Algorithmic Aspects of Wireless Networks
Algorithmic Aspects of Wireless Networks

Guest Editors: Xiuzhen Cheng, Wei Li, and Taieb Znati
Editor-in-Chief
Phillip Regalia, Institut National des Télécommunications, France

Associate Editors

Thushara Abhayapala, Australia
Mohamed H. Ahmed, Canada
Farid Ahmed, USA
Alagan Anpalagan, Canada
Anthony Boucouvalas, Greece
Lin Cai, Canada
Biao Chen, USA
Pascal Chevalier, France
Chia-Chin Chong, USA
Huaiyu Dai, USA
Soura Dasgupta, USA
Ibrahim Develi, Turkey
Petar Djuric, USA
Mischa Dohler, France
Abraham O. Fapojuwo, Canada
Michael Gastpar, USA
Alex Gershman, Germany
Wolfgang Gerstacker, Germany
David Gesbert, France
Fary Z. Ghassemlooy, UK
Christian Hartmann, Germany
Stefan Kaiser, Germany
G. K. Karagiannidis, Greece
Hyung-Myung Kim, Korea
C. C. Ko, Singapore
Visa Koivunen, Finland
Richard Kozick, USA
Bhaskar Krishnamachari, USA
S. Lambotharan, UK
Vincent Lau, Hong Kong
David I. Laurenson, UK
Tho Le-Ngoc, Canada
Wei Li, USA
Yonghui Li, Australia
Tongtong Li, USA
Zhiqiang Liu, USA
Stephen McLaughlin, UK
Sudip Misra, Canada
Marc Moonen, Belgium
Eric Moulines, France
Sayandev Mukherjee, USA
A. Nallanathan, Singapore
Kameswara Rao Namuduri, USA
A. Pandharipande, The Netherlands
Athina Petropulu, USA
Brian Sadler, USA
Ivan Stojmenovic, Canada
Lee Swindlehurst, USA
Sergios Theodoridis, Greece
Lang Tong, USA
Luc Vandendorpe, Belgium
Athanasios V. Vasilakos, Greece
Yang Xiao, USA
Xuexhi Yang, USA
Lawrence Yeung, Hong Kong
Dongmei Zhao, Canada
Weihua Zhuang, Canada
Contents

Algorithmic Aspects of Wireless Networks, Xiuzhen Cheng, Wei Li, and Taieb Znati
Volume 2007, Article ID 52861, 2 pages

A Utility-Based Downlink Radio Resource Allocation for Multiservice Cellular DS-CDMA Networks, Mahdi Shabany, Keivan Navaie, and Elvino S. Sousa
Volume 2007, Article ID 76193, 11 pages

Fault-Tolerant Target Localization in Sensor Networks, Min Ding, Fang Liu, Andrew Thaeler, Dechang Chen, and Xiuzhen Cheng
Volume 2007, Article ID 96742, 9 pages

Physical Layer Built-In Security Analysis and Enhancement Algorithms for CDMA Systems, Tongtong Li, Qi Ling, and Jian Ren
Volume 2007, Article ID 83589, 7 pages

Tree-Based Distributed Multicast Algorithms for Directional Communications and Lifetime Optimization in Wireless Ad Hoc Networks, Song Guo, Oliver W. W. Yang, and Victor C. M. Leung
Volume 2007, Article ID 98938, 10 pages

Rate-Based Active Queue Management for TCP Flows over Wired and Wireless Networks, Jun Wang and Min Song
Volume 2007, Article ID 54038, 8 pages

Radar Sensor Networks: Algorithms for Waveform Design and Diversity with Application to ATR with Delay-Doppler Uncertainty, Qilian Liang
Volume 2007, Article ID 89103, 9 pages

Communication Timing Control with Interference Detection for Wireless Sensor Networks, Yuki Kubo and Kokuke Sekiyama
Volume 2007, Article ID 54968, 10 pages

Bandwidth Optimization in Centralized WLANs for Different Traffic Types, R. J. Haines, N. Fanning, T. Lewis, and J. Coon
Volume 2007, Article ID 23917, 14 pages
Recent advances in electronic and computer technologies have paved the way for the proliferation of ubiquitous wireless networks. Fast deployment of these communication networks for the users is preferred under many situations. Topics that are related to ad hoc and sensor networking, mobile computing, and wireless and mobile security have been extensively studied recently. Potential applications of these networks include search and rescue, smart homes, battlefield surveillance, environment monitoring and control, and so forth.

In response to the above demand for wireless networks, this special issue aims at providing a timely and concise reference of the current activities and findings in the relevant technical fields, as well focuses on the state-of-the-art and up-to-date efforts in the algorithmic aspects of wireless networks include location management, topology control and coverage, security and privacy, scalable design, cross-layer design, resource optimization, QoS, to just name a few. We believe that almost all papers included in this special issue not only provide novel ideas, new analytical models, simulation and experimental results, and handful experience in this field, but also simulate the future research activities in the area of the quality of service for mobile ad hoc networks. A brief summary of each paper is listed as follows.

The first paper, by M. Shabany et al., proposes a novel framework to model downlink resource allocation problem in multiservice direct sequence code division multiple access (DS-CDMA) cellular networks. This framework is based on a defined utility function, which leads to utilizing the network resources in a more efficient way. This utility function quantifies the degree of utilization of resources. As a matter of fact, using the defined utility function, users’ channel fluctuations and their delay constraints along with the load conditions of all BSs are all taken into consideration. Unlike previous works, the authors solve the problem with the general objective of maximizing the total network utility instead of maximizing the achieved utility of each base-station (BS). It is shown that this problem is equivalent to finding the optimum BS assignment throughout the network, which is mapped to a multidimensional multiple-choice Knapsack problem (MMKP). Since MMKP is NP-hard, a polynomial-time suboptimal algorithm is then proposed to develop an efficient base-station assignment. Simulation results indicate a significant performance improvement in terms of achieved utility and packet-drop ratio.

The second paper, by M. Ding et al., introduces the authors’ exploratory work toward identifying the targets in sensor networks with faulty sensors. They explore both spatial and temporal dimensions for data aggregation to decrease the false alarm rate and improve the target position accuracy. To filter out extreme measurements, the median of all readings in a close neighborhood of a sensor is used to approximate its local observation to the targets. The sensor whose observation is a local maximal computes a position estimate at each epoch. Results from multiple epochs are combined together to further decrease the false alarm rate and improve the target localization accuracy. Their algorithms have low computation and communication overheads. Simulation study demonstrates the validity and efficiency of their design.

The third paper, by T. Li et al., analyzes security weakness of the operational and proposed CDMA systems and presents an encryption-based secure scrambling process. First, instead of using the long code sequences generated by the LFSR directly, the scrambling sequences are generated through AES operations. As a result, the physical layer built-in security of the CDMA system is significantly increased with very limited complexity load. Second, it is shown that by scrambling the training sequence and the message sequence separately with two independent scrambling sequences, both information privacy and system performance, can be further...
improved. Finally, error-tolerant decryption can be achieved through secure scrambling. The proposed scheme is very feasible and can be readily implemented for security enhancement in wireless networks.

The fourth paper, by S. Guo et al., considers the problem of maximizing the network lifetime for both single and multiple multicast connections in a wireless ad hoc network (WANET) that use omnidirectional/directional antennas and have limited energy resources. Unlike most centralized multicast algorithms, the authors provide a globally optimal solution to this problem in a distributed manner for the special case of single multicast session in a WANET with omnidirectional antennas. This graph-theoretic approach provides us with insights into more general case of using directional antennas, and inspires us to produce a group of distributed algorithms. Experimental results show that our distributed algorithms outperform other centralized multicast algorithms significantly in terms of network lifetime for both single session and multiple session scenarios.

The fifth paper, by J. Wang and M. Song, analyzes existing AQM schemes and proposes a rate-based exponential AQM (REAQM) scheme. The proposed REAQM scheme uses input rate as the primary metric and queue length as the secondary metric. The objectives are to stabilize the system with low packet delay, low packet loss, and high link utilization regardless the dynamic of network conditions. The authors prove the global asymptotic stability of the equilibrium based on Lyapunov theory. Simulation results indicate that REAQM is capable of performing well for TCP flows over both wired and wireless links, and has comparable implementation complexity as other AQM schemes.

The sixth paper, by Q. Liang, firstly performs some theoretical studies on radar sensor network (RSN) design based on linear frequency modulation (LFM) waveform: (1) the conditions for waveform coexistence, (2) interferences among waveforms in RSN, (3) waveform diversity in RSN. Then the author applies RSN to ATR with delay-doppler uncertainty and proposes maximum-likelihood (ML) ATR algorithms for fluctuating target and nonfluctuating target. Simulation results show that the author’s RSN vastly reduces the ATR error comparing to a single radar system in ATR with delay-doppler uncertainty. The proposed waveform design and diversity algorithms can also be applied to active RFID sensor networks and underwater acoustic sensor networks.

The seventh paper, by Y. Kubo and K. Sekiyama, deals with a novel communication timing control for wireless networks and radio interference problem. Communication timing control is based on the mutual synchronization of coupled phase oscillatory dynamics with a stochastic adaptation, according to the history of collision frequency in communication nodes. Through local and fully distributed interactions in the communication network, the coupled phase dynamics self-organizes collision-free communication. In wireless communication, the influence of the interference wave causes unexpected collisions. Therefore, they propose a more effective timing control by selecting the interaction nodes according to received signal strength.

The last paper, by R. J. Haines et al., reviews and compares different mechanisms, to achieve this end, and a number of different means of obtaining the configurations themselves. They describe an analytical model of the system under consideration and present two mathematical approaches to derive solutions for any system configuration and deployment, along with an adaptive feedback-based solution. They also describe a comprehensive simulation-based model for the problem, and a prototype that allows comparison of these approaches. Their investigations demonstrate that a self-adaptive dynamic approach far out-performs any static scheme, and that using a mathematical model to produce the configurations themselves confers several advantages.

In closing, we would like to thank the support from the Editor-in-Chief, Phillip Regalia, and the contributions from authors and reviewers, to make this special issue possible.

Xiuzhen Cheng
Wei Li
Taieb Znati
Research Article

A Utility-Based Downlink Radio Resource Allocation for Multiservice Cellular DS-CDMA Networks

Mahdi Shabany,1 Keivan Navaie,2,3 and Elvino S. Sousa1

1Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada M5S 3G4
2Department of Electrical and Computer Engineering, Faculty of Engineering, Tarbiat Modares University, P.O. Box 14155–4838, Tehran, Iran
3The Broadband Communications and Wireless Systems (BCWS) Center, Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Received 30 May 2006; Revised 1 December 2006; Accepted 8 January 2007

Recommended by Wei Li

A novel framework is proposed to model downlink resource allocation problem in multiservice direct-sequence code division multiple-access (DS-CDMA) cellular networks. This framework is based on a defined utility function, which leads to utilizing the network resources in a more efficient way. This utility function quantifies the degree of utilization of resources. As a matter of fact, using the defined utility function, users’ channel fluctuations and their delay constraints along with the load conditions of all BSs are all taken into consideration. Unlike previous works, we solve the problem with the general objective of maximizing the total network utility instead of maximizing the achieved utility of each base station (BS). It is shown that this problem is equivalent to finding the optimum BS assignment throughout the network, which is mapped to a multidimensional multiple-choice knapsack problem (MMKP). Since MMKP is NP-hard, a polynomial-time suboptimal algorithm is then proposed to develop an efficient base-station assignment. Simulation results indicate a significant performance improvement in terms of achieved utility and packet drop ratio.

Copyright © 2007 Mahdi Shabany et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Third generation wireless cellular networks provide a variety of services ranging from multimedia to Internet access. In order to enable these services cellular networks are required to support multiple classes of traffic with diverse quality-of-service (QoS) requirements. Due to the limited availability of radio resources, designing a resource control mechanism to utilize the network resources efficiently is a crucial task for the next generation cellular communication systems. However, designing an optimal resource allocation scheme in CDMA cellular networks is a challenging problem especially when different parameters are involved in the system such as the rate, QoS, and delay requirements of various services.

The optimization can be performed either in the network level or in the cell level. Conventional methods for resource allocation in wireless networks are based on the characterization of traffic flows. In these methods the objective is either to minimize base-station power consumption or to maximize the system capacity [1–4]. There are two major limitations in these approaches: they require the traffic characteristics of each flow, which may be difficult to obtain unless standard assumptions such as Poisson traffic are made. Furthermore, admission and access control must be considered in conjunction with the resource allocation mechanism. Moreover, these classical approaches fail to address the throughput-delay tradeoff efficiently [5].

For the multirate delay-constrained services, as in 3G, the conventional approaches are not effective enough in terms of the optimization of the network resources. Therefore, an alternative approach that avoids the above limitations is required. An efficient approach, which surmounts this challenge, is to assign a utility function to each user based on its QoS requirements and channel status. This utility function represents the benefit that the network can earn by serving that user. In other words, by introducing the utility function, no matter how many various services are involved in the network, each service is specified and integrated in the system modeling via a utility function. This implies that the system
treats multiclass services in a unified way. The utility function then can be used as a tool to design an optimal resource allocation scheme. The objective of the allocation scheme is to optimize the total network utility, which is defined as the summation of all the users’ utility functions.

There is no clear way to define the utility function for multirate delay-constrained services. It is a complicated task because a comprehensive and yet meaningful utility function requires to take all the various aspects of the network and service types into account. Some of these aspects include the channel status, required data rates, and delay constraints of the services.

In this paper, we define a novel utility function for each user that is a function of its channel status, its required service as well as the load condition of the corresponding serving base station. This new definition of the utility function incorporates the information of both the network side (channel) and the user side (rate and delay) in a unified way for radio resource allocation. We focus our attention on the downlink resources (i.e., power and bandwidth), which is considered to be the bottleneck in multiservice systems [6]. To design such a scheme, we take into account the system variations in the physical layer as well as the traffic load of the base stations.

In other words, we propose a utility-based base station assignment and resource scheduling scheme for the downlink in multiservice cellular DS-CDMA networks. Unlike previous works, we solve the problem with the general objective of maximizing the total network utility (multiple base) instead of maximizing the utility of each base station (BS) individually. The scheme can be considered as a scheduler determining the set of users that should be served within each time slot. For the special case of having only packet traffic the work in this paper is a general case of the work in [7, 8].

2. LITERATURE REVIEW

Radio resource allocation for the downlink in a DS-CDMA cellular network is considered in [9, 10] based on the joint power allocation and base-station assignment. A pricing framework based on the utility concept has been introduced in [11]. Using this concept, the uplink resource allocation for power and spreading gain control for one type of non-real-time service is studied in [12]. Utility-based modeling is also utilized for uplink power control in a single service multicell data network in [13]. In the proposed method in [13], QoS for data users is modeled through a utility function that indicates the value of information per assigned power level (bits per Joule). Using the utility function the problem is solved by modeling it as a noncooperative game where each user tries to maximize its own utility.

For multiservice cellular networks with a mixture of symmetric and asymmetric services, it has been shown that in most cases the downlink performance is more critical than that of the uplink [6]. For the downlink, the power allocation problem for multiservice DS-CDMA wireless networks is studied in [14], where the downlink power control problem for multicell wireless networks is formulated as a noncooperative game, although they do not consider downlink power limitation. In practice, transmission power limitation in DS-CDMA cellular systems is a major concern. Therefore, it is necessary to develop algorithms for the power-constrained case as it is presented in this paper.

The pricing framework is also used in [15] to develop a distributed joint power allocation and base-station assignment with the objective of maximization of the total network utility. However, in the strategy adopted in [15], each base station tries to maximize its total utility without considering the status of others. Therefore, the proposed scheme does not necessarily result in maximum total network utility. Furthermore, other QoS parameters such as delay constraint is not discussed. An opportunistic transmission scheduling with resource-sharing constraints has been proposed in [16], which exploits time-varying channel conditions in a single cell. However, the user’s delay constraint is not taken into account in [16]. Moreover, their proposed utility function only depends on the channel status in the time slot that the user is being served.

Downlink resource allocation problem for multicell multiservice DS-CDMA system is also studied in our previous works [17, 18]. Both papers, besides per-user throughput, take into account delay requirements of data services as well. The optimum power allocation scheme in a multiservice environment, which supports both data and real-time services, is then modeled using the multiple-choice multidimensional knapsack problem (MMKP); however, the detailed analysis of the problem as well as corresponding heuristic algorithm for MMKP has not been presented in [17, 18].

In our later work [7], we show that optimal packet scheduling in a packet-oriented cellular CDMA/TDMA network can also be modeled as an MMKP. Exploiting delay tolerance of data traffic, we then introduced the notion of multiaccess-point diversity, which is a potential form of diversity in cellular networks, where a signal can be transmitted to the corresponding mobile user via multiple base stations. In [8] we derived analytical performance gain bound on multiaccess-point diversity.

3. SYSTEM MODEL

We consider a time hierarchy for wireless cellular systems where there are three main types of temporal variations in the system.

1. Small-scale variation that is mainly due to the fast fading effect of wireless channel. Fast fading is a consequence of multipath propagation due to reflections of the signal by physical obstacles. We consider $T_f$ second as the time-scale of small-scale variations, that is, fading is assumed to be constant during each $T_f$ seconds.

2. Medium-scale variations that is because of the shadowing effect. Shadowing is the result of the existence of some obstacles between the transmitter and the receiver, usually modeled by a log-normal distribution. Here $T_w$ indicates the time-scale of the medium-scale variations.

3. Large-scale variations that is due to the mobility of users in the network, which results in variations in the system
connectivity. In this paper, $T_p$ indicates the time scale of such variations.

In each time scale, appropriate mechanisms should be utilized to manage the above variations. In this paper, a multiservice DS-CDMA cellular network is considered. Base-stations and users are nonuniformly distributed in the network coverage area. This system supports both real-time and nonreal-time (data) services. Real-time services include voice while being in the coverage of base-station $i$, $\Gamma_{i,sj}$, can be written as

$$\Gamma_{i,sj} = \frac{W}{R_j \sum_{k=1, k \neq i}^{M} P_{Tk} g_{k,sj} + (1 - \alpha)(P_{Ti} - P_{i,sj}) g_{i,sj} + \eta}$$

(1)

for all $i$ in $B$, $s$ in $RT$, and $j$ in $N_i$, where $W$ is the chip rate, $r_s$ is the data rate of user $j$, and $\eta$ is the spectral density of the additive white Gaussian noise. The term in the numerator represents the desired received power at the location of the user $j$, while being in the coverage of base-station $i$, and $g_{i,sj}$ is the gain between the base-station $i$ and user $j$ of the class $s$, which accounts for the effect of path loss, as well as the large scale fading (shadowing). A fast power control is assumed to be running with a separate mechanism, and the outer loop power control is performed within each $T_p$ seconds.

The first term in denominator represents the total received interference from the other base stations, inter cell interference, while the second term shows the intra cell interference, resulted from the portion of the power of base-station $i$ that is allocated to the other users within the coverage area of the base-station $i$, $P_{Ti} - P_{i,sj}$. The parameter $\alpha$ is the orthogonality factor that is due to the effect of the multi-path fading.

Based on (1), the achieved rate of each user, $r_j$, depends directly on the amount of allocated power to that user by its base station, $P_{i,sj}$, as well as its received interference. Basically these are the two main factors that enable us to manage the total capacity of the system. Using the above definitions, the problem of optimal power allocation to real-time users is formulated as the following classic downlink power control.

### 4. Base-Station Assignment

The system is time slotted and at any time slot each base station first allocates power to the real-time users.

#### 4.1. Real-time users

We consider a system with hexagonal cells including a central cell and the cells in its first and second tier. The received bit-energy-to-interference-plus-noise-spectral-density ratio of user $j$ served by service $s$ while being in the coverage of base-station $i$, $\Gamma_{i,sj}$, can be written as

$$\Gamma_{i,sj} = \frac{W}{R_j \sum_{k=1, k \neq i}^{M} P_{Tk} g_{k,sj} + (1 - \alpha)(P_{Ti} - P_{i,sj}) g_{i,sj} + \eta}$$

(1)
problem:

\[
\begin{align*}
\min \left\{ \sum_{i=1}^{M} \sum_{j=1}^{N_R} P_{ij}, \right. \\
\text{s.t. } 0 \leq \sum_{j=1}^{N_R} P_{ij} \leq P_{Ti},
\end{align*}
\]

where (4) denotes the constraint for the maximum allowable BS transmit power that can be assigned based on an upper layer mechanism (i.e., managed by RNC). Constraint (4) indicates the air interface QoS satisfaction of the real-time users. The allocated power based on the downlink power control is the solution of (3), (e.g., see [19–21]).

### 4.2. Nonreal-time traffic

After power allocation to the real-time users, the available power for allocation to the nonreal-time data users is upper bounded by the remaining power of each base-station, which comes from the hardware limitation. We denote this available upper power for allocation to the nonreal-time data users is

\[
P_{Ri}(n) = P_{Ti}(n) - \sum_{s \in RT} \sum_{j=1}^{N_R} P_{ij}(n).
\]

The solution of (3) results in maximum available power. Note that all of the remaining resource of the system because of the more interference generated in the system by admitting more and more nonreal-time users. Therefore, to prevent real-time users’ call degradation after power allocation to nonreal-time users, someone may allocate powers to the real-time users’ call degradation after power allocation to nonreal-time users. Therefore, at the end all real-time users will experience an acceptable level of QoS. The bit energy to the interference spectral density ratio for user \(j\) of the base-station \(i\) served by the service \(s\) is

\[
\Gamma_{ij} = \frac{W p_{ij} g_{ij}}{R_j(\Gamma_{ij} + \eta_j)} \geq \Gamma_s,
\]

where \(\Gamma_s\) is the minimum required \(E_b/I_0\) of the service \(s\), \(W\) is the chip rate, \(\eta_j\) is the additive white Gaussian noise at the user \(j\)’s receiver, and \(I_{ij}\) is the total received interference at the location of user \(j\) served by the base-station \(i\) calculated by RNC as follows:

\[
I_{ij}(n) = \sum_{k=1,k \neq i}^{M} P_{Tk}g_{kj} + (1 - \alpha) P_{Ti}g_{ij}.
\]

Based on (6), data rate of each user depends on its allocated power, \(p_{ij}\), channel gain, \(g_{ij}\), and received interference, \(I_{ij}\). Hereafter, we simply refer to \(g_{ij}(n)/I_{ij}(n)\) as the channel status and drop subscript \(s\) for the brevity of discussion.

Providing service to a user with poor channel status would require more air interface resources such as transmission power, \(p_{ij}\), or longer transmission time due to a lower data rate. As a result, providing the service to a user with better channel status leads to an efficient system resource utilization. On the other hand, among users with the same channel status, providing service to users with less remaining tolerable delay leads to QoS satisfaction of these users while does not degrade the service level of the others. Therefore, utility-based resource allocation is the technique of choice, where the user’s service and channel quality is jointly integrated and considered by a utility function, which is used as a tool to optimize the resource allocation scheme.

### 4.3. Utility-based resource allocation

Considering the delay tolerance of a nonreal-time data user, the network can wait for a good channel status and then send to that user. This idea has been used in recently proposed methods based on utility-based resource control [13, 15]. In these methods, the total network throughput is maximized subject to a set of QoS and resource constraints. For each user, a utility function is defined as an indicator of user’s achieved throughput.

In the case where each user has a finite delay constraint, the user’s throughput can only indicate the user’s satisfaction if it is served in its predetermined tolerable delay period. Taking a network side insight, for a data user with a given maximum delay tolerance, serving that user can be done during its maximum delay tolerance period. This is an opportunity for the network to postpone serving that user and serve other users with better channel status, which corresponds to the less air interface resource to be allocated, and/or a worse delay condition. In this paper, we define a novel utility function that shows the network’s benefit due to the above mentioned opportunity.

For user \(j\) being served by the BS \(i\) in time-slot \(n\), we propose the utility function as

\[
\begin{align*}
\Phi_{ij}(n) &= \left\{ \begin{array}{ll}
\Phi(d_j(n))\Psi(I_{ij}(n)), & i \in AS_j, \\
0, & \text{otherwise}
\end{array} \right.
\end{align*}
\]

where \(d_j(n)\) is the remaining tolerable delay of user \(j\), \(\Phi(\cdot)\) is an increasing function of \(1/d_j(n)\), and \(\Psi(\cdot)\) is the probability of success in packet transmission that is assumed to be an increasing function of \(\Gamma_{ij}(n)\), defined in (6). The function \(\Phi(d_j(n))\) manages the delay-throughput tradeoff by increasing the priority of the users with a given minimum delay tolerance, while \(\Psi(\Gamma_{ij}(n))\) characterizes multiaccess-point and multiuser diversity gains. For instance, from two users with the same channel status, the one with less \(d_j(n)\) has the higher priority to be served by the network, while between two users with the same delay constraint, the one with a better channel status is served first. In brief, the utility function
defined in (8) is a decreasing function of $d_j(n)$, which has its maximum value at $d_j(n) = 0$.

Total network utility, $Q : \overrightarrow{u}$, is defined as a function of the individual utilities of the users that are assigned to the BSs, where $\overrightarrow{u} \triangleq (u_{1b_1}, u_{2b_2}, \ldots, u_{N_b})$ is the utility vector, index $b_i$ shows the assigned BS to the user $j$, and $Q(\cdot)$ is a casual policy defined based on the network performance perspective.

The mathematical definition of $Q(\cdot)$ is related to the service provider’s resource management strategy and generally is as follow:

$$Q(\overrightarrow{u}) \triangleq \sum_{j=1}^{N} \sum_{i=1}^{M} u_{ij}(n)x_{ij}(n), \quad (9)$$

where $x_{ij}(n)$ is the assignment indicator in time-slot $n$, that is, $x_{ij}(n) = 1$ if BS $i$ is assigned to user $j$ and $x_{ij}(n) = 0$, otherwise. If a specific user is not assigned to the network at time-slot $n$, this means that a BS that is out of its active set is selected for serving. Therefore, by the definition in (8), its corresponding utility would be zero. The total network utility represents the total benefit that network earns by serving the users while their delay requirements are also being satisfied.

In this paper, the total network utility is defined as the sum of all individual user’s utility. In other words, the higher network utility shows the more resource control efficiency in terms of providing service to the users with the maximum achievable utility.

5. BASE-STATION ASSIGNMENT ALGORITHM

In this paper, our objective is to maximize the total network utility. Such optimization leads to maximizing the total allocated data rate in the network while considering the channel status, and the delay constraints of all users. In other words, maximizing the total network utility shows that the network waits intelligently for a better accessible channel status for each user while considering its maximum tolerable delay. Based on (8), the utility function of a user depends on its assigned base station. Therefore, for a given set of available powers for nonreal-time users, the problem of maximizing the total utility of the network leads to the problem of finding the optimum base-station assignment, which is implemented by RNC.

In DS-CDMA networks, for each user, the base-station assignment is performed based on the selection of a base-station whose corresponding received $E_c/I_0$, the bit energy of pilot channel to the total received interference spectral density, is the maximum. In other words, each user has an active set of base stations from which it chooses its best server. This active set is defined as a set of base stations whose corresponding received $E_c/I_0$ are greater than a performance threshold, that is,

$$AS_j = \{i \mid i \in B, \ (E_c/I_0)_{ij} \geq \gamma_{\text{min}}\}, \quad (10)$$

where $\gamma_{\text{min}}$ is the minimum required $E_c/I_0$.

In this case, in selecting the best server for each user, the traffic profile of the network and the target base station is not taken into account while in our scheme it is possible for a specific user, whose best server is overloaded, to be served by another base station in its active set with better load condition. Therefore, the total utility of the network can be improved.

Here, we propose a base-station assignment mechanism, which selects the best server of each user to maximize the total network utility. The input of the algorithm consists of the values of the utility functions of all users, which can be defined in an arbitrary but meaningful way. Therefore, our proposed modeling can be applied in a more general case by any definition of utility. Let $P_R = [P_{R1}, \ldots, P_{RM}]$ be the vector of base-stations’ remaining powers. Therefore, the optimal base-station assignment in the time-slot $n$ is a solution of the following optimization problem:

$$\max_{x_{ij}} \left( \sum_{j=1}^{N} \sum_{i=1}^{M} u_{ij}(n)x_{ij}(n) \right), \quad (11)$$

s.t. $\sum_{j \in A_i} p_{ij}(n)x_{ij}(n) \leq P_{Ri}(n), \ \forall i \in B, \quad (12)$

$$\sum_{i=1}^{M} x_{ij}(n) = 1, \ x_{ij}(n) \in \{0, 1\} \ \forall j = 1, \ldots, N, \quad (13)$$

where $x_{ij}(n)$ is one if the user $j$ is assigned to the base-station $i$ at the time-slot $n$, and zero, otherwise. For the brevity of discussion in the following we drop the time index $n$.

Let $MS_i = \{ j \mid i \in AS_j \}$ be the set of nonreal-time users that base station $i$ is in their active sets. The total required power to serve a valid subset of $MS_i$ should be smaller than or equal to $P_{Ri}$. Each user is assumed to be served by only one base-station. Therefore, a feasible base-station assignment, $\Omega_m$, is the combination of nonintersect valid subsets of $MS_{\Omega_m}, i = 1, \ldots, M$. A valid subset means a subset whose sum of required powers of its individual users is less than or equal to the total remaining power of its corresponding base-station. Our objective is to find $\Omega_m$ as its corresponding total utility, $U(\Omega_m^*)$, such that

$$m^* = \arg \max_m U(\Omega_m^*). \quad (14)$$

The base-station assignment scheme is summarized in Algorithm 1.

In the following, we map the downlink resource allocation problem in (12) to a multidimensional multiple-choice knapsack problems (MMKP).
Definition 1 (MMKP). An MMKP is the problem where there is an $M$-dimensional knapsack with $M$ total allowable volumes of $W_1, W_2, \ldots, W_M$ and there are $N$ groups of items. Group $j$ has $n_j$ items. Each item has a value and $M$ volumes corresponding to the knapsack’s $M$ dimensions. The objective of the MMKP is to pick up exactly one item from each group for the maximum total value of the selected items, subject to the volume constraints of the knapsack’s dimensions. In mathematical representation, let $w_{kj}$ be the value of the $k$th item of the $j$th group, let $\omega_{kj} = (w_{k1}, \ldots, w_{kM})$ be the required volume of the $k$th item of the $j$th group corresponding to $M$ dimensions, and let $\omega = (W_1, \ldots, W_M)$ be the volume constraints of different knapsack’s dimensions. Then the problem can be written as

$$\begin{align*}
\max_{x_{kj}} & \sum_{j=1}^{N} n_j \sum_{k=1}^{M} x_{kj} u_{kj}, \\
\text{s.t.} & \sum_{j=1}^{N} n_j x_{kj} \omega_{kj} \leq W_i \quad \forall i \in \{1, \ldots, M\}, \\
& \sum_{k=1}^{M} x_{kj} = 1 \quad \forall j \in \{1, \ldots, N\}, x_{kj} \in \{0, 1\}.
\end{align*}$$

(15)

5.1. Algorithm for optimal base-station assignment

Problem (12) is mapped to a multidimensional multiple-choice knapsack problem (MMKP) as follows. We consider $M$ base stations as a knapsack with $M$ dimensions and each user as a group. Each group has $n_j$ (here $M$) items equal to the number of base stations. Item $k$ of the user $j$ has a value $u_{kj}$ defined in (8), that is, the utility of user $j$ when it is assigned to the base-station $k$, and $M$ volumes $p_{kj} = (p_{1jk}, \ldots, p_{Mjk})$, which is defined as

$$p_{ijk}(n) = \begin{cases} p_{ijk}, & k \in AS_j, i = k, \\ 0, & \text{otherwise,} \end{cases}$$

(16)

which ensures that item $k$ of any group (user), that corresponds to base-station $k$, can only be assigned to base-station $k$, which is meaningful.

Therefore, if item $k$ of group $j$ is selected in the optimal solution, it means that the user $j$ has been assigned to the base-station $k$, its corresponding achieved utility is $u_{kj}$, and the amount of power it takes from the base-station $k$ is $p_{ijk}$. We have to choose exactly one item from each group, meaning that each user can be assigned to at most one base station. It is worth mentioning that by the definition of MMKP we have to choose exactly one item from each group. However, the selection of all users is not feasible in many cases. Therefore, if user $j$ does not exist in the optimal solution it means that one of its items whose corresponding value and volumes are zero has been selected. This indirectly implies that user $j$ has not been assigned to the network.

Using above mapping, problem (12) can be rewritten as

$$\begin{align*}
\max_{x_{kj}} & \sum_{j=1}^{N} \sum_{k=1}^{M} x_{kj} u_{kj}, \\
\text{s.t.} & \sum_{j=1}^{N} \sum_{k=1}^{M} x_{kj} p_{ijk} \leq P_{Ri} \quad \forall i \in B, \\
& \sum_{k=1}^{M} x_{kj} = 1 \quad \forall j \in \{1, \ldots, N\}, x_{kj} \in \{0, 1\},
\end{align*}$$

(17)

where $x_{kj}$ is one when the item $k$ of user $j$ is selected.

Since the problem was formulated as an MMKP, any technique available to solve MMKP can be used. Generally, there are two approaches to solve an MMKP; exact and heuristic. The exact solution is based on the branch-and-bound algorithm [22]. The computational complexity of these algorithms is $O(2^M N)$. Therefore, branch-and-bound linear programming approach (BBLP) is often too slow to be useful for radio resource allocation. The alternative is to use a heuristic approach. There are some heuristic algorithms in the literature like the ones in [23, 24]. We use the modified version of [24] to solve our MMKP. Here, we briefly outline some of the known theory on Lagrange multipliers and the algorithm for solving our MMKP to simplify the understanding of our approach.

Theorem 1 (see [25]). Let $\lambda_1, \ldots, \lambda_M$ be $M$ nonnegative Lagrange multipliers, and let $x_{kj}^* \in \{0, 1\}$ be the solution of

$$\max \left\{ \left( \sum_{j=1}^{N} \sum_{k=1}^{M} x_{kj} u_{kj} \right) - \sum_{i=1}^{M} \lambda_i \sum_{j=1}^{N} \sum_{k=1}^{M} x_{kj} p_{ijk} \right\},$$

(20)

then the binary variables $x_{kj}^*$ are also the solution to

$$\max_{x_{kj}} \sum_{j=1}^{N} \sum_{k=1}^{M} x_{kj} u_{kj},$$

(21)

$$\sum_{j=1}^{N} \sum_{k=1}^{M} x_{kj} p_{ijk} \leq \sum_{i=1}^{M} \lambda_i x_{kj}^* p_{ijk}.$$  

(22)

Theorem 1 is the fundamental result that makes Lagrange multipliers applicable to discrete optimization problems such as the MMKP. According to this theorem, the solution to the unconstrained optimization problem (20) is also the solution to the constraint optimization problem (22), which is our MMKP with the constraint values $P_{Ri}$ replaced by $\sum_{j=1}^{N} \sum_{k=1}^{M} x_{kj}^* p_{ijk}$. Therefore, if the multipliers $\lambda_i$ are known, the optimization problem is easily solved, because by a simple manipulation equation (20) can be written as

$$\max \left\{ \sum_{j=1}^{N} \sum_{k=1}^{M} \left( u_{kj} - \sum_{i=1}^{M} \lambda_i p_{ijk} \right) x_{kj} \right\},$$

(23)

which in turn implies that the solutions are

$$x_{kj}^* = \begin{cases} 1, & \text{if } u_{kj} - \sum_{i=1}^{M} \lambda_i p_{ijk} > 0, \\
0, & \text{otherwise.} \end{cases}$$

(24)
I. INITIALIZATION PHASE

\[
\begin{align*}
\lambda_i &\leftarrow 0 \quad \forall i = 1, \ldots, M; \\
p_{ij} &\leftarrow p_{ijk}/p_i \quad \forall j = 1, \ldots, N; \forall k = 1, \ldots, n_j; \\
\hat{K}_j &\leftarrow \text{argmax}_k (u_{ij} - \lambda_i (p_{ij} - p_{ijk})) / p_{ijk} \\
T_i &\leftarrow \sum_{j=1}^{N} p_{ij} \hat{K}_j \quad \forall i = 1, \ldots, M; \\
\end{align*}
\]

II. DROP PHASE

While \((T_i > 1\) for any \(i\)) do

\[
\begin{align*}
\hat{I} &\leftarrow \text{argmax}_i \{T_i\} \\
\{ j \mid \hat{K}_j = \hat{I} \} &\leftarrow \{ j \mid \hat{K}_j = \hat{I} \} \\
\hat{K}_j &\leftarrow \text{argmin}_j \{\lambda_j\} \quad \forall j, k \\
\Delta \lambda_j &\leftarrow (u_{ij} - \lambda_j (p_{ij} - p_{ijk})) / p_{ijk} \\
\hat{K}^{*} &\leftarrow \text{argmax}_{j \in \{ j \mid \hat{K}_j = \hat{I} \}} \{\Delta \lambda_j\} \\
\lambda^{*} &\leftarrow \lambda_j + \Delta \lambda^{*} \\
x^{*}_{\hat{K}^{*}} &\leftarrow 0 \\
x^{*}_{\hat{K}^{*}} &\leftarrow 1 \quad (i.e., \hat{K}^{*} = K^{*}) \\
T^{*} &\leftarrow T^{*} - p_{\hat{K}^{*}} \\
T^{*} &\leftarrow T^{*} + p_{K^{*}} \\
\end{align*}
\]

III. ADD PHASE

While more items can be exchanged

\[
\begin{align*}
\mu_{ik} &= \begin{cases} 
(u_{ij} - x^{*}_{\hat{K}_j} - x_{\hat{K}_j} & \text{if } (u_{ij} - u_{\hat{K}_j} > 0, T_{ik} + p_{ijk} \leq 1) \\
0 & \text{otherwise} 
\end{cases} \\
\end{align*}
\]

\[
\begin{align*}
\hat{K}^{*}' &\leftarrow \text{argmax}_{j \in \{ j \mid \hat{K}_j = \hat{I} \}} \{\mu_{ik}\} \\
\hat{T}^{*}_j &\leftarrow \hat{T}^{*}_j - p_{\hat{K}^{*}_j} \\
\hat{T}^{*}_j &\leftarrow \hat{T}^{*}_j + p_{\hat{K}^{*}_j} \\
x^{*}_{\hat{K}^{*}_j} &\leftarrow 0 \\
x^{*}_{\hat{K}^{*}_j} &\leftarrow 1 \quad (i.e., \hat{K}^{*}_j = K^{*}) \\
\end{align*}
\]

Algorithm 2: Heuristic algorithm for base-station assignment.

Since we have another constraint in (19), among the solutions in (24), we have to look for the one which satisfies (19) and is optimal at the same time.

Therefore, the only step to do so is to compute the Lagrange multipliers \(\lambda_i\). It is worth noting that if these multipliers are computed such that the terms \(P_{Ri} - \sum_{j=1}^{N} x_{\hat{K}_j}^{*} p_{ijk}\) are nonnegative, the solution is feasible. The solution is optimal, if the following condition holds:

\[
\sum_{i=1}^{M} \lambda_i \left(P_{Ri} - \sum_{j=1}^{N} \sum_{k=1}^{n_j} x_{\hat{K}_j}^{*} p_{ijk}\right) = 0
\]

(i.e., the case whereby error is zero). The MMKP algorithm is given in Algorithm 2.

5.2. Heuristic algorithm

The algorithm starts with the most valuable item of each user \(j\) as the selected item \((\hat{K}_j)\), and the Lagrange multipliers initialized to zero such that the constraints in (19) and (24) are satisfied, Initialization Phase. In general, however, the volume constraints will now be violated. The initial choice of selected items is adapted to obey the volume constraints by repeatedly improving on the most violated constraint, \(\hat{I}\). This step is done in DROP phase.

Consider the users whose selected items correspond to base-station \(I\) (i.e., \(\{ j \mid \hat{K}_j = \hat{I} \}\)). For each item \(k\) of these users, the increase \(\Delta \lambda_{kj}\) of multiplier \(\lambda_j\), that results from exchanging the selected item of group \(j\), is computed. Eventually, the item \(K^{*}\) of user \(j^{*}\) causing the least increase of multiplier \(\lambda_j\) is chosen for exchange. This choice minimizes the widening of the gap between the optimal solution characterized by (25) and the solution returned by MMKP algorithm. The process is repeated until for each user an item has been selected such that the volume constraints are satisfied. Since each user has always an item whose value and \(M\)-dimension volume are zero corresponding to the base station that is not in its active set, the solution is always feasible.

After completion of Drop Phase, there may be some space left in the knapsack. This space may be utilized to improve the solution by replacing some selected items with more valuable ones. Therefore, in the Add Phase of the algorithm, each item \(k\) of every user \(j\) is checked against the selected item of that user \((\hat{K}_j)\). It is tested whether item \(k\) is more valuable than the selected item, and if \(k\) can replace the selected item without violating the volume constraints. Among all exchangeable items, the item \(K'\) of user \(j'\) causing the largest increase of the knapsack value is exchanged with the selected item of that user \((\hat{K}'_J)\). This process is repeated until no more exchanges are possible. The resulting solution comprised of the selected items is feasible, and even optimal, if (25) is satisfied.

Proposition 1. The maximum difference between the total achieved throughput using above suboptimal algorithm and globally optimal solution is

\[
\sum_{i=1}^{M} \lambda_i \left(P_{Ri} - \sum_{j=1}^{N} \sum_{k=1}^{n_j} x_{\hat{K}_j}^{*} p_{ijk}\right)
\]

where \(x_{\hat{K}_j}^{*}\) are the outputs of the heuristic algorithm.

Proof. See the appendix.

5.3. Computational complexity

Step I is just the initialization whose effect on the time complexity of the algorithm is negligible \(O(M + 3NM + M^2N)\). Drop phase is the determining factor in the complexity of the algorithm. Basically this step can be repeated at most \(NM\) times until no infeasible knapsack \((T_i > 1)\) remains. At each iteration, there are \(NM^2 + NM + 2M\) additions and/or comparisons, which means that the complexity of this phase
is at most \(O(MN(NM^2 + NM + 2M))\). Therefore, ignoring the negligible terms, we end up to the total complexity of \(O(N^2M^3)\), which is polynomial time. For detailed complexity analysis, see [17].

### 6. Simulation Results

We consider a two-tier hexagonal cell configuration with a wrap-around technique [26]. A universal mobile telecommunication system (UMTS), with a fast power controller running at 1500 updates per second, is simulated. Cross-correlation between the codes in a cell at the mobile receiver is assumed to be equal to 0.3. We simulate a mixture of voice and data users; voice services with 12.2 kbps, activity factor of 0.67 and minimum required \(E_b/I_0 = 5\) dB, while data services have minimum required \(E_b/I_0\) of 3 dB. Packet arrival is modeled by a Poisson process.

In this paper, we define

\[
\Phi(d_j(n)) = \begin{cases} 
\exp\left(\frac{1}{T_w + d_j(n)}\right), & 0 \leq d_j(n) \leq \tau_j, \\
0, & \text{otherwise}. 
\end{cases} 
\tag{27}
\]

In fact, any function that is a decreasing function of \(d_j(n)\) will result in the same performance result. It is seen that if \(d_j(n)\) of a user approaches zero, its corresponding \(\Phi(\cdot)\) becomes very high, and overrides channel considerations in (8). Note that when all services have no delay constraint, the problem is simply reduced to the conventional SIR-based base-station assignment.

Channel fading is based on the Gudmundson model with fading standard deviation equal to 6.5 dB. A distance-dependent channel loss with path exponent of \(-4\) is considered. We focus on the central cell and use the delay constraint and channel status of users to determine the utility function for each user relative to the base stations in its active set.

We now compare the gain of our proposed base-station assignment to the conventional SIR-based assignment. Initially, \(N_{\text{uni}}\) users were distributed uniformly throughout all the cells. After that, \(N_{\text{nonuni}}\) users were added to the boundary of the central cell. All users have the same delay constraint. The ratio of total achieved utility of our scheme to that of SIR-based scheme versus the number of added nonuniform users in an 8-set cell corresponding to the central cell and seven cells in its first tier is shown in Figure 1.

It is seen that our proposed scheme performs better for small values of \(N_{\text{uni}}\), which means more total utility is gained when neighboring cells are lightly loaded or have users with more relaxed delay constraints. Therefore, the rate of increase in total utility is maximum for \(N_{\text{uni}} = 2\). This idea is seen more clearly in Figure 2, where the rate of increase in achieved utility for different cases is shown.

It is seen by increasing the number of added nonuniform users in the boundary of the central cell, the performance is better when the number of uniform users is smaller. This is because adjacent cells can serve more users of the central cell when they have a smaller number of users. Moreover, by increasing the number of nonuniform users, \(N_{\text{nonuni}}\), the total achieved gain approaches a steady-state value, which is the maximum capacity that can be obtained using our scheme.

In another scenario, we distributed 5 users in all cells like before, but limited the number of base stations in the active set of each user. Moreover, we considered the results for the two different patterns of nonuniform users’ distributions. In the first case (pattern A), we distributed more users throughout the central cell randomly, while in the second one (pattern B) the users were grouped in subcells located at the cell boundary in the corner of three adjacent cells. The result is shown in Figure 3. It is seen that by increasing the number of allowable BSs in the active set of each user the performance is improved slightly. Moreover, if all nonuniform users are located in the cell boundary for large values of \(N_{\text{nonuni}}\), the total achieved utility is improved while for small values of \(N_{\text{nonuni}}\) the results are almost the same.

We also consider the total network utility as in (12) and compare the system performance for three distinct resource control schemes: SIR-based (SIR-BSA), the individual BS utility maximization (IU-BSA) [15], and the proposed JBATS. Nonuniform user distribution in the network coverage area is expressed by the nonuniformity factor \(\mu_D\), which is the ratio of the users that are distributed nonuniformly to the total number of users. The result is shown in Figure 4.

In order to study the run-time performance of the algorithm, we implemented it along with the optimal algorithm based on branch and bound search using linear programming for upper bound computation. Although branch-and-bound is infeasible in practical application for larger data sets, we run this algorithm to determine the optimality of the heuristics by finding an upper bound using the linear programming approach. We have performed experiments on an extensive set of problem sets where we used randomly generated MMKP instances for our tests. For each set of parameters \(N\) and \(M\), we run the algorithm ten times and tabulated.
the averages of achieved throughput and execution time. Table 2 shows the percentage of the achieved throughput using our heuristic method compared to the value achieved in the optimal case. Moreover, the third column of the table shows the required execution time in the heuristic method compared to that of branch-and-bound method. It shows that the performance is really good for large sets (greater than 95% most of the time), while the execution time is just a few percent of the time required for optimal solution (less than 5%).

### Table 2: Performance comparison of branch-and-bound and a heuristic algorithm in terms of total achieved throughput and execution time.

| N   | Value % | Time % |
|-----|---------|--------|
| 40  | 92.5    | 15.3   |
| 70  | 95.6    | 4.2    |
| 100 | 97.3    | 3.9    |
| 130 | 98.1    | 2.7    |
| 160 | 97.7    | 2.7    |
| 190 | 98.1    | 2.9    |
| 220 | 98.5    | 3.1    |
| 250 | 98.7    | 3.1    |
| 280 | 97.5    | 3.9    |
| 310 | 97.4    | 3.0    |
| 340 | 98.3    | 2.4    |
| 370 | 99.3    | 1.9    |
| 400 | 99.2    | 2.6    |

### 7. CONCLUSION

In this paper, we propose a novel comprehensive scheme, which leads to utilizing the network resources more efficiently. To design such a scheme we take a multi time scale approach. Then in large time scales, we adaptively adjust base-station coverage area based on the corresponding traffic profile of the users in the coverage area. Then in medium time-scales we utilize a utility-based platform to formulate downlink resource allocation based on a novel defined utility function. This utility function quantifies the degree of utilization of network resources. Unlike previous works, we solve the problem with the general objective of maximizing...
the total network utility instead of achieved utility of each base station. We then map this problem to multidimensional multiple-choice knapsack Problems (MMKP). Since MMKP is NP-hard, a polynomial-time suboptimal algorithm was then modified to develop an efficient base-station assignment. Simulation results indicate significant performance improvement using the proposed scheme.

**APPENDIX**

*Proof of Proposition 1.* Assume \( X^* = \{x_{kj}^*\} \) is the output of the algorithm, and \( Y^* = \{y_{kj}\} \) is the result of the globally optimum solution. Let denote \( T_j^* = \sum_{j=1}^{N} \sum_{k=1}^{M} x_{kj}^* p_{ijk} \). Therefore, the total achieved throughput using the heuristic algorithm can be written as (A.1)-(A.2). For the optimal solution, \( Y^* \), we can rewrite the same expression as in (A.2) as

\[
\sum_{j=1}^{N} \sum_{k=1}^{M} x_{kj}^* u_{kj} = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{M} \lambda_i x_{kj}^* p_{ijk} + \sum_{j=1}^{N} \sum_{k=1}^{M} x_{kj}^* u_{kj} \tag{A.1}
\]

\[
- \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{M} \lambda_i x_{kj}^* p_{ijk} = \sum_{k=1}^{M} \lambda_i T_i^* + \sum_{j=1}^{N} \sum_{k=1}^{M} \left( u_{kj} - \sum_{i=1}^{N} \lambda_i p_{ijk} \right) x_{kj}^*, \tag{A.2}
\]

where \( T_j^* = \sum_{j=1}^{N} \sum_{k=1}^{M} y_{kj}^* p_{ijk} \). By definition, we know that all \( T_j^* \leq P_{Ri} \). Therefore, the upper limit for (27) can be written as

\[
\sum_{j=1}^{N} \sum_{k=1}^{M} y_{kj}^* u_{kj} \leq \sum_{k=1}^{M} \lambda_i P_{Ri} + \sum_{j=1}^{N} \sum_{k=1}^{M} \left( u_{kj} - \sum_{i=1}^{N} \lambda_i p_{ijk} \right) y_{kj}^*. \tag{A.3}
\]

Using (A.3) and (A.4), the difference between total achieved throughput using the sub-optimal algorithm and the global optimal solution is

\[
\sum_{j=1}^{N} \sum_{k=1}^{M} u_{kj} ( y_{kj}^* - x_{kj}^* ) \leq \sum_{k=1}^{M} \lambda_i (P_{Ri} - T_i^*) + \left\{ \sum_{j=1}^{N} \sum_{k=1}^{M} \left( u_{kj} - \sum_{i=1}^{N} \lambda_i p_{ijk} \right) y_{kj}^* - \sum_{j=1}^{N} \sum_{k=1}^{M} \left( u_{kj} - \sum_{i=1}^{N} \lambda_i p_{ijk} \right) x_{kj}^* \right\}. \tag{A.5}
\]

Let us denote the wireless communication link as \( \lambda_i \), and the wireless channel as \( \beta_{kj} \), where \( \beta_{kj} = (u_{kj} - \sum_{i=1}^{N} \lambda_i p_{ijk}) \). We define the following sets \( H_1 = (X^* \cup Y^*) - Y^* \), \( H_2 = (X^* \cup Y^*) - X^* \), and \( H_3 = (X^* \cap Y^*) \).

For the elements of \( H_3 \), it is clear that \( W \) is equal to zero. For the elements of \( H_1 \), \( \sum_{j=1}^{N} \sum_{k=1}^{M} \beta_{kj} y_{kj}^* = 0 \) and \( \sum_{j=1}^{N} \sum_{k=1}^{M} \beta_{kj} x_{kj}^* \geq 0 \), hence \( W \leq 0 \). For the elements of \( H_2 \), \( \sum_{j=1}^{N} \sum_{k=1}^{M} \beta_{kj} y_{kj}^* \leq 0 \) (since \( \beta_{kj} \leq 0 \)) and \( \sum_{j=1}^{N} \sum_{k=1}^{M} \beta_{kj} x_{kj}^* = 0 \), thus, again \( W \leq 0 \). Therefore, in all cases, we have \( W \leq 0 \), which in conjunction with (A.5) meaning that

\[
\sum_{j=1}^{N} \sum_{k=1}^{M} u_{kj} ( y_{kj}^* - x_{kj}^* ) \leq \sum_{k=1}^{M} \lambda_i (P_{Ri} - T_i^*) \]

\[
+ \sum_{k=1}^{M} \lambda_i \left( P_{Ri} - \sum_{j=1}^{N} \sum_{k=1}^{M} x_{kj}^* p_{ijk} \right), \tag{A.6}
\]

which completes the proof. \( \square \)

**ACKNOWLEDGMENTS**

This paper was partly presented at the IEEE/Sarnoff Symposium on Advances in Wired and Wireless Communication, April 2004, and the IEEE Ninth International Symposium on Computers and Communications (ISCC ’04), June 2004. This work was supported in part by Bell University Laboratory, University of Toronto.

**REFERENCES**

[1] M. L. Honig and J. B. Kim, “Resource allocation for packet data transmission in DS-CDMA,” in Proceedings of the 33rd Annual Allerton Conference on Communication, Control, and Computing, pp. 925–934, Monticello, Ill, USA, October 1995.

[2] M. L. Honig and J. B. Kim, “Allocation of DS-CDMA parameters to achieve multiple rates and qualities of service,” in Proceedings of IEEE Global Telecommunications Conference (GLOBECOM ’96), vol. 3, pp. 1974–1978, London, UK, November 1996.

[3] J. B. Kim and M. L. Honig, “Resource allocation for multiple classes of DS-CDMA traffic,” IEEE Transactions on Vehicular Technology, vol. 49, no. 2, pp. 506–519, 2000.

[4] J. B. Kim, M. L. Honig, and S. Jordan, “Dynamic resource allocation for integrated voice and data traffic in DS-CDMA,” in Proceedings of the 54th IEEE Vehicular Technology Conference (VTC ’01), vol. 1, pp. 42–46, Atlantic City, NJ, USA, October 2001.

[5] A. E. Gamal, J. Mammen, B. Prabhakar, and D. Shah, “Throughput-delay trade-off in energy constrained wireless networks,” in Proceedings of IEEE International Symposium on Information Theory (ISIT ’04), p. 439, Chicago, Ill, USA, June–July 2004.

[6] H. Holma and A. Toskala, WCDMA for UMTS: Radio Access for Third Generation Mobile Communications, John Wiley & Sons, New York, NY, USA, 2000.

[7] K. Navaie and H. Yanikomeroglu, “Optimal downlink resource allocation for non-realtime traffic in cellular CDMA/TDMA networks,” IEEE Communications Letters, vol. 10, no. 4, pp. 278–280, 2006.
[8] K. Navaie, D. Y. Montuno, H. Yanikomeroglu, and Y. Q. Zhao, “Optimal downlink resource allocation for cellular CDMA networks,” in Adaptation Techniques in Wireless Multimedia Networks, Y. Xiao and W. Li, Eds., Nova Science, New York, NY, USA, 2006.

[9] R. D. Yates and C.-Y. Huang, “Integrated power control and base station assignment,” IEEE Transactions on Vehicular Technology, vol. 44, no. 3, pp. 638–644, 1995.

[10] S. V. Hanly, “An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity,” IEEE Journal on Selected Areas in Communications, vol. 13, no. 7, pp. 1332–1340, 1995.

[11] F. P. Kelly, “Charging and rate control for elastic traffic,” European Transactions on Telecommunications, vol. 8, no. 1, pp. 33–37, 1997.

[12] S.-J. Oh and K. M. Wasserman, “Optimality of greedy power control and variable spreading gain in multi-class CDMA mobile networks,” in Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (Mobicom ’99), pp. 102–112, Seattle, Wash, USA, August 1999.

[13] C. U. Saraydar, N. B. Mandayam, and D. J. Goodman, “Pricing and power control in a multiscell wireless data network,” IEEE Journal on Selected Areas in Communications, vol. 19, no. 10, pp. 1883–1892, 2001.

[14] M. Xiao, N. B. Shroff, and E. K. P. Chong, “Utility-based power control in cellular wireless systems,” in Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM ’01), vol. 1, pp. 412–421, Anchorage, Alaska, USA, April 2001.

[15] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, “Joint resource allocation and base-station assignment for the downlink in CDMA networks,” IEEE/ACM Transactions on Networking, vol. 14, no. 1, pp. 1–14, 2006.

[16] X. Liu, E. K. P. Chong, and N. B. Shroff, “Opportunistic transmission scheduling with resource-sharing constraints in wireless networks,” IEEE Journal on Selected Areas in Communications, vol. 19, no. 10, pp. 2053–2064, 2001.

[17] M. Shabany, K. Navaie, and E. S. Sousa, “Downlink resource allocation for data traffic in heterogeneous cellular CDMA networks,” in Proceedings of the 9th International Symposium on Computers and Communications (ISCC ’04), vol. 1, pp. 436–441, Alexandria, Egypt, June-July 2004.

[18] M. Shabany and K. Navaie, “Joint pilot power adjustment and base station assignment for data traffic in cellular CDMA networks,” in Proceedings of IEEE/Sarnoff Symposium on Advances in Wired and Wireless Communication, pp. 179–183, Princeton, NJ, USA, April 2004.

[19] R. Vannithamby and E. S. Sousa, “An optimum rate/power allocation scheme for downlink in hybrid CDMA/TDMA cellular system,” in Proceedings of the 52th IEEE Vehicular Technology Conference (VTC ’00), vol. 4, pp. 1734–1738, Boston, Mass, USA, September 2000.

[20] T.-H. Lee, J.-C. Lin, and T. S. Yu, “Downlink power control algorithms for cellular radio systems,” IEEE Transactions on Vehicular Technology, vol. 44, no. 1, pp. 89–94, 1995.

[21] S. A. Grandhi and J. Zander, “Constrained power control in cellular radio systems,” in Proceedings of the 44th IEEE Vehicular Technology Conference (VTC ’94), vol. 2, pp. 824–828, Stockholm, Sweden, June 1994.

[22] W. Shih, “A branch and bound method for the multiconstraint zero-one knapsack problem,” The Journal of the Operational Research Society, vol. 30, no. 4, pp. 369–378, 1979.
Fault-Tolerant Target Localization in Sensor Networks

Min Ding, Fang Liu, Andrew Thaeler, Dechang Chen, and Xiuzhen Cheng

1 Computer Science Department, The George Washington University, Washington, DC 20052, USA
2 Division of Epidemiology and Biostatistics, Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, Bethesda, MD 20817, USA

Received 21 July 2006; Revised 8 November 2006; Accepted 13 November 2006

Recommended by Wei Li

Fault-tolerant target detection and localization is a challenging task in collaborative sensor networks. This paper introduces our exploratory work toward identifying the targets in sensor networks with faulty sensors. We explore both spatial and temporal dimensions for data aggregation to decrease the false alarm rate and improve the target position accuracy. To filter out extreme measurements, the median of all readings in a close neighborhood of a sensor is used to approximate its local observation to the targets. The sensor whose observation is a local maxima computes a position estimate at each epoch. Results from multiple epoches are combined together to further decrease the false alarm rate and improve the target localization accuracy. Our algorithms have low computation and communication overheads. Simulation study demonstrates the validity and efficiency of our design.

Copyright © 2007 Min Ding et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

The development of wireless sensor networks provides many exciting applications, including roadway safety warning [1], habitat monitoring [2], smart classroom [3], and so forth. Such networks rely on the collaboration of thousands of resource-constrained error-prone sensors for monitoring and control. One important task of a typical sensor network is to detect and report the locations of targets (e.g., tanks, land mines, etc.) with the presence of faulty sensor measurements. In our study, we seek fault-tolerant algorithms to identify the region containing targets and the position of each target.

Filtering faulty sensor measurements and locating targets are not trivial. Due to the stingy energy budget within each sensor, we have to seek localized and computationally efficient algorithms such that a sensor can determine whether a target presents and whether it needs to report the target information to the base station (to determine whether and where a target presents). The existence of faulty sensors exacerbates the “hardness” of the problem. False alarms waste network resource. They may mislead users to make wrong decisions. Therefore, target identification and localization algorithms must be fault-tolerant, must have a low false alarm rate, and must be robust.

In this paper, we propose fault-tolerant algorithms to detect the region containing targets and to identify possible targets within the target region. Here only the same kind of targets are considered. To avoid the disturbance of extreme measurements at faulty sensors, each sensor collects neighboring readings and computes the median, representing its local observation on the targets. Median is proved to be an effective robust nonparametric operator that requires no strong mathematical assumptions [4]. A median exceeding some threshold indicates the occurrence of a possible target. Whether a real target exists or not must be jointly determined by neighboring sensors at the same time. To localize a target within the target region, a sensor whose observation is a local maxima computes the geometric center of neighboring sensors with similar observations. We also explore time dimension to reduce the false alarm rate. Results from multiple epoches are combined to refine the target position estimates. Our algorithms have low computation and communication overheads because only simple numerical operations (maximum, median, and mean) are involved at each sensor. The protocol has a low communication overhead too, since only sensors in charge of the location estimation report to the base station. Simulation study indicates that in most cases our algorithms can identify all the targets and only one report for one target is sent to the base station per epoch when up to 20% of the sensors are faulty, and when the network is moderately dense.

This paper is organized as follows. Related work and network model are sketched in Sections 2 and 3, respectively. Fault-tolerant target identification and localization
algorithms are proposed in Section 4. Simulation results are reported in Section 5. We conclude our paper in Section 6.

2. RELATED WORK

Target detection and localization [5–8], target classification [9–11] and target tracking [12–15] have attracted many research activities in sensor networks. In this section, we focus on the related works in target localization and target identification.

Clouqueur et al. [5] seek algorithms to collaboratively detect a target region. Each sensor obtains the target energy (or local decision) from other sensors, drops extreme values if faulty sensors exist, computes the average, and then compares it with a predetermined threshold for final decisions. For these algorithms, the challenge is the determination of the number of extreme values. This is unavoidable when using “mean” for data aggregation. As a comparison, we explore the utilization of “median” to effectively filter out extreme values for target region detection.

Zou and Chakrabarty [6–8] propose an energy-aware target detection and localization strategy for cluster-based wireless networks. The cluster head collects event notification from sensors within the cluster and then executes a probabilistic localization algorithm to determine candidate nodes to be queried for target information. This algorithm is designed only for cluster-based sensor networks. The cluster head must keep a pregenerated detection probability table constructed from sensor locations. Each sensor reports the detection of an object to the cluster head based on its own measurements. This work does not consider fault-tolerance at all, thus the decision by cluster head may be based on incorrect information.

Fang et al. [9, 10] provide the algorithms for target counting and enumeration in sensor networks. A spanning tree is constructed to locate a possible target. The root of each tree has the maximal sensed signal power among all the nodes in the tree cluster. The tree structures which define the target region are formed step by step. Each node in the tree must relay its root information. Fault-tolerance is not considered in their protocols, therefore a faulty sensor may be elected as a leader and reports wrong target information.

Li et al. [16] estimate target position by solving a nonlinear least squares problem. Target localization based on the time-of-arrival (TOA) [17] or the direction-of-arrival (DOA) [18] of acoustical/seismic signals has also been explored. Locating victims through emergency sensor networks in a centralized fashion has been studied in [19]. In [14, 15], a spanning tree rooted at the sensor node close to a target is used for target tracking, with target position estimated by the location of the root sensor. We propose much simpler algorithms for target identification and localization in this paper.

3. NETWORK MODEL

In this paper, we assume that $N$ sensors are deployed uniformly in a $b \times b$ square field located in the two dimensional Euclidean plane $\mathcal{R}^2$, with a base station residing in the boundary. Sensors are powered by batteries and have a fixed radio range. The base station has a strong computational capability with an unlimited power supply. Power conservation and fault-tolerance are the major goals when designing algorithms for target localization.

Let $R(s_i)$ or $R_i$ denote the reading of sensor $s_i$. Instead of a 0–1 binary variable, $R(s_i)$ is assumed to represent signal strength measurements on factors such as vibration, light, sound, and so on. A target region, denoted by $T \mathcal{R}$, is a subset of $\mathcal{R}^2$ such that it contains all the sensors that can detect the presence of the target. A sensor’s reading is faulty if it reports inconsistent and arbitrary values to the neighboring sensors [5]. Sensors with faulty readings are called faulty sensors. In this paper, we will use $s_i$ to refer to either the $i$th sensor or the location of the $i$th sensor.

We assume that each sensor can compute its physical position through either GPS or some GPS-less techniques [20–22]. In this paper, we focus on the fault-tolerant target identification and localization, and thus the delivery of the target location will not be considered. We assume there exists a robust routing protocol in charge of the transmission of the target information to the base station.

All targets emit some kinds of signals (vibration, acoustical, light, etc.) when present. These signals will be propagated to the surrounding area with a decayed intensity. The following model is used to quantify the signal strength at location $s_i$ for a target at location $L$ [5]:

$$
S_s(s_i) = \begin{cases} 
P_0, & \text{if } d < d_0, \\
\frac{P_0}{(d/d_0)^k}, & \text{otherwise}, 
\end{cases}
$$

(1)

where $P_0$ is the signal intensity at $L$, $d = \|L - s_i\|$ is the Euclidean distance between the target and the sensor at $s_i$, $d_0$ is a constant that accounts for the physical size of the target, and $k \in [2.0, 5.0]$ [23] is a decay factor determined by the environment. The signal strength measured by a sensor at $s_i$ is then

$$
R(s_i) = S_s(s_i) + N(s_i),
$$

(2)

where $N(s_i)$ represents the noise level at $s_i$. We assume $N(s_i)$ follows $N(\mu, \sigma^2)$, a Gaussian distribution with mean $\mu$ and variance $\sigma^2$. For Gaussian white noise, $\mu = 0$. When more than one target present in the network, signals of multiple targets are summed at each sensor.

In this paper, we assume sensors can properly execute our algorithms even though their readings are faulty. In other words, we assume there is no fault in processing and transmitting/receiving neighboring measurements.

4. FAULT-TOLERANT TARGET DETECTION AND LOCALIZATION

In this section, we first describe an algorithm for target region detection. Then we present a procedure to estimate the locations of the targets from the sensors within the target region. We also propose an algorithm for data aggregation along temporal dimension to decrease the false alarm rate and improve the target position accuracy.
For any given sensor $s_i$,
(1) Obtain signal measurements $R^{(i)}_1, R^{(i)}_2, \ldots, R^{(i)}_n$ from all sensors in $\mathcal{N}(s_i)$.
(2) Compute med. of the set $\{R^{(i)}_1, R^{(i)}_2, \ldots, R^{(i)}_n\}$ as the estimated reading $\tilde{R}_i$ at location $s_i$.
(3) Determine event sensors. A sensor $s_i$ is an event sensor if the estimated value $\tilde{R}_i$ is larger than a predefined threshold $\theta_1$.

**Algorithm 1:** For target region detection.

### 4.1. Target region detection

Our target region detection algorithm aims at finding all sensors that can detect the presence of the targets. Nodes closer to the targets usually have higher measurements. Faulty sensors may report arbitrary values.

Let $\mathcal{N}(s_i)$ denote a bounded closed set of $\mathbb{R}^2$ that contains a sensor $s_i$ and additional $n - 1$ sensors. The set $\mathcal{N}(s_i)$ represents a closed neighborhood of the sensor $s_i$. An example of $\mathcal{N}(s_i)$ is the closed disk centered at $s_i$ with its radius equal to the radio range. Let $R^{(i)}_1, R^{(i)}_2, \ldots, R^{(i)}_n$ denote the signal strength measured by the nodes in $\mathcal{N}(s_i)$. A possible estimate of signal strength at location $s_i$ is

$$\tilde{R}_i = \text{med}_i,$$  \hspace{1cm} (3)

where $\text{med}_i$ denotes the median of the set $\{R^{(i)}_1, R^{(i)}_2, \ldots, R^{(i)}_n\}$. In other words, one could estimate $R_i$ by the “center” of $\{R^{(i)}_1, R^{(i)}_2, \ldots, R^{(i)}_n\}$.

Note that $\text{med}_i$, in (3) should not be replaced by the mean $(R^{(i)}_1 + R^{(i)}_2 + \cdots + R^{(i)}_n)/n$ of the set $\{R^{(i)}_1, R^{(i)}_2, \ldots, R^{(i)}_n\}$. This is because the sample mean cannot represent well the “center” of a sample when some values of the sample are extreme. Nevertheless, median is widely used to estimate the “center” of samples with outliers. Its conditional correctness is proved in [4]. Faulty sensors may have extreme values, representing outliers in the sample set. Faulty readings have little influence on med, as long as most sensors behave properly.

The procedure of target region detection is described as follows.

Intuitively, an event sensor is a sensor that can detect the presence of the targets. Compared to the value fusion method for target region detection in [5], which computes the mean after dropping $\eta$ highest and $\eta$ lowest values, Algorithm 1 employs the robust operator median so that it effectively eliminates the effects of faulty sensors without exploiting any complicated algorithm for the estimation of $\eta$.

### 4.2. Target Localization

Algorithm 1 is used to detect the presence of targets. It does not tell how many targets exist and where they are. Shifting the task of target localization to the base station by sending the measurements of all sensors in the target region is too expensive in terms of energy consumption. Therefore, we consider to delegate one sensor to communicate with the base station for each target and compute the position of the target locally. The following algorithm is employed to locate the targets in the target region.

Note that in step (1) of Algorithm 2, $m$ can be smaller than $n$. A sensor is selected as a root sensor if its estimated signal strength is a maxima among event sensors in $\mathcal{N}(s_i)$. Nodes closer to the targets usually have larger measurements and thus have a higher probability to become root sensors. Furthermore, the number of root sensors is constrained by (4). A root sensor uses (5) to compute the location of a target based on the locations of some neighboring nodes. As a comparison, most related works in literature [9, 10, 14, 15] utilize the position of the root sensor as an approximation of the target position.

**Algorithm 2:** For target localization.

We observe that the two algorithms proposed in Sections 4.1 and 4.2 explore only spatial information for data aggregation. In reality, sensors sample their observations periodically. By investigating along the temporal dimension, performance for target detection and localization can be improved, as verified by simulation study in Section 5. In this section, we discuss how the base station can identify false alarms and improve the target position accuracy by using location estimates obtained at $T$ epochs from root sensors. For better elaboration, we call the location estimates by root sensors the raw data.

Assume both Algorithms 1 and 2 are executed once per epoch. The base station receives a sequence of raw data, denoted by $\{\tilde{L}^{(1)}, \tilde{L}^{(2)}, \ldots, \tilde{L}^{(t)}\}$, from root sensors, where
gets. The degree of fault-tolerance has been considered in our
includes two tasks: evaluating the degree of fault-tolerance
Evaluation of the target detection and localization algorithms

\section{Simulation}
Section 5, in most cases only one message per target will be
sent to the base station per epoch in moderately dense sensor
networks. As indicated by the simulation study in
we have studied the accuracy for target localization when only
one target presents in the network. We note that for a low
network density and a high sensor fault probability, the base

5. Performance metrics

Evaluation of the target detection and localization algorithms
includes two tasks: evaluating the degree of fault-tolerance and
evaluating the accuracy of the estimated positions of targets. The degree of fault-tolerance has been considered in our
prior work [24].

To evaluate the accuracy of the estimated positions of the
targets, we first define position error \( e(\hat{L}_w) \) for the target at
location \( L_w \) to be the Euclidean distance between \( \hat{L}_w \) and the
real target location \( L_w \), that is,

\[
e(\hat{L}_w) = ||\hat{L}_w - L_w||.
\]

We use the average of the position errors for all targets to
evaluate the accuracy of our algorithms,

\[
e(\mathcal{L}) = \frac{e(\hat{L}_1) + e(\hat{L}_2) + \cdots + e(\hat{L}_p)}{p},
\]

where \( p \) is the total number of the targets in the network. Obviously, smaller \( e(\mathcal{L}) \) indicates higher position accuracy.

\subsection{Simulation setup}
MATLAB is used to perform all simulations. The sensor
nodes are deployed in a 32 \( \times \) 32 square region, which resides
in the first quadrant such that the lower-left corner and the
origin are colocated. Sensor coordinates are defined accord-
ingly. We fix the transmission range of each sensor to be 3, and vary the number of sensor nodes to get different network
densities. Network density is defined as the average number
of one-hop neighbors for each sensor. Sensors are randomly
deployed according to the uniform distribution. We choose
\( \mathcal{N}(s_i) \) to be the set containing all one-hop neighbors of \( s_i \).

In the simulation for multiple targets detection and local-
ization, three targets are located in the network region, where
the coordinates of each target position are randomly sampled
from \([8, 10]\). The distance between each target pair is not less
than \( 4d_0 = 8 \). We also evaluate the performance of the al-
gorithms when one target is deployed. In this scenario, the
target coordinates are chosen in the similar way.

In this paper, we consider identification and localization
problem for targets of the same kind, thus we assume all the
targets to have the same signal intensity. The signal intensity
\( P_0 \) from each target is set to 30. Signal model follows (1) with
\( d_0 = 2 \) and \( k = 2 \). (We have simulated the cases of \( k = 3, 4, 5 \),
and obtained similar results. We only report the result for
\( k = 2 \) in this paper.) For sensor \( s_i \), its noise level \( N(s_i) \) is
drawn from \( N(\mu, \sigma^2) \) with \( \mu = 0 \) and \( \sigma = 1 \), characteriz-
ning both environment disturbance and sensor measurement
error. The readings of a faulty sensor are randomly chosen from
\([0, 60]\).

The base station classifies the position estimates from different
epochs into different groups based on the distances of
pairwise estimates and \( d_0 \). A group indicates the existence of a
target only if its cardinality is not less than half of the num-
er of epochs under consideration.

Note that two thresholds (\( \theta_1 \) in Algorithm 1 and \( \theta_2 \) in
Algorithm 2) are needed to make decisions. Throughout the
simulation, we choose \( \theta_1 = 3\sigma = 3 \), showing that a normal
sensor has a low probability \((1 - 99.7\%)\) to report a noise value that is larger than \( 3\sigma \). To estimate the locations of the
detected targets, we set \( \theta_2 = 4 \). This means that sensors in
close proximity of a root sensor will contribute to the target
position estimation if the deviation of their (estimated) sig-
nal strengths from that of the root sensor is at most 4.

5.3. Simulation results

In this section, we report our simulation results, with
each representing an averaged summary over 100 runs. In our prior work [24], we have evaluated the degree of fault-tolerance of our algorithms through two parameters:
the correction accuracy and the false correction rate. We also have
studied the accuracy for target localization when only one target
presents in the network. We note that for a low
network density and a high sensor fault probability, the base

Algorithm 3: For target identification.
station fails to locate the real target with a reasonably high false alarm rate. Furthermore, the simulation results in [24] also indicate that it is sufficient to overcome the disturbance of the Byzantine behavior of faulty sensors using the readings from 9 epoches. Thus in this paper we choose to use $p \leq 0.35$ and $T = 1$ or $9$ for multiple target detection and localization, where $p$ is the sensor fault probability.

We first study the number of targets detected when three targets present in the network. The targets are apart enough so that different targets can be identified. Figures 1 and 2 illustrate the number of targets detected by the base station when position estimates from 1 epoch and from 9 epoches are exploited, respectively.

First, we observe that in moderately and high dense networks, the probability of reporting the existence of three targets is high. The false alarm rate is less than $0.1$ for $p \leq 0.20$ and density $= 30, 50$ when aggregating over 9 epoches, as shown in Figure 2. By comparing Figure 1 with Figure 2, we observe that the number of reported targets contributing to the false alarm rate can be reduced by increasing $T$. We also notice that the average numbers of position estimates sent to the base station at each epoch are $3.18$ and $3.05$ for $p = 0.20$ and density $= 30, 50$, respectively (as shown in Figures 1(b) and 1(c)). This indicates that in many cases, only three root sensors need to send their target location estimation to the base station at each epoch. Therefore, the communication overhead of our algorithms is low. In Figure 1(a), we observe that false alarm exists under density $= 10$ and $T = 1$ when faulty sensors do not exist. It is possible for some sensors to be a local maxima due to the accumulation of the signal strength from all targets. Therefore, median is not robust enough under low density.
Figure 3: The number of targets detected when one target is deployed. Here, $T = 9$ and density = 10, 30, 50, respectively.

Figure 4: Position error versus $p$ with different network densities when three targets are deployed. In this scenario, $T = 1$.

Figure 5: Position error versus $p$ with different network densities when three targets are deployed. In this scenario, $T = 9$.

For comparison, we study the performance in the scenario when only one target exists. Similarly, the number of reported target leading to the false alarm rate can be reduced by temporal aggregation. Here, we only report the number of targets detected by the base station for $T = 9$. As shown in Figures 3(b) and 3(c), the false alarm rate equals to 0 for $p \leq 0.20$ and density = 30, 50 by aggregating over 9 epoches. Our algorithms have better performance for one target identification since there is no interference of signal strengths from multiple targets.

Figures 4 and 5 illustrate the position error in units versus $p$ for multiple target localization under different network densities. Both figures demonstrate that our algorithms obtain a high accuracy for target localization. As shown in Figure 5, position errors are less than 0.5 unit when density $\geq 30$ and $p \leq 0.25$. By comparing Figures 4 and 5, we observe that position errors are decreased when position estimates from multiple epoches are exploited. Note that position errors generally increase with higher $p$ when the network density is fixed. We also note that a higher density could decrease position errors. This is reasonable since in higher density networks, more sensors are involved in the computation, which brings in more information, and thus results in more accurate results. For the case when only one target presents, the position errors show the similar trends when position estimates from 1 epoch and from 9 epoches.
are exploited. The results are not shown here for space constraint.

5.4. Discussion

The simulation results reported in the previous section reveal the high performance of our algorithms for target detection and localization in moderate and high density networks when \( p \leq 0.20 \). The false alarm rate is decreased and the target position accuracy is increased by exploring both temporal and spatial aggregation.

We notice that two targets may be identified as a single one when their locations are very close. It is necessary to study the sensitivity of our algorithms for targets that are close to each other. Thus, we evaluate our algorithms for the scenarios when two targets are deployed at different positions.

Figures 6 and 7 illustrate the frequency of one target being detected and two targets being detected, versus the variable distance between the two targets under different sensor fault probabilities for density = 10, 30, 50, respectively. We observe that the frequency of one target detected normally decreases and the frequency of two targets detected increases when their distance gets larger. Two targets are distinguishable when their distance is equal or larger than 8. Note that in our simulation we consider targets with size \( d_0 = 2 \) and the decay factor \( k = 2 \). These two parameters are key factors to the sensitivity of our algorithms. Under these settings, for moderately or highly dense networks, the probability that two targets are ambiguous is high when their distance is less than 6. It is also interesting to notice that the two targets are more easily to be distinguished with higher \( p \) when density = 10, due to the relatively high disturbance of fault sensors.

Our algorithms may fail when the locations of the two targets are very close. One and only one local maxima may be formed at a sensor that has roughly the same distance to both targets, due to the accumulation of the target signal strength.

Figure 6: Frequency of one target detected versus target distance when two targets are deployed, \( T = 9 \), and density = 10, 30, 50, respectively.

Figure 7: Frequency of two targets detected versus target distance when two targets are deployed, \( T = 9 \), and density = 10, 30, 50, respectively.
In this case, the energy level at the root sensor may be explored. We target this as our future research.

6. CONCLUSION

In this paper, we present fault-tolerant algorithms for target identification and localization in sensor networks. In this study, data aggregation is conducted along both temporal and spatial dimensions for decreasing the false alarm rate and increasing the target position accuracy. Simulation results verify the efficiency and effectiveness of our design.

This paper is exploratory in that we use “median” instead of “mean” to locally aggregate neighboring readings to filter out faulty measurements. We report the simulation results when the target region contains multiple targets. We believe that this idea can be extended to target classification and target tracking, and decide to explore along this direction in the future.

ACKNOWLEDGMENTS

The research of the fourth author is supported by the NSF CAREER Award no. CNS-0347674.

REFERENCES

[1] K. Xing, M. Ding, X. Cheng, and S. Rotenstreich, “Safety warning based on roadway sensor networks,” in IEEE Wireless Communications and Networking Conference (WCNC ’05), vol. 4, pp. 2355–2361, New Orleans, LA, USA, March 2005.

[2] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, “Wireless sensor networks for habitat monitoring,” in Proceedings of the ACM International Workshop on Wireless Sensor Networks and Applications (WSNA ’02), pp. 88–97, Atlanta, GA, USA, September 2002.

[3] S. S. Yau, S. K. S. Gupta, F. Karim, S. I. Ahamed, Y. Wang, and B. Wang, “Smart classroom: enhancing collaborative learning using pervasive computing technology,” in Proceedings of the 6th WFEO World Congress on Engineering Education and the 2nd ASEE International Colloquium on Engineering Education (ASEE ’03), pp. 13633–13642, Nashville, Tenn, USA, June 2003.

[4] E. L. Lehmann, Nonparametrics: Statistical Methods Based on Ranks, Prentice-Hall, Englewood Cliffs, NJ, USA, 1998.

[5] T. Clouqueur, K. K. Saluja, and P. Ramanathan, “Fault tolerance in collaborative sensor networks for target detection,” IEEE Transactions on Computers, vol. 53, no. 3, pp. 320–333, 2004.

[6] Y. Zou and K. Chakrabarty, “Energy-aware target localization in wireless sensor networks,” in Proceedings of the 1st IEEE International Conference on Pervasive Computing and Communications (PerCom ’03), pp. 60–67, Dallas–Fort Worth, TX, USA, March 2003.

[7] Y. Zou and K. Chakrabarty, “Target localization based on energy considerations in distributed sensor networks,” in Proceedings of the 1st IEEE International Workshop on Sensor Network Protocols and Applications (SNPA ’03), pp. 51–58, Anchorage, Alaska, USA, May 2003.

[8] Y. Zou and K. Chakrabarty, “Sensor deployment and target localization in distributed sensor networks,” ACM Transactions on Embedded Computing Systems, vol. 3, no. 1, pp. 61–91, 2004.

[9] Q. Fang, F. Zhao, and L. Guibas, “Counting targets: building and managing aggregates in wireless sensor networks,” Tech. Rep. P2002-10298, Palo Alto Research Center (PARC), Palo Alto, Calif, USA, 2002.

[10] Q. Fang, F. Zhao, and L. Guibas, “Lightweight sensing and communication protocols for target enumeration and aggregation,” in Proceedings of the International Symposium on Mobile Ad hoc Networking and Computing (MobiHoc ’03), pp. 165–176, Annapolis, MD, USA, June 2003.

[11] J. Shin, L. Guibas, and F. Zhao, “A distributed algorithm for managing multi-target identities in wireless ad-hoc sensor networks,” in Proceedings of the 2nd International Workshop on Information Processing in Sensor Networks (IPSN ’03), pp. 223–238, Palo Alto, Calif, USA, April 2003.

[12] J. Aslam, Z. Butler, F. Constantin, V. Crespi, G. Cybenko, and D. Rus, “Tracking a moving object with a binary sensor network,” in Proceedings of the 1st International Conference on Embedded Networked Sensor Systems (SenSys ’03), pp. 150–161, Los Angeles, Calif, USA, November 2003.

[13] W.-P. Chen, J. C. Hou, and L. Sha, “Dynamic clustering for acoustic target tracking in wireless sensor networks,” in Proceedings of the 11th IEEE International Conference on Network Protocols (ICNP ’03), pp. 284–294, Atlanta, Georgia, USA, November 2003.

[14] W. Zhang and G. Cao, “Optimizing tree reconfiguration for mobile target tracking in sensor networks,” in Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM’04), vol. 4, pp. 2434–2445, Hong Kong, March 2004.

[15] W. Zhang and G. Cao, “DCTC: dynamic convoy tree-based collaboration for target tracking in sensor networks,” IEEE Transactions on Wireless Communications, vol. 3, no. 5, pp. 1689–1701, 2004.

[16] D. Li, K. D. Wong, Y. H. Hu, and A. M. Sayeed, “Detection, classification, and tracking of targets,” IEEE Signal Processing Magazine, vol. 19, no. 2, pp. 17–29, 2002.

[17] D. Friedlander, C. Griffin, N. Jacobson, S. Phoha, and R. R. Brooks, “Dynamic agent classification and tracking using an ad hoc mobile acoustic sensor network,” EURASIP Journal on Applied Signal Processing, vol. 2003, no. 4, pp. 371–377, 2003.

[18] L. Yip, K. Comanor, J. C. Chen, R. E. Hudson, K. Yao, and L. Vandenberghe, “Array processing for target DOA, localization, and classification based on AML and SVM algorithms in sensor networks,” in Proceedings of the 2nd International Workshop on Information Processing in Sensor Networks (IPSN ’03), pp. 269–284, Palo Alto, Calif, USA, April 2003.

[19] S. Ray, R. Ungrangsi, F. De Pellegrini, A. Trachtenberg, and D. Starobinski, “Robust location detection in emergency sensor networks,” in Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM ’03), vol. 2, pp. 1044–1053, San Francisco, Calif, USA, March-April 2003.

[20] M. Y. Al-Laho, M. Song, and J. Wang, “Mobility-pattern based localization update algorithms for mobile wireless sensor networks,” in Proceedings of the 1st International Conference on Mobile Ad-hoc and Sensor Networks (MSN ’05), vol. 3794 of Lecture Notes in Computer Science, pp. 143–152, Wuhan, China, December 2005.
[21] X. Cheng, A. Thaeler, G. Xue, and D. Chen, “TPS: a time-based positioning scheme for outdoor wireless sensor networks,” in Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM ’04), vol. 4, pp. 2685–2696, Hong kong, March 2004.

[22] A. Thaeler, M. Ding, and X. Cheng, “iTPS: an improved location discovery scheme for sensor networks with long-range beacons,” Journal of Parallel and Distributed Computing, vol. 65, no. 2, pp. 98–106, 2005.

[23] M. Hata, “Empirical formula for propagation loss in land mobile radio services,” IEEE Transactions on Vehicular Technology, vol. 29, no. 3, pp. 317–325, 1980.

[24] M. Ding, D. Chen, A. Thaeler, and X. Cheng, “Fault-tolerant target detection in sensor networks,” in Proceedings of IEEE Wireless Communications and Networking Conference (WCNC ’05), vol. 4, pp. 2362–2368, New Orleans, La, USA, March 2005.
Historically developed for secure communication and military use, CDMA has been identified as a major modulation and multiple-access technique for 3G systems and beyond. In addition to the wide bandwidth and low power-spectrum density which make CDMA signals robust to narrowband jamming and easy to be concealed within the noise floor, the physical layer built-in information privacy of CDMA system is provided by pseudorandom scrambling. In this paper, first, security weakness of the operational and proposed CDMA airlink interfaces is analyzed. Second, based on the advanced encryption standard (AES), we propose to enhance the physical layer built-in security of CDMA systems through secure scrambling. Performance analysis demonstrates that while providing significantly improved information privacy, CDMA systems with secure scrambling have comparable computational complexity and overall system performance with that of conventionally scrambled systems. Moreover, it is shown that by scrambling the training sequence and the message sequence separately with two independent scrambling sequences, both information privacy and system performance can be further improved. The proposed scheme can readily be applied to 3G systems and beyond.

Copyright © 2007 Tongtong Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

As people are relying more and more on wireless communication networks for critical information transmission, security has become an urgent issue and a bottleneck for new wireless communication services such as wireless mobile Internet and e-commerce [1]. Due to user mobility and the fact that there is no physical boundary in wireless environment, wireless communication networks are facing much more significant challenges compared to their data network counterparts.

Direct sequence spread-spectrum system, also known as code-division multiple access (CDMA), was historically developed for secure communication and military use. In CDMA systems, each user is assigned a specific spreading sequence to modulate its message signal. The spreading process increases the bandwidth of the message signal by a factor \( N \), known as spreading factor or the processing gain, and meanwhile reduces the power-spectrum density of the signal also by a factor \( N \). With large bandwidth and low power spectrum density, CDMA signals are resistant to malicious narrowband jamming and can easily be concealed within the noise floor, preventing from being detected by an unauthorized person. Moreover, the message signal cannot be recovered unless the spreading sequence is known, this makes it difficult for unauthorized person to intercept the signal.

Due to high spectrum efficiency and simplicity in system planning, CDMA is now finding widespread civilian and commercial applications such as cellular phones, personal communications, and position location. As it is well known, CDMA is used in the US digital cellular standard IS-95 and has been identified as the major modulation technique for third generation (3G) wireless communications and beyond.

Relying on the long pseudorandom spreading sequence generator, the operational CDMA system (IS-95) and the proposed 3G UMTS system can provide a near-satisfactory physical layer built-in security solution to voice centric wireless communications, which generally last only a very short period of time. However, the security features provided by these systems are far from being adequate and being acceptable when used for data communications. In literature, wireless security is generally considered from MAC layer and network layer, see [2], for example, and few thoughts have been given to the physical layer security enhancement. In this paper, we show that by combining cryptographic techniques and modulation techniques in the transmitter and receiver.
Since the channelization codes are chosen to be Walsh codes, which are easy to generate, the physical layer built-in security of CDMA systems mainly relies on the long pseudorandom scrambling sequence, also known as long code. In the following, we will analyze the maximum complexity to recover the long code of the IS-95 system and the 3GPP UMTS system.

### 2.1. Scrambling code recovery of the IS-95 system

In IS-95, the long-code generator consists of a 42-bit number called *long-code mask* and a 42-bit linear feedback shift register (LFSR) specified by the following characteristic polynomial:

\[
\begin{align*}
    x^{42} + x^{35} + x^{33} + x^{31} + x^{27} + x^{26} + x^{25} \\
    + x^{22} + x^{21} + x^{19} + x^{18} + x^{17} + x^{16} \\
    + x^{10} + x^{7} + x^{6} + x^{5} + x^{2} + x + 1,
\end{align*}
\]

where the 42-bit long code mask is shared between the mobile and the base station.

As shown in Figure 2, each chip of the long code is generated by the modulo-2 inner product of a 42-bit mask and the 42-bit state vector of the LFSR.

Let \( M = \{m_1, m_2, \ldots, m_{42}\} \) denote the 42-bit mask and \( S(t) = \{s_1(t), s_2(t), \ldots, s_{42}(t)\} \) denote the state of the LFSR at time instance \( t \). The long-code sequence \( c(t) \) at time \( t \) can thus be represented as

\[
c(t) = m_1s_1(t) + m_2s_2(t) + \cdots + m_{42}s_{42}(t),
\]

where the additions are modulo-2 additions.

As it is well known, for a sequence generated from an \( n \)-stage linear feedback shift register, if an eavesdropper can intercept a \( 2n \)-bit sequence segment, then the characteristic polynomial and the entire sequence can be reconstructed according to the Berlekamp-Massey algorithm [3]. This leaves an impression that the maximum complexity to recover the long-code sequence \( c(t) \) is \( O(2^{42}) \). However, for IS-95, since the characteristic polynomial is known to the public, an eavesdropper only needs to obtain 42 bits of the long-code sequence to determine the entire sequence. That is, the maximum complexity to recover the long-code sequence \( c(t) \) is only \( O(2^{42}) \).

In fact, since \( s_1(t), s_2(t), \ldots, s_{42}(t) \) are the outputs of the same LFSR, they should all be the same except for a phase difference, that is,

\[
s_{42}(t) = s_{41}(t - 1) = \cdots = s_1(t - 41).
\]

Let \( a = [a_1, a_2, \ldots, a_{42}] \) denote the coefficient vector of the characteristic polynomial in (5), then it follows from (7) that

\[
s_i(t) = a_1s_{i-1}(t) + a_2s_{i-2}(t) + \cdots + a_{42}s_{i-42}(t) \\
    = a_is_i(t - 1) + a_is_i(t - 2) + \cdots + a_{42}s_i(t - 42).
\]
as the maximum complexity to recover it is only \(O(2^{42})\).

Let \(c(t)\) be the continuous long-code sequence bits, then the whole long-code sequence can be regenerated. Therefore, the long code sequence of IS-95 is vulnerable under ciphertext-only attacks as the maximum complexity to recover it is only \(O(2^{42})\).

Substituting (8) into (6), we have

\[
c(t) = \sum_{i=1}^{42} m_i s_i(t) = \sum_{i=1}^{42} m_i \left( \sum_{j=1}^{42} a_j s_i(t-j) \right) = \sum_{j=1}^{42} a_j s_i(t-j)
\]

(9)

Define

\[
A = \begin{bmatrix}
a_1 & 0 & \cdots & 0 \\
a_2 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
a_{41} & 0 & 0 & \cdots & 1 \\
a_{42} & 0 & 0 & \cdots & 0
\end{bmatrix},
\]

(10)

then it follows that

\[
[c(t), c(t-1), \ldots, c(t-41)] = [c(t-1), c(t-2), \ldots, c(t-42)] \ast A.
\]

(11)

Let \(C(t) = [c(t), c(t-1), \ldots, c(t-41)]\), then for any \(n \geq t\), from (11) we have

\[
C(n) = C(t) \ast A^{n-t}.
\]

(12)

Therefore, as long as \(C(t)\) for a time instance \(t\) is known, then the entire sequence can be recovered. In other words, as long as an eavesdropper can intercept/recover up to 42 continuous long-code sequence bits, then the whole long-code sequence can be regenerated. Therefore, the long code sequence of IS-95 is vulnerable under ciphertext-only attacks as the maximum complexity to recover it is only \(O(2^{42})\).

2.2. Scrambling code recovery of the 3GPP UMTS system

In the 3GPP UMTS standard, Gold codes generated from two generator polynomials of degree 18 are used as scrambling code, as shown in Figure 3.

Denote the states for the two LFSRs at time instance \(t\) as \(r(t) = [r_{17}(t), r_{16}(t), \ldots, r_1(t), r_0(t)]\) and \(s(t) = [s_{17}(t), s_{16}(t), \ldots, s_1(t), s_0(t)]\), where

\[
\begin{align*}
    r_{17}(t) &= r_7(t-1) + r_0(t-1), \\
    s_{17}(t) &= s_{10}(t-1) + s_7(t-1) + s_5(t-1) + s_0(t-1).
\end{align*}
\]

(13)

Then at time instance \(t\), sequence \(I\) can be written as

\[
I(t) = r_0(t-1) + s_0(t-1),
\]

(14)

while sequence \(Q\) can be expressed as

\[
Q(t) = \sum_{i=0}^{17} a_i r_i(t-1) + \sum_{i=0}^{17} b_i s_i(t-1),
\]

(15)

where \(a_i\) and \(b_i\) are either 0 or 1 as shown in Figure 3.

Note that \(r_0(t) = r_1(t-1) = \cdots = r_{17}(t-17)\) and \(s_0(t) = s_1(t-1) = \cdots = s_{17}(t-17)\), we have

\[
Q(t) = \sum_{i=0}^{17} a_i r_0(t+i-1) + \sum_{i=0}^{17} b_i s_0(t+i-1).
\]

(16)

From (14) and (16), it follows that the maximum complexity to recover the scrambling code of the 3GPP UMTS system based on ciphertext-only attack is \(O(2^{256})\).

This implies that the physical layer built-in security of the 3GPP UMTS is actually weaker than that of the IS-95 system, therefore, in the subsequent sections, we will focus on the IS-95 system and the results can be directly applied to 3GPP systems.

Once the long-code sequence is recovered, the desired user’s signal can be recovered through signal separation and extraction techniques. If the training sequence is known, simple receivers, for example, the Rake receiver, can be used to extract the desired user’s signal. Even if the training sequence is unknown, the desired user’s signal can still be recovered through blind multiuser detection and signal separation algorithms, see [4–6], for example.

3. AES-BASED SECURITY ENHANCEMENT OF THE SCRAMBLING PROCESS

As can be seen from the previous sections, the physical layer security of CDMA systems relies on the scrambling process,
and the built-in information privacy provided by the operational and proposed CDMA systems is far from being adequate. In this paper, to enhance the physical layer built-in security of CDMA systems, we propose to generate the scrambling sequence using the advanced encryption standard (AES), also known as Rijndael.

Rijndael was identified as the new AES in October 2, 2000. Rijndael’s combination of security, performance, efficiency, ease of implementation, and flexibility makes it an appropriate selection for the AES. Rijndael is a good performer in both hardware and software across a wide range of computing environments. Its low memory requirements make it very well suited for restricted-space environments such as mobile handset to achieve excellent performance. More details on AES can be found in [7].

As mentioned before, we will focus our discussion on IS-95 system as it has a stronger physical layer security and the results can be directly applied to 3GPP systems. The proposed secure scrambling scheme aims to increase the physical layer built-in security of CDMA systems, to prevent exhaustive key search attack, while minimizing the changes required to the existing standards. As shown in Figure 4, the proposed secure scrambling is essentially a countermode AES. In Figure 4, $s_0, s_1, s_2, \ldots$ represents the output of the LFSR characterized by (5) as in the IS-95 system, $K$ is the 128-bit common secret encryption key shared between the base station and the mobile station ($K$ can also be 192 bits or 256 bits, as specified in the AES algorithm), and $M_0, M_1, \ldots, M_i$ denote successive message blocks with the same size as $K$, $d$ is the shift between the successive inputs to the AES engine. If the input to the $i$th encryption block is $s_{t+id}, s_{t+1+id}, \ldots, s_{t+127+id}$ with initial delay $t$, then the input to the $(i+1)$th block is $s_{t+(i+1)d}, s_{t+1+(i+1)d}, \ldots, s_{t+127+(i+1)d}$. The selection of $d$ should maximize the diversity between different inputs to the AES engine, which can be achieved by requiring $d$ and $2^d - 1$ to be relatively prime. In other words, $d$ should not be divided by 3, 7, 43, and 127.

The secure scrambling process can be summarized as follows:

1. The base station and the mobile station share a common initial state for the LFSR and an L-bit ($L = 128, 192$ or 256) common secret encryption key $K$.

2. The long scrambling sequence is generated through encryption of a particular segment of the sequence generated from the LFSR using the shared secret key $K$.

3. The scrambling process is realized by adding the scrambling sequence to the chip-rate spread signal.

For the 3GPP system, secure scrambling can be performed in the same manner by applying AES to the $i, Q$ scrambling sequences separately. As described in [8, 9], the shared secret data between the mobile station and base station can be updated from time to time. To prevent malicious key reload, the key update request can only be initiated from the base station.

4. SECURITY OF THE PROPOSED SCRAMBLING PROCESS

In this section, we use data encryption standard (DES) [10] as a benchmark to evaluate the security of the proposed secure scrambling, which is essentially ensured by AES. We compare the number of possible keys of AES and that of IS-95 scrambling sequence. The number of keys determine the effort required to crack the cryptosystem by trying all possible keys.

The most important reason for DES to be replaced by AES is that it is becoming possible to crack DES by exhaustive key search. Single DES uses 56-bit encryption key, which means that there are approximately $7.2 \times 10^{16}$ possible DES keys. In the late 1990s, specialized “DES cracker” machines were built and they could recover a DES key after a few hours. In other words, by trying all possible key values, the hardware could determine which key was used to encrypt a message [11]. Compared with DES, IS-95 has only 42-bit shared secret. The approximate number of keys is about $4.4 \times 10^{41}$, which is less than $10^4$ of the number of DES 56-bit keys. This makes it possible to break the IS-95 long-code mask almost in real time through exhaustive key search.

On the other hand, AES specifies three key sizes: 128, 192, and 256 bits. In decimal terms, this means that approximately there are

- (i) $3.4 \times 10^{38}$ possible 128-bit keys;
- (ii) $6.2 \times 10^{57}$ possible 192-bit keys;
- (iii) $1.1 \times 10^{77}$ possible 256-bit keys.
Thus, if we choose $L = 128$, then there are on the order of $10^{21}$ times more AES 128-bit keys than DES 56-bit keys. Assuming that one could build a machine that could recover a DES key in a second (i.e., try $2^{55}$ keys per second), as we can see, this is a very ambitious assumption and far from what we can do today, then it would take that machine approximately 149 thousand-billion (149 trillion) years to crack a 128-bit AES key. To put that into perspective, the universe is believed to be less than 20-billion-year old.

Security measurement through the number of all possible keys is based on the assumption that the attacker has no easy access to the secret encryption key, therefore, the attacker has to perform an exhaustive key search in order to break the system. As it is well known, the security of AES is based on the infeasible complexity in recovering the encryption key. Currently, no weakness has been detected for AES, thus, exhaustive key search is still being recognized as the most effective method in recovering the encryption key and breaking the cryptosystem. In our case, in order for the attacker to obtain the scrambling sequence, the attacker needs to know the input sequence and encryption key. It is reasonable to require that the 42-bit initial secret of the LFSR in Figure 4 to be kept a secret together with the 128-bit encryption key. And the attacker will only have access to the scrambled message sequence, for which the secure scrambling sequence is generated from encryption of a 128-bit segment of the LFSR sequence using 128-bit shared secret key between the mobile station and the base station.

As pointed out in Section 2, for the IS-95 system, the entire scrambling sequence can be regenerated as long as 42 successive bits of the scrambling sequence are recovered. In the proposed procedure, even if one block of the scrambling sequence is intercepted, the attacker still needs to recover the secret key $K$ and the input segments $[s_1..s_{127}]$ in order to regenerate the entire scrambling sequence, that is, the attacker still needs to break AES.

The key update technique currently used can reduce the risk for the opponent to maliciously reload a new key since the process is controlled by the base station. However, it is still essential to protect the encryption key and to protect the mobile station from being hacked by the malicious attackers.

5. PERFORMANCE ANALYSIS OF CDMA SYSTEMS WITH SECURE SCRAMBLING

Pseudorandom scrambling in CDMA systems provides physical layer built-in user privacy for information transmission. However, from communication point of view, scrambling was originally designed to reduce interference of mobiles that use the same channelization code in different cells, and to ensure performance stability among user population by providing the desired wideband spectral characteristics, since the Walsh functions may not spread each symbol’s power spectrum uniformly in the available frequency band [12, 13]. When applying secure scrambling, two natural questions are the following.

1. What effect does it have on system performance?
2. Will it introduce significant computational complexity?

In this section, it will be demonstrated that while providing strong physical layer built-in security, secure scrambling has comparable computational complexity and system performance with that of the conventional scrambling process. It is also shown that by scrambling the training sequence and the message sequence separately with two independent scrambling sequences, both information privacy and system performance can be further improved.

5.1. Computational complexity

In this section, we compare the computational complexity of the proposed secure scrambling and conventional
scrambling. For this purpose, we only need to compare the complexity of the two scrambling sequence generation methods. Note that they both use the same 42-bit LFSR as specified in (5). In IS-95, each bit of the long scrambling code is generated through

$$c(t) = m_1s_1(t) + m_2s_2(t) + \ldots + m_{42}s_{42}(t).$$  \hspace{1cm} (17)

For the proposed secure scrambling, every 128-bit block of the scrambling sequence is generated through one AES encryption process. Here, we compare the number of instructions required by each method for every 128 bits, and also the time required for every 128 bits using a Dell computer with 1024 M RAM and 2.8 GHz CPU speed. The results are provided in Table 1. As can be seen, the computational complexity of secure scrambling is comparable with that of the scrambling process used in IS-95.

### 5.2. System performance and further improvement using separately scrambled training

Under the same spectral efficiency, in this section, we compare the input-output BER (bit-error-rate) performance of CDMA systems with conventional scrambling and secure scrambling, respectively. In practical systems, after spreading and scrambling, passband PAM (pulse amplitude modulation) is performed. Mapping information bearing bits to symbols, passband PAM is equivalent to a complex-valued baseband PAM system [14]. When BPSK or QPSK is chosen, the modulo-two addition between the message bits and the spreading sequence or the scrambling sequence is now equivalent to multiplying the message symbols using binary (±1) sequences. In this paper, our discussion is based on the equivalent discrete-time baseband PAM model of CDMA systems, for which the spreading sequences and scrambling sequences are both binary antipodal sequences.

Based on (4), desired user’s signal can be extracted through a two-stage procedure. First, training-based channel estimation is performed through correlation. Second, Rake receiver is applied to combine multipath components. It should be pointed out that currently, it is a common practice in industry to choose the chip rate training sequence be all 1’s. The training sequence is put as a prefix to the chip-rate message sequence, and then it is scrambled using the long scrambling sequence. Channel estimation is therefore carried out based on the correlation property of the front part of the scrambling sequence.

*This practice has two drawbacks:* first, from security point of view, the front part of the scrambling sequence is exposed to attackers, which makes it possible to recover the whole scrambling sequence right away if secure scrambling is not used. This, at the meantime, illustrates the importance of secure scrambling, which can prevent the whole scrambling sequence being recovered based on the knowledge of part of it. Second, from the performance point of view, the correlation property of part of the scrambling sequence may not be ideal, and it can decrease the system performance due to nonaccurate channel estimation.

#### Separately scrambled training

To overcome these shortcomings, we propose to scramble the training sequence with an independent short scrambling sequence. The training sequence and its scrambling sequence are designed subject to the following constraints.

1. The short scrambling sequence is independent of the long scrambling sequence.
2. The short scrambling sequence has the same length as that of the training sequence.
3. The scrambled training sequence is a Gold sequence.

Or equivalently, we can choose the training sequence be a Gold sequence and then no scrambling is necessary for it. At the meantime, the information sequence is scrambled with the long scrambling sequence. In other words, training sequence is separated from the information sequence in the scrambling procedure. As a result, the long scrambling sequence will not be exposed to malicious attackers and the channel estimation can be performed based on the low cross-correlation of Gold sequences. We term the proposed approach as “separated training,” and denote the conventional practice by “non-separated training.”

In the simulation, we choose the processing gain to be $N = 16$, and consider the single receiver case. It is assumed that QPSK signals are transmitted over four-ray multipath channels for each user, with the first path to be the dominant path. The multipath delays are uniformly distributed over the interval $[0, N - 1]$. That is, the maximum multipath delay $L$ is allowed to be up to one symbol period, a reasonable assumption for wideband CDMA systems. The short scrambling sequence is chosen to be Gold sequences of length 63, and the training sequence is chosen to be a sequence of all 1’s of the same length. Without loss of generality, user 1 is chosen to be the desired user. Figure 5 shows the bit error rate (BER) versus different signal-to-noise ratio (SNR) levels, assuming 4 equal power users in the system. SNR is defined as the chip SNR with respect to user 1. Multipath channels and information sequence consist of 1024 QPSK symbols are generated randomly in each Monte Carlo run, and the result is averaged over 100 runs.

As can be seen, system with secure scrambling has comparable performance with that of IS-95, and “separated training” delivers much better results compared to that of “non-separated training.”

#### 5.3. Discussions and extension to other wireless systems

From the previous two sections, we can see that with a slight increase in complexity, the physical layer built-in security of the CDMA systems can be improved significantly. Moreover, secure scrambling has the error-tolerant feature, that is, an individual error in the received message will have a limited local effect, it will not prevent the decryption of other parts of the message. This feature is very helpful under scenarios where retransmission is difficult or even impossible.
| Method                | Number of operations required for every 128 bits | Time (in seconds) |
|----------------------|-----------------------------------------------|-------------------|
|                      | AND                                           | OR                | BIT-SHIFT     | TOTAL           |                   |
| IS-95                | 5376                                          | 5248              | 5376          | 16000           | 0.0226            |
| Secure scrambling    | 7096                                          | 6644              | 8640          | 22380           | 0.0536            |

Extension of the physical layer built-in security from CDMA systems to other wireless systems is partially possible. For example, the secure scrambling block can be implemented after the channel encoder in any wireless systems to introduce physical layer security. However, nonspread-spectrum system may not have the same antijamming features as the spread-spectrum systems, since the frequency domain diversity is not available anymore.

6. CONCLUSION

In this paper, security weakness of the operational and proposed CDMA systems is analyzed and an encryption-based secure scrambling process is presented. First, instead of using the long-code sequences generated by the LFSR directly, the scrambling sequences are generated through AES operations. As a result, the physical layer built-in security of the CDMA system is significantly increased with very limited complexity load. Second, it is shown that by scrambling the training sequence and the message sequence separately with two independent scrambling sequences, both information privacy and system performance can be further improved. Finally, error-tolerant decryption can be achieved through secure scrambling. The proposed scheme is very feasible and can readily be implemented for security enhancement in wireless networks.

ACKNOWLEDGMENTS

This paper is supported in part by MSU IRGP and Cyper Security Initiative.

REFERENCES

[1] R. Nichols and P. Lekkas, Wireless Security: Models, Threats, and Solutions, McGraw-Hill Telecom Professional Series, McGraw-Hill, New York, NY, USA, 2002.
[2] IEEE, “IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications,” November 1999.
[3] J. Massey, “Shift-register synthesis and BCH decoding,” IEEE Transactions on Information Theory, vol. 15, no. 1, pp. 122–127, 1969.
[4] S. Bhashyam and B. Aazhang, “Multiuser channel estimation and tracking for long-code CDMA systems,” IEEE Transactions on Communications, vol. 50, no. 7, pp. 1081–1090, 2002.
[5] C. J. Escudero, U. Mitra, and D. T. M. Slock, “A Toeplitz displacement method for blind multipath estimation for long code DS/CDMA signals,” IEEE Transactions on Signal Processing, vol. 49, no. 3, pp. 654–655, 2001.
[6] A. J. Weiss and B. Friedlander, “Channel estimation for DS-CDMA downlink with aperiodic spreading codes,” IEEE Transactions on Communications, vol. 47, no. 10, pp. 1561–1569, 1999.
[7] National Bureau of Standards, “FIPS Publication 197: Advanced Encryption Standard (AES),” November 2001, http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf.
[8] TIA/EIA/IS-95-B, “Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System,” 1998.
[9] V. K. Garg, IS-95 CDMA and cdma2000: Cellular/PCS Systems Implementation, Pearson Education, Upper Saddle River, NJ, USA, 1999.
[10] National Bureau of Standards, “FIPS Publication 81: DES Modes of Operation,” December 1980, http://www.itl.nist.gov/fipspubs/fip81.htm.
[11] Electronic Frontier Foundation (EFF), “EFF DES Cracker Project,” http://www.eff.org/Privacy/Crypto/Crypto_misc/DESCracker/.
[12] S. Parkvall, “Variability of user performance in cellular DS-CDMA-long versus short spreading sequences,” IEEE Transactions on Communications, vol. 48, no. 7, pp. 1178–1187, 2000.
[13] T. S. Rappaport, Wireless Communications: Principles and Practices, Prentice-Hall, Upper Saddle River, NJ, USA, 2nd edition, 2002.
[14] J. Proakis, Digital Communications, McGraw-Hill, New York, NY, USA, 4th edition, 2000.
Research Article

Tree-Based Distributed Multicast Algorithms for Directional Communications and Lifetime Optimization in Wireless Ad Hoc Networks

Song Guo, 1 Oliver W. W. Yang, 2 and Victor C. M. Leung 1

1 Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada V6T 1Z4
2 School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada K1N 6N5

Received 1 June 2006; Revised 29 October 2006; Accepted 30 October 2006

Recommended by Xiuzhen Cheng

We consider the problem of maximizing the network lifetime in WANETs (wireless ad hoc networks) with limited energy resources using omnidirectional or directional antennas. Unlike most solutions that use a centralized multicast algorithm, we use graph-theoretic approach to solve the problem in a distributed manner. After providing a globally optimal solution for the special case of single multicast session using omnidirectional antenna, this approach leads us to a group of distributed algorithms for multiple multicast in WANETs using directional antennas. Experimental results show that our distributed multicast algorithms for directional communications outperform other centralized multicast algorithms significantly in terms of network lifetime for both single-session and multiple-session scenarios.

Copyright © 2007 Song Guo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

There is an increasing interest in wireless ad hoc networks in many application domains where instant infrastructure is needed and no central backbone system and administration (like base stations and wired backbone in a cellular system) exist. Each communicating node in these networks acts as a router in addition to a host in order to communicate with each other over a limited number of shared radio channels. A communication session can be achieved either through a single-hop transmission if the communicating nodes are close enough to each other, or through multiple hops by relaying through intermediate nodes. Since each node in such a network is usually powered by a battery with limited amount of energy, the wireless ad hoc network will become unusable after the batteries are drained. Consequently, energy efficiency is an important design consideration for wireless ad hoc networks.

Over the last few years, energy efficient communication in wireless ad hoc networks with directional antennas has received more and more attention. This is because directional communications can save transmission power by concentrating RF energy where it is needed [1, 2]. On the other hand, the broadcast/multicast communication is also an important issue as many routing protocols for wireless ad hoc networks need this mechanism to maintain the routes between nodes. Therefore, one would be interested in finding an algorithm that would provide the maximum lifetime to the multicast session. The optimization metric is typically defined as the duration of the network operation time until the battery depletion of the first node in the network.

Some work has considered maximizing the network lifetime in a WANET with omnidirectional antennas for a single broadcast session, for example, [3–6], or a single multicast session, for example, [6–10]. The same problem with directional antennas has been studied in [1, 2, 11–14]. It has been proven to be an NP-hard problem [13]. The only exact solution for such difficult problem is the MILP formulation presented in [12]. In [1, 2], the authors extend the minimum energy metric by incorporating residual battery energy based on the observation that long-lived multicast/broadcast trees should consume less energy and should avoid nodes with small residual energy as well. The MLR-MD (for maximum lifetime routing for multicast with directional antenna) algorithm has been proposed recently in [13]. The basic idea of the MLR-MD algorithm is to start with a multicast routing solution first (e.g., a single beam from the source covering all multicast destination nodes) and then iteratively improve...
lifetime performance of the current solution by identifying the node with the smallest lifetime and revising routing topology as well as corresponding beamforming behavior for an increased network lifetime. All existing solutions are centralized, meaning that at least one node needs global network information in order to construct an energy efficient multicast tree.

In this paper, we explore the energy conservation offered by directional communications for providing long-lived broadcasting/multicasting in wireless ad hoc networks. Our focus is on establishing source-initiated multicast trees to maximize network operating time in energy-limited wireless ad hoc networks with single or multiple multicast sessions. Similar to previous research on the same problems \cite{1–14}, we only consider static networks because mobility adds a whole new dimension to the problem and it is out of the scope of this paper.

Unlike the previous work, we would like to design the distributed algorithms that can run on the wireless nodes with limited resources (i.e., bandwidth, memory, computational capacity, and power). We first use graph-theoretic approach to solve the special case of single multicast session using omnidirectional antenna. This graph-theoretic approach provides us insights into more general case of using directional antennas, and inspires us to produce a group of distributed algorithms. We will extend these solutions to maximize the network lifetime over multiple sessions as well as in more realistic scenarios for a wide range of potential civil and military applications. A straightforward approach is that the same trees that were optimized for single session operation are used for the multiple session operations.

The main contribution of this paper is that we present a group of distributed multicast algorithms for the network lifetime maximization problem in WANETs with omnidirectional antennas or directional antennas. In particular, we prove that our distributed algorithm for a single multicast session using omnidirectional antennas is globally optimal. Experimental results also show that our distributed multicast algorithms for directional communications outperform other centralized multicast algorithms significantly in terms of network lifetime for both single-session and multiple-session scenarios.

The rest of this paper is organized as follows. Section 2 develops the system model. Section 3 exploits some important properties of a min-max tree and proposes a group of distributed algorithms for both omnidirectional and directional antenna scenarios. Section 4 demonstrates the performance of our algorithms through a simulation study. Section 5 gives the conclusion on the results.

The following symbols and notations listed in Table 1 will pertain to the remainder of this paper.

2. SYSTEM MODEL

We model our wireless ad hoc network as a simple directed graph $G$ with a finite node set $N$ and an arc set $A$ corresponding to the unidirectional wireless communication links. Each node is equipped with a directional antenna, which concentrates RF transmission energy to where it is needed. We assume an ideal MAC layer that provides bandwidth availability, that is, frequency channels, time slots, or CDMA orthogonal codes, depending on the access schemes.

Assuming the transmitted energy at node $v$ to be uniformly distributed across the beamwidth $\theta_v$ ($\theta_{\min} \leq \theta_v \leq \theta_{\max}$), the minimal transmitted power required by node $v$ to support a link between two nodes $v$ and $u$ separated by a distance $r_{vu}$ ($r_{vu} > 1$) is proportional to $r_{vu}^{\alpha}$ and beamwidth $\theta_v$, where the propagation loss exponent $\alpha$ typically takes on a value between 2 and 4. Without loss of generality, all receivers

| Table 1: Symbols and notations. |
|----------------------------------|
| $G(A, N)$ | A directed graph modeling the wireless ad hoc network with a node set $N$ and an arc set $A$ corresponding to the unidirectional wireless communication link |
| $A(T_s)$ | The arc set of a multicast tree $T_s$ |
| $C_v$ | The child node set of node $v$ |
| $D$ | The set of destination nodes of a multicast session |
| $M$ | The set of multicast members including source node and all destination nodes |
| $N(T_s)$ | The node set of a multicast tree $T_s$ |
| $N_v$ | A set of neighboring nodes of node $v$ located within its maximum transmission range |
| $TN_v$ | A tree node set in which each node belongs to the multicast tree $T_s$ and lies in the maximum transmission range of node $v$ |
| $T_s$ | A multicast tree rooted at a source node $s$ |
| $p_{vu}$ | The RF transmission power needed for the link from node $v$ to node $u$ |
| $p_{\max}$ | The maximum RF transmission power level that a node can choose |
| $p_{\text{recv}}$ | The minimum power needed for reception processing |
| $p_{\text{trans}}$ | The minimum power needed for transmission processing |
| $r_{vu}$ | The distance between node $v$ and node $u$ |
| $w_{vu}$ | The weight for an arc $(v, u)$ in graph $G$ |
| $\alpha$ | The propagation loss exponent |
| $\delta(T_s)$ | The maximum weight of the arc in $T_s$ |
| $\delta_{\min}$ | The minimum $\delta(T_s)$ for all $T_s$ over $\Omega_M$ |
| $\delta_{\min}^\theta$ | The lower bound of $\delta_{\min}$ estimated at node $v$ |
| $\delta_{\max}^\theta$ | A lower bound of $\delta_{\min}$ |
| $\varepsilon_v$ | The residual battery energy of node $v$ |
| $\theta_v$ | The antenna beamwidth of node $v$ ($\theta_{\min} \leq \theta_v \leq \theta_{\max}$) |
| $\theta_v(C_v)$ | The minimum possible antenna beamwidth for node $v$ to cover a node set $C_v$ |
| $r_{\tau}$ | The maximal lifetime of a tree arc |
| $\Omega_M$ | The family of trees $T_s$ of $G$ spanning all the nodes in $M$ |
are assumed to have the same signal detection threshold, which is typically normalized to one. Then the transmission power $p_{vu}$ needed by node $v$ to reach node $u$ can be expressed as

$$p_{vu} = \frac{w_{vu}^* \cdot \theta_v}{360}. \quad (1)$$

Any node $v \in N$ can choose its power level, not to exceed some maximum value $p_{\text{max}}$. In addition to RF propagation, energy may also be expended for transmission processing (on modulation, encoding, etc.) and reception processing (on demodulation, decoding, etc.). For simplicity, these quantities are the same for any node, denoted as $p_{\text{tran}}$ and $p_{\text{recv}}$, respectively.

We consider a source-initiated multicast with a multicast set $M = \{s\} \cup D$, where $s$ is the source node and $D$ is the set of destination nodes. All the nodes involved in the multicast form a multicast tree rooted at the node $s$, that is, a rooted tree $T_s$, with a node tree set $N(T_s)$, and a tree arc set $A(T_s)$. We define a rooted tree as a directed acyclic graph with a source node with no incoming arcs, and each other node $v$ has exactly one incoming arc. A node with no out-going arcs is called a leaf node, and all other nodes are internal nodes (also called relay nodes). An important property of a rooted tree is that for any node $v$ in the rooted tree $T_s$, there must exist a single directed acyclic path in the tree.

Let the energy supply $\epsilon = \{\epsilon_u \mid u \in N\}$ be the initial energy level associated with each node in $G$. The residual lifetime $\tau_{vu}$ of a tree arc $(v, u)$ is therefore

$$\tau_{vu} = \begin{cases} \frac{\epsilon_v}{p_{vu} + p_{\text{tran}} + p_{\text{recv}}}, & v \neq s, \\ \frac{\epsilon_v}{p_{vu} + p_{\text{tran}}}, & v = s, \end{cases} \quad (2)$$

### 3. DISTRIBUTED MIN-MAX TREE ALGORITHMS

We first consider the graph representation of the WANET with omnidirectional antennas ($\theta_v = 360$), and assign

$$w_{vu} = \frac{1}{\tau_{vu}} = \begin{cases} \frac{w_{vu}^* + p_{\text{tran}} + p_{\text{recv}}}{\epsilon_v}, & v \neq s, \\ \frac{w_{vu}^* + p_{\text{tran}}}{\epsilon_v}, & v = s, \end{cases} \quad (3)$$

as the arc weight in the graph. It has been shown in [11] that the single session-based maximum lifetime multicast problem is equivalent to the min-max tree problem, which is to determine a directed tree $T_s$ spanning all the multicast members (i.e., $M \subseteq A(T_s)$) such that the maximum of the tree arc weight $\delta(T_s)$ is minimized, where

$$\delta(T_s) = \max \{w_{vu} \mid (v, u) \in A(T_s)\}. \quad (4)$$

Due to their equivalence, we will only investigate the properties of the min-max tree in this section. In the following, we will provide a related theorem that is used to derive our efficient algorithms.

![Figure 1: Illustration of the proof for Theorem 1. (The arrow line denotes the directed tree link and arrow curve denotes the directed path.)](image)

#### 3.1. A min-max tree theorem

Let $T_s^*$ be the min-max tree and $\Omega_M$ is the family of the trees spanning all the nodes in $M$, we therefore have

$$\delta_{\text{min}} \equiv \delta(T_s^*) \leq \delta(T_s), \quad \forall T_s \in \Omega_M. \quad (5)$$

A tree link $(v, u)$ is called the bottleneck link of the tree $T_s$ if $w_{vu} = \delta(T_s)$.

**Theorem 1.** Let $(v, u)$ be the bottleneck link of the multicast tree $T_s \in \Omega_M$. If there exists a node set $X, s \in X$ and $D \cap (N - X) \neq \phi$, such that $w_{vu} \leq w_{xy}$ for any $x \in X$ and $y \in N - X$, then $T_s$ is a min-max tree.

**Proof.** For any multicast tree $T_s^* \subseteq \Omega_M$, let $(v', u')$ be its bottleneck link. Note that there is at least one multicast member $z (z \neq s)$ belonging to $N - X$, that is, $z \in D \cap (N - X)$, since otherwise it contradicts the fact $D \cap (N - X) \neq \phi$. Therefore, there must exist an arc $(a, b) \in A(T_{s}^*)$, as shown in Figure 1, connecting $X$ and $N - X$ (i.e., $a \in X$ and $b \in N - X$) in order to satisfy the requirement that there exists a directed path from $s$ to the multicast member $z$.

From the given condition in Theorem 1, we have $w_{vu} \leq w_{ab}$. Furthermore, since $(a, b) \in A(T_{s}^*)$, the bottleneck link weight $\delta(T_{s}^*)$ of tree $T_{s}^*$ must be equal to or greater than the weight of any other tree link, for example, link $(a, b)$. That is, $w_{ab} \leq \delta(T_{s}^*)$. We thus can derive that $\delta(T_s) = w_{vu} \leq w_{ab} \leq \delta(T_{s}^*)$ for any $T_s^* \in \Omega_M$, that is, $T_s$ is a min-max tree.

#### 3.2. Min-max tree algorithm

Theorem 1 immediately suggests an MMT (min-max tree) algorithm for the maximum lifetime multicast problem as follows.

Initially, the multicast tree $T_s$ only contains the source node. It then iteratively performs the following search-and-grow procedure until the tree contains all the nodes in $M$. 

![Diagram](image)
The MMT(G, s) algorithm
(1) Initialize T₀ by setting N(T₀) = {s} and A(T₀) = φ.
(2) Repeat
   (i) Search phase:
       Find the arc(v, u) connecting N(Tᵢ) and N − N(Tᵢ)
       with minimum value wᵥᵤ, and then add (v, u) into
       the tree by setting N(Tᵢ+₁) = N(Tᵢ) ∪ {u} and
       A(Tᵢ+₁) = A(Tᵢ) ∪ {(v, u)}.
   (ii) Grow phase:
       while (exist link (x, y) connecting N(Tᵢ)
       and N − N(Tᵢ) such that wₓᵧ ≤ wᵥᵤ)
       Add (x, y) into the tree by setting
       N(Tᵢ+₁) = N(Tᵢ) ∪ {x} and
       A(Tᵢ+₁) = A(Tᵢ) ∪ {(x, y)}.
       until (M ⊆ N(Tᵢ)).
(3) Obtain the final multicast tree by pruning
    the broadcast tree Tᵢ.

Algorithm 1: The MMT algorithm.

Search-and-grow procedure

Find the link (v, u) connecting tree node set and nontree
node set with minimum weight wᵥᵤ, and then include it into
the multicast tree. Consequently, the tree Tᵢ would grow by
including as many nontree links (x, y) as possible into the
multicast tree if wₓᵧ ≤ wᵥᵤ until no more such links can be
found.

A pseudocode of the MMT algorithm is given in Algo-
rithm 1.

We will use a ten-node network as a simple example to
illustrate the basic tree construction steps in MMT. All nodes
are multicast members and node 0 is the source. Each node
has the same initial energy supply in a 10 × 10 square as
shown in Figure 2. The maximum transmission range is set
to 5 and a propagation loss exponent is α = 2.

Step 1. Initially, the tree consists of only the source node 0.

Step 2. In the first iteration, the link (0, 4) connecting node
sets {0} and {1, 2, 3, 4, 5, 6, 7, 8, 9} is found with minimum
weight, and then added into the tree as shown by the dark
arc in Figure 2(a). There is no any other link included in the
tree in the following grow operation.

Step 3. In the second iteration, the link (0, 7) connecting
node sets {0, 4} and {1, 2, 3, 5, 6, 7, 8, 9} is found with mini-
imum weight and added into the tree. The tree then grows by
including link (7, 9) as shown by the light arcs in Figure 2(b)
since w₇₉ < w₉₇.

Step 4. In the third iteration, the link (9, 1) connecting node
sets {0, 4, 7, 9} and {1, 2, 3, 5, 6, 8} is found with minimum
weight and added into the tree. The tree then grows by
including links (1, 3), (1, 5), (1, 6), (3, 8), and (6, 2) since their
weights are all less than w₉₁. The min-max tree is eventu-
ally obtained as shown in Figure 2(c) with the bottleneck link
(9, 1) that is found in the last iteration.

We have the following observations for the search-and-
grow process.

(1) Only one link is chosen in search phase, for example,
link (v, u) as shown in Figure 3, where Tᵢ is a par-
tially constructed multicast tree at the beginning of
this search phase.
(2) The weight wᵥᵤ denoted as δᵢ,LB, must be a lower bound
of δᵢ,min and it is given by
\[ \deltaᵢ,LB = \min \{ wₓᵧ | (x, y) ∈ A, x ∈ N(Tᵢ), y ∈ N − N(Tᵢ) \}. \]

(3) There would be multiple links to be included into the
multicast tree in a subsequent grow phase. A larger
constructed multicast tree Tᵢ is then obtained by the
end of the search-and-grow process.
(4) The new added links grow from certain nodes (e.g.,
node v), called grow points, by absorbing as many new
links as possible denoted as the tree branches in the
darker shaded area in Figure 3. It is interesting to note
that there would be multiple such grow points in Tᵢ,
for example, node v′, if wᵥᵤ = wᵥᵤ′.
(5) The sequence of the weight wᵥᵤ in the min-max tree
formation is in an increasing order and the final one
in the sequence is equal to δᵢ,min.
(6) After the multicast members are all in the tree, all re-
dundant links, indicated by the dotted arrows in Figure
3, should be pruned from the tree.
Finally, it remains to show that the multicast tree discovered by the MMT algorithm is a min-max tree. This is stipulated as follows.

**Lemma 1.** At least one bottleneck link of the tree constructed by MMT is included in the tree in a search operation.

**Proof.** We prove it by contradiction. Suppose that each bottleneck link, for example, \((x, y)\), of the tree constructed by MMT is added in the tree in a grow operation, and the link \((v, u)\) is included into the tree just in the preceding search operation. From the search-and-grow procedure, we have \(w_{xy} \leq w_{yu}\). On the other hand, \(w_{yu} \leq w_{xy}\) because \((x, y)\) is a bottleneck link of the tree. Therefore, we derive \(w_{xy} = w_{yu}\), that is, \((v, u)\) is also a bottleneck link, which contradicts the above assumption that all bottleneck links are included in grow operations.

**Theorem 2.** MMT constructs a min-max tree.

**Proof.** From the conclusion of Lemma 1, there exists a bottleneck link that is added into the tree in a search operation. Let \(T_s\) be the partially constructed multicast tree before entering such search operation. At this situation, the node set \(X = N(T_s)\) satisfies the conditions in Theorem 1 and therefore we conclude that the final tree obtained from the MMT algorithm is a min-max tree.

### 3.3. The DMMT-OA algorithm

The above analysis would allow us to design distributed algorithm. Our DMMT-OA (distributed MMT algorithm for omnidirectional antenna) uses search-and-grow cycles to discover a min-max tree. Such feature is beneficial to implement it in a distributed fashion. We have formulated a data structure to maintain locally the multicast forwarding state at each tree node \(v\): a membership status and the neighborhood table \(N_v\). The membership status indicates if this node is a source, receiver, or forwarder. A node can be both a receiver and forwarder. The neighborhood table \(N_v\) contains one entry for each neighbor \(u\) within its maximum transmission range. Each entry in the table includes a flag to indicate if the node \(u\) is a tree node or a nontree node. More specifically, if \(u\) is a tree node, the relationship to node \(v\) is further indicated as parent, child, or other (neither parent nor child). All tree nodes within \(N_v\) are denoted as \(TN_v\).

The distributed algorithm assumes an underlying bea- coning protocol which allows each node to be aware of the existence of all its neighbors and the information \(w_{xy}\) between any two neighbor nodes \(x\) and \(y\). After the neighbor discovery, any node \(v\) will create an entry for each neighbor \(u\) and set node \(u\) as nontree. When there is a multicast request, the source will begin to construct a min-max tree as follows.

In a search operation, each tree node \(v\) (initially only source node \(s\)) first locally calculates an estimation of the lower bound of \(\delta_{\min}\) as follows:

\[
\delta_{LB}^v = \min \{w_{yu} \mid u \in N_v - TN_v\}. \tag{7}
\]

It would unicast a multicast-join-reply (MJREP) message back to its parent with the parameter \(\delta_{LB}\) if \(v\) is a leaf node, or with the parameter \(\min[\delta_{LB}^v, \delta_{LB}^x] \mid x\) is a child node of \(v\)
after collecting all MJREPs from its children if \( v \) is a relay node. Note that node \( v \) does not send this message if the parent flag is not set yet. Furthermore, if \( v \) is a multicast member, it also attaches its own address in the MJREP message, which will be propagated to the source to notify its attendance to the multicast.

In this manner, the source will eventually obtain the lower bound \( \delta_{LB} \) just as given in (6) once all MJREPs are received from its children. If not all multicast members are included in the tree, the source will initiate the grow operation by propagating the multicast-join-request (MJREQ) messages with the parameter \( \delta_{LB} \) all over the network.

When receiving the first MJREQ message, each intermediate node \( v \) will first set the transmitting node (from which MJREQ is received) as parent in its neighborhood table, then send back an acknowledgment message which allows its parent node to set itself as a child. Node \( v \) would also forward MJREQ to any node \( u \) only if \( w_{vu} \leq \delta_{LB} \). All subsequent duplicate MJREPs (with the same request ID) from other nodes are simply dropped, while the corresponding relationship flag is set as other for each of these nodes in node \( v \)'s neighborhood table. The multicast forwarding state at each tree node \( v \) is set as follows. If node \( v \) is a destination, it will set it as receiver. In addition, if node \( v \) is a relay node (i.e., there is at least one entry with a child flag in its neighborhood table), it will set its membership status as forwarder.

After a short period of time, no more MJREqs would be received at node \( v \). This means that the grow operation completes around node \( v \), and it then goes to the search operation again as described earlier. Finally, a forwarding tree is created in these search-and-grow cycles until all members join the tree. After that, a min-max multicast tree is obtained by pruning all the unnecessary links in a distributed fashion from the nonmember leaf nodes.

The above DMMT-OA algorithm for the omnidirectional antenna networks can be straightforward applied for directional communications. Figure 4 shows the result by running the DMMT-OA algorithm for the scenario with \( \theta_{min} \) and \( \theta_{max} \) in which the shaded sectors indicate the areas covered by the directional antennas. This simple process is to reduce the antennas beamwidth of each internal node \( v \) to the smallest possible value that provides beam coverage of all its downstream neighbors in the tree, subject to the constraint \( \theta_{min} \leq \theta_v \leq \theta_{max} \).

### 3.4. The DMMT-DA algorithm

The DMMT-DA (distributed MMT algorithm for directional antennas) algorithm is similar in principle to DMMT-OA for the formation of min-max tree, in the sense that new nodes are added into the tree in search-and-grow cycles. We must first incorporate the antenna beamwidth into the arc weight as follows:

\[
\begin{align*}
  w_{vu} &= \begin{cases} 
    \frac{r_{vu}^2 \cdot \theta_v(C_v)}{360 \cdot \varepsilon_v} + \frac{P_{tran} + P_{doc}}{\varepsilon_v}, & v \neq s, \\
    \frac{r_{vu}^2 \cdot \theta_v(C_v)}{360 \cdot \varepsilon_v}, & v = s,
  \end{cases}
\end{align*}
\]

where \( \theta_v(C_v) \in [\theta_{min}, \theta_{max}] \) is the minimum possible antenna beamwidth for node \( v \) to cover all its children \( C_v \) in the tree.

Let \( T_s \) be the partially constructed tree obtained at the beginning of a search phase. In order to obtain the lower bound provided by (7) in this search phase, each tree node \( v \) needs to recalculate the weight \( w_{vu} \) using (8), in which the node set \( C_v \) is given as follows:

\[
C_v = \{x \mid (v, x) \in A(T_s)\} \cup \{u\}.
\]

In a grow operation, the new children, for example, node \( x \), of each tree node \( v \), should be included into the tree as many as possible if a tree structure is still maintained and \( w_{vu} \) is not greater than the lower bound \( \delta_{LB} \) that is obtained from the previous search operation, that is,

\[
C_v = \arg \max_{C_v} \left| \{x \mid x \in N_v - TN_v \wedge w_{vx} \leq \delta_{LB}\} \right|.
\]

Finally, we use the same network configuration in Figure 2 to illustrate the tree construction steps in DMMT-DA.

**Step 1.** Initially, the tree consists of only the source node 0.

**Step 2.** In the first iteration, the link \((0, 4)\) is found and added into the tree with minimum beamwidth \( \theta_0(4) = 30 \) as shown by the shaded sector in Figure 5(a). There is no any other link included in the tree in the following grow operation.

**Step 3.** In the second iteration, the link \((4, 1)\) is found and added into the tree with minimum beamwidth in the search operation. The tree then grows by including links \((1, 3), (1, 6), (3, 8), \) and \((6, 2)\) as shown in Figure 5(b), where
Figure 5: Examples of min-max tree construction using DMMT-DA algorithm.

Table 2: Parameter values for simulation.

| Parameters | Description                  | Values                        |
|------------|------------------------------|-------------------------------|
| \( n \)    | Network size                 | 100                           |
| \( \theta_{\text{min}} \) | Minimum antenna beamwidth    | \(10^\circ, 30^\circ, 60^\circ, 90^\circ, 180^\circ\), and \(360^\circ\) |
| \( \theta_{\text{max}} \) | Maximum antenna beamwidth    | \(360^\circ\)                 |
| \( p_{\text{max}} \) | Maximum RF power level       | 100                           |
| \( p_{\text{tran}} \) | Minimum power needed for transmission processing | 0.1*                         |
| \( p_{\text{recv}} \) | Minimum power needed for reception processing | 1*                           |
| \( E(\varepsilon) \) | Mean of the initial energy   | 500**                         |
| \( D(\varepsilon) \) | Variance of the initial energy | 200**                      |
| \( \alpha \) | Propagation loss exponent    | 2                             |

*We have also used other values of \((p_{\text{tran}}, p_{\text{recv}}) = (0, 0)\) and \((0.01, 0.1)\), and have observed similar simulation results.
**Can be arbitrary units that are consistent with the units of distance.

\[ \theta_i(\{3, 6\}) = \angle 316^\circ, \theta_i(\{8\}) = 30, \text{ and } \theta_i(\{2\}) = 30, \text{ since } w_{13}, w_{16}, w_{38}, \text{ and } w_{62} \text{ are all less than } w_{41}. \]

**Step 4.** In the third iteration, the link \((8, 5)\) is found and added into the tree with minimum beamwidth. The tree then grows by including links \((5, 9)\), and \((9, 7)\). The min-max tree is eventually obtained as shown in Figure 5(c) with the bottleneck link \((8, 5)\) that is found in the last iteration.

4. Performance Evaluation

We have evaluated the performance of our distributed algorithms in many network examples. The evaluation is done via simulation written in C++ for the set of heuristic algorithms \(I = \{\text{DMMT-OA, DMMT-DA, RB-MIP-}\beta, \text{D-MIP-}\beta\}\), where \(\beta\) is a parameter that reflects the importance assigned to the impact of residual energy\(^2\) [2]. We use RB-MIP-\(\beta\) and D-MIP-\(\beta\) to denote algorithms RB-MIP and D-MIP with different values of \(\beta\), respectively. We have only considered \(\beta = 0, 1, \text{ and } 2\). In each network example, a number of nodes are randomly generated within a square region \(10 \times 10\). The values of parameters used in simulation are given in Table 2.

We use the metric normalized network lifetime to evaluate and compare algorithm performance. It is defined as the ratio of actual network lifetime obtained using heuristic algorithm to the best solution obtained by choosing the maximum lifetime from all heuristic algorithms. Such metric provides a measure of how close each algorithm comes to provide the longest lifetime tree. Thus allows us to facilitate the comparison of different algorithms over a wide range of network examples.

4.1. Performance in single session scenarios

In experiments based on single sessions, multicast groups of a specified size \(m (m = 5, 25, 50, 100)\) are chosen randomly from the overall set of nodes. One of the nodes is randomly chosen to be the source. We randomly generated 100 different network examples, and we present here the average over those examples for all cases.

\(^1\)The symbol \(\angle abc\) indicates the degree of angle between \(\text{arc}(b, a)\) and \(\text{arc}(b, c)\).
\(^2\)The cost of a link \((v, u)\) is defined as \(c_{vu} = p_{vu} \cdot (E_v(0)/E_v(t))^\beta\), where \(E_v(t)\) is the residual energy at node \(v\) at time \(t\).
Figure 6 illustrates the mean normalized network lifetime as a function of multicast group size and minimal antenna beamwidth for all algorithms. In all cases, DMMT-DA provides much better performance than other algorithms, and its superiority is even greater in network examples with larger $\theta_{\text{min}}$, for example, always within 5% close to the best solution when $\theta_{\text{min}} \geq 30^\circ$. In fact, as guaranteed by Theorem 2, DMMT-DA degenerates into DMMT-OA and therefore both achieve the globally optimal solutions for the case of using omnidirectional antennas.

4.2. Performance in multiple session scenarios

In multiple session-based experiments, multicast requests arrive with interarrival times that are exponentially distributed with rate $1/n$ at each node. Session durations are exponentially distributed with mean 1. Multicast groups are chosen randomly for each session request; the number of destinations is uniformly distributed between 1 and $n - 1$. Similarly, we randomly generated a sequence of multicast requests in each scenario and the experimental results are obtained from
5. CONCLUSION

We have presented a group of distributed multicast algorithms for static WANETs with omnidirectional/directional antennas. The correctness of our algorithm in providing a maximum lifetime multicast tree has been proved as well for WANETs with omnidirectional antennas and single session. The performance of our algorithms in terms of network lifetime has been also validated using the simulations over a large number of network examples.

ACKNOWLEDGMENTS

This research was supported in part by the NSERC (Canada) Discovery Grant no. OGP004286, NSERC Research Grant no. OGP0042878 and an NSERC Postdoctoral Fellowship Award.

REFERENCES

[1] J. E. Wieselthier, G. D. Nguyen, and A. Ephremides, “Energy-limited wireless networking with directional antennas: the case of session-based multicasting,” in Proceedings of IEEE 21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM ’02), vol. 1, pp. 190–199, New York, NY, USA, June 2002.

[2] J. E. Wieselthier, G. D. Nguyen, and A. Ephremides, “Energy-aware wireless networking with directional antennas: the case of session-based broadcasting and multicasting,” IEEE Transactions on Mobile Computing, vol. 1, no. 3, pp. 176–191, 2002.

[3] I. Kang and R. Poovendran, “On the lifetime extension of energy-efficient multihop broadcast networks,” in Proceedings of the International Joint Conference on Neural Networks (IJCNN ’02), vol. 1, pp. 365–370, Honolulu, Hawaii, USA, May 2002.

[4] I. Kang and R. Poovendran, “Maximizing static network lifetime of wireless broadcast ad hoc networks,” in Proceedings of IEEE International Conference on Communications (ICC ’03), vol. 3, pp. 2256–2261, Anchorage, Alaska, USA, May 2003.

[5] A. K. Das, R. J. Marks, M. El-Sharkawi, P. Arabshahi, and A. Gray, “MDLT: a polynomial time optimal algorithm for maximization of time-to-first-failure in energy constrained wireless broadcast networks,” in Proceedings of IEEE Global Telecommunications Conference (GLOBECOM ’03), vol. 1, pp. 362–366, San Francisco, Calif, USA, December 2003.

[6] M. X. Cheng, J. Sun, M. Min, and D.-Z. Du, “Energy-efficient broadcast and multicast routing in ad hoc wireless networks,” in Proceedings of the 22nd IEEE International Performance, Computing and Communications Conference (IPCCC ’03), pp. 87–94, Phoenix, Ariz, USA, April 2003.

[7] B. Wang and S. K. S. Gupta, “On maximizing lifetime of multicast trees in wireless ad hoc networks,” in Proceedings of the International Conference on Parallel Processing (ICPP ’03), pp. 333–340, Taiwan, China, October 2003.

[8] P. Floreen, P. Kaski, J. Kohonen, and P. Orponen, “Multicast time maximization in energy-constrained wireless networks,” in Proceedings of the Joint Workshop on Foundations of Mobile Computing, pp. 50–58, San Diego, Calif, USA, September 2003.

[9] L. Georgiadis, “Bottleneck multicast trees in linear time,” IEEE Communications Letters, vol. 7, no. 11, pp. 564–566, 2003.
[10] S. Guo, V. C. M. Leung, and O. W. W. Yang, “A scalable distributed multicast algorithm for lifetime maximization in large-scale resource-limited multihop wireless networks,” in Proceedings of the International Wireless Communications and Mobile Computing Conference (IWCMC ’06), pp. 419–424, Vancouver, BC, Canada, July 2006.

[11] S. Guo and O. W. W. Yang, “Multicast lifetime maximization for energy-constrained wireless ad hoc networks with directional antennas,” in Proceedings of IEEE Global Telecommunications Conference (GLOBECOM ’04), vol. 6, pp. 4120–4124, Dallas, Tex, USA, November-December 2004.

[12] S. Guo and O. W. W. Yang, “Formulation of optimal tree construction for maximum lifetime multicasting in wireless ad hoc networks with adaptive antennas,” in Proceedings of IEEE International Conference on Communications (ICC ’05), vol. 5, pp. 3370–3374, Seoul, Korea, May 2005.

[13] Y. T. Hou, Y. Shi, H. D. Sherali, and J. E. Wieselthier, “Online lifetime-centric multicast routing for ad hoc networks with directional antennas,” in Proceedings of IEEE the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM ’05), vol. 1, pp. 761–772, Miami, Fla, USA, March 2005.

[14] S. Guo, V. C. M. Leung, and O. W. W. Yang, “Distributed multicast algorithms for lifetime maximization in wireless ad hoc networks with omni-directional and directional antennas,” in Proceedings of IEEE Global Telecommunications Conference (GLOBECOM ’06), San Francisco, Calif, USA, November-December 2006.
Research Article

Rate-Based Active Queue Management for TCP Flows over Wired and Wireless Networks

Jun Wang and Min Song

Department of Electrical and Computer Engineering, Old Dominion University, 231 Kaufman Hall, Norfolk, VA 23529, USA

Received 21 July 2006; Revised 18 November 2006; Accepted 22 November 2006

Recommended by Xiuzhen Cheng

Current active queue management (AQM) and TCP protocol are designed and tuned to work well on wired networks where packet loss is mainly due to network congestion. In wireless networks, however, communication links suffer from significant transmission bit errors and handoff failures. As a result, the performance of TCP flows is significantly degraded. To mitigate this problem, we analyze existing AQM schemes and propose a rate-based exponential AQM (REAQM) scheme. The proposed REAQM scheme uses the input rate as a primary metric and queue length as the secondary metric. The objectives of REAQM are to stabilize networks with low packet loss, low packet delay, and high link utilization regardless the dynamic of network conditions. We prove the global asymptotic stability of the equilibrium based on Lyapunov theory. Simulation results suggest that REAQM is capable of performing well for TCP flows over both wired and wireless networks, and has comparable implementation complexity as other AQM schemes.

Copyright © 2007 J. Wang and M. Song. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

The essence of network congestion control is that a sender adjusts its transmission rate according to the congestion measure of the underline networks. There are two components to accomplish this. One is a source algorithm that dynamically adjusts the transmission rate in response to the congestion; the other one is a link algorithm that implicitly or explicitly conveys information about the current congestion measure to sources using that link. In the current Internet, the source algorithm is carried out by TCP, and the link algorithm is performed by active queue management (AQM) schemes at routers. Examples of AQM schemes are RED [1], REM [2], AVQ [3], and Yellow [4]. TCP defines how the source rates are adjusted while AQM schemes define how the congestion measure is defined and updated.

According to the type of metrics used to measure congestion, AQM schemes can be classified into three catalogs: queue-based, rate-based, and schemes based on concurrent queue and rate metrics. In queue-based schemes, congestion is observed by average or instantaneous queue length and the control aim is to stabilize the queue length. The drawback of queue-based schemes is that a backlog is inherently necessitated. Rate-based schemes predict the utilization of the link, determine the level of congestion, and take actions based on the packet arrival rate. Rate-based schemes can provide early feedback for congestion [3]. Other AQM schemes deploy a combination of queue length and input rate to measure congestion and achieve a tradeoff between queue stability and responsiveness.

Current TCP/AQM algorithms assume that packet loss is mainly due to network congestion. However, these TCP/AQM algorithms are insufficient for the hybrid wired and wireless networks. In wireless networks, communication links have intrinsic characteristics that affect the performance of transport protocols including variable bandwidth, corruption, channel allocation delays, and asymmetry [5]. All of these cause significant noncongestion packet loss. The generic TCP protocol notifies sources to reduce their transmission rates only when facing packet losses due to network congestion. Therefore, the TCP protocol in hybrid wired and wireless networks needs not only to detect packet losses, but also to detect the reason of packet losses. Moreover, the degree of statistical multiplexing number of flows over wireless links is different from that over wired links [5]. For example, a wireless link only has several flows instead of hundreds of flows over a wired link. While existing TCP/AQM schemes are tuned well for wired links which have
high degree of statistical multiplexing, they may not work well for wireless links with only several flows.

In this paper, we present a new rate-based exponential AQM (REAQM) with explicit congestion notification (ECN) marking, and use REAQM to enhance the TCP performance over wireless links. The main idea of REAQM is to use the mismatch between input rate and link capacity as the primary metric, which exploits the early feedback benefit of rate-based marking. Furthermore, REAQM uses queue length to compute a coefficient of rate mismatch and use it as a secondary metric. The rationale of the secondary metric is to achieve a tradeoff between queue stability and responsiveness of the system. Using ECN marking to notify senders of incipient congestion, REAQM overcomes the packet loss problem of burst data and errors over wireless links, and also performs well for high degree of multiplex of flows over wired links.

The rest of the paper is structured as follows. In Section 2, we introduce some related work, including the analytical model for TCP/AQM algorithms, the state-of-the-art research results on AQM schemes, and the characteristics of wireless links. We present the design of REAQM and analytical results in Section 3. In Section 4, we discuss simulation design and results, and compare our REAQM scheme with other AQM schemes. Finally, we provide concluding remarks in Section 5.

2. RELATED WORK

We start this section by introducing three representative AQM schemes. RED is a queue-based, most popular AQM schemes used [1]. In RED, packets are randomly dropped before the buffer is completely full, and the drop probability increases with the average queue length. As source rates increase, queue length grows and more packets are marked/dropped. This prompts the sources to reduce their rates. RED configuration has been a problem since its first proposal, and many studies have tried to address this issue [6]. While the presence of a persistent queue indicates a congestion, its length gives very little information as to the severity of congestion. Therefore, decoupling queue length from congestion management has some benefits.

REM is a queue-and-rate-based scheme. The objective of REM is to stabilize both the input rate around link capacity and the queue length around a predetermined target, regardless of the number of users sharing the link [2]. In REM, each output queue maintains a price function as a congestion measure. The price is updated based on rate mismatch (i.e., difference between input rate and link capacity) and queue mismatch (i.e., difference between queue length and the predetermined target). Correspondingly, REM uses an exponential marking probability function. REM decouples congestion measure from performance measure such as packet loss, queue length, or packet delay. Since REM tries to maintain a target queue length regardless the number of flows, it limits its ability to handle burst traffic.

AVQ is primarily a rate-based scheme, as opposed to queue length or average queue length-based marking [3, 7]. AVQ maintains a virtual queue whose capacity is less than the actual capacity of the link. When a packet arrives in the real queue, the virtual queue is also updated to reflect the new arrival. Packets in the real queue are marked/dropped when the virtual buffer overflows. The virtual capacity at each link is then adapted to ensure that each link achieves a desired utilization of the link. Another rate-based scheme is called Yellow [4]. Yellow uses the load utilization as a primary metric to manage congestion, and a queue control function as a secondary metric to improve congestion control performance. According to Yellow, a load factor is calculated from link incoming rate and virtual available capacity, where the virtual available capacity is updated based on the queue control function. The queue control function is introduced to achieve a stable and smooth response. The main idea of Yellow is to perform queue management based on the link load factor, and to predict incipient congestion timely and accurately with controlled queuing delays, stability, and robustness. Packet marking probability is based on the load factor to achieve high link utilization and avoid high packet loss ratio.

Rate-based marking provides early feedback and responds fast when there exists a rate mismatch between input rate and link capacity. The advantages of this approach have been explored in [8]. The early feedback is more robust to the presence of extremely short flows or variability in the number of long flows in the network. When utilization is close to 100%, the variance introduced by short flows seems to lead to an undesirable transient behavior where excessively large queue lengths persist over long periods of time. Since queue length is a cumulative difference value of rate mismatch, queue metric is insensitive to current queue arrival and drain rates. Low [7] proposed a duality model for TCP/AQM algorithms to explain the equilibrium properties of a large network under TCP/AQM control, such as throughput, delay, queue length, loss probability, and fairness. Duality model considers the process of congestion control as a distributed computation by sources and links over a network to solve a global optimization problem in real time.

Gurtov and Floyd [5] consider the interplay between wireless links and transport protocols, and present the rules to appropriately model wireless links. Wireless links suffer from significant packet losses due to bit errors and handoffs. A TCP flow can only reduce its sending rate on losses due to network congestion, not on those due to wireless effects. New AQM algorithms have been proposed specifically for wireless links. Saghfo et al use instantaneous queue length to measure congestion and propose a deterministic dropping strategy, packet discard prevention counter, which is tailored to an estimate of the pipe capacity and TCP’s rate halving policy [9]. Li and Liu present an explicit feedback scheme with AQM to handle large bandwidth delay product and burst packet losses over wireless links [10]. The main problem in such scenarios is preventing the slow start overshoot of TCP connections. The overshoot happens because TCP detects congestion up to one RTT after filling the buffer and the first packet drop. The TCP’s sending rate at this point can be twice the available bandwidth of the path and thus generate many packet drops.
Most wireless networks suffer from bursty errors. A two-state Markov error model is used to capture the packet loss characteristics [11]. The two states represent a good state with a low-bit-error-rate (BER) value and a bad state with a high BER value. The link switches between good and bad states, and the lifetime of each state is exponentially distributed with a different mean value at each time. Moreover, the degree of statistical multiplexing number of flows over wireless links is different from that over wired links. For example, a wireless link only has several flows instead of hundreds of flows over a wired link. When the degree of statistical multiplexing is low and the buffer size is small, existing AQM algorithms may not perform sufficiently well. While many link technologies include forward error correction (FEC) and local retransmission for addressing corruption at the link layer, these mechanisms can introduce their own complications, and result in a high variability of bandwidth and delay on wireless links.

3. REAQM SCHEME

3.1. Scheme description

The objectives of REAQM scheme are to regulate the link utilization and stabilize the networking system regardless the changing network conditions. Similar to REM, REAQM maintains a variable, price, as a congestion measure and uses an exponential marking probability function. On the other hand, REAQM differs from REM in the definition of congestion measure. Denote \( p_l(t) \) as the price at queue \( l \) in period \( t \). The marking probability function of REAQM is

\[
m_l(t) = 1 - \phi^{-p_l(t)},
\]

(1)

where \( \phi \) is a constant larger than 1. The parameter \( \phi \) determines the range of loss or marking probability, which also depends on the range of price \( p_l(t) \). Ideally, \( \phi \) should be chosen in a way that the end-to-end probability observed at hosts is small, especially for the AIMD algorithm of TCP Reno and its variants.

At each packet arrival epoch, the price is updated according to the following equation:

\[
p_l(t+1) = \left[ p_l(t) + y \left( f(q_l(t)) y_l(t) - c_l \right) \right]^+.
\]

(2)

Here, \( p_l(0) = 0 \), and \([z]^+ = \max\{z, 0\}\). \( q_l(t) \) is the aggregate queue length at queue \( l \) in period \( t \), \( y_l(t) \) is the aggregate input rate to queue \( l \) in period \( t \), and \( c_l \) is the link capacity at queue \( l \). The smoothing parameter \( y > 0 \) is a step size which determines the speed of convergence of the algorithm. A larger \( y \) results in a faster convergence, but it also incurs a higher risk of oscillatory queue. The coefficient of rate mismatch is a function of queue mismatch at queue \( l \) in period \( t \):

\[
f(q_l(t)) = \frac{\alpha_l q_l(t)}{c_l} + 1,
\]

(3)

where weight parameter, \( \alpha_l \), is a constant. Obviously, \( f(q_l(t)) \geq 1 \), and it is equal to 1 only if \( q_l(t) = 0 \). If the value of \( \alpha_l \) is small, the queue length has less effect on the coefficient. Substituting (3) into (2), we get

\[
p_l(t+1) = \left[ p_l(t) + y \left( \frac{\alpha_l q_l(t) y_l(t)}{c_l} + y_l(t) - c_l \right) \right]^+.
\]

(4)

Price is updated, periodically or asynchronously, mainly based on rate mismatch. The price is incremented if rate mismatches is positive, and decremented otherwise. Rate mismatch is positive when the input rate exceeds the link capacity and negative otherwise. Queue length is used to compute the coefficient of rate mismatch and adjust the scale of rate mismatch. If queue length is small, the impact of rate mismatch decreases and REAQM is less aggressive. Otherwise, REAQM is more aggressive. The rationale behind price update is that the smaller queue length, the more rate mismatch is allowed. This makes a tradeoff between the system stability and utilization. If current input rate exceeds link capacity, the packet marking probability will increase; otherwise, it will decrease. Second, the larger the current queue length, the larger the coefficient of rate mismatch is; thus the larger the price value increases or decreases as well. While rate mismatch is positive, that is, \( y_l(t) > c_l \), the price always increases for any value of queue length. If \( \alpha_l \) is larger enough, frequently positive rate mismatch results in a small queue size and may lower system throughput. We modify the price update equation in this case as

\[
p_l(t+1) = \left[ p_l(t) + y \left( \frac{\alpha_l' q_l(t) (y_l(t) - y_0)}{c_l} + y_l(t) - c_l \right) \right]^+,
\]

(5)

where \( y_0 \) is target queue length while rate mismatch is positive, and \( \alpha_l' \) is relative weight parameter larger than \( \alpha_l \). REAQM has comparable implementation complexity as REM, since the mainly difference of REAQM and REM is the price function.

When the number of TCP flow increases, the rate mismatches grow and thus push up the price and hence marking probability. This increases the intensity of congestion signal to sources, which then reduce transmission rates. When transmission rates are too small, the mismatches will be negative. This lowers the price and marking probability and raises transmission rates, until eventually the mismatches are driven to zero yielding high utilization and negligible loss and delay in equilibrium. Recall that the mean queue length steadily increases as the number of flows increases in RED. In contrast, the price steadily increases while rate mismatches grows and the mean queue length is stabilized under REAQM.

Generally, ECN bit is used to inform the source which losses are due to congestion or wireless effects [2, 7]. If a packet is marked by setting its ECN bit, the mark is carried to the destination and then conveyed back to the source via acknowledgments. We set ECN bit to 1 while packets are probabilistically marked according REAQM scheme, and drop packets only when they arrive at a full buffer. Although this method cannot completely prevent buffer from overflowing and thus some packets are dropped by congestion, it differentiates most of error losses and congestion losses.
We conclude this section by stating that AQM must be more aggressive to avoid buffer overflow for high degree of statistical multiplexing and less aggressive to avoid underutilization for low degree of statistical multiplexing. Consider a single link accessed by many TCP sources with the same round-trip time (RTT). Assume fixed packet size and the link is equally shared amongst n TCP long-live flows. A congestion notification to one flow reduces the offered load by a factor of \((1 - nf)\). Therefore, the larger \(n\) is, the less impact of individual marking.

### 3.2. Stability analysis

In this section, we present a global stability analysis for REAQM based on the duality model for TCP/AQM system [7]. The global asymptotic stability of the equilibrium for REAQM scheme can be proved based on Lasalle’s invariance principle applied to a suitable Lyapunov function as in [12].

Let \((q^*, p^*)\) be an equilibrium of our system, \(q^*_l\) and \(x^*_l\) be the equilibrium source rate and buffer size, respectively. Further, let \(y^*_l\) be the equilibrium link rate for queue \(l\).

The queue length is taken as follows:

\[
q_l(t + 1) = [q_l(t) + y_l(t) - c_l]^+.
\]

From (2) and (6), the price and queue dynamic are taken to be

\[
\dot{p}_l(t) = \begin{cases} 
\gamma f(q_l(t)) y_l(t) - c_l, & p_l(t) > 0, \\
\gamma f(q_l(t)) y_l(t) - c_l^+, & p_l(t) = 0,
\end{cases}
\]

\[
\dot{q}_l(t) = \begin{cases} 
(y_l(t) - c_l^+), & q_l(t) > 0, \\
(y_l(t) - c_l)^+, & q_l(t) = 0,
\end{cases}
\]

**Theorem 1.** Given the system defined by (7), assume \(f(q_l)\) is strictly decreasing in \(q_l > 0\) and that \(R\) is of full row rank, then the unique equilibrium point \(q^* = 0\), \(p^*\) is globally asymptotically stable.

First, we introduce the candidate Lyapunov function \(V(q, p)\):

\[
V(q, p) = \sum_{l=1}^{L} \left( y_l a_l q_l^2 + (c_l - y_l^+)^+ p_l \right) + \sum_{i=1}^{S} \phi_i(q_i),
\]

where \(\phi_i(q_i) = \int_{q_i}^{q^*} (x^* - f_i(\sigma)) d\sigma\). Function \(V\) is nonnegative and radially unbounded.

Take the derivative of \(V(q, p)\) along trajectories of our system:

\[
\dot{V} = \sum_{l=1}^{L} \left[ y_l a_l q_l \dot{q}_l + (c_l - y_l^+)^+ \dot{p}_l \right] + \sum_{i=1}^{S} \dot{\phi}_i(q_i).
\]

The last term above can be rewritten as

\[
\sum_{i=1}^{S} \dot{\phi}_i(q_i) = (x^* - x)^T \dot{q} = (x^* - x)^T R^T \dot{p} = (y^* - y)^T \dot{p} = \sum_{l=1}^{L} (y_l^+ - y_l) \dot{p}_l.
\]

Substituting back, we have

\[
\dot{V} = \sum_{l=1}^{L} \left[ y_l a_l q_l \dot{q}_l + (c_l - y_l)^+ \dot{p}_l \right] = \sum_{l=1}^{L} v_l,
\]

where \(v_l = y_l a_l q_l \dot{q}_l + (c_l - y_l)^+ \dot{p}_l\). We will show that \(\dot{V} \leq 0\) for each \(l\). We apply the dynamic equation (7), and discuss the four cases:

1. \(q_l > 0\), \(p_l > 0\). Here

\[
\dot{v}_l = y_l a_l q_l \dot{q}_l + (c_l - y_l)^+ \dot{p}_l = y_l (y_l - c_l) [a_l q_l - (f(q_l) y_l - c_l)]
\]

\[
= y_l (y_l - c_l) \left[ a_l q_l - \left( \frac{a_l}{c_l} q_l y_l + y_l - c_l \right) \right] = -y_l \left( \frac{a_l q_l}{c_l} + 1 \right) (y_l - c_l)^2 = -y f(q_l) (y_l - c_l)^2.
\]

2. \(q_l > 0\), \(p_l = 0\). Here

\[
\dot{v}_l = (y_l - c_l) y_l a_l q_l - y_l \dot{q}_l = -y f(q_l) (y_l - c_l)^2.
\]

3. \(q_l = 0\), \(p_l > 0\). Here \(f(q_l) = 1\), so

\[
\dot{v}_l = (c_l - y_l) y_l a_l q_l - y_l \dot{q}_l = -y_l (y_l - c_l)^2.
\]

4. \(q_l = 0\), \(p_l = 0\). Here \(f(q_l) = 1\), so

\[
\dot{v}_l = (c_l - y_l) y_l a_l q_l - y_l \dot{q}_l = (c_l - y_l) [y_l (y_l - c_l)]^+.
\]

If \(y_l - c_l < 0\), \(y_l (y_l - c_l)]^+ = 0\), so \(\dot{v}_l = 0\). If \(y_l - c_l \geq 0\), \(\dot{v}_l = -y_l (y_l - c_l)^2\).

We thus confirm that \(\dot{V} \leq 0\) for every \(l\), thus \(\dot{V} \leq 0\). Based on Lyapunov’s stability theorem, the trajectory \((q(t), p(t))\) must remain bounded over the time, and the equilibrium point \((q^*, p^*)\) is stable in the sense of Lyapunov. Note that \(V = 0\) when either \(y_l - c_l < 0\), or \(y_l - c_l < 0\) and \(q_l = p_l = 0\), which is the desired stable zone. The set of states \((q, p)\) where the Lyapunov derivative is zero is as same as that in [12]. The system is globally asymptotically stable by means of Lasalle’s invariance principle.

### 4. SIMULATIONS AND ANALYSIS

Simulations are conducted using ns-2 simulator. Figure 1 shows the network topology, where \(n\) TCP flows share a bottleneck link that marks or drops packets according to some AQM scheme.
In Figure 1, the two routers (R₁ and R₂) are connected by a link with a capacity of 10 Mbps, which could be a wired or wireless link. The capacity of all other links is 100 Mbps. The propagation delay of the bottleneck link (between two routers) is set to be 10 milliseconds, and those of the other links are set to be uniformly distributed between 10 milliseconds and 30 milliseconds. NewReno is used as default transport protocol with the TCP data packet size 1000 bytes. The performance metrics are average queue length, link utilization, and packet loss ratio.

In the first experiment, we study the parameter setting of REAQM, especially, for the parameters $\alpha_l$. We assume that the bottleneck link is reliable and run the simulation for different number of TCP flows. We start the experiment with 50 FTP flows in the system, and new 50 FTP flows are added every 50 seconds until the total number of flows reach 250. The queue size is 50. The objective is to study the sensitivity for the parameter $\alpha_l$ and the thumb of rule for parameter configuration.

The performance influence of parameter $\alpha_l$ is given in Figure 2. The value of $\alpha_l$ determines the prominence given to the queue length when determining the level of congestion. As can be seen the average queue length (Figure 2(a)) is reduced with increasing $\alpha_l$. The smaller $\alpha_l$ is, the smaller the coefficient of rate mismatch is. Therefore, queue length has less effect on the rate mismatch. If only (4) is used, the price is always increased with positive rate mismatch. While small $\alpha_l$ is chosen, average queue length will be increased. The coefficient is small while we use small $\alpha_l$. The price increases lightly even with large queue length and positive rate mismatch. Eventually, it results in very large queue length. On the other hand, large $\alpha_l$ indicates that even a small queue length has relative effect on the coefficient of rate mismatch. If both the number of flows and $\alpha_l$ are large enough, average queue length is very small and link utilization is also low. Therefore, price is updated based on (5) in the experiment. For the same $\alpha_l$, average queue length decreases mostly while the number of flows increases. The reason is larger number of flows results in a higher probability of positive rate mismatch, which reduces the allowable queue length. It should be noticed that packet loss ratio (Figure 2(b)) is increased lightly with increasing $\alpha_l$. For different $\alpha_l$, the larger average queue length is, the smaller packet loss ratio is. Therefore, choosing different value of $\alpha_l$ makes a tradeoff between queue length and loss ratio. Packet loss ratio is increased with the increasing of number of flows. Link utilization (Figure 2(c)) is very stable.
for all $\alpha_l$ while the number of flows is larger than 50, and has oscillation while number of flows is 50. We can use different values of $\alpha_l$ to achieve small queue length or maintain small packet loss ratio.

In the following experiments, we compare REAQM with other AQM schemes over wired and wireless networks. In these experiments, we use the “gentle” version of RED. The parameters of REM were chosen as recommended in [2], and the parameters of AVQ were chosen as recommended in [3]. The queue size is 50 for all AQM schemes. The parameters of REAQM are set as follows: $\alpha_l = 0.11$, $\gamma = 0.001$, and $\Phi = 1.001$.

In the second experiment, we consider the bottleneck link is reliable, and FTP flows are added or dropped to the network. We start the experiment with only small number of FTP flows in the system, and new 50 FTP flows are added every 50 seconds until the total number of flows reach the maximal value, 300. Then 50 flows are dropped every 50 seconds until the total number of flows is equal to the initial number. For every 50 seconds, we calculate each performance metric. A long transient period is always with an increasing average queue length while new flows are added before the scheme is able to converge at new network condition. Therefore, average queue length over each 50 seconds interval captures persistent transients.

Figure 3 shows the performance of each AQM scheme under varying degree of flow multiplexing over the wired link. The overall performance of REAQM is comparable with REM and RED. Average queue length (Figure 3(a)) of REAQM is comparable with that of REM, and is larger than that of RED and AVQ. Packet loss ratio (Figure 3(b)) of REAQM is smaller than other schemes. Link utilization (Figure 3(c)) of REAQM is comparable with that of REM, and is higher than RED and AVQ. The average queue length and link utilization of REAQM are also very stable for different number of flows. REAQM uses rate mismatch as main metric to update the price, which makes it very robust to the present of variability in the number of flows. The performance of AVQ is different with other schemes. It has the smallest average queue length and link utilization.

In the third experiment, we compare the properties of various AQM schemes while the bottleneck link is wireless. We modify NewReno so that a source halves its window when it receives a mark or detects a loss through timeout, but retransmits without halving its window when it detects a loss through duplicate acknowledgments. The number of flows added or dropped at each time interval is 10 due to the low degree of flow multiplexing over wireless link.

Figure 4 shows the performance of each AQM schemes under varying degree of flow multiplexing over the wireless link. The average queue length (Figure 4(a)) of REAQM is slight smaller than that of REM, and higher than that of other two schemes. Packet loss ratios (Figure 4(b)) of all schemes are comparable. The link utilization (Figure 4(c)) of REAQM and REM are better than RED and AVQ. While the number of flows is small, link utilization of AVQ is lower than that of other schemes.

Figure 3: Performance test versus number of FTP flows for the different AQM schemes over wired link.
5. CONCLUSIONS

In wireless networks, packets are lost mainly because of bit errors and intermittent connectivity. In this paper, we have presented a new rate-based exponential AQM (REAQM) with explicit congestion notification (ECN) marking, and use REAQM to enhance the TCP performance over both wired and wireless links. The main idea of REAQM is to use the mismatch between input rate and link capacity as the primary metric, which exploits the early feedback benefit of rate-based marking. Furthermore, REAQM uses queue length to compute a coefficient of rate mismatch and uses it as a secondary metric. The rationale of the secondary metric is to achieve a tradeoff between queue stability and responsiveness of the system. The global asymptotic stability of REAQM has been proved using Lyapunov theory. Simulation results suggest that REAQM is capable of performing well for TCP flows over both wired and wireless networks, and has comparable implementation complexity as other AQM schemes.

ACKNOWLEDGMENT

The research of Dr. Min Song is partially supported by U.S. National Science Foundation (NSF) Career Award CNS-0644247.

REFERENCES

[1] S. Floyd and V. Jacobson, “Random early detection gateways for congestion avoidance,” IEEE/ACM Transactions on Networking, vol. 1, no. 4, pp. 397–413, 1993.
[2] S. Athuraliya, S. H. Low, V. H. Li, and Q. Yin, “REM: active queue management,” IEEE Network, vol. 15, no. 3, pp. 48–53, 2001.
[3] S. S. Kunniyur and R. Srikant, “An adaptive virtual queue (AVQ) algorithm for active queue management,” IEEE/ACM Transactions on Networking, vol. 12, no. 2, pp. 286–299, 2004.
[4] C. Long, B. Zhao, X. Guan, and J. Yang, “The yellow active queue management algorithm,” Computer Networks, vol. 47, no. 4, pp. 525–550, 2005.
[5] A. Gurtov and S. Floyd, “Modeling wireless links for transport protocols,” ACM SIGCOMM Computer Communication Review, vol. 34, no. 2, pp. 85–96, 2004.
[6] H. Ohsaki and M. Murata, “Steady state analysis of the RED gateway: stability, transient behavior, and parameter setting,” IEICE Transactions on Communications, vol. E85-B, no. 1, pp. 107–115, 2002.
[7] S. H. Low, “A duality model of TCP and queue management algorithms,” IEEE/ACM Transactions on Networking, vol. 11, no. 4, pp. 525–536, 2003.
[8] S. Deb and R. Srikant, “Rate-based versus queue-based models of congestion control,” in Proceedings of Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS /Performance ’04), pp. 246–257, New York, NY, USA, June 2004.
[9] M. Sago, F. Ludwig, M. Meyer, and J. Peisa, “Queue management for TCP traffic over 3G links,” in Proceedings of IEEE Wireless Communications and Networking Conference (WCNC ’03), vol. 3, pp. 1663–1668, New Orleans, La, USA, March 2003.

Figure 4: Performance test versus number of FTP flows for the different AQM schemes over wireless link.
[10] V. H. Li and Z.-Q. Liu, “PET: enhancing TCP performance over 3G & beyond networks,” in Proceedings of the 58th IEEE Vehicular Technology Conference (VTC ’03), vol. 4, pp. 2302–2306, Orlando, Fla, USA, October 2004.

[11] H. Wang and N. Moayeri, “Finite-state Markov channel—a useful model for radio communication channels,” IEEE Transactions on Vehicular Technology, vol. 44, no. 1, pp. 163–171, 1995.

[12] F. Paganini, “A global stability result in network flow control,” Systems and Control Letters, vol. 46, no. 3, pp. 165–172, 2002.
Research Article

Radar Sensor Networks: Algorithms for Waveform Design and Diversity with Application to ATR with Delay-Doppler Uncertainty

Qilian Liang

Department of Electrical Engineering, University of Texas at Arlington, Room 518, 416 Yates Street, Arlington, TX 76019-0016, USA

Received 30 May 2006; Revised 28 November 2006; Accepted 29 November 2006

Recommended by Xiuzhen Cheng

Automatic target recognition (ATR) in target search phase is very challenging because the target range and mobility are not yet perfectly known, which results in delay-Doppler uncertainty. In this paper, we firstly perform some theoretical studies on radar sensor network (RSN) design based on linear frequency modulation (LFM) waveform: (1) the conditions for waveform coexistence, (2) interferences among waveforms in RSN, (3) waveform diversity in RSN. Then we apply RSN to ATR with delay-Doppler uncertainty and propose maximum-likelihood (ML) ATR algorithms for fluctuating targets and nonfluctuating targets. Simulation results show that our RSN vastly reduces the ATR error compared to a single radar system in ATR with delay-Doppler uncertainty. The proposed waveform design and diversity algorithms can also be applied to active RFID sensor networks and underwater acoustic sensor networks.

Copyright © 2007 Qilian Liang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION AND MOTIVATION

The goal for any target recognition system is to give the most accurate interpretation of what a target is at any given point in time. There are two classes of motion models of targets, one for maneuvering targets and one for nonmaneuvering (constant velocity and acceleration) targets. The area that is still lacking in target recognition is the ability to detect reliably when a target is beginning a maneuver where its speed and range are uncertain. The tracking system can switch the algorithms applied to the problem from a nonmaneuvering set to the maneuvering set when a target is beginning a maneuver. But when the tracker does finally catch up to the target after the maneuver and then perform ATR, the latency is too high. In time-critical mission situation, such latency in ATR is not tolerable. In this paper, we are interested in studying automatic target recognition with range and speed uncertainty, that is, delay-Doppler uncertainty, using radar sensor networks (RSN). The network of radar sensors should operate with multiple goals managed by an intelligent platform network that can manage the dynamics of each radar to meet the common goals of the platform rather than each radar to operate as an independent system. Therefore, it is significant to perform signal design and processing and networking cooperatively within and between platforms of radar sensors and their communication modules. In this paper, we are interested in studying algorithms on radar sensor network (RSN) design based on linear frequency modulation (LFM) waveform: (1) the conditions for waveform coexistence, (2) interferences among waveforms in RSN, (3) waveform diversity in RSN. Then we apply RSN to ATR with delay-Doppler uncertainty. In nature, diverse waveforms are transmitted by animals for specific applications. For example, when a bat and a whale are in the search mode for food, they emit a different type of waveform than when they are trying to locate their prey. The Doppler-invariant waveforms that they transmit are environment dependent [1]. Hence, in RSN, it may be useful to transmit different waveforms from different neighbor radars and they can collaboratively perform waveforms diversity for ATR. Sowelam and Tewfik [2] developed a signal selection strategy for radar target classification, and a sequential classification procedure was proposed to minimize the average number of necessary signal transmissions. Intelligent waveform selection was studied in [3, 4], but the effect of Doppler shift was not considered. In [5], the performance of constant
The LFM waveform can be defined as will also consider the bandwidth constraint. Besides, the radar channel is narrowband, so we are not properly designed. We will introduce orthogonality as the criterion for waveform design in RSN to make them coexist. Although, the radar channel is narrowband, so we will also consider the bandwidth constraint.

In our radar sensor networks, we choose LFM waveform. The LFM waveform can be defined as

\[ x(t) = \sqrt{\frac{E}{T}} \exp(j2\pi \beta t^2), \quad -\frac{T}{2} \leq t \leq \frac{T}{2}, \]  

(1)

In radar, ambiguity function (AF) is an analytical tool for waveform design and analysis that succinctly characterizes the behavior of a waveform paired with its matched filter. The ambiguity function is useful for examining resolution, side lobe behavior, and ambiguities in both range and Doppler for a given waveform [10]. For a single radar, the matched filter for waveform \( x(t) \) is \( x^*(t) \), and the ambiguity function of LFM waveform is [10]

\[ A(\tau, F_D) = \left| \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) \exp(j2\pi F_D t) x^*(t - \tau) dt \right| = \frac{E}{T} \sin \left[ \pi \frac{\left| (F_D + \beta \tau) (T - |\tau|) \right|}{F_D + \beta \tau} \right], \quad -T \leq \tau \leq T. \]  

(2)

Three special cases can simplify this AF:

(1) when \( \tau = 0 \),

\[ A(0, F_D) = \left| \frac{E \sin(\pi F_D T)}{T \pi (F_D)} \right|; \]  

(3)

(2) when \( F_D = 0 \),

\[ A(\tau, 0) = \left| \frac{E \sin(\pi \beta \tau (T - |\tau|))}{T \pi \beta \tau} \right|, \quad -T \leq \tau \leq T; \]  

(4)

(3) and

\[ A(0, 0) = E. \]  

(5)

However, the above ambiguity is for one radar only (no co-existing radar).

For radar sensor networks, the waveforms from different radars will interfere with each other. We choose the waveform for radar \( i \) as

\[ x_i(t) = \sqrt{\frac{E}{T}} \exp\left[j2\pi(\beta t^2 + \delta_i t)\right], \quad -\frac{T}{2} \leq t \leq \frac{T}{2} \]  

(6)

which means there is a frequency shift \( \delta_i \) for radar \( i \). To minimize the interference from one waveform to the other, optimal values for \( \delta_i \) should be determined to have the waveforms orthogonal to each other, that is, let the cross-correlation between \( x_i(t) \) and \( x_n(t) \) be 0,

\[ \int_{-\frac{T}{2}}^{\frac{T}{2}} x_i(t)x_n^*(t)dt = \frac{E}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} \exp\left[j2\pi(\beta t^2 + \delta_i t)\right] \exp\left[-j2\pi(\beta t^2 + \delta_n t)\right] dt \]  

\[ = E \sin \left[ \pi (\delta_i - \delta_n) T \right]. \]  

(7)

If we choose

\[ \delta_i = \frac{i}{T}, \]  

(8)

where \( i \) is a dummy index, then (7) can have two cases:

\[ \int_{-\frac{T}{2}}^{\frac{T}{2}} x_i(t)x_n^*(t)dt = \begin{cases} E, & i = n, \\ 0, & i \neq n. \end{cases} \]  

(9)

So, choosing \( \delta_i = i/T \) in (6) can have orthogonal waveforms, that is, the waveforms can coexist if the carrier spacing is \( 1/T \) between two radar waveforms. That is, orthogonality amongst carriers can be achieved by separating the carriers by an integer multiple of the inverse of waveform pulse duration. With this design, all the orthogonal waveforms can work simultaneously. However, there may exist time delay and Doppler shift ambiguity which will have interferences to other waveforms in RSN.
3. INTERFERENCES OF LFM WAVEFORMS IN RADAR SENSOR NETWORKS

3.1. RSN with two radar sensors

We are interested in analyzing the interference from one radar to another if there exist time delay and Doppler shift. For a simple case where there are two radar sensors (i and n), the ambiguity function of radar i (considering interference from radar n) is

\[ A_i(t_i, t_n, F_{D_i}, F_{D_n}) \]

(10)

\[ = \left| \int_{-\infty}^{\infty} x_i(t) \exp\left(2\pi F_{D_i} t\right) dt \right| \]

(11)

\[ + x_n(t-t_n) \exp\left(2\pi F_{D_n} t\right) x_i^*(t-t_i) dt \]

\[ \leq \int_{T/2+\min(t_i, t_n)}^{T/2+\max(t_i, t_n)} x_n(t-t_n) \exp\left(2\pi F_{D_n} t\right) x_i^*(t-t_i) dt \]

(12)

\[ + \int_{T/2+\max(t_i, t_n)}^{T/2} x_i(t) \exp\left(2\pi F_{D_i} t\right) x_i^*(t-t_i) dt \]

\[ = \int_{T/2+\min(t_i, t_n)}^{T/2+\max(t_i, t_n)} x_n(t-t_n) \exp\left(2\pi F_{D_n} t\right) x_i^*(t-t_i) dt \]

(13)

To make analysis easier, we assume \( t_i = t_n = \tau \) which is a reasonable assumption because radar sensors can be coordinated by the clusterhead to send out LFM waveforms. Then (13) can be simplified as

\[ A_i(\tau, F_{D_i}, F_{D_n}) \approx \left| E \sin \left( \pi (n-i+F_{D_i} T) \right) \right| \]

(14)

\[ + \left| \frac{E \sin \left( \pi (F_{D_i} + \beta \tau) (T-|\tau|) \right)}{T \pi (F_{D_i} + \beta \tau)} \right| . \]

Some special cases of (14) are listed as follows.

1. If \( F_{D_i} = F_{D_n} = 0 \), then (14) becomes

\[ A_i(\tau, 0, 0) \approx \left| \frac{E \sin \left( \pi \beta \tau (T-|\tau|) \right)}{\pi \beta T \tau} \right| . \]

(15)

2. If \( \tau = 0 \), then (14) becomes

\[ A_i(0, F_{D_i}, F_{D_n}) \approx \left| E \sin \left( \pi (n-i+F_{D_i} T) \right) \right| \]

(16)

\[ + \left| E \sin \left( \pi F_{D_i} T \right) \right| . \]

3. If \( F_{D_i} = F_{D_n} = 0 \), \( \tau = 0 \), and \( \delta_i \) and \( \delta_n \) follow (8), then (14) becomes

\[ A_i(0, 0, 0) \approx E. \]

(17)

3.2. RSN with M radar sensors

It can be extended to an RSN with M radars. Assuming time delay \( \tau \) for each radar is the same, then the ambiguity function of radar 1 (considering interferences from all the other \( M-1 \) radars with CF pulse waveforms) can be expressed as

\[ A_1(\tau, F_{D_1}, \ldots, F_{D_M}) \approx \left| \sum_{i=2}^{M} E \sin \left( \pi (i-1+F_{D_i} T) \right) \right| \]

\[ + \left| \frac{E \sin \left( \pi (F_{D_i} + \beta \tau) (T-|\tau|) \right)}{T \pi (F_{D_i} + \beta \tau)} \right| . \]

(18)

Similarly, we can have three special cases.

1. If \( F_{D_1} = F_{D_2} = \cdots = F_{D_M} = 0 \), then (18) becomes

\[ A_1(\tau, 0, 0, \ldots, 0) \approx \left| \frac{E \sin \left( \pi \beta \tau (T-|\tau|) \right)}{\pi \beta T \tau} \right| . \]

(19)

Comparing it against (4), it shows that our derived condition in (6) can have a radar in RSN and it gets the same signal strength as that of a single radar (no coexisting radar) when the Doppler shift is 0.

2. If \( \tau = 0 \), then (18) becomes

\[ A_1(0, F_{D_1}, F_{D_2}, \ldots, F_{D_M}) \approx \left| \sum_{i=1}^{M} E \sin \left( \pi (i-1+F_{D_i} T+\beta \tau T) \right) \right| . \]

(20)

Comparing to (3), a radar in RSN has more interferences when unknown Doppler shifts exist.

3. If \( F_{D_1} = F_{D_2} = \cdots = F_{D_M} = 0 \), \( \tau = 0 \), and \( \delta_i \) in (6) follows (8), then (18) becomes

\[ A_1(0, 0, 0, \ldots, 0) \approx E. \]

(21)

4. APPLICATION TO ATR WITH DELAY-DOPPLER UNCERTAINTY

In RSN, the radar sensors are networked together in an ad hoc fashion. They do not rely on a pre-existing fixed infrastructure, such as a wireline backbone network or a base station. They are self-organizing entities that are deployed on demand in support of various events surveillance, battlefield, disaster relief, search and rescue, and so forth. Scalability concern suggests a hierarchical organization of radar sensor networks with the lowest level in the hierarchy being a cluster. As argued in [11–14], in addition to helping with scalability and robustness, aggregating sensor nodes into clusters has additional benefits:

1. conserving radio resources such as bandwidth;
2. promoting spatial code reuse and frequency reuse;
3. simplifying the topology, for example, when a mobile radar changes its location, it is sufficient for only the nodes in attended clusters to update their topology information;
(4) reducing the generation and propagation of routing information; and,
(5) concealing the details of global network topology from individual nodes.

In RSN, each radar can provide their waveform parameters such as $\delta_i$ to their clusterhead radar, and the clusterhead radar can combine the waveforms from its cluster members.

In RSN with $M$ radars, the received signal for clusterhead (assume it is radar 1) is

$$r_1(u,t) = \sum_{i=1}^{M} a(u)x_i(t - t_i) \exp(j2\pi F_{D_i} t) + n(u,t), \quad (22)$$

where $a(u)$ stands for radar cross section (RCS) and can be modeled using nonzero constants for nonfluctuating target and four Swerling target models for fluctuating target \cite{10}; $F_{D_i}$ is the Doppler shift of target relative to waveform $i$; $t_i$ is delay of waveform $i$ and $n(u,t)$ is additive white Gaussian noise (AWGN). In this paper, we propose a RAKE structure for waveform diversity combining, as illustrated by Figure 1.

According to this structure, the received $r_1(u,t)$ is processed by a bank of matched filters, then the output of branch 1 (after integration) is

$$\left|Z_{1}(u; t_1, \ldots, t_M, F_{D_1}, \ldots, F_{D_M})\right|$$

\[= \int_{-T/2}^{T/2} \left| r_1(u,t)x_1^*(t - t_1) ds \right| \]

\[= \int_{-T/2}^{T/2} \left[ \sum_{i=1}^{M} a(u)x_i(t - t_i) \exp(j2\pi F_{D_i} t) + n(u,t) \right] \]

\[\times x_1^*(t - t_1) dt \]

\[\approx \left| \sum_{i=1}^{M} a(u)x_i(t - t_i) \exp(j2\pi F_{D_i} t) + n(u,t) \right| \quad (23)

where $\int_{-T/2}^{T/2} n(u,t)x_1^*(t - t_1)dt$ can easily be proved to be AWGN, let

$$n(u,t_1) \triangleq \int_{-T/2}^{T/2} n(u,t)x_1^*(t - t_1) dt \quad (24)$$

follow a white Gaussian distribution. Assuming $t_1 = t_2 = \cdots = t_M = t_1$, then based on (18),

$$\left|Z_{1}(u; \tau, F_{D_1}, \ldots, F_{D_M})\right|$$

\[= \left| \sum_{i=2}^{M} a(u)x_i(t - t_i) \exp(j2\pi F_{D_i} t) + n(u,t) \right| \]

\[\approx \left| \sum_{i=2}^{M} a(u)x_i(t - t_i) \exp(j2\pi F_{D_i} t) + n(u,t) \right| \quad (25)

Similarly, we can get the output for any branch $m (m = 1, 2, \ldots, M)$,

$$|Z_m(u; \tau, F_{D_1}, \ldots, F_{D_M})|$$

\[= \left| \sum_{i=1, i \neq m}^{M} a(u)x_i(t - t_i) \exp(j2\pi F_{D_i} t) + n(u,t) \right| \quad (26)

So, $|Z_m(u; \tau, F_{D_1}, \ldots, F_{D_M})|$ consists of three parts, signal (reflected signal from radar $m$ waveform):

$$\left| \sum_{i=1, i \neq m}^{M} a(u)x_i(t - t_i) \exp(j2\pi F_{D_i} t) + n(u,t) \right| \quad (27)

and noise: $|n(u, \tau)|$. Delay-Doppler uncertainty happens quite often in target search and recognition where target range and velocity are not yet perfectly known.

We can also have three special cases for

$$|Z_m(u; \tau, F_{D_1}, \ldots, F_{D_M})| \quad (29)

(1) When $F_{D_1} = \cdots = F_{D_M} = 0$,

$$|Z_m(u; \tau, 0,0,\ldots, 0)\right| \approx \left|a(u)\sin(\pi(\tau - |\tau|)) + n(u, \tau)\right| \quad (30)

(2) If $\tau = 0$, then (26) becomes

$$|Z_m(u; 0, F_{D_1}, \ldots, F_{D_M})| \approx \left| \sum_{i=1}^{M} a(u)x_i(t - t_i) \exp(j2\pi F_{D_i} t) + n(u,t) \right| \quad (31)

(3) If $\tau = 0$ and $F_{D_1} = \cdots = F_{D_M} = 0$, then (26) becomes

$$|Z_m(u; 0, 0,0,\ldots, 0)\right| \approx |a(u) + n(u)| \quad (32)$$
How to combine all the $Z_m's (m = 1, 2, \ldots, M)$ is very similar to the diversity combining in communications to combat channel fading, and the combination schemes may be different for different applications. In this paper, we are interested in applying RSN waveform diversity to ATR, for example, recognition that the echo on a radar display is that of an aircraft, ship, motor vehicle, bird, person, rain, chaff, clear-air turbulence, land clutter, sea clutter, bare mountains, forested areas, meteors, aurora, ionized media, or other natural phenomena. Early radars were “blob” detectors in that they detected the presence of a target and gave its location in range and angle, and radar began to be more than a blob detector and could provide recognition of one type of target from another [7]. It is known that small changes in the aspect angle of complex (multiple scatter) targets can cause major changes in the radar cross section (RCS). This has been considered in the past as a means of target recognition, and is called fluctuation of radar cross section with aspect angle, but it has not had much success [7]. In this paper, we propose a maximum-likelihood automatic target recognition (ML-ATR) algorithm for RSN. We will study both fluctuating targets and nonfluctuating targets.

4.1. ML-ATR for fluctuating targets with delay-Doppler uncertainty

Fluctuating target modeling is more realistic in which the target RCS is drawn from either the Rayleigh or chi-square of degree four pdf. The Rayleigh model describes the behavior of a complex target consisting of many scatters, none of which is dominant. The fourth-degree chi-square models targets having many scatters of similar strength with one dominant scatter. Based on different combinations of pdf and decorrelation characteristics (scan-to-scan or pulse-to-pulse correlation), four Swerling models are used [10]. In this paper, we will focus on “Swerling 2” model which is Rayleigh distribution with pulse-to-pulse correlation. This pulse-to-pulse decorrelation implies that each individual pulse results in an independent value for RCS $\alpha$.

For Swerling 2 model, the RCS $|a(u)|$ follows Rayleigh distribution and its I and Q subchannels follow zero-mean Gaussian distributions with variance $\gamma^2$. Assume

$$a(u) = a_I(u) + ja_Q(u)$$

and $n(u) = n_I(u) + jn_Q(u)$ follows zero-mean complex Gaussian distribution with variance $\sigma^2$ for the I and Q subchannels. Observe (26), for given $\tau, F_D, (i = 1, \ldots, M),$

$$\sum_{i=1, i \neq m}^{M} a(u)E \sin \left[ \pi (i - m + F_D T) \right] + \frac{a(u)E \sin \left[ \pi (F_{Dn} + \beta T)(T - |\tau|) \right]}{T\pi(F_{Dn} + \beta T)}$$

$$= a(u)E \left[ \sum_{i=1, i \neq m}^{M} \sin \left[ \pi (i - m + F_D T) \right] + \frac{\sin \left[ \pi (F_{Dn} + \beta T)(T - |\tau|) \right]}{T\pi(F_{Dn} + \beta T)} \right]$$

follows zero-mean complex Gaussian distributions with variance $E^2 y^2 \left[ \sum_{i=1, i \neq m}^{M} \sin \left[ \pi (i - m + F_D T) \right] + \frac{\sin \left[ \pi (F_{Dn} + \beta T)(T - |\tau|) \right]}{T\pi(F_{Dn} + \beta T)} \right]^2$ for the I and Q subchannels. Since $n(u, \tau)$ also follows zero-mean Gaussian distribution, so $|Z_m(u, \tau, F_D, \ldots, F_{Dn})|$ of (26) follows Rayleigh distribution. In real world, the perfect values of $\tau$ and $F_D$ are not known in the target search phase and the mean values of $\tau$ and $F_D$ are 0, so we just assume the parameter of this Rayleigh distribution $b = \sqrt{E^2 y^2 + \sigma^2}$ (when $T$ and $F_D$ equal to 0).

Let $y_m \triangleq |Z_m(u; \tau, F_D, \ldots, F_{Dn})|$, then

$$f(y_m) = \frac{y_m}{E^2 y^2 + \sigma^2} \exp \left( - \frac{y_m^2}{2(E^2 y^2 + \sigma^2)} \right).$$

The mean value of $y_m$ is $\sqrt{\pi(E^2 y^2 + \sigma^2)/2}$ and the variance is $(4 - \pi)(E^2 y^2 + \sigma^2)/2$. The variance of signal is $(4 - \pi)E^2 y^2/2$ and the variance of noise is $(4 - \pi)\sigma^2/2$.

Let $y \triangleq [y_1, y_2, \ldots, y_M]$, then the pdf of $y$ is

$$f(y) = \prod_{m=1}^{M} f(y_m).$$

4.2. ML-ATR for nonfluctuating targets with delay-Doppler uncertainty

In some sources, the nonfluctuating target is identified as “Swerling 0” or “Swerling 5” model [15]. For nonfluctuating target, the RCS $a(u)$ is just a constant $a$ for a given target. Observe (26), for given $\tau, F_D, (i = 1, \ldots, M),$

$$\sum_{i=1, i \neq m}^{M} \sin \left[ \pi (i - m + F_D T) \right] + \frac{\sin \left[ \pi (F_{Dn} + \beta T)(T - |\tau|) \right]}{T\pi(F_{Dn} + \beta T)}$$

is just a constant. Since $n(u, \tau)$ follows zero-mean Gaussian distribution, so $|Z_m(u; \tau, F_D, \ldots, F_{Dn})|$ of (26) follows...
Rician distribution with direct path value

\[ \lambda = aE \left[ \sum_{i=1, i \neq m}^{M} \text{sinc} \left( \pi (i - m + F_{Di}, T) \right) \right. \]
\[ \left. + \frac{\sin \left( \pi (F_{Di} + \beta \tau) (T - |\tau|) \right)}{T \pi (F_{Di} + \beta \tau)} \right] \]

(39)

Since \( \tau \) and \( F_{Di} \) are uncertain and zero-mean, so we just use the approximation

\[ \lambda = aE \]

(40)

which is obtained when \( \tau \) and \( F_{Di} \) equal to 0.

Let \( y_m \triangleq |Z_m(u; \tau, F_{Di}, \ldots, F_{Dm})| \), then the probability density function (pdf) of \( y_m \) is

\[ f(y_m) = \frac{2y_m}{\sigma^2} \exp \left[ -\frac{(y_m^2 + \lambda^2)}{\sigma^2} \right] I_0 \left( \frac{2\lambda y_m}{\sigma^2} \right), \]

(41)

where \( \sigma^2 \) is the noise power (with I and Q subchannel power \( \sigma^2/2 \)), and \( I_0(\cdot) \) is the zero-order modified Bessel function of the first kind. Let \( y \triangleq [y_1, y_2, \ldots, y_M] \), then the pdf of \( y \) is

\[ f(y) = \prod_{m=1}^{M} f(y_m). \]

(42)

The ML-ATR algorithm to decide a target category \( C \) based on \( y \) can be expressed as,

\[ C = \arg \max_{n=1, \ldots, N} \max_{\lambda} \left| f(y \mid \lambda = E|a_n|) \right| \]
\[ = \arg \max_{n=1, \ldots, N} \left( \frac{2ym}{\sigma^2} \right) \prod_{m=1}^{M} \left[ I_0 \left( \frac{2E|a_n|y_m}{\sigma^2} \right) \right] \]
\[ \times \exp \left[ -\frac{(y_m^2 + E^2|a_n|^2)}{\sigma^2} \right] I_0 \left( \frac{2E|a_n|y_m}{\sigma^2} \right). \]

(43)

\section{5. Simulations}

Radar sensor networks will be required to detect a broad range of target classes. In this paper, we applied our ML-ATR to automatic target recognition with delay-Doppler uncertainty. We assume that the domain of target classes is known a priori (\( N \) in Sections 4.1 and 4.2), and that the RSN is confined to work only on the known domain.

For fluctuating target recognition, our targets have 6 classes with different RCS values, which are summarized in Table 1 [7]. We assume the fluctuating targets follow “Swerling 2” model (Rayleigh with pulse-to-pulse decorrelation), and assume the RCS value listed in Table 1 to be the standard deviation (sd) \( y_n \) of RCS \( a_n(w) \) for target \( n \). We applied the ML-ATR algorithm in Section 4.1 (for fluctuating target case) for target recognition within the six targets domain. We chose \( T = 0.1 \) ms and \( \beta = 10^6 \). At each average SNR value, we ran Monte-Carlo simulations for \( 10^5 \) times for each target. In Figures 2(a), 2(b), 2(c), we plotted the average ATR error for fluctuating targets with different delay-Doppler uncertainty and compared the performances of single-radar system, 5-radar RSN, and 10-radar RSN. Observe these three figures.

(1) The two RSNs vastly reduce the ATR error comparing to a single-radar system in ATR with delay-Doppler uncertainty, for example, the 10-radar RSN can achieve ATR error 9% comparing against the single-radar system with ATR error 37% at SNR = 32 dB with delay-Doppler uncertainty \( \tau \in [-0.1T, 0.1T] \) and \( F_{Di} \in [-200 \text{ Hz}, 200 \text{ Hz}] \).

(2) Our LFM waveform design can tolerate reasonable delay-Doppler uncertainty which are testified by Figures 2(b), 2(c).

(3) According to Skolnik [7], radar performance with probability of recognition error \( (p_r) \) less than 10% is good enough. Our 10-radar RSN with waveform diversity can have probability of ATR error much less than 10% for the average ATR for all targets. However, the single-radar system has probability of ATR error much higher than 10%. Our RSN with waveform diversity is very promising to be used for real-world ATR.

(4) Observe Figures 2(a), 2(c), the average probability of ATR error in Figure 2(c) is not as sensitive to the SNR as that in Figure 2(a), that is, ATR error curve slope becomes flat with higher delay-Doppler uncertainty, which means that the delay-Doppler uncertainty can dominate the ATR performance when it is too high.

For nonfluctuating target recognition, our targets have 6 classes with different RCS values, which are summarized in Table 1 [7]. We applied the ML-ATR algorithms in Section 4.2 (for nonfluctuating target case) to classify an unknown target as one of these 6 target classes. We chose \( T = 0.1 \) ms and \( \beta = 10^6 \). At each average SNR value, we ran Monte-Carlo simulations for \( 10^5 \) times for each target. In Figures 3(a), 3(b), 3(c), we plotted the probability of ATR error with different delay-Doppler uncertainty. Observe these figures.

(1) The two RSNs tremendously reduce the ATR error comparing to a single-radar system in ATR with delay-Doppler uncertainty, for example, the 10-radar RSN can achieve ATR error 9% comparing against the single-radar system with ATR error 22% at SNR = 22 dB with delay-Doppler uncertainty \( \tau \in [-0.2T, 0.2T] \) and \( F_{Di} \in [-500 \text{ Hz}, 500 \text{ Hz}] \).

(2) Comparing Figures 2(a), 2(b), 2(c) against Figures 3(a), 3(b), 3(c), the gain of 10-radar RSN for fluctuating target recognition is much larger than that for nonfluctuating
target recognition, which means our RSN has better capacity to handle the fluctuating targets. In real world, fluctuating targets are more meaningful and realistic.

(3) Comparing Figures 3(a), 3(b), 3(c) against Figures 2(a), 2(b), 2(c), the ATR needs much lower SNR for nonfluctuating target recognition because Rician distribution has direct path component.

6. CONCLUSIONS AND FUTURE WORKS

We have studied LFM waveform design and diversity in radar sensor networks (RSN). We showed that the LFM waveforms can coexist if the carrier frequency spacing is $1/T$ between two radar waveforms. We made analysis on interferences among waveforms in RSN and proposed a RAKE structure for waveform diversity combining in RSN. We applied the RSN to automatic target recognition (ATR) with delay-Doppler uncertainty and proposed maximum-likelihood (ML)-ATR algorithms for fluctuating targets and nonfluctuating targets. Simulation results show that RSN using our waveform diversity-based ML-ATR algorithm performs much better than single-radar system for fluctuating targets and nonfluctuating targets recognition. It is also demonstrated that our LFM waveform-based RSN can handle the delay-Doppler uncertainty which quite often happens for ATR in target search phase.

The waveform design and diversity algorithms proposed in this paper can also be applied to active RFID sensor networks and underwater acoustic sensor networks because LFM waveforms can also be used by these active sensor...
networks to perform collaborative monitoring tasks. In this paper, the ATR is for single-target recognition. We will continuously investigate the ATR when multiple targets coexist in RSN and each target has delay-Doppler uncertainty. In our waveform diversity combining, we have used spatial diversity combining in this paper. We will further investigate spatial-temporal-frequency combining for RSN waveform diversity.

ACKNOWLEDGMENTS

This work was supported by the US Office of Naval Research (ONR) Young Investigator Program Award under Grant no. N00014-03-1-0466. The author would like to thank ONR Program Officer Dr. Rabinder N. Madan for his direction and insightful discussion on radar sensor networks.

REFERENCES

[1] R. A. Johnson and E. L. Titlebaum, “Range-doppler uncoupling in the doppler tolerant bat signal,” in Proceedings of IEEE Ultrasonics Symposium, pp. 64–67, Boston, Mass, USA, October 1972.
[2] S. M. Sowelama and A. H. Tewfik, “Waveform selection in radar target classification,” IEEE Transactions on Information Theory, vol. 46, no. 3, pp. 1014–1029, 2000.
[3] P. M. Baggenstoss, “Adaptive pulselength correction (APLE-CORR): a strategy for waveform optimization in ultrawideband active sonar,” IEEE Journal of Oceanic Engineering, vol. 23, no. 1, pp. 1–11, 1998.
[4] D. J. Kershaw and R. J. Evans, “Optimal waveform selection for tracking systems,” IEEE Transactions on Information Theory, vol. 40, no. 5, pp. 1536–1550, 1994.
[5] R. Niu, P. Willett, and Y. Bar-Shalom, “Tracking considerations in selection of radar waveform for range and range-rate measurements,” IEEE Transactions on Aerospace and Electronic Systems, vol. 38, no. 2, pp. 467–487, 2002.

[6] Y. Sun, P. Willett, and R. Lynch, “Waveform fusion in sonar signal processing,” IEEE Transactions on Aerospace and Electronic Systems, vol. 40, no. 2, pp. 462–477, 2004.

[7] M. I. Skolnik, Introduction to Radar Systems, McGraw Hill, New York, NY, USA, 3rd edition, 2001.

[8] H. Deng, “Synthesis of binary sequences with good auto-correlation and cross-correlation properties by simulated annealing,” IEEE Transactions on Aerospace and Electronic Systems, vol. 32, no. 1, pp. 98–107, 1996.

[9] Q. Liang, “Waveform design and diversity in radar sensor networks: theoretical analysis and application to automatic target recognition,” in Proceedings of International Workshop on Wireless Ad Hoc and Sensor Networks (IWWAN ’06), New York, NY, USA, June 2006.

[10] M. A. Richards, Fundamentals of Radar Signal Processing, McGraw-Hill, New York, NY, USA, 2005.

[11] C. R. Lin and M. Gerla, “Adaptive clustering for mobile wireless networks,” IEEE Journal on Selected Areas in Communications, vol. 15, no. 7, pp. 1265–1275, 1997.

[12] A. Iwata, C.-C. Chiang, G. Pei, M. Gerla, and T.-W. Chen, “Scalable routing strategies for ad hoc wireless networks,” IEEE Journal on Selected Areas in Communications, vol. 17, no. 8, pp. 1369–1379, 1999.

[13] T.-C. Hou and T.-J. Tsai, “An access-based clustering protocol for multihop wireless ad hoc networks,” IEEE Journal on Selected Areas in Communications, vol. 19, no. 7, pp. 1201–1210, 2001.

[14] M. Steenstrup, “Cluster-based networks,” in Ad Hoc Networking, C. Perkins, Ed., chapter 4, pp. 75–138, Addison-Wesley, Reading, Mass, USA, 2001.

[15] P. Swerling, “Probability of detection for fluctuating targets,” IEEE Transactions on Information Theory, vol. 6, no. 2, pp. 269–308, 1960.
Research Article

Communication Timing Control with Interference Detection for Wireless Sensor Networks

Yuki Kubo1,2 and Kokuke Sekiyama3

1 Ubiquitous System Laboratory, Corporate Research and Development Center, OKI Electric Industry Co., Ltd., 2-5-7 Honmachi, Chuo-Ku, Osaka-Shi, Osaka 541-0053, Japan
2 Department of System Design Engineering, University of Fukui, 3-9-1 Bunkyo, Fukui-Shi, Fukui 910-8507, Japan
3 Department of Micro-Nano Systems Engineering, Nagoya University, Furo-Cho, Chikusa-Ku, Nagoya 464-8603, Japan

Received 31 May 2006; Revised 16 October 2006; Accepted 18 October 2006

Recommended by Xiuzhen Cheng

This paper deals with a novel communication timing control for wireless networks and radio interference problem. Communication timing control is based on the mutual synchronization of coupled phase oscillatory dynamics with a stochastic adaptation, according to the history of collision frequency in communication nodes. Through local and fully distributed interactions in the communication network, the coupled phase dynamics self-organizes collision-free communication. In wireless communication, the influence of the interference wave causes unexpected collisions. Therefore, we propose a more effective timing control by selecting the interaction nodes according to the received signal strength.

Copyright © 2007 Y. Kubo and K. Sekiyama. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

In recent years, research on wireless sensor networks has been promoted rapidly [1]. The sensor networks are composed of distributed sensor devices connected with wireless communication and sensing functions. Potential application fields of the sensor networks include stock-management systems, road traffic surveillance systems, and air-conditioning control systems of a large-scale institution and so on. There are many technical issues in the sensor networks. In this paper, we deal with two problems. One of them is a communication timing control for collision avoidance. Another is the influence of interference wave on the communication timing control. In order to cope with malfunctions and changes of the number of active sensor nodes, a distributed autonomous communication timing control is preferable to centralized approaches which must rely on a fixed base station in general.

In order to avoid the collision issue, TDMA [2] system has been presented, which is a multiplexing technology in the time domain that makes it possible to avoid collisions by assigning a communication slot to one frame. Hence, no collision occurs, and any node can obtain impartial communication right in TDMA. TDMA is widely used in cellular telephone systems. However, TDMA is fundamentally a centralized management technique depending on a base station and is applicable to a star link network. Meanwhile, distributed slot assignment TDMA approach for ad hoc networks has been proposed. In Ephremides and Truong algorithm [3], allocation of one transmission slot is assured for each node by preparing N slots for N nodes. In addition, it is possible to add more slot allocations by referring to information of the slot allocation within the two hop nodes for the collision avoidance based on the distributed algorithm. However, this algorithm requires total number of the node. Hence, this algorithm has a limitation in changing the number of nodes flexibly. USAP-MA [4] deals with a distributed slot assignment in TDMA for changes of the number of nodes. This method provides a dynamic change of frame length corresponding to the number of nodes and network topology, and improves bandwidth efficiency. Also, the other methods of slot reservation have been proposed for TDMA [4–6]. However, these TDMA-based approaches require a global time synchronization.

As another collision avoidance technique, CSMA [7, 8] has been widely used. CSMA is a simple and scalable protocol. In the case of low-traffic situation, CSMA works efficiently. However, according to the increase of nodes, communication throughput sharply declines due to occurrence of
frequent packet collisions. Such collisions should be avoided for not only improvement of the throughput efficiency, but also saving the electric energy consumption required in the retransmissions. Furthermore, several problems are pointed out with regard to the cost of carrier sense and hidden terminals. Also, with the CSMA-based approach, it is difficult to ensure impartial communication right because of the high contention of nodes that share communication channel.

Other research in the wireless sensor networks includes SMAC, SMACS. SMAC is based on CSMA, where each node broadcasts a sleep timing schedule to the neighbor nodes. The nodes receiving this message are to adjust the schedule of sleep, by which a node can save energy consumption. Although the problem of collision is inevitable, the aim of this research is focused on a timing control for energy saving. Hence, fundamental problems in CSMA remain unsolved. SMACS realizes an efficient communication based on synchronization between two nodes. These nodes attempt to schedule a communication timing with each other. Additionally, each node utilizes a different frequency band for a different link for collision avoidance. In this method, the risk of collisions can be reduced by random sharing of the frequency band. SMACS is different from the basic TDMA in that synchronization is required between two corresponding nodes while TDMA requires global synchronization. In general, global synchronization without a base station is hard to achieve. We have proposed a distributed communication timing control for collision avoidance named phase diffusion time-division method (PDTD) [12]. This method is a distributed communication timing control based on the dynamics of coupled phase oscillator among the peripheral nodes. Through local and fully distributed interactions, the coupled phase dynamics self-organizes the effective phase synchronous state that allows collision-free communication.

On the other hand, radio interference is an important problem in the wireless communication. Interference problems include two kinds of problems. One of them is to reduce influence of interference. Another problem concerns the communication timing under the influence of interference. Radio interference greatly influences the communication protocol [13]. Decentralized scheduling TDMA is based on the graph structure of the node connection within communication range. The issue of radio interference is not considered in decentralized scheduling TDMA. Therefore, in the presence of interference wave, it may not be an appropriate schedule method when considering the issue of interference. Also, in the case of CSMA-based protocol, hidden terminal collision avoidance mechanism based on RTS and CTS messages will not work appropriately [14]. In the previous timing control based on PDTD, we did not deal with radio interference problems. Therefore, unexpected collisions may occur in the real environment. In this paper, we propose the extended version of PDTD with interference detection (PDTD/ID). Each node exchanges the received signal strength and specifies the interference source node. This has to be incorporated for interaction nodes for collision avoidance in PDTD. We verify the efficiency of the proposed method by simulation experiments.

2. COMMUNICATION TIMING CONTROL

2.1. Outline of PDTD

In this section, we will review a basic concept of PDTD. We assume a situation in which a node periodically transmits data. The node is modeled as an oscillator that periodically repeats the states of the communication and noncommunication. Hence, mutual adjustment of the communication timing is formulated based on the coupled oscillator dynamics. The communication timing state of the node is expressed as a phase. The phase of the oscillator for node $i$ is denoted as $\theta_i$, and angular velocity is $\omega_i$. We suppose that each node can transmit data only within the phase interval $0 < \theta_i < \phi_c$ as depicted in Figure 1. If other nodes do not transmit in the interval $0 < \theta_i < \phi_c$, no collision occurs. Figure 1 shows the phase relation from the viewpoint of node 0. Figure 1(left) depicts initial state. In this case, the phase difference is not large enough, hence a collision occurs. If each node forms appropriate communication timing like Figure 1(right), collision does not occur. The node transmits the control message.
at $\theta_i = 0$ for communication timing control. Each node is assumed to know the phase value of the neighbor nodes by receiving the control message, and to calculate phase dynamics.

### 2.2. Node interaction

We explain the method of exchanging phase value with each other by the control message. The control message from node $i$ includes the following information:

1. one-hop neighbor node ID $j = (0, 1, 2, \ldots)$;
2. phase value of one-hop neighbor ($\tilde{\theta}_0, \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_j$);
3. received signal strength value from one-hop neighbor ($P_{i-0}, P_{i-1}, P_{i-2}, \ldots, P_{i-j}$).

The phase value of one-hop neighbor is used for calculation of communication timing control. The received signal strength value is used for selection of interference nodes. These are detailed in Sections 2.3 and 3. Since the control messages are transmitted by the same channel with the data messages, there is possibility that the control messages might be occasionally lost by collisions. However, the transmission of the control messages is executed periodically, it is unlikely that the control message is lost every time.

The process to convey node information to the neighboring nodes is explained as follows. The node is assumed to be able to know only its self phase value when calculating phase dynamics. However, the node estimates the phase value of the neighboring nodes from their control messages. In this paper, the neighbor node of which information is temporarily generated based on this estimation is called a virtual node. The node controls communication timing by the interaction with a virtual node. Figures 2(a) and 2(b) show the case that node 2 transmits the control message.

At $\theta_2 = 0$, then the control message includes information of nodes 1, 3, 4, and 5 that exist in one-hop neighbor. Node 1 that received this message generates virtual nodes corresponding to nodes 1, 2, 3, 4, and 5 listed in control message from node 2. The phase with dashed circle in Figure 2(b) denotes the corresponding node. A virtual node corresponding to node 2 (sender of the control message) is registered as one-hop neighbor node. Nodes 3, 4, and 5 (the other nodes contained in the control message) are registered as two-hop neighbor nodes. In this regard, node 3 is classified as the two-hop neighbor node from node 1. However, if node 1 is able to communicate directly with node 3, node 3 is registered as the one-hop neighbor node. Through sending and receiving of a periodic control message, each node has node information within two-hop neighbor nodes as a virtual node.

### 2.3. Communication timing control based on PDTD

#### Coupled phase dynamics

PDTD provides communication timing control based on phase dynamics for collision avoidance. Node $i$ interacts with a virtual node and forms appropriate phase-difference pattern. Let $\hat{\theta}_{ij}$ denote phase value of virtual node $j$ for node $i$. Then the governing equation is given by the following equations:

$$
\frac{d\theta_i}{dt} = \omega_i + \sum_{j \in K_i} k_j R(\Delta \hat{\theta}_{ij}) + \xi(S_i),
$$

$$
\Delta \hat{\theta}_{ij} = \hat{\theta}_{ij} - \theta_i,
$$

$$
\frac{d\hat{\theta}_{ij}}{dt} = \hat{\omega}_{ij},
$$

where $\omega_i$ and $\hat{\omega}_{ij}$ denote the angular velocity of node $i$ and virtual node $j$, respectively, and $k_j$ is the coupled strength value. $K_i$ is a virtual node set of node $i$. Every node is allowed to transmit data for $\phi_c/\omega_i(s)$ every cycle. $\xi(S_i)$ is a stochastic term, details of which are explained in Section 2.3. Interaction with the neighbor nodes is governed by phase-response
function \( R(\Delta \theta_{ij}) \) which is a repulsive function as follows:

\[
R(\Delta \theta_{ij}) = \begin{cases} 
\Delta \theta_{ij} - \phi, & \Delta \theta_{ij} \leq \phi, \\
0, & \phi < \Delta \theta_{ij} < 2\pi - \phi, \\
\Delta \theta_{ij} - 2\pi + \phi, & 2\pi - \phi \leq \Delta \theta_{ij}.
\end{cases}
\]

### Stochastic adaptation

When relying only on the repulsive interaction, the phase-difference pattern often fails to converge to the desired stationary state. Therefore, a stochastic adaptation term \( \xi(S_i) \) is introduced, which is determined by the estimated risk of the collision. As an evaluation index, phase overlap rate is defined. Node communication state is defined such that \( O_i = 1 \) denotes that node \( i \) is allowed to communicate, and \( O_i = 0 \) denotes that node \( i \) is prohibited to communicate, which is given by

\[
O_i(\theta_i(t)) = \begin{cases} 
1, & 0 \leq \theta_i < \phi, \\
0, & \phi \leq \theta_i < 2\pi.
\end{cases}
\]

Flag function to indicate phase overlap of communication timing between node \( i \) and virtual nodes is given by

\[
x_i(t) = \begin{cases} 
1, & O_i(\theta_i) = 1, \sum_{j \in K_i} O_j(\tilde{\theta}_{ij}) > 0, \\
0, & \text{else}.
\end{cases}
\]

\( x_i = 1 \) indicates that there is a phase overlap that would cause a collision. If \( \sum_{i=1}^{T_i} x_i(t) \neq 0 \), then one collision is counted for one cycle. Let \( n \) indicate the occurrence time of phase overlap for past \( n \) cycles. overlap rate \( c_i \) is given by

\[
c_i(t) = \frac{n}{N_i}.
\]

The stress of being exposed to the risk of collision is accumulated by the following mechanism:

\[
S_i(t) = 2S_i(t - \tau) + s(c_i),
\]

\[
s(c_i) = \begin{cases} 
0.0, & 0 \leq c_i < 0.2, \\
0.03, & 0.2 \leq c_i < 0.5, \\
0.05, & 0.5 \leq c_i < 0.8, \\
0.1, & 0.8 \leq c_i < 0.9, \\
0.3, & 0.9 \leq c_i.
\end{cases}
\]

where \( \tau = n \cdot T_i \) is a stress accumulating time scale. Random phase jump is implemented every \( n \cdot T_i \) cycles with probability \( S_i \), where if \( S_i > 1 \), then \( S_i = 1 \). After random phase jump, then \( S_i = 0 \). The destination of phase jump is decided as follows. Assume that node \( i \) has \( N_i \) virtual nodes, the phase of which is denoted as \( \tilde{\theta}_{ij} \). Sorting the phase value \( \tilde{\theta}_{ij} \) in ascending order, such as \( \tilde{\theta}_{ij}^{(1)} < \cdots < \tilde{\theta}_{ij}^{(k)} < \cdots < \tilde{\theta}_{ij}^{(N_i)} \), the corresponding node to \( k \)th phase value is \( v_k \). The destination of stochastic jump is depicted as shown in Figure 3. The list of destination \( u_k \) is given by

\[
\label{eq:9}
u_k = \frac{v_k + v_{k+1}}{2} \quad (k = 1, 2, \ldots, N_i - 1).
\]

The preferential selection probability \( u_k \) is decided by the equation

\[
\label{eq:10}
\begin{align*}
\frac{p_k}{\sum_{l=1}^{N_i-1} \exp \left( \beta(v_{k+l} - v_k) \right)} & \quad (l = 1, 2, \ldots, N_i - 1),
\end{align*}
\]

where \( \beta \) is a sensitivity parameter of the selection.

### 3. Communication Timing Control with Interference Node Detection

#### 3.1. Radio interference problem

In a wireless communication, even in the presence of weak interference wave, a node may fail to communicate if the desired wave strength from the node is weak. On the other hand, if the desired wave strength is sufficiently strong, the node may be able to receive data from the other node successfully despite presence of a strong interference wave. The reception error caused by an interference wave is estimated by signal-to-interference ratio (SIR). The threshold of SIR to correctly receive a signal is determined by modulation methods and spec of the receiver. In the communication timing control described in Section 2.3, however, the influence of interference wave was not taken into account in our model. In spite of the assumption that the interaction range is within the two-hop neighbors, interference waves can be reached beyond the interaction range, and hence this could cause unexpected collisions. Therefore, each node has to select the interaction nodes based on the relation between received signal wave strength and interference wave strength.

#### 3.2. Radio interference model

In this section, we discuss how the interference source is specified based on the received electric power. As shown in Figure 4, nodes \( i, j, k \) and \( k \) are placed, where the internode distance between nodes \( i \) and \( j \) and the one between nodes \( j \) and \( k \) are denoted by \( d_{ij}, d_{ij} \) respectively. The interference occurs in node \( j \) when node \( i \) transmits to node \( j \). Also assume that all nodes transmit in the same electric power \( p_t \). The received electric power \( p_d \) is assumed available by the following equation [14]:

\[
\label{eq:11}
p_d = \frac{c \cdot \tau}{d^a},
\]

where \( d \) is the distance between the sender node and the receiver node. \( a \) is the signal attenuation coefficient. \( c \) is the combined parameter that is related to the reception strength. Assume that node \( i \) is the transmitting source, and node \( k \) is
an interference source. With (11), SIR is defined as the ratio of the electric power between the desired signal from node $i$ and the interference wave from node $k$;

$$SIR = \frac{p(d_i)}{p(d_j)} = \left( \frac{d_i}{d_j} \right) ^ \alpha . \quad (12)$$

SIR has to be bigger than the threshold $e_{air}$ in order for the transmission from node $i$ to be successfully received in node $j$. Otherwise, in the case of $SIR \leq e_{air}$, the interference would occur in node $j$, and node $k$ is referred to as the interference source node for node $j$. In general, the existence range of interference source node is given by the following equation:

$$d_i \leq \sqrt{e_{air}} \cdot d_j . \quad (13)$$

We call the existence range of interference source node as ERIS in the following section. It can be said that ERIS is proportional to the distance $d_i$ by (13). In order for node $i$ to be able to communicate with node $j$ successfully, node $i$ has to specify which node can be the interference node for node $j$. Such nodes are referred to as the interference source nodes. Node $i$ is not allowed to transmit at the same time as the interference source node.

### 3.3. Interference node detection

#### Existence range of interference source

As mentioned in the previous section,

$$SIR = \frac{p(d_i)}{p(d_j)} > e_{air} \quad (14)$$

is required for successful communication in the presence of interference waves. Taking logarithm in (14), we obtain

$$P_k - P_j > E_{air} , \quad (15)$$

where $P_k = 10 \log_{10} p(d_i) , \; P_j = 10 \log_{10} p(d_j) , \; \text{and} \; E_{air} = 10 \log_{10} e_{air}$. Figure 5 shows the existence range of interference source (ERIS). Let $P_{min}$(dBm) be the minimum received signal strength for a successful communication. In the case that node 1 transmits to node 2 that is located on the boundary of communication range from node 1, the received signal strength on the boundary positions will become $P_{min}$(dBm). Hence, it is supposed that $P_i = P_{min}$ in (15), then $P_{min} - E_{air} > P_i$ is derived, which indicates that node 2 will fail to receive the transmission from node 1, if the strength of interference wave is larger than $P_i = P_{min} - E_{air}$(dBm). The ERIS, the corresponding range for $P_i$, will become larger than the communication range of node 2. Therefore, some extension is required for the timing control with two-hop neighbor nodes based on the PDTD because the interference wave may cause another collision. On the other hand, when node 1 transmits to node 3, which is closer than node 2, assume that node 3 receives the signal of strength $P_3 = P_{min} + E_{air}$(dBm). This is the case of $P_i = P_{min}$ in (15), where since $P_i - E_{air} > P_i$ and $P_{min} > P_i$ is obtained. This implies that the ERIS ($P_i$) is the same or inside of the communication range of node 3. Therefore, if the communication range is redefined as $P_i$ instead of $P_{min}$, it is possible to avoid the problem caused by the interference wave in PDTD.

#### Detection process of interference node

In this section, the process of interference node detection is addressed. This method is based on the evaluation of the received signal strength, where two different scenarios can be considered. The first case is that when node $a$ transmits to node $b$, the interference occurs in the destination node $b$ because of transmissions from some other nodes. In this case, node $a$ needs to specify which nodes are causing the interference to node $b$ (detection of the interference nodes), in an attempt to execute the timing control with such interference nodes. On the other hand, the second case is that the transmission from node $a$ to a destination node $c$ is causing an interference to node $b$, where node $a$ is becoming an interference node for node $b$ unintentionally, and such a node could exist many around node $a$. Hence, node $a$ is asked to specify the node set that can be interfered by the transmission of node $a$, and conduct a timing control with those nodes to avoid a potential collision. The first case is exemplified in more detail in Figure 6(a), where node 1 receives a control message from node 2 with the signal strength larger than $P_i$(dBm) in an attempt to
specify the interference nodes for node 2. As described in Section 2.2, the control message from node 2 includes the signal strength data which had been received by node 2 from the other nodes. In Figure 6(a), this control message includes data from nodes 1, 3, 4, 5, 6, 7, 8, and 10.

Let \( P_{b \rightarrow a} \) denote the received signal strength of node \( b \) from node \( a \), then node 1 compares \( P_{2 \rightarrow 1} \) (the desired signal) with \( P_{2 \rightarrow x} \), \( x = 3, 4, 5, 6, 8, 10 \) in order to judge as to whether each node \( x \) would become the interference source. From (15), if \( P_{2 \rightarrow 1} - P_{2 \rightarrow x} \leq E_{\text{air}} \), node \( x \) may cause the interference to node 2. Such a node set is defined as

\[
L_i(b \rightarrow a) = \{ x \mid P_{b \rightarrow x} \geq P_{b \rightarrow a} - E_{\text{air}}, x \neq a \}. \tag{16}
\]

Equation (16) represents the node set that could cause the interference to node \( b \) when node \( a \) transmits to node \( b \). It should be noted that the node set \( L_i(b \rightarrow a) \) is determined by node \( a \) based on the control message from node \( b \), hence node \( a \) is excluded from the set \( L_i(b \rightarrow a) \). As depicted in Figure 6(a), \( L_i(2 \rightarrow 1) = \{ 3, 4, 5, 6, 7 \} \) that are the nodes inside the range of dashed circle, \( P_{2 \rightarrow 1} - E_{\text{air}} \). While, the second scenario is exemplified in Figure 6(b) where there is no direct communication between nodes 1 and 9 but node 1 can receive the control message from node 9 with the signal strength of less than \( P_e \) (dBm) for the sake of the interaction in PDTD. In other words, node 1 is outside the communication range \( P_e \) though it is within the interaction range \( P_{\text{air}} \). Node 9 will have a direct communication with node \( x \), the signal strength of which is \( P_{9 \rightarrow x} > P_e \). When node 1 transmits to a peripheral node, such as node 2, the transmission from node 1 may interfere with the desired signal for node 9 from node \( x \), for instance, \( x = 12 \). Also, if \( P_{9 \rightarrow x} - P_{9 \rightarrow 1} \leq E_{\text{air}} \) holds, node 1 becomes an interference node to the desired signal for node 9. Therefore, the node set comprising the nodes that are interfered with the transmission of node 9 and prevented from receiving a desired signal from node 9 is defined as follows:

\[
C_i(b \rightarrow a) = \{ x \mid P_{9 \rightarrow x} \leq P_{b \rightarrow a} + E_{\text{air}}, P_{9 \rightarrow x} \geq P_e, x \neq a \}. \tag{17}
\]

It should be noted that since \( C_i(b \rightarrow a) \) is estimated by node \( a \) based on the received control message from node \( b \), node \( a \) is excluded from the node set \( C_i(b \rightarrow a) \). As an example, \( C_i(9 \rightarrow 1) = \{ 5, 12 \} \) is depicted in the confined colored area of Figure 6(b).

In this method, the parameters associated with necessary SIR threshold \( E_{\text{air}} \) and the minimum reception electric power \( P_{\text{air}} \) have to be preassigned in order to abstract the interference nodes. After every node specifies the interference nodes, it conducts a communication timing control with those included in \( L_i \) and \( C_i \). That is, the interaction nodes (the virtual node set for node \( i \) in (1) are adaptively specified as \( L_i(j - i) \cup C_i(j - i) \).

4. SIMULATION

4.1. Simulation setting

Simulations are conducted to illustrate performance of PDTD/ID. As a simulation setting, 10×10 nodes are assigned as follows.

Case 1 (regular grid model (Figure 7(a))). 10×10 nodes are assigned on the regular grid, where the internode distance is assumed as \( d = 25 \) (m).

Case 2 (perturbed grid model (Figure 7(b))). Node allocation is perturbed by the uniform random value in \([ -d/2, d/2] \) from the regular grid allocation.

The radio parameters and the node parameters are listed in Tables 1 and 2, respectively. Also, the node arrangement and communication range are depicted in Figure 7. The initial value of the phase \( \theta_i \) is randomly assigned in \([0, 2\pi] \) for both Cases 1 and 2. Since the purpose of this simulation is to verify the proposed timing control and interference node selection, we focus our argument on the timing control, hence the traffic model is simplified. Each node transmits packets in the phase interval \( 0 < \theta_i < \phi_i \) every cycle. It is preferable that the node decides \( \phi_i \) as autonomous. However, we decide \( \phi_i \)
Figure 7: Node arrangement and interference node.

Table 1: Radio parameters.

| Parameter | Value              |
|-----------|--------------------|
| $c \cdot p_t$ | Radio parameter     | 0.01135 |
| $\alpha$  | Signal attenuation coefficient | 4       |
| $E_{\text{sir}}$ | Necessary SIR       | 10 (dB) |
| $P_{\text{min}}$ | Lowest reception electric power | $-90$ (dBm) |

Table 2: Node parameters.

| Parameter | Value                   |
|-----------|-------------------------|
| $\phi_c$  | Available communication interval | $2\pi/15$ (Case 1) (rad) |
| $\omega$  | Eigenfrequency of node   | $2\pi/5$ (rad/s) |
| $\beta$   | Sensitivity of stochastic jump | 10       |

as a fixed value in this simulation. We evaluate the successful transmission rate that is defined as available communication time(s) per cycle normalized by the maximum communication time(s) per cycle ($\phi_c / \omega$). Collision rate is the collision state time(s) per cycle normalized by the maximum communication time(s) per cycle.

4.2. Simulation results

The results of node selection for interaction are shown in Figures 7(a) and 7(b), where the large circle indicates the communication range of node 34, and the small circle indicates the equivalent curve of the signal strength $P_c$ from node 34. The encircled nodes in Figure 7 imply the interference nodes in the case that node 34 transmits to a node within the small circle $P_c$ curve (or communication range); hence node 34 has to interact with encircled nodes for collision avoidance. Table 3 shows a specific example for signal strength values in the case of Figure 7(b). Table 3(a) shows the list of signal strength in the case that node 34 receives the control message from node 35, the information gathered by node 35. Node 34 specifies the interaction nodes based on (16). Because the value of SIR is less than the desired threshold $E_{\text{sir}} = 10$ (dB) as listed in Table 1 for successful reception, node 34 has to avoid the overlap of communication timing with nodes 25, 44, and 45. Table 3(b) shows the table of signal strength, when node 34 receives a control message from node 33, and node 34 selects interaction node based on (17). Because node 34 interferes with reception of node 33, node 34 has to avoid overlap of communication timing with 24 and 43. Thus, interaction nodes (encircled nodes in Figure 7) are selected autonomously.

As mentioned in Section 2.3, each node evaluates the overlap rate of communication phase by (7). It can be said that the phase-difference pattern for the communication timing control is completed when the overlap rate of all nodes converged to 0. The time series of average overlap rate is shown in Figures 8(a) and 9(a), and it can be seen that it took around 60–100 cycles to complete the timing control. Also, average successful transmission rate increased according to decline of the average overlap rate as shown in Figures 8(b) and 9(b). Because of the overhead of the control message for interactions, the average success transmission rate is inevitably below 1. After having converged to the stationary state, the successful transmission rate remained steady in the high value, and any collision did not occur as shown in Figures 8(c) and 9(c). Hence, it is confirmed that every node correctly specified the interference source nodes and effectively conducted the communication timing control with interaction nodes. During the timing formation, it was possible...
Table 3: Signal strength and interaction node selection.

| P_{35} \rightarrow P_{34} | Strength (dBm) | SIR (dB) |
|--------------------------|----------------|----------|
| P_{35} \rightarrow P_{34} | -83.6          | 49.9     |
| P_{35} \rightarrow P_{34} | -89.5          | 55.2     |
| P_{35} \rightarrow P_{34} | -82.9          | 58.6     |
| P_{35} \rightarrow P_{34} | -85.8          | 59.9     |
| P_{35} \rightarrow P_{34} | -85.8          | 59.9     |
| P_{35} \rightarrow P_{34} | -80.8          | 59.9     |
| P_{35} \rightarrow P_{34} | -67.2          | 80.9     |
| P_{35} \rightarrow P_{34} | -74.5          | 87.9     |
| P_{35} \rightarrow P_{34} | -68.5          | 78.9     |
| P_{35} \rightarrow P_{34} | -87.0          | 87.9     |
| P_{35} \rightarrow P_{34} | -80.0          | 87.9     |
| P_{35} \rightarrow P_{34} | -89.1          | 87.9     |
| P_{35} \rightarrow P_{34} | -85.0          | 87.9     |
| P_{35} \rightarrow P_{34} | -74.3          | 87.9     |
| P_{35} \rightarrow P_{34} | -81.5          | 87.9     |
| P_{35} \rightarrow P_{34} | -86.7          | 87.9     |

(b) Control message from 33, receiver node 34, corresponding to Figure 7(b)

Average collision rate indicates 0.1. That collision is caused by influence of nodes outside two hops. Additionally, available phase interval $\phi_c$ becomes small (with ID $2\pi/27$, without ID $2\pi/34$) so that a lot of interaction nodes exist. However, interference node detection has the limitation of range of destination node (Figure 5).

Figures 10(a) and 10(b) show the spatial distribution of the successful transmission rate and the collision rate. After having completed the timing control, the inequality of transmission right was prevented. In the conventional contention-based access control, the equal transmission right is difficult to achieve. Thus, the communication timing control which can also cope with the interference wave is realized in a static radio condition. However, the reception signal strength may change dynamically due to the influence of fading effect, a problem remaining to be dealt with in our future work.
5. CONCLUSION

In this paper, we proposed a novel communication timing control method for the wireless networks, named phase diffusion time-division method with interference detection, PDT/D-ID. Without interference detection, PDT/D may be faced with difficulty to operate in real environment. Through the local exchanging of received signal strength value, every node selects the interaction nodes for collision avoidance in the presence of interference wave. PDT/D-ID realizes a fully distributed timing control with the interference node detection. A model of the interference wave was examined for the simulation, and the simulation experiments illustrated satisfactory results in the large-scale network. Interaction node selecting method based on the reception signal strength is expected to be effective in the real environment.

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “Wireless sensor networks: a survey,” Computer Networks, vol. 38, no. 4, pp. 393–422, 2002.
[2] D. D. Falconer, F. Adachi, and B. Gudmundson, “Time division multiple access methods for wireless personal communications,” IEEE Communications Magazine, vol. 33, no. 1, pp. 50–57, 1995.
[3] A. Ephremides and T. V. Truong, “Scheduling broadcasts in multihop radio networks,” IEEE Transactions on Communications, vol. 38, no. 4, pp. 456–460, 1990.
[4] C. D. Young, “USAP multiple access: dynamic resource allocation for mobile multihop multichannel wireless networking,” in Proceedings of IEEE Military Communications Conference (MILCOM ’99), vol. 1, pp. 271–275, Atlantic City, NJ, USA, October-November 1999.

[5] M. K. Marina, G. D. Kondylis, and U. C. Kozat, “RBRP: a robust broadcast reservation protocol for mobile ad hoc networks,” in Proceedings of IEEE International Conference on Communications (ICC ’01), vol. 3, pp. 878–885, Helsinki, Finland, June 2001.

[6] Z. Tang and J. J. Garcia-Luna-Aceves, “A protocol for topology-dependent transmission scheduling in wireless networks,” in Proceedings of IEEE Wireless Communications and Networking Conference (WCNC ’99), vol. 3, pp. 1333–1337, New Orleans, La, USA, September 1999.

[7] L. Kleinrock and F. Tobagi, “Packet switching in radio channels: part I—carrier sense multiple-access modes and their throughput-delay characteristics,” IEEE Transactions on Communications, vol. 23, no. 12, pp. 1400–1416, 1975.

[8] F. Tobagi and L. Kleinrock, “Packet switching in radio channels: part II—the hidden terminal problem in carrier sense multiple-access and the busy-tone solution,” IEEE Transactions on Communications, vol. 23, no. 12, pp. 1417–1433, 1975.

[9] S. G. Glisic, “1-persistent carrier sense multiple access in radio channels with imperfect carrier sensing,” IEEE Transactions on Communications, vol. 39, no. 3, pp. 458–464, 1991.

[10] W. Ye, J. S. Heidemann, and D. Estrin, “An energy-efficient MAC protocol for wireless sensor networks,” in Proceedings of 21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM ’02), pp. 1567–1576, New York, NY, USA, June 2002.

[11] K. Sohrabi, J. Gao, V. Ailawadhi, and G. J. Pottie, “Protocols for self-organization of a wireless sensor network,” IEEE Personal Communications, vol. 7, no. 5, pp. 16–27, 2000.

[12] K. Sekiyama, Y. Kubo, S. Fukunaga, and M. Date, “Distributed time division pattern formation for wireless communication networks,” International Journal of Distributed Sensor Networks, vol. 1, no. 3-4, pp. 283–304, 2005.

[13] J. Li, C. Blake, D. S. J. De Couto, H. I. Lee, and R. Morris, “Capacity of ad hoc wireless networks,” in Proceedings of the 7th ACM International Conference on Mobile Computing and Networking, pp. 61–69, Rome, Italy, July 2001.

[14] F. Ye, S. Yi, and B. Sikdar, “Improving spatial reuse of IEEE 802.11 based ad hoc networks,” in Proceedings of IEEE Global Telecommunications Conference (GLOBECOM ’03), vol. 2, pp. 1013–1017, San Francisco, Calif, USA, December 2003.
Research Article

Bandwidth Optimization in Centralized WLANs for Different Traffic Types

R. J. Haines, N. Fanning, T. Lewis, and J. Coon

Telecommunications Research Laboratory, Toshiba Research Europe Ltd., 32 Queen Square, Bristol BS1 4ND, UK

Received 31 May 2006; Revised 24 November 2006; Accepted 10 January 2007

Recommended by Wei Li

Allocating bandwidth between different forms of coexisting traffic (such as web-browsing, streaming, and telephony) within a wireless LAN is a challenging and interesting problem. Centralized coordination functions in wireless LANs offer several advantages over distributed approaches, having the benefit of a system overview at the controller, but obtaining a stable configuration of bandwidth allocation for the system is nontrivial. We present, review, and compare different mechanisms to achieve this end, and a number of different means of obtaining the configurations themselves. We describe an analytical model of the system under consideration and present two mathematical approaches to derive solutions for any system configuration and deployment, along with an adaptive feedback-based solution. We also describe a comprehensive simulation-based model for the problem, and a prototype that allows comparison of these approaches. Our investigations demonstrate that a self-adaptive dynamic approach far outperforms any static scheme, and that using a mathematical model to produce the configurations themselves confers several advantages.

Copyright © 2007 R. J. Haines et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

The IEEE802.11 protocols [1] have become the dominant standard for wireless local area networks (WLANs). These protocols have evolved to support a variety of traffic types, which benefit from different scheduling and control mechanisms. There are two common forms of traffic encountered by WLANs. The first is sporadic, bursty data traffic, which is most efficiently served by highly distributed contention-based access schemes. The second is traffic with stringent quality-of-service (QoS) requirements, such as bandwidth, delay and/or jitter, which needs a more structured approach to provide guarantees of access.

There are two complementary approaches to serving QoS-sensitive traffic: distributed and centralized access [2]. Distributed approaches have largely focused on differentiated access that prioritizes different traffic types but then relies on statistical guarantees of access for each priority level, although more recent developments also incorporate a distributed reservation mechanism [3]. Centralized approaches, where a central controller allocates resources, benefit from having a global view of the entire system, and from being able to concentrate complexity in a single (more feature-rich, more expensive, higher-powered) device.

The IEEE802.11 standards offer both centralized and distributed controls. In this work, we concentrate on the centralized point coordination function (PCF), which can be seen as a specialized case of the more flexible and complex hybrid coordination function (HCF) of IEEE802.11e [4]. These centralized approaches are most attractive in single-access-point scenarios such as commonly found in the home, as there are scheduling complexities that arise with multiple-access-point scenarios. The PCF allows the coexistence of both QoS-sensitive traffic and bursty data traffic through the polling of the former and the direct contention of the latter. This is achieved by overlaying a repeating time-division superframe onto the medium, with distinct phases for polled and contending traffic.

The configuration of this superframe directly affects the system’s ability to support the two types of traffic effectively. If the configuration is badly wrong, then the QoS requirements may be missed, or the data traffic starved of access. Balancing these two competing classes of traffic in an optimal way is the fundamental subject of this work.

Published work in this specific area of configuring the superframe has, to date, relied on empirical, simulation-based studies of different scenarios to derive lookup tables.
of superframe configurations [5]. This work improves upon these studies with a more comprehensive and accurate simulation model, and then goes on to propose novel solutions to this problem that have sound mathematical foundations and offer a more dynamic approach. This more flexible and adaptable approach allows a continuous optimized set of superframe parameters to be derived and the more theoretical basis permits greater confidence in the optimal nature of the values being employed than is possible with purely experimental results.

This paper is structured as follows: in Section 2 the IEEE802.11 PCF is explained to give a background to this problem area. In Section 3 we examine related work in this area and highlight how this contribution differs from, and improves upon, what has gone before. Section 4 presents our simulation model that improves upon that in the literature, whilst Sections 5 and 6 describe our mathematical approaches to this problem. In Section 7 we describe a simulation prototype that allows direct comparison of all of these approaches. Finally, in Section 8, we conclude this paper.

2. IEEE802.11 CENTRALIZED CONTROL

The IEEE802.11 standard [1] was created as a wireless alternative to wired local area networks (LANs), which at that time were predominately deployed in office environments to carry internet data traffic. Nonetheless, even at that time, it was recognized that support for QoS-sensitive traffic would be required. To achieve this, two complementary access schemes were specified, the best-effort contention-based distributed coordination function (DCF) for delay-insensitive traffic, and the optional centralized polling-based point coordination function (PCF) for time-bounded traffic, such as audio/video streams and voice over internet protocol (VoIP) traffic.

DCF is the mandatory access mechanism in IEEE802.11. For sporadic bursty data traffic, this offers a very efficient means of access: devices (stations, STA, in IEEE802.11 parlance) can compete for access to the medium as soon as they have a packet to transmit. The underlying access scheme is carrier-sense multiple access with collision avoidance (CSMA/CA). Multiple access and collision avoidance are achieved with a combination of prerequisite quiet periods on the medium (hence the carrier sense) followed by random backoffs to avoid collisions. The durations of the quiet periods (termed interframe spaces) prioritize access onto the medium. For example, the shortest interframe space (short interframe space, SIFS) is used between the transmission of a packet and the transmission by the receiving station of its acknowledgment. Transmission of this acknowledgement has the highest priority of any packet (as it is the only means by which the transmitting station can be aware of successful delivery, and therefore not retransmit the original packet), so it is allowed onto the medium with the shortest possible interframe space following the end of the original packet transmission. Stations newly contending for access must wait for a much longer interframe space (the DCF interframe space, DIFS) before even being able to contend for access with the random backoff procedure.

However, for QoS-sensitive traffic where a packet must be sent at a guaranteed time, contending for access (and potentially losing) with every packet quickly becomes impossible under all but the lightest of network loads. To guarantee packet transmission, reservation and polling schemes must be considered. In these cases, the additional overhead of reserving a transmission in advance becomes acceptable. The centralized PCF of the original IEEE802.11 standard, and its progeny, the hybrid coordination function (HCF) in the IEEE802.11e standard [4], both introduce centralized coordination of resources to allow this QoS-sensitive traffic to coexist alongside contention-based data exchanges. The difference between the two is that HCF allows a more flexible allocation of transmission opportunities compared to PCF, although this is at the cost of increased complexity.

Centralized coordination imposes a time-based repeating superframe onto the medium (as illustrated in Figure 1), characterized by the transmission of a broadcast beacon, followed by a contention-free (pollled) period (CFP) and then a contention-based access period (CP). The process of overlaying this structure onto the otherwise anarchic access mechanism of DCF is possible through the aforementioned interframe spaces: the central controller is able to use the PCF interframe space (PIFS), of shorter duration than the DIFS, to preempt contending stations and seize the medium to begin the superframe.

The structure of the superframe is determined by two parameters, its duration and the proportion of time spent in the contention-free phase. This duration (i.e., the beacon and CFP repetition rate) and the relative size of the CFP to the rest of the superframe, typically termed CFPREP and CFPMAX, respectively, are both configurable by the point controller (PC) entity located at the access point (AP). These two values are broadcast in the beacon to all stations.

These parameters determine the success of a given WLAN deployment from the perspective of the polled traffic, the contention-based traffic, or both. A badly configured system will fail to deliver the performance that the end user has the right to expect, irrespective of the headline data rate of the product.

3. RELATED WORK

The distributed approach to serving QoS-sensitive traffic has been closely studied in recent years, both in the guise of the enhanced distributed channel access (EDCA) subset of the
IEEE802.11e HCF [4] and in the WiMedia MAC [3] (formed from one of the survivors of the now-defunct IEEE802.15.3a standard). The latter offers extensions to the IEEE802.11e EDCA subset including a fully distributed solution including both hard and soft reservations of slots (soft reservation being the ability for a station to tentatively reserve a slot, and for it to be made available for other stations if unused). The performance of the WiMedia MAC has been evaluated, and the soft-reservation scheme is found to be particularly efficient [6]. A number of extensions and enhancements to these distributed schemes have been proposed from a number of different perspectives, the sheer number of which suggesting that there are several shortcomings to this approach. These extensions have included the use of admission control [7] by the higher layers, the addition of hybrid automatic repeat request (ARQ) mechanisms [8] and variable backoffs (contention windows) [9, 10] to the MAC protocol, and cross-layer schemes linking the differentiated access categories to the modulation and coding schemes of the physical layer [11].

The centralized approach has been less well studied, often because a distributed solution is viewed as being inherently more scalable and less complex [12]. However, under heavy and asymmetric loads such as would result from streaming high-definition television and similar demanding applications, it has been observed that the distributed approach results in a severe impact on the coexisting traffic streams [13, 14]. The complexity of the 802.11e HCF scheme has been highlighted as an issue, and an enhanced PCF (EPCF) has been proposed [15] to address some issues with PCF that HCF also addresses, whilst not imposing all of the complexity of HCF.

A self-adaptive scheme to configure the PCF superframe has been proposed [5]. This proposed scheme selects parameters from predefined lookup tables indexed by a quantized number of active polled stations and stepped values for the maximum allowable delay of the applications. The values populating the lookup tables are derived through experimental simulation results, which result in values of an almost random nature, as depicted in Figure 2.

These results do not take into account the minimum CFP and CP sizes mandated by the standard [1, 16], and crucially, there is no means of generating values outside of the simulation scenarios considered. Nonetheless, these values provide a valuable benchmark for the approaches considered herein. The traffic considered in this benchmark study is a combination of data and VoIP flows, an important area for investigation as internet telephony applications continue to gain popularity.

4. IMPROVED EMPIRICAL RESULTS

An improved (standard-compliant) simulation model, using the configuration proposed in [5], has been developed in OPNET.

The network model is constrained to 16 STAs and an AP throughout the study presented herein, with all stations located within a 300 m diameter. All 16 STAs produce voice traffic but only 6 of them produce data traffic. The PC function is performed in the AP which is the destination for all transmissions. The AP transmits only MAC control and management frames, such as ACKs, polls, and beacons.

An STA based on a generic node model (Figure 3) generates voice and potentially data application traffic along with the necessary MAC control frames. The different traffic streams are buffered in individual queues until the frames are transmitted. The data queue is served during the CP and the voice queue is served during the CFP. The interval between successive data MAC service data unit (MSDU) generations varies exponentially with a mean of 7.5 frames per second (fps). The data MSDUs vary exponentially in size with a mean of 1000 bytes. Brady’s model [17] is employed for the voice traffic generator, which produces 200-byte MSDUs. To prevent idle CFPs and sudden traffic surges, the start times of the voice generators are random over the first two seconds of the simulation.

The AP model, which is based on the generic node model, controls the CFP with the transmission of beacons, polls, and CFP end (CF-END) frames using the PC function.
The AP responds to received data MPDUs with acknowledgements (ACKs) during the CPs. The QoS performance is also measured in the AP model as it provides sinks for the two types of traffic. The polling list, which consists of all 16 STAs, is cycled through continuously during the CFP. When a voice MPDU has been received in response to a poll frame, the AP acknowledges its reception in the proceeding poll frame by setting the frame type field to be a combined poll and acknowledgment. If a node does not have any voice packets queued when polled, it responds with a null data frame. At the beginning of a CFP, the polling is resumed where the previous CFP ended. If sufficient time remains in a CFP after all nodes have been polled, the polling cycle begins again. Intelligent polling schemes, such as biasing the polling to nodes that did not previously respond [18–20], are not utilized in this study. A check is made to ensure that sufficient time remains in the CFP to accommodate a polled voice frame exchange (i.e., poll + voice MSDU + 2SIFS + CF-END) prior to every poll transmission. An early CF-END is transmitted if insufficient time remains.

No check is made during the CP to ensure that the DCF access mechanism frame exchange sequence (DIFS + CW + data + SIFS + ACK) will be complete before the next expected beacon transmission. This will occasionally result in CP stretching which will shorten the duration of the proceeding CFP.

An IEEE802.11b physical layer (PHY) is assumed as this provides a fair comparison with the referenced work in this area. The fundamental behavior of a MAC is largely independent of the PHY technology, and when performing comparisons between different MAC solutions, the specifics of the PHY are not particularly relevant. The physical layer is modeled so that packet losses due to link errors do not occur. Packet losses occur due to collisions only, and so observations on the performance can be described purely in terms of MAC behavior. It is also assumed that there are no hidden stations, the capture effect does not occur, and none of the stations are in power-saving mode. The model parameters are summarized in Table 1.

| Parameter       | Value   | Parameter       | Value   |
|-----------------|---------|-----------------|---------|
| Slot            | 20 μs   | Mean data MSDU | 1 kbyte |
| SIFS            | 10 μs   | Mean data rate | 7.5 fps |
| PIFS            | 30 μs   | Voice MSDU     | 200 bytes |
| DIFS            | 50 μs   | Voice mean on : off | 1 s : 1.35 s |
| CWMIN           | 31 slots| Voice on rate  | 64 kbps |
| PLCP time       | 192 μs  | Beacon         | 160 bytes |
| MAC header      | 28 bytes| ACK            | 14 bytes |
| Data rate       | 2 Mbps  | Poll\CF_end    | 20 bytes |
| Control rate    | 1 Mbps  | Queue sizes    | 250 Kbits |

Table 1: Summary of model parameters.

| Parameter | Value |
|-----------|-------|
| CFP MAX   | 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 |
| CFP REF   | 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250 |

Table 2: Simulated CFP MAX and CFP REF values.

Generally, the voice traffic has the more stringent performance requirements of the two traffic types. Therefore, the performance of the CFP, and that of the associated voice traffic, is focused on in the presentation of the results. Failure to satisfy these requirements results in wasted transmissions as packets received outside of the QoS constraints will probably be dropped at the transport or application layer. The approach taken is to determine how to configure the system so that the time-dependent voice traffic is satisfied whilst ensuring that the maximum possible amount of medium time remains for data traffic.

Simulations have been performed for all permutations of the CFP MAX and CFP REF settings contained in Table 2. This provided 399 simulations each covering 5 minutes of simulated time. However, some of the CFP MAX and CFP REF combinations will result in CP and CFP durations that are less than the minimum mandated by the standard. These invalid permutations can be discounted at a later stage.

The first set of simulation results is the mean voice traffic throughputs, which are illustrated in Figure 4. Sixteen STAs produce approximately 435 kbps of voice traffic within the network. The voice traffic throughput results show that the CFP_MAX value has to be around 45% and above so that all of the voice traffic generated can be accommodated.

It is not sufficient to concentrate solely on providing the necessary resource to accommodate all of the voice traffic to produce a successful system. The delay that is experienced is arguably more important for time-dependent voice services. The mean delays experienced by the voice traffic during the
simulations are illustrated in Figure 5. CFPMAX values below 60% incur significant delays so only a subset of the CFPMAX results is included. The CFPMAX value of 45% suggested by the mean voice throughput results will result in voice delays in excess of three seconds, which is unacceptable for telephony services. Voice transmission requires delays below 25 milliseconds if echo cancellation is not available, 150 milliseconds for high quality with echo cancellation, and 400 milliseconds for acceptable quality with echo cancellation [21]. The results show that CFPMAX values in the region of 70% and above are required to achieve mean delays below 150 milliseconds.

The mean voice delay results can generate a lookup table to select CFPMAX and CFPREP values that result in a given delay. They can also predict the performance of a particular superframe configuration generated by an optimization algorithm. This allows different optimization techniques to be compared. The most interesting observation of Figure 5 is the apparent immunity to CFPREP variations that the near horizontal contours suggest.

Despite having similar mean delays, the probability density functions (PDFs) of instantaneous voice packet delays for given CFPREP values are quite different. Figure 6 illustrates the distribution of delays that were experienced for a subset of the CFPREP values with a constant CFPMAX of 70%. This value of CFPMAX provides mean delays in the region of 150 milliseconds. The distributions contain two peaks, the first occurring at \((\text{nodes}/2) \times \text{polling-exchange duration}\) and the second occurring at \(\text{CFP REP} \times (1 - \text{CFPMAX})\). The former occurs due to the average wait experienced during a polling period, equal to half the time to poll all twelve stations and the latter due to packets having to wait for a CP to pass.

Figure 7 presents the cumulative distribution functions (CDFs) of voice packet delays, and shows the percentage that satisfies a given delay constraint. A CDF is required if the maximum instantaneous delay is the important performance parameter. The CDF can predict the percentage of frames that may be dropped due to the delay constraints not being met. For a 400-millisecond instantaneous delay threshold, a CFPMAX setting of 70% requires a CFPREP in the region of 170 milliseconds and above. This will provide a voice service of acceptable quality only if echo cancellation is included [21, 22]. The CDFs illustrate that delay distributions can be highly CFPREP sensitive in certain regions. Figure 7 shows that the percentage of packets within the constraint of 100-millisecond maximum delay varies from 65 to 85 depending on the CFPREP setting.

Focusing on the CFP and its associated voice traffic prevents valuable medium time from being wasted. However, it is also important to understand the effect of superframe configuration on the CP and the associated data traffic. Biasing resource allocation to the voice traffic is only sensible to the point where the voice services have their QoS constraints satisfied. Further biasing in the direction of voice traffic provides no noticeable improvements in the performance of voice services but it results in a noticeable degradation of the data services.

The data traffic throughput results, illustrated in Figure 8, show that values of CFPMAX below approximately 80% are required to support all of the data traffic (360 Kbps) generated in the given scenario. The CDF of instantaneous voice traffic delays, Figure 7, has demonstrated that for 70% CFPMAX, a minimum CFPREP of 170 milliseconds is required for acceptable voice transmission. This superframe configuration provides sufficient CP capacity to fully accommodate the generated data traffic. Higher-quality voice transmissions demanding delays in the region of 150 milliseconds will require the superframe configuration to be biased further in favor of the CFP. CFPMAX values in excess of 80% will reduce the amount of data traffic that can be supported. Reducing the proportion of medium time available for the CP increases the likelihood of CP stretching as there is a greater probability that data packets will be awaiting transmission at the end of the CP. This CP stretching will have a negative impact on the CFP albeit smaller than the positive impact of increasing the amount of resource allocated to the CFP.

5. NONLINEAR OPTIMIZATION

The first mathematical technique we propose as a candidate solution as a verifiable theoretical model is that of nonlinear optimization of an abstracted model of the data exchanges on the superframe [23]. Nonlinear optimization theory provides a number of means to optimize a number of variable parameters to provide a stable system solution. These techniques have been applied to a number of areas within communications, including wireless sensor network access [24] and deriving training sequences for orthogonal-frequency division multiplexing (OFDM) systems [25]. We use the barrier method [26] in this work.

No matter how robust the mathematical analysis technique adopted, its success is, of course, dependent on how closely the model being analyzed resembles reality. In the case of nonlinear optimization, this means that the formation of the objective and constraint functions is crucial. Our approach is to maximize the utilization of the contention-free and contending phases simultaneously within a number of
constraints, such that the two phases’ utilizations are traded-off against each other. Therefore, expressions for these two phases must be carefully developed to represent the efficiency of the resource allocation in each phase, such that the resulting objective function can determine how far from the ideal each component is.

Before the model is developed, as with the preceding simulation study, we make assumptions of a reliable physical layer channel (no link errors, no collisions), and exclude hidden terminals, the capture-effect, and the power-saving mechanism, and assume that all stations are fully backlogged (i.e., they always have data to send).

Each phase is affected by two inefficiency components. The first is the efficiency of an individual exchange (which scales linearly with the number of exchanges) and the second is the efficiency of the whole phase, taking into account any unused airtime at the end of the phase.

Firstly, consider the QoS-sensitive polled traffic in the CFP (as illustrated in Figure 9). In the case of the first component, due to the assumption that all stations are fully backlogged, no poll is wasted, so each packet polled from station incurs an overhead comprising just the interframe spaces between contention-free packets (SIFS):

\[ C_a = 2 \times \text{SIFS} \tag{1} \]

The second component is the wastage at the end of the CFP if it is configured to any size not divisible exactly by the frame exchange duration (although note that, in practice, the central controller can terminate the CFP early and make this “wasted” period available to the CP). The overall efficiency for the CFP can be calculated as

\[ V(N_p) = \left( 1 - \frac{N_p (C_b - C_a)}{xy} \right), \tag{2} \]

where \( C_b \) is the entire polled exchange duration (ms) and \( C_a \) is the polled exchange overhead from (1), and \( x \) and \( y \) are CFPMAX and CFPREP, respectively. These parameters are tabulated for convenience in Table 3.

Figure 6: PDFs of voice packet delays at 70% CFPMAX.
The number of polled terminals, $N_p$, is a parameter that the AP can reasonably be expected to know as all stations must associate with the AP if polling service is required.

For the CP (as illustrated in Figure 10), recall the operation of the DCF. Stations must wait for the DIFS period of silence on the medium (with the 802.11b physical layer, this is 50 μs). If this period has elapsed without any activity on the medium, the station then performs a random backoff for a random number of slots (each of 20-microsecond duration in 11b) drawn from the range $[0, CW]$, where $CW$ (contention window) begins at 31 (11b again) and can increase as a binary exponential up to the limit 1023.

If the station detects a transmission during the contention window before its backoff has finished, then the station has lost this particular contention to another station (which happened to choose a smaller backoff this time around), and it must suspend the countdown, and resume it on a later attempt. If a station gains access but experiences a collision on transmission, it will increase the size of $CW$ for the next attempt. However, the “no collisions” assumption can be used to simplify this mechanism by freezing $CW$ at its smallest value of 31, and taking the mean $CW$ value of 15.5 for every contention. If every contention is assumed to win without any other terminal transmitting during the $CW$ phase (although in reality the probability of seeing another terminal transmit is going to increase with the number of terminals present), then a single DIFS per contention can be assumed.
This gives the first efficiency component of the CP as

\[ M_S = \text{DIFS} + \text{backoff} + \text{SIFS} + \text{ACK frame}. \quad (3) \]

The second efficiency component (wastage at the end of the phase) can be determined from the effective number of contending stations. This in turn depends on the traffic level and the total number of contending stations, \( N_c \). If we know the approximate packet rate of this traffic, \( P_s \), the effective number of concurrently sending stations will be \( y \times P_s \times N_c \). Hence, the overall efficiency of the CP simplifies to

\[ L(N_c) = \frac{P_s \times N_c (M_s - H_s)}{(1 - x)}, \quad (4) \]

where \( H_s \) is the entire standardized contended exchange duration (ms), \( M_s \) is the standardized contended exchange overhead (ns) from (3), \( N_c \) is the number of contending stations, \( y \) is \( \text{CFP}_{\text{REP}} \), and \( x \) is \( \text{CFP}_{\text{MAX}} \). We must further constrain this expression by the frame-generation rate of the traffic, otherwise this becomes almost a “self-optimizing” model that will always fill the CP to capacity. We can use the utilization functions \( L \) and \( V \) in the following objective function:

\[ f_0(x, y) = (1 - L(N_c))^2 + (V(N_p))^2. \quad (5) \]

We use the \( 1 - L(N_c) \) term since higher values of \( L \) correspond to good performance (in contrast to high values of \( V \), which indicate poorer performance), and square both terms to ensure that both are positive and continuously differentiable over the whole domain of interest. Substituting the expressions for \( V \) and \( L \) given in (2) and (4), respectively, and simplifying gives

\[ f_0(x, y) = \left(1 - \frac{P_s N_c (M_s - H_s)}{1 - x} \right)^2 + \left(1 - \frac{N_p (C_b - C_a)}{x y} \right)^2. \quad (6) \]

A number of constraints on this solution can be identified. \( \text{CFP}_{\text{MAX}} \) is a ratio of two time periods, so it must be positive and less than one. \( \text{CFP}_{\text{REP}} \) is bounded by the worst-case polling frequency (“delay,” \( D \)) specified by the application. Additionally, both the CFP and CP are subject to minimum duration constraints (“\( \text{CFP}_{\text{MIN}} \)” and “\( \text{CP}_{\text{MIN}} \)” respectively) according to the standard [1]. The CFP has to be at least big enough to contain one polled exchange comprising the largest payload possible in each direction, plus a beacon and a CF-end. The CP has to be large enough to contain an acknowledged exchange of the largest payload possible.

Mathematically, the problem reduces to an optimization problem over two variables, \( x \) and \( y \): minimize \( f_0(x, y) \) from (6), subject to the set of constraints:

\[ \begin{align*}
\text{CFP}_{\text{MIN}} - x y & \leq 0, \\
\text{CP}_{\text{MIN}} - (1 - x) y & \leq 0, \\
0 & \leq x \leq 1, \\
0 & \leq y \leq D.
\end{align*} \quad (7) \]

5.1. Nonlinear vector optimization of model

Before standard optimization techniques can be unleashed on the problem, the objective function must be first reformulated in vector form with a single variable. Let \( z = (x, y)^T \), and define the two unit vectors \( e_1 = (1, 0)^T \) and \( e_2 = (0, 1)^T \). We can then rewrite the objective function as

\[ f_0(z) = \left(1 - \frac{\alpha}{1 - e_1^T z} \right)^2 + \left(1 - \frac{\beta}{z^T E z} \right)^2. \quad (8) \]

Here \( \alpha = P_s N_c (M_s - H_s), \beta = N_p (C_b - C_a), \) and \( E = e_1 e_1^T \). Other parameters are defined in Table 3, along with the values used in the application of this model. The constants are determined by the physical layer under consideration and the characteristics of the traffic flows.

In vector notation, the constraints can be restated as follows:

(i) \( \text{CFP}_{\text{MIN}} - z^T E z \leq 0 \); first constraint;
(ii) \( \text{CFP}_{\text{MIN}} - e_1^T z + z^T E z \leq 0 \); second constraint;
(iii) \( e_1^T z - 1 \leq 0 \); third constraint, upper bound;
(iv) \( -e_1^T z \leq 0 \); third constraint, lower bound;
(v) \( e_2^T z - D \leq 0 \); forth constraint, upper bound;
(vi) \( -e_2^T z \leq 0 \); forth constraint, lower bound.

Before the barrier method [26] can be used to solve this problem, there is one more hurdle to overcome. This objective function is not convex, and furthermore may have multiple solutions (local minima). Two of these minima may occur at the extreme values of the feasible set, with a third local minimum from the objective function. Feasible starting points must be determined to guide the solution in the right direction. By examining the inequality constraints of the original problem, it is possible to find feasible starting points \( x_0 \) and \( y_0 \) that can be used to initialize the barrier method. Consider the following two inequalities:

\[ \begin{align*}
\text{CFP}_{\text{MIN}} & \leq x y, \\
\text{CP}_{\text{MIN}} & \leq (1 - x) y = y - x y.
\end{align*} \quad (9) \]

These are obtained by rearranging the first two inequalities of the original problem statement. Solving the second inequality for \( x y \) enables the composite inequality to be written as \( \text{CFP}_{\text{MIN}} \leq x y \leq y - \text{CP}_{\text{MIN}} \).

Thus, for a given \( y = y_0 \), a feasible \( x = x_0 \) can be taken from the interval

\[ x_0 \in \left( \frac{\text{CFP}_{\text{MIN}}}{y}, 1 - \frac{\text{CP}_{\text{MIN}}}{y} \right), \quad (10) \]

and the following feasible starting point constraint must be met:

\[ \text{CFP}_{\text{MIN}} > y_0 - \text{CP}_{\text{MIN}}. \quad (11) \]

5.2. Application of model

The assumptions and parameters used in [5] and the simulation model in Section 4 can be adopted by this model to give
some concrete values. These parameters are given in Tables 1, and 3 gives the resulting concrete values for the constants in the model. The starting point constraint in (11) can be met for these values when, for example, $\text{CFP}_{\text{MIN}} = 39.922$, CP$_{\text{MIN}} = 21.404$, and $y_0 = 48$.

Three local minima were discovered using the following set of initial $x$ values:

1. $1.2(CFP_{\text{MIN}})/y$;
2. $0.5(1 - CP_{\text{MIN}} - CFP_{\text{MIN}})/y$;
3. $0.8(1 - (CP_{\text{MIN}} - CFP_{\text{MIN}}))/y$.

The first of these is a point near the lower end of the feasible set, the second a point in the middle, and the third a point towards the top end of the feasible set for $x$. For many values of $D$ and $N_p$, all of these local minima were found to be identical, indicating that the local minimum is a global minimum. In the case where a number of local minima were found, the objective function was evaluated at each one and the minimum. In the case where a number of local minima were discovered using the following values:

$$
\phi(s) = \int_0^\infty e^{-st}dF(t). (12)
$$

These CDFs (and corresponding LSTs) are used to capture the distributions of service times and waiting times. A central result [28] in queueing theory analysis for a queue with exponential arrival times (mean rate $\lambda$) and general service time distribution (with LST $\eta(s)$ and mean $\tau$) is that the LST of the waiting time is given by

$$
w(s) = \frac{s(1 - \lambda\tau)}{s - \lambda(1 - \eta(s)). (13)}
$$

The optimum values of $\text{CFP}_{\text{MAX}}$ are fairly variable, especially for larger values of $D$ and the smaller values of $N_p$. This variability seems to occur mainly when the objective is most flat: in that it does not vary much over a wide range of $\text{CFP}_{\text{MAX}}$ values. This means that the instability happens in exactly the situations where choosing a precise value of $\text{CFP}_{\text{MAX}}$ is least important. The $\text{CFP}_{\text{REF}}$ optima tend to be close to the maximum $D$, especially for smaller $D$ where the constraints do not permit much variation anyway. For larger $D$, the optimum values are significantly smaller than $D$, this is in line with the fact that there is much more potential to fit the polled and contention periods within a smaller repetition time.

6. Queuing Theory Approach

Queuing theory models can be used to analyze the performance of many aspects of wireless networks. Here we apply this approach to the polling phase of the PCF procedure. In these models, the system is thought of as a queue which is filled with packets by an arrival process and is emptied by a serving process. In this application, the arrival process is the voice packet generation system, and the serving process is the polling mechanism as implemented by the AP. Queuing models aim to provide information about the distributions of the time spent in the queue (the waiting time) and queue length distributions. The waiting time depends on the mixture of arrival time distribution and service time distribution. The arrival time model for this application is a simple Poisson process when the voice stream is in “on” mode; we assume here that the switch from “on” to “off” occurs sufficiently infrequently to not influence the waiting time distribution. The service time distribution is dependent on the exact polling process used by the AP.

A specific use of this technique to packet delay of polled protocols can be found in [27]. The technique of Laplace-Stieltjes transforms (LST) allows the treatment of the service time distributions to be as general as possible and provides more detailed information about the full distribution of the waiting times. We present the analysis in this form here primarily for the first reason, since we do not use information beyond the mean waiting time explicitly in this paper. The service time distribution is given either as a cumulative distribution function (CDF), or its derivative, the probability density function (PDF). The LST of a CDF of a random variable $F(t)$ is given by

$$
\phi(s) = \int_0^\infty e^{-st}dF(t). (12)
$$

These CDFs (and corresponding LSTs) are used to capture the distributions of service times and waiting times. A central result [28] in queueing theory analysis for a queue with exponential arrival times (mean rate $\lambda$) and general service time distribution (with LST $\eta(s)$ and mean $\tau$) is that the LST of the waiting time is given by

$$
w(s) = \frac{s(1 - \lambda\tau)}{s - \lambda(1 - \eta(s)). (13)}
$$
This is known as the Pollaczek-Khintchine (PK) formula. Inverting the corresponding LST to get back to the more useful PDF of the waiting times is often intractable. However, we can readily extract the set of moments ($M_n$) of the PDF distribution using the following formula:

$$M_n(F) = (-1)^n \left( \frac{d^n}{ds^n} \phi(s) \right)_{s=0}. \quad (14)$$

All the properties of a distribution can be deduced from its full set of moments, but this may require computation of a large number of them. The mean ($\mu$) and variance ($\sigma^2$) can be calculated directly from just the first and second moments:

$$\mu = M_1(F),$$
$$\sigma^2 = M_2(F) - \mu^2. \quad (15)$$

### 6.1. Application to PCF delay model

This theory can be applied to analyze the delay times of the polling procedure in 802.11 PCF. The polling procedure that the AP runs flips between two states, polling and contending. We make two assumptions in this model.

1. Service times of the polling mechanism are independent.
2. The time to poll and receive responses from the complete set of stations is constant.

The first is not strictly the case here since there is a deterministic switch between polling and CP modes. This means that the short delay that occurs in polling mode is very likely to be followed by an equally short delay, and similarly longer delays will tend to follow longer delays when the system is in CP mode. In practice, this assumption should only restrict the range of parameters over which the results are valid, since the deterministic process is likely to be more stable in the face of configurations that would otherwise cause the polling mechanism to break down with unacceptably large delays.

The second is an approximation since if a station has a packet, its response will take longer than if it is returning an empty queue. In the model, we approximate such a delay by looking at the expected number of stations that has packets and combining it with the with-packet and without-packet polling times, building a weighted average for the polling time, which we denote by $r$. This constant rate assumption will have greatest effect on large superframe configurations since the variation in total time to poll will be the largest across the whole frame in these configurations.

Next we construct a CDF for the service time for the polling traffic. Each station gets polled a total number $n_{\text{poll}} = \lfloor xy/r \rfloor$ of times each superframe. As in the previous section, we use $x$ to denote CFP$_{\text{MAX}}$ and $y$ to denote CFP$_{\text{REP}}$. In each of these occasions, the service time is $r$. In the following time slot, the CFP ends and the service time is equal to the length of the CP, $y(1 - x)$. So the service time has value $r$ with probability $n_{\text{poll}}/(n_{\text{poll}} + 1)$, and value $y(1 - x)$ with probability $1/(n_{\text{poll}} + 1)$. This translates to a PDF for the service times of

$$\operatorname{ServPDF}(t) = \frac{\delta(t - r) n_{\text{poll}}}{n_{\text{poll}} + 1} + \frac{\delta(t - y(1 - x))}{n_{\text{poll}} + 1}. \quad (16)$$

Here we use $\delta(t)$ the Dirac delta function to represent in the PDF what would be discontinuities in the CDF. The required CDF is given by the integral of this function. This service time has corresponding LST given by

$$\operatorname{LST}_{\text{Serv}}(s) = e^{rs} - \frac{\lambda}{\lambda - (1 - r)} e^{y(x - 1)} - \frac{\lambda}{\lambda - (1 - r)} n_{\text{poll}} + 1. \quad (17)$$

We insert this in the PK formula (13), assuming that the voice source is in talk-spurt mode with a Poisson arrival rate of packets with mean $\lambda$. If we compute the first moment using (14), we obtain the following formula for the mean packet delay:

$$D_{\text{(x,y,r)}} = \frac{\lambda n_{\text{poll}} r^2 (x - 1)^2 y^2}{2(\lambda (x - 1) y + (1 - \lambda r) n_{\text{poll}} + 1)}. \quad (18)$$

Once suitable values of $\lambda$ and $r$ are set from the scenario parameters, the mean packet delay can be computed. Figure 13 shows the mean delay predicted by this method compared to the mean delay observed from OPNET simulation. Here we fix CFP$_{\text{REP}}$ to be 120 milliseconds and show the delays for a range of CFP$_{\text{MAX}}$ values from 50% to 90%. For 12 voice stations, there is very close agreement between the model and the simulations. For larger numbers of stations, there is more discrepancy for smaller values of CFP$_{\text{MAX}}$, but this is where both model and simulation tend to break down anyway due to high packet delays.
Generally though, the agreement is close enough to answer the kind of questions that are required to optimize the performance of this system. For example, the value of CFP\textsubscript{MAX} that is required to bring the average delay below 50 milliseconds for 14 voice stations is approximately 0.67 for both model predictions and simulations.

### 6.2. Selection of optimal CFP\textsubscript{MAX}

This approach does not yet provide enough information to optimize for the values of CFP\textsubscript{REP}. Future work would look at the higher moments provided by the LST approach and uses these to more accurately predict the percentage of packets satisfying delay requirements. However, we can still use the mean packet delay to select CFP\textsubscript{MAX} values.

We can combine this packet delay model with a DCF throughput model derived from the one described in [29] to produce a combined performance objective function. The polled component of the objective is constructed from the required and predicted delay in the following way:

\[
\text{Obj}_{\text{polled}} = U\left(\frac{\text{delay}_{\text{req.}}}{\text{delay}_{\text{predicted}}}\right). \tag{19}
\]

Here \(U(x)\) is a utility function designed to map the region \([0, \infty)\) onto \([0, 1)\) (here we use \((Ax^k + x)/(Ax^k + x^2 + 1)\), with \(A = 18, k = 6\)). Similarly for the data traffic, with predicted and required levels of throughput,

\[
\text{Obj}_{\text{data}} = U\left(\frac{\text{throughput}_{\text{predicted}}}{\text{throughput}_{\text{req.}}}\right). \tag{20}
\]

The total objective is simply the sum of these two components; maximizing the value of this function provides optimum CFP\textsubscript{MAX} values.

Figure 14 illustrates the objectives for different numbers of voice stations and delay requirements. The curves for the

![Figure 14: Objectives for a set of different network sizes and delay requirements. The blue dotted line shows the polled traffic utility, the green gives the data utility. The black line is the combined objective; red dotted line is the optimum CFP\textsubscript{MAX}.](image)
set of 16 voice stations are only partially given since the delay model is not applicable to values of \( C_{FP\text{MAX}} \) less than 0.6 with this number of stations. Note also that with the level of data traffic described in Section 4 (with 6 data nodes), the performance for these devices does not drop off significantly until \( C_{FP\text{MAX}} \) reaches beyond 0.8. This causes the optimum \( C_{FP\text{MAX}} \) values to be driven chiefly by the polled traffic. The optima do change in an intuitive fashion, since increasing the number of voice stations and tightening the delay restriction both force a higher value of \( C_{FP\text{MAX}} \) in order to give higher priority to the voice traffic.

7. APPLICATION

So far in this paper, we have presented a number of alternative means for configuring the PCF superframe. Clearly, what is needed is a means for fairly comparing these different approaches. To achieve this, a prototype has been developed in OPNET, embodying an adaptive mechanism that reconfigures the superframe as needed. This model can be configured to use a fixed superframe structure (and thereby, replicate the results of the aforementioned simulations), but can also be configured to react to changes in traffic conditions. Additionally, the model can provide benchmark results, it can be configured to use the results from Li et al. [5], or can carry all traffic via the contention-based DCF mechanism.

7.1. Adaptive model

The key to any adaptive model is the provision of a feedback path, in this case to feed back information on the degree to which QoS requirements are being met, such as the end-to-end delays are experienced. In this model, this feedback path is abstracted as a statistic wire in OPNET, although a real implementation would, of course, require additional signaling to transfer this information.

The adaptation mechanism comprises a check as to whether the QoS requirements are being met as the superframe is started. This check is performed periodically, the period dictated by the tolerance to jitter. If the requirements have drifted, then the reconfiguration process is triggered.

The reconfiguration process supports multiple data sets. These data sets can either be a fixed lookup table (as it is the case with the Li results), or can be generated dynamically by an online optimization algorithm. This approach models the range of possible implementations that could be considered, from simple products with a limited number of lookup tables to more complex adaptive solutions.

The entire adaptation algorithm can be disabled (thereby supporting a predefined fixed superframe configuration), and the traffic streams can all be diverted to DCF to show the effect of not having a polling mechanism at all.

7.2. Scenarios

In this paper, we concentrate on two specific scenarios of interest. To allow comparison with the benchmark results, the physical layer rates are those supported by 802.11b physical layer (2 Mbps for data and 1 Mbps for control messages). The first scenario comprises an AP plus:

(i) 6 STA: local file transfer (DCF, 1500-byte mean);
(ii) 10 STA: carrying voice traffic (30-millisecond maximum delay, fixed MSDU 200 bytes, 64 kbps to correspond to a G.711-style voice CODEC, duty cycle 1 : 2.35).

The second scenario is an extension of the first. In this second scenario, an additional “glut” of QoS-sensitive traffic is initiated halfway through the simulation, in the form of an additional six voice traffic streams.

7.3. Results

The first scenario is a static scenario in terms of the traffic stream profile, as no new traffic streams begin and no existing traffic streams end in the course of the simulation. It would be expected that measurements of performance parameters will soon reach a fairly steady state, and this is borne out by results shown in Figures 15 and 16.

The results are for a DCF-only configuration (i.e., all traffic having to contend for access), a fixed superframe scheme and adaptive schemes using data from the benchmark (Li), and the results from Sections 4, 5, and 6. In the following traces, in all but the DCF case, there are separate traces for the polling and contending traffic flows, the polling results are marked by solid lines, the contending are marked by dashed lines. Specific traces of interest are highlighted in the figures.

The DCF benchmark configuration shows increasing instantaneous end-to-end delay, but offers the best receive rate of all the contention schemes as no time is spent polling. The other schemes all achieve the required end-to-end delay requirements of the voice traffic, including the fixed superframe configuration because it has an “ideal” configuration selected for this scenario (\( C_{FP\text{MAX}} \) of 85% and \( C_{FP\text{REP}} \) of 30 milliseconds). The received data rates (Figure 16) have all clustered in a similar way, with the polled traffic getting considerably greater throughput.
The benefits of PCF and of having an adaptive scheme soon become apparent when the second scenario is considered. Firstly, let us consider the disadvantages of a DCF-only system. As can be seen in Figure 18, the received data rate for DCF does not change when the system is further loaded with additional traffic, and, in Figure 17, the end-to-end delay increases.

The adaptive schemes are able to respond to the change in traffic stream demand and reconfigure to provide a nearly constant end-to-end delay for the polling traffic, sacrificing some of the end-to-end delay performance of the contending traffic, which is an acceptable and even sensible tradeoff.

A more detailed examination of the adaptive schemes reveals that the nonlinear optimization approach offers the most stable configurations, but the queuing-theory-based approach offers comparable results and has the benefit of having more potential for distributed solutions in this area. The nonlinear optimization approach does well on the polled delays, but that is at the expense of the contention traffic, which incurs a greater penalty than with the other approaches. There is the clear benefit with models that cater for all of the constraints of the IEEE802.11 specification, making any solution based on those results fully compliant with the standard.

8. CONCLUSIONS

This paper has described the IEEE802.11 centralized control schemes, concentrating on the PCF. There has been a considerable amount of research into the support of QoS-sensitive traffic in more distributed aspects of IEEE802.11, but much less investigation into centralized solutions. An existing superframe configuration solution has been described and opportunities for improvement have been identified.

A number of solutions for configuring the PCF superframe have been presented. Firstly, an improved simulation model has been used to provide an accurate set of results for any lookup-table-oriented solution. This model confers the advantage over the literature available to date of being fully compliant with the standard. This approach demonstrates the need to focus on the time-dependent services and shows the importance of considering several performance measurements.

Secondly, two mathematical models have been developed, resulting in optimized sets of values for a given configuration, and, critically, general purpose algorithms that provide optimal results for any set of model constraints.

Finally, an adaptive prototype has been presented that can show each approach in active use, highlighting the effects of changes in traffic requirements. This prototype has highlighted the consistency of the more mathematically based approaches, as well as demonstrating the benefits of both centralized control and adaptive solutions.

In terms of future work, we hope to extend this solution to the more general case of the IEEE802.11e HCF, as well as investigating the benefits (and disadvantages) of distributed methods of handling mixed traffic networks such as the distributed reservation protocol [3].
REFERENCES

[1] IEEE (Institute of Electrical Electronics Engineers), “IEEE Standard 802.11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications,” 1999.

[2] Y. Xiao, “QoS guarantee and provisioning at the contention-based wireless MAC layer in the IEEE 802.11e wireless LANs,” IEEE Wireless Communications, vol. 13, no. 1, pp. 14–21, 2006.

[3] WiMedia-Alliance (Ecma International (Ecma)), “Standard ECMA-368 High Rate Ultra Wideband PHY and MAC Standard,” 2005, http://www.wimedia.org/en/resources/eis.asp?
id=res.

[4] IEEE (Institute of Electrical Electronics Engineers), “IEEE Standard 802.11e—Part 11 Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements,” 2005.

[5] C. Li, J. Li, and X. Cai, “A study of self-adaptive transmission for integrated voice and data services over an IEEE 802.11 WLAN,” in Proceedings of 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC’04), vol. 3, pp. 1922–1926, Barcelona, Spain, September 2004.

[6] Y. Zang, G. Hiertz, J. Habetha, B. Otal, H. Sirin, and H.-J. Reumerman, “Towards high speed wireless personal area network - efficiency analysis of MBOA MAC,” in Proceedings of International Workshop on Wireless Ad-Hoc Networks (IWAN’05), pp. 10–20, London, UK, May 2005.

[7] Y. Xiao and H. Li, “Evaluation of distributed admission control for the IEEE 802.11 e EDCA,” IEEE Communications Magazine, vol. 42, no. 9, pp. S20–S24, 2004.

[8] J. Wall and J. Y. Khan, “An adaptive ARQ enhancement to support multimedia traffic using 802.11 wireless LANs,” in Proceedings of IEEE Global Telecommunications Conference (GLOBECOM’04), vol. 5, pp. 3037–3041, Dallas, Tex, USA, November 2004.

[9] F. Chatzimisios, A. C. Boucouvalas, and V. Vitsas, “IEEE 802.11 wireless LANs: performance analysis and protocol refinement,” EURASIP Journal on Wireless Communications and Networking, vol. 2005, no. 1, pp. 67–78, 2005.

[10] L. Gannoune and S. Robert, “Dynamic tuning of the contention window minimum (CWmin) for enhanced service differentiation in IEEE 802.11 wireless ad-hoc networks,” in Proceedings of 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC’04), vol. 1, pp. 311–317, Barcelona, Spain, September 2004.

[11] M. Bandinelli, F. Chifi, R. Fantacci, D. Tarchi, and G. Vannucchi, “A link adaptation strategy for QoS support in IEEE 802.11e-based WLANs,” in Proceedings of IEEE Wireless Communications and Networking Conference (WCNC’05), vol. 1, pp. 120–125, New Orleans, La, USA, March 2005.

[12] A. Iera, G. Ruggieri, and D. Tripodi, “Providing throughput guarantees in 802.11e WLAN through a dynamic priority assignment mechanism,” Wireless Personal Communications, vol. 34, no. 1-2, pp. 109–125, 2005.

[13] D. Chen, D. Gu, and J. Zhang, “Supporting real-time traffic with QoS in IEEE 802.11e based home networks,” in Proceedings of 1st IEEE Consumer Communications and Networking Conference (CCNC’04), pp. 205–209, Las Vegas, Nev, USA, January 2004.

[14] G. Smith and D. Dillon, “QOS over IEEE 802.11: the need for HCCA for video applications,” Bemai, pp. 1–13, 2004.

[15] J. N. Al-Karaki and J. M. Chang, “Quality of service support in IEEE 802.11 wireless ad hoc networks,” Ad Hoc Networks, vol. 2, no. 3, pp. 265–281, 2004.

[16] S. Mangold, S. Choi, G. R. Hiertz, O. Klein, and B. Walke, “Analysis of IEEE 802.11 e for QoS support in wireless LANs,” IEEE Wireless Communications, vol. 10, no. 6, pp. 40–50, 2003.

[17] P. Brady, “A model for generating on-off speech patterns in two-way conversation,” Bell System Technical Journal, vol. 48, pp. 2445–2472, 1969.

[18] J. Zheng and E. Regentova, “An improved polling scheme for voice support in IEEE 802.11 wireless network,” in Proceedings of International Conference on Information Technology: Coding and Computing (ITCC’05), vol. 2, pp. 603–608, Las Vegas, Nev, USA, April 2005.

[19] X. Ma, C. Du, and Z. Niu, “Adaptive polling list arrangement scheme for voice transmission with PCF in wireless LANs,” in Proceedings of Joint Conference of the 10th Asia-Pacific Conference on Communications and the 5th International Symposium on Multi-Dimensional Mobile Communications Proceedings (APCC/MDMC’04), vol. 1, pp. 21–25, Beijing, China, August-September 2004.

[20] R. Y. W. Lam, V. C. M. Leung, and H. C. B. Chan, “Polling-based protocols for packet voice transport over IEEE 802.11 wireless local area networks,” IEEE Wireless Communications, vol. 13, no. 1, pp. 22–29, 2006.

[21] ITU-T (International Telecommunications Union - Telecommunication Standardisation Sector), “Transmission systems and media: general characteristics of international telephone connections and international telephone circuits: one-way transmission time,” 1996.

[22] X. Ma, Y. Wu, Z. Niu, and T. Sato, “Performance analysis of the packetized voice transmission with PCF in an IEEE802.11 infrastructure wireless LAN,” in Proceedings of 9th Asia-Pacific Conference on Communications (APCC ’03), vol. 2, pp. 571–575, Penang, Malaysia, September 2003.

[23] R. J. Haines, T. Lewis, J. Coon, and N. Fanning, “Non-linear optimization of IEEE 802.11e super-frame configuration,” in Proceedings of 63rd IEEE Vehicular Technology Conference (VTC ’06), vol. 3, pp. 1211–1215, Melbourne, Australia, May 2006.

[24] B. Krishnamachari and F. Ordóñez, “Analysis of energy-efficient, fair routing in wireless sensor networks through non-linear optimization,” in Proceedings of 58th IEEE Vehicular Technology Conference (VTC ’03), vol. 3, pp. 2844–2848, Orlando, Fla, USA, October 2003.

[25] M. Sandell and J. Coon, “Near-optimal training sequences for MIMO OFDM systems with nullized subcarriers,” in Proceedings of IEEE Global Telecommunications Conference (GLOBECOM’05), pp. 2244–2249, St. Louis, Mo, USA, November–December 2005.

[26] S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge University Press, Cambridge, Mass, USA, 2004.

[27] L. Wang, M. Hamdi, R. Manivasakan, and D. H. K. Tsang, “Multimedia-MAC protocol: its performance analysis and applications for WDM networks,” IEEE Transactions on Communications, vol. 54, no. 3, pp. 518–531, 2006.

[28] R. B. Cooper, Introduction to Queueing Theory, Elsevier/North Holland, Amsterdam, The Netherlands, 1981.

[29] 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.