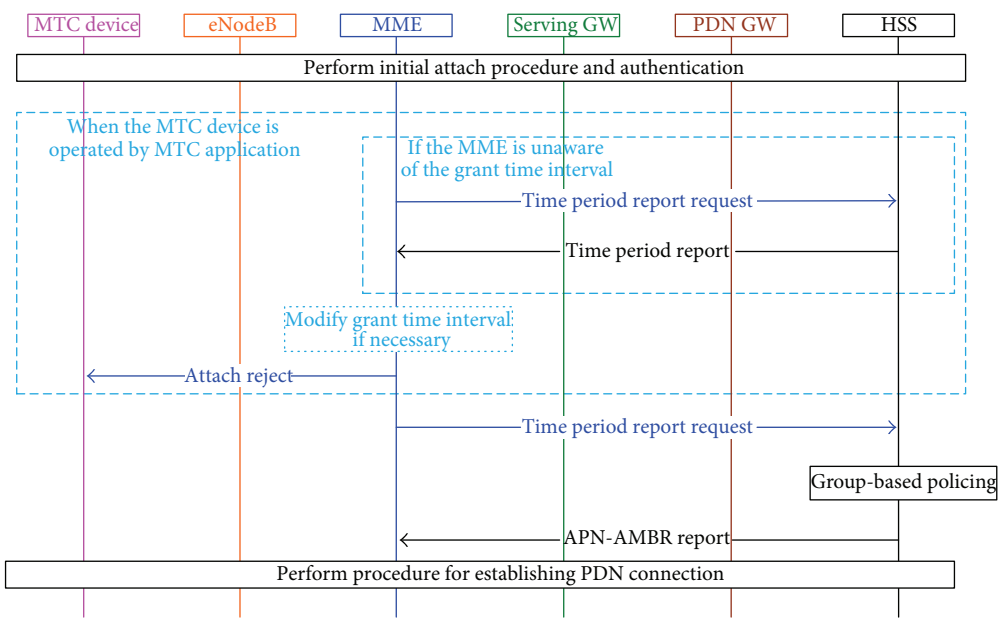


Figure 3: Sequence diagram of MTC application in an LTE system.

3. Proposed Model and Mechanism

This section presents an analysis of the blocking probabilities of a MTC application that distinguishes multiple classes of equal mean grant time interval. Based on the group-based policing, the HSS uses extended APN-AMBR report fields in 3GPP networks. As the APN-AMBR parameter travels through an MME, UE-AMBR losses occurrence for various APN-AMBR types because of inaccessible bandwidth after the group-based policing. Let T be the grant time interval. Let G be the group-APN-AMBR to group MTC devices when they are served, which is the parameter to fulfill

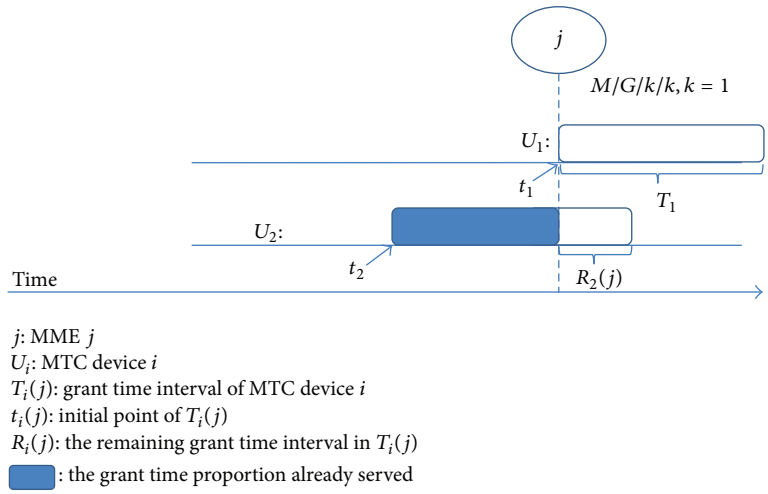


Figure 4: Corresponding queuing model for MTC application.

the satisfaction of the bandwidth requirement of application. The terms T and G are the two operational parameters of the proposed framework.

3.1. Queuing Model for MTC Application. The MTC Application being considered is Event-Trigger MTC. Assume the requests arrive at the MME following a Poisson process. Let N be the number of all MTC devices in the network. If the MME can modify the service time, the system behaves as an $M/G/k/k$ system, and the blocking probability can be obtained using the Erlang B formula [15]. However, when the system includes more than one priority class, the application of $M/G/k/k$ becomes more complex. Assume that the system contains two MTC devices. Figure 4 shows the case when MTC Device 2 (U_2) arrives at the MME before MTC Device 1 (U_1). However, U_2 is stopped because the priority ensures the grant time interval of U_1 by modifying the grant time interval of U_2. The remaining portion of the overlapping grant time interval at MME j of MTC device i is denoted by $R_i(j)$, and this represents the residual service time. If the basic MTC device and all QoS MTC devices are constant, the degree of isolation between two arbitrary classes depends only on their effective overlapping grant time interval. This is because a basic MTC device can be interpreted as a constant shift in time of the reservation process, and thus, neither arrival nor reservation events are reordered in the grant time interval. This result has also been proven by simulation for various arrival and service time distributions. Hence, assume that $R_i(j) = 0$ without loss of generality and consider the overlapping grant time interval between MTC Device 1 and MTC Device 2 as $R_2(j) = T_2 - (t_2 - t_1) > 0$. In this case, $T_1(j)$ has preemptive priority over $R_2(j)$. The blocking probability of U_1 is simply obtained using the Erlang B formula. This study uses the Erlang B formula to evaluate the blocking probability of the all the $T_i(j)$ traffic of MTC devices.

Consider a time period report request arriving at a certain time instant in the HSS and requesting to reserve a grant time interval q time slots after its arrival time. Without loss of generality, consider the arrival time of the time period report request to be slot 0, and the start of the MTC device requested duration to be slot q. To calculate the blocking probability of this MTC device request, consider the traffic load at slot q, as seen at the time of the request. Any MTC device requested duration generated in the future (from time slots after slot 0) for slot q will not affect the probability of accepting/blocking the MTC device request. The number of grant time intervals with a start time within slot q, as seen at the instant in which the time period report request arrives, is Poisson-distributed with mean

$$\lambda^{(q)} = \begin{cases} \lambda, & \text{for } q \leq 0, \\ \lambda \left\{ 1 - \sum_{a=0}^{q-1} f(a) \right\}, & \text{for } q > 0. \end{cases} \quad (1)$$

The number of MTC device requests for time slot q from all previous slots, including slot q, is λ. Given a time period report request arriving in time slot 0, the probability that this MTC device requested duration requests that time slot q is $f(q)$. Therefore, the number of MTC device requests from slot 0 for slot n is $\lambda f(q)$. From a time period report request viewpoint, the number of grant time interval arrivals in slot n is the sum of MTC device requests from all time slots before slot 0, in addition to the MTC device requests from slot 0. Thus, the number of arrivals in slot q, as seen by the time period report request, is the sum of all possible MTC device requests (λ) minus any future MTC device requests to be made in time slots 1 to q for slot q.

Based on this assumption, define $\mu_i = 1/E[T_i]$, where $E[T_i]$ is the expected value of the MTC device $i(U_i)$ grant time interval at the same MME. To determine the effective service time of a grant time interval under the MTC application, refer to Figure 4. For a predefined time, a bandwidth is reserved for a length of time that is equal to the sum of two time intervals. The duration of the first interval is equal to that of the grant time interval and is distributed according to $T_i(j)$ with a mean $1/\mu$. The conventional Markov model was implemented to analyze the MTC requests queue model. The Markov model

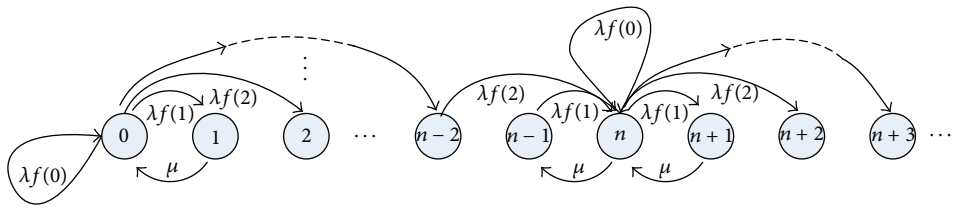


Figure 5: Markov model for the MTC requests queue.

was constructed for the MTC requests queue. The state n represents the number of MTC requests within the frame duration. The process of the MTC requests queue assumes that time slot q is working. The Markov process is illustrated in Figure 5, based on the assumptions in our study. The state n represents the number of requests in the MTC system.

The duration of the second interval is equal to that of the forbidden time interval and is distributed according to $R_i(j)$ with a mean $\overline{R}$. Based on these observations, an output port of an MME node using a MTC application behaves as an $M/G/k/k$ loss system. The traffic intensity $\rho^{(\mathrm{TC})}$ of the queue is

$$\rho^{(\mathrm{TC})} = \lambda^{(q)}\left(\frac{1}{\mu} + \overline{R}\right). \tag{2}$$

This study first presents modeling the MME using an $M/M/k/k$ queue with preemptive priorities, where the arrival rate of t_i is λ_i and the service rate is μ_i. Let k be the number of classes in the MTC application. Denote $\rho_i = (\lambda_i/\mu_i)$ as the traffic load of t_i. If t_1 has an absolute priority (i.e., all other MTC devices have signaling overhead), the blocking probability of one class can be obtained using the Erlang B formula in the $M/M/k/k$ queue, expressed as

$$P(\rho_1, k) = \frac{\left(\rho_1^k/k!\right)}{\left(\sum_{m=0}^{k} \rho_1^m/m!\right)}. \tag{3}$$

Similar to (3), the blocking probability of the superposition of the two classes with total traffic load $\rho_1 + \rho_2$ can be calculated as follows:

$$P(\rho_{1,2}, k) = \frac{\left(\rho_{1,2}^k/k!\right)}{\left(\sum_{m=0}^{k} \rho_{1,2}^m/m!\right)}, \tag{4}$$

where $\rho_{1,2} = \rho_1 + \rho_2$. In the multiclass case for the $M/G/k/k$ queue, the blocking probabilities for service classes with different QoS values can be obtained by heuristically generalizing (3) and (4) to an arbitrary number of k classes. A conservation law can be formulated for every set of classes $E_x = \{0, \ldots, x\}$ with $0 < x \le k - 1$:

$$\left(\sum_{y=0}^{x} \lambda_i\right) P\left(\rho_{1,2,\ldots,E_x}, k\right) = \sum_{y=0}^{x} \lambda_y \cdot P\left(\rho_{1,2,\ldots,y}, k\right), \tag{5}$$

where $P(\rho_{1,2,\ldots,E_x}, k)$ is the total blocking probability of all classes in E_x. It describes the probability that a low-priority MTC device that started the grant time interval prior to the grant time interval of other MTC devices has not finished its grant time interval. In general, the blocking probability of a MTC application can be calculated based on (3)–(5) as follows:

$$\begin{aligned} P\left(\rho_{1,2,\ldots,E_x}, k\right) &= \left(\frac{1}{\lambda_{E_x}}\right)\left(\sum_{y=0}^{E_x} \lambda_y\right) \\ &\quad \times \left[P(1, 2, \ldots, E_x) - p_{1,2,\ldots,E_x-1} P(1, 2, \ldots, E_x - 1)\right], \end{aligned} \tag{6}$$

where $p_{1,2,\ldots,E_x-1} = (\sum_{y=1}^{E_x-1} \lambda_y)/(\sum_{y=1}^{E_x} \lambda_y)$ and $P(1, 2, \ldots, E_x) = ((\sum_{y=1}^{E_x} \rho_y)^k/k!)/(\sum_{m=0}^{k} (\sum_{y=1}^{E_x} \rho_y)^m/m!)$.

3.2. Multistage Control Mechanism. The attach report message is selected from each MTC device after reaching the HSS because the MTC application identifies acceptable attach MTC devices. The goal of the MSC mechanism is to improve throughput requirement with a QoS guarantee. Figure 6 shows a flowchart of the APN-AMBR allocation process. The MME receives an attach request and determines whether a MTC application operates the attached MTC device. If a MTC application operates the MTC device, the MME also determines whether to modify the grant time interval. The MME sends an attach report to the HSS, and the HSS collects all MTC devices requesting attachment. The HSS determines whether a new attach report has been received from the MME. When the HSS knows the already attached MTC device, it can process the MSC mechanism without waiting for a duration. The HSS sends the APN-AMBR report information after the MSC mechanism. The current AMBR definitions enable system operators to differentiate the service level provided for each of these services. In the proposed architecture, rate policing prevents the network from becoming overloaded and ensures that the services send data in accordance with the specified maximum bit rates. Because MTC devices are encouraged to adapt their APN-AMBR to starvation, resource starvation is confined to those who ignore starvation. The uplink and downlink scheduling functions implemented by the LTE system are largely responsible for fulfilling the QoS characteristics associated with the different bearers.

These APN-AMBRs are APN-level quantities and are therefore known at the HSS. These APN-AMBRs propagate through UE attach procedures down to the MME to enforce

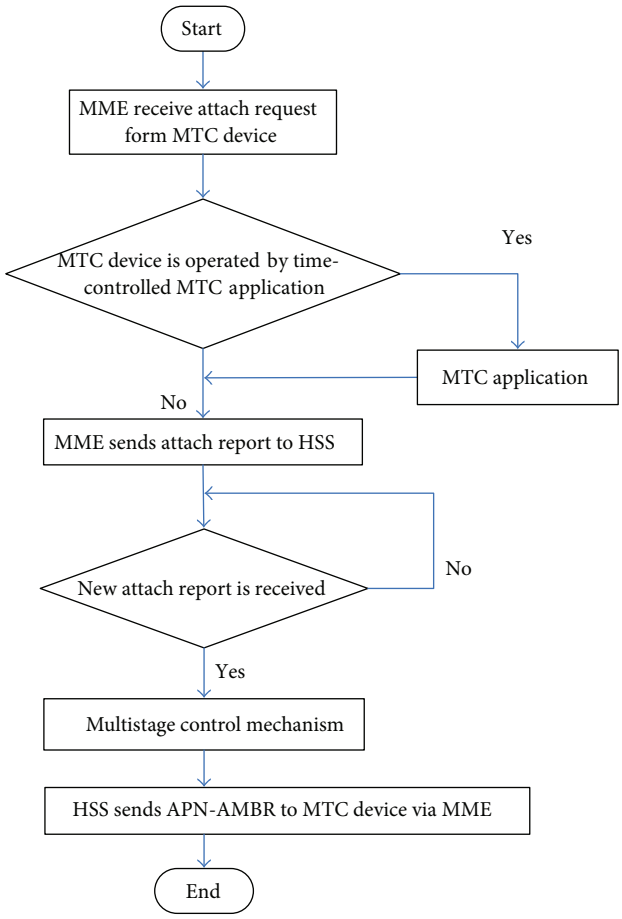


Figure 6: Flowchart of the APN-AMBR allocation process.

the data rate from the specific APN through rate policing. Each MTC device may have various packets, and the stream types of these packets may have various UE-AMBRs. This mechanism also uses the following parameters:

(i) N: the number of MTC devices that must be buffered within an MME in the LTE system;

(ii) Q: the number of MMEs in a MME pool area;

(iii) S: the number of stages in the MSC mechanism;

(iv) $G_{i,j}$: the group APN-AMBR size of the ith MTC device at the jth MME;

(v) $R_{i,j}$: the UE-AMBR size of the ith MTC device at the jth MME;

(vi) D_i: the default APN-AMBR size of the ith MTC device;

(vii) M_i: the minimum APN-AMBR size of the ith MTC device;

(viii) A_i: the allocation APN-AMBR size of the ith MTC device.

Consider the MSC mechanism in which resource use can be calculated. Let $C_{i,j}$ be the resource use of the ith MTC device at the jth MME, where $C_{i,j} = (R_{i,j}/G_{i,j})$. The challenge of this utility maximization problem is to determine the APN-AMBR of each subscribed MTC device to be served and the amount of resources to be allocated to each APN-AMBR of all subscribed MTC devices. When resources are limited, the HSS must determine the number of APN-AMBRs for each MTC device to receive resources, in order to maximize the total utility of the system. We could find an allocation $\Gamma = \{A_1, A_2, \ldots, A_N\}$, where $A_i = \{a_{i,j} : a_{i,j} \geq 0, j = 1, 2, \ldots, Q\}$ denotes the set of allocated resource to each APN-AMBR of the ith MTC device. Formally, the objective of this problem is to find a Γ to:

$$\begin{aligned} \text{Maximize :} \quad & \sum_{j=1}^{Q}\sum_{i=1}^{N} R_{i,j}C_{i,j} \\ \text{Subject to :} \quad & \sum_{j=1}^{Q}\sum_{i=1}^{N} a_{i,j} \leq A, \\ & 0 \leq C_{i,j} \leq 1, \\ & R_{i,j} \geq 0, \quad \forall i, j. \end{aligned} \tag{7}$$

By solving this equality, the MTC application can easily be obtained for the user. Resource use can be maximized with the optimal values, which can be obtained by $R_{i,j}$ and $C_{i,j}$ parameters. However, the solution of this optimization equation is not explicit because the MME does not know the group APN-AMBR size and the number of MMEs in the MME pool area in this stage. How to evaluate the APN-AMBR for each MTC device is important. One way to address this is to refer to resource use, by employing the utilization range to evaluate the APN-AMBR for each MTC device at the HSS. The utilization ranges for various stages exhibit up-bound and low-bound utilization. Let $H_{s'}$ and $L_{s'}$ be the up-bound and low-bound utilization of the s'th stage, respectively. Define up-bound utilization as

$$H_{s'} = \begin{cases} \sum_{i=1}^{s'} \left(\frac{1}{e}\right)^{i}, & \text{if } s' < S, \\ 1, & \text{if } s' = S. \end{cases} \tag{8}$$

The up-bound utilization and low-bound utilization can be determined by the HSS. Hence, low-bound utilization can be represented as

$$L_{s'} = \begin{cases} 0, & \text{if } s' = 1, \\ \sum_{i=2}^{s'} \left(\frac{1}{e}\right)^{i-1}, & \text{if } 1 < s' = S. \end{cases} \tag{9}$$

This method focuses on improving resource utilization. The MSC mechanism helps the MME allocate the APN-AMBR for each MTC device to maximize the resource use and satisfy the target UE-AMBR. The granted bit rate of the APN-AMBR is a critical parameter in the MSC mechanism. First, the MME assigns the initial resource use levels, $H_{s'}$ and $L_{s'}$, based on the required resource use of up bound and low bound, respectively. Therefore, it is necessary to check all utilization ranges $(H_{s'}, L_{s'})$ to find the utilization ranges

```
Input: N, S, G_{i,j}, R_{i,j}, D_i, M_i,
Output: A_i
for j = 1 to Q do
        for i = 1 to N do                                   //allocate resource
        {
                C_{i,j} = (R_{i,j}/G_{i,j});
                Find (C_{i,j} < H_{s'} and C_{i,j} ≥ L_{s'});      // use (8) and (9) with given s'
                Switch (s')
                {
                        case (s' == S): A_i = M_i; break;
                        case(s' ∈ Φ): MME sends attach reject message; break;
                        default: A_i = D_j/e^{(s'-1)}; break;
                }
        }
        end for;
end for;
```

Algorithm 1: Multistage control mechanism.

that maximize resource use as candidate solutions. Among these candidates, the candidate with the maximum resource utilization is the solution for the MSC decision. If multiple combinations produce the same maximum resource use, select the one with the maximum number of MTC device-rejected attachments. Algorithm 1 describes the assignment of the APN-AMBR by using the MSC mechanism in an MME pool area.

To improve fairness, the MTC devices should be allocated their APN-AMBR from the group APN-AMBR in order. If the UE-AMBR of the signaling overhead type is incomplete, signaling nonoverhead UE-AMBR will occupy the resource of the group APN-AMBR in the MTC applications. The MTC device allocation process ensures that FCFS scheduling is based on fairness if the stream types have the same priorities. The frame allocation process guarantees higher throughput and achieves fairness between several MTC device stream types. For this reason, MTC devices with a better resource condition enjoy better perceived quality between several stream types. Jain's fairness index is a conventional method of assessing the quality of the traffic type [16]. The term f represents the total number of QCIs for the duration of the MTC application. The value of Jain's fairness index is generally between $1/f$ and 1. When the LTE system indicates an increase in Jain's fairness index value, the system has higher fairness for all CQIs. This model can be written as follows:

$$\mathrm{FI} = \frac{1}{f} \times \frac{\left(\sum_{n=1}^{f} X_n\right)^2}{\sum_{n=1}^{f} X_n^2}, \tag{10}$$

where X_n is the APN-AMBR for the n MTC devices. In this description, the bit rate of all traffic generated by a group of MTC devices can be controlled by determining the APN-AMBR of each MTC device according to current resource use of the MTC group. Group-based policing regulates the maximum bit rate of all traffic generated, and the total bit rate of traffic generated through all non-GBR bearers connects to the same APN. The MME sends the APN-AMBR to the MTC device requesting attachment by using the PDN connection establishment procedure. The MME may store $G_{i,j}$, D_i, M_i, $H_{s'}$, and $L_{s'}$ in individual groups of MTC devices by using a multistage controlled bit rate feature.

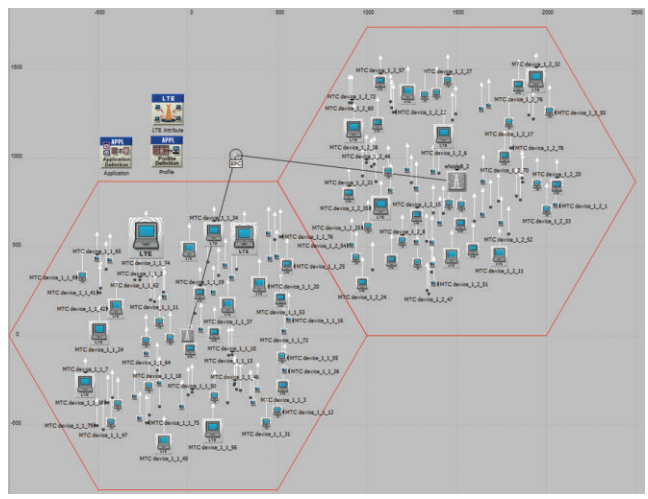


Figure 7: Network topology of the simulation.

4. Simulation Environment

The experiments in this study were performed using OPNET Modeler 17.1 with LTE module capability to simulate an MTC environment [20]. The simulation was conducted for 3,600 seconds to investigate the stable state result for all MTC device nodes.

This simulation tests the performance of the proposed mechanism in a typical network consisting of one EPC and 160 randomly distributed MTC device nodes. As Figure 7 shows, this experimental environment is based on a simulated OPNET MTC network topology containing two cells. Each cell has an eNB, and each eNB has 80 MTC

Table 1: LTE simulation system parameters [17, 18].

Parameters	Value
PHY profile	FDD
Bandwidth	20 MHz
Packet size	1024 byte
Cycle prefix	Normal
UL SC-FDMA channel (base frequency)	1920 MHz
DL OFDMA channel (base frequency)	2110 MHz
Max. retransmission (HARQ)	3
Handover type	Intrafrequency
Path-loss parameter	Free space

devices. The transmission powers of the MTC device and eNB Node were set to 0.006 watts and 0.012 watts, respectively. The OPNET node models of the eNB and MTC devices are *lte_enodeb_atm4_ethernet4_slip4_router* and *lte_wkstn_adv*, respectively. These simulations assume that the moving mode of the MTC devices follows the random waypoint model. The movement speed of each MTC device was uniformly distributed between 1 and 5 m/s. The fixed eNB node model featured router functionality. The UE node model featured workstation functionality [21]. The global configuration object was used to configure the parameters, such as EPS bearer definitions and PHY profiles, in the LTE attributes node [17, 18].

The UEs have a CQI index to provide QoS awareness. The services are grouped into different QoS classes. The main contribution of the proposed mechanism is the MAC layer, whereas the actual physical transmission is adopted from the OPNET LTE model. Table 1 presents a summary of the simulation parameters. This system considers both the downlink and uplink. The MAC layer scheduler implies that the GBR bearers are always allocated radio resources before the non-GBR bearers. The eNB module implements priority scheduling for GBR and non-GBR bearers. Table 2 presents the experimental data for QoS class services. The MTC traffic on the EPS bearer is generated between the eNB and MTC devices [19]. The EPS bearer configuration attribute defines four bearers: platinum, gold, silver, and bronze.

5. Results and Discussion

Figure 8 shows a comparison of the blocking probability and traffic load of the MTC application, validating the analytical results by simulation. The general distribution is set according to the assumption that 20% of grant time intervals are long (with a service rate of 1.2), 30% of grant time intervals are short (with a service rate of 0.9), and the remaining 40% grant time intervals have a service rate of 1. Simulation results indicate that saturation occurs when the number of MTC devices reaches 80 and $\rho > 0.002$ (using the MTC application in the LTE system). Saturation means that all grant time intervals of MTC devices have been blocked. However, the lower-priority grant time intervals cannot be scheduled when the number of MTC devices increases, and the blocking probability is more than 1. To prevent the transmission of

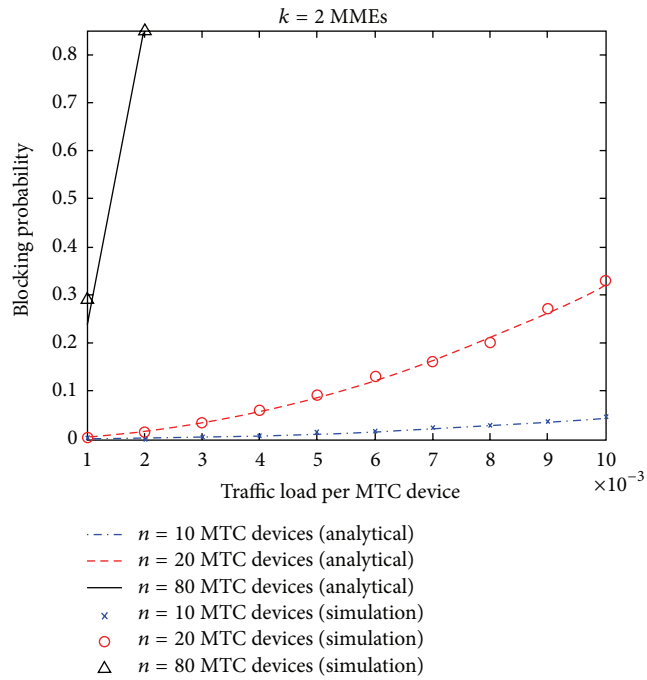


Figure 8: Comparisons of blocking probability, and traffic load.

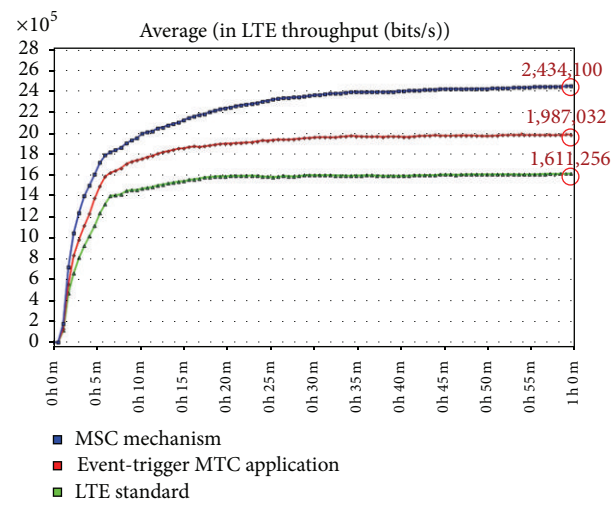


Figure 9: System throughput in the simulation (in bits/second).

blocked MTC devices from wasting AMBR, it is necessary to periodically verify the inaccessible bandwidth from the MME.

Figure 9 shows the system throughput for the LTE standard, Event-Trigger MTC application, and our proposed MSC mechanism. The simulation results in Figure 9 indicate that saturation occurs when the simulation time reaches 1,800 seconds. Saturation indicates that all the MTC devices of one eNB have been scheduled. The Event-Trigger MTC application reserves the extra bandwidth and reallocates the remaining bandwidth. Using the Event-Trigger scheme, the maximal average throughput reaches 2 Mbps before the system becomes saturated. The MSC mechanism with a MTC

Table 2: Standardized QCI characteristics [19].

Service class	Priority	Type	Packet delay budget	Packet error loss rate	Guaranteed bit rate
Platinum	1	Non-GBR	100 ms	10^{-6}	—
	2	GBR	100 ms	10^{-2}	128 kbps
Gold	3	GBR	50 ms	10^{-3}	858 kbps
	4	GBR	150 ms	10^{-3}	384 kbps
Silver	5	GBR	300 ms	10^{-6}	384 kbps
	6	Non-GBR	300 ms	10^{-6}	—
Bronze	7	Non-GBR	100 ms	10^{-3}	—
	8	Non-GBR	300 ms	10^{-6}	—

application and a multistage controlled process achieves higher performance because it has a higher average throughput. This is the main reason for obtaining the time period report information and an appropriate bandwidth-controlled feature. After the bandwidth allocation reaches an improved state because of the resource use, the MTC device can transmit at a higher data rate using previously inaccessible bandwidth. Because the MSC mechanism considers the method of allocation of the APN-AMBR, the MTC device can select the appropriate APN-AMBR with a superior MTC application to the data transmission, thereby achieving a maximal average throughput of 2.43 Mbps. This phenomenon demonstrates that the proposed MSC mechanism markedly improves throughput compared to the LTE standard because of the extra bandwidth consumed and the inferior group-based control method of the LTE standard. The final values of throughput after 3,600 seconds in the simulation are 1,611,256 bps for LTE standard system, 1,987,032 bps for Event-Trigger MTC application, and 2,434,100 bps for MSC mechanism. Compared to the LTE system with Event-Trigger MTC application, the MSC mechanism achieves a 22.5% higher throughput in this simulation.

Figure 10 shows the relationships among the packet delay times of the MTC application. For the Event-Trigger MTC application, the MME requests an additional period without modifying resource use at the HSS. Additional periods cause the time period report time to increase as the delay time increases. The final values of packet delay time after 3,600 seconds in the simulation are 0.123 seconds for LTE standard system, 0.132 seconds for Event-Trigger MTC application, and 0.100 seconds for MSC mechanism. Compared to the LTE standard and Event-Trigger MTC application, the packet delay time of the MSC mechanism achieves higher performance in the LTE system because each MME continuously monitors its APN-AMBR. The HSS gathers the APN-AMBR information received through the create session request messages sent by the MMEs. When the group-based policing is optimal or the resource use is low, the MTC application should also be served using the MSC mechanism.

Figure 11 shows the simulation results of Jain's fairness index value at various MTC traffic applications for 80 nodes, indicating that all of the MTC traffic applications in this paper have relatively high Jain's fairness index values (0.62 < FI). This chart shows that the MSC mechanism has a higher fairness than the LTE standard and Event-Trigger

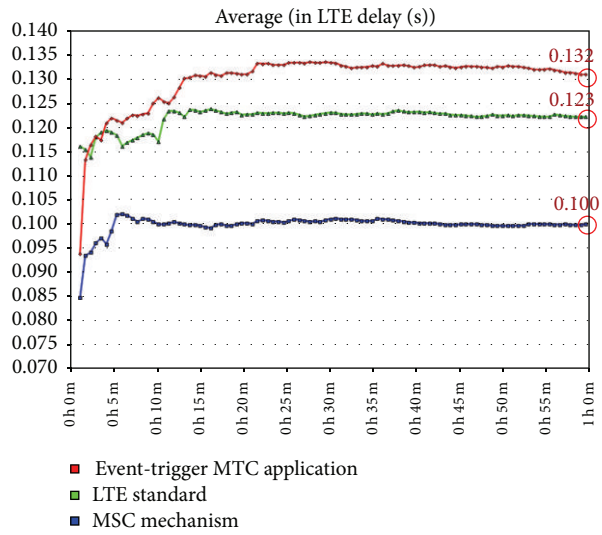


Figure 10: Packet delay time of the MTC application in the simulation (in seconds).

MTC application. For the MSC mechanism, the fairness of MTC traffic applications depends on the amount of traffic transmitted by the selected APN-AMBR. This is the main reason the MSC mechanism has more resource utilization. For the transmission model of an Event-Trigger MTC application, a number of MTC traffic applications are reallocated a portion of bandwidth by the attach reject message. In the MTC traffic type, the fairness index of Event-Trigger MTC application is higher than that of the LTE standard because the Event-Trigger MTC application considers the method of allocating AMBR based on the modified period and the amount of controlled time.

Markov chain with $M/G/k/k$ and Jain's fairness index are used to analyze machine-type communication in a 3GPP network. Table 3 compares the results between analysis and simulation of Jain's fairness index (FI) values, which shows that analysis can be validated by simulation.

6. Conclusion

This study proposes an MSC mechanism in an LTE system to process the operations of bandwidth allocation based

Table 3: Comparison of results between analysis and simulation of Jain's fairness index (FI) value.

Service class	Gold		Silver		Bronze		Platinum	
Simulation or analysis	Sim.	Ana.	Sim.	Ana.	Sim.	Ana.	Sim.	Ana.
MSC mechanism	0.89	0.89	0.80	0.79	0.72	0.70	0.71	0.70
Event-Trigger MTC application	0.89	0.89	0.79	0.79	0.70	0.70	0.70	0.70
LTE standard	0.89	0.89	0.76	0.75	0.67	0.66	0.63	0.63

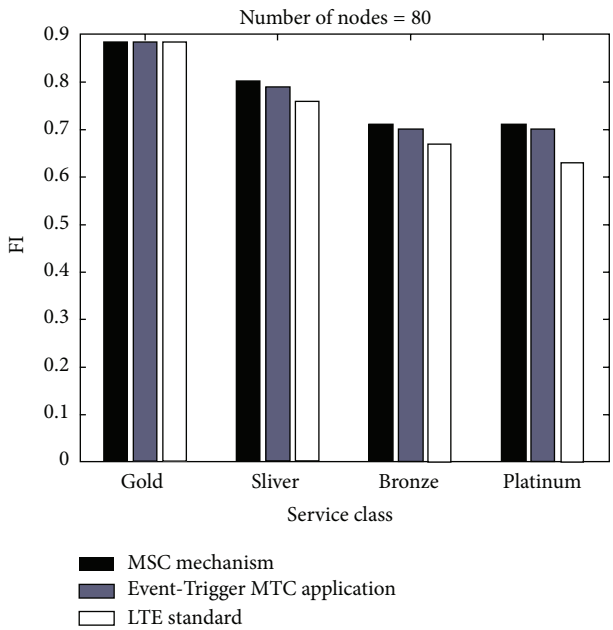


Figure 11: The comparison of Jain's fairness index (FI) values.

on MTC application, QoS parameters, and the bandwidth requirements of MTC devices and the eNB. This study presents a comparison of the performance of the proposed MSC mechanism and the LTE standard mechanism of MTC application. The service flow simulations of the OPNET modeler indicate the MSC mechanism achieves a higher system throughput, a lower delay time, and greater long-term fairness for multiple MTC devices. Experimental results indicate that the throughput of the MSC mechanism is 22.5% higher than that of the LTE standard model using the group-based policing, which implies that the proposed MSC mechanism is an effective bandwidth allocation method in an LTE system with MTC devices.

Conflict of Interests

No conflict of interests exists between the authors and all of the mentioned organizations or corporations, including 3GPP and OPNET Modeler. The 3rd Generation Partnership Project (3GPP) is a nonprofit international telecommunication standard development organization, which defines new-generation telecommunication standards including LTE (long-term evolution). In this study, we perform the simulation by OPNET Modeler ver. 17.1, which is a popular simulation tool for analyzing and designing communication networks, devices, protocols, and applications. They have paid the licensing fee for legal use of OPNET Modeler ver. 17.1 in our research.

Acknowledgment

A special thanks goes to Institute for Information Industry in Taiwan, for its great help in supporting the simulation environment and key LTE simulation modules to this study.

References

[1] P. Jain, P. Hedman, and H. Zisimopoulos, "Machine type communications in 3GPP systems," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 28–35, 2012.

[2] T. Taleb and A. Kunz, "Machine type communications in 3GPP networks: potential, challenges, and solutions," *IEEE Communications Magazine*, vol. 50, no. 3, pp. 178–184, 2012.

[3] O. Del Rio Herrero and R. De Gaudenzi, "High efficiency satellite multiple access scheme for machine-to-machine communications," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 4, pp. 2961–2989, 2012.

[4] K. Chang, A. Soong, M. Tseng, and Z. Xiang, "Global wireless machine-to-machine standardization," *IEEE Internet Computing*, vol. 15, no. 2, pp. 64–69, 2011.

[5] 3GPP, "Service Requirements for Machine-Type Communications, TS 22.368 V12.0.0," 2012.

[6] A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, "Cellular-based machine-to-machine: overload control," *IEEE Network*, vol. 26, no. 6, pp. 54–60, 2012.

[7] 3GPP, "System Improvements for Machine-Type Communications, TR 23.888 V11.0.0," 2012.

[8] K. Zheng, F. Hu, W. Wang, W. Xiang, and M. Dohler, "Radio resource allocation in LTE-advanced cellular networks with M2M communications," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 184–192, 2012.

[9] C. Songyean, R. Heetae, L. Jangwon, B. Beomsik, L. Chaegwon, and J. Sangsoo, "Group-based control method and apparatus for MTC devices in mobile communication system," United States patent US, 2012/0209978, 2012.

[10] S. Lien and K. Chen, "Massive access management for QoS guarantees in 3GPP machine-to-machine communications," *IEEE Communications Letters*, vol. 15, no. 3, pp. 311–313, 2011.

[11] Y. Chen and W. Wang, "Machine-to-machine communication in LTE-A," in *Proceedings of IEEE 72nd Vehicular Technology Conference Fall (VTC '10-Fall)*, pp. 1–4, Ottawa, Canada, September 2010.

[12] S. Y. Lien, K. C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 66–74, 2011.

[13] 3GPP, "GPRS enhancements for E-UTRAN access, TS 23.401 V11.4.0," 2012.

[14] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. D. Johnson, "M2M: from mobile to embedded internet," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 36–43, 2011.

[15] D. Gross and C. M. Harris, *Fundamentals of Queuing Theory*, Wiley, New York, NY, USA, 3rd edition, 1998.

[16] R. Jain, D. M. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared system," DEC Research Report TR-301, 1984.

[17] M. Torad, A. El Qassas, and H. Al Henawi, "Comparison between LTE and WiMAX based on system level simulation using OPNET modeler (release 16)," in *Proceedings of 28th National Radio Science Conference (NRSC '11)*, pp. 1–9, Cairo, Egypt, April 2011.

[18] K. Andersson, S. A. M. Mostafa, and R. Ui-Islam, "Mobile VoIP user experience in LTE," in *Proceedings of IEEE 36th Conference on Local Computer Networks (LCN '11)*, pp. 785–788, Bonn, Germany, October 2011.

[19] 3GPP, "Policy and charging control architecture, TS 23.203 V11.8.0," 2012.

[20] "OPNET LTE Specialized Model," http://www.opnet.com/LTE/.

[21] B. H. Lee and S. L. Kim, "Mobility control for machine-to-machine LTE systems," in *Proceedings of Wireless Conference 2011-Sustainable Wireless Technologies (European Wireless)*, pp. 1–5, Vienna, Austria, April 2011.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 214814, 14 pages
http://dx.doi.org/10.1155/2013/214814

Research Article

A Mutual-Evaluation Genetic Algorithm for Numerical and Routing Optimization

Chih-Hao Lin and Jiun-De He

Department of Information Management, Chung Yuan Christian University, Jhongli 320, Taiwan

Correspondence should be addressed to Chih-Hao Lin; linch@cycu.edu.tw

Received 9 May 2013; Accepted 22 July 2013

Academic Editor: Anyi Chen

Many real-world problems can be formulated as numerical optimization with certain objective functions. However, these objective functions often contain numerous local optima, which could trap an algorithm from moving toward the desired global solution. To improve the search efficiency of traditional genetic algorithms, this paper presents a mutual-evaluation genetic algorithm (MEGA). A novel mutual-evaluation approach is employed so that the merit of selected genes in a chromosome can be determined by comparing the fitness changes before and after interchanging with those in the mating chromosome. According to the determined genome merit, a therapy crossover can generate effective schemata to explore the solution space efficiently. The computational experiments for twelve numerical problems show that the MEGA can find near optimal solutions in all test benchmarks and achieve solutions with higher accuracy than those obtained by eight existing algorithms. This study also uses the MEGA to find optimal flow-allocation strategies for multipath-routing problems. Experiments on quality-of-service routing scenarios show that the MEGA can deal with these constrained routing problems effectively and efficiently. Therefore, the MEGA not only can reduce the effort of function analysis but also can deal with a wide spectrum of real-world problems.

1. Introduction

Many engineering optimization issues can be formulated as global optimization problems with numerical functions. When solving a complex problem, the particular challenge is that algorithms may be trapped in local optima and fail in finding global optima. Recently, genetic algorithms (GAs) have received considerable attention for solving complex and unstructured problems [1, 2]. However, traditional genetic algorithm (TGA) often suffers from the drawbacks of premature convergence and weak exploitation capabilities [3].

To overcome the deficiencies of the TGA, this paper proposes a mutual-evaluation approach to incorporate with the TGA as a mutual-evaluation genetic algorithm (MEGA). The proposed therapy crossover can implicitly generate effective schemata to efficiently exploit the given search space without explicitly analyzing the solution space. The performance of the proposed MEGA is experimented on 12 well-known numerical functions and compared with four well-known evolutionary algorithms (EAs) and four existing modified GAs. The experimental results show that the proposed MEGA is able to increase the accuracy by several orders of magnitude in almost all the cases. That is, MEGA can effectively approach the global optimum without being trapped in many local optima.

Because of the simplification property of the mutual-evaluation approach, the MEGA is suitable to deal with a wide spectrum of real world problems. In this paper, the proposed MEGA is also realized to deal with multipath routing problems in a multicommodity network, where more than one routes will be connected for each origin-destination (OD) pair. The goal is to find an optimal flow-allocation strategy to minimize the total transmission cost and satisfy all quality-of-service (QoS) requirements at the same time. The performance of the proposed MEGA is experimented by optimizing several multipath routing problems. The experiment results show that the MEGA outperforms the TGA dramatically. Furthermore, the search ability of the MEGA is robust in obtaining consistent results.

The rest of the paper is organized as follows. Section 2 describes the proposed mutual-evaluation approach. The main operations of the MEGA are introduced in Section 3.

In Section 4, the experimental studies on 12 numerical optimization problems are described. The results are compared with eight existing optimization algorithms. In Section 5, the MEGA is realized to deal with a real-world problem: multipath routing problem. A novel representation of chromosomes for routing problems is also proposed. Finally, conclusions and contributions are offered in Section 6.

2. Mutual Evaluation Approach

To overcome the deficiencies of the TGA, modified GAs should keep evolutional population as diverse as possible to improve algorithms' exploration capability for discovering new solution area. In regard to evaluation approaches, existing GAs can be classified into three categories: (1) chromosome-oriented, (2) gene-oriented, and (3) schema-oriented.

2.1. Chromosome-Oriented Approaches. Chromosome-oriented GAs concentrate on the chromosome fitness in the evolutionary process. The TGA adopts chromosome fitness to evaluate the quality of whole individual chromosome [4, 5]. A number of researches recently focus on incorporating mathematical methods with the chromosome-oriented GAs to alleviate the deficiency of premature convergence [4]. For example, Yao et al. proposed fast evolutionary programming (FEP) with Cauchy mutation to solve the premature convergence deficiency [6]. To improve slow finishing deficiency, Tu and Lu developed a stochastic genetic algorithm (StGA) with mathematical methods, which not only can find global optima, but also can reduce the computational effort [7]. By merging niche techniques and Nelder-Mead's simplex method, Wei and Zhao proposed the niche hybrid genetic algorithm (NHGA) to alleviate premature convergence and weak exploitation deficiencies of TGA [3].

2.2. Gene-Oriented Approaches. In gene-oriented GAs, the compact genetic algorithm (cGA) is a common representative [8]. The cGA represents the population as a probability vector over the set of individuals to mimic the order-one behavior of TGA. The cGA manipulates the gene distribution and essentially evolves each gene individually [8]. To enhance cGA's performance, Ahn and Ramakrishna proposed a strong elitism version of cGA [9] and then Rimeharoen et al. introduced a moving average technique to update the probability vector [10]. Although the cGA reduces memory requirement and offers many advantages, its limitation is the assumption of the independency between individual genes.

2.3. Schema-Oriented Approaches. The schema-oriented GAs explore the exact schemata by borrowing from the schema theorem proposed by Holland [11]. A schema is a pattern within a chromosome defined by fixing the values of specific chromosome loci. The increase of effective schemata enables the efficient search within a solution space and guides the evolution of the population in approaching the global optimal solution [12]. Yen and Shyu proposed a statistical gene evaluation method that uses simple statistical quantities to investigate the individual gene influence which suggests better choices for a gene evolution [13]. Kubota et al. proposed the virus-evolutionary genetic algorithm (VEGA) [14] that simulates coevolution of a virus population and a host population. VEGA applies horizontal propagation and vertical inheritance in a population with virus infection operators and genetic operators [14]. In this paper, the proposed MEGA is a new schema-oriented GA in which an innovative mutual-evaluation approach is used to achieve highly efficient evolution with necessary robustness. The MEGA does not only evaluate a chromosome by its fitness but also analyze genome's merit to improve the population's quality.

2.4. The Proposed Mutual-Evaluation Approach. According to the biological concept of the genetic engineering, a gene splicing is a process that manipulates genes outside the traditional random reproductive process. The proposed MEGA integrates a mutual-evaluation approach in a novel therapy crossover operation to produce offspring by introducing isolation, manipulation, and reintroduction of gene splicing techniques to improve the chromosome's fitness. The algorithm randomly selects two parents from a mating pool of generation t. Let the parent with superior fitness be named as the good parent $\vec{x}_{\text{good}}(t)$ and the other one is the bad parent $\vec{x}_{\text{bad}}(t)$. The MEGA generates a therapy mask to indicate which gene loci in a chromosome are chosen for crossover points. For each bit in a chromosome, we uniformly generate a random number in interval [0, 1] and compare the number with a predefined therapy rate p. If the random number is less than p, its mask bit of the corresponding locus is set to value 1, which means that this gene locus belongs to the therapy genome. Otherwise, the mask bit is 0 that means the gene in this locus will not change during crossover operation.

Step 1. According to Darwin's evolution theory, two crossover parents are combined to produce new offspring in the hope that the fitness of next generation may improve gradually. In this paper, the therapy crossover wants to preserve parents' advantage and enhance population's diversity at the same time. Thus, offspring inherits the majority of parents' properties from the good parent than the bad one. Each mating parent has a different therapy rate according to its fitness; for example, good parent $\vec{x}_{\text{good}}(t)$ has a lower therapy rate (e.g., $p_{\text{good}} = 0.45$ in this paper) than that of the bad one $\vec{x}_{\text{bad}}(t)$ (e.g., $p_{\text{bad}} = 0.9$). On average, 45% of genes in the good parent will be merged with those genes in the bad parent.

Example 1. A minimization function $f(x) = \vec{x}^T \times \vec{x}$ is adopted here to illustrate the rationale of the therapy crossover. Let two parents with five genes at generation t be $\vec{x}_{\text{good}}(t) = [1.0\ \ 1.0\ \ 1.0\ \ 1.0\ \ 1.0]^T$ and $\vec{x}_{\text{bad}}(t) = [2.0\ \ 2.0\ \ 2.0\ \ 2.0\ \ 2.0]^T$, where $\vec{x}_{\text{good}}(t)$ is the good parent with better fitness $f(\vec{x}_{\text{good}}(t)) = 5$ and $\vec{x}_{\text{bad}}(t)$ is the bad one with worst fitness $f(\vec{x}_{\text{bad}}(t)) = 20$. Two therapy masks for $\vec{x}_{\text{good}}(t)$ and $\vec{x}_{\text{bad}}(t)$ are randomly generated by comparing five random numbers with therapy rates $p_{\text{good}} = 0.45$ and $p_{\text{bad}} = 0.9$, respectively. Without loss of generality, we assume that these two masks are $\vec{m}_{\text{good}} = [0\ \ 1\ \ 0\ \ 1\ \ 0]^T$ and $\vec{m}_{\text{bad}} = [1\ \ 1\ \ 1\ \ 1\ \ 0]^T$. That

is, the second and fourth genes of $\vec{x}_{\text{good}}(t)$ should be merged, and most of the genes in $\vec{x}_{\text{bad}}(t)$ should be merged except for the fifth one.

Step 2. For comparison purpose, two auxiliary chromosomes $\vec{s}_{\text{good}}$ and $\vec{s}_{\text{bad}}$ are generated for $\vec{x}_{\text{good}}(t)$ and $\vec{x}_{\text{bad}}(t)$, respectively. The auxiliary chromosome clones genes from the corresponding parent and replaces the selected therapy loci with those genes in the other parent. Thus, in (1), $\vec{s}_{\text{good}}$ copies all genes from $\vec{x}_{\text{good}}(t)$ and then replaces the genes (its locus marked by $\vec{m}_{\text{good}}$) by those genes in $\vec{x}_{\text{bad}}(t)$. On the other hand, (2) describes the value of each gene in $\vec{s}_b$ according to its therapy mask $\vec{m}_{\text{bad}}$:

$$\vec{s}_{\text{good}} = \left[s_{gi}\right], \quad \text{where } s_{gi} = x_{gi} \times \left(\neg m_{gi}\right) + x_{bi} \times m_{gi}, \quad (1)$$

$$\vec{s}_{\text{bad}} = \left[s_{bi}\right], \quad \text{where } s_{bi} = x_{bi} \times \left(\neg m_{bi}\right) + x_{gi} \times m_{bi}. \quad (2)$$

Notation ($\neg$) denotes logic negation (usually expressed by "NOT") which operates on one Boolean value and returns its complement as result.

Example 2. Thus, (1) guides $\vec{s}_{\text{good}}$ to copy the majority of genes (i.e., 1st, 3rd, and 5th) from $\vec{x}_{\text{good}}(t)$ and the minority of genes (i.e., 2nd and 4th) from $\vec{x}_{\text{bad}}(t)$, that is, $\vec{s}_{\text{good}} = [1.0 \;\; 2.0 \;\; 1.0 \;\; 2.0 \;\; 1.0]^T$, with respect to $\vec{m}_{\text{good}} = [0 \;\; 1 \;\; 0 \;\; 1 \;\; 0]^T$. For $\vec{x}_{\text{bad}}(t)$ in this example, $\vec{s}_{\text{bad}}$ also can be produced as $\vec{s}_{\text{bad}} = [1.0 \;\; 1.0 \;\; 1.0 \;\; 1.0 \;\; 2.0]^T$ by (2).

Step 3. Because it is difficult to determine the merit of a set of genes in a chromosome, a simple but effective method proposed in this paper is the mutual-evaluation approach, which measures the merit of two selected genomes by comparing the fitness changes before and after interchanging the genome with the other mating chromosome. Comparing the fitness change before and after the genome replacement (i.e., $f(\vec{x}_{\text{good}}(t))$ versus $f(\vec{s}_{\text{good}})$ and $f(\vec{x}_{\text{bad}}(t))$ versus $f(\vec{s}_{\text{bad}})$) can realize the substitution effect and can be used to represent the relative merit of these genomes.

Example 3. Because the fitness of $\vec{s}_{\text{good}}$ (i.e., $f(\vec{s}_{\text{good}}) = f([1.0 \;\; \mathbf{2.0} \;\; 1.0 \;\; \mathbf{2.0} \;\; 1.0]^T) = 11$) is worse than that of $\vec{x}_{\text{good}}(t)$ (i.e., $f(\vec{x}_{\text{good}}(t)) = f([1.0 \;\; \mathbf{1.0} \;\; 1.0 \;\; \mathbf{1.0} \;\; 1.0]^T) = 5$), we can intuitively deduce that genome $[{}^* \;\; 1.0 \;\; {}^* \;\; 1.0 \;\; {}^*]^T$ may perform better than genome $[{}^* \;\; 2.0 \;\; {}^* \;\; 2.0 \;\; {}^*]^T$ with respect to $[1.0 \;\; {}^* \;\; 1.0 \;\; {}^* \;\; 1.0]^T$. Therefore, the offspring $\vec{x}_{\text{good}}(t+1)$ inherits more genetic material from $[{}^* \;\; 1.0 \;\; {}^* \;\; 1.0 \;\; {}^*]^T$ than $[{}^* \;\; 2.0 \;\; {}^* \;\; 2.0 \;\; {}^*]^T$ for the second and forth genes. Comparing the fitness of $\vec{s}_{\text{bad}}$ (i.e., $f(\vec{s}_{\text{bad}}) = f([\mathbf{1.0} \;\; \mathbf{1.0} \;\; \mathbf{1.0} \;\; \mathbf{1.0} \;\; 2.0]^T) = 8$) and that of $\vec{x}_{\text{bad}}(t)$ (i.e., $f(\vec{x}_{\text{bad}}(t)) = f([\mathbf{2.0} \;\; \mathbf{2.0} \;\; \mathbf{2.0} \;\; \mathbf{2.0} \;\; 2.0]^T) = 20$), one can deduce that genome $[1.0 \;\; 1.0 \;\; 1.0 \;\; 1.0 \;\; {}^*]^T$ performs better than $[2.0 \;\; 2.0 \;\; 2.0 \;\; 2.0 \;\; {}^*]^T$ with respect to $[{}^* \;\; {}^* \;\; {}^* \;\; {}^* \;\; 2.0]^T$.

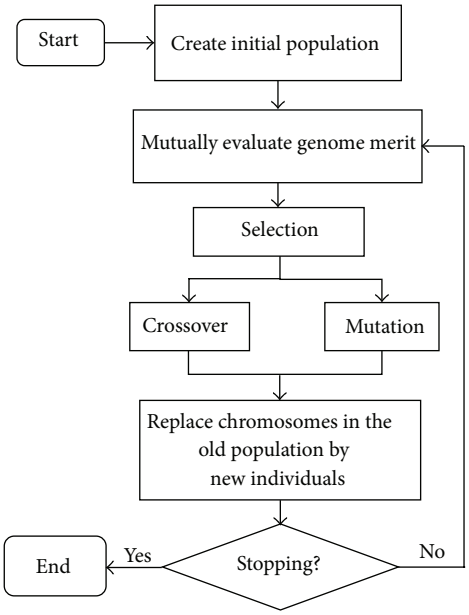


Figure 1: Flowchart of the proposed MEGA.

3. Mutual-Evaluation Genetic Algorithm (MEGA)

The main operations of the MEGA are initialization, mutual evaluation, selection, crossover, mutation, and replacement. We depict the flowchart of the MEGA in Figure 1 and describe their functionality in this section.

3.1. Encoding and Initialization. For illustration, the following minimization problem with fixed boundaries is considered:

$$\begin{aligned} &\text{Minimize} \quad f(\vec{x}) \\ &\text{subject to} \quad \vec{l} \le \vec{x} \le \vec{u}. \end{aligned} \quad (3)$$

Notation $\vec{x} = [x_1 \;\; x_2 \;\; \cdots \;\; x_N]^T \in \Re^N$ is the variable vector and $f(\vec{x})$ denotes the objective function. Because the lower bound $\vec{l} = [l_1 \;\; l_2 \;\; \cdots \;\; l_N]^T$ and the upper bound $\vec{u} = [u_1 \;\; u_2 \;\; \cdots \;\; u_N]^T$ define the feasible solution space, the domain of each x_i is denoted as interval $[l_i, u_i]$.

For numerical problems, each decision variable is treated as a gene and encoded by a floating-point number. Each chromosome representing a feasible solution is encoded as a vector of genes $\vec{x} = [x_1 \;\; x_2 \;\; \cdots \;\; x_N]^T$, where x_i denotes the value of the ith gene and N is total number of variables in an optimization problem. An initial population of M chromosomes is randomly generated within the feasible solution space $[\vec{l}, \vec{u}]$.

3.2. Selection Operation. Fitness of each chromosome represents the objective function value of this solution, denoted as $f_j = f(\vec{x}_j) = f([x_{j1} \;\; x_{j2} \;\; \cdots \;\; x_{jN}]^T)$ for the jth chromosome. The MEGA employs a traditional roulette selection

method as a discriminator of the solution quality that uses chromosome's fitness to create a selective pressure towards global optimal solution. Those chromosomes with higher fitness should have a greater selection chance, thus creating a selective pressure towards high-fitness solutions [1].

3.2.1. Therapy Crossover. The proposed MEGA incorporates the mutual-evaluation approach with a therapy crossover to enhance the exploitation ability and speed up the convergence rate. According to the determined relative merit of genome, the parents linearly combine their genomes to generate a new genome for their offspring like (4) for $\vec{x}_{\text{good}}(t+1)$ and (5) for $\vec{x}_{\text{bad}}(t+1)$, respectively. If a mask bit in its therapy mask is 0 (e.g., $m_{g1} = 0$), the gene at the locus of the good parent does not change (i.e., $x_{g1} = s_{g1}$). Otherwise, if a mask bit is 1 (e.g., $m_{g2} = 1$), the locus in the crossover child (e.g., $\vec{x}_{\text{good}}(t+1)$) inherits from both genetic materials from two parents (i.e., $x_{g1} \times \text{coef} + s_{g1} \times (1 - \text{coef})$). The principle of this linear combination is to retain the favorable schemata in the evolution process. Therefore, each gene of crossover child can be reproduced by the following equations:

$$\vec{x}_{\text{good}}(t+1) = \begin{cases} \vec{x}_{\text{good}}(t) \times \text{coef} + \vec{s}_{\text{good}} \times (1-\text{coef}), & \text{if } f\left(\vec{x}_{\text{good}}\right) \le f\left(\vec{s}_{\text{good}}\right) \\ \vec{x}_{\text{good}}(t) \times (1-\text{coef}) + \vec{s}_{\text{good}} \times \text{coef}, & \text{if } f\left(\vec{x}_{\text{good}}\right) > f\left(\vec{s}_{\text{good}}\right), \end{cases} \tag{4}$$

$$\vec{x}_{\text{bad}}(t+1) = \begin{cases} \vec{x}_{\text{bad}}(t) \times \text{coef} + \vec{s}_{\text{bad}} \times (1-\text{coef}), & \text{if } f\left(\vec{x}_{\text{bad}}\right) \le f\left(\vec{s}_{\text{bad}}\right) \\ \vec{x}_{\text{bad}}(t) \times (1-\text{coef}) + \vec{s}_{\text{bad}} \times \text{coef}, & \text{if } f\left(\vec{x}_{\text{bad}}\right) > f\left(\vec{s}_{\text{bad}}\right). \end{cases} \tag{5}$$

Coefficient coef = 0.8 + (rand/2) is a random value in interval [0.8, 1.3] to reduce the deficiency of premature convergence. The rand function returns a uniformly distributed pseudo-random number between 0 and 1.

Example 4. In this example, we assume the random coefficient coef = 1.2. Because $f(\vec{x}_{\text{good}}(t)) = 5$ is better than $f(\vec{s}_{\text{good}}) = 11$, (4) guides us to get $\vec{x}_{\text{good}}(t+1) = \vec{x}_{\text{good}}(t) \times 1.2 + \vec{s}_{\text{good}} \times (1 - 1.2) = \begin{bmatrix} 1 & 0.8 & 1 & 0.8 & 1 \end{bmatrix}^T$. Thus, this child's fitness $f(\vec{x}_{\text{good}}(t+1)) = 4.28$ is better than $f(\vec{x}_{\text{good}}(t)) = 5$. Similarly, (5) calculates $\vec{x}_{\text{bad}}(t+1) = \vec{x}_{\text{bad}}(t) \times (1 - 1.2) + \vec{s}_{\text{bad}} \times 1.2 = \begin{bmatrix} 0.8 & 0.8 & 0.8 & 0.8 & 2 \end{bmatrix}^T$ because fitness $f(\vec{x}_{\text{bad}}(t)) = 20$ is worse than $f(\vec{s}_{\text{bad}}) = 8$. The child's fitness $f(\vec{x}_{\text{bad}}(t+1)) = 6.56$ is better than $f(\vec{x}_{\text{bad}}(t)) = 20$. Therefore, both children have better fitness values than that of their parents.

The exclusive features of the therapy crossover include that (1) the merit of each genome is evaluated individually; and (2) the gene merit facilitates the MEGA to perform an efficient search by adaptively shifting emphasis on significant genome without explicit functional analysis [15]. That is, the therapy crossover can avoid frequently throwing away potential schemata in inferior chromosomes and inherit the genetic advantages of superior chromosomes without loss of genetic diversity.

3.2.2. Partial-Gaussian Mutation. In this paper, a partial-Gaussian mutation used in the MEGA can increase population diversity to enhance its exploration ability. The original Gaussian mutation proposed by Hinterding in 1995 can converge to near-optimal solutions of some multimodal optimization problems [16]. Our partial-Gaussian mutation concentrates on exploiting potential optimal areas and speeds up the convergent effect. At the beginning of evolution, a high mutation rate is assigned to sample the search space extensively. And then, the mutation rate decreases with time to fine tune solutions and concentrates on exploiting potential optimal areas. Equation (6) calculates the mutation rate p_m as

$$p_m = 0.5 \times \left(1 - \frac{\text{Current Generation}}{\text{Maximal Generation}}\right), \tag{6}$$

where maximal generation is 3000. A random number in interval [0, 1] is generated for each gene and then compared with the mutation rate p_m. If the mutation rate is greater than or equal to the random number, this gene value will be flipped by adding a unit Gaussian distributed random value to the chosen gene. Otherwise, no mutation occurs at this gene. This mutation operation can only be used for integer and float genes.

3.3. Reproduction Operation. The MEGA adopts a replacement-with-elitism method to monotonously enhance the solution quality. A number of the elite parents can survive into next generation for preventing good solutions from being lost through a nondeterministic selection operation. Successive population consists of three evolutionary sources: (1) the elite 10% chromosomes can survive to next generation; (2) 80% offsprings are produced by the crossover operation; and, (3) the rest 10% chromosomes are produced by the mutation operation.

4. Performance Analyses for Numerical Optimization Problems

4.1. Numerical Test Functions. Numerical experiments are conducted to demonstrate the robustness and reliability of the proposed MEGA. This work selects 12 well-known benchmark functions, which cover broad range functionality characteristics with two categories: unimodal functions (Functions f_1–f_6) and multimodal functions (Function f_7–f_{12}). Table 1 depicts these test functions with their formulation, problem dimension (N), prescribed search domain (D), and their global optimum function value ($f_{\min}$) in each column. The high-dimension unimodal functions are the Sphere function (f_1), the Schwefel's Problem 2.22 (f_2), the Schwefel's Problem 1.2 (f_3), the Schwefel's Problem 2.21 (f_4), a modified Rosenbrock function (f_5), and the noisy quadratic function (f_6). Unimodal functions are relatively easy to solve but the difficulty increases as the problem dimension goes high. The high-dimension multimodal functions are a generalized Schwefel's function 7 (f_7), the Rastrigin's function 6 (f_8),

Table 1: Twelve benchmark functions.

Benchmark functions	N	D	$f_{\min}$
$f_1(x) = \sum_{i=1}^{N} x_i^2$	30	$[-100, 100]$	0
$f_2(x) = \sum_{i=1}^{N} \lvert x_i \rvert + \prod_{i=1}^{N} \lvert x_i \rvert$	30	$[-10, 10]$	0
$f_3(x) = \sum_{i=1}^{N} \left(\sum_{j=1}^{i} x_j\right)^2$	30	$[-100, 100]$	0
$f_4(x) = \max\{\lvert x_i \rvert,\ i = 1, 2, \ldots, N\}$	30	$[-100, 100]$	0
$f_5(x) = \sum_{i=1}^{N-1} \left[100\left(x_i^2 - x_{i+1}\right)^2 + \left(x_i - 1\right)^2\right]$	30	$[-5, 10]$	0
$f_6(x) = \sum_{i=1}^{N} x_i^4 + \text{random}[0, 1)$	30	$[-1.28, 1.28]$	0
$f_7(x) = \sum_{i=1}^{N} \left(-x_i \sin\left(\sqrt{\lvert x_i \rvert}\right)\right)$	30	$[-500, 500]$	-12596.5
$f_8(x) = \sum_{i=1}^{N} \left[x_i^2 - 10\cos(2\pi x_i) + 10\right]$	30	$[-5.12, 5.12]$	0
$f_9(x) = -20\exp\left(-0.2\sqrt{1/n \sum_{i=1}^{n} x_i^2}\right) - \exp\left((1/N)\sum_{i=1}^{N} \cos(2\pi x_i)\right) + 20 + \exp(1)$	30	$[-32, 32]$	0
$f_{10}(x) = (1/4000)\sum_{i=1}^{N} x_i^2 - \prod_{i=1}^{N} \cos(x_i/\sqrt{i}) + 1$	30	$[-600, 600]$	0
$f_{11}(x) = \frac{\pi}{N}\left\{10\sin^2(\pi y_1) + \left(y_N - 1\right)^2 + \sum_{i=1}^{N-1}\left(y_i - 1\right)^2 \times \left[1 + 10\sin^2(\pi y_{i+1})\right]\right\} + \sum_{i=1}^{N} u(x_i, 10, 100, 4)$ where $y_i = 1 + (1/4)(x_i + 1)$ and $u(x_i, a, k, m) = \begin{cases} k(x_i - a)^m & x_i > a \\ 0 & -a \le x_i \le a \\ k(-x_i - a)^m & x_i < -a \end{cases}$	30	$[-50, 50]$	0
$f_{12} = (1/10)\left\{\sin^2(3\pi x_1) + \left(x_N - 1\right)^2\left[1 + \sin^2(2\pi x_N)\right] + \sum_{i=1}^{N-1}\left(x_i - 1\right)^2\left[1 + \sin^2(3\pi x_{i+1})\right]\right\} + \sum_{i=1}^{N} u(x_i, 5, 100, 4)$	30	$[-50, 50]$	0

N: problem dimension; D: prescribed search domain; $f_{\min}$: global optimum value.

Table 2: Experimental results obtained by the MEGA over 50 independent trails.

Test function	$f_{\min}$	Mean best value	Standard deviation	Mean required generations	Computational effort (MNFE)
f_1	0	2.44×10^{-19}	1.33×10^{-18}	64	17,280
f_2	0	2.00×10^{-7}	4.27×10^{-7}	725	195,750
f_3	0	5.49×10^{-30}	2.74×10^{-29}	650	175,500
f_4	0	1.99×10^{-8}	2.42×10^{-8}	185	49,950
f_5	0	0.03762	0.02816	88	23,760
f_6	0	3.78×10^{-3}	1.84×10^{-3}	96	25,920
f_7	−12596.5	−12596.5	9.98×10^{-3}	132	35,640
f_8	0	0.0	0.0	155	41,850
f_9	0	2.78×10^{-8}	5.15×10^{-8}	400	108,000
f_{10}	0	0.0	0.0	23	6,210
f_{11}	0	7.85×10^{-7}	4.42×10^{-7}	215	58,050
f_{12}	0	1.11×10^{-5}	8.31×10^{-6}	70	18,900

Table 3: Comparison of the mean best value (and Std Dev) with other EAs.

Test function	PSO (Variance)	EO (Variance)	CEP	FEP	MEGA
f_1	11.175 (1.3208)	9.8808 (0.9444)	2.2×10^{-4} (5.9×10^{-4})	5.7×10^{-4} (1.3×10^{-4})	2.44×10^{-19} (1.33×10^{-18})
f_2	N/A	N/A	2.6×10^{-3} (1.710^{-4})	8.1×10^{-3} (7.7×10^{-4})	2.00×10^{-7} (4.27×10^{-7})
f_3	N/A	N/A	0.05 (0.066)	0.016 (0.014)	5.49×10^{-30} (2.74×10^{-29})
f_4	N/A	N/A	2.0 (1.2)	0.3 (0.5)	1.99×10^{-8} (2.42×10^{-8})
f_5	1911.598 (374.2935)	1610.39 (293.5783)	6.17 (13.61)	5.06 (5.87)	0.03762 (0.02816)
f_6	N/A	N/A	0.018 (6.4×10^{-3})	7.6×10^{-3} (2.6×10^{-3})	3.78×10^{-3} (1.84×10^{-3})
f_7	N/A	N/A	−7917.1 (634.5)	−12554.5 (52.6)	−12596.5 (9.98×10^{-3})
f_8	47.1354 (1.8782)	46.4689 (2.4545)	89.0 (23.1)	0.046 (0.012)	0.0 (0.0)
f_9	N/A	N/A	9.2 (2.8)	0.018 (0.0021)	2.78×10^{-8} (5.15×10^{-8})
f_{10}	0.4498 (0.0566)	0.4033 (0.0436)	2.52×10^{-7}	0.016 (0.022)	0.0 (0.0)
f_{11}	N/A	N/A	1.76 (2.4)	9.2×10^{-6} (3.6×10^{-6})	7.85×10^{-7} (4.42×10^{-7})
f_{12}	N/A	N/A	1.4 (3.7)	1.6×10^{-4} (7.3×10^{-5})	1.11×10^{-5} (8.31×10^{-6})

a modified Ackley's Path Function 10 (f_9), the Griewank's function 8 (f_{10}), a generalized Penalized Function 1 (f_{11}), and a generalized Penalized Function 2 (f_{12}). Multimodal functions represent the most difficult class of problems, which possess many local optima and could trap an algorithm into one of its local optimal solutions.

4.2. Algorithm Implementation and Parameter Settings. In all cases, the population size is 150, in which the number of elite individuals is 15; the therapy crossover produces 120 individuals, and the mutation produces 15 ones. The therapy rates for good parent and bad parent are $p_{\text{good}} = 0.45$ and $p_{\text{bad}} = (1 - p_{\text{good}}) = 0.55$, respectively. For each test function, 50 independent trials with different seeds are performed using the MATLAB environment.

For complexity analysis, the mean number of function evaluations serves as a measure of required computational effort for an algorithm. Different from the crossover in traditional GAs, the exclusive feature of MEGA is the usage of the mutual-evaluation approach for genome therapy. Because

Table 4: Comparison of the computational effort (MNFE) with other EAs.

Test function	PSO	EO	CEP	FEP	MEGA
f_1	250,000	500,000	100,500	100,500	17,280
f_2	N/A	N/A	200,000	200,000	195,750
f_3	N/A	N/A	500,000	500,000	175,500
f_4	N/A	N/A	500,000	500,000	49,950
f_5	250,000	500,000	2000,000	2000,000	23,760
f_6	N/A	N/A	300,000	300,000	25,920
f_7	N/A	N/A	900,000	900,000	35,640
f_8	250,000	500,000	500,000	500,000	41,850
f_9	N/A	N/A	150,000	150,000	108,000
f_{10}	250,000	500,000	200,000	200,000	6,210
f_{11}	N/A	N/A	150,000	150,000	58,050
f_{12}	N/A	N/A	150,000	150,000	18,900

only an auxiliary chromosome should be extraevaluated for one crossover child, the number of required evaluations in each generation is (function evaluations per generation) = [(population size) × (1 + crossover fraction)] = [150 × (1 + 80%)] = 270. Therefore, the total number of function evaluations in each experimental run is equal to (functional evaluations per generation) × (no. of terminal generations).

4.3. Experimental Results. Table 2 summarizes the experimental results of the MEGA in 50 trials, which include (1) the global optimum ($f_{\min}$), (2) the mean best function value, (3) the standard deviation of the obtained function values, (4) the mean required generation, and (5) the mean number of function evaluations (MNFEs).

The first thing we can observe from Table 2 is that the obtained results are equal or really close to the "known" optimal solutions. Particulary, this paper uses a computational precision of 60 digits after point. Thus, the results "0" on f_8 and f_{10} in Table 2 mean that they are less than 10^{-60}. The standard deviation with respect to the functions f_8 and f_{10} is equal to zero; that is, the results of all 50 runs reach the optimum. All the obtained results approach the "known" optimal values with small differences. Secondly, the small standard deviations for all test functions also indicate that the MEGA consistently converges to the near-optimal solutions in all 50 trails. Finally, all of the mean numbers of function evaluations are relatively small. Therefore, the proposed MEGA can address a variety of numerical optimization functions effectively. To further analyze the solution capability of the proposed MEGA, the following sections describe the comparisons between the MEGA and two groups of optimization algorithms for the 12 benchmark functions.

4.4. Comparison with Other Evolutionary Algorithms. The performance of the MEGA is compared with four state-of-the-art EAs: particle swarm optimization (PSO) [17], evolutionary optimization (EO) [17], conventional evolutionary programming (CEP) [6], and FEP [6]. The comparison of the experimental results for 12 test functions is shown in Table 3.

For all the unimodal functions (f_1–f_6) in Table 3, we can observe that the MEGA can achieve dramatically the highest accuracy than others, while other four algorithms experience premature convergence on functions f_3, f_4, and f_5. For all the multimodal functions (f_7–f_{10}), the results shown in Table 3 clearly indicate that the MEGA can identify the actual optima of these functions with the highest accuracy. The MEGA can achieve better solution accuracy than other four EAs for all 12 benchmark functions.

The computational efforts required for the algorithms are measured by their mean numbers of function evaluations and depicted in Table 4. Obviously, the comparison demonstrates that the MEGA outperformed all the four algorithms for all the 12 functions with respect to the convergent ability. Therefore, the comparison indicates that the MEGA is both efficient and effective in solving the unimodal and multimodal benchmark functions.

4.5. Comparison with Other Genetic Algorithms. The performance of the MEGA is compared with those of four well-known GAs: cluster-based adaptive mutation genetic algorithm (CMGA) [18], orthogonal genetic algorithm with quantization (OGA/Q) [19], hybrid taguchi genetic algorithm (HTGA) [20], and StGA [7]. The OGA/Q is a quantized-version of the OGA, which incorporates with the Taguchi method to minimize the effect of chromosome variation without eliminating the population diversity [21]. The HTGA enhanced the Taguchi method as a new operation to adapt a dynamically extended precision method from a low-precision solution space to a high-precision one [20]. The above algorithms have been executed to solve the test functions and the results were reported in the literature. We will use these existing results for a direct comparison in Tables 5 and 6.

As the termination criteria used in these four algorithms are different, to make a fair comparison basis, we let the solution qualities obtained by our MEGA be slightly better than those of the four algorithms (in Table 5), and then, compared the mean computational effort at the given accuracy (in Table 6). For the unimodal functions (f_1–f_6), the convergence rate of an algorithm is a more important issue

Table 5: Comparison of the mean best value (and Std Dev) with other GAs.

Test function	CMGA	OGA/Q	HTGA	StGA (Variance)	MEGA
f_1	N/A	0 (0)	0 (0)	2.45×10^{-15} (5.25×10^{-16})	2.44×10^{-19} (1.33×10^{-18})
f_2	N/A	0 (0)	0 (0)	2.0×10^{-7} (2.95×10^{-8})	2.00×10^{-7} (4.27×10^{-7})
f_3	N/A	0 (0)	0 (0)	9.98×10^{-29} (6.90×10^{-29})	5.49×10^{-30} (2.74×10^{-29})
f_4	N/A	0 (0)	0 (0)	2.01×10^{-8} (3.42×10^{-9})	1.99×10^{-8} (2.42×10^{-8})
f_5	N/A	0.7520 (0.1140)	0.7 (0)	0.04435 (0.0)	0.03762 (0.02816)
f_6	N/A	6.301×10^{-3} (4.069×10^{-4})	1.000×10^{-3} (0)	8.4×10^{-4} (1.00×10^{-3})	3.78×10^{-3} (1.84×10^{-3})
f_7	−8722.16	−12569.4537 (6.447×10^{-4})	−12569.4600 (0)	−12569.5 (0.0)	−12596.5 (9.98×10^{-3})
f_8	157.76	0 (0)	0 (0)	4.42×10^{-13} (1.14×10^{-13})	0.0 (0.0)
f_9	N/A	4.440×10^{-16} (3.989×10^{-17})	0 (0)	3.52×10^{-8} (3.51×10^{-9})	2.78×10^{-8} (5.15×10^{-8})
f_{10}	0.3283	0 (0)	0 (0)	2.44×10^{-17} (4.54×10^{-17})	0.0 (0.0)
f_{11}	N/A	6.019×10^{-6} (1.159×10^{-6})	1.000×10^{-6} (0)	8.03×10^{-7} (1.96×10^{-14})	7.85×10^{-7} (4.42×10^{-7})
f_{12}	N/A	1.869×10^{-4} (2.615×10^{-5})	1.000×10^{-4} (0)	1.13×10^{-5} (4.62×10^{-13})	1.11×10^{-5} (8.31×10^{-6})

Table 6: Comparison of the computational effort (MNFE) with other GAs.

Test function	CMGA	OGA/Q	HTGA	StGA	MEGA
f_1	N/A	112,559	20,844	30,000	17,280
f_2	N/A	112,612	14,285	17,600	195,750
f_3	N/A	112,576	26,469	23,000	175,500
f_4	N/A	112,893	21,261	32,000	49,950
f_5	N/A	167,863	60,737	45,000	23,760
f_6	N/A	112,652	20,065	25,500	25,920
f_7	600,000	302,166	163,468	1,500	35,640
f_8	600,000	224,710	16,267	28,500	41,850
f_9	N/A	112,421	16,632	10,000	108,000
f_{10}	600,000	134,000	20,999	52,500	6,210
f_{11}	N/A	134,556	66,457	8,000	58,050
f_{12}	N/A	134,143	59,003	16,000	18,900

than the achieved satisfactory accuracy. Because of different problem dimensions used in the HTGA (i.e., $N = 100$), this study cannot compare its performance with other algorithms for these six unimodal functions. In Table 6, we can observe that the proposed MEGA requires fewer MNFEs than the OGA/Q for four of the six test functions. The convergence rate of the MEGA was similar to that of the StGA for a half of the six test functions. Because of the narrow valleys of f_2 and f_3, the MEGA was forced to change its searching direction continually; thus, it approached the high-accuracy optimum slowly. That is why the convergence rates of the MEGA were lower than those of other algorithms for f_2 and f_3.

For the multimodal functions (f_7–f_{12}), the quality of the solutions is more crucial than the required computational effort because the solution quality reveals the algorithm's ability to escape from local optima and achieves near-global solutions. From the obtained results of the numerical experiments in Tables 5 and 6, we can see that the MEGA was superior to the CMGA with respect to both solution accuracy and convergence rate. The solution accuracies achieved by

the MEGA were similar to those of the OGA/Q, HTGA, and StGA for all the six multimodal functions. However, the MEGA converged slower than the HTGA and StGA for the multimodal function f_9. This finding implies that the proposed MEGA is robust and effective in solving the multimodal functions and can perform as good as the four state-of-the-art GAs.

5. Performance Evaluation for Multipath Routing Problems

5.1. Multipath Routing Problems. Multipath traffic engineering becomes more attractive for ubiquitous networks to satisfy the required QoS of mobile network applications [22]. For a given network topology with available link capacities, it is required to determine the optimal distribution of traffic requests on multipath routing subject to the constraints imposed by QoS specifications. The service network is modeled as a connected weighted, directed graph $G = (V, E)$, where the $V = \{v_1, v_2, \ldots, v_n\}$ is the vertex set of G, and the $E = \{e_1, e_2, \ldots, e_m\}$ is a finite set of edges. We consider a link-state routing environment, where each node has knowledge of its neighbor links and each source node knows (1) traffic load from node i to node j ($i \neq j$), (2) QoS requirements/constraints (e.g., delay constraint), (3) residual bandwidth of each link, (4) link cost for the traffic, and (5) time delay for the traffic to pass through each link at a certain time period.

The link $e = (i, j) \in E$ connects a source node $i \in V$ to a destination node $j \in V$ with positive cost c_e, capacity b_e, and a delay function (i.e., $D_e : \Re^+ \rightarrow \Re^+$). The cost function represents the traffic-related delay of each edge and limit to positive cost. The delay requirement specifies the upper bound of delay tolerance, denoted as Δ_w.

In multipath routing problems, messages are routed for a set of OD pairs W, where t_w denotes the traffic demand volume (or bandwidth) for OD pair w. The set of all permissible paths for the wth OD pair is denoted as P_w. Let indicator function f_{wp} be a nonnegative continuous variable denoting the traffic flow allocated on path p of request w. Indicator function y^e_{wp} is 1 if link e belongs to path p for request w, and 0 otherwise; that is, $y_{pi} = \begin{cases} 1, & \text{if } e_i \in p \\ 0, & \text{otherwise,} \end{cases}$ where $i = 1, 2, \ldots, m$. The aggregate flow on link e is denoted as g_e, which must satisfy the capacity feasibility constraint; that is, $g_e = \sum_{w \in W} \sum_{p \in P_w} f_{wp} y^e_{wp} \leq b_e$, for all $e \in E$.

The multipath routing problem is to find multiple paths to transmit the traffic demand for all OD pairs such that the total routing cost is minimal and the QoS constraints should be satisfied. The multipath routing can be modeled as a combinatorial linear programming problem in the following:

objective function,

$$\text{minimize} \quad \sum_{e \in E} c_e g_e \tag{7}$$

$$\text{subject to} \quad g_e = \sum_{w \in W} \sum_{p \in P_w} f_{wp} y^e_{wp} \quad \forall e \in E \tag{8}$$

$$g_e \leq b_e \quad \forall e \in E \tag{9}$$

$$\sum_{e \in E} y^e_{wp} f_{wp} D_e(g_e) \leq \Delta_w \quad \forall p \in P_w,\ w \in W \tag{10}$$

$$\sum_{p \in P_w} f_{wp} = t_w \quad \forall w \in W \tag{11}$$

$$f_{wp} \leq t_w r_{wp} \quad \forall p \in P_w,\ w \in W \tag{12}$$

$$\sum_{p \in P_w} r_{wp} = x_w \quad \forall w \in W \tag{13}$$

$$r_{wp} \text{ binary} \quad \forall p \in P_w,\ w \in W. \tag{14}$$

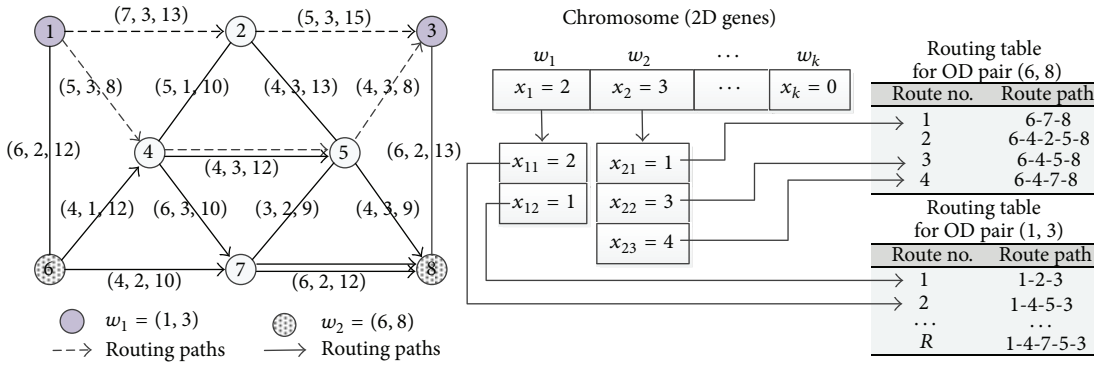
The objective function (in (7)) is to minimize the total routing cost in the multipath routing problem. The first constraint (in (8)) calculates the aggregate traffic on link e. The second constraint assures the capacity constraint on each link by (9). The third constraint (in (10)) ensures the total delay of each OD pair to satisfy a prespecified path-delay bound Δ_w. The constraint in (11) enforces that traffic demand of each OD pair should be satisfied. Equation (12) lets an auxiliary binary variable r_{wp} be 1 if the traffic f_{wp} along path p is larger than zero, and 0 otherwise. Equation (13) sums up the used routes and assigns the number to the wth gene x_w.

5.2. Algorithm Implementation for Multipath Routing Problems

5.2.1. Encoding. The encoding method is the most important steps towards solving real world problems using EAs. In this paper, the encoding method maps all multipath OD pairs into a chromosome based on route aspect. Since there are so many candidate paths between two nodes in the network graph, traditional GAs may consume considerable computational effort in searching infeasible solutions because genetic operations do not always preserve feasibility. Therefore, to reduce the search space, this work uses the K shortest path routing algorithm to precalculate the first R shortest paths and record in a routing table.

The MEGA maintains a population of chromosomes to optimize a given objective function for the multipath routing problem. Each chromosome can be represented by a two-dimensional array of integers. The first-dimension genes $\vec{x} = [x_1 \ x_2 \ \cdots \ x_N]^T$ represent the number of used routes for each OD pair. The second-dimension genes record the route number of each used path for realizing the demand of each OD pair. Thus, if the gene value of the wth gene is $x_w = k$, the vector $[x_{w1}, x_{w2}, \ldots, x_{wk}]^T$ represents k route numbers of subflows for OD pair w. Gene x_{wp} is an integer in interval $[1, R]$ to represent a route number for the OD pair w in its routing table. We use an example of network topology to illustrate the relationship between a chromosome, 2D genes, and routing tables in Figure 2.

Figure 2(a) depicts an example of a network graph, link parameters, and two multipath routings. Parameters along links are triple (cost, delay, and bandwidth). The first OD pair, that is, $w_1 = (1, 3)$, has two routing paths $p_{12} = \{(1, 4), (4, 5), (5, 3)\}$ and $p_{11} = \{(1, 2), (2, 3)\}$ to transmit traffic from node 1 to node 3. Thus, the first-dimension gene $x_1 = 2$.



(a) Network topology example with two OD pairs (b) Representation of a chromosome (c) Routing tables

Figure 2: Example of genotype coding: (a) network graph with link parameter (cost, delay, and bandwidth), (b) representation of chromosome by using 2D genes, and (c) two routing tables for paths for node 6 to node 8 and paths for node 1 to node 3.

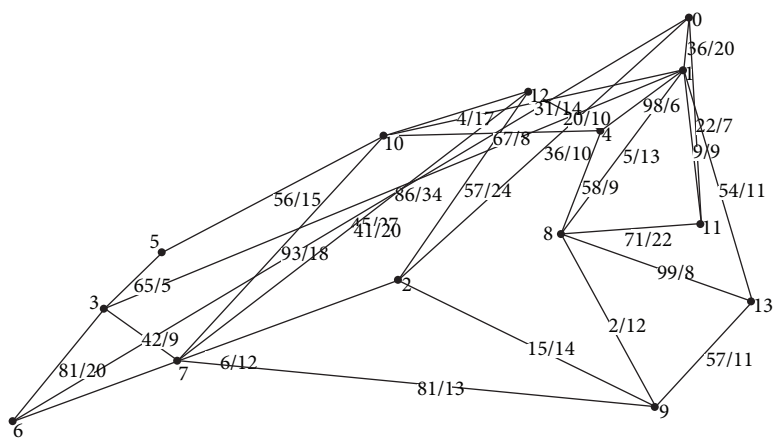


Figure 3: A randomly generated network with 14 nodes and average degree 4.

The route numbers in its routing table (in Figure 2(c)) corresponding to paths p_{12} and p_{11} are 2 and 1; thus, the second-dimension genes for w_1 (in Figure 2(b)) record $x_{11} = 2$ and $x_{12} = 1$, respectively. In Figure 2(a), the second OD pair, that is, $w_2 = (6, 8)$ routes its traffic along three subflows along paths p_{21}, p_{23}, and p_{24}. Therefore, the first-dimension gene is $x_2 = 3$ and the corresponding second-dimension genes are $x_{21} = 1$, $x_{22} = 3$, and $x_{23} = 4$, respectively.

5.2.2. Therapy Crossover for Multipath Routing. Each time the selection operation chooses two crossover parents from the population. The proposed mutual-evaluation approach calculates the merit of two selected genomes by comparing the changes in the chromosome fitness before and after interchanging the first- and second-dimension genomes with the other mating chromosome. The first-dimension genes can linearly combine with those in the other chromosome by using the proposed therapy crossover (in Section 3.2.1 (4) and (5)). The obtained results should be transformed into integer type, and therefore, the second-dimension genes should be modified. If the obtained result of the first-dimension gene is larger than before, we randomly select a suitable number of genes from the other parent to add into the second-dimension genes in this chromosome. If the result equals to the original value, we randomly select a small number of genes from the other parent to replace the original genes. Otherwise, a suitable number of genes should be randomly selected to remove from this chromosome.

5.2.3. Mutation for Multipath Routing. Mutation operation performs on an individual chromosome to flip one or more genes with a small probability (typically 0.001) and ensures that no point in the search space has a zero probability of being searched. According to a mutation probability, the mutation randomly selects a subset of genes and chooses new paths from its routing table. Thus, the route numbers of these new paths replace the original values of selected genes. The resulting chromosome is a new multipath routing plan that can increase population diversity.

5.3. Test Platform and Performance Metrics. In this paper, we use the well-known network generation tool [23] to create an asynchronous network based on the Waxman's techniques [24]. The network in Figure 3 illustrates a random generated

Table 7: The mean experimental results for the low-rate cases (0.25 Mbps) obtained by (a) MEGA and (b) TGA.

(a) MEGA

OD pairs	3	4	5	6	7	8	9	10
Case 1	116.16	160.66	193.84	234.28	267.5	298.12	334.64	368.47
Case 2	132.4	154.1	176.45	226.84	254.5	287.3	335.1	374.12
Case 3	100.84	141.1	171.5	210.56	254.67	291.6	321.64	341.6
Case 4	114.5	157.64	189.67	231.99	260.46	290.14	327.45	360.41
Case 5	125.1	170.14	203.46	248.82	176.45	308.2	343.71	386.46
Avg.	117.8	156.73	186.98	230.50	242.72	295.07	332.51	366.21

(b) TGA

OD pairs	3	4	5	6	7	8	9	10
Case 1	138.31	173.62	235.81	249.45	284.65	301.64	355.14	380.39
Case 2	154.96	184.60	193.59	236.07	282.32	321.73	363.22	392.32
Case 3	129.25	174.79	188.36	233.30	271.43	309.39	345.11	371.40
Case 4	137.98	192.39	203.39	250.22	266.99	313.95	354.95	384.30
Case 5	165.01	186.55	220.27	256.96	190.42	311.60	354.38	392.94
Avg.	145.10	182.39	208.28	245.20	259.16	311.66	354.56	384.27

Table 8: The standard deviation of results for the low-rate cases (0.25 Mbps) obtained by (a) MEGA and (b) TGA.

(a) MEGA

OD pairs	3	4	5	6	7	8	9	10
Case 1	4.28	5.12	7.01	12.80	11.79	14.31	15.26	16.82
Case 2	4.79	6.63	7.98	11.82	12.92	14.50	17.06	14.94
Case 3	3.49	6.89	6.76	10.58	13.01	14.30	15.13	16.92
Case 4	4.41	4.86	6.36	10.52	12.91	12.50	15.80	17.84
Case 5	5.27	5.60	6.23	9.62	13.47	14.18	15.71	15.24
Avg.	4.45	5.82	6.87	11.07	12.82	13.96	15.79	16.35

(b) TGA

OD pairs	3	4	5	6	7	8	9	10
Case 1	7.56	7.63	7.57	15.15	14.99	16.55	17.17	18.40
Case 2	7.04	9.34	8.35	15.36	14.23	17.49	18.59	19.31
Case 3	7.14	7.69	9.28	14.81	15.77	15.55	17.78	21.26
Case 4	8.24	9.47	9.05	14.66	15.37	16.50	18.72	18.06
Case 5	6.85	10.47	11.21	14.30	14.28	17.43	19.91	16.61
Avg.	7.37	8.92	9.09	14.86	14.93	16.70	18.43	18.73

graph with 14 nodes and the average degree of each node is four. Figure 3 only shows the cost/delay information along one direction link (from a smaller-ID node to a larger-ID one) to reduce the complexity of the graph representation. All links are assumed to have 1.5 Mbps of bandwidth capacity. For each test scenario, the OD pairs are randomly generated five times for each scenario to decrease the selection bias. Two kinds of transmission rates are assigned for the OD pairs: 0.25 Mbps (lowrate) and 0.5 Mbps (highrate).

The performance of the proposed MEGA is evaluated based on the following ways.

(1) *The total routing cost*: This cost reflects the algorithm's ability to construct multiple paths for all OD pairs by using low-cost and lightly utilized links.

(2) *The maximum end-to-end delay for OD pairs*: It indicates the algorithm's ability to satisfy the delay bound imposed by the service level agreement of applications.

An algorithm's effectiveness in allocating network resources can be judged by monitoring how frequently that algorithm fails to construct a set of acceptable OD pairs. There are two kinds of failure. One is that the created OD pair does not satisfy its delay bound. The other one is that the algorithm cannot find unsaturated links to create a path for OD pairs.

5.4. Computational Experiments for Multipath Routing Problems. The performance of the proposed MEGA is compared

Table 9: The mean experimental results for the high-rate cases (0.5 Mbps) obtained by MEGA and TGA.

Algorithm	MEGA				TGA			
OD pairs	3	4	5	6	3	4	5	6
Case 1	116.50	159.69	193.06	231.94	124.30	166.01	204.70	248.78
Case 2	110.84	142.02	180.71	209.44	120.51	149.94	200.26	223.59
Case 3	134.10	165.12	201.36	235.76	144.18	176.07	213.18	255.57
Case 4	100.94	136.34	170.46	192.10	124.78	150.85	179.52	209.35
Case 5	121.34	158.49	187.54	224.62	134.62	172.35	204.96	228.07
Avg.	116.74	152.33	186.63	218.77	129.68	163.04	200.52	233.07

Table 10: The standard deviation of results for the high-rate cases (0.5 Mbps) obtained by MEGA and TGA.

Algorithm	MEGA				TGA			
OD pairs	3	4	5	6	3	4	5	6
Case 1	4.58	5.50	7.59	10.04	5.79	6.55	8.55	11.13
Case 2	5.14	6.49	9.12	11.24	5.71	6.62	10.39	11.90
Case 3	3.72	4.98	7.12	9.56	5.21	6.47	9.12	11.14
Case 4	4.30	6.30	8.84	10.56	5.67	7.71	9.68	10.67
Case 5	5.04	6.81	9.40	10.54	6.93	8.43	10.66	12.20
Avg.	4.56	6.02	8.41	10.39	5.86	7.16	9.68	11.41

with the TGA for two kinds of multipath routing scenarios with respect to different transmission rates and different number of total OD pairs in the network. In the first test scenario, the experiments are performed for 100 independent runs with 10 generations in the MATLAB environment. The evolution curves of total routing costs with respect to lowrate and highrate are compared in Figures 4(a) and 4(b), respectively. The evolution curves in Figure 4(a) show that the efficiency of the MEGA is better than that of the TGA in both two transmission rates. Particulary, even though the transmit rate increases two times than the low-rate case, the growing ratio of the total routing cost in Figure 4(b) is still less than that obtained by the TGA.

The second scenario simulates the stress test to measure the robustness of these two algorithms. We increase the number of OD pairs from 3 to 10 and conduct 100 independent trials for each test case. The OD pairs are randomly generated five times for each scenario to simulate the great diversity of OD-pair selection for performance evaluation.

Tables 7 and 8 are the "mean" and "standard deviation" experimental results obtained by the proposed MEGA and TGA for five cases in the low-rate scenario (with 0.25 Mbps transmission rate). The mean experimental results obtained by the proposed MEGA are better than the results of the TGA on all the test cases. That is, the MEGA can find better routing paths to serve all multipath transmission requirements than TGA (in Table 7). For the standard deviations in 100 independent runs, the proposed algorithm achieved superior results compared with the TGA in all the test cases (in Table 8). This finding implies that the proposed algorithm is robust in solving multipath routing problems and can perform better than the TGA.

In the high-rate scenario, the test network cannot afford the data flows if the number of OD pairs is larger than 6. Thus, the following experiments only increase the number of OD pairs from 3 to 6 and the transmission rate is 0.5 Mbps for the high-rate cases. We simulate 100 independent trials for each test case and depict the "mean" and "standard deviation" of the experimental results obtained by the MEGA and TGA in Tables 9 and 10, respectively. Experimental results indicate that the proposed MEGA outperforms the TGA with respect to the "mean" and "standard deviation" of routing costs for all high-rate scenarios. Furthermore, the search ability of the MEGA is robust in obtaining consistent results and performs better than the TGA.

6. Conclusions and Future Works

To the best of our knowledge, the proposed MEGA is the first mutual-evaluation approach, which calculates each genome merit by interchange-compare-replace method. The genome evaluation facilitates the MEGA to perform an efficient search by dynamically shifting emphasis to significant genomes in the feasible space without abdicating any portion of the candidate schemata. The therapy crossover is also proposed to preserve better-performance schema patterns. Simpler than other modified approaches, the proposed MEGA can preserve high quality genomes during evolution period without using extra analyzing techniques.

The performance of the proposed algorithm was measured using 12 benchmark functions. The performance was compared with four existing EAs and four well-known GAs. The experiment results show that the MEGA is able to find near-optimal solutions, even though other algorithms experience difficulties in approaching the global optima on some functions. The behavior of the algorithm is also consistent as indicated by a small standard deviation among the 50 trials for each test function. Furthermore, the MEGA can increase the accuracy by several orders of magnitude than other algorithms in almost all test functions. That is, the MEGA can

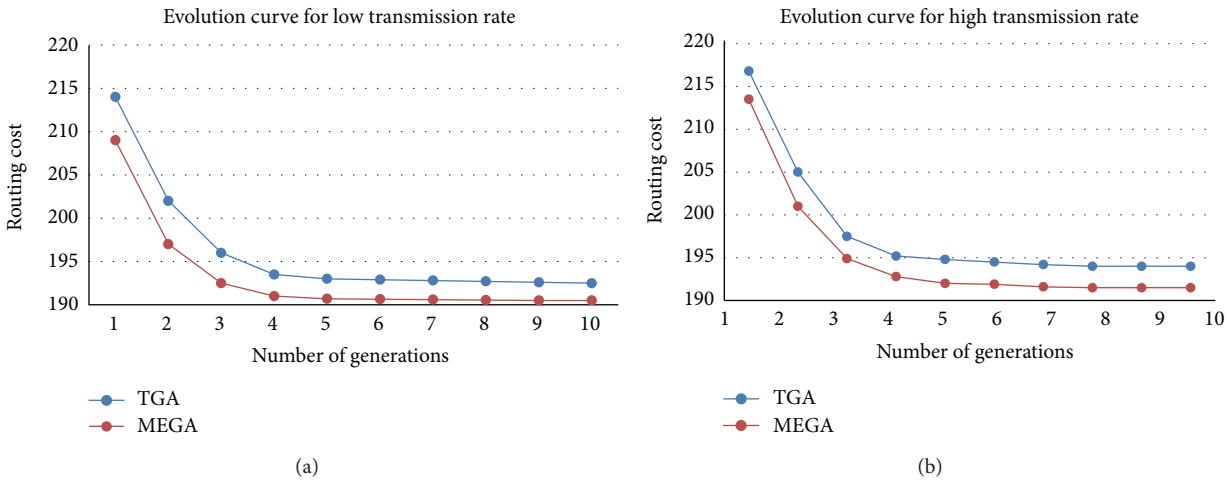


Figure 4: Evolution curves of two different transmission rates (a) 0.25 Mbps and (b) 0.5 Mbps.

outperform the existing global optimization algorithms with a dramatic improvement in terms of effectiveness.

Furthermore, because of the simplification property of the mutual-evaluation approach, this study slightly modified the encoding method of the MEGA to solve the multipath routing problems in multicommodity networks. The bandwidth-delay constraints are introduced to enforce the service quality of multimedia applications. The experimental results show that the MEGA not only can solve QoS constrained multipath routing problems, but also can outperform the TGA. That is, the proposed MEGA not only can reduce the effort of explicit function analysis but also can deal with a wide spectrum of real world problems.

The contributions of the paper are as follows. (1) We develop a novel mutual-evaluation approach which incorporates with a GA that has never been jointly considered in the literature. (2) We introduce a novel therapy crossover, which not only can evolve superior genomes but also can achieve global optima. (3) We introduce a novel chromosome representation for the MEGA to address multipath routing problems in a multicommodity network. (4) Experimental results show that the proposed MEGA can achieve significant performance gain over several well-known algorithms under the considered scenarios. The proposed MEGA has high exploration and exploitation abilities as a robust, statistically sound, and quickly convergent algorithm.

We have observed that there are many researches for routing problems. In the future, we will further compare with more state-of-the-art methods on QoS multipath routing. Particulary, the simplicity property of the MEGA can help to route dynamic services across heterogeneous environments. Therefore, developing a distributed MEGA to enhance its scalability for highly dynamic environments is also our future work.

References

[1] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, Springer, Berlin, Germany, 2nd edition, 1994.

[2] M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, Mass, USA, 1996.

[3] L. Wei and M. Zhao, "A niche hybrid genetic algorithm for global optimization of continuous multimodal functions," *Applied Mathematics and Computation*, vol. 160, no. 3, pp. 649–661, 2005.

[4] H. R. Mashhadi, H. M. Shanechi, and C. Lucas, "A new genetic algorithm with Lamarckian individual learning for generation scheduling," *IEEE Transactions on Power Systems*, vol. 18, no. 3, pp. 1181–1186, 2003.

[5] J. A. Vasconcelos, J. A. Ramírez, R. H. C. Takahashi, and R. R. Saldanha, "Improvements in genetic algorithms," *IEEE Transactions on Magnetics*, vol. 37, no. 5 I, pp. 3414–3417, 2001.

[6] X. Yao, Y. Liu, and G. Lin, "Evolutionary programming made faster," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 2, pp. 82–102, 1999.

[7] Z. Tu and Y. Lu, "A robust stochastic genetic algorithm (StGA) for global numerical optimization," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 5, pp. 456–470, 2004.

[8] G. R. Harik, F. G. Lobo, and D. E. Goldberg, "The compact genetic algorithm," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 4, pp. 287–297, 1999.

[9] C. W. Ahn and R. S. Ramakrishna, "Elitism-based compact genetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 7, no. 4, pp. 367–385, 2003.

[10] S. Rimeharoen, D. Sutivong, and P. Chongstitvatana, "Updating strategy in compact genetic algorithm using moving average approach," in *Proceedings of IEEE Conference on Cybernetics and Intelligent Systems*, pp. 1–6, June 2006.

[11] J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, Mich, USA, 1975.

[12] J. McCall, "Genetic algorithms for modelling and optimisation," *Journal of Computational and Applied Mathematics*, vol. 184, no. 1, pp. 205–222, 2005.

[13] E. C. Yeh and Y.-Y. Shyu, "New genetic algorithm with statistical gene evaluation," in *Proceedings of the 1st International Joint Conference of NAFIPS/IFIS/NASA*, pp. 409–410, December 1994.

[14] N. Kubota, K. Shimojima, and T. Fukuda, "Role of virus infection in virus-evolutionary genetic algorithm," in *Proceedings*

of IEEE International Conference on Evolutionary Computation (ICEC '96), pp. 182–187, May 1996.

[15] C. H. Lin, "A rough penalty genetic algorithm for constrained optimization," *Information Sciences*, vol. 241, pp. 119–1137, 2013.

[16] R. Hinterding, "Gaussian mutation and self-adaption for numeric genetic algorithms," in *Proceedings of IEEE International Conference on Evolutionary Computation*, pp. 384–388, December 1995.

[17] P. J. Angeline, "Evolutionary optimization versus particle swarm optimization: philosophy and performance differences," in *Proceedings of the Evolutionary Programming VII*, pp. 601–610, 1998.

[18] T.-Y. Sun, C.-C. Liu, S.-T. Hsieh, C.-G. Lin, and K.-Y. Lee, "Cluster-based adaptive mutation mechanism to improve the performance of genetic algorithm," in *Proceedings of the 6th International Conference on Intelligent Systems Design and Applications (ISDA '06)*, pp. 461–466, October 2006.

[19] Y.-W. Leung and Y. Wang, "An orthogonal genetic algorithm with quantization for global numerical optimization," *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 1, pp. 41–53, 2001.

[20] J.-T. Tsai, T.-K. Liu, and J.-H. Chou, "Hybrid Taguchi-genetic algorithm for global numerical optimization," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 4, pp. 365–377, 2004.

[21] Q. Zhang and Y.-W. Leung, "An orthogonal genetic algorithm for multimedia multicast routing," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 1, pp. 53–62, 1999.

[22] E. M. El-Alfy, S. N. Mujahid, and S. Z. Selim, "A Pareto-based hybrid multiobjective evolutionary approach for constrained multipath traffic engineering optimization in MPLS/GMPLS networks," *Journal of Network and Computer Applications*, vol. 36, pp. 1196–1207, 2013.

[23] H. Salama, "The multicast routing simulator," The Real-Time Communication Project, 1997, http://rtcomm.csc.ncsu.edu/index.htm.

[24] B. M. Waxman, "Routing of multipoint connections," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1617–1622, 1988.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 876191, 9 pages
http://dx.doi.org/10.1155/2013/876191

Research Article

Interference Control for Cognitive Network with High Mobility

Yuanxuan Li, Gang Zhu, Siyu Lin, Ke Guan, and Bo Ai

State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China

Correspondence should be addressed to Yuanxuan Li; g.liyuanxuan@gmail.com

Received 19 April 2013; Revised 9 July 2013; Accepted 23 July 2013

Academic Editor: Chih-Hao Lin

Interference control (IC) between the secondary system and the primary system is an important issue for underlay cognitive radio network (CRN). The secondary system should limit the interference power to primary system by adjusting its transmission power. Many relevant works have been done based on the assumption of the quasistatic channel which is not suitable for the fast time-varying fading channel; the performance of IC in underlay CRN will become worse when the channel varies fast. This paper studies the IC issue in high mobility environment. By considering the channel state information (CSI) outdatedness, a short frame structure scheme and a mean interference power constraint scheme are proposed to reduce the influence of CSI outdatedness on IC performance. Furthermore, by considering the channel estimation error, a spherical error region model based robust IC scheme is designed as well. The proposed IC schemes of the secondary system are converted to the power allocation problems, and then they are formulated to optimization problem whose objects are to maximize the capacity of the secondary system with the interference constraints. The above optimization problems are solved by the water-filling style method. The simulation results show that the proposed IC schemes can effectively control the interference power to the primary system.

1. Introduction

Traditionally, the different spectrum bands have been allocated to different mobile communication systems, such as Wideband Code Division Multiple Access (WCDMA) system, Long-Term Evolution (LTE) system, and Long-Term Evolution Advanced (LTE-A) system, by the spectrum authorization. The fixed spectrum allocation policy is the most convenient way for spectrum management avoiding the interference among different communication systems. However, the current spectrum management policies have seriously limited the development of mobile communication systems [1]. Cognitive network technology has been considered as a promising way to improve the spectrum efficiency. In cognitive radio networks (CRNs), the users are divided into two classes: primary users (PUs) and secondary users (SUs). There are two operation types of CRNs: underlay and interweave (overlay) CRNs. For interweave CRNs, the SUs can use the spectrum band only if the PUs are inactive in this band [2]. If the PUs are under heavy-load status over long term, the SUs can hardly use the spectrum bands. To meet the increasing quality of service (QoS) requirements of SUs, the underlay CRN has been proposed, which permits the SUs to utilize the spectrum whenever the PUs are active or idle [3]. In this case, SUs can get more freedom and better spectrum utilization. However, to guarantee the PUs' QoS, the SUs should control their transmission power to avoid interfering with the PUs.

The underlay CRNs require that the interference from SUs does not disturb the signal detection and decoding of PUs, so the interference power should be limited under a predefined threshold at primary receiver (PR). The interference control (IC) plays a key role on the CRNs operation. How to maximize the capacity of the secondary system with interference constraint is an open issue for the CRNs optimization. In [4–7], there were several power allocation schemes with IC constraints in underlay CRNs. The design of interference constraints for SUs has been proposed in [8]. The distribution of interference power from SU has been discussed in [9]. In [10], Rabbachin et al. presented a statistical IC model considering the influence of channel fading. However, most IC schemes assumed that the wireless channel is quasistatic, which cannot fully reflect the nature of fading channel. The IC model with quasistatic channel assumption especially in the

high mobility environment [11] will became invalid over time. Recently, some researchers started to concern the impact of the mobile fading channel for IC at SUs. Using the channel correlation coefficient to model the channel variation, Qiu et al. designed the interference constraints for the transmission schemes of SUs [12]. The time variation property of channel fading brings more challenges for IC in the secondary system. For the SU, the IC design should consider the impact of time-varying channel in different movement speed situations. Furthermore, since the IC constraints design relies on the channel state information (CSI), the channel estimation error will lead to the failure of IC. Therefore, the CSI outdatedness and channel estimation error should be considered for IC in the high mobility environment.

This paper focuses on the design of IC schemes for the secondary system in the high mobility environment. The IC schemes of the secondary system are converted to the power allocation problems. These problems can be formulated to optimization problem whose objects are to maximize the capacity of the secondary system with the interference constraints. The contributions are fourfold. First, in order to mitigate the influence of the CSI outdatedness, a short frame structure for IC scheme in normal mobility environment is proposed. The power allocation results are given by water-filling style solutions. Second, for high mobility environment, the IC schemes with mean interference power constraint are proposed; then the power allocation problem is transformed to a convex problem by Jensen's inequality. Third, in order to cope with the channel estimation error problem, the robust IC scheme based on spherical error region model is proposed as well. Similar to the second IC scheme, the solution can be given through solving the approximated convex problem. Finally, the above three IC schemes are evaluated in different environments. Compared with the existing schemes, the simulation results show that the proposed IC schemes can control the interference power to the primary system effectively.

The rest of the paper is organized as follows: Section 2 describes the underlay system model. Section 3 proposes three interference control schemes. The secondary user power allocation schemes for IC will be analyzed in detail. The simulation results of interference control with mobile environment are presented in Section 4. The conclusions are made in Section 5.

2. System Model

The system model of underlay CRN is shown in Figure 1. The primary network has the primary base station (PBS) and the primary user (PU) which are interfered by the secondary base station (SBS) and the SU. The SU and the SBS should limit their transmission power to meet the requirement of signal to interference plus noise ratio (SINR) at the PU and the PBS. The CSIs between the SU(SBS) and the PBS(PU) are denoted by $H_{I_{\mathrm{UL}}}$ and $H_{I_{\mathrm{DL}}}$. The $H_{s_{\mathrm{UL}}}$ and $H_{s_{\mathrm{DL}}}$ represent uplink and downlink CSIs of the secondary system. The secondary devices (SU and SBS) are assumed to be able to control their transmission power by interference constraint using the $H_{I_{\mathrm{UL}}}$ and $H_{I_{\mathrm{DL}}}$. Both primary system

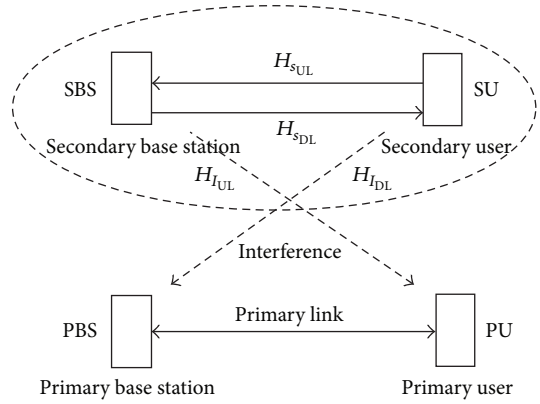


Figure 1: The underlay CR network topology.

and secondary system are orthogonal frequency division multiplexing (OFDM) systems with time division duplexing (TDD) and well synchronization characters. When the PU uploads the information to the PBS in the uplink, the SU transmits its signal to the SBS in the uplink at the same time slot. Similar to the uplink, the downlink at the secondary system is synchronized with the downlink at the primary system. The SUs access the CRN by the carrier sense multiple access with the collision avoidance (CSMA/CA) method which prevents the collisions between SUs. Assuming that the SU is moving (e.g., SU in a vehicle or a high speed train) and the PU is in low mobility status, the time-varying channel fading follows a zero-mean complex Gaussian process. The channel state of the primary network is quasistatic, which means that the channel state keeps steady in a frame duration. Secondary system utilizes the CSI ($H_{I_{\mathrm{UL}}}$ and $H_{I_{\mathrm{DL}}}$) to design its IC scheme to control the interference. However, between the primary system and the secondary system, the channel state varies from symbol to symbol. Therefore, the CSI will be outdated in a frame duration. The relationship between $H(t, f)$ and $H(t+\tau, f)$ can be described as Clark-Jakes' model whose autocorrelation function is denoted by $\rho(\tau) = \sigma_0^2 J_0(2\pi f_d \tau)$, where σ_0^2 is the variance of the channel, $J_0(\cdot)$ denotes the zero-order Bessel function of the first kind, f_d is the maximum doppler frequency, and τ is the outdated time [13].

The channel capacity of SU at the tth symbol can be calculated as

$$C(t) = \sum_{f=1}^{M} B \log_2 \left(1 + \frac{P_s(t, f) \left| H_s(t, f) \right|^2}{N_0 + I_p} \right), \qquad (1)$$

where B is the bandwidth of the secondary system, I_p is the interference power from the primary system, N_0 is the noise power, $P_s(t, f)$ is the transmission power of the secondary system at the fth subband in the tth symbol time, and $H_s(t, f)$ is the CSI of secondary link. The total transmission power constraint is $\sum_{f=1}^{M} P_s(t, f) \leq \Phi$, where $P_s(t, f) \geq 0$, for all f. The secondary system should control its transmission power with interference constraints while maximizing the channel capacity $C(t)$. The influence of CSI outdatedness is different in various movement speeds.

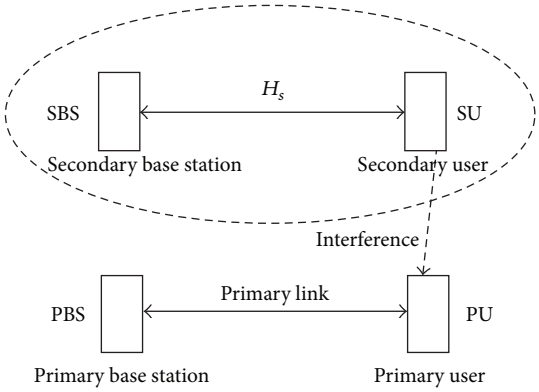


Figure 2: An underlay CRN with the primary system and the secondary system.

Furthermore, the channel estimation error cannot be avoided in time-varying fading channel, so it may lead to the failure of IC. Hence, how to control the interference in mobile environment is an open issue for the secondary system. In the following, several interference constraints are proposed to limit the interference power to the primary system. First, for the influence of CSI outdatedness, two IC schemes (a short frame structure in normal mobility environment and a mean interference power constraint) are designed in high mobility environment. Second, to cope with the channel estimation error, the channel estimation error on the basis of the spherical error region model for IC is formulated.

3. The Power Allocations at Secondary System with Interference Constraints

3.1. Traditional Interference Control Scheme. As shown in Figure 2, the existing works assume that the channel state keeps steady in one frame duration which does not vary from symbol to symbol. The $H_I(-\tau_I, f)$ represents the CSI between the SU and the PU which is delayed by τ_I symbols. The $H_s(-\tau_s, f)$ represents the CSI with the secondary system which is delayed by τ_s symbols.

The interference constraint in the fth subband at the tth symbol time is

$$I(t,f) = P_s(t,f)\left|H_I(-\tau_I,f)\right|^2 \le \phi, \quad \forall f, \tag{2}$$

where ϕ is the interference power threshold, which is the maximum interference power that the primary system can tolerate. The object of the optimization is maximizing the channel capacity of the secondary system. Therefore, the optimization problem with the interference power constraints can be formulated as

$$\min\left\{-C(t) = \sum_{f=1}^{M} B\log_2\left(1+\frac{P_s(t,f)\left|H_s(-\tau_s,f)\right|^2}{N_0+I_p}\right)\right\},$$

$$\text{s.t.} \quad \sum_{f=1}^{M} P_s(t,f) \le \Phi, \qquad -P_s(t,f) \le 0,$$
$$P_s(t,f)\left|H_I(-\tau_I,f)\right|^2 \le \phi. \tag{F-1}$$

Because it is a convex optimization problem, the Lagrangian formulation is given by

$$\begin{aligned} L(P_s(t,f)) = &-\sum_{f=1}^{M} B\log_2\left(1+\frac{P_s(t,f)\left|H_s(-\tau_s,f)\right|^2}{N_0+I_p}\right) \\ &+\sum_{f=1}^{M}\alpha(-P_s(t,f)) \\ &+\sum_{f=1}^{M}\beta\left(P_s(t,f)-\frac{\phi}{\left|H_I(-\tau_I,f)\right|^2}\right) \\ &+\gamma\left(\sum_{f=1}^{M}P_s(t,f)-\Phi\right), \end{aligned} \tag{3}$$

where α, β, and γ are Lagrangian multipliers. The Karush-Kuhn-Tucker (KKT) conditions are used to solve this optimization problem [12]

$$\begin{aligned} &\frac{\partial L(P_s(t,f))}{P_s} \\ &\quad = -\frac{B}{\ln 2}\frac{\left|H_s(-\tau_s,f)\right|^2}{N_0+I_p+\left|H_s(-\tau_s,f)\right|^2 P_s(t,f)} - \alpha + \beta + \gamma \\ &\quad = 0, \\ &\qquad \alpha \ge 0, \quad \alpha P_s(t,f) = 0, \\ &\qquad \beta \ge 0, \quad \beta\left(P_s - \frac{\phi}{\left|H_I(-\tau_I,f)\right|^2}\right) = 0. \end{aligned} \tag{4}$$

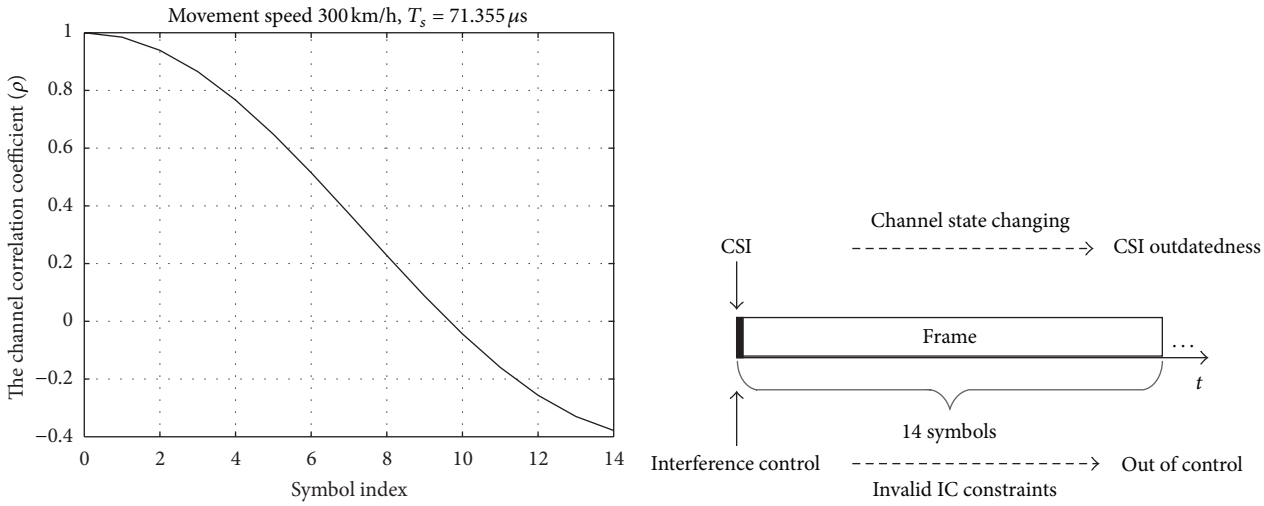
The water-filling style solution is

$$P_s(t,f) = \begin{cases} P_{A1}, & \dfrac{B}{\mu\ln 2} \ge P_{A1}+P_{A2}, \\ \dfrac{B}{\mu\ln 2} - P_{A2}, & P_{A2} < \dfrac{B}{\mu\ln 2} < P_{A1}+P_{A2}, \\ 0, & \dfrac{B}{\mu\ln 2} \le P_{A2}, \end{cases} \tag{5}$$

where $P_{A1} = \phi/|H_I(-\tau_I,f)|^2$, $P_{A2} = (N_0+I_p)/|H_s(-\tau_s,f)|^2$, and μ is the Lagrange multiplier.

The constant constraint with the assumption of the quasistatic channel is a simplified way to control the interference power to the primary system. The actual interference power in the fth subband at the τth symbol is

$$I(t,f) = P_s(t,f)\left|H_I(t+\tau,f)\right|^2, \tag{6}$$



(a) The change of channel correlation

(b) The CSI outdatedness process

Figure 3: The CSI outdatedness in standard LTE transmission frame.

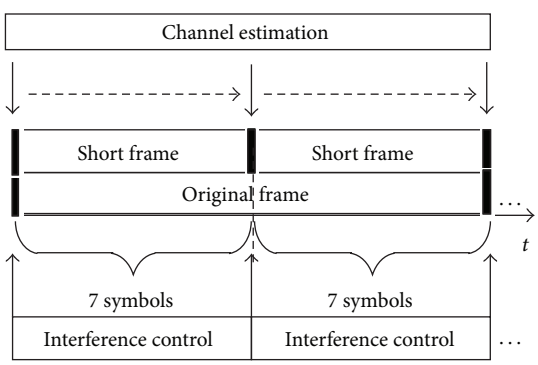


Figure 4: The short frame structure.

where τ is the number of outdated symbols. Taking the LTE system configuration, for example, there are 14 symbols in one frame. As shown in Figure 3, the CSI outdatedness process can be explained by the change of the channel correlation. The channel correlation coefficient ρ decreases from symbol to symbol, which means that the gap between CSI and channel gain of current symbol becomes larger. The IC schemes with the quasistatic assumption cannot track the variation of fading channel, and, therefore, it will be invalid over time. This transmission scheme can only ensure the interference power under the predefined threshold in the beginning of a frame duration. In the rest of the time, the interference power may exceed the predefined threshold caused by CSI outdatedness. Thus, the CSI outdatedness should be considered in the high mobility environment.

3.2. Interference Control with Short Frame Structure. Only considering the interference between SU and PU is not proper for IC in underlay CRNs. As shown in Figure 1, the secondary system needs to control the interference power to PU and PBS. Both the primary system and the secondary system are the TDD systems with well synchronization. For the primary system, the SU disturbs the PBS in uplink, and the SBS disturbs the PU in downlink. For the channel between primary devices and secondary devices, $H_{I_{\mathrm{UL}}}(-\tau_1, f)$ and $H_{I_{\mathrm{DL}}}(-\tau_2, f)$ denote that the CSI knowledge is outdated by τ_1 and τ_2 symbols at subband f in uplink and downlink, respectively. For the channel between SU and SBS, $H_{s_{\mathrm{UL}}}(-\tau_{s_1}, f)$ denotes that the CSI of uplink channel is outdated by τ_{s_1} symbols at subband f, and $H_{s_{\mathrm{DL}}}(-\tau_{s_2}, f)$ denotes that the CSI of downlink channel is outdated by τ_{s_2} symbols at subband f.

The CSI in the high mobility environment will be changed faster, which leads to more serious CSI outdatedness in time-varying fading channel. The time-varying fading channel varies in a small range; in a short time interval, its dynamic characteristics can be described by Clark-Jakes' model. Therefore, a new frame structure is required to fit the mobile environment. So, intuitively, the shorter frame duration means the smaller outdated time with CSI. In Figure 4, a short frame structure is proposed for IC in mobile environment which has 7-symbol duration. Shorting the frame length can improve the performance of IC and reduce the interference power from SU

$$\begin{aligned}
&\min \left\{ -C(t) \right. \\
&\quad = \left. \sum_{f=1}^{M} B \log_2 \left(1 + \frac{P_{s_{UL}}(t,f) \left| H_{s_{\mathrm{UL}}}(-\tau_{s_1}, f) \right|^2}{N_0 + I_p} \right) \right\} \\
&+ \min \left\{ -C(t) \right. \\
&\quad = \left. \sum_{f=1}^{M} B \log_2 \left(1 + \frac{P_{s_{\mathrm{DL}}}(t,f) \left| H_{s_{\mathrm{DL}}}(-\tau_{s_2}, f) \right|^2}{N_0 + I_p} \right) \right\}
\end{aligned}$$

$$\begin{aligned}
\text{s.t.}\quad & P_{s_{\mathrm{UL}}}(t,f)\left|H_{I_{\mathrm{UL}}}(-\tau_1,f)\right|^2 \le \phi,\\
& P_{s_{\mathrm{DL}}}(t,f)\left|H_{I_{\mathrm{DL}}}(-\tau_2,f)\right|^2 \le \phi,\\
& \sum_{f=1}^{M} P_{s_{\mathrm{UL}}}(t,f) \le \Phi, \qquad -P_{s_{\mathrm{UL}}}(t,f) \le 0,\\
& \sum_{f=1}^{M} P_{s_{\mathrm{DL}}}(t,f) \le \Phi, \qquad -P_{s_{\mathrm{DL}}}(t,f) \le 0.
\end{aligned} \tag{F-2}$$

The solution is

$$P_{s_{\mathrm{UL}}} = \begin{cases} P_{B1}, & \dfrac{B}{\mu \ln 2} \ge P_{B1}+P_{B2},\\ \dfrac{B}{\mu \ln 2} - P_{B2}, & P_{B2} < \dfrac{B}{\mu \ln 2} < P_{B1}+P_{B2},\\ 0, & \dfrac{B}{\mu \ln 2} \le P_{B2}, \end{cases}$$
$$P_{s_{\mathrm{DL}}} = \begin{cases} P_{B3}, & \dfrac{B}{\mu \ln 2} \ge P_{B3}+P_{B4},\\ \dfrac{B}{\mu \ln 2} - P_{B4}, & P_{B4} < \dfrac{B}{\mu \ln 2} < P_{B3}+P_{B4},\\ 0, & \dfrac{B}{\mu \ln 2} \le P_{B4}, \end{cases} \tag{7}$$

where $P_{B1} = \phi/|H_{I_{\mathrm{UL}}}(-\tau_1,f)|^2$, $P_{B3} = \phi/|H_{I_{DL}}(-\tau_2,f)|^2$, $P_{B2} = (N_0+I_p)/|H_s(-\tau_{s_1},f)|^2$, and $P_{B4} = (N_0+I_p)/|H_s(-\tau_{s_2},f)|^2$.

3.3. Interference Control with Mean Interference Power Constraint. Considering high speed environments, such as high speed railway system, it is much harder to control the interference power $I(t,f)$ under a predefined threshold all the time due to the fast time-varying nature of the channel. Therefore, a mean interference power constraint based interference control scheme is proposed to maximize the mean capacity of the channel of the secondary system. ϕ is the interference threshold, and I_p is the interference power from the primary system. The uplink and the downlink correlation coefficients between the SU to the PBS and the SBS to the PU are denoted by $\rho_{I_{\mathrm{UL}}}(t) = J_0(2\pi f_d(t+\tau_1))$ and $\rho_{I_{\mathrm{DL}}}(t) = J_0(2\pi f_d(t+\tau_2))$. And the $\rho_{s_{\mathrm{UL}}}(t) = J_0(2\pi f_d(t+\tau_{s_1}))$ and $\rho_{s_{DL}}(t) = J_0(2\pi f_d(t+\tau_{s_2}))$ denote the uplink and downlink channel correlation coefficients between the SU and the SBS. Assuming that the uplink and the downlink channels are independent rayleigh channel, the CSI can be denoted as $|H_{I_{\mathrm{UL}}}(t)| = |\mu_{U1}(t) + j\mu_{U2}(t)|$ and $|H_{I_{\mathrm{DL}}}(t)| = |\mu_{D1}(t) + j\mu_{D2}(t)|$, which can be described as a complex Gaussian random processes and the variance is $2\sigma^2$ [13]. The interference constraints of the uplink and the downlink can be described as

$$E\left[I(t,f)|_{H_{I_{\mathrm{UL}}}}\right] = P_{s_{\mathrm{UL}}}(t,f)\,E\left[\left|H_{I_{\mathrm{UL}}}(t,f)\right|^2\right] \le \phi, \tag{8}$$

$$E\left[I(t,f)|_{H_{I_{\mathrm{DL}}}}\right] = P_{s_{\mathrm{DL}}}(t,f)\,E\left[\left|H_{I_{\mathrm{DL}}}(t,f)\right|^2\right] \le \phi. \tag{9}$$

For time-varying fading channel, the CSI can be expressed as

$$\begin{aligned}
H_{s_{\mathrm{UL}}}(t,f) &= \rho_{s_{\mathrm{UL}}}(t)\,H_{s_{\mathrm{UL}}}\left(-\tau_{s_1},f\right) + \widetilde{H}_{s_{\mathrm{UL}}}(t,f),\\
H_{s_{\mathrm{DL}}}(t,f) &= \rho_{s_{\mathrm{DL}}}(t)\,H_{s_{\mathrm{DL}}}\left(-\tau_{s_2},f\right) + \widetilde{H}_{s_{\mathrm{DL}}}(t,f),\\
H_{I_{\mathrm{UL}}}(t,f) &= \rho_{I_{\mathrm{UL}}}(t)\,H_{I_{\mathrm{UL}}}(-\tau_1,f) + \widetilde{H}_{I_{\mathrm{UL}}}(t,f),\\
H_{I_{\mathrm{DL}}}(t,f) &= \rho_{I_{\mathrm{DL}}}(t)\,H_{I_{\mathrm{DL}}}(-\tau_2,f) + \widetilde{H}_{I_{\mathrm{DL}}}(t,f),
\end{aligned} \tag{10}$$

where $\widetilde{H}_{I_{\mathrm{UL}}} \sim \mathcal{N}(0, 1-\rho^2_{I_{\mathrm{UL}}}(t))$, $\widetilde{H}_{I_{\mathrm{DL}}} \sim \mathcal{N}(0, 1-\rho^2_{I_{\mathrm{DL}}}(t))$, $\widetilde{H}_{s_{\mathrm{UL}}} \sim \mathcal{N}(0, 1-\rho^2_{s_{\mathrm{UL}}}(t))$, and $\widetilde{H}_{s_{\mathrm{DL}}} \sim \mathcal{N}(0, 1-\rho^2_{s_{\mathrm{DL}}}(t))$.

Then, we obtain the expected value of CSI

$$E\left[\left|H_{s_{\mathrm{UL}}}\right|^2\right] = \rho^2_{I_{\mathrm{UL}}}(t)\left|H_s\left(-\tau_{s_1},f\right)\right|^2 + \left(1-\rho^2_{s_{\mathrm{UL}}}(t)\right)2\sigma^2, \tag{11}$$

$$E\left[\left|H_{s_{\mathrm{DL}}}\right|^2\right] = \rho^2_{I_{\mathrm{DL}}}(t)\left|H_s\left(-\tau_{s_2},f\right)\right|^2 + \left(1-\rho^2_{s_{\mathrm{DL}}}(t)\right)2\sigma^2, \tag{12}$$

$$E\left[\left|H_{I_{\mathrm{UL}}}\right|^2\right] = \rho^2_{I_{\mathrm{UL}}}(t)\left|H_I(-\tau_1,f)\right|^2 + \left(1-\rho^2_{I_{\mathrm{UL}}}(t)\right)2\sigma^2, \tag{13}$$

$$E\left[\left|H_{I_{\mathrm{DL}}}\right|^2\right] = \rho^2_{I_{\mathrm{DL}}}(t)\left|H_I(-\tau_2,f)\right|^2 + \left(1-\rho^2_{I_{\mathrm{DL}}}(t)\right)2\sigma^2. \tag{14}$$

By substituting (13) and (14) into (8), power allocation with interference constraint can be written as

$$\begin{aligned}
P_{s_{\mathrm{UL}}} &\le \frac{\phi}{\rho^2_{I_{\mathrm{UL}}}(t)\left|H_I(-\tau_1,f)\right|^2 + \left(1-\rho^2_{I_{\mathrm{UL}}}(t)\right)2\sigma^2},\\
P_{s_{\mathrm{DL}}} &\le \frac{\phi}{\rho^2_{I_{\mathrm{DL}}}(t)\left|H_I(-\tau_2,f)\right|^2 + \left(1-\rho^2_{I_{\mathrm{DL}}}(t)\right)2\sigma^2}.
\end{aligned} \tag{15}$$

Because the instantaneous channel capacity of the secondary system is changed over time, in the long run, the optimization object is the expected value of channel capacity which can be formulated as

$$C = E\left[C_{\mathrm{UL}} + C_{\mathrm{DL}}\right]. \tag{16}$$

The optimization formulation at the tth symbol is

$$\begin{aligned}
&\min\left\{-E\left[\sum_{f=1}^{M} B\log_2\left(1+\frac{P_{s_{\mathrm{UL}}}(t,f)\left|H_{s_{\mathrm{UL}}}\left(-\tau_{s_1},f\right)\right|^2}{N_0+I_p}\right)\right]\right\}\\
&\quad+\min\left\{-E\left[\sum_{f=1}^{M} B\log_2\left(1+\frac{P_{s_{\mathrm{DL}}}(t,f)\left|H_{s_{\mathrm{DL}}}\left(-\tau_{s_2},f\right)\right|^2}{N_0+I_p}\right)\right]\right\},
\end{aligned}$$

$$\begin{aligned}
\text{s.t.}\quad & P_{s_{\mathrm{UL}}} \le \frac{\phi}{\rho_{I_{\mathrm{UL}}}^2(t)\,H_I(-\tau_1,f)+\left(1-\rho_{I_{\mathrm{UL}}}^2(t)\right)2\sigma^2},\\
& P_{s_{\mathrm{DL}}} \le \frac{\phi}{\rho_{I_{\mathrm{DL}}}^2(t)\,H_I(-\tau_2,f)+\left(1-\rho_{I_{\mathrm{DL}}}^2(t)\right)2\sigma^2},\\
& \sum_{f=1}^{M} P_{s_{\mathrm{UL}}}(t,f) \le \Phi, \qquad -P_{s_{\mathrm{UL}}}(t,f) \le 0,\\
& \sum_{f=1}^{M} P_{s_{\mathrm{DL}}}(t,f) \le \Phi, \qquad -P_{s_{\mathrm{DL}}}(t,f) \le 0.
\end{aligned} \tag{17}$$

Unfortunately, the optimization function $E[-\log(\cdot)]$ is non-convex. Because the $-\log(\cdot)$ is a convex function, there has $E[-\log(\cdot)] \ge -\log(E[\cdot])$ according to Jensen's inequality. Then the approximate function can be obtained $E[-\log(\cdot)] \approx -\log(E[\cdot])$. The approximate channel capacity of the secondary system is

$$\begin{aligned}
& E\left[C_{\mathrm{UL}}+C_{\mathrm{DL}}\right]\\
& \approx \min\left\{-\sum_{f=1}^{M} B\log_2\left(1+\frac{P_{s_{\mathrm{UL}}}(t,f)\,E\left[\left|H_{s_{\mathrm{UL}}}(-\tau_{s_1},f)\right|^2\right]}{N_0+I_p}\right)\right\}\\
& \quad +\min\left\{-\sum_{f=1}^{M} B\log_2\left(1+\frac{P_{s_{\mathrm{DL}}}(t,f)\,E\left[\left|H_{s_{\mathrm{DL}}}(-\tau_{s_2},f)\right|^2\right]}{N_0+I_p}\right)\right\},\\
\text{s.t.}\quad & P_{s_{\mathrm{UL}}} \le \frac{\phi}{\rho_{I_{\mathrm{UL}}}^2(t)\,H_I(-\tau_1,f)+\left(1-\rho_{I_{\mathrm{UL}}}^2(t)\right)2\sigma^2},\\
& P_{s_{\mathrm{DL}}} \le \frac{\phi}{\rho_{I_{\mathrm{DL}}}^2(t)\,H_I(-\tau_2,f)+\left(1-\rho_{I_{\mathrm{DL}}}^2(t)\right)2\sigma^2},\\
& \sum_{f=1}^{M} P_{s_{\mathrm{UL}}}(t,f) \le \Phi, \qquad -P_{s_{\mathrm{UL}}}(t,f) \le 0,\\
& \sum_{f=1}^{M} P_{s_{\mathrm{DL}}}(t,f) \le \Phi, \qquad -P_{s_{\mathrm{DL}}}(t,f) \le 0.
\end{aligned} \tag{F-3}$$

The solutions are

$$P_{s_{\mathrm{UL}}} = \begin{cases} P_{C1}, & \dfrac{B}{\mu\ln 2} \ge P_{C1}+P_{C2},\\ \dfrac{B}{\mu\ln 2}-P_{C2}, & P_{C2} < \dfrac{B}{\mu\ln 2} < P_{C1}+P_{C2},\\ 0, & \dfrac{B}{\mu\ln 2} \le P_{C2}, \end{cases}$$

$$P_{s_{\mathrm{DL}}} = \begin{cases} P_{C3}, & \dfrac{B}{\mu\ln 2} \ge P_{C3}+P_{C4},\\ \dfrac{B}{\mu\ln 2}-P_{C4}, & P_{C4} < \dfrac{B}{\mu\ln 2} < P_{C3}+P_{C4},\\ 0, & \dfrac{B}{\mu\ln 2} \le P_{C4}, \end{cases} \tag{18}$$

where $P_{C1} = \phi/(\rho_{I_{\mathrm{UL}}}^2(t)H_I(-\tau_1,f)+(1-\rho_{I_{\mathrm{UL}}}^2(t))2\sigma^2)$, $P_{C2} = (N_0+I_p)/(\rho_{s_{\mathrm{UL}}}^2(t)H_s(-\tau_{s_1},f)+(1-\rho_{s_{\mathrm{UL}}}^2(t))2\sigma^2)$, $P_{C3} = \phi/(\rho_{I_{\mathrm{DL}}}^2(t)H_I(-\tau_2,f)+(1-\rho_{I_{\mathrm{DL}}}^2(t))2\sigma^2)$, and $P_{C4} = (N_0+I_p)/(\rho_{s_{\mathrm{DL}}}^2(t)H_s(-\tau_{s_2},f)+(1-\rho_{s_{\mathrm{DL}}}^2(t))2\sigma^2)$.

3.4. Interference Control with the Channel Estimation Error. Since the channel estimation error of time-varying channel cannot be neglected, the estimated CSI between the secondary system and the primary system is imperfect. If the transmission power allocation scheme for the secondary system relies on the imperfect channel estimation result, the interference constraint will be

$$I(t,f) = P_s(t,f)\left|\widehat{H}_I(t,f)\right|^2 < \phi. \tag{19}$$

Then, the optimized transmission power of the secondary system (P_s^*) with the channel estimation error may lead to the interference problem for primary system. The variance value of channel estimation error is σ_e^2.

The actual interference power is

$$I(t,f) = P_s^*(t,f)\left\{\left|\widehat{H}_I(t,f)\right|^2+\sigma_e^2\right\}. \tag{F-4}$$

The interference power may exceed the interference threshold with the imperfect channel estimation result.

The estimation error is modeled as the spherical region $\|e_I\| < \sigma_e$, where the σ_e is the maximum radius of error bound [14]. The channel model is described by $H_{I_{\mathrm{UL}}}(t,f) = \widehat{H}_{I_{\mathrm{UL}}} + e_{I_{\mathrm{UL}}}$ and $H_{I_{\mathrm{DL}}}(t,f) = \widehat{H}_{I_{\mathrm{DL}}} + e_{I_{\mathrm{DL}}}$, where $\widehat{H}_{I_{\mathrm{UL}}}$ and $\widehat{H}_{I_{\mathrm{DL}}}$ are the estimated values of uplink and downlink channels, and $e_{I_{\mathrm{UL}}}$ and $e_{I_{\mathrm{DL}}}$ are the channel estimation errors. The variance value of the channel estimation error in uplink is assumed to be equal to the variance value of downlink channel estimation error, which can be uniformly denoted by σ_e^2.

The mean interference power constraint with imperfect CSI knowledge can be expressed as

$$\begin{aligned}
E\left[I(t,f)\,|_{H_{I_{\mathrm{UL}}}}\right] &= P_s(t,f)\,E\left[\left|H_{I_{\mathrm{UL}}}(t,f)\right|^2\right]\\
&= P_s(t,f)\,E\left[\left|\widehat{H}_{I_{\mathrm{UL}}}(t,f)\right|^2+\sigma_e^2\right]\\
&= P_s(t,f)\left\{\rho_{I_{\mathrm{UL}}}^2(t)\left|\widehat{H}_{I_{\mathrm{UL}}}(-\tau_1,f)\right|^2\right.\\
&\qquad \left.+\left(1-\rho_{I_{\mathrm{UL}}}^2(t)\right)2\sigma^2+\sigma_e^2\right\}\\
&\le \phi,
\end{aligned}$$

$$E\left[I(t,f)\,|_{H_{I_{\mathrm{DL}}}}\right] = P_s(t,f)\left\{\rho^2_{I_{\mathrm{DL}}}(t)\left|\widehat{H}_{I_{\mathrm{DL}}}(-\tau_2,f)\right|^2 + \left(1-\rho^2_{I_{\mathrm{DL}}}(t)\right)2\sigma^2+\sigma_e^2\right\} \le \phi. \tag{20}$$

The mean capacity of the SU can be described as

$$\begin{aligned} &E\left\{C_{\mathrm{UL}}(t)+C_{\mathrm{DL}}(t)\right\} \\ &\approx \min\left\{-B\sum_{f=1}^{M}\log_2\left(1+\frac{P_{s_{\mathrm{UL}}}(t,f)\,E\left[\left|\widehat{H}_{s_{\mathrm{UL}}}(-\tau_{s_1},f)\right|^2\right]}{N_0+I_p}\right)\right\} \\ &\quad+\min\left\{-B\sum_{f=1}^{M}\log_2\left(1+\frac{P_{s_{\mathrm{DL}}}(t,f)\,E\left[\left|\widehat{H}_{s_{\mathrm{DL}}}(-\tau_{s_2},f)\right|^2\right]}{N_0+I_p}\right)\right\}, \end{aligned} \tag{21}$$

$$\begin{aligned} \text{s.t.}\quad & P_{s_{\mathrm{UL}}}\left\{\rho^2_{I_{\mathrm{UL}}}(t)\left|\widehat{H}_{I_{\mathrm{UL}}}(-\tau_1,f)\right|^2+\left(1-\rho^2_{I_{\mathrm{UL}}}(t)\right)2\sigma^2+\sigma_e^2\right\}\le\phi, \\ & P_{s_{\mathrm{DL}}}\left\{\rho^2_{I_{\mathrm{DL}}}(t)\left|\widehat{H}_{I_{\mathrm{DL}}}(-\tau_2,f)\right|^2+\left(1-\rho^2_{I_{\mathrm{DL}}}(t)\right)2\sigma^2+\sigma_e^2\right\}\le\phi, \\ & \sum_{f=1}^{M}P_{s_{\mathrm{UL}}}(t,f)\le\Phi, \qquad -P_{s_{\mathrm{UL}}}(t,f)\le 0, \\ & \sum_{f=1}^{M}P_{s_{\mathrm{DL}}}(t,f)\le\Phi, \qquad -P_{s_{\mathrm{DL}}}(t,f)\le 0. \end{aligned} \tag{F-5}$$

The solution is

$$P_{s_{\mathrm{UL}}}(t,f)=\begin{cases} P_{D1}, & \dfrac{B}{\mu\ln 2}\ge P_{D1}+P_{D2}, \\ \dfrac{B}{\mu\ln 2}-P_{D2}, & P_{D2}<\dfrac{B}{\mu\ln 2}<P_{D1}+P_{D2}, \\ 0, & \dfrac{B}{\mu\ln 2}\le P_{D2}, \end{cases}$$

$$P_{s_{\mathrm{DL}}}(t,f)=\begin{cases} P_{D3}, & \dfrac{B}{\mu\ln 2}\ge P_{D3}+P_{D4}, \\ \dfrac{B}{\mu\ln 2}-P_{D4}, & P_{D2}<\dfrac{B}{\mu\ln 2}<P_{D3}+P_{D4}, \\ 0, & \dfrac{B}{\mu\ln 2}\le P_{D4}, \end{cases} \tag{22}$$

where $P_{D1}=\phi/(\rho^2_{I_{\mathrm{UL}}}(t)|\widehat{H}_{I_{\mathrm{UL}}}(-\tau_1,f)|^2+(1-\rho^2_{I_{\mathrm{UL}}}(t))2\sigma^2+\sigma_e^2)$, $P_{D2}=(N_0+I_p)/(\rho^2_{I_{\mathrm{UL}}}(t)|\widehat{H}_{s_{\mathrm{UL}}}(-\tau_{s_1},f)|^2+(1-\rho^2_{s_{\mathrm{UL}}}(t))2\sigma^2+\sigma_e^2)$, $P_{D3}=\phi/(\rho^2_{I_{\mathrm{DL}}}(t)|\widehat{H}_{I_{\mathrm{DL}}}(-\tau_2,f)|^2+(1-\rho^2_{I_{\mathrm{DL}}}(t))2\sigma^2+\sigma_e^2)$, $P_{D4}=(N_0+I_p)/(\rho^2_{I_{\mathrm{DL}}}(t)|\widehat{H}_{s_{\mathrm{DL}}}(-\tau_{s_2},f)|^2+(1-\rho^2_{s_{\mathrm{DL}}}(t))2\sigma^2+\sigma_e^2)$, and μ is the Lagrange multiplier.

4. Simulation Results

In this section, we compare the interference control performance using different interference constraints. The primary system and secondary system have 512 subcarriers which are divided into 16 subbands. The carrier frequency is 2 GHz, and the subcarrier spacing is 15 kHz. Each OFDM symbol of duration is 71.355 μs. We set the time delay $\tau_{s_1}=0$, $\tau_{s_2}=0$, $\tau_1=0$, and $\tau_2=0$.

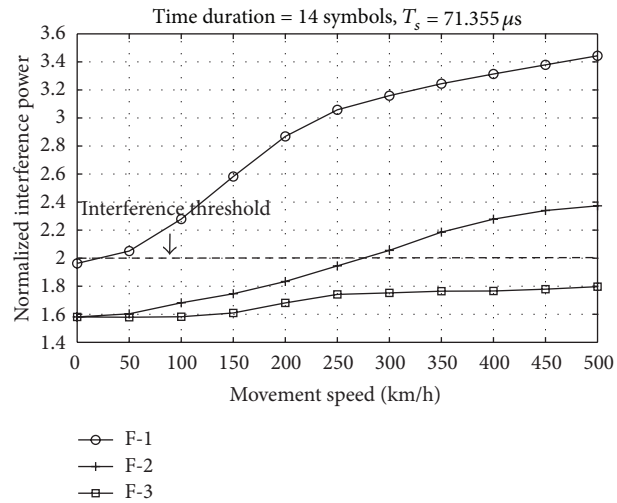


Figure 5: The mean interference power with different movement speeds.

The time-varying fading channel has 16 uncorrelated Rayleigh fading taps; each tap has 3 dB decay factor. The time-varying fading channel is described by Clark-Jakes' model. The noise power is $N_0=1$. We set $\sigma^2=0.5$, and the maximum transmission power with secondary system is $\Phi=64$. The predefined interference threshold is $\phi=2$, $I_p=10$. The variance of channel estimation error is $\sigma_e^2=0.1$.

We compare the performance of the different IC schemes in mobile environment. The mean interference power is shown in Figure 5. Since the traditional IC (F-1) is designed based on the assumption of quasistatic channel, the interference constraints are only validated in motionless or slow movement condition (less than 30 km/h). As the increased movement speed, the interference power will exceed the interference threshold rapidly. The short frame structure (F-2) can guarantee the interference power under the interference threshold under 250 km/h, but the performance of IC becomes worse rapidly when speed approaches 500 km/h. The (F-3) curve shows the IC performance of the mean interference power constraint in time-varying fading channel; the secondary system limits their transmission power in uplink and downlink. The higher movement speed will lead to worse IC performance. Comparing the three different constraints, the secondary system should adopt the mean interference constraint in high mobility environment to satisfy the QoS demand primary system.

The interference power changes from symbol to symbol in the time-varying fading channel at 300 km/h as shown in Figure 6. The channel correlationship can be described by Clark-Jakes' model which follows the 0th order Bessel function J_0; the channel correlation coefficient ρ is a variable of time which has been shown in Figure 3. Since the proposed

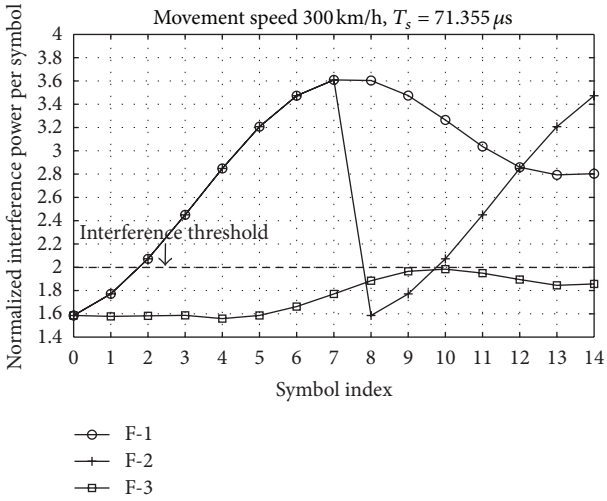


Figure 6: The interference power per symbol at primary receiver.

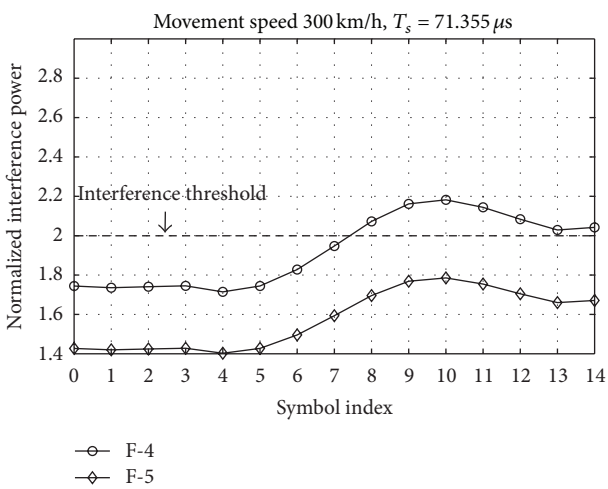


Figure 7: Mean interference power at the primary receiver.

interference constraints are impacted by ρ^2, the performance of the interference constraint is related to the change of the channel correlation coefficient. The value of ρ^2 will decrease from 1 to 0 in the first 9 symbols and then increase from the 10th symbol. Therefore, as shown in Figure 6, the trend of the interference power has the convex and wavy features. The traditional IC constraint with quasistatic assumption (F-1) cannot reflect the variation of the time-varying channel. The interference constraint cannot ensure the interference power under threshold after 2 symbols, and the maximum interference power is above 3.6. As the curve (F-2) is shown, the interference power seriously decreases at the 8th symbol which means that the short frame structure can lead to more frequently IC than standard frame length. The IC performance with short frame structure is better than traditional way to constrain the interference power. However, the communication overhead of short frame IC is twice as much as standard frame structure, which spends more time on channel estimation, thus wasting the opportunity of data transmission. When the channel state changes fast, it is impossible to control the interference by shorting the frame length unlimitedly. To cope with the IC problem in high mobility environment, the IC with mean interference constraint can be seen as a more effective way to control the interference power. The performance of IC with mean interference power constraint (F-3) shows that it perfectly limits the interference power under the interference threshold.

Considering the influence of the channel estimation error, the performance of different IC schemes at 300 km/h is shown in Figure 7. Because of the IC relies on the accuracy CSI, the influence of estimation error with CSI cannot be ignored. The actual interference power may exceed the interference threshold with the error-existed CSI. It is observed that the IC scheme which considers the mean interference constraint with imperfect channel estimation (F-4) may become invalid after several symbols. From the results shown in Figure 7, it can be found that constraining the interference power with considering the maximum radius of error bound is a conservative IC scheme. The interference power can be limited under the interference threshold within error-existed CSI. The results show that our proposed scheme is robust when the channel estimation error exists.

5. Conclusion

In underlay CRNs, limiting the interference power to the primary system is an important topic for the secondary system optimization. The existing IC schemes were based on the quasistatic channel assumption, but the variation of channel in a frame duration cannot be neglected in the high mobility environment. In this paper, we have developed the IC schemes in mobile environment considering the influence of the time-varying fading channel and the channel estimation error. The proposed approaches can limit the interference power under the predefined interference threshold in the mobile environment. Future directions include the IC schemes with the assumption of nonsynchronous transmission process between the primary system and the secondary system and the multiple antennas IC schemes in mobile environment.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities under Grant no. 2012YJS017, Key Project of State Key Laboratory of Rail Traffic Control and Safety under Grant no. RCS2012ZZ004, Key Grant Project of Chinese Ministry of Education no. 313006, and Program for New Century Excellent Talents in University under Grant NCET-09-0206.

References

[1] FCC, *Spectrum Policy Task Force*, Proceedings of the Federal Communications Commission, Washington, DC, USA, 2002.

[2] J. Mitola, "Cognitive radio architecture evolution," *Proceedings of the IEEE*, vol. 97, no. 4, pp. 626–641, 2009.

[3] O. Simeone, Y. Bar-Ness, and U. Spagnolini, "Stable throughput of cognitive radios with and without relaying capability," *IEEE Transactions on Communications*, vol. 55, no. 12, pp. 2351–2360, 2007.

[4] J. Mietzner, L. Lampe, and R. Schober, "Distributed transmit power allocation for multihop cognitive-radio systems," *IEEE Transactions on Wireless Communications*, vol. 8, no. 10, pp. 5187–5201, 2009.

[5] M. Gastpar, "On capacity under receive and spatial spectrum-sharing constraints," *IEEE Transactions on Information Theory*, vol. 53, no. 2, pp. 471–487, 2007.

[6] A. Goldsmith, S. A. Jafar, I. Maric, and S. Srinivasa, "Breaking spectrum gridlock with cognitive radios: an information theoretic perspective," *Proceedings of the IEEE*, vol. 97, no. 5, pp. 894–914, 2009.

[7] H. Yao, Z. Zhou, H. Liu, and L. Zhang, "Optimal power allocation in joint spectrum underlay and overlay cognitive radio networks," in *Proceedings of the 4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM '09)*, pp. 1–5, June 2009.

[8] L. Le and E. Hossain, "Resource allocation for spectrum underlay in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5306–5315, 2008.

[9] M. Z. Win, P. C. Pinto, and L. A. Shepp, "A mathematical theory of network interference and its applications," *Proceedings of the IEEE*, vol. 97, no. 2, pp. 205–230, 2009.

[10] A. Rabbachin, T. Q. S. Quek, H. Shin, and M. Z. Win, "Cognitive network interference," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 2, pp. 480–493, 2011.

[11] K. Guan, Z. Zhong, and B. Ai, "Assessment of LTE-R using high speed railway channel model," in *Proceedings of the 3rd IEEE International Conference on Communications and Mobile Computing (CMC '11)*, pp. 461–464, April 2011.

[12] W. Qiu, B. Xie, H. Minn, and C. Chong, "Interference-controlled transmission schemes for cognitive radio in frequency-selective time-varying fading channels," *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 142–153, 2012.

[13] M. Patzold, *Mobile Fading Channels*, John Wiley & Sons, New York, NY, USA, 2002.

[14] M. B. Shenouda and T. N. Davidson, "Convex conic formulations of robust downlink precoder designs with quality of service constraints," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 714–724, 2007.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 986985, 11 pages
http://dx.doi.org/10.1155/2013/986985

Research Article

A Rough Penalty Genetic Algorithm for Multicast Routing in Mobile Ad Hoc Networks

Chih-Hao Lin and Chia-Chun Chuang

Department of Information Management, Chung Yuan Christian University, Jhongli City 32023, Taiwan

Correspondence should be addressed to Chih-Hao Lin; linch@cycu.edu.tw

Received 26 April 2013; Accepted 13 July 2013

Academic Editor: Anyi Chen

Multicast routing is an effective way to transmit messages to multiple hosts in a network. However, it is vulnerable to intermittent connectivity property in mobile ad hoc network (MANET) especially for multimedia applications, which have some quality of service (QoS) requirements. The goal of QoS provisioning is to well organize network resources to satisfy the QoS requirement and achieve good network delivery services. However, there remains a challenge to provide QoS solutions and maintain end-to-end QoS with user mobility. In this paper, a novel penalty adjustment method based on the rough set theory is proposed to deal with path-delay constraints for multicast routing problems in MANETs. We formulate the problem as a constrained optimization problem, where the objective function is to minimize the total cost of the multicast tree subject to QoS constraints. The RPGA is evaluated on three multicast scenarios and compared with two state-of-the-art methods in terms of cost, success rate, and time complexity. The performance analyses show that this approach is a self-adaptive method for penalty adjustment. Remarkably, the method can address a variety of constrained multicast routing problems even though the initial routes do not satisfy all QoS requirements.

1. Introduction

Multicasting is a service method in which a source node can deliver copies of messages to multiple recipients at different locations in a communication network. Multicasting techniques play a critical role in many applications such as video conference, internet games, and web-based learning. In this paper, multicast routing problems mainly focus on finding a minimum Steiner tree and satisfying quality-of-service (QoS) requirements. Unfortunately, the problem of finding a Steiner tree is known to be a NP-complete problem [1], even if links have unit costs.

The multicast tree in mobile ad hoc networks (MANETs) is vulnerable to intermittent connectivity property during the transmission period [2]. Due to the nodal mobility and the dynamic topology of mobile networks, a service provider should find a cost-effective tree for its multicast customers in real time and assure certain QoS requirements [3]. Without the support of communication infrastructure, the challenge of the dynamic routing on a changing topology is how to decide a multicast tree as soon as possible. To enhance the effectiveness and efficiency, this work uses a routing representation scheme to encode the multicast tree. Therefore, this paper models the multicast routing problem with QoS constraints to support multimedia transmission in MANETs [4, 5].

The battery limitation of a mobile node is a critical constraint while developing multicast routing protocols. Genetic algorithm (GA) presents a potential solution for the multiconstrained multicast routing problem [6]. Traditionally, penalty-function methods are the most popular constraint-handling techniques. According to the degree of constraint violation, penalty coefficient should be determined carefully [7]. The static-penalty (SP) method applies a static coefficient for each constraint and then adjusts penalty coefficient manually [8]. To adjust penalty coefficient automatically, the dynamic-penalty (DP) method combines the generation number and a scaling constant to adjust coefficient automatically [9]. Furthermore, the adaptive-penalties (AP) method tries to avoid infeasible solutions by adjusting the penalty coefficient according to the convergent situation [10].

Different from several related researches [2, 4, 5], this paper synthesizes the rough set theory (RST) and penalized techniques as a rough penalty genetic algorithm (RPGA).

The emphasis of the proposed RPGA uses a rough-penalty (RP) method to releases/enforces some penalties on inefficient/efficient constraints during evolution. To facilitate the effectiveness of multiple penalties, the RPGA incorporates with a therapeutic crossover to enlarge the genetic diversity in a population and guide to find the optimal solution. The performance of the RPGA is evaluated by three kinds of constrained multicast routing scenarios. Experimental results show that the proposed RPGA not only can find near-optimal solutions but also can obtain robust feasible results for QoS-based multicast routing problems.

The rest of paper is organized as follows. Section 2 models the multicast routing problems with QoS constraints in MANETs. In Section 3, the operations of the proposed RPGA are described in detail. Section 4 introduces the proposed RP method. Section 5 reports the experimental results, algorithm analyses, and performance comparisons for three test networks. Finally, the paper is summarized in Section 6.

2. Problem Description and Formulation

2.1. Network Modeling. At a certain time period in a MANET, we assume that the service provider knows (1) traffic load from node i to node j ($i \neq j$), (2) QoS requirements/constraints (e.g., delay constraint), (3) bandwidth available of each link, (4) link cost for the traffic to pass through each link, and (5) time delay for the traffic to pass through each OD pair. We consider the multicast routing problem with bandwidth and delay constraints. The communication network is modeled as a connected weighted, directed graph $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_n\}$ is a finite set of network nodes and $E = \{e_1, e_2, \ldots, e_m\}$ is the set of network links. The link $e = (u, v) \in E$ connects node $u \in V$ to node $v \in V$ with positive cost function (i.e., $\text{cost}(e) : E \rightarrow R^+$), available bandwidth (i.e., $\text{BW}(e) : E \rightarrow R^+$), and delay function (i.e., $\text{delay}(e) : E \rightarrow R^+$). The number of nodes and links (i.e., the cardinalities of V and E) is n and m, respectively. For each multicast session, messages are routed from a source node s to a set of multicast destination group $D = \{d_1, d_2, \ldots, d_k\} \subseteq V$. A multicast tree $T(s, D) = (V_T, E_T)$ represents a solution to the multicast routing problem, where $V_T \subseteq V$ and $E_T \subseteq E$. Thus, this tree $T(s, D)$ is a subgraph of G with root s and a set of nodes D. Let $P_T(s, d)$ represent a path in the tree, $T(s, D)$, from source node s to a destination node $d \in D - \{s\}$. We have the following definitions.

Definition 1. The cost of multicast tree $T(s, D)$ is the sum of the links' cost in the tree. The link cost may represent its monetary cost or resource utilization:

$$\text{cost}(T) = \sum_{e \in E_T} \text{cost}(e). \tag{1}$$

Definition 2. The bottleneck bandwidth of path $P_T(s, d)$ is the minimum value of links' bandwidth along the path, which represents the residual bandwidth of a communication path:

$$\begin{gathered} \text{BW}\left(P_T(s,d)\right) = \min\left(\text{BW}(e) \mid \forall e \in P_T(s,d)\right), \\ \forall P_T(s,d) \subset T, \quad \forall d \in D. \end{gathered} \tag{2}$$

Definition 3. The delay of path $P_T(s, d)$ is the sum of links' delay along the path from s to d. The link delay may include its nodal processing, queueing, transmission, and propagation delays:

$$\begin{gathered} \text{delay}\left(P_T(s,d)\right) = \sum_{e \in P_T(s,d)} \text{delay}(e), \\ \forall P_T(s,d) \subset T, \quad \forall d \in D. \end{gathered} \tag{3}$$

Definition 4. The delay of multicast tree T is the maximum value of paths' delay in the tree, that is,

$$\begin{aligned} &\text{delay}(T) \\ &\quad = \max\left(\text{delay}\left(P_T(s,d)\right) \mid \forall P_T(s,d) \subset T, \forall d \in D\right). \end{aligned} \tag{4}$$

Figure 1 depicts an example of a network graph, link parameters, and a multicast tree. Parameters along links are triple (cost, delay, bandwidth). In Figure 1, the source node is Node 1 (i.e., $s = 1$). The destination nodes are nodes 3, 7, and 8 (i.e., $D = \{3, 7, 8\}$). The Steiner tree $T(s, D)$ consists of three paths $P_T(1, 3) = \{(1, 4), (4, 5), (5, 3)\}$, $P_T(1, 7) = \{(1, 4), (4, 5), (5, 7)\}$, and $P_T(1, 8) = \{(1, 4), (4, 5), (5, 8)\}$. The total cost of the Steiner tree can be calculated by (1), that is, $\text{cost}(T) = 20$. By using (2), we can calculate the bottleneck bandwidth of each path along the Steiner tree, that is, $\text{BW}(P_T(1, 3)) = 8$, $\text{BW}(P_T(1, 7)) = 8$, and $\text{BW}(P_T(1, 8)) = 7$. We can also apply (3) to calculate $\text{delay}(P_T(1, 3)) = 9$, $\text{delay}(P_T(1, 7)) = 8$, and $\text{delay}(P_T(1, 7)) = 10$. Finally, we can derive the total delay of the Steiner tree $\text{delay}(T) = 9$ by (4).

2.2. Problem Definition. The optimal multicast tree T depends on the operator objectives, such as network cost, transmission delay, or target QoS [11]. Therefore, solving the multicast routing problem is equivalent to find the optimal distribution tree on the basis of a certain cost function under a given set of constraints. In this paper, the multicast routing in MANETs can be modeled as a combinatorial optimization problem in the following:

$$\begin{aligned} &\text{minimize} \quad \text{cost}(T(s,D)) \\ &\text{subject to:} \quad \text{delay}\left(P_T(s,d)\right) \leq \Delta_d \quad \forall d \in D, \\ &\qquad\qquad\quad \text{BW}\left(P_T(s,d)\right) \geq B_d \quad \forall d \in D. \end{aligned} \tag{5}$$

The objective function is to minimize the total cost of the multicast tree. The path-delay constraint enforces that the total delay of each OD pair must be smaller than or equal to its delay bound Δ_d. The minimum bandwidth requirement is denoted as B_d. Therefore, the multicast routing problem is to determine a multicast tree connecting the source node to every destination node such that the cost of this tree is minimum, while the path-delay and bottleneck-bandwidth from the source node to any destination node satisfy the prescribed QoS requirements.

2.3. Routing Table. Since there are so many candidate paths between two nodes in the network graph $G = (V, E)$,

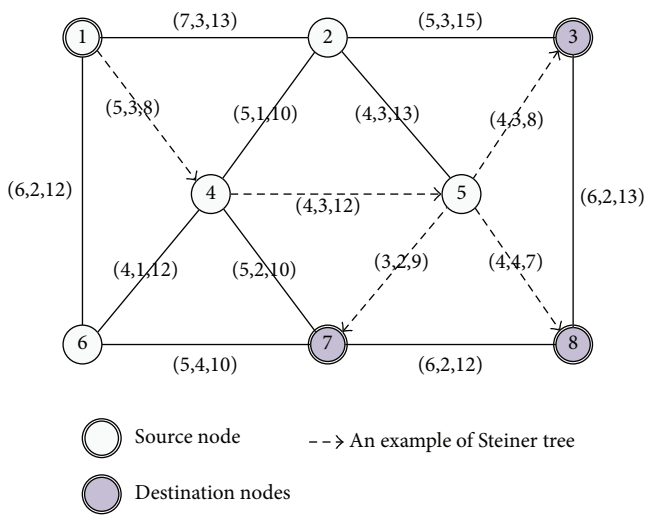


Figure 1: An example of a network graph, link parameters, and a Steiner tree.

Table 1: An example of routing table for OD pair (1, 7) in Figure 1.

Routing table of path from node 1 to node 7				
Route no.	Route path	Cost of the path	Delay along the path	Bottleneck bandwidth
0	1-4-7	10	5	8
1	1-6-7	11	6	10
2	1-4-5-7	12	8	8
⋮	⋮	⋮	⋮	⋮
$R-1$	1-2-3-8-7	24	10	12

traditional GAs may consume considerable computational effort in searching infeasible solutions because genetic operations do not always preserve feasibility. Therefore, to reduce the search space, this work uses the K shortest path routing algorithm to precalculate the first R shortest paths and record in a routing table. For the network topology in Figure 1, an example of its routing table for OD pair (1,7) will look like Table 1, which includes the first R shortest paths from node 1 to node 7 with the route path, total cost, aggregate delay, and bottleneck bandwidth.

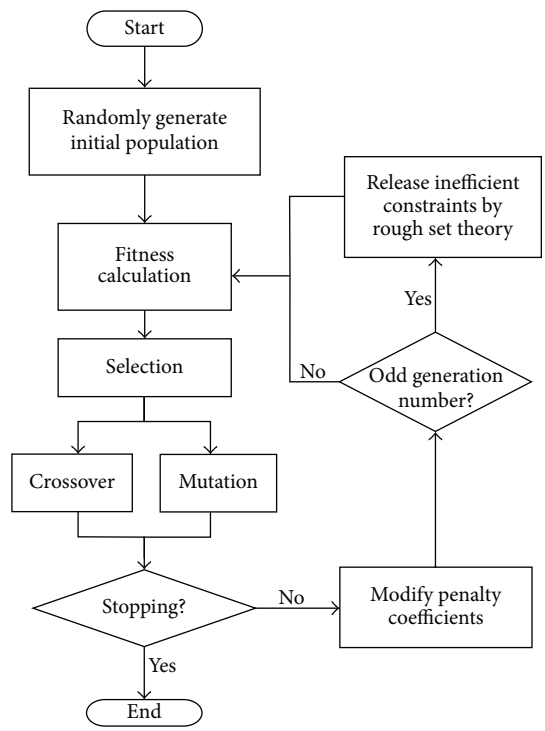


Figure 2: Flow chart of the RPGA.

3. RPGA for Multicasting Routing Problem

The proposed RPGA adopts a RP method to enhance the searching abilities of original GAs for handling constrained multicasting routing problems. To enhance the exploration ability, the RPGA adopts the RST to enlarge the genetic diversity by releasing inefficient constraints and also enforcing efficient ones when the generation number is odd. The flow chart of RPGA (in Figure 2) consists of several genetic operations, such as initialization, selection, crossover, mutation, replacement, and RP method. To enhance the exploitation ability, the proposed RPGA applies the therapeutic crossover to improve the convergence rate during the evolution.

3.1. Encoding and Initialization. In this paper, the encoding method is based on a routing representation for multicast trees. The RPGA maintains a population of chromosomes, which represent a candidate set of Steiner trees for the multicast routing problem. Given a source node s and a set of destination nodes $D = \{d_1, d_2, \ldots, d_k\}$, a chromosome can be represented by a string of integers with length k. The chromosome is denoted as $\vec{X} = (x_1, x_2, \ldots, x_k)$, where x_i is an integer in interval $[0, R-1]$ to represent a route number for the OD pair from s to u_i in the routing table. In the

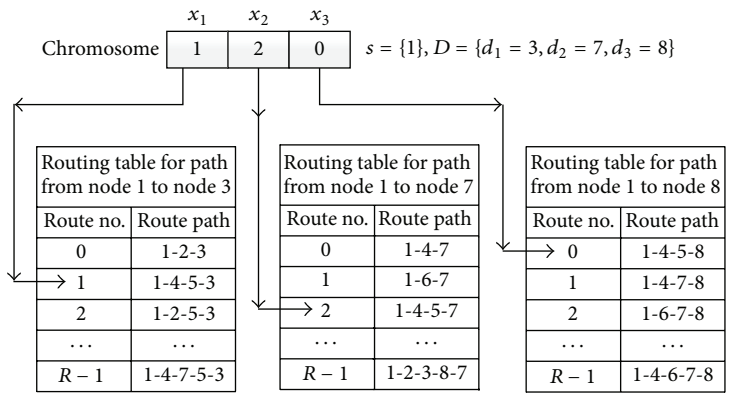


Figure 3: Representation of chromosome, genes, and routing table.

example of Figure 1, the second OD pair (i.e., source node 1 to destination node 7) is routed along Path 1-4-5-7. The route number of this path in its routing table (in Table 1) is 2. Thus, we should assign the route number as the value of the second gene, that is, $x_2 = 2$. Therefore, the relationship between chromosome, gene, and routing table for the example in Figure 1 can be illustrated as Figure 3. The RPGA starts with a random population within the gene value interval $[0, R-1]$, no matter whether these multicast trees satisfy the QoS constraints or not. The emphasis is that the RPGA can find the global optimum for constrained problems even though the initial population is infeasible.

3.2. Fitness Function with Self-Adaptive Penalty Adjustment. The fitness value of each chromosome represents the quality of the corresponding multicast tree (i.e., $\vec{X} = T(s, D)$). However, the penalty adjustment for QoS constraints is difficult to adapt suitably. In this paper, we mix the aspects of the Joines and Houck's DP method [9] with the RST to find a Steiner tree that satisfies the QoS constraints and minimizes total routing cost as well.

The proposed RP method adjusts penalty terms according to both violation magnitude and evolution time. To solve constrained optimization problems effectively, each individual in generation t is evaluated using an expanded objective function (6):

$$\psi\left(\vec{X}\right) = \operatorname{cost}\left(\vec{X}\right) + \sum_{k=1}^{m}\left((C \times t)^{\pi(k,t)} \times \max\left(0, \Phi_k\left(\vec{X}\right)\right)^2\right), \tag{6}$$

where C is a "severity" factor, m is the total number of constraints, and Φ_k is the violation magnitude of constraint k. This fitness function combines a *coefficient* $(C \times t)$ with an *exponent* $\pi(k, t)$ to increase penalty pressure over time. For constraint k in generation t, the exponent $\pi(k, t)$ is a representative penalty multiplier that is initially assigned as 2. And then, the penalty multiplier is tuned iteratively according to the discernible mask $\vec{\mu}$ and the representative attribute value γ_k of superior class X_{good}. The exponent $\pi(k, t)$ is defined as

$$\begin{gathered}\pi(k,t) = \begin{cases}\pi(k,t-1) \times \gamma_k, & \text{if } \mu_k = 1,\\ \pi(k,t-1), & \text{if } \mu_k = 0,\end{cases}\\ \forall k = 1,\ldots,m; \quad \forall t = 1,\ldots,\text{MaxGeneration},\\ \pi(k,0) = 2 \quad \forall k = 1,\ldots,m.\end{gathered} \tag{7}$$

Remarkably, the discernible mask $\vec{\mu}$ can be used to enable $\pi(k, t)$ by differentiating significant characteristics between classes X_{good} and X_{bad}. If the kth constraint is discernible (i.e., $\mu_k = 1$), the exponent $\pi(k, t)$ is adjusted by the representative attribute value (γ_k); otherwise, the exponent retains the same value as in the previous generation. All these RP coefficients will be introduced in Section 4.

3.3. Selection Operation. A selection operation uses fitness to determine the solution quality and to select high-quality chromosomes for the recombination operation [12]. The RPGA employs a stochastic universal selection to create selection pressure towards the global optimal solution. The measurement of a chromosome's fitness is its value of the expanded objective function in (6).

3.4. Crossover Operation. The crossover operation represents the mixing of genetic material from two selected parents to produce one child chromosome. The RPGA proposes a therapeutic crossover that incorporates a gene-therapy method with a conventional crossover scheme to enhance the exploitation ability and speed up the convergence rate [13].

Each time the selection operation chooses two crossover parents from the population. The therapeutic crossover gives each gene locus an equal chance of being a crossover point (i.e., belongs to a therapeutic genome $i \in G_c$ where $G_c \in \{1, 2, \ldots, k\}$). The proposed gene-therapy method evaluates the merit of two selected genomes by comparing the changes in the chromosome fitness before and after interchanging the genomes with the other mating chromosome [13].

```
user is Number of OD pair;
Node is all nodes in the network;
Parents is the chromosomes which be selected;
nKids is the numbers of chromosomes create by crossover;
r1, r2 is the chromosomes which are crossover;
coeC is the number of gene will be crossover;
thisScore is the multi path score;
thisGeneScore is the path score in the multi path;
xoverPoint is genes in the chromosome which be select to crossover;
lengthXP is length of xoverPoint;

Begin
for i = 1:nKids
    Pick chromosomes r1 and r2 from parents;
    coeC is decrease by generations from 1 to 0.2;
    Randomly select xoverPoint which probability small than coeC;

    if thisScore(r1) ≤  thisScore(r2)
        xoverKids(i) = thisPopulation(r1);
        for j = 1 to lengthXP
            if thisGeneScore(r2(j)) < thisGeneScore(r1(j))
                Record path r1(j) as a;
                Record path r2(j) as b;
                if the route paths in a and b have the same intermediate node (m)
                    Interchange the sub-routes (s, m) or (m, d) in a and b;
                else
                    xoverKids(i, j) = r2(j);
                end if
            end if
        end for
    else

        xoverKids(i) = thisPopulation(r2);
        for j = 1 to lengthXP
            if thisGeneScore(r1(j)) < thisGeneScore(r2(j))
                Record path r1(j) as b;
                Record path r2(j) as a;
                if the route paths in a and b have the same intermediate node (m)
                    Interchange the sub-routes (s, m) or (m, d) in a and b;
                else
                    xoverKids(i, j) = r1(j);
                end if
            end if
        end for
    end if
end for
End
```

Pseudocode 1: Pseudocode of the crossover operation for routing problem.

According to their relative merit, these two genomes combine to generate a new genome for their offspring. Therefore, offspring inherit more genetic material from the superior genome than the inferior one. We depict the pseudocode of the therapeutic crossover in Pseudocode 1.

3.5. Mutation Operation. A mutation operation used in GAs can increase population diversity to enhance its exploration ability [13]. This work uses the bit-flip mutation with a fixed small probability p_m. According to this mutation probability p_m, the mutation operation randomly selects a subset of genes and chooses new paths from its routing table. Thus, the route numbers of these new paths in its routing table replace the original values of selected genes. The resulting chromosome is a new multicast tree and can increase population diversity.

3.6. Replacement and Termination. The proposed RPGA adopts a replacement-with-elitism method to prevent best solutions from being lost through a selection process. A successive population is produced from three sources:

(1) the replacement-with-elitism method selects the best 10% chromosomes to join the new population; (2) the crossover operation recombines 80% of child chromosomes; and (3) the mutation operation constructs other child chromosomes for the next generation. The RPGA will stop when it reaches the predefined maximum iterations.

4. RP Method for Constraint Handling

To address the multicast routing problem with QoS constraints, the challenge is how to optimize the objective function value against its constraint violations. Traditional GAs are ineffective in searching feasible solutions because genetic operations do not always preserve feasibility [14]. Therefore, penalty-function methods are the most popular constraint-handling techniques for constrained optimization [15].

The novel RP method has been proposed in our previous work for numerical constrained problems [13]. This proposed RPGA inspired by the Pawlak's RST [16, 17] has been proved better than several existing algorithms for solving a variety of numerical optimization problems [13]. The RP method can automatically adjust penalty coefficients during the evolution. Furthermore, the method does not depend on extra functional analyses for its solution space. Therefore, this study aims to effectively extend the original RPGA as a new constraint-handling technique to address the multicast routing problem with QoS constraints in MANETs. During the genetic evolution, the proposed RP method uses the attribute reduction concept to find appropriate penalizing strategies and release some inefficient constraints.

4.1. Flow of RP Method. The pseudocode of the RP method is depicted in Pseudocode 2. The proposed RP method not only penalizes constraint violations to exploit feasible space but also releases ineffective constraints to explore infeasible space. Therefore, the RP method is a self-adaptive approach that can measure infeasibility and can adjust each penalty coefficient automatically.

4.2. Rough Penalty Classification. This work uses information granulation as a key function for implementing a divide-and-conquer strategy. Elementary information granules are indiscernibility classes of constraint violations. The information system is an information table of attribute values containing rows labeled by objects and columns labeled by attributes [18].

Remark 5. A partition granularity (ρ) is defined for classifying the magnitude of constraint violations. The design principle is that solution quality increases as its constraint penalty moves closer to zero. Therefore, this study uses a smaller range in near-zero regions than in other regions.

4.3. Rough Decision System. A decision system is an IS with the form $DT = (U, A \cup \{d\}, D)$ in which each individual is treated as an object of a nonempty finite set U. Attribute set $A = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ is a nonempty finite set of conditional attributes, where each *penalty multiplier* (α_k) corresponds to a conditional attribute. The supervised knowledge is expressed by a decision attribute (denoted by $d \notin A$). An information function D maps each object to a decision attribute, that is, $D : U \rightarrow V_d$ for $d \notin A$, $V_d = \{0, 1\}$. For a minimization problem, the information function is designed as follows:

$$D\left(\vec{x}_j\right) = d_j = \begin{cases} 1, & \text{if } f\left(\vec{x}_j\right) < f_{\text{average}}, \\ 0, & \text{if } f\left(\vec{x}_j\right) \geq f_{\text{average}}, \end{cases} \tag{8}$$

where

$$f_{\text{average}} = \frac{1}{p} \sum_{j=1}^{p} f\left(\vec{x}_j\right). \tag{9}$$

Remark 6. Because the penalty multiplier should be adjusted according to the region of its constraint violation, this work enlarges the penalty multiplier when its violation level increases. In the illustration in Figure 4, the penalty multiplier will be assigned as $(\alpha)^{-2}$, $(\alpha)^{-1}$, $(\alpha)^{0}$, $(\alpha)^{+1}$, $(\alpha)^{+2}$, and $(\alpha)^{+3}$ for constraint regions 1, 2, 3, 4, 5, and 6, where α denotes the coefficient and the partition granularity (ρ) is six.

4.4. Significant Penalty and Attribute Reduction. Based on the concept of attribute reduction, attributes may not be equally important, and some of them can be eliminated from a decision table without degrading information quality. Attribute reduction can be generalized by introducing attribute evaluation, which can express the merit of each attribute in the information table [19].

Remark 7. Decision attribute d in DT determines a partition CLASS(d) of object set U, where CLASS(d) is the object classification with respect to decision attribute d. The minimal subset of penalized constraints is applied to distinguish above-average individuals (i.e., their decision attributes are "1") and below-average ones (i.e., those values are "0"). The representative value of each relevant attribute is assigned as the attribute value with the maximum cardinality in the same class.

5. Computational Experiments

5.1. Test Platform and Parameter Setting. In this paper, the proposed RPGA is evaluated by solving multicast routing problems in MANETs. We use the well-known network generation tool [20] to create an asynchronous network based on the Waxman's techniques [21] and depicted in Figure 5. The network illustrates a random graph in which 40 nodes are connected, and each node has average of four connections to other nodes (i.e., average degree of a node is 4). Each link has its own cost, delay, and bandwidth information. In order to reduce the complexity of the graph representation, Figure 5 only shows the cost/delay information along one direction link (from a smaller ID node to a larger ID one). For example, traffic along the link from node 20 to node 34 will spend 54 units cost and delay 14 msec. In all cases, the maximum number of iterations is 40, the population size is 20, and the number of elite individuals is 2. For each test problem, 30 independent runs with different seeds are performed using the MATLAB environment.

```
conN is mean constraint numbers
popnum is population numbers
IT is information table, which is (conN + 1) * popnum matrix
gp is good population which decision variable = 1 from IT
bp is bad population which decision variable = 0 from IT
gc_k is good characteristic from gp
bc_k is bad characteristic from bp
t is the number of current generation
π_k(t) is a RP exponent for the t generation
λ_k is a RP coefficient
C is constraints weight
τ is next generation penalty coefficient
Begin
        Create information table IT;
        //Divide IT to two group gp and bp
        For i = 1 to popnum
            IF decision variable = 1 Pick IT(i) to good population gp;
            Else Pick IT(i) to bad population bp;
            End if
        End for
        //Find characteristic from gp and bp
        For k = 1 to conN
            gc_k = mod all gp_k;
            bc_k = mod all bp_k;
        End for
        //Find RP coefficient μ
        For k = 1 to conN
            If gc_k = bc_k λ_k = gc_k;
            Else λ_k = 1;
            End if
        End for
        //Modify RP penalty exponent π(k, t)
        For k = 1 to conN
            π_k(t) = π_k(t − 1) * λ_k;
            τ_k = (C × t) ∧ π_k(t), for all k = 1 to conN;
        End for
End
```

Pseudocode 2: Pseudocode of the RP method.

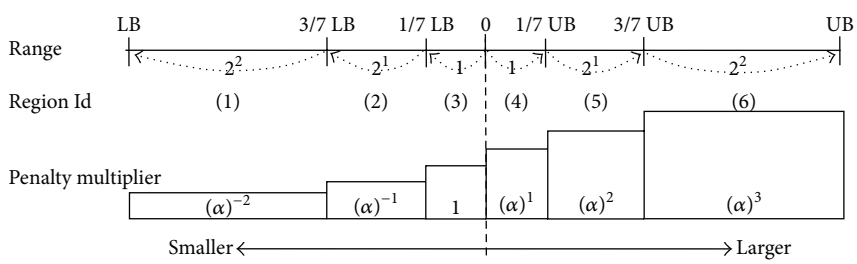


Figure 4: Penalty multiplier classification.

5.2. Performance Metrics. In this paper, the performance metrics of solution algorithms consist of (1) the total cost of the obtained multicast tree; (2) the success rate with respect to the QoS constraints; and (3) the required CPU time for computing the multicast routing problems. The success rate (θ_{req}) concerns about the percentage of feasible routes with respect to the QoS requirements. We can define the success rate as follows [22]:

$$\theta_{\text{req}} = \frac{N_{\text{ack}}}{N_{\text{req}}}, \tag{10}$$

where N_{ack} is the number of OD pairs that satisfy all QoS constraints and N_{req} is the total number of OD pairs in this multicast group.

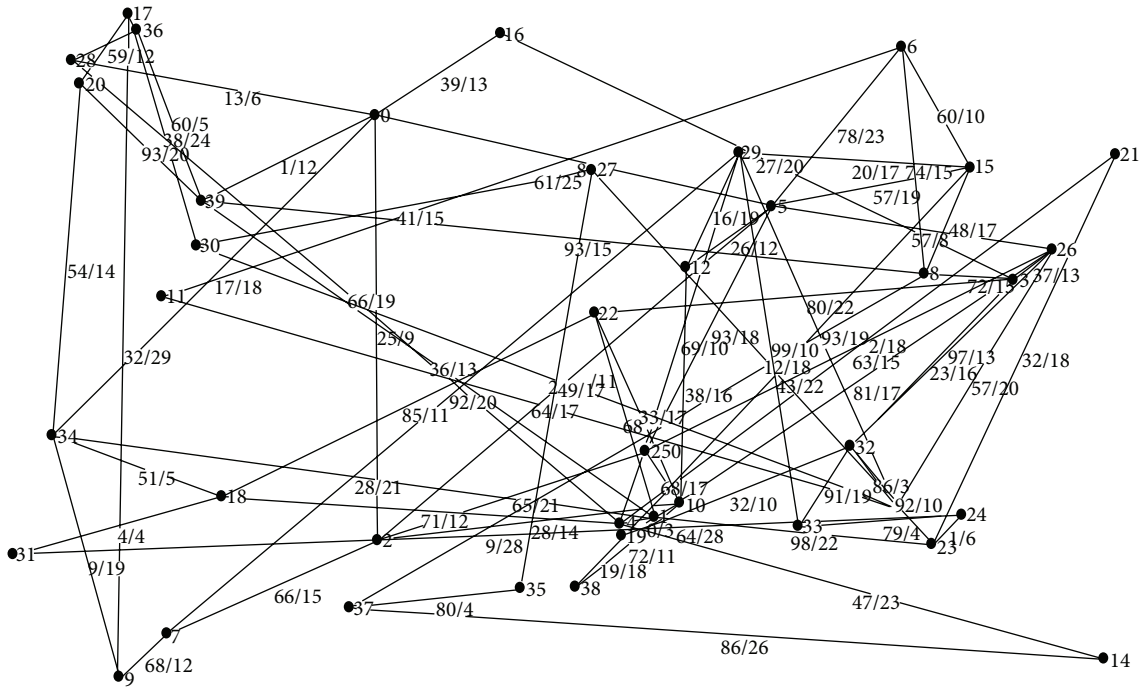


Figure 5: A randomly generated network with 40 nodes and average degree four.

Table 2: The results obtained by different partition granularities (ρ) when $\alpha = 1.01$.

Partition (ρ)	Cost	Success rate	CPU time
4	1032.9	0.966666667	1.089747617
6	1006.4	0.98245614	1.075849966
8	1021.5	0.970175439	1.085799014

5.3. Algorithm Analyses for Different Parameter Settings. It is well known that the performance of GAs significantly depends on the configuration of its operating parameters. To investigate the impact of various parameter settings in the RP method, this study experiments on two parameters: the partition granularity (ρ) and the penalty coefficient (α). The test network (in Figure 5) has 40 nodes with average degree of 4. We randomly select 20 nodes (50% of total nodes) as the test multicast group, that is, one source node and 19 destination nodes. All OD pairs have the same delay bound for 60 msec.

Firstly, this study conducts 30 runs to find the appropriate partition granularity (ρ) for this problem. When the penalty coefficient is fixed ($\alpha = 1.01$), the experimental results for different partition granularities changing from $\rho = 4$, $\rho = 6$, to $\rho = 8$ are shown in Table 2. We normalize these results relative to that of $\rho = 6$ and show the percentage comparison in Figure 6. Obviously, the RPGA with 6 partitions can achieve better results than other partition settings with respect to all the three performance metrics.

Secondly, when the partition granularity is given ($\rho = 6$), the experiments on different penalty coefficients ($\alpha = 1.0001$, 1.001, 1.01, 1.1, and 10) are tested for 30 runs. The experimental results on the average of cost, success rate and computing time are depicted in Table 3. The normalized percentages relative

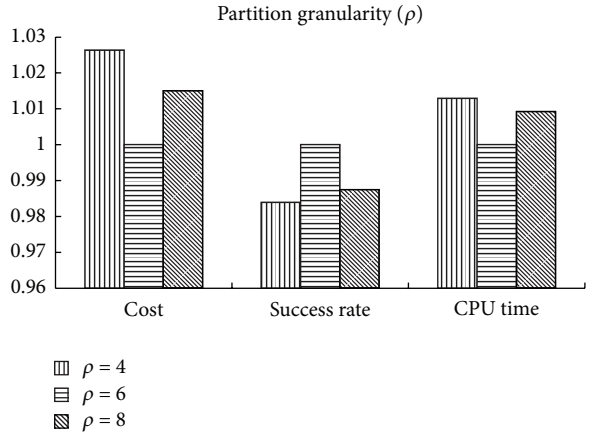


Figure 6: Comparison with different partition granularities (in percentage relative to $\rho = 6$).

Table 3: The results obtained by different penalty coefficients (α) when $\rho = 6$.

Coefficient (α)	Cost	Success rate	CPU time
10	1022.833333	0.989473684	1.084000564
1.1	1031.966667	0.978947368	1.068427877
1.01	1006.4	0.98245614	1.075849966
1.001	1031.7	0.975438596	1.065008048
1.0001	1022.333333	0.973684211	1.059503391

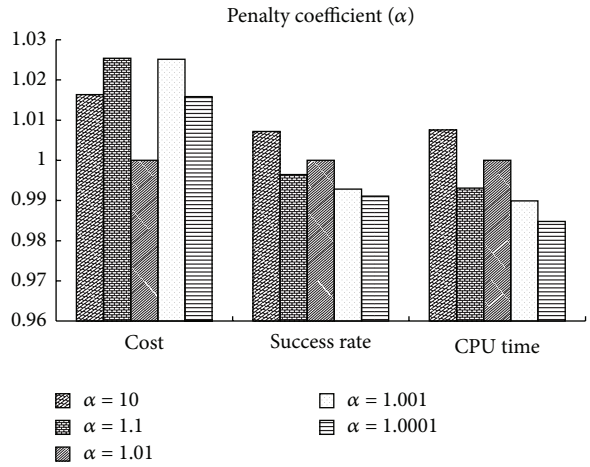


Figure 7: Comparison with different penalty coefficients (in percentage relative to $\alpha = 1.01$).

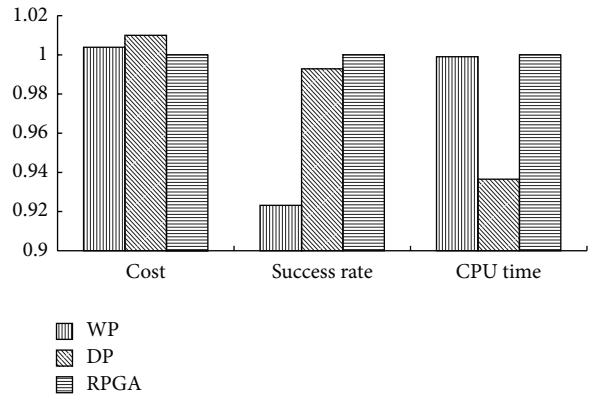


Figure 8: Comparison with different penalty methods (in percentage relative to the RPGA).

to $\alpha = 1.01$ are shown in Figure 7. Noticeably, the RPGA with $\alpha = 1.01$ can achieve better results than other settings on all the cost, success rate, and computing time metrics.

5.4. Comparison with Other Existing Methods. The performance of the proposed RPGA is compared with two well-known penalty methods, which are the Wang's penalty (WP) method [22] and the DP method [9] for three kinds of multicast scenario. In the first test scenario, the test network has 40 nodes with average degree of 4, which is the same as Figure 5. The multicast group includes 20 OD pairs. All OD pairs have the same delay bound for 60 msec. We execute each method for 30 independent runs and report the experimental results in Table 4.

For the mean cost in Table 4, the RPGA can find the best result in average 30 runs. Furthermore, the standard deviation of the routing cost obtained by the RPGA is smaller than that by the DP and similar to that of the WP. That is, the RPGA can reliably find the minimum-cost Steiner tree. Compared with the success rate, the RPGA is the best method with respect to the mean and standard deviation of the success rate, even though the WP has a high probability of converging on infeasible solutions. That is, the proposed RPGA can succeed in finding feasible and minimum-cost solutions. From the CPU-time metric, the comparison can help us to realize how much time complexity is needed to compute the RP method because the RPGA is enhanced from the DP method. In Table 4, the RPGA only spends a little computing effort (about extra 6.7% computing time) to obtain better results than the DP method.

For comparison, all the results are normalized as percentages relative to the results of the RPGA (in Figure 8). Obviously, the performance comparison shows that the RPGA can find the minimum-cost multicast tree effectively (with higher success rate) and efficiently (with lower computing effort) than the other two methods.

The evolution curves of all three methods on the total cost and path delay are shown in Figures 9(a) and 9(b), respectively. In Figure 9(a), we can observe that both the DP and RPGA methods rapidly converge on the low-cost results after 8 generations; however, the WP method takes 13 generations to slowly converge on a high-cost one. From the path delay aspect in Figure 9(b), all the three methods can satisfy the delay bound. The RPGA, DP, and WP methods take 2, 2, and 4 generations to achieve feasible solutions, respectively. Therefore, the proposed RPGA has a similar convergent effect with the DP. Both the RPGA and DP method outperform the WP with respect to the effectiveness and efficiency abilities.

In the second test scenario, we randomly generate 8 test networks with the numbers of nodes from 10 to 80 to mimic the stress test for these three methods. In those tests, all delay constraint bounds are 60 msec and the multicast group size is 50% of network nodes. When nodes number increases, the network overhead increases obviously and end-to-end delay increases at the same time. The experimental results depicted in Figure 10 are normalized as percentages relative to the result of the RPGA for comparison. Compared with Figures 10(a) and 10(b), we can observe that RPGA can find the minimum cost of feasible multicast tree, even though the success rates of the DP in the 60-node network and the WP in almost all networks are less than 90%. The WP is the worst method in these three methods on the success rate. However, the computing effort of the RPGA is higher than the other two methods in Figure 10(c). We can also find that the more number of network nodes in the test scenarios, the closer computing time needed for all three penalty methods.

In the third test scenario, we change the delay-bound requirements from 20 msec to 90 msec in a test network, which has 40 nodes and its multicast group size is 20. The comparisons between the success rates and the delay bounds are shown in Figure 11. The success rate of the RPGA is similar to that of the DP and is better than that of the WP, especially, when the delay bounds are lower than 60 msec.

6. Conclusions and Future Works

The proposed RP method cooperates with GAs for dealing with QoS-based multicast routing problems. The principle of the RP method is that the RPGA releases/enforces some

Table 4: The results of different penalty methods.

Penalty methods	Cost (mean)	Cost (st. dev.)	Success rate (mean)	Success rate (st. dev.)	CPU time
WP	1010.366667	82.23745406	0.907017544	0.077799684	1.074814229
DP	1016.5	101.1191683	0.975438596	0.033095275	1.00755095
RPGA	1006.4	83.13910238	0.98245614	0.028772225	1.075849966

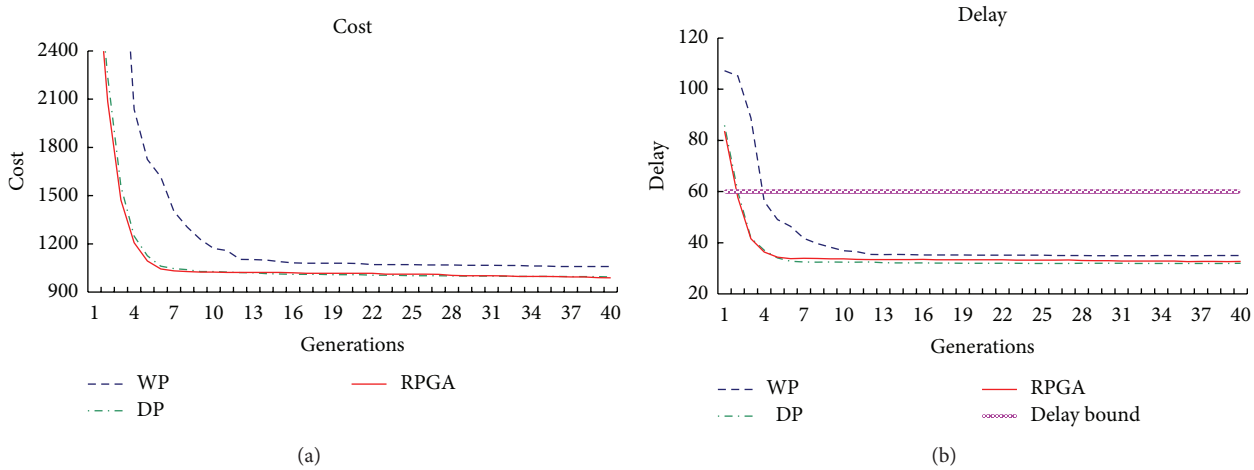


Figure 9: Evolution curves of three methods on the performance metrics: (a) cost and (b) delay.

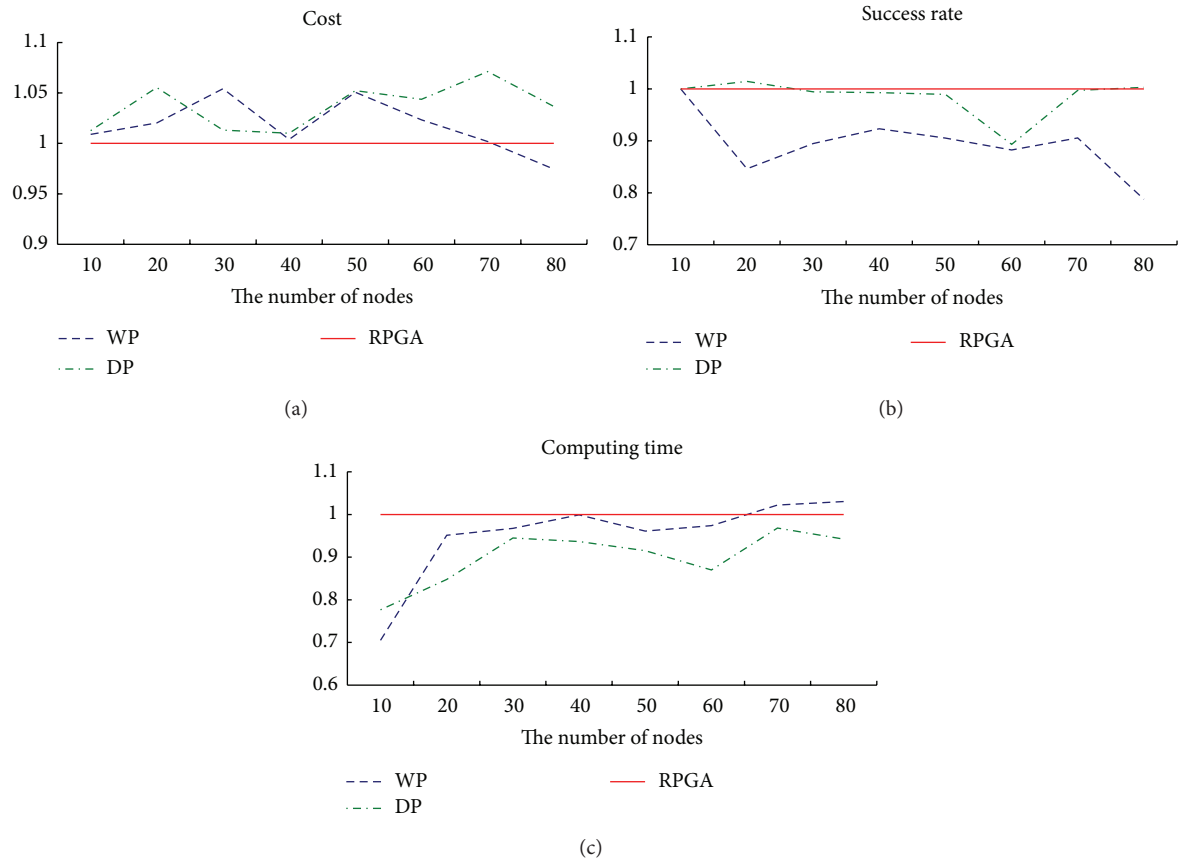


Figure 10: Comparison with three methods for different numbers of network nodes in percentage relative to the RPGA on (a) cost, (b) success rate, and (c) CPU time.

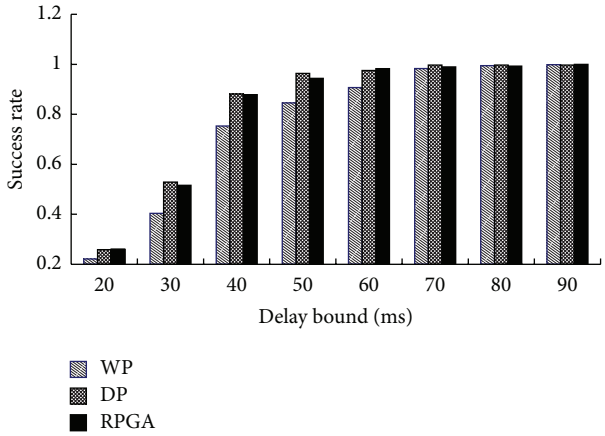


Figure 11: Comparison with three methods on the success rates for different delay bounds.

penalties on inefficient/efficient constraints during evolution. Importantly, this approach can find the optimum or near-optimal solutions even though the initial population includes infeasible solutions. The performance of the proposed algorithm was measured using three kinds of test scenarios and compared with two state-of-the-art methods. Experimental results indicate that the proposed RPGA can find near-optimal solutions and outperforms two existing methods for constrained multicast routing problems. The proposed algorithm is also robust in obtaining feasible solutions of all the test functions even though the WP method has smaller success rates for some difficult problems. In conclusion, the performance assessment also demonstrates that the proposed RP method has a remarkable capability to balance the objective function and the constraint violations as an effective and efficient method for solving a variety of QoS-based multicast routing problems.

We have observed that the computing effort of the RPGA is higher than that of other two penalty methods in small-scale networks. In the future, we will study the scalability of the proposed RPGA in finding multicast routes for dynamic MANETs with large-scale dimension. Since MANETs allow ubiquitous service access without any fixed infrastructure, developing a distributed algorithm for high mobility environments is also our future work.

References

[1] M. R. Gareg and D. S. Johnson, *Computer and Intractability: A Guide to the Theory of NP-Completeness*, W.J. Freeman, New York, NY, USA, 1979.

[2] C.-H. Liu, T.-C. Chiang, and Y.-M. Huang, "A near-optimal multicast scheme for mobile ad hoc networks using a hybrid genetic algorithm," in *Proceedings of the 20th International Conference on Advanced Information Networking and Applications*, vol. 1, pp. 465–470, April 2006.

[3] S. Gangwar, S. Pal, and K. Kumar, "Mobile ad hoc networks: a comparative study of QoS routing protocols," *International Journal of Computer Science & Engineering Technology*, vol. 2, no. 1, pp. 771–775, 2012.

[4] M. Hamdan and M. E. El-Hawary, "A Novel Genetic Algorithm Searching Approach for Dynamic Constrained Multicast Routing," in *Proceedings of the Canadian Conference on Electrical and Computer Engineering: Toward a Caring and Humane Technology (CCECE '03)*, pp. 1127–1130, May 2003.

[5] H. T. Tran and R. J. Harris, "Solving QoS multicast routing with genetic algorithms," in *Proceedings of the International Conference on Information Communications and Signal Processing*, vol. 3, pp. 1944–1948, 2003.

[6] S. Sumathy and E. Sri Harsha, "Survey of genetic based approach for multicast routing in MANET," *International Journal of Engineering and Technology*, vol. 4, no. 6, pp. 474–485, 2013.

[7] M. Gen and R. Cheng, "Survey of penalty techniques in genetic algorithms," in *Proceedings of the IEEE International Conference on Evolutionary Computation (ICEC '96)*, pp. 804–809, May 1996.

[8] A. Homaifar, C. X. Qi, and S. H. Lai, "Constrained optimization via genetic algorithms," *Simulation*, vol. 62, no. 4, pp. 242–254, 1994.

[9] J. A. Joines and C. R. Houck, "On the use of non-stationary penalty functions to solve nonlinear constrained optimization problems with GA's," in *Proceedings of the 1st IEEE Conference on Evolutionary Computation*, vol. 2, pp. 579–584, June 1994.

[10] A. B. Hadj-Alouane and J. C. Bean, "A genetic algorithm for the multiple-choice integer program," *Operations Research*, vol. 45, no. 1, pp. 92–101, 1997.

[11] A. T. Haghighat, K. Faez, M. Dehghan, A. Mowlaei, and Y. Ghahremani, "GA-based heuristic algorithms for bandwidth-delay-constrained least-cost multicast routing," *Computer Communications*, vol. 27, no. 1, pp. 111–127, 2004.

[12] M. Zhang, W. Luo, and X. Wang, "Differential evolution with dynamic stochastic selection for constrained optimization," *Information Sciences*, vol. 178, no. 15, pp. 3043–3074, 2008.

[13] C. H. Lin, "A rough penalty genetic algorithm for constrained optimization," *Information Science*, vol. 241, pp. 119–137, 2013.

[14] D. Powell and M. M. Skolnick, "Using genetic algorithms in engineering design optimization with non-linear constraints," in *Proceedings of the 5th International Conference on Genetic Algorithms*, pp. 424–431, 1993.

[15] Z. Michalewicz, "A survey of constraint handling techniques in evolutionary computation methods," in *Proceedings of the 4th Annual Conference on Evolutionary Programming*, pp. 135–155, 1995.

[16] Z. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.

[17] Z. Pawlak and A. Skowron, "Rudiments of rough sets," *Information Sciences*, vol. 177, no. 1, pp. 3–27, 2007.

[18] Z. Pawlak and A. Skowron, "Rough sets and Boolean reasoning," *Information Sciences*, vol. 177, no. 1, pp. 41–73, 2007.

[19] S. O. Kimbrough, G. J. Koehler, M. Lu, and D. H. Wood, "On a Feasible-Infeasible Two-Population (FI-2Pop) genetic algorithm for constrained optimization: Distance tracing and no free lunch," *European Journal of Operational Research*, vol. 190, no. 2, pp. 310–327, 2008.

[20] H. Salama, "The multicast routing simulator," The Real-Time Communication Project, Version 2, 1997, http://rtcomm.csc.ncsu.edu/index.htm.

[21] B. M. Waxman, "Routing of multipoint connections," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1617–1622, 1988.

[22] Z. Wang, B. Shi, and E. Zhao, "Bandwidth-delay-constrained least-cost multicast routing based on heuristic genetic algorithm," *Computer Communications*, vol. 24, no. 7-8, pp. 685–692, 2001.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 518213, 11 pages
http://dx.doi.org/10.1155/2013/518213

Research Article

Effective Proactive and Reactive Defense Strategies against Malicious Attacks in a Virtualized Honeynet

Frank Yeong-Sung Lin, Yu-Shun Wang, and Ming-Yang Huang

Department of Information Management, National Taiwan University, Taipei, Taiwan

Correspondence should be addressed to Yu-Shun Wang; yu.shun.wang.tw@gmail.com

Received 11 April 2013; Accepted 22 July 2013

Academic Editor: Anyi Chen

Virtualization plays an important role in the recent trend of cloud computing. It allows the administrator to manage and allocate hardware resources flexibly. However, it also causes some security issues. This is a critical problem for service providers, who simultaneously strive to defend against malicious attackers while providing legitimate users with high quality service. In this paper, the attack-defense scenario is formulated as a mathematical model where the defender applies both proactive and reactive defense mechanisms against attackers with different attack strategies. In order to simulate real-world conditions, the attackers are assumed to have incomplete information and imperfect knowledge of the target network. This raises the difficulty of solving the model greatly, by turning the problem nondeterministic. After examining the experiment results, effective proactive and reactive defense strategies are proposed. This paper finds that a proactive defense strategy is suitable for dealing with aggressive attackers under "winner takes all" circumstances, while a reactive defense strategy works better in defending against less aggressive attackers under "fight to win or die" circumstances.

1. Introduction

The vision for most service providers is to provide high-quality service and improve customer satisfaction, thus maximizing profit. From an infrastructure perspective, the evolution of computing architecture has shifted from mainframe, cluster computing, distributed computing, and grid computing to cloud computing. As a recent and increasingly noteworthy trend in information technology (IT), cloud computing describes a new service model based on the Internet. From a managerial perspective, one of the key success factors of cloud computing is its ability to promise and achieve high quality and availability of service.

Applying virtualization technology makes the job of providing enterprise security more difficult. As observed in the IBM X-Force Mid Year Trend and Risk Report conducted in August 2010 [1], attackers continue to take advantage of security flaws. The rate of vulnerability disclosures in 2010 is higher than any point from 2000 to 2009. The number of virtualization vulnerability disclosures from 1999 through the end of 2009 ascended rapidly, peaking in 2008 at 100. While this number fell by 12 percent to 88 in 2009, this drop indicates that virtualization vendors have recognized the threat these flaws pose and have increased their attention to security. Although the ratio of virtualization vulnerability disclosures increased by only 1 percent from 2007 through 2009, these vulnerabilities still represent a notable security threat. Therefore, it is increasingly important to understand the security implications of virtualization technology.

In addition to external malicious attackers, another weak component of networks is insiders. They insensibly assist people with bad intentions in their legal use of computer systems or networks. Such insider-assisted malicious attacks adopt social engineering as a method to exploit common human behavior. These attacks require a lower degree of technical ability than standard malicious attacks but can cause higher degrees of damage. As a result, organizations emphasize the importance of policy enforcement and increase employee education to mitigate the risks posed by social engineering.

In response to these problems, this paper considers both proactive defense resources, such as firewalls, IDS, IPS, and reactive defense techniques. All types of resources considered

Table 1: Given parameters.

Notation	Description
N	The index set of all nodes
C	The index set of all core nodes
L	The index set of all links
M	The index set of all levels of virtual machine monitors (VMMs)
H	The index set of all types of honeypots
P	The index set of candidate nodes equipped with false target function
Q	The index set of candidate nodes equipped with fake traffic generating function
R	The index set of candidate nodes equipped with false target and fake traffic generating function
S	The index set of all kinds of services
B	The defender's total budget
w	The cost of constructing one intermediate node
o	The cost of constructing one core node
P	The cost of each virtual machine (VM)
k_i	The maximum number of virtual machines on VMM level i, where $i \in M$
α_i	The weight of ith service, where $i \in S$
E	All possible defense configurations, including resources allocation and defending strategies
Z	All possible attack configurations, including attacker attributes, strategies, and transition rules
$\overrightarrow{A_{ij}}$	An attack configuration, including the attributes, strategies, and transition rules of the attacker launches jth attack on ith service, where $i \in S$, $1 \le j \le F_i$
F_i	The total attacking times on ith service for all attackers, where $i \in S$
γ	The cost of constructing a reconfiguration function to one node
λ	The minimum number of hops from core node attackers can start to compromise
c_i	The number of hops from core node category i attackers starts to compromise, where $i \in Z$

in this work one is quantified not only by monetary, but also by time, labor, and other possible factors. As defenders seek out an appropriate defense resource allocation within budget limitations and observe the Quality of Service (QoS) requirement, it becomes increasingly important to best determine how to find a defense mechanism that can detect risky attack behavior early and mislead attackers to routes distant from the server before attackers are at the gates.

In this paper, we assume that organizations may encounter a great diversity of threats. Whether these threats are external or internal, they can bring about vast loss to finances and reputation. However, budgets for security and training are often inadequate. Hence, it is much more important for a system or network to enhance robustness in order to satisfy QoS requirements for service users, than to prevent all categories of malicious attacks. This symbiotic concept to security is called survivability, which is widely defined and applied in previous works [2–6].

Table 2: Decision variables.

Notation	Description
$\overrightarrow{D_i}$	A defense configuration, including resource allocation and defending strategies on ith service, where $i \in S$
$T_{ij}(\overrightarrow{D_i}, \overrightarrow{A_{ij}})$	1 if the attacker achieves his goal successfully and 0 otherwise, where $i \in S$, $1 \le j \le F_i$
n_i	The proactive defense resource allocated to node i, where $i \in N$
l_i	The number of VMM level i purchased, where $i \in M$
δ_i	The number of services that honeypot i can simulate, where $i \in H$
ε_i	The interactive capability of false target honeypot i, where $i \in P$
θ_i	The maximum throughput of fake traffic of honeypot i, where $i \in Q$
$V(l_i)$	The cost of VMM level i with l_i VMMs, where $i \in M$
$h(\delta_i, \varepsilon_i)$	The cost of constructing a false target honeypot, where $i \in P$
$f(\delta_i, \theta_i)$	The cost of constructing a fake traffic generator honeypot, where $i \in Q$
$t(\delta_i, \varepsilon_i, \theta_i)$	The cost of constructing a honeypot equipped with false target and fake traffic functions, where $i \in R$
B_{NL}	The budget of constructing nodes and links
$B_{\mathrm{proactive}}$	The budget of proactive defense resource
B_{special}	The budget of reactive defense resource
$B_{\mathrm{virtualized}}$	The budget of virtualization
B_{honeypot}	The budget of honeypots
B_{reconfig}	The budget of reconfiguration functions
e	The total number of intermediate nodes
q_{ij}	The capacity of direct link between nodes i and j, where $i \in N$, $j \in N$
$g(q_{ij})$	The cost of constructing a link from node i to node j with capacity q_{ij}, where $i \in N$, $j \in N$
x_i	1 if node i is equipped with false target function and 0 otherwise, where $i \in N$
y_i	1 if node i is equipped with fake traffic generating function and 0 otherwise, where $i \in N$
z_i	1 if node i is equipped with reconfiguration function and 0 otherwise, where $i \in N$

Survivability is a typical metric that measures network performance under intentional attacks or other failures. Traditionally, network security status is divided into two discrete types: compromised and safe. Typically, when providing services, the evaluation criterion is the quality of service, but applying this dichotomy of compromise and safety ignores the intermediate status. For instance, while under an attack, the performance level of each service continually declines. Therefore, network status should be represented in a continuous form [2]. Hence, in this paper, survivability is chosen as the metric for describing network status. According to [2], survivability is defined as: "The capability of a system to fulfill its mission, in a timely manner, in the presence of

Table 3: Verbal notations.

Notation	Description
G_{core_i}	Loading of each residual core node i, where $i \in C$
U_{link_i}	Link utilization of each link i, where $i \in L$
K_{effect}	Negative effect caused by applying fake traffic adjustment
I_{effect}	Negative effect caused by applying dynamic topology reconfiguration
J_{effect}	Negative effect caused by applying local defense
O_{tocore}	The number of hops legitimate users experienced from one boundary node to core nodes
Y	The total compromise events
$W_{\text{threshold}}$	The predefined threshold about QoS
W_{final}	The QoS level at the end of attack
$W(\cdot)$	The value of QoS determined by G_{core_i}, U_{link_j}, K_{effect}, I_{effect}, J_{effect}, and O_{tocore}, where $i \in C$, $j \in L$
ρ_{defense_i}	The total defense resource of the shortest path from compromised nodes detected to core node i divided by total defense resource, where $i \in C$
τ_{hops_i}	The number of hops from compromised nodes detected to core node i divided by the number of hops from attacker's starting point, where $i \in C$
ω_{degree_i}	The linking number of core node i divided by the maximum number in the topology, where $i \in C$
$s^i{}_{\text{priority}_j}$	The priority of service j provided by core node i divided by the maximum service priority in the topology, where $i \in S$, $j \in C$
$\beta_{\text{threshold}}$	The risk threshold of core nodes
$\beta(\cdot)$	The risk status of each core node, which is the aggregation of defense resource, number of hops, link degree, and service priority

Table 4: Environment Parameters.

Testing platform	
Programming language	C
Compiler	GNU GCC 4.6.2
Total evaluation times	60,000
Distributions applied to describe attackers' attributes	Normal distribution

attacks, failure, or accidents. . .including networks and large-scale systems of systems."

Along with the concept of survivability, the vulnerability of each node is determined by the Contest Success Function (CSF), which is also applied in [7–10]. CSF originates from the economic rent seeking problem found in Economic Theory. This method also applies a continuous approach to the problem. The form of the CSF is success probability $= T^m/(T^m + t^m)$, when applied to attack and defense scenarios, where T represents the resources invested by the attacker and t stands for resources deployed by the defender. Further, m is known as contest intensity, which illustrates the nature of the battle, while success probability is the probability of a node being compromised. When the value m is between 0 and 1, it represents "fight to win or die" circumstance [11], which means the effectiveness of resources is insignificant. For $m \geq 1$, the effectiveness of resources invested by both sides is exponentially increasing. If m closes to ∞, it stands for a "winner takes all" circumstance; significant advantage is granted to the stronger side, even if that side is stronger by only one invested resource [12].

Table 5: Parameters for defender.

Parameters	Value
Topology type	Scale-free network
Number of nodes	49
Number of core nodes (each service)	6 (1, 2, 3)
Number of terminal nodes	5
Number of services	3
Weight of each service	1 : 2 : 3
Number of users	30
Defender total budget	1,700,000
Topology construction budget	700,000
Proactive defense budget	400,000
Reactive defense budget	600,000

Table 6: Parameters for attacker.

Parameters	Value
Total attack budget	A normal distribution with lower bound 300,000 and upper bound 1,500,000
Capability	A normal distribution with lower bound ε and upper bound 1
Aggressiveness	A normal distribution with lower bound 0.1 and upper bound 0.9
Attacker's objective	Service disruption or steal confidential information

There are many popular methodologies applied for solving survivability problems. In recent years, game theory is a widely used one. Nevertheless, in [3], the authors point out that this solution approach is limited for deterministic scenarios. Even the emerging branch of game theory, stochastic game, is still confined with this assumption that all values of probabilistic variables have to be determined before the attack and defense starts. This feature has a negative effect on creating cyber attack and defense scenario since, in real world, decisions made during attack and defense depend on current status. Given all values of variables makes the scenario far from reality.

For example, when choosing a victim from candidate nodes, an attacker should apply information like the loading of each node, traffic amount on each link, and/or number of users on each node to evaluate the importance of all candidates. Then, choosing the most appropriate one as the target, yet the restriction of game theory enforces the choosing probability which should be determined at the beginning of the cyber warfare. In other words, those variations happened during attack and defense, like traffic reroute, link status, node conditions, are ignored. Consequently, in this work, Monte Carlo simulation is applied to consider hopefully and cover every angle in the attack and defense scenario.

2. Problem Formulation

2.1. Problem Description. In order to improve system survivability, the defender deploys both proactive and reactive defense resources to confront different attacks. Proactive defense resources are deployed before an attack is launched. In this paper, proactive resources include a firewall system, antivirus software, detection techniques, such as Intrusion Detection System (IDS) and Intrusion Protection System (IPS).

Alternatively, reactive defense resources are activated during an attack as an immediate action for the defender. The mechanisms considered in this paper can strengthen defenses, provide deception, and provide resource concentration.

For strengthening defenses, since the scenario considered in this paper is constructed in a virtualized environment, a number of Virtual Machines (VMs) are governed by a Virtual Machine Monitor (VMM) that controls all information details for all VMs. When an attack event is detected, local defense functions for each VMM are activated automatically to raise the defense capability of the virtualized nodes belonging to the same VMM. However, this mechanism does not always create positive effects. Once an attacker determines that the target network is a virtualized environment and discovers the existence of a VMM, he can compromise the VMM through vulnerabilities in APIs [13]. If the VMM is compromised, all VMs belonging to this VMM are also compromised.

Deception mechanisms are widely applied in defense, and in this paper honeypots are considered to distract attackers. According to [14–17], honeypots not only serve as a passive decoy fooling attackers into believing they have achieved their goal and preemptively terminating the attack but also as an active lure that acts as a service-providing node to attract attackers. The former is known as a false target, and the latter can be implemented by a fake core node that spreads service-like traffic to attract attackers. When facing different attackers, the deceiving probability of each honeypot is different. This probability is jointly determined by attackers' capability and the interaction level of a honeypot.

As for resource concentration, by modifying the concept of "rotation" discussed in [18–20] and adapting it to our scenario, the defender can adopt dynamic responsive strategies to improve system survivability. Hence, while under an attack, the defender can apply dynamic topology reconfiguration to exchange the neighbor of one core node, which has the strongest defensive resources, with a node that is close to the attacker's current location.

However, since dynamic topology reconfiguration requires node rotation, it negatively impacts QoS. Therefore, unless the risk level of a core node exceeds a predefined threshold, the defender will not activate this mechanism. Furthermore, false positive and false negative situations for every defense mechanism are considered.

In addition to the defense mechanisms described above, the following attributes are also considered for creating a realistic attack-defense scenario.

2.1.1. Goal. Generally, attackers target a network to either disrupt services or to compromise servers and steal sensitive information. Therefore in this paper an attacker's goal may be service disruption or the theft of confidential information. Service disruption is achieved by ensuring that the minimal level of service quality is not fulfilled. In the case of information theft, attackers usually establish their target before launching an attack, and once core nodes with the desired information are compromised, the defenders lose.

2.1.2. Budget. Budget stands for the primary resources for an attack, including money, manpower, computing effort, time, and other important factors. Without sacrificing generality, an attackers' budget follows a general distribution. When determining the result of a compromising event, for example, one attacker invests a certain amount of attack resources on compromising the target node, and the CSF is applied to decide the compromised probability. In contrast with [21] and [22], if an attacker invests more resources than the defender on a given target node, it is not guaranteed that the attacker wins; it only raises the compromised probability.

2.1.3. Capability. This criterion depicts an attackers' proficiency and is also described by a general distribution. For highly proficient attackers, there is a high probability that they will see through reactive defense mechanisms, such as honeypots.

2.1.4. Attack Type. In general, a malicious attacker launches an attack from one of the boundary nodes, which is commonly considered an external attack. In contrast, other attacks can be launched by malicious insiders and cause more severe consequences [13]. Malicious insiders are able to choose an internal node as their starting position for compromising the network.

In the real world, external attackers may apply social engineering to escalate their access privileges. This mechanism allows the attacker to bypass some proactive defense facilitates, like a firewall. As a result, attackers have an edge on compromising the network.

In order to best mimic real-world conditions, all attack types discussed above are considered in this paper.

2.1.5. Aggressiveness. This metric describes the preferred compromised probability of an attacker when attacking a target node. This attribute is highly dependent on budget, since the compromised probability is the left-hand side of the CSF. In other words, the attack resources required for compromising a target node are calculated by the given defense resources, contest intensity (m), and attacker's aggressiveness. Highly aggressive attackers prefer a high success probability and like to spend a larger amount of resources to compromise the target node than less aggressive attackers. An attacker's aggressiveness is determined by a general distribution.

2.1.6. Next Hop Selection Criterion. In this paper, the attackers are assumed to have incomplete information and imperfect knowledge. Since they can only gather local information,

(1) Attackers only have incomplete information.
(2) The defender only has incomplete information regarding the network since there are unaware vulnerabilities.
(3) A service is provided by multiple core nodes.
(4) Each service has different weights.
(5) One virtual machine only provides one service.
(6) Only malicious nodal attacks are considered.
(7) The compromise probability of a target node is determined by the Contest Success Function (CSF).

Box 1: Problem assumptions.

such as the defense level of one hop neighbors, link traffic, or link utilization, attackers must use available information wisely to choose a proper strategy to select the next victim. Inspired by the seminal work of [23], possible attack strategies may be developed based on the quantity of proactive defense resources, link utilization, or a portion of the target service traffic of each candidate node. Additionally, this paper also considers a more irrational strategy of a blind attack.

By applying general distributions to describe attacker-related attributes, the total number of attacker categories is nearly infinite. This feature increases the generalizability of our model. The detailed assumptions are described in Box 1.

To describe the attack procedures in more detail, we develop the following idea to explain the interaction between the defender and attackers. The following figures are drawn from the attacker's view for clarity. In other words, all figures are a logical topology that has been already virtualized, since one physical machine may represent many VMs. The explanations of components are listed in Figure 1.

First, the defender deploys proactive defense resource on each node, and the attacker starts to compromise the network from the edge nodes (S). Before launching an attack, the attacker probes all the candidate nodes to gather sufficient information and determine the next hop selection criterion for this compromise event. By applying the next hop selection criterion, the victim is chosen. The result of this compromise event is determined by the CSF. If the target is compromised, it becomes a spring board for attackers to assault other uncompromised neighbors. Corresponding information is shown in Figure 2. The topology demonstrated in Figures 2 and 3 for presentation purposes; the topology structure implemented in the simulations is a scale-free network.

The risk level of each core node is evaluated by minimum defense resources, the number of hops, the link degree, and service priority. Minimum defense resource stands for the shortest path from compromised nodes to one core node, divided by total defense resources. The number of hops is determined by the minimum number of hops from compromised nodes to one core node, divided by the maximum number of hops from the attackers' starting point to one core node. Link degree is the linking number of each core node divided by the maximum linking number in the topology. Service priority is the weight of the service that is provided by the target core node divided by the highest weight of service in the topology. If any of the core nodes is in danger, the defender can activate defense mechanisms, such as a fake traffic generator and a dynamic topology reconfiguration under QoS limitations.

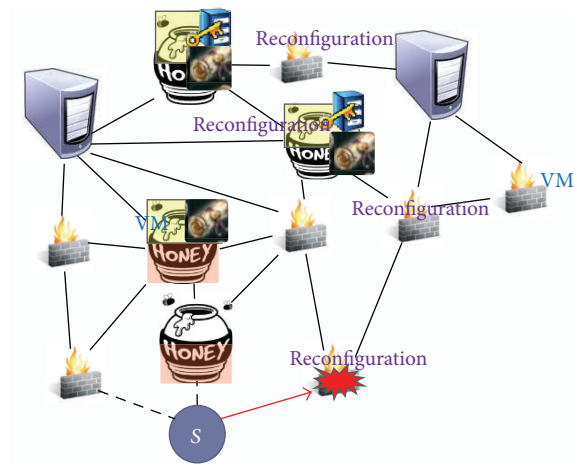


Figure 1: Explanations of Components.

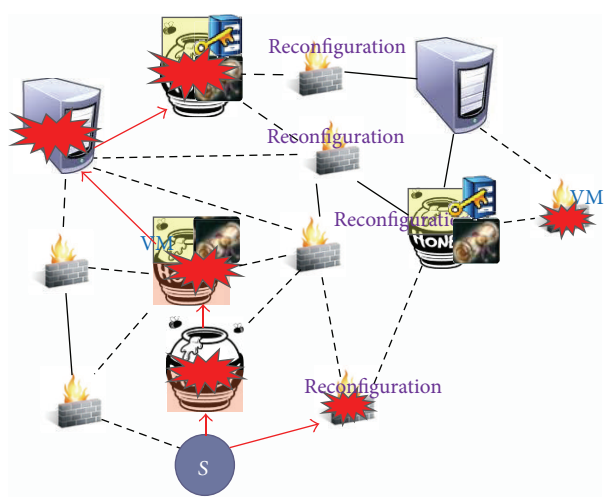


Figure 2: Sample network and resource allocation scheme.

When the defender activates a dynamic topology reconfiguration, only nodes implemented with the reconfiguration function are considered. First, the defender chooses the least proactively defended neighbor with the reconfiguration function of a risky core node. The defender then selects another node that is both not a neighbor of the core node and the most proactively defended nearby the node selected from the previous step. If there is a fake traffic generating honeypot, the defender is also able to activate the mechanism for influencing an attacker's next hop selecting criterion.

In the event that attackers attack the virtualized nodes and this malicious event is detected by the defender, the local

Given
(1) all possible defense configurations set, including defense resource allocation and defending strategies,
(2) all possible attack configurations set, including attacker attributes, strategies, and selection criterion,
(3) the total attack times on each service.
Objective
to minimize the service compromised probability of the target network.
Subject to
(1) budget constrain for both the defender and attackers,
(2) the minimum QoS requirement for legitimate users.
To determine
the effective defense strategies to allocate resources.

Box 2: Problem description.

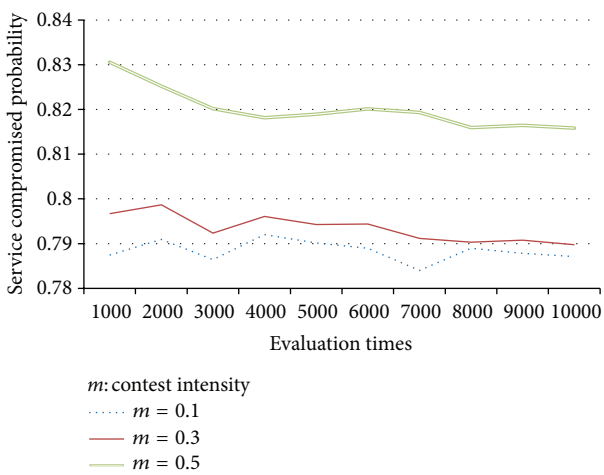


Figure 3: A possible result of an attack and defense scenario.

defense mechanism is activated automatically. Consequently, the defense level increases for the attacked node, the VMM, and the rest of this virtualized group. If attackers see that the node is under a virtualized environment, they may continue to compromise the VMM.

When a core node is compromised, the defender must evaluate whether the performance level still fulfills the minimum QoS requirement and update the degree of risk regarding each core node. If an attacker targets multiple services, once they compromise a false target honeypot and are deceived by it, they will change their target to the next service. Finally, if attackers exhaust their budget, the attack is terminated and the defender wins the battle. One possible result is shown in Figure 3. A structural description of the problem is presented in Box 2.

2.2. Mathematical Formulation. The scenario discussed previously is modeled as a minimization problem, given the parameters and decision variables shown in Tables 1 and 2, respectively. Furthermore, since the high amount of randomness involved renders the nature of this problem nondeterministic, it is quite difficult to formulate this problem purely through mathematics. To explain the nondeterministic nature, the following example is used: considering two adversaries with the same attack strategy, since the problem is nondeterministic, the consequences can be totally different. One may be distracted by a honeypot and the other may successfully achieve the goal. This feature dramatically expands the width in survivability studies. Further, the scenario considered in this work gives the defender the average survivability analysis of a network.

There are three major reasons that an average survivability analysis is more meaningful for the defender. First, in the real world, there are only few adversaries holding "complete" information regarding the target network. This kind of "worst case" rarely happen. Secondly, analyzing the worst case though gives a defender the lower bound of network survivability. Nevertheless, it overestimates the budget required for defending a network. An average survivability analysis makes conclusions closer to reality. Last but not the least, in average case analysis, if the defender wants to evaluate the survivability when facing a stronger adversary, it can be achieved by simply tuning the parameter. Average case analysis is flexible and adjustable according to the demand of a defender.

Thus, to well describe the problem, some verbal notations and constraints are included, which are shown in Table 3.

Objective function:

$$\min_{\overrightarrow{D_i}} \frac{\sum_{i\in S}\left[\alpha_i \times \sum_{j=1}^{F_i} T_{ij}\left(\overrightarrow{D_i}, \overrightarrow{A_{ij}}\right)\right]}{\sum_{i\in S}\left(\alpha_i \times F_i\right)}, \quad \text{(IP 1)}$$

subject to
mathematical constraints:

$$\overrightarrow{D_i} \in E \quad \forall i \in S, \quad \text{(IP 1.1)}$$

$$\overrightarrow{A_{ij}} \in Z \quad \forall i \in S,\ 1 \le j \le F_i, \quad \text{(IP 1.2)}$$

$$q_{ij} \ge 0 \quad \forall i, j \in N, \quad \text{(IP 1.3)}$$

$$x_i + y_j \ge 1 \quad \forall i, j \in H, \quad \text{(IP 1.4)}$$

$$c_i \ge \lambda \quad \forall i \in Z, \quad \text{(IP 1.5)}$$

$$B_{\text{NL}} + B_{\text{proactive}} + B_{\text{reactive}} \le B, \quad \text{(IP 1.6)}$$

$$B_{\text{virtualized}} + B_{\text{honeypot}} + B_{\text{reconfig}} \le B_{\text{reactive}}, \quad \text{(IP 1.7)}$$

$$w \times e + o \times \|C\| + \frac{\sum_{i\in N}\sum_{j\in N} g\left(q_{ij}\right)}{2} \le B_{\mathrm{NL}}, \quad \text{(IP 1.8)}$$

$$\sum_{i\in N} n_i \le B_{\mathrm{proactive}}, \quad \text{(IP 1.9)}$$

$$\sum_{i\in M} v\left(l_i\right) + p \times \sum_{i\in M} l_i \times k_i \le B_{\mathrm{virtualized}}, \quad \text{(IP 1.10)}$$

$$\sum_{i\in P} x_i \times h\left(\delta_i, \varepsilon_i\right) + \sum_{j\in Q} y_j \times f\left(\delta_j, \theta_j\right) + \sum_{i\in N}\sum_{j\in N} x_i \times y_j \times t\left(\delta_i, \varepsilon_i, \theta_j\right) \le B_{\mathrm{honeypot}}, \quad \text{(IP 1.11)}$$

$$\sum_{i\in N} z_i \times r \le B_{\mathrm{reconfig}}, \quad \text{(IP 1.12)}$$

$$w \times e \ge 0, \quad \text{(IP 1.13)}$$

$$g\left(q_{ij}\right) \ge 0 \quad \forall i, j \in N, \quad \text{(IP 1.14)}$$

$$n_i \ge 0 \quad \forall i \in N, \quad \text{(IP 1.15)}$$

$$v\left(l_i\right) \ge 0 \quad \forall i \in M, \quad \text{(IP 1.16)}$$

$$h\left(\delta_i, \varepsilon_i\right) \ge 0 \quad \forall i \in P, \quad \text{(IP 1.17)}$$

$$f\left(\delta_j, \theta_j\right) \ge 0 \quad \forall j \in Q, \quad \text{(IP 1.18)}$$

$$t\left(\delta_i, \varepsilon_i, \theta_j\right) \ge 0 \quad \forall i \in N, \quad \text{(IP 1.19)}$$

$$x_i = 0 \text{ or } 1 \quad \forall i \in N, \quad \text{(IP 1.20)}$$

$$y_i = 0 \text{ or } 1 \quad \forall i \in N, \quad \text{(IP 1.21)}$$

$$z_i = 0 \text{ or } 1 \quad \forall i \in N, \quad \text{(IP 1.22)}$$

verbal constraints:

$$\frac{\int_{y=1}^{Y}\left[W\left(G_{\mathrm{core}_i}, U_{\mathrm{link}_j}, K_{\mathrm{effect}}, I_{\mathrm{effect}}, J_{\mathrm{effect}}, O_{\mathrm{tocore}}\right)\right] dy}{Y} \ge W_{\mathrm{threshold}}, \quad \text{where } i \in C,\ j \in L, \quad \text{(IP 1.23)}$$

$$W_{\mathrm{final}} \ge W_{\mathrm{threshold}}. \quad \text{(IP 1.24)}$$

For each core node, when

$$\beta\left(\rho_{\mathrm{defense}}, \tau_{\mathrm{hops}}, \omega_{\mathrm{degree}}, s_{\mathrm{priority}_i}\right) \ge \beta_{\mathrm{threshold}}, \quad \text{where } i \in S, \quad \text{(IP 1.25)}$$

the defender can activate the reactive defense mechanisms.

The goal of the objective function is to minimize the compromised service probability, which is modeled as the weighted total of successful attacks, divided by the total weighted number of attacks. The result of an attack is determined by $T_{ij}(\overrightarrow{D_i}, \overrightarrow{A_{ij}})$. For each service, the defender constructs a defense configuration, including defense resource allocation and defending strategies to oppose attackers targeting the service with different attack configurations, including attacker attributes, strategies and transition rules. Not only malicious attacks but also QoS issues are taken into consideration.

Equation (IP 1.1) stands for the feasibility of the defense configuration of each service. For the attacking side, (IP 1.2) denotes that the attack configuration should be feasible. Equation (IP 1.6) denotes that the summation of defense resources spent should not exceed total budget B. Equations (IP 1.7)~(IP 1.19) jointly restrain the budget of each type of defense resource individually. Equations (IP 1.20)~(IP 1.22) impose binary restrictions on decision variables. Finally, (IP 1.23)~(IP 1.25) jointly represent that the performance reduction caused by either malicious attacks or activating reactive defense mechanisms should not violate the QoS requirement.

3. Numerical Analysis

In this paper, a scale-free network is constructed for evaluation, since this structure is similar to real-world forms, such as the Internet. The implementation algorithm is referenced from [24].

3.1. Simulation Environment. Table 4 presents the system experiment parameters. Defender-related parameters are illustrated in Table 5. The unit of budget is dollar. For attackers, the parameters are shown in Table 6.

The value of each attacker's budget, capability and aggressiveness is governed by a normal distribution with different lower and upper bounds.

3.2. Numerical Result. As mentioned previously, the Contest Success Function is applied for quantifying vulnerability. The value of contest intensity is classified into two groups: scores greater than 1 and scores smaller than 1.

3.2.1. Convergence. The first issue to address before constructing meaningful simulations is convergence. In this paper, the convergence of data is considered as numerical stability. While the magnitude of data vibrations is within the acceptable interval, for example, 0.2%, the corresponding number of simulation times (M) is set as the number of evaluations for each attack and defense scenario.

For rigorousness, in convergence experiments, the effect of contest intensity is jointly considered. Three different magnitudes are simulated on a 49 node scale-free network. For the following experiments in this subsection, the horizontal axis represents the evaluation of the number of attacks, and the vertical axis stands for the network system compromised probability, which is the objective function of the proposed mathematical model.

Figures 4 and 5 show the result of 10,000 simulations with different contest intensity. Here, it is clear that the value of service compromised probability is unstable. In Figures 6 and 7, the vibration becomes alleviative but is still not convergent. When the number of simulations is raised to 100,000, as

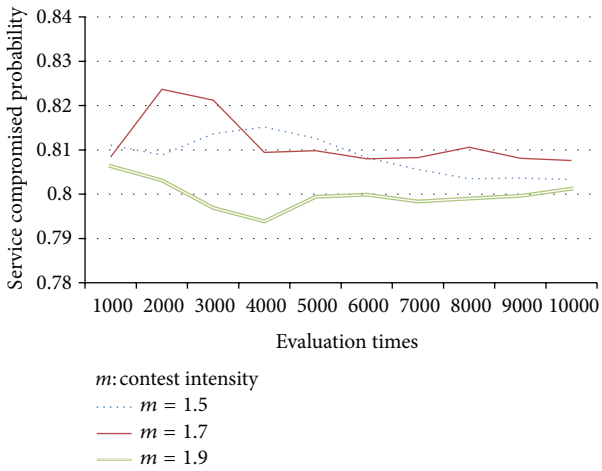


Figure 4: Convergence test on $m < 1$ group for 10,000 simulations.

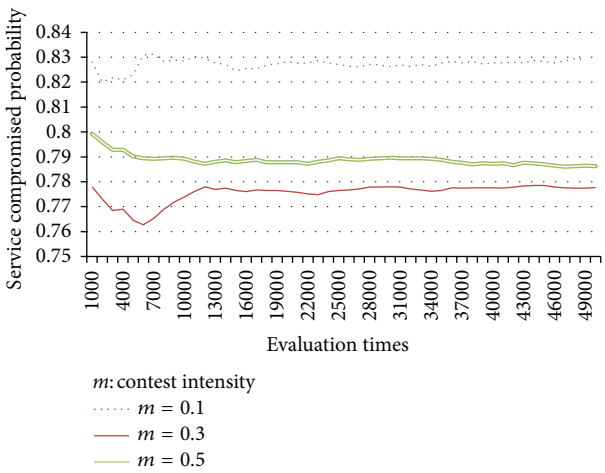


Figure 5: Convergence test on $m > 1$ group for 10,000 simulations.

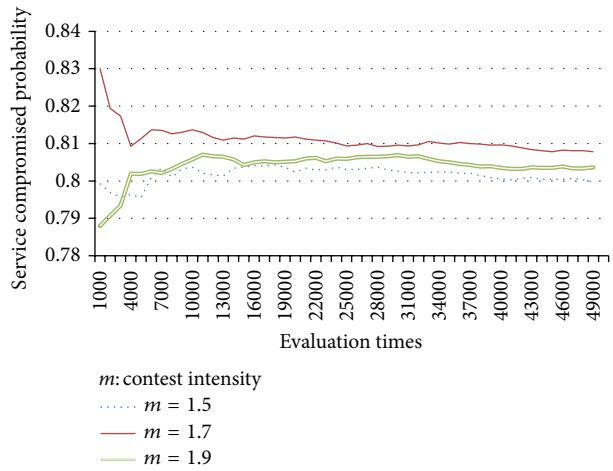


Figure 6: Convergence test on $m < 1$ group for 50,000 simulations.

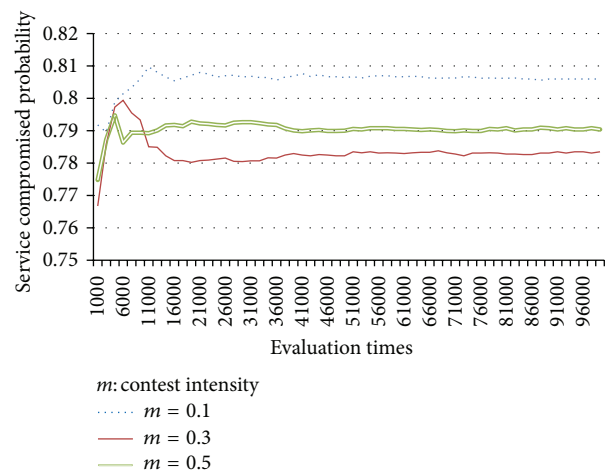


Figure 7: Convergence test on $m > 1$ group for 50,000 simulations.

shown in Figures 8 and 9, there is a stable trend after 60,000 among all different values of contest intensity.

From Figures 4 to 9, it is clear that the 60,000 simulations are a large enough number to give converging results among all values of contest intensity. Therefore, M is set to be 60,000.

3.2.2. Analysis of Key Factors Influencing Service Compromised Probability. Contest intensity has great influence on the nature of network attack and defense. However, as shown in Figure 10, there is no consistent tendency between contest intensity and service compromised probability. The same conclusions also appear in [8–10].

If the effects of contest intensity and attacker's aggressiveness are jointly considered, there are some meaningful and explainable results.

In Figure 11, three value intervals of aggressiveness including 0.1 to 0.9 (average), 0.1 to 0.5 (less aggressive), and 0.5 to 0.9 (aggressive) are simulated among different values of contest intensity. It is clear that when both the effect of contest intensity and an attack's aggressiveness are considered, there are consistent trends. For aggressive attackers, it is more advantageous to have higher values of contest intensity. Alternatively, less aggressive attackers have leverage when the value of contest intensity is small. However, for average attacks there is no clear tendency.

The value of contest intensity is separated into high-value and low-value groups. Here Figure 12 illustrates the variation of service compromised probability in low-level contest intensity groups when facing less aggressive attackers. While in Figure 12 the objective function value presents a linear decreasing form, Figure 13 demonstrates an exponentially decreasing from.

Figures 14 and 15 also present a similar phenomenon; when facing aggressive attackers, the variation of service compromised probability shows an exponentially increasing trend for contest intensity belonging to the low-level group. For the high-level group the tendency is more linear.

According to the previous results, the key factors influencing service compromised probability include contest intensity and attack aggressiveness. These results are summarized in Figure 16.

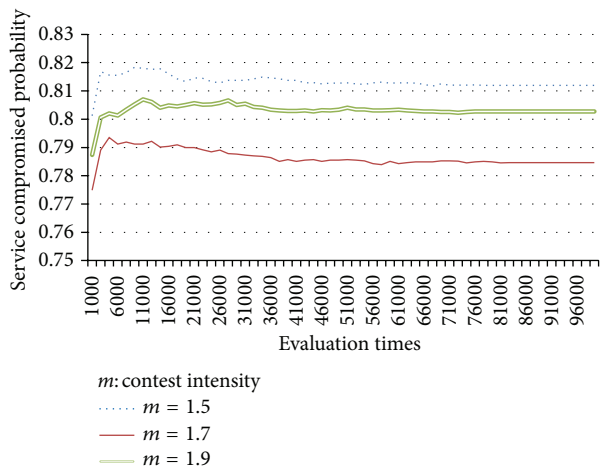


FIGURE 8: Convergence test on $m < 1$ group for 100,000 simulations.

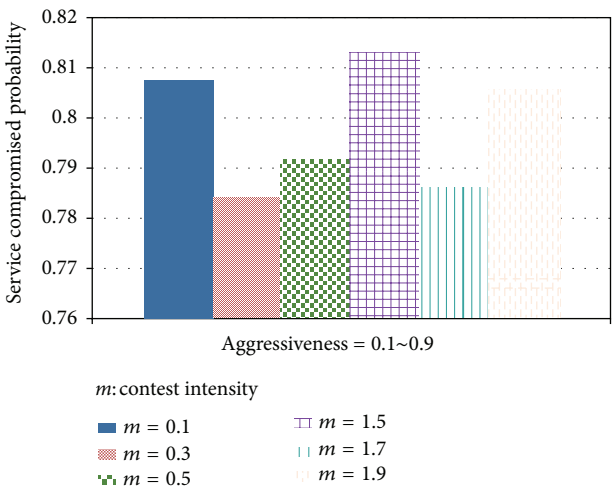


FIGURE 9: Convergence test on $m > 1$ group for 100,000 simulations.

4. Discussion of Results

Based on the simulation results, some interesting and meaningful arguments are presented in this section.

4.1. The Effectiveness of Resources Invested by the Defender and Attackers Is Highly Dependent on the Nature of Battle. The objective function value shown in Figure 12 presents a linear decreasing form when the contest intensity belongs to a low-level group and the defender is dealing with less aggressive attackers. The reason is that the effectiveness of resources invested by both players is insignificant. Thus, the service compromised probability is less sensitive with contest intensity under "fight to win or die" circumstances.

Furthermore, less aggressive attackers only invest few resources into compromising a target. With the same total budget, they are capable of launching more attacks than aggressive attackers. Therefore, the service compromised probability of less aggressive attackers is much higher than more aggressive attackers.

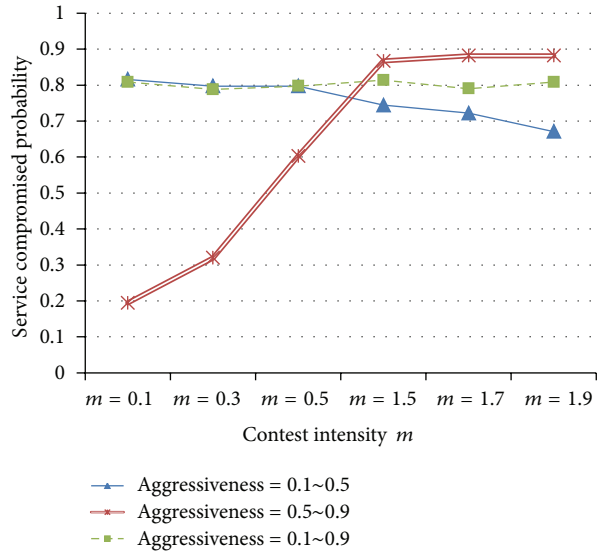


FIGURE 10: Service compromised probability under different contest intensity.

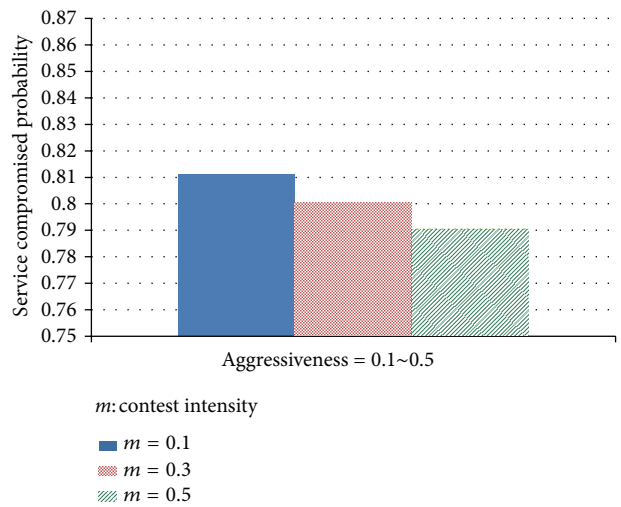


FIGURE 11: Effect of Aggressiveness among Different Contest Intensity.

In the case of "winner takes all," Figure 13 illustrates an exponentially decreasing trend among service compromised probability; this is because the effectiveness of resources invested is significant to the result of a battle. With a larger value of contest intensity, the vantage of a defender against less aggressive attackers is more obvious.

On the contrary, for aggressive attackers, there is an exponentially increasing tendency under "fight to win or die" circumstances. Since the contest intensity serves as the exponent of the Contest Success Function, when the value increases the effectiveness of resources invested grows exponentially. This is further shown in Figure 14.

Nevertheless, the exponential tendency does not appear through all values of contest intensity. For the "winner takes all" circumstance, in Figure 15, the increasing rate of service

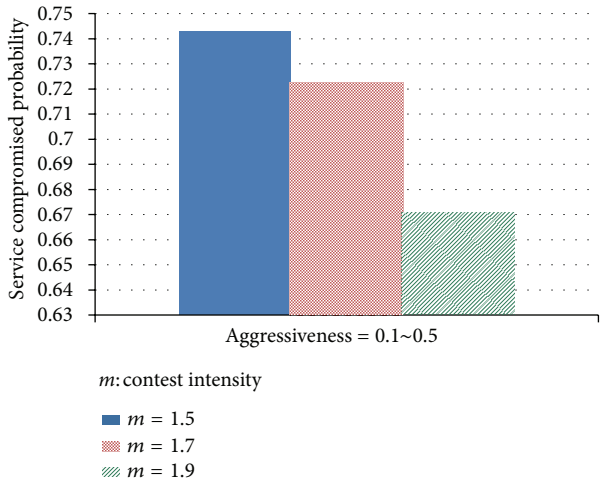


Figure 12: Service compromised probability of less aggressive attacker under $m < 1$ circumstance.

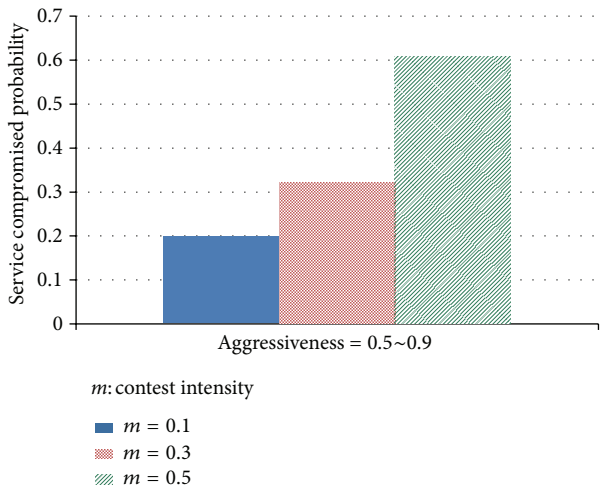


Figure 13: Service compromised probability of less aggressive attacker under $m > 1$ circumstance.

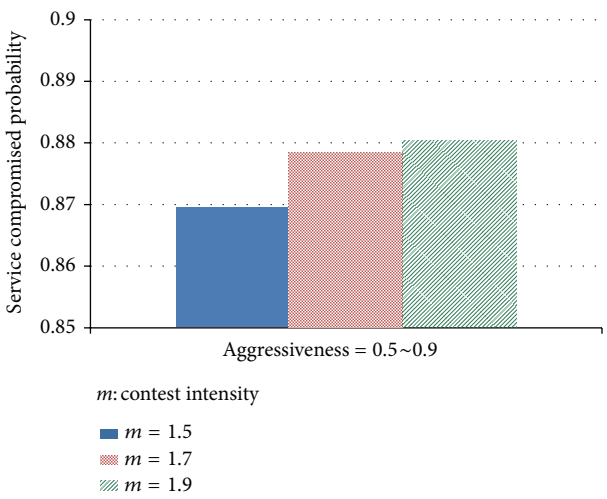


Figure 14: Service compromised probability of aggressive attacker under $m < 1$ circumstance.

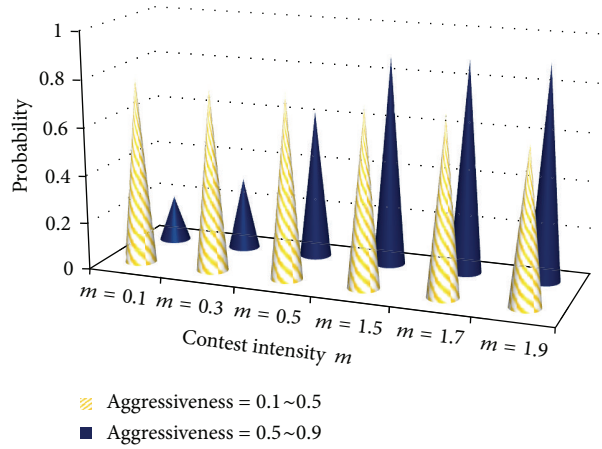


Figure 15: Service compromised probability of aggressive attacker under $m > 1$ circumstance.

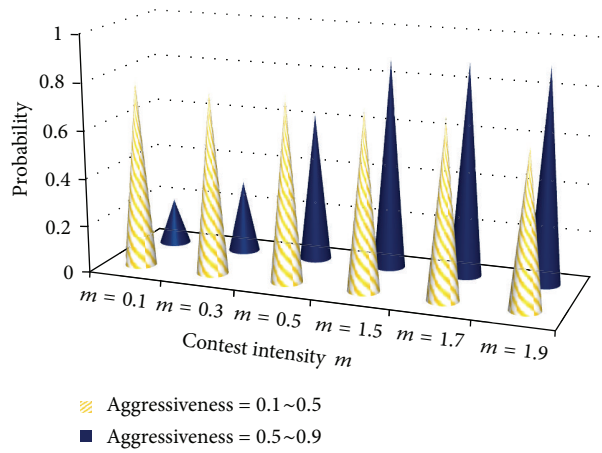


Figure 16: Comparison of service compromised probability under diverse contest intensity and aggressiveness.

compromised probability slows down when contest intensity becomes higher. This is because although aggressive attackers prefer to gain an edge by spending more attack resources to compromise a target, each compromise costs significant resources. Thus, aggressive attackers run out of budget very quickly, and the increasing trend is not exponential.

4.2. Proactive Defense Is Advantageous under "Winner Takes All" Circumstance. According to previous discussions, the effectiveness of resources invested is significant under "winner takes all" circumstances. For less aggressive attackers, the performance is poor since they only invest few resources on an attack. Aggressive attackers tend to spend a larger quantity to compromise a target. If a defender raises the quantity of defense resources on important nodes, it effectively weakens attackers.

Consequently, regardless of whether defenders face either aggressive or less aggressive attackers, a proactive defense performs better in deterring attackers.

4.3. Reactive Defense Is Advantageous under "Fight to Win or Die" Circumstance. Similarly, under the "fight to win

or die" circumstance the effectiveness of defense resources is insignificant. No matter how many proactive defense resources are invested, there is only limited influence on the objective function value. In other words, less aggressive attackers can compromise a target even by only investing scarce attack resources. For aggressive attackers, since they prefer spending a large quantity of resources on compromising a victim, they easily run out of resources before a target service is compromised.

Based on this analysis, it is advantageous for the defender to apply reactive defense mechanisms against malicious attackers under "fight to win or die" circumstance.

5. Conclusions

This paper models an attack and defense scenario that involves high amounts of randomness as mathematical formulations where attackers are assumed to have incomplete information. It further considers a virtualization environment for helping relate these results to recent trends in cloud computing.

According to the simulation results, the outcome of a contest is influenced not only by the quantity of defense resources invested on each node but also by the contest intensity. An attacker's aggressiveness is introduced as a new dimension, and meaningful results are discovered. Effective defense strategies are proposed in the discussion of the results.

For future works, collaborative attacks should be taken into consideration since the attack strategy discussed in this paper is for individual attacks. In other words, there is only one attack targeting a victim's network at a time. Thus, no synergy is considered. A more complicated collaborative attack pattern is worth further study.

Acknowledgment

This work was supported by the National Science Council, Taiwan (Grant nos. NSC 102-2221-E-002-104 and NSC 101-2218-E-011-009).

References

[1] IBM Internet Security Systems X-Force research and development team, "IBM X-Force 2010 Mid-Year Trend and Risk Report," *IBM*, August 2010.

[2] R. J. Ellison, D. A. Fisher, R. C. Linger, H. F. Lipson, T. Longstaff, and N. R. Mead, "Survivable network systems: an emerging discipline," Tech. Rep. CMU/SEI-97-TR-013, 1997.

[3] S. Roy, C. Ellis, S. Shiva, D. Dasgupta, V. Shandilya, and Q. Wu, "A survey of game theory as applied to network security," in *Proceedings of the 43rd Annual Hawaii International Conference on System Sciences (HICSS '10)*, January 2010.

[4] M. N. Lima, A. L. D. Santos, and G. Pujolle, "A survey of survivability in mobile Ad hoc Networks," *IEEE Communications Surveys and Tutorials*, vol. 11, no. 1, pp. 66–77, 2009.

[5] Z. Ma, "Towards a unified definition for reliability, survivability and resilience (I): the conceptual framework inspired by the handicap principle and ecological stability," in *Proceedings of the IEEE Aerospace Conference*, pp. 1–12, March 2010.

[6] F. Xing and W. Wang, "On the survivability of wireless ad HOC networks with node misbehaviors and failures," *IEEE Transactions on Dependable and Secure Computing*, vol. 7, no. 3, pp. 284–299, 2010.

[7] S. Skaperdas, "Contest success functions," *Economic Theory*, vol. 7, no. 2, pp. 283–290, 1996.

[8] G. Levitin and K. Hausken, "False targets efficiency in defense strategy," *European Journal of Operational Research*, vol. 194, no. 1, pp. 155–162, 2009.

[9] K. Hausken and G. Levitin, "Protection vs. false targets in series systems," *Reliability Engineering and System Safety*, vol. 94, no. 5, pp. 973–981, 2009.

[10] G. Levitin and K. Hausken, "Preventive strike vs. false targets and protection in defense strategy," *Reliability Engineering and System Safety*, vol. 96, no. 8, pp. 912–924, 2011.

[11] J. Hirshleifer, "Conflict and rent-seeking success functions: ratio vs. difference models of relative success," *Public Choice*, vol. 63, no. 2, pp. 101–112, 1989.

[12] J. Hirshleifer, "The paradox of power," *Economics and Politics*, vol. 3, pp. 177–200, 1993.

[13] J. Archer, A. Boehme, D. Cullinane, P. Kurtz, N. Puhlmann, and J. Reavis, "Top Threats to Cloud Computing V 1.0," *Cloud Security Alliance*, March 2010.

[14] H. Debar, F. Pouget, and M. Dacier, "White paper: 'Honeypot, Honeynet, Honeytoken: Terminological issues,'" Institut Eurécom Research Report RR-03-081, 2003.

[15] B. Cheswick, "An evening with berferd in which a cracker is lured, endured, and studied," in *Proceedings of the USENIX Conference*, pp. 163–174, USENIX, 1992.

[16] C. Seifert, I. Welch, and P. Komisarczuk, "Taxonomy of honeypots," Tech. Rep. CS-TR-06/12, 2006.

[17] M. H. y López and C. F. L. Reséndez, "Honeypots: basic concepts, classification and educational use as resources in information security education and courses," in *Proceedings of the Informing Science and IT Education Conference*, 2008.

[18] Y. Huang, D. Arsenault, and A. Sood, "Closing cluster attack windows through server redundancy and rotations," in *Proceedings of the 6th IEEE International Symposium on Cluster Computing and the Grid (CCGRID '06)*, May 2006.

[19] Y. Huang, D. Arsenault, and A. Sood, "Incorruptible self-cleansing intrusion tolerance and its application to DNS security," *Journal of Networks*, vol. 1, no. 5, pp. 21–30, 2006.

[20] M. Smith, C. Schridde, and B. Freisleben, "Securing stateful grid servers through virtual server rotation," in *Proceedings of the 17th International Symposium on High Performance Distributed Computing (HPDC '08)*, pp. 11–22, June 2008.

[21] F. Y.-S. Lin, Y.-S. Wang, and P.-H. Tsang, "Efficient defense strategies to minimize attackers' success probabilities in honeynet," in *Proceedings of the 6th International Conference on Information Assurance and Security (IAS '10)*, pp. 80–85, August 2010.

[22] F. Y.-S. Lin, Y.-S. Wang, P.-H. Tsang, and J.-P. Lo, "Redundancy and defense resource allocation algorithms to assure service continuity against natural disasters and intelligent attacks," in *Proceedings of the 5th International Conference on Broadband Wireless Computing, Communication and Applications (BWCCA '10)*, pp. 206–213, November 2010.

[23] F. Cohen, "Managing network security: attack and defence strategies," *Network Security*, vol. 1999, no. 7, pp. 7–11, 1999.

[24] S. Nagaraja and R. Anderson, "Dynamic topologies for robust scale-free networks," *Bio-Inspired Computing and Communication*, vol. 5151, pp. 411–426, 2008.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 930316, 10 pages
http://dx.doi.org/10.1155/2013/930316

Research Article

Efficient Periodic Broadcasting for Mobile Networks at Small Client Receiving Bandwidth and Buffering Space

Hsiang-Fu Yu,[1] Yao-Tien Wang,[2] Jong-Yih Kuo,[3] and Chu-Yi Chien[1]

[1] *Department of Computer Science, National Taipei University of Education, Taipei 10671, Taiwan*
[2] *Department of Computer Science and Information Engineering, Hungkuang University, Taichung 43302, Taiwan*
[3] *Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei 10608, Taiwan*

Correspondence should be addressed to Hsiang-Fu Yu; yu@tea.ntue.edu.tw

Received 7 January 2013; Accepted 18 March 2013

Academic Editor: Chih-Hao Lin

Periodic broadcasting is an effective approach for delivering popular videos. In general, this approach does not provide interactive (i.e., VCR) functions, and thus a client can tolerate playback latency from a video server. The concept behind the approach is partitioning a video into multiple segments, which are then broadcast across individual communication channels in terms of IP multicast. The method improves system throughput by allowing numerous clients to share the channels. For many broadcasting schemes, client receiving bandwidth must equal server broadcasting bandwidth. This limitation causes these schemes to be infeasible in mobile networks because increasing receiving bandwidth at all client sites is expensive, as well as difficult. To alleviate this problem, the fibonacci broadcasting (FiB) scheme allows a client with only two-channel bandwidth to receive video segments. In comparison with other similar schemes, FiB yields smallest waiting time. Extending FiB, this work proposes a new scheme (called FiB+) to achieve smaller client buffering space and the same waiting time under two-channel receiving bandwidth. Extensive analysis shows that FiB+ can yield 34.5% smaller client buffer size than that of FiB. Further simulation results also indicate that FiB+ requires lower client buffering space than several previous schemes.

1. Introduction

Video-on-demand (VOD) services have become popular due to advances in network and computer technology [1, 2]. A VOD system may easily run out of bandwidth since the growth in bandwidth can never keep up with the growth in the number of clients. To alleviate the problem, one way is to simply broadcast popular videos. According to the study in [3, 4], a few very popular videos constitute most client requests. Data broadcasting is thus suitable to transfer popular videos that may be interesting to many clients in a particular period of time. An efficient method for broadcasting a popular video is to divide it into segments, which are simultaneously and periodically transmitted across individual communication channels in terms of IP multicast [5]. Because video broadcasting does not provide VCR functions, a client is able to tolerate playback latency. To ensure continuous playback, clients must simultaneously download and save the video segments from these channels. The clients usually have to wait for the occurrence of the first segment before they can start playing the video. Since the clients cannot watch the video immediately, the broadcasting schemes provide near VOD services.

The fast broadcasting (FB) scheme [6] improves segment partition and arrangement to yield shorter waiting time. To achieve near-minimum waiting time, the recursive frequency-splitting (RFS) scheme [7] broadcasts each segment at the frequency that can keep continuous video playback. A scalable binomial broadcasting scheme [8] transfers a variable-length video using constant bandwidth. To simplify the implementation of multiple channels, the PAS scheme [9] broadcasts video segments over a single channel. The reverse-order scheduling (ROS) scheme [10] transmits segments of the same group in reverse order over a single channel to save buffering space.

With the fast growth of wireless networks, mobile video services become more and more popular. Broadcasting videos under rather restricted client resources is increasingly important. The following schemes address the savings on client buffer size and bandwidth. Modifying the FB scheme [6], the reverse fast broadcasting (RFB) scheme [11] buffers 25% of video size, just half of what is required by FB. By combining RFS and RFB, the hybrid broadcasting scheme (HyB) [12] achieves small client buffering space and waiting time. Different RFB-based hybrid schemes were proposed in [13, 14]. The skyscraper broadcasting (SkB) scheme [15] allows a client to download video data using only a bandwidth of two channels. The client-centric approach (CCA) [16] also permits a client downloading video data via a small number of channels, and CCA+ [17] further yields smaller waiting time than SkB. Like SkB and CCA+, the fibonacci broadcasting (FiB) scheme [18] supports a client with two-channel bandwidth but achieves the minimum waiting time. The authors in [19] proposed an FB-based scheme for heterogeneous clients. The studies in [20, 21] deploy a proxy in VoD systems to serve heterogeneous clients.

The contributions of this study are summarized as follows.

(1) Extending FiB, this work proposes a promising scheme, called FiB+, to deliver near VOD services to clients with small receiving bandwidth and buffering space. In comparison with FiB, FiB+ still yields the minimum waiting time under two-channel receiving bandwidth; moreover, FiB+ can save about 34.5% of buffer size.

(2) The paper investigates the total client buffer requirements for FiB+ and explains why this scheme requires smaller buffering space than FiB. We further derive the maximum number of segments buffered by an FiB+ client mathematically. Extensive performance analysis has been conducted on FiB+ by comparing a number of past reported counterparts. The results indicate that FiB+ yields relatively lower client buffer requirements than most schemes.

The remainder of this study is organized as follows. The FiB scheme is introduced in Section 2. Section 3 presents FiB+. This section also verifies the on-time video delivery under two-channel client bandwidth. Section 4 evaluates the performance of FiB+. Brief conclusions are drawn in Section 5.

2. Review of FiB

Let k be the number of server channels throughout the paper. The FiB scheme [18] unequally divides a video of length L into k segments, denoted by $S_1, S_2, \ldots, S_k$ in sequence. The length of a segment S_i is based on the following equation and equals $Ln_i/\sum_{p=1}^{k} n_p$ as follows:

$$n_i = \begin{cases} 1, & i = 1 \\ 2, & i = 2 \\ n_{i-1} + n_{i-2}, & 3 \le i \le k. \end{cases} \quad (1)$$

Table 1: List of terms used in the proposed scheme and their respective definitions.

Term	Definition
L	Video length
k	Number of broadcasting channels for each video on the server side
C_i	ith broadcasting channel, $i = 1, \ldots, k$
N	Number of segments of a video
S_i	ith video segment, $i = 1, \ldots, N$
n_i	Number of segments transferred on channel C_i
m_i	Total segments transferred on channels C_1 through C_i, $m_i = \sum_{p=1}^{i} n_p$
b	Video playback rate assumed to equal the data transmission rate of each channel
T_0	Starting time to watch the first segment
Time unit	Basic unit on time axis, whose length equals the length of a segment (i.e., L/N)

Assume that the data transmission rate of each channel equals the playback rate b. The server then periodically broadcasts segment S_i on channel C_i, as illustrated in Figure 1. In the figure, segments downloaded and played by a client are gray. When a client wants to watch a video, the client first downloads segments S_1 and S_2 on the first two channels C_1 and C_2. Once finishing receiving the segment S_1, the client continuously accepts segment S_2 and newly downloads segment S_3 on channel C_3. The client repeats the process, which starts downloading segment S_i on channel C_i once finishing receiving segment S_{i-2} from channel C_{i-2}, until all the segments are loaded. An FiB client thus requires a bandwidth of only two channels to download video segments.

3. FiB+

Some of the frequently used terms and their definitions are listed in Table 1. On the server side, the FiB+ scheme includes the following steps.

(1) The server equally divides a video into N segments, denoted by $S_1, S_2, \ldots, S_N$ in sequence, where $N = \sum_{p=1}^{k} n_p$, where n_p is based on (1). The length of each segment thus equals L/N. From (1), we further yields

$$m_i = \sum_{p=1}^{i} n_p = n_{i+2} - 2. \quad (2)$$

See Appendix A for details. Thus, $N = n_{k+2} - 2$. The FiB+ scheme then assembles segments $S_{n_{i+1}-1}$ to $S_{n_{i+2}-2}$ (i.e., $S_{m_{i-1}+1}$ to S_{m_i}) into group G_i sequentially, as illustrated in Figure 2(a). The number of segments of group G_i thus equals n_i. For instance, group G_4 includes segments S_7 to S_{11}, and $n_4 = 5$.

(2) Channel C_i, where $1 \le i \le k-2$, periodically broadcasts the segments of group G_i in sequence, as shown in Figure 2(b). That is, segments $S_{n_{i+1}-1}$ to $S_{n_{i+2}-2}$ are transferred one by one on channel C_i.

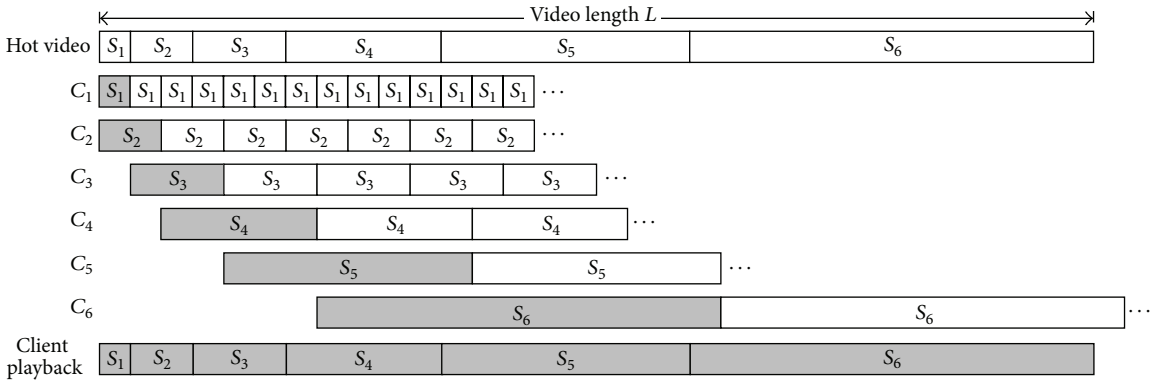


Figure 1: Illustration of segment downloading for the FiB scheme, where $k = 6$.

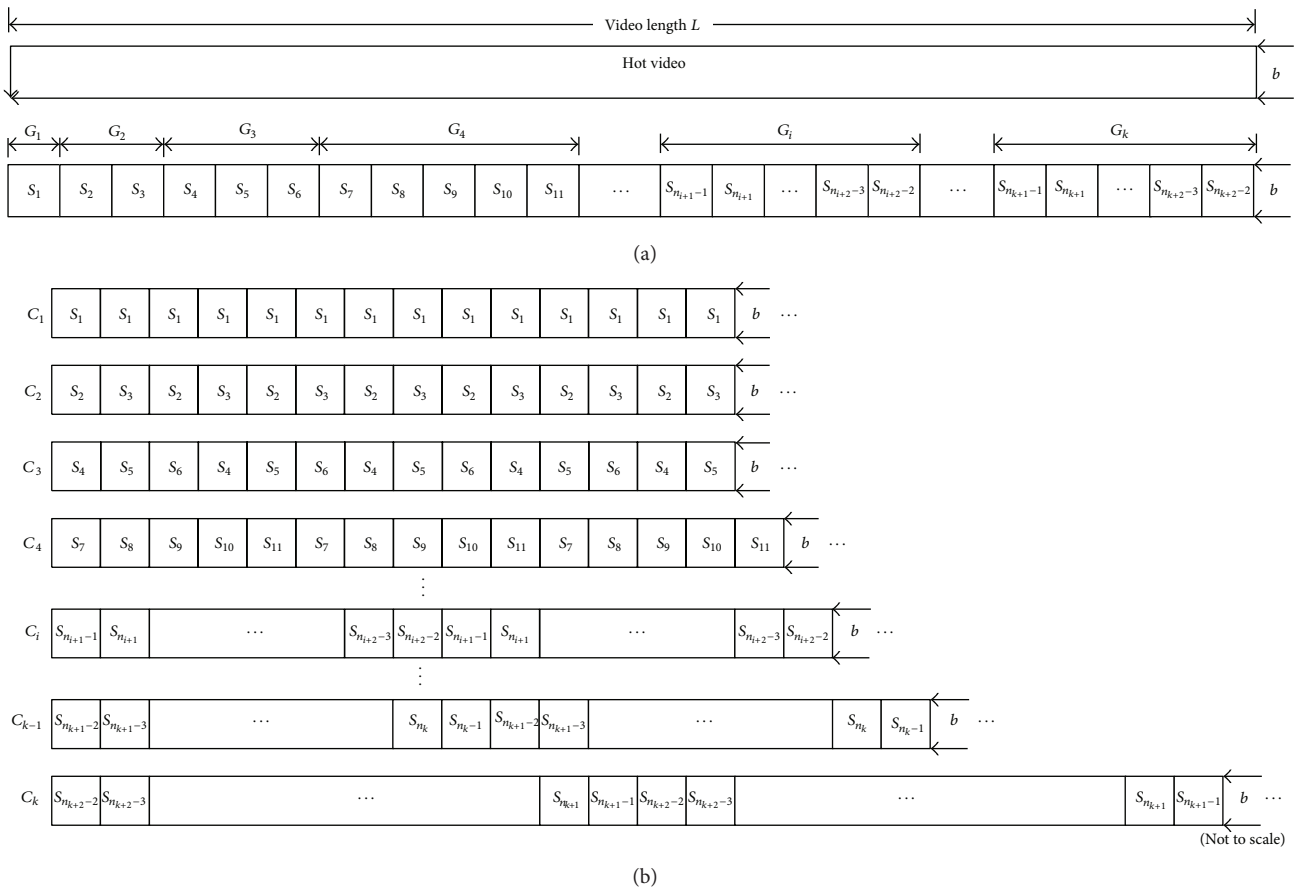


Figure 2: Channel allocation for the FiB+ scheme.

(3) The segments of groups G_{k-1} and G_k are cyclically transmitted on channels C_{k-1} and C_k in reverse order, respectively. Figure 2(b) shows that the scheme repeatedly broadcasts the segments of group G_{k-1} in the order of $S_{n_{k+1}-2}$ to S_{n_k-1} and the segments of group G_k in the order of $S_{n_{k+2}-2}$ to $S_{n_{k+1}-1}$.

Figure 3 demonstrates the segment broadcasting and downloading for FiB+, where the segments downloaded and played by a client are gray. Let T_0 be the time that the client starts receiving video segments and be the origin (i.e., the first time unit) of the time axis. Due to $k = 6$, FiB+ equally divides a video into 32 segments, which are then classified into six groups. The segments of groups G_1 to G_4 are broadcast sequentially on channels C_1 to C_4, respectively. In addition, FiB+ transmits segments of groups G_5 and G_6 on channels C_5 and C_6 in reverse order.

A client is assumed to have enough buffers to store video segments downloaded. We further suppose that one time unit equals the length of a segment throughout this study. Playing a video on the client side includes the following steps.

(1) Download segments of group G_i on channel C_i during time units n_{i-1} to $n_{i-1} + n_i - 1 = n_{i+1} - 1$, where

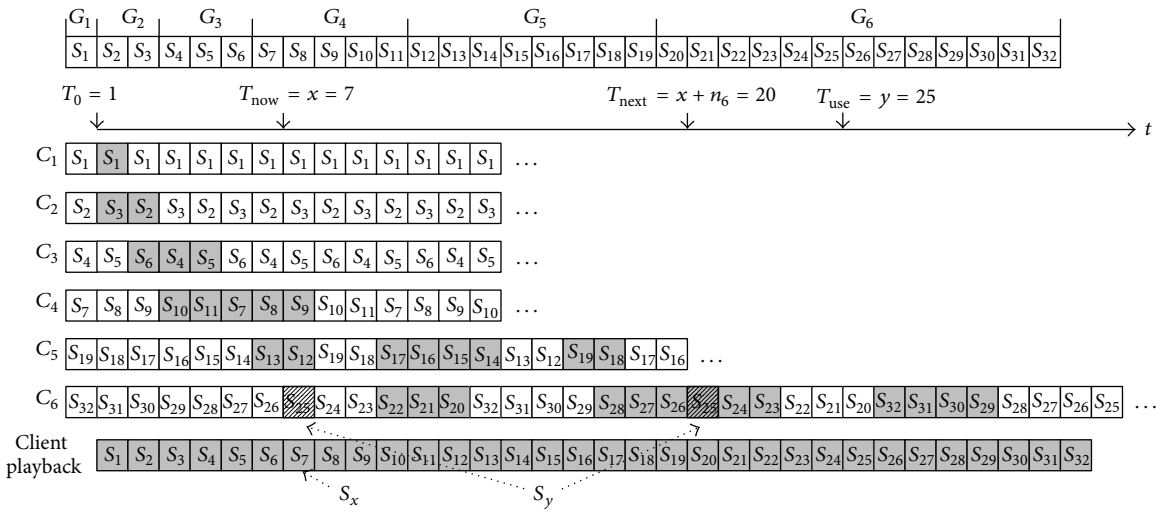


Figure 3: Illustration of segment broadcasting and downloading for the FiB+ scheme, where $k = 6$ and $N = 32$.

$1 \le i \le k-2$ and $n_0 = 1$. For example, the client accepts segments from channel C_4 during time units $n_{4-1} = 3$ to $n_{4+1} - 1 = 7$, as shown in Figure 3.

(2) The paper next presents how a client receives segments from channels C_{k-1} and C_k. (In Figure 3, refer to channels C_5 and C_6.) Suppose that a client first sees a segment S_y on a channel C_i at time T_{now} and sees the next segment S_y at time T_{next}, where $i = k-1$ or $i = k$. (In Figure 3, $i = 6$ and $y = 25$.) The client is also assumed to play segments S_x and S_y at time T_{now} and T_{use}, respectively. Clearly, if $T_{next} \le T_{use}$, the client can delay downloading segment S_y at time T_{now}, without interrupting the playback. Substituting $T_{use} = y$th time unit, $T_{now} = x$th time unit, and $T_{next} = T_{now} + n_i$ into $T_{next} \le T_{use}$, we obtain

$$x + n_i \le y. \tag{3}$$

If the inequality is true, the client does not receive segment S_y at time T_{now}, otherwise, performs the downloading immediately. For instance, when the client first sees segment S_{25} with only diagonal lines on channel C_6 at the 7th time unit (i.e., T_{now}) in Figure 3, (3) is true, $7 + 13 \le 25$, and the client does not download the segment. Afterwards, the client sees next segment S_{25} with gray color and diagonal lines at the 20th time unit. The client must receive the segment because (3) does not hold, $20 + 13 > 25$.

(3) The client plays the video in the order of $S_1, S_2, \ldots, S_N$ at time T_0.

(4) The client stops loading data from networks when all the segments have been received.

FiB+ and FiB differ in three areas.

(i) *Equal-length segment partition versus variable-length segment partition*. FiB+ divides a video into multiple equal-length segments, while FiB partitions a video into variable-length segments. For example, given $k = 6$, FiB divides a video into six segments, whose lengths are $L/32$, $2L/32$, $3L/32$, $5L/32$, $8L/32$, and $13L/32$. On the other hand, FiB+ partitions a video into 32 segments, whose lengths all equal $L/32$.

(ii) *Multiple segments on each channel versus single segment*. FiB+ cyclically broadcasts several segments on each channel except the first channel, and FiB transmits only one.

(iii) *Segment transmission in reverse order*. The FiB+ scheme broadcasts segments on the last two channels in reverse order. For example, the scheme transmits segments S_{19} to S_{12} on channel C_5 and segments S_{32} to S_{20} on channel C_6, as illustrated in Figure 3.

3.1. Analysis of Segment Playing and Downloading on a Single Channel. We next analyze the segment downloading on channel C_i, where $1 \le i \le k$.

For $1 \le i \le k-2$, a client receives segments $S_{n_{i+1}-1}$ to $S_{n_{i+2}-2}$ from channel C_i during time units n_{i-1} to $n_{i+1}-1$ and plays the segments during time units $n_{i+1} - 1$ to $n_{i+2} - 2$, as mentioned previously. Suppose that a client sees a segment S_j at the $(n_{i+1}-1)$th time unit, where $n_{i+1}-1 \le j \le n_{i+2}-2$. Figure 4(a) shows how a client downloads and plays segments, where the segments downloaded and played by the client are gray.

For $i = k-1$ or k, a client downloads segments according to (3). We also assume that a client sees a segment S_j at the $(n_{i+1}-1)$th time unit, where $n_{i+1}-1 \le j \le n_{i+2}-2$. A complete segment-downloading diagram for channel C_i is based on (3), as indicated in Figure 4(b). The explanation is as follows.

The client always downloads segment S_j since the inequality of (3) does not hold for $x = n_{i+1} - 1$ and $y = j$ (i.e., $n_{i+1} - 1 + n_i = n_{i+2} - 1 > j$). In addition, because the segments of group G_i are transmitted once on channel C_i every n_i time units and are played during time units $n_{i+1} - 1$

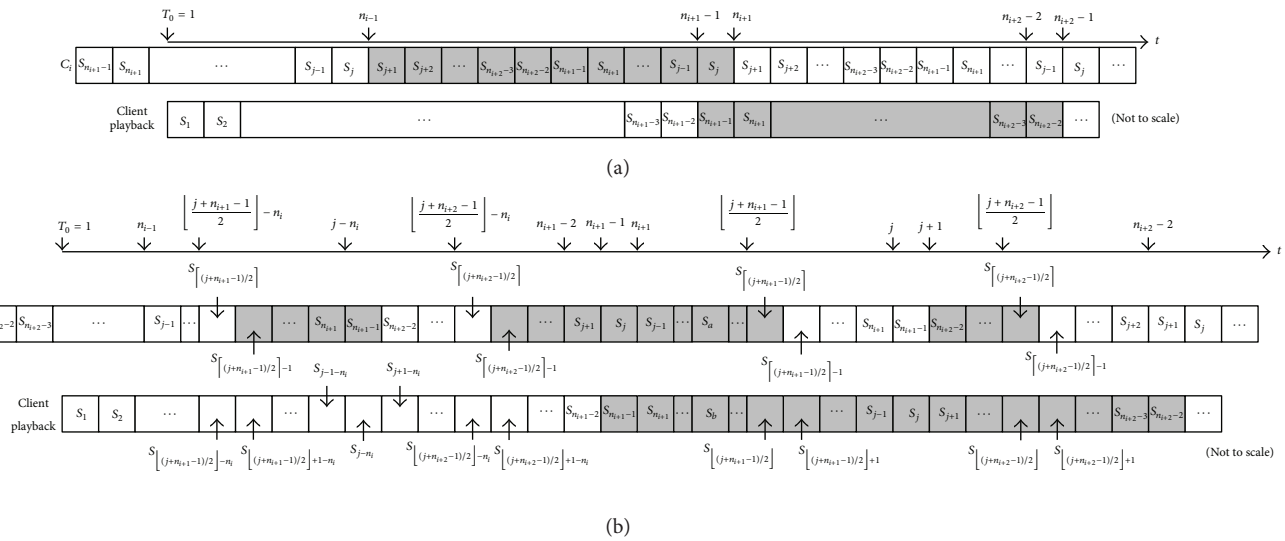


Figure 4: Segment downloading on channel C_i for FiB+.

to $n_{i+2}-2$, the client downloads a segment of group G_i either during time units $n_{i+1}-1$ to $n_{i+2}-2$ or during time units n_{i-1} to $n_{i+1}-2$ (i.e., before the downloading of segment S_j).

3.1.1. Segment Downloading within $[n_{i+1}-1, n_{i+2}-2]$. Figure 4(b) shows that segments $S_{n_{i+1}-1}$ to S_j are broadcast on channel C_i in descending order during time units $n_{i+1}-1$ to j, while a client plays these segments in turn. Let S_a be a segment broadcast on channel C_i during this period, and let S_b be the client-playback segment corresponding to S_a. Clearly, if $a \geq b$, a client must download segment S_a to ensure continuous playing. Because the number of segments between S_j and S_a on channel C_i equals the number of segments between $S_{n_{i+1}-1}$ and S_b on the client playback, $j-a=b-(n_{i+1}-1)$ and $a=j+n_{i+1}-1-b$. Substituting this equation to $a \geq b$ yields $(j+n_{i+1}-1)/2 \geq b$. The maximum value of b is $\lfloor (j+n_{i+1}-1)/2 \rfloor$, and the corresponding value of a equals $j+n_{i+1}-1-\lfloor (j+n_{i+1}-1)/2 \rfloor = \lceil (j+n_{i+1}-1)/2 \rceil$. Figure 4(b) illustrates that the client downloads segments S_j to $S_{\lceil (j+n_{i+1}-1)/2 \rceil}$ during time units $n_{i+1}-1$ to $\lfloor (j+n_{i+1}-1)/2 \rfloor$ and does not download any segment during time units $\lfloor (j+n_{i+1}-1)/2 \rfloor + 1$ to j. Similarly, this study obtains that a client receives segments $S_{n_{i+2}-2}$ to $S_{\lceil (j+n_{i+2}-1)/2 \rceil}$ during time units $j+1$ to $\lfloor (j+n_{i+2}-1)/2 \rfloor$ and does not download any segment during time units $\lfloor (j+n_{i+2}-1)/2 \rfloor + 1$ to $n_{i+2}-2$.

3.1.2. Segment Downloading within $[n_{i-1}, n_{i+1}-2]$. From (3), the client must download segments $S_{\lceil (j+n_{i+2}-1)/2 \rceil - 1}$ to S_{j+1} during time units $\lfloor (j+n_{i+2}-1)/2 \rfloor + 1 - n_i$ to $n_{i+2}-2-n_i = n_{i+1}-2$ because the client does not download these segments during time units $\lfloor (j+n_{i+2}-1)/2 \rfloor + 1$ to $n_{i+2}-2$, as shown in Figure 4(b). The figure further indicates that the client does not accept segments $S_{n_{i+2}-2}$ to $S_{\lceil (j+n_{i+2}-1)/2 \rceil}$ during time units $j+1-n_i$ to $\lfloor (j+n_{i+2}-1)/2 \rfloor - n_i$ since the client will perform their downloading during time units $j+1$ to $\lfloor (j+n_{i+2}-1)/2 \rfloor$. Similarly, the client must receive segments $S_{\lceil (j+n_{i+1}-1)/2 \rceil - 1}$ to $S_{n_{i+1}-1}$ during time units $\lfloor (j+n_{i+1}-1)/2 \rfloor + 1 - n_i$ to $j-n_i$ because the client does not download these segments during time units $\lfloor (j+n_{i+1}-1)/2 \rfloor + 1$ to j. Furthermore, since the client will download segments S_{j-1} to $S_{\lceil (j+n_{i+1}-1)/2 \rceil}$ during time units n_{i+1} to $\lfloor (j+n_{i+1}-1)/2 \rfloor$, the client does not load any segment during time units $n_{i+1}-n_i = n_{i-1}$ to $\lfloor (j+n_{i+1}-1)/2 \rfloor - n_i$, as illustrated in Figure 4(b).

3.2. Workable Verification. This section shows that FiB+ guarantees continuous playback and two-channel bandwidth demand on the client side.

3.2.1. Continuous Playback. To keep on-time video delivery, the study in [7] indicates that a video server must broadcast a segment S_j on a channel C_i at least once in every j time units. For FiB+, a server transmits a segment S_j once every n_i time units, where $n_{i+1}-1 \leq j \leq n_{i+2}-2$. This paper thus needs to prove $j \geq n_i$. We then evaluate

$$\begin{aligned} j-n_i &\geq (n_{i+1}-1)-n_i, \quad \text{due to } n_{i+1}-1 \leq j \\ &= n_{i-1}-1 \\ &\geq 0. \end{aligned} \tag{4}$$

For FiB+, the segment broadcasting frequency is large enough to let clients receive video data in time.

3.2.2. Two-Channel Bandwidth Demand. From the previous analysis in Figure 4, we make a temporal-channel map of segment downloading for each channel, as indicated in Figure 5. In this figure, segments downloaded and played by a client are gray. This work divides client playback time t (in terms of time units) into multiple successive durations for ease of explanation.

For $T_0 \leq t \leq n_{k-2}-1$, the client merely receives segments from channels C_1 to C_{k-2} because the client starts the segment

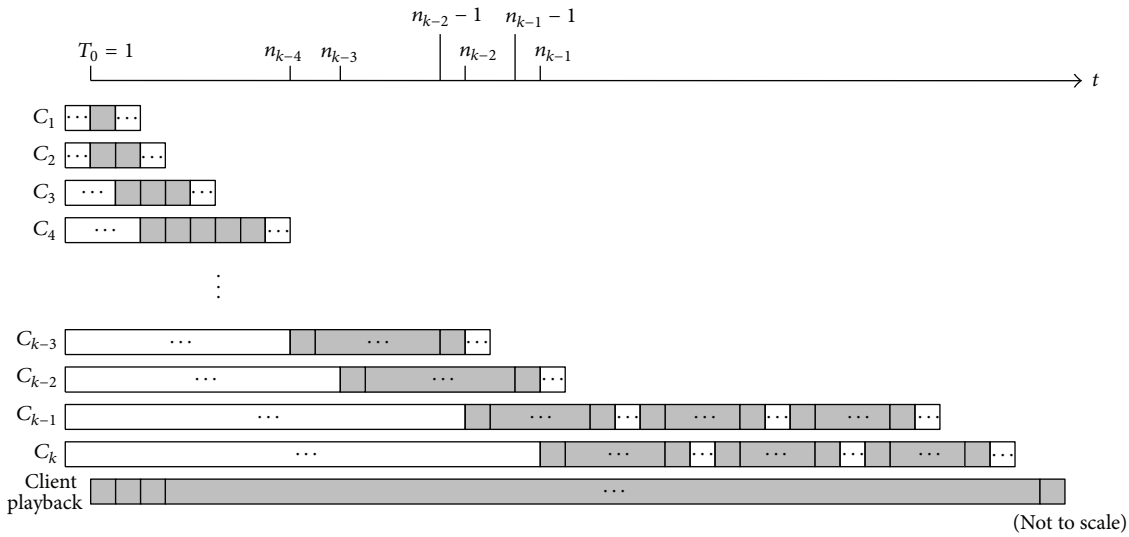


FIGURE 5: Segment downloading using client bandwidth $2b$.

downloading on channel C_{k-1} at time unit n_{k-2}, as shown in Figure 5. According to Step 1 of segment downloading on the client side, the client uses two-channel bandwidth to load segments in this period.

For $n_{k-2} \le t \le n_{k-1} - 1$, the client receives segments only from channels C_{k-2} and C_{k-1} because the client finishes receiving segments from channel C_1 to C_{k-3} before time unit n_{k-2} and starts downloading segments on channel C_k at time unit n_{k-1}, as indicated in Figure 5.

For $n_{k-1} \le t$, the client simply receives the remaining segments from channels C_{k-1} and C_k because the segment downloading on channel C_{k-2} completes at time unit $n_{k-1} - 1$.

Accordingly, an FiB+ client can download segments using two-channel bandwidth.

4. Performance Analysis and Comparison

When a client just misses segment S_1 of a requested video, the maximum waiting time δ equals the access time of the segment from the first channel. Thus, $\delta = L/N = L/\sum_{p=1}^{k} n_p$, the same as that of FiB. According to the previous studies [15, 17], this work has calculated the values of N offered by these schemes at different numbers of channels, as listed in Table 2. The larger the value is, the smaller the waiting time is. The table reveals that FiB and FiB+ yield far bigger values than other schemes. Figure 6 shows maximum waiting time versus server channels. FiB and FiB+ thus achieve much smaller waiting time than SkB and CCA+ under two-channel client bandwidth. For example, when the server bandwidth equals 10 channels, FiB and FIB+ reduce the broadcast latency to less than 32 seconds. By contrast, SkB and CCA+ incur more than 51 and 48 seconds, respectively.

Before analyzing the entire buffer requirements, we first investigate the number of the buffered segments when a client performs segment downloading on a single channel C_i.

Lemma 1. *Let $B(i,t)$ be the function of the number of the segments buffered by an FiB+ client on channel C_i at the tth time unit.*

For $1 \le i \le k-2$,

$$\begin{cases} B(i,t) = 0, & t < n_{i-1}, \\ B(i,t) = t - n_{i-1} + 1, & n_{i-1} \le t \le n_{i+1} - 2, \\ B(i,t) = n_{i+2} - 2 - t, & n_{i+1} - 2 < t \le n_{i+2} - 2, \\ B(i,t) = 0, & n_{i+2} - 2 < t. \end{cases} \tag{5a}$$

For $i = k-1$ or k,

$$\begin{cases} B(i,t) = 0, & t < n_{i-1}, \\ B(i,t) \le \left\lfloor \dfrac{t - n_{i-1} + 2}{2} \right\rfloor, & n_{i-1} \le t \le n_{i+1} - 2, \\ B(i,t) \le \left\lfloor \dfrac{n_i}{2} \right\rfloor, & n_{i+1} - 2 < t \le n_{i+2} - 2 - \left\lceil \dfrac{n_i}{2} \right\rceil, \\ B(i,t) \le n_{i+2} - 2 - t, & n_{i+2} - 2 - \left\lceil \dfrac{n_i}{2} \right\rceil < t \le n_{i+2} - 2, \\ B(i,t) = 0, & n_{i+2} - 2 < t. \end{cases} \tag{5b}$$

Proof. See Appendix B. □

Equation (5b) shows that a client buffers at most $\lfloor n_i/2 \rfloor$ segments from channel C_i for $i = k-1$ or k. On the other hand, an FiB client needs to buffer $n_i - 1$ segments [18]. Such a difference leads to the result that FiB+ requires much smaller buffering space.

Theorem 2. *Let $B(t)$ be the maximum number of segments buffered by an FiB+ client. Then, $B(t) \le \lceil n_{k-1}/4 \rceil + \lfloor n_k/2 \rfloor$.*

Proof. See Appendix C. □

Due to $\lim_{i\to\infty}(n_{i+1}/n_i) \approx ((1+\sqrt{5})/2)$ [22] and $N = n_{k+2} - 2$, $\lim_{k\to\infty}((\lceil n_{k-1}/4 \rceil + \lfloor n_k/2 \rfloor)/N) \approx 0.25$. This result

Table 2: The values of N offered by different schemes.

k	1	2	3	4	5	6	7	8	9	10
FiB/FiB+	1	3	6	11	19	32	53	87	142	231
SkB	1	3	5	10	15	27	39	64	89	141
CCA+2	1	3	5	10	15	27	39	64	89	149

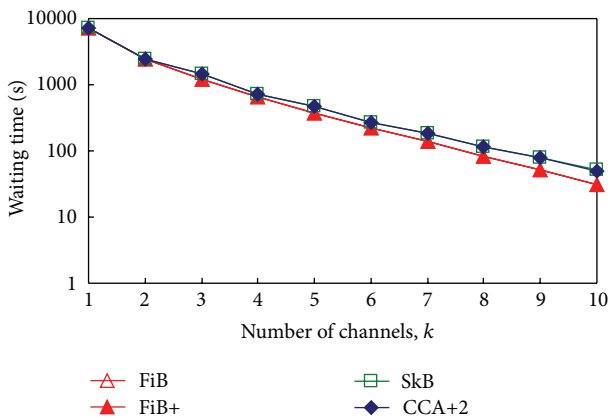


Figure 6: Maximum waiting time incurred on new clients at different numbers of channels ($L = 120$ minutes).

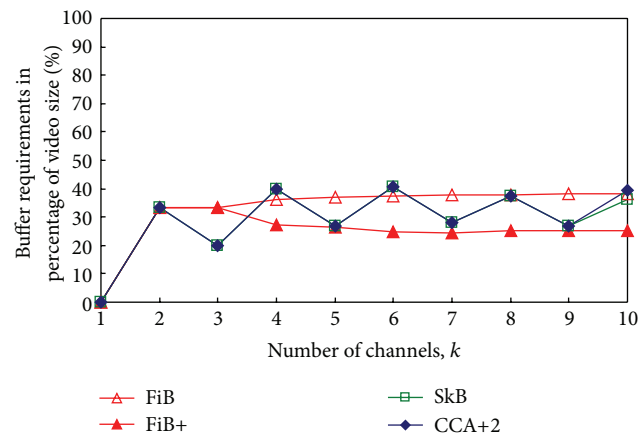


Figure 7: Comparison of required buffers.

indicates that an FiB+ client can buffer only 25% of video size, like RFB. We used the Perl language to implement a simulator, which could estimate the buffer requirements for FiB+. The results are listed in Table 3. The buffer size required by FiB+ is quite close to the bound when $k \geq 6$. Because FiB has to buffer $n_k - 1$ segments, we can derive the buffer reduction rate of FiB+ versus FiB as follows: $\lim_{k\to\infty}(1 - (\lceil n_{k-1}/4\rceil + \lfloor n_k/2\rfloor)/(n_k - 1)) \approx 0.345$. FiB+ can reduce buffer requirements by 34.5%, when compared with FiB. Table 3 shows that with the growth of k, the reduction rate is close to the bound. For example, for $k = 10$, an FiB client buffers 38.1% of video size. By contrast, an FiB+ client buffers only 25.1%. The reduction rate is 34.1%. According to the previous studies [15, 17], this work presents the buffer sizes required by different broadcasting schemes at different numbers of server channels, as indicated in Figure 7. For $k > 3$, the FiB+ scheme outperforms all the schemes.

5. Conclusions

With the advance of mobile computing technology, many clients access VOD services through their mobile devices. Delivering videos under rather restricted client resources is increasingly important. To fulfill this requirement, several schemes, such as SkB, FiB, and CCA+, are proposed to allow a client to watch a video using two-channel bandwidth. Extending FiB, this work devises FiB+, which exhibits the merits of small client waiting time and buffering space. The scheme still guarantees on-time video delivery under two-channel receiving bandwidth. According to the performance analysis, FiB+ yields the minimum waiting time and requires smaller client buffer size, when compared with most existing schemes.

Appendices

A. Proof of Equation (2)

For $i \bmod 2 = 1$, let $i = 2q + 1$. Then,

$$\begin{aligned}\sum_{p=1}^{i} n_p &= n_1 + n_2 + n_3 + n_4 + \cdots + n_i \\ &= n_3 + n_5 + \cdots + n_{2q+1} + n_{2q+1}, \quad \text{from (1)} \\ &= n_2 + n_3 + n_5 + \cdots + n_{2q+1} + n_{2q+1} - n_2 \\ &= n_4 + n_5 + \cdots + n_{2q+1} + n_{2q+1} - n_2, \quad \text{from (1)} \\ &= n_{2q+2} + n_{2q+1} - n_2, \quad \text{from (1)} \\ &= n_{2q+3} - n_2, \quad \text{from (1)} \\ &= n_{i+2} - 2.\end{aligned} \tag{A.1}$$

For $i \bmod 2 = 0$, let $i = 2q$. Then,

$$\begin{aligned}\sum_{p=1}^{i} n_p &= n_1 + n_2 + n_3 + n_4 + \cdots + n_i \\ &= n_3 + n_5 + \cdots + n_{2q+1}, \quad \text{from (1)} \\ &= n_2 + n_3 + n_5 + \cdots + n_{2q+1} - n_2 \\ &= n_4 + n_5 + \cdots + n_{2q+1} - n_2, \quad \text{from (1)}\end{aligned}$$

$$= n_{2q+2} - n_2, \quad \text{from (1)}$$
$$= n_{i+2} - 2. \tag{A.2}$$

Accordingly, $\sum_{p=1}^{i} n_p = n_{i+2} - 2$.

B. Proof of Lemma 1

This study first proves (5a). For easy understanding, please refer to Figure 4(a).

(1) For $t < n_{i-1}$, a client downloads no segment on channel C_i because these segments will appear again during time units n_{i-1} to $n_{i+1} - 1$. Thus, $B(i,t) = 0$.

(2) For $n_{i-1} \le t \le n_{i+1} - 2$, a client continuously accepts one segment every time unit but consumes no segment. Thus, the number of buffered segment equals $t - n_{i-1} + 1$. When $t = n_{i+1} - 2$, the client buffers the maximum segments and $B(i,t) = n_i - 1$.

(3) For $n_{i+1} - 2 < t \le n_{i+2} - 2$, Figure 4(a) shows that a client stops loading data but plays the video in this period. The client consumes one segment every time unit, and thus the buffered segments decrease, $B(i,t) = n_i - (t - (n_{i+1} - 2)) = n_{i+2} - 2 - t$.

(4) For $n_{i+2} - 2 < t$, a client has finished playing all the segments on channel C_i, and thus $B(i,t) = 0$.

The proof for (5b) is as follows. For easy understanding, please refer to Figure 4(b).

(1) For $t < n_{i-1}$, a client does not download any segment on channel C_i, and thus, $B(i,t) = 0$.

(2) For $n_{i-1} \le t \le n_{i+1} - 2$, the paper divides the value range of t into four successive subranges for ease of proof.

(a) For $n_{i-1} \le t \le \lfloor (j + n_{i+1} - 1)/2 \rfloor - n_i$, Figure 4(b) shows that the client downloads no segment on channel C_i, and thus $B(i,t) = 0 \le \lfloor (t - n_{i-1} + 2)/2 \rfloor$.

(b) For $\lfloor (j + n_{i+1} - 1)/2 \rfloor - n_i < t \le j - n_i$,

$$\begin{aligned} B(i,t) &= t - \left(\left\lfloor \frac{j + n_{i+1} - 1}{2} \right\rfloor - n_i \right), \quad \text{see Figure 4(b)} \\ &\le t + n_i - \left\lfloor \frac{(t + n_i) + (n_{i+1} - 1)}{2} \right\rfloor, \quad \text{due to } t \le j - n_i \\ &\le \left\lfloor \frac{t - n_{i-1} + 2}{2} \right\rfloor. \end{aligned} \tag{B.1}$$

(c) For $j - n_i < t \le \lfloor (j + n_{i+2} - 1)/2 \rfloor - n_i$,

$$\begin{aligned} B(i,t) &= (j - n_i) - \left(\left\lfloor \frac{j + n_{i+1} - 1}{2} \right\rfloor - n_i \right), \quad \text{see Figure 4(b)} \\ &\le \left\lfloor \frac{t - n_{i-1} + 2}{2} \right\rfloor, \quad \text{due to } j - n_i < t. \end{aligned} \tag{B.2}$$

(d) For $\lfloor (j + n_{i+2} - 1)/2 \rfloor - n_i < t \le n_{i+1} - 2$,

$$\begin{aligned} B(i,t) &= \left(j - \left\lfloor \frac{j + n_{i+1} - 1}{2} \right\rfloor \right) \\ &\quad + \left(t - \left(\left\lfloor \frac{j + n_{i+2} - 1}{2} \right\rfloor - n_i \right) \right), \quad \text{see Figure 4(b)} \\ &\le \left\lfloor \frac{j + 2t - n_{i+1} + 2 + 2n_i}{2} \right\rfloor - \left\lfloor \frac{j + n_{i+2} - 1}{2} \right\rfloor \\ &\le \left\lfloor \frac{t - n_{i-1} + 2}{2} \right\rfloor, \quad \text{due to } t \le n_{i+1} - 2. \end{aligned} \tag{B.3}$$

Thus, $B(i,t) \le \lfloor n_i/2 \rfloor$ when $t = n_{i+1} - 2$.

(3) For $n_{i+1} - 2 < t \le n_{i+2} - 2 - \lceil n_i/2 \rceil$, the client plays one segment every time unit while downloading at most one segment. The number of buffered segments is not larger than that at time unit $t = n_{i+1} - 2$, and thus $B(i,t) \le \lfloor n_i/2 \rfloor$.

(4) For $n_{i+2} - 2 - \lceil n_i/2 \rceil < t \le n_{i+2} - 2$, the client has played $t - (n_{i+1} - 2)$ segments, and the number of the remaining segments is

$$n_i - (t - (n_{i+1} - 2)) = n_{i+2} - 2 - t. \tag{B.4}$$

Thus, $B(i,t) \le n_{i+2} - 2 - t$.

(5) For $n_{i+2} - 2 < t$, the client finishes playing all the segments on channel C_i so $B(i,t) = 0$.

The proof is complete.

C. Proof of Theorem 2

This work divides client playing time t (in terms of time units) into multiple successive durations for ease of proof.

(1) For $t \le n_{k-2} - 2$, Figure 5 shows that the client simply receives segments from channels C_1 to C_{k-2}. In addition, the client downloads two segments but plays only one every time unit. The number of buffered segments thus increases with time and achieves the maximum at time unit $n_{k-2} - 2$. At this time, the client has finished playing segments from channels C_1 to C_{k-4}, and thus simply buffers segments from channels C_{k-3} and C_{k-2}. Accordingly,

$$\begin{aligned} B(t) &= B(k-3, n_{k-2} - 2) + B(k-2, n_{k-2} - 2) \\ &= ((n_{k-2} - 2) - n_{k-4} + 1) \\ &\quad + ((n_{k-2} - 2) - n_{k-3} + 1), \quad \text{from (5a)} \\ &= n_{k-2} - 2, \quad \text{from (1)} \\ &\le \left\lceil \frac{n_{k-1}}{4} \right\rceil + \left\lfloor \frac{n_k}{2} \right\rfloor. \end{aligned} \tag{C.1}$$

(2) For $n_{k-2} - 2 < t \le n_{k-1} - 2$, the client has finished playing all the segments received from channels C_1 to C_{k-4} and downloads no segment from channel C_k. We thus merely consider the segments on channels C_{k-3} to C_{k-1}. The client downloads at least one segment from channel C_{k-2} but plays

Table 3: Comparison of buffering space in the percentage of video size using k server channels.

k	1	2	3	4	5	6	7	8	9	10
FiB (%)	0	33.3	33.3	36.4	36.8	37.5	37.7	38	38	38.1
FiB+ (%)	0	33.3	33.3	27.3	26.3	25	24.5	25.3	25.4	25.1
Reduction rate (%)	0	0	0	25	28.6	33.3	35	33.3	33.3	34.1

only one every time unit. Thus, the maximum number of buffered segments appears at time unit $n_{k-1} - 2$ as follows:

$$\begin{aligned} B(t) &= B(k-3, n_{k-1}-2) + B(k-2, n_{k-1}-2) \\ &\quad + B(k-1, n_{k-1}-2) \\ &\le n_{k-2} - 1 + \left\lfloor \frac{(n_{k-1}-2) - n_{k-2} + 2}{2} \right\rfloor, \\ &\qquad\qquad \text{from (5a) and (5b)} \\ &= n_{k-2} + \left\lfloor \frac{n_{k-3}}{2} \right\rfloor - 1 \\ &\le \left\lceil \frac{n_{k-1}}{4} \right\rceil + \left\lfloor \frac{n_k}{2} \right\rfloor. \end{aligned} \tag{C.2}$$

(3) For $n_{k-1} - 2 < t \le n_k - 2$, the client has finished playing all the segments received from channels C_1 to C_{k-3}, and thus the client only buffers segments from channels C_{k-2} to C_k as follows:

$$\begin{aligned} B(t) &= B(k-2, t) + B(k-1, t) + B(k, t) \\ &\le n_k - 2 - t + \left\lfloor \frac{t - n_{k-2} + 2}{2} \right\rfloor + \left\lfloor \frac{t - n_{k-1} + 2}{2} \right\rfloor, \\ &\qquad\qquad \text{from (5a) and (5b)} \\ &\le \left\lceil \frac{n_{k-1}}{4} \right\rceil + \left\lfloor \frac{n_k}{2} \right\rfloor. \end{aligned} \tag{C.3}$$

(4) For $n_k - 2 < t \le n_{k+1} - 2 - \lceil n_{k-1}/2 \rceil$, the client has finished playing all the segments received from channels C_1 to C_{k-2} and only performs segment downloading on channels C_{k-1} and C_k as follows:

$$\begin{aligned} B(t) &\le B(k-1, t) + B(k, t) \\ &\le \left\lfloor \frac{n_{k-1}}{2} \right\rfloor + \left\lfloor \frac{t - n_{k-1} + 2}{2} \right\rfloor, \quad \text{from (5b)} \\ &\le \left\lceil \frac{n_{k-1}}{4} \right\rceil + \left\lfloor \frac{n_k}{2} \right\rfloor, \quad \text{due to } t \le n_{k+1} - 2 - \left\lceil \frac{n_{k-1}}{2} \right\rceil. \end{aligned} \tag{C.4}$$

(5) For $n_{k+1} - 2 - \lceil n_{k-1}/2 \rceil < t \le n_{k+1} - 2$, similarly, the client merely downloads segments on channels C_{k-1} and C_k as follows:

$$\begin{aligned} B(t) &\le B(k-1, t) + B(k, t) \\ &\le n_{k+1} - 2 - t + \left\lfloor \frac{t - n_{k-1} + 2}{2} \right\rfloor, \quad \text{from (5b)} \\ &\le \left\lceil \frac{n_{k-1}}{4} \right\rceil + \left\lfloor \frac{n_k}{2} \right\rfloor, \quad \text{due to } n_{k+1} - 2 - \left\lceil \frac{n_{k-1}}{2} \right\rceil < t. \end{aligned} \tag{C.5}$$

(6) For $n_{k+1} - 2 < t$, the client simply performs data downloading on channel C_k, and thus $B(t) = B(k, t)$. From (5b), $B(t) = B(k, t) \le \lfloor n_k/2 \rfloor \le \lceil n_{k-1}/4 \rceil + \lfloor n_k/2 \rfloor$.

The proof is complete.

Acknowledgment

This work was financially supported by National Science Council, Taiwan under a research grant numbered NSC 101-2221-E-152-004.

References

[1] TechNavio, "Global video on demand market 2011–2015," August 2012.

[2] Digital TV Research, "A sustained boom forecast for global online TV and video," October 2012.

[3] M. Vilas, X. G. Pañeda, R. García, D. Melendi, and V. G. García, "User behaviour analysis of a video-on-demand service with a wide variety of subjects and lengths," in *Proceedings of the 31st EUROMICRO Conference on Software Engineering and Advanced Applications (EUROMICRO-SEAA '05)*, pp. 330–337, September 2005.

[4] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, "Understanding user behavior in large-scale video-on-demand systems," in *Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems (EuroSys '06)*, pp. 333–344, October 2006.

[5] J. Choi, A. S. Reaz, and B. Mukherjee, "A survey of user behavior in VoD service and bandwidth-saving multicast streaming schemes," *IEEE Communications Surveys and Tutorials*, 2008.

[6] L.-S. Juhn and L.-M. Tseng, "Fast data broadcasting and receiving scheme for popular video service," *IEEE Transactions on Broadcasting*, vol. 44, no. 1, pp. 100–105, 1998.

[7] Y.-C. Tseng, M.-H. Yang, and C.-H. Chang, "A recursive frequency-splitting scheme for broadcasting hot videos in VOD service," *IEEE Transactions on Communications*, vol. 50, no. 8, pp. 1348–1355, 2002.

[8] Z. Y. Yang, Y. M. Chen, and L. M. Tseng, "A seamless broadcasting scheme with live video support," *International Journal of*

Digital Multimedia Broadcasting, vol. 2012, Article ID 373459, 8 pages, 2012.

[9] Y.-W. Chen and C.-Y. Chen, "PAS: a new scheduling scheme for broadcasting a video over a single channel," *IET Communications*, vol. 5, no. 7, pp. 951–960, 2011.

[10] B. S. Wu, C. C. Hsieh, and Y. W. Chen, "A reverse-order scheduling scheme for broadcasting continuous multimedia data over a single channel," *IEEE Transactions on Broadcasting*, vol. 57, no. 3, pp. 721–728, 2011.

[11] H.-F. Yu, H.-C. Yang, and L.-M. Tseng, "Reverse Fast Broadcasting (RFB) for video-on-demand applications," *IEEE Transactions on Broadcasting*, vol. 53, no. 1, pp. 103–110, 2007.

[12] H.-F. Yu, "Hybrid broadcasting with small buffer demand and waiting time for video-on-demand applications," *IEEE Transactions on Broadcasting*, vol. 54, no. 2, pp. 304–311, 2008.

[13] Y. N. Chen and L. M. Tseng, "An efficient periodic broadcasting with small latency and buffer demand for near video on demand," *International Journal of Digital Multimedia Broadcasting*, vol. 2012, Article ID 717538, 7 pages, 2012.

[14] Y. W. Chen, C. C. Lin, and C. Y. Huang, "Hybrid broadcasting scheme with low waiting time and buffer requirement for video-on-demand services," *IET Communications*, vol. 6, no. 17, pp. 2949–2956, 2012.

[15] K. A. Hua and S. Sheu, "Skyscraper broadcasting: a new broadcasting scheme for metropolitan video-on-demand systems, ," in *Proceedings of the ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM '97)*, pp. 89–100, September 1997.

[16] Y. Cai, A. Hua, and S. Sheu, "Leverage client bandwidth to improve service latency of distributed multimedia applications," *Journal of Applied Systems Studies*, vol. 2, no. 3, pp. 686–704, 2001.

[17] A. Natarajan, Y. Cai, and J. Wong, "An enhanced client-centric approach for efficient video broadcast," *Multimedia Tools and Applications*, vol. 43, no. 2, pp. 179–193, 2009.

[18] Y. Guo, L. Gao, D. Towsley, and S. Sen, "Smooth workload adaptive broadcast," *IEEE Transactions on Multimedia*, vol. 6, no. 2, pp. 387–395, 2004.

[19] C.-J. Wu, Y.-W. Chen, and Y.-L. Wang, "The minimum bandwidth required at each time slot of the fast broadcasting scheme," *Information Processing Letters*, vol. 111, no. 20, pp. 1014–1018, 2011.

[20] J. B. Kwon, "Proxy-assisted scalable periodic broadcasting of videos for heterogeneous clients," *Multimedia Tools and Applications*, vol. 51, no. 3, pp. 1105–1125, 2011.

[21] H. Febiansyah and J. B. Kwon, "Dynamic proxy-assisted scalable broadcasting of videos for heterogeneous environments," *Multimedia Tools and Applications*, 2012.

[22] J. Kepler, *A New Year Gift: On Hexagonal Snow*, Oxford University Press, Oxford, UK, 1966.

Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 629350, 8 pages
http://dx.doi.org/10.1155/2013/629350

Research Article

Single-Channel Data Broadcasting under Small Waiting Latency

Hsiang-Fu Yu

Department of Computer Science, National Taipei University of Education, Taipei 10671, Taiwan

Correspondence should be addressed to Hsiang-Fu Yu; yu@tea.ntue.edu.tw

Received 5 March 2013; Accepted 18 March 2013

Academic Editor: Pin-Han Ho

Due to the advancement of network technology, video-on-demand (VoD) services are growing in popularity. However, individual stream allocation for client requests easily causes a VoD system overload; when its network and disk bandwidth cannot match client growth. This study thus presents a fundamentally different approach by focusing solely on a class of applications identified as latency tolerant applications. Because video broadcasting does not provide interactive (i.e., VCR) functions, a client is able to tolerate playback latency from a video server. One efficient broadcasting method is periodic broadcasting, which divides a video into smaller segments and broadcasts these segments periodically on multiple channels. However, numerous practical systems, such as digital video broadcasting-handheld (DVB-H), do not allow clients to download video data from multiple channels because clients usually only have one tuner. To resolve this problem in multiple-channel broadcasting, this study proposes a novel single-channel broadcasting scheme, which leverages segment-broadcasting capability further for more efficient video delivery. The comparison results show that, with the same settings of broadcasting bandwidth, the proposed scheme outperforms the alternative broadcasting scheme, the hopping insertion scheme, SingBroad, PAS, and the reverse-order scheduling scheme for the maximal waiting time.

1. Introduction

Due to the advancement of network technology, video-on-demand (VoD) services are growing in popularity. Clients can watch their desired videos at anytime without waiting or visiting a video rental store. Because of the online access provided by VoD services, several studies have predicted the success of VoD [1, 2]. VoD is inherently a personalized service because of its characteristic one-to-one interaction. Therefore, a VoD system typically allocates a dedicated stream for each incoming video request [3]. However, individual stream allocation easily causes a VoD system overload when its network and disk bandwidth cannot match client growth. This study thus presents a fundamentally different approach by focusing solely on a class of applications identified as latency tolerant applications. The key feature of latency tolerant applications is that they are unconcerned with latency between video servers and clients. Broadcast video streaming is perhaps the most important example of this class of applications [4]. Because video broadcasting does not provide interactive (i.e., VCR) functions, a client is able to tolerate playback latency from a video server. One efficient broadcasting method is periodic broadcasting, which divides a video into smaller segments and broadcasts them periodically on a set of communication channels. This method enhances bandwidth usage by allowing various clients to share the same channel bandwidths. Because periodic broadcasting typically requires a client to wait for the beginning of the first segment before starting playback, this scheme cannot support real-time VoD services.

The fast broadcasting (FB) [5] scheme divides a video into a geometrical series of $1, 2, 4, \ldots, 2^{k-1}$, where k is the number of broadcasting channels. An implementation of the FB scheme on IP multicasting was reported in [6]. To achieve minimal latency, the harmonic broadcasting (HB) scheme [7] partitions a video into multiple segments, and each segment S_i is divided into i subsegments. The subsegments of the same segment are then broadcast on the same channel. The recursive frequency-splitting (RFS) scheme [8] achieves a near-minimal waiting time by periodically broadcasting each segment at a frequency that can guarantee continuous video

playback. In modifying the FB scheme, the reverse fast broadcasting (RFB) scheme [9] buffers 25% of video length, merely half of what is required by the FB scheme. By combining RFB and RFS, the hybrid broadcasting scheme (HyB) [10] requires the same client buffering space as that of RFB; however, it achieves smaller waiting time. The study in [11] integrates the fixed-delay pagoda broadcasting scheme [12] and RFB to reduce client waiting time and buffer demand. A generalized reverse sequence-based model [13] was proposed to clarify why broadcasting segments in reverse order can reduce buffer requirements. A scalable binomial broadcasting scheme [14] transfers live videos using constant bandwidth, regardless of video length.

The mentioned schemes transfer video segments on multiple channels simultaneously and periodically, and a client typically must receive segments from these channels concurrently. To perform multiple-channel segment broadcasting, a server must multiplex video segments into multiple channels and synchronize these segments across these channels. Segment multiplexing and synchronization are difficult, because packet transmissions are varied with network traffic. In addition, numerous practical systems, such as digital video broadcasting-handheld (DVB-H) and integrated services digital broadcasting-handheld (ISDBH), do not permit a client to download video data from multiple channels, because the client typically has only one tuner [15, 16]. To solve these problems caused by multiple-channel broadcasting, many studies were proposed to broadcast segments over a single channel, such as the alternative broadcasting (AB) scheme [15], the hopping insertion (HI) scheme [16], SingBroad [17], PAS [18], and the reverse-order scheduling (ROS) scheme [19]. The basic concept behind these schemes is to partition a video into equal-sized segments, which are classified into several groups and transferred over a single channel according to a predefined arrangement.

This study proposes a single-channel broadcasting scheme to yield short waiting time. Let kb be the bandwidth of a single channel, where k is a positive integer and b is the playback rate of a video. The proposed scheme partitions the single channel as an infinite set of time slots. Each time slot is further composed of smaller subslots. A video of length L is equally divided into $2^k - 1$ segments, which are then arranged to k groups, denoted by $G_0, G_1, \ldots, G_{k-1}$. A segment of group G_i is split into 2^i equal subsegments, which are then placed to individual subslots. The mathematical analysis shows that the maximal client waiting time of the scheme is $(k+1)L/k(2^k - 1)$. This study also verifies the workability of the scheme and compares it with several current approaches. The comparison results show that, with the same settings of broadcasting bandwidth, the proposed scheme outperforms AB, HI, SingBroad, PAS, and ROS for the maximal waiting time. Extensive simulations also indicate that the proposed scheme requires smaller client buffering space than AB and SingBroad for $k > 4$.

The remainder of this study is organized as follows. Section 2 reviews AB, HI, SingBroad, PAS, and ROS. Section 3 introduces the proposed scheme and verifies its accuracy. Section 4 shows the evaluations of the performance of the scheme, and Section 5 makes a brief conclusion.

Table 1: List of terms used in this study and their respective definitions.

Term	Definition
L	Video length
b	Video playback rate
kb	Bandwidth of a single channel, where k is a positive integer
N	Number of segments
S_i	ith video segment
d	Segment length of S_i
$S_{i,j}$	jth subsegment of S_i
t_a	Client arrival time
T_i	ith time slot, $i \geq 0$
w	Maximal waiting time for playback

2. Related Work

This section introduces AB [15], HI [16], SingBroad [17], PAS [18], and ROS [19]. Table 1 defines the terms used in this study. As mentioned previously, this study divides a single channel into an infinite set of time slots.

The AB scheme [15] splits a video into N segments. This scheme proposes two modes for determining the value of N. One is the mechanism-dominant (MD) mode, and the other is the waiting time-dominant (WD) mode. In the MD mode, $N = \lfloor (k+3)/2 \rfloor$, and clients start playing video data when they receive segment S_1. The AB scheme with WD obtains $N = \lceil (k+3)/2 \rceil$. In this mode, the starting time of video playback is determined by whether each segment can be played continuously, rather than the downloading of segment S_1. For both modes, the AB scheme broadcasts segment S_1 on the single channel at time slot T_i if $i \bmod 2 = 0$. The rest of segments are broadcast sequentially and periodically on the remaining time slots. Figure 1 shows an example to demonstrate the segment downloading and playing for AB, where $k = 4$. The segments downloaded and played by a client are gray. The AB scheme with MD divides a video into three segments, as shown in Figure 1(a). A client starts to play video data at the beginning of segment S_1. In addition, the AB scheme with WD partitions the same video into four segments, as presented in Figure 1(b). Note that if the client started playing segment S_1 on time slot T_2, segment S_2 would not be played continuously. Therefore, the client begins to play video data on time slot T_3 to guaranteecontinuous playback.

The HI scheme [16] divides a video into N even segments, where N is an arbitrary positive integer. This scheme then classifies the segments into $\lceil N/Q \rceil$ groups, where $Q = \lceil N/\exp(0.57(k-1)) \rceil$. Group G_j contains segments S_{jQ+1} to $S_{(j+1)Q}$, where $0 \leq j \leq \lceil N/Q \rceil - 2$. The last group includes the remaining segments. Initially, HI puts the segments of group G_0 together in order. The segments of the remaining groups are then inserted into the segments of G_0 in a hopping way to obtain the final broadcasting schedule [16].

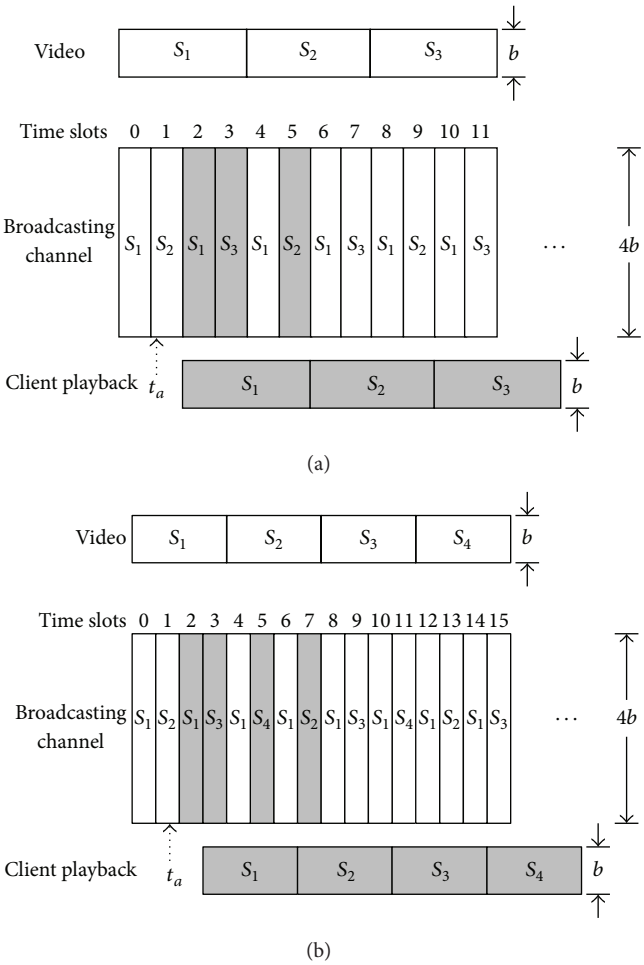


Figure 1: Segment partition and arrangement for AB.

SingBroad [17] partitions a video into $2^{k-1} - 1$ segments that are arranged into $k - 1$ groups. Group G_j contains segments S_{2^j} to $S_{2^{j+1}-1}$, where $0 \le j \le k - 2$. Segment S_{2^j+i} of group G_j is broadcast on time slot $T_{j+i(k-1)+2^j(k-1)y}$, where $0 \le i \le 2^j - 1$ and y is zero or a positive integer. For example, for $k = 4$, SingBroad divides a video into seven segments, which are then arranged to three groups. Group G_2 contains segments S_4 to S_7. Segment S_5 is broadcast on time slot T_{5+12y} (e.g., time slots T_5, T_{17}, T_{29}, and so on), where $j = 2$ and $i = 1$. When a video request arrives, the client must wait for the beginning of the nearest segment S_1 to start video downloading and playing.

Like the SingBroad scheme, the PAS scheme [18] splits a video into $2^{k-1} - 1$ segments and classifies these segments into $k - 1$ groups. Group G_j contains segments S_{2^j} to $S_{2^{j+1}-1}$, where $0 \le j \le k - 2$. Unlike SingBroad, PAS further divides each segment of G_j into 2^j even subsegments. Each time slot T_i is split into $2^{i \bmod (k-1)}$ subslots that are used to place subsegments. For instance, for $k = 4$, PAS partitions a video into segments S_1 to S_7, which are arranged to three groups. Segment S_5 of G_2 is divided into four subsegments $S_{5,1}$, $S_{5,2}$, $S_{5,3}$, and $S_{5,4}$ that are broadcast across various subslots. A client must wait for the nearest segment S_1 to begin video downloading and playing.

The ROS scheme [19] divides a video into $3 \times 2^{k-2}$ segments that are classified into k groups. Groups G_0 and G_1 contain $\{S_1\}$ and $\{S_2, S_3\}$, respectively. The remaining group G_j includes segments $S_{3\times 2^{j-2}+1}$ to $S_{3\times 2^{j-1}}$ where $2 \le j \le k - 1$. Let y be zero or a positive integer. Segment of G_0 is broadcast on time slot T_{ky}. This scheme then puts segments S_3 and S_2 of G_1 on time slots T_{1+2ky} and $T_{k+1+2ky}$, respectively. The ROS scheme transmits segment $S_{3\times 2^{j-1}-x}$ of the remaining group G_j on time slot $T_{j+xk+3\times 2^{j-2}\times ky}$, where $2 \le j \le k-1$ and $0 \le x \le 3 \times 2^{j-2} - 1$. For example, segment S_6 of group G_2 is broadcast on time slot T_{2+3ky} because $j = 2$ and $x = 0$. For $k = 4$, ROS puts segment S_6 on time slots T_2, T_{14}, T_{26}, and others. When a client wants to watch a video, the client must wait for the beginning of the nearest segment S_1 to start downloading. In addition, segments S_2 and S_3 must be received in order. For the segments of the remaining groups, the client downloads them according to the following process. Suppose that S_p is the segment that a client is currently playing, and segment S_i of G_j is the segment that appears on the channel and is not received by the client. If $p + 3 \times 2^{j-2} < i$, the client does not download segment S_i. Otherwise, the client receives it. When downloading segment S_1 is complete, the client starts video playback.

3. Proposed Scheme

According to the mentioned schemes [15–19], the number of video segments mainly determines client waiting time. Therefore, the key to minimizing the waiting time is to partition a video into as many segments as possible, under the condition that ensures continuous playback. To maximize the segment number, the proposed scheme broadcasts video data over a single channel according to the following step.:

(1) Divide a video into $2^k - 1$ (i.e., $N = 2^k - 1$) equal-length segments, denoted by $S_1, S_2, \ldots, S_{2^k-1}$ in sequence. The length of each segment, d, thus equals $L/(2^k - 1)$. For example, in Figure 2, a server allocates a single channel with a bandwidth of $3b$ to broadcast a video of length L. The video is equally divided into $2^3 - 1$ segments, denoted by $S_1, S_2, \ldots, S_7$. The length of each segment equals $L/7$.

(2) Classify these segments into k groups, denoted by $G_0, G_2, \ldots, G_{k-1}$. Assemble segments S_{2^j} to $S_{2^{j+1}-1}$ into group G_j sequentially. Figure 2 shows that the segments are then classified into three groups $G_0 = \{S_1\}$, $G_1 = \{S_2, S_3\}$, and $G_2 = \{S_4, S_5, S_6, S_7\}$. Each segment S_i of group G_j is further partitioned into 2^j subsegments, denoted by $S_{i,1}, S_{i,2}, \ldots, S_{i,2^j}$. As shown in Figure 2, segment S_5 of group G_2 is split into four subsegments $S_{5,1}$, $S_{5,2}$, $S_{5,3}$, and $S_{5,4}$.

(3) Partition a single channel as an infinite set of time slots, denoted by T_0, T_1, T_2, and so on. Each time slot

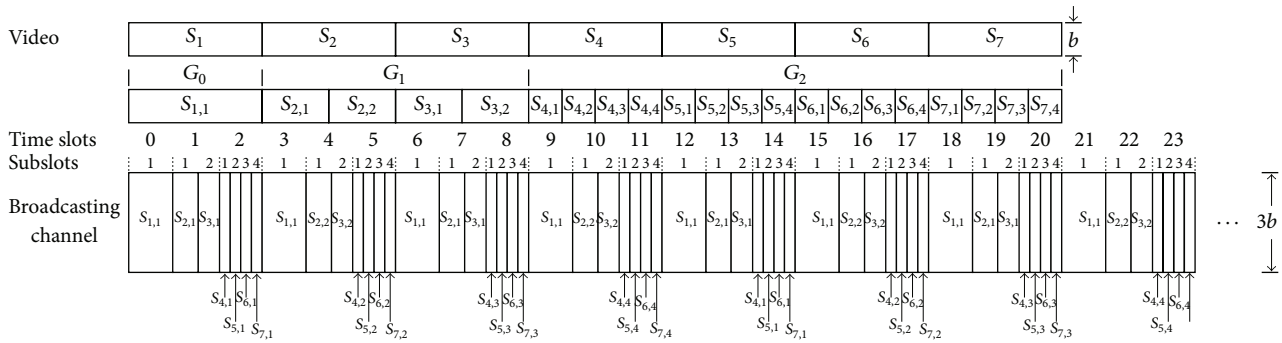


FIGURE 2: Segment partition and arrangement for the proposed scheme.

is used to deliver segment data at a bandwidth of kb, and the length of each time slot equals

$$\frac{L}{(kN)} = \frac{d}{k}. \tag{1}$$

(4) A time slot T_i is further divided into 2^j subslots, denoted by $T_{i,1}, T_{i,2}, \ldots, T_{i,2^j}$, if $i \bmod k = j$. The length of a subslot of time slot T_i thus equals $d/(2^j k)$. For example, Figure 2 shows that the length of each time slot equals $d/3$ because $k = 3$. Time slot T_2 is further partitioned into four subslots $T_{2,1}$, $T_{2,2}$, $T_{2,3}$, and $T_{2,4}$ because 2 mod 3 = 2.

(5) Put the segment data of each group on each time slot in sequence. For example, the segment data of groups G_0, G_1, and G_2 are sequentially broadcast on time slots T_0, T_1, T_2, and so on, as indicated in Figure 2. In general, the segment data of group G_j are put on time slot T_{j+ky}, because there are k groups, where y is zero or a positive integer. Furthermore, the scheme sequentially broadcasts the subsegments of the segments of group G_j on the subslots of time slot T_{j+ky}. For example, Figure 2 shows that the subsegments of segments S_4 to S_7 of group G_2 are sequentially put on the subslots of T_2, T_5, T_8, T_{11}, and others. Note that only a subsegment of a segment of the same group is put on a subslot of a time slot. Because the segment data of group G_j are broadcast once every k time slots and each segment consists of 2^j subsegments, each subsegment is transmitted once every $2^j k$ time slots. Therefore, the scheme broadcasts subsegment $S_{i,x}$ of group G_j on subslot

$$T_{j+(x-1)k+2^j ky, i-2^j+1}, \tag{2}$$

where $y \in$ int and $y \geq 0$.

For example, Figure 2 shows that the proposed scheme puts subsegment $S_{5,3}$ of group G_2 on subslot $T_{8+12y,2}$ (e.g., $T_{8,2}$, $T_{20,2}$, and $T_{32,2}$) because $k = 3$, $j = 2$, $i = 5$, and $x = 3$.

This study next presents how to download video segments on the client side. A client is assumed to have a sufficient buffer to store downloaded segments. Suppose that a client can download and play the same segment concurrently, because the downloading bandwidth is equal to or larger than the playback rate. This study also assumes that a client desires to watch a video at time t_a. Let $T_{u,v}$ be the subslot that is nearest to time t_a. The segment downloading and playing are as the following.

(1) The client must wait for subslot $T_{u,v}$ before receiving subsegments. Once the subslot is up, the client starts downloading the subsegment from this subslot.

(2) After this downloading is complete, the client continues to receive the remaining subsegments from the following subslots. If a subsegment has been downloaded, the client simply skips it.

(3) When all the subsegments are received, the client stops the segment downloading.

(4) The client assembles the received subsegments to form complete segments and starts playing them at the beginning of subslot $T_{u+k,v}$.

Figure 3 shows an example for demonstrating how to download and play video segments, where the subsegments downloaded and played by a client are gray. Because subslot $T_{1,2}$ is closest to the client arrival time t_a, the client starts downloading subsegment $S_{3,1}$ on subslot $T_{1,2}$. The client then continues to receive subsegments from subslots $T_{2,1}$ to $T_{5,4}$. Because subsegment $S_{1,1}$ has been downloaded on subslot $T_{3,1}$, the client does not receive it again on subslot $T_{6,1}$. Similarly, the client does not download subsegments $S_{3,1}$, $S_{1,1}$, $S_{2,2}$, and $S_{3,2}$ on subslots $T_{7,2}$, $T_{9,1}$, $T_{10,1}$, and $T_{10,2}$, respectively. When the client finishes receiving all the subsegments at the end of subslot $T_{11,4}$, the client stops downloading subsegments. The client assembles the received subsegments to form complete segments and plays them at the start of subslot $T_{4,2}$, as shown in Figure 3.

3.1. Workable Verification. Suppose that segment S_i is in group G_j, where $2^j \leq i \leq 2^{j+1} - 1$. The mentioned broadcasting process transfers a subsegment of segment S_i once every k time slots. Because the number of subsegments of segment S_i equals 2^j, the broadcasting process can transmit all the subsegments of segment S_i once every $2^j k$ time slots.

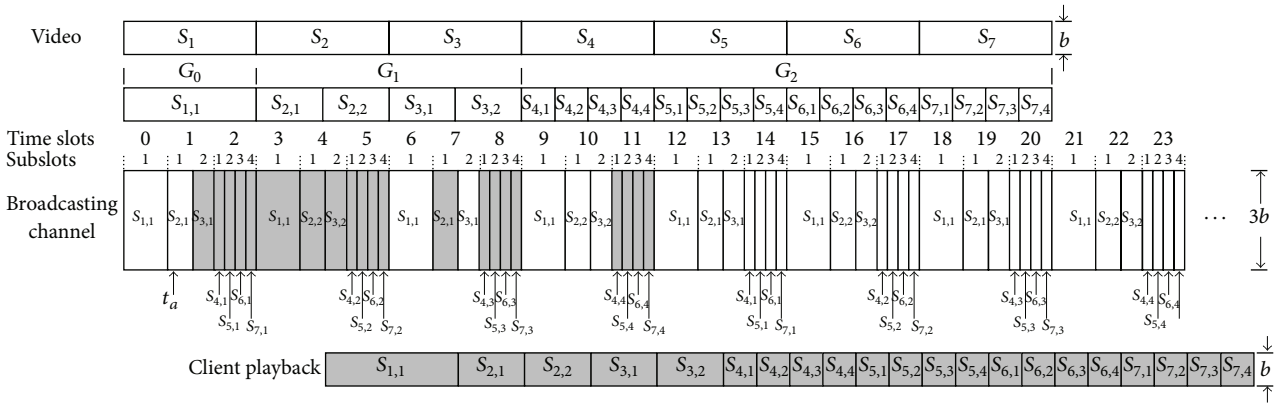


Figure 3: Segment downloading and playing for the proposed scheme.

According to the downloading process, a client starts segment downloading at the beginning of subslot $T_{u,v}$. Therefore, the client can receive all the subsegments of S_i at the beginning of subslot $T_{u+2^j k,v}$. In addition, the client begins segment playback at the beginning of subslot $T_{u+k,v}$. Because the playback length of a segment equals k time slots according to (1), the start time to play segment S_i is the beginning of subslot $T_{u+k+ik,v}$. To guarantee continuous playback for the client, the end time of downloading segment S_i must be earlier than the start time of its playback. That is, the beginning of subslot $T_{u+k+ik,v}$ must be later than the beginning of subslot $T_{u+2^j k,v}$. This study evaluates

$$(u+k+ik)-\left(u+2^j k\right)=\left(i+1-2^j\right)k>0,\ \text{due to } 2^j\le i. \tag{3}$$

The end time of downloading segment S_i is earlier than the start time of its playback. Therefore, the proposed scheme ensures continuous video playback on the client side.

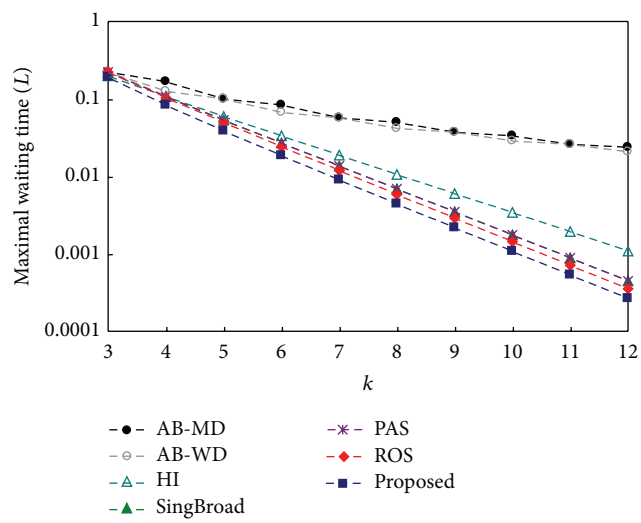


Figure 4: Maximal waiting time (in terms of L) incurred on new clients at different broadcasting bandwidth.

4. Performance Analysis and Comparison

This study primarily selected client waiting time and buffer demand as the performance criteria. The proposed scheme was compared with AB, HI, SingBroad, PAS, and ROS. According to the downloading process, when a client exactly arrives at the beginning of a subslot, the waiting time equals k timeslots (i.e., d) because of (1). If the client just misses the startup of a subsegment on the channel, the client must additionally wait for the length of the subsegment. Because subsegment $S_{1,1}$ is the longest subsegment, the maximal waiting time w equals $k+1$ time slots. That is,

$$w=\frac{(k+1)\,d}{k}=\frac{(k+1)\,L}{k\left(2^k-1\right)}. \tag{4}$$

Table 2 summarizes the maximal waiting time incurred by AB [15], HI [16], SingBroad [17], PAS [18], ROS [19], and the proposed scheme. The results show that the number of segments mainly determines the maximal waiting times for all the schemes. The increase of the server bandwidth (i.e., the value of k) enlarges the number of segments and thus reduces the waiting time.

To clarify the performance advantages of the proposed scheme, this study calculated the maximal waiting times of AB, HI, SingBroad, PAS, ROS, and the proposed scheme at various values of k, where the value of N for HI equals 10000. Figure 4 shows the performance results. As the server bandwidth increases, the waiting times under all the schemes are sharply reduced. In addition, the proposed scheme yields the shortest waiting time. For example, when the server bandwidth equals $7b$ (i.e., $k=7$), the scheme reduces the broadcast latency to less than $0.009L$. In contrast, AB-MD, AB-WD, HI, SingBroad, PAS, and ROS yield 0.057, 0.057, 0.019, 0.014, 0.014, and $0.012L$, respectively. The proposed scheme reduces the waiting times by 84%, 84%, 53%, 36%, 36%, and 25%. Assume that the video length L is 120 min. Figure 5 shows the maximal waiting time for all the schemes in seconds. For $k=6$, the waiting times of AB-MD, AB-WD, HI, SingBroad, PAS, ROS, and the proposed scheme are 600,

Table 2: Maximal waiting time for different schemes in terms of video length L.

Scheme	AB-MD	AB-WD	HI	SingBroad	PAS	ROS	Proposed
Maximal waiting time	$\frac{2L}{\lfloor (k+3)/2 \rfloor k}$	$\frac{2L}{\lceil (k+3)/2 \rceil k}$	$\frac{(Q \times H(\lfloor N/Q \rfloor) + ((N - Q\lfloor N/Q \rfloor)/(\lfloor N/Q \rfloor + 1)))L}{kN}$, where $Q = \left\lceil \frac{N}{\exp(0.57(k-1))} \right\rceil$ $H(i) = \sum_{j=1}^{i} \frac{1}{j}$	$\frac{(k-1)L}{(2^{k-1}-1)k}$	$\frac{(k-1)L}{(2^{k-1}-1)k}$	$\frac{(k+1)L}{3 \times 2^{k-2}k}$	$\frac{(k+1)L}{(2^k-1)k}$

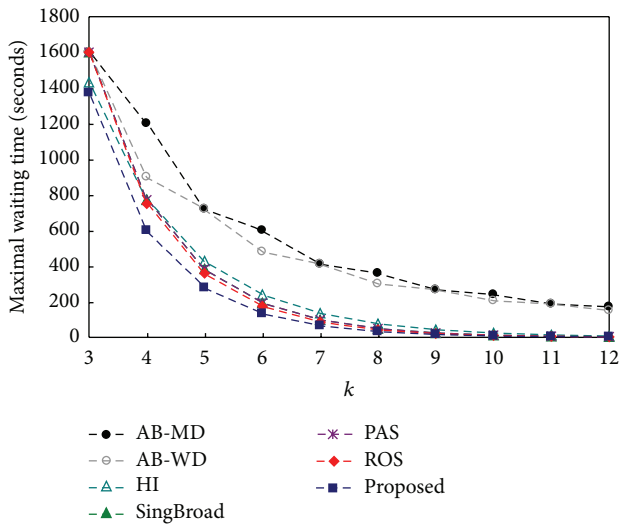


Figure 5: Maximal waiting time yielded by different schemes, where $L = 120$ min.

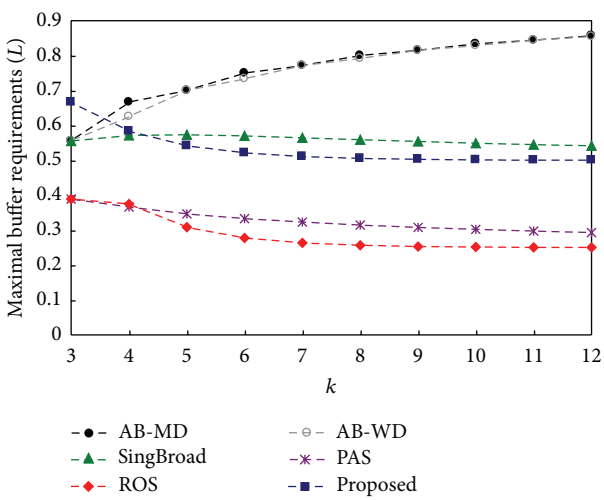


Figure 6: Maximum buffer requirements for AB-MD, AB-WD, SingBroad, PAS, ROS, and the proposed scheme.

480, 240, 194, 194, 175, and 133 s, respectively. In this case, the waiting times for the proposed scheme are 78%, 72%, 45%, 31%, 31%, and 24% smaller than those of AB-MD, AB-WD, HI, SingBroad, PAS, and ROS, respectively.

With low cost and large capacity of storage disks, client buffer demand is no longer a substantial concern. However, for completeness, this work studies the required buffer size under AB-MD, AB-WD, SingBroad, PAS, ROS, and the proposed scheme (the comparison does not include HI, because its buffer requirements are not provided in [16]). Because this study did not derive a close formula for the required buffering space of the proposed scheme, a simulator in Perl [20] was developed to exhaustively search all possibilities to determine the maximum buffering space required at various broadcasting bandwidths. Figure 6 shows the client buffer requirements regarding video length L, where the server bandwidth is varied from $3b$ to $12b$. The proposed scheme initially requires the largest buffering space. However, as the server bandwidth increases, the client buffer requirements drop and approach 50% of video size. Therefore, the proposed scheme yields smaller buffer requirements than AB and SingBroad.

5. Conclusion

A VoD system typically allocates a dedicated stream for each incoming video request; however, individual stream allocation easily causes the system overloaded. This study thus presents a fundamentally different approach by focusing solely on a class of applications identified as latency tolerant applications. Because video broadcasting does not provide interactive functions, a client is able to tolerate playback latency. One efficient broadcasting method is periodic broadcasting, which divides a video into smaller segments and broadcasts them periodically on multiple channels. However, the implementation of multiple-channel broadcasting is difficult and complicated. Therefore, this study proposes a novel single-channel broadcasting scheme for more efficient video delivery. The correctness of the scheme is verified mathematically. The performance comparisons show that, with the same settings of broadcasting bandwidth, the proposed scheme yields the shortest waiting time when compared with AB, HI, SingBroad, PAS, and ROS.

Acknowledgment

This work was financially supported by National Science Council, Taiwan, under a Research Grant no. NSC 101-2221-E-152-004.

References

[1] TechNavio, "Global Video on Demand Market 2011–2015," August 2012.

[2] Digital TV Research, "A sustained boom forecast for global online TV and video," October 2012.

[3] J. Choi, A. S. Reaz, and B. Mukherjee, "A survey of user behavior in VoD service and bandwidth-saving multicast streaming schemes," *IEEE Communications Surveys and Tutorials*, vol. 14, no. 1, pp. 156–169, 2012.

[4] K. Mayer-Patel and A. Jones, "StrandCast: peer-to-peer content distribution for latency tolerant applications," in *Proceedings of the 2nd International Conference on Communication Systems and Networks, (COMSNETS '10)*, pp. 1–10, Bangalore, India, January 2010.

[5] L.-S. Juhn and L. M. Tseng, "Fast data broadcasting and receiving scheme for popular video service," *IEEE Transactions on Broadcasting*, vol. 44, no. 1, pp. 100–105, 1998.

[6] Z.-Y. Yang, *The telepresentation system over internet with latecomers support [Ph.D. thesis]*, Department of Computer Science and Information Engineering, National Central University, Taiwan, 2000.

[7] L.-S. Juhn and L.-M. Tseng, "Harmonic broadcasting for video-on-demand service," *IEEE Transactions on Broadcasting*, vol. 43, no. 3, pp. 268–271, 1997.

[8] Y.-C. Tseng, M.-H. Yang, and C.-H. Chang, "A recursive frequency-splitting scheme for broadcasting hot videos in VOD service," *IEEE Transactions on Communications*, vol. 50, no. 8, pp. 1348–1355, 2002.

[9] H.-F. Yu, H.-C. Yang, and L.-M. Tseng, "Reverse Fast Broadcasting (RFB) for video-on-demand applications," *IEEE Transactions on Broadcasting*, vol. 53, no. 1, pp. 103–111, 2007.

[10] H.-F. Yu, "Hybrid broadcasting with small buffer demand and waiting time for video-on-demand applications," *IEEE Transactions on Broadcasting*, vol. 54, no. 2, pp. 304–311, 2008.

[11] Y. N. Chen and L. M. Tseng, "An efficient periodic broadcasting with small latency and buffer demand for near video on demand," *International Journal of Digital Multimedia Broadcasting*, vol. 2012, Article ID 717538, 7 pages, 2012.

[12] J. F. Paris, "A fixed-delay broadcasting protocol for video-on-demand," in *Proceedings of the 10th International Conference on Computer Communications and Networks (ICCCN '01)*, pp. 418–423, Scottsdale, Ariz, USA, October 2001.

[13] H. F. Yu, P. H. Ho, and H. C. Yang, "Generalized sequence-based and reverse sequence-based models for broadcasting hot videos," *IEEE Transactions on Multimedia*, vol. 11, no. 1, pp. 152–165, 2009.

[14] Z.-Y. Yang, Y.-M. Chen, and L.-M. Tseng, "A seamless broadcasting scheme with live video support," *International Journal of Digital Multimedia Broadcasting*, vol. 2012, Article ID 373459, 8 pages, 2012.

[15] T. Yoshihisa, M. Tsukamoto, and S. Nishio, "A scheduling scheme for continuous media data broadcasting with a single channel," *IEEE Transactions on Broadcasting*, vol. 52, no. 1, pp. 1–10, 2006.

[16] T. Yoshihisa, M. Tsukamoto, and S. Nishio, "A broadcasting scheme considering units to play continuous media data," *IEEE Transactions on Broadcasting*, vol. 53, no. 3, pp. 628–635, 2007.

[17] Y. W. Chen and C. Y. Hsieh, "SingBroad: a scheduling scheme for broadcasting continuous multimedia data over a single channel," *Computer Networks*, vol. 53, no. 9, pp. 1546–1554, 2009.

[18] Y.-W. Chen and C.-Y. Chen, "PAS: a new scheduling scheme for broadcasting a video over a single channel," *IET Communications*, vol. 5, no. 7, pp. 951–960, 2011.

[19] B. S. Wu, C. C. Hsieh, and Y. W. Chen, "A reverse-order scheduling scheme for broadcasting continuous multimedia data over a single channel," *IEEE Transactions on Broadcasting*, vol. 57, no. 3, pp. 721–728, 2011.

[20] http://www.perl.org/.