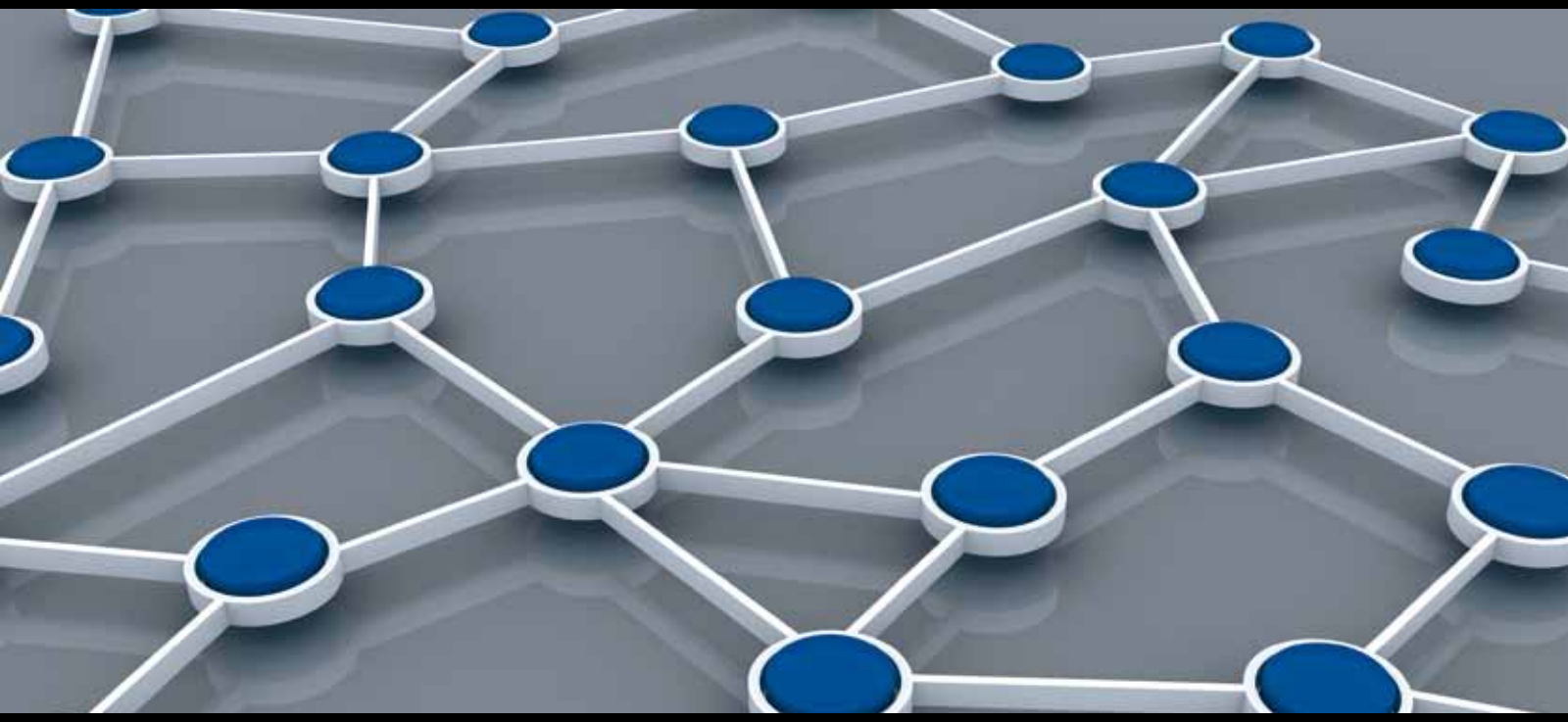
International Journal of Distributed Sensor Networks

Wireless Machine-to-Machine Networks

Guest Editors: JianHua He, Yan Zhang, Zhong Fan, Hsiao-Hwa Chen, and Lin Bai



Wireless Machine-to-Machine Networks

International Journal of Distributed Sensor Networks

Wireless Machine-to-Machine Networks

Guest Editors: Jianhua He, Yan Zhang, Zhong Fan, Hsiao-Hwa Chen, and Lin Bai

Editorial Board

Contents

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 535927, 2 pages
doi:10.1155/2012/535927

Editorial
Wireless Machine-to-Machine Networks

Jianhua He,[1] Yan Zhang,[2] Zhong Fan,[3] Hsiao-Hwa Chen,[4] and Lin Bai[5]

[1] *School of Electronics and Applied Science, Aston University, Birmingham B4 7ET, UK*
[2] *Simula Research Laboratory, Fornebu, 1325 Lysaker, Norway*
[3] *Telecommunications Research Laboratory, Toshiba Research Europe Ltd., Bristol BS1 4ND, UK*
[4] *Department of Engineering Science, National Cheng Kung University, Tainan City 70101, Taiwan*
[5] *School of Electronic and Information Engineering, Beihang University, Beijing 100191, China*

Correspondence should be addressed to Jianhua He, j.he7@aston.ac.uk

Received 30 October 2012; Accepted 30 October 2012

Machine-to-Machine (M2M) technology allows events captured by sensors in machine devices to be communicated directly with each other through wireless and/or wired systems and processed by the machine devices without human intervention. With worldwide expansion of wireless networks information can be exchanged between machine devices much easier and faster at lower cost. M2M technology opens unique opportunities to businesses as well as consumers on cost reduction and services improvement. It holds huge potential for applications in a wide range of industries.

Over the next few years there will be a huge increase in the number of machines enabled by M2M technology. The development of wireless M2M networks supporting machine devices in that scale is pivotal to the success of M2M. Sensor nodes in wireless M2M networks could be connected by a wide range of wireless network technologies including cellular networks, WPAN, RFID, Wi-Fi, Wimax, and satellite networks. However, these heterogeneous network technologies combined with the low-power machine devices and diverse QoS requirements of M2M applications present big challenges to the wireless M2M networks, including support of possibly huge number (up to billions) of M2M devices, supporting a wide range of M2M applications with significantly different traffic patterns and QoS requirements, supporting different types of sensors with low power and low mobility.

These challenges are not so often faced by general wireless sensor networks and can cause problems on almost all layers of wireless M2M networks. To ensure success of M2M applications under these challenges, many issues remain to be solved for wireless M2M networks. For example, wireless M2M networks need to consistently provide reliable communication services (in terms of effective QoS and security) for the M2M applications. Scalable, smart, and adaptive protocols and algorithms should also be designed and implemented for M2M networks as the networks are more likely to malfunction without human intervention. These protocols and algorithms should intelligently adapt to dynamic network environments characterized by a large amount of low-power M2M devices, heterogeneous networks with limited bandwidth, and many coexisting applications with diverse QoS requirements.

This special issue aims to gather researchers and engineers from different relevant areas, such as wireless sensor networks, RFID, WPAN, cellular networks, and smart grid, to present the latest research on wireless M2M networks. Thirteen papers have been accepted in this special issue. They are mainly focused on four different research areas: (1) tackling the challenge of supporting a large number of M2M devices by developing advanced wireless communication technologies to improve network capacity; (2) design and analysis of wireless channel access and resource management for M2M applications which have specific QoS requirements; (3) energy efficiency issue and scalability issue of networking/topology control; (4) scalable network architecture and flexible resource description that can be used for M2M networks.

The paper "*On the achievable user number of the downlinks in cellular-based machine-to-machine communications*"

by Y. Li et al. is concentrated on analyzing the number of M2M users that can be admitted into cellular networks with certain preconfigured QoS.

In the paper "*Distributed energy-efficient topology control algorithm in home M2M networks*" by C.-Y. Lee et al. a distributed energy-efficient topology control algorithm was proposed to enhance energy efficiency and prolong network lifetime.

The paper "*Energy-efficient chain formation algorithm for data gathering in wireless sensor networks*" by S.-J. Lim et al. investigated chain-based routing protocols for improving energy efficiency of wireless sensor networks.

The paper "*An energy-efficient and fault-tolerant convergecast protocol in wireless sensor networks*" by T. Yang et al. explored the hyper-graph theory to compute energy efficient data delivery for convergecast applications over large scale WSNs.

The paper "*Collaborative relay beamforming strategies for multiple destinations with guaranteed QoS in wireless machine-to-machine networks*" by D. Wang et al. studied the amplify-and-forward (AF) relay beamforming to guarantee the QoS requirements of M2M devices. Two iterative strategies were proposed to jointly optimize the source antenna selection and the collaborative relay beamforming weights.

In the paper "*Resource description language: a unified description language for network embedded resources*" A. C. Santos et al. proposed a resource description language that represents a uniform way of describing embedded resources, which can contribute to overcome performance issues in dense or highly dynamic M2M networks.

In the paper "*Service-oriented radio architecture: a novel M2M network architecture for cognitive radio systems*" X. Dong et al. proposed a flexible M2M network architecture, which can be used for the development of cognitive radio networks.

The paper "*A dynamic pricing scheme for congestion game in wireless machine-to-machine networks*" by Z. Mao et al. developed a distributed dynamic pricing scheme for wireless M2M networks with objective of minimizing the maximum latency over all source nodes, in which routing congestion was formulated as a dynamic game.

The paper "*Optimal routing control in disconnected machine-to-machine networks*" by Y. Wu et al. investigated incentive policies for optimal routing in M2M networks which may not be always connected.

The paper "*Adaptive message rate control of infrastructured DSRC vehicle networks for coexisting road safety and non-safety applications*" by W. Guan et al. is concentrated on the QoS provisioning in vehicle network for both road safety and non-safety applications. Adaptive rate control schemes were proposed to improve the system performances.

In the paper "*Distributed beamforming for relay assisted multiuser machine-to-machine networks*" C. Chen et al. developed distributed algorithms for relay assisted beamforming, by which each relay node individually learns its own beamforming weights with local channel state information (CSI). Two suboptimal relay beamforming schemes were proposed that only require local CSI to minimize mean square error (MSE) for all the users with non-orthogonal channels.

The paper "*Multilayer orthogonal beamforming for priority-guaranteed wireless communications*" by J. Xie et al. studied efficient spectrum sharing between M2M devices and normal mobile users over cellular networks. A multilayer orthogonal beamforming (MOBF) scheme is proposed by which frequency resource could be efficiently reused by both normal mobile users and M2M devices, and QoS of normal users is not affected by the admission of M2M devices.

The paper "*Investigation of uncoordinated coexisting IEEE 802.15.4 networks with sleep mode for machine-to-machine communications*" by C. Ma et al. investigated the impact of uncoordinated operations on the QoS performances of coexisting IEEE 802.15.4 based M2M networks.

The thirteen papers included in this special issue presented the latest research on both protocol design and theoretic analysis for wireless M2M networks, which are based on various wireless communication technologies (e.g., cellular networks, wireless sensor networks, DSRC, and IEEE 802.15.4). We hope that this special issue can help readers to get a better understanding about the breadth and depth of the current research on wireless M2M networks and boost further research and practical developments in the field of wireless M2M networks.

Acknowledgments

The Guest Editors would like to thank all the authors for their contributions to this special issue. Our special thanks go to all the reviewers for their hard work in providing timely responses and constructive comments. We also thank the journal editorial staff who provided great support to us on this special issue.

Jianhua He
Yan Zhang
Zhong Fan
Hsiao-Hwa Chen
Lin Bai

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 830371, 12 pages
doi:10.1155/2012/830371

Research Article

Challenges for Coexistence of Machine to Machine and Human to Human Applications in Mobile Network: Concept of Smart Mobility Management

Rajarshi Sanyal,[1] Ernestina Cianca,[2] and Ramjee Prasad[3]

[1] *Engineering Expert (Roaming and Signalling), Belgacom International Carrier Services, Rue Lebeau 4, 1000 Brussels, Belgium*
[2] *Department of Electronic Engineering, University of Rome Tor Vergata, via del Politecnico 1, 00133 Rome, Italy*
[3] *CTIF, Aalborg University, Niels Jernes Vej 12, 9220 Aalborg, Denmark*

Correspondence should be addressed to Rajarshi Sanyal, rajarshi.sanyal@bics.com

Received 23 April 2012; Revised 9 August 2012; Accepted 13 August 2012

Academic Editor: Jianhua He

A key factor for the evolution of the mobile networks towards 4G is to bring to fruition high bandwidth per mobile node. Eventually, due to the advent of a new class of applications, namely, Machine-to-Machine, we foresee new challenges where bandwidth per user is no more the primal driver. As an immediate impact of the high penetration of M2M devices, we envisage a surge in the signaling messages for mobility and location management. The cell size will shrivel due to high tele-density resulting in even more signaling messages related to handoff and location updates. The mobile network should be evolved to address various nuances of the mobile devices used by man and machines. The bigger question is as follows. Is the state-of-the-art mobile network designed optimally to cater both the Human-to-Human and Machine-to-Machine applications? This paper presents the primary challenges for the coexistence of M2M and H2H devices in a mobile network and draws emphasis for revisiting the mobility management aspects and congestion control in the-state-of-the-art network. Further, we set out a mobile network architecture with smart mobility management which aims to reduce the signaling interaction between the device and the network to optimise the power and bandwidth.

1. Introduction

The concept of Machine to Machine (M2M) communications fits into the new trend of devices that we see around us disseminating information though the mobile network to the remote peer or to the cloud. These objects have their own Internet protocol addresses. They are embedded in complex systems and interfaced with sensors to obtain information from their environment (e.g., food products that record the temperature along the supply chain).

Some guidelines are given for particular aspects of this M2M technology, such as ETSI's TS 102.689 [1]. But a unified approach contemplating the coexistence and seamless interoperability between the different device types is still "work in progress." This poses a risk, as the state-of-the-art mobile networks were designed keeping in mind Human to Human (H2H) communication. Our contention is that the state-of-the-art mobile network may not be resilient to the data surge likely to be caused by M2M applications. It is probable that a group of M2M devices for a particular industry application may initiate services and access the mobile network at the same time, causing network congestion and degrading the quality of other services. For example, a fleet management company may have moving carriers in different geographical locations trying to access the network and upload data at the same instance. Considering the high proliferation of these devices, an avalanche effect on signaling and data traffic can be generated, thus impacting the gross QoS (Quality of Service) offered by the network. It is imperative that the mobility management processes need to be revisited in order to cope up with the surge in the mobility management messages in the mobile network and to maintain the same QoS without being heavy on the bandwidth. Attempts should be made to

reduce the interaction with the network and to simplify the mobility and location management-related processes.

This paper addresses the technical architecture of a Smart Mobile Network Access Topology (SMNAT). We call it smart because it is

(i) lite and does not engage complex application layer processes for mobility management and thereby reduces signaling overhead and wards off congestion,

(ii) agnostic to the device type: no intricate network adaptations, scheduling, topology control for M2M applications to thwart congestion,

(iii) identifying and addressing the mobile nodes directly at the physical layer.

SMNAT covers broad implementation scenarios, mainly in the cellular network domain and is not confined to explicit setups that we witness in ad hoc or sensor networks comprising of nodes with self-organisation capabilities. Hence there are no tailor-made amendments in the framework or operational procedures for specific scenarios (like swarm networks as an example). Topology control (physical or graphical) is not implemented as we assume in this paper that the M2M devices will acquire the mobile network like any other H2H device and will not establish direct multipath/multihop communication between the wireless nodes. The purpose of SMNAT is to ensure that the coexistence of H2H and M2M devices in the same network does not impose any collateral impairment. It is designed to abridge the intricate mobility and location management processes as well as the addressing principle to render more operational efficiency and cater the needs of the M2M devices.

The rest of the paper is organised as follows. In Section 2 we dwell on the issues in addressing and location management due to the coexistence of the M2M and H2H devices in the "state-of-the-art" network. In Section 3 we describe the smart mobile network essentials. The unique addressing scheme of SMNAT implementing the physical layer, the mobility and location management processes are subsequently discussed. In Section 4, comparative analysis of the signaling processes and the overhead due to mobility management processes between SMNAT and state-of-the-art is made; Section 5 addresses the key issues with the available mobile network topologies and how they are addressed/resolved by the smart approach. Conclusions are drawn in Section 6.

2. Challenges in Mobility Management in the Homogeneous State-of-the-Art Mobile Networks for H2H and M2M Applications

Challenges faced by the present mobile network in the light of the M2M scenarios are as follows.

(1) Number of M2M devices can be fairly large compared to the H2H devices. So the length of the addressing parameters that are used today (MSISDN, IMSI, or even the IPV6 addresses) seems to be insufficient to identify the high volume of devices in the network.

(2) Power management will play an important role. For example, an M2M device meant for location-based application in a freight container will need to be sustained for weeks, if not months, may be without an external constant power source. So the battery power should be carefully conserved, and one of the effective ways to do so is to reduce the signaling interaction between the handset and the network to maintain location and presence information.

(3) Variation of QoS requirements: some real-time applications have high demands on latency and reliability (as e.g., accident sensors). While the other applications like reporting services may be contended with a lesser priority scheme but with a higher bandwidth requirement.

(4) Different traffic patterns: some M2M devices may not transmit any data for months, for example, a crash sensor. While other sensors may transmit data continuously or periodically.

The foremost challenge for 4G is to support seamless interoperability, handover and roaming for the H2H and M2M devices without compromising the "grade of service," "round trip delay," and latency. It has to meet the desired reliability for the M2M application, while supporting the high throughput demand of a wide variety of H2H multimedia services such as multimedia web browsing, video, online gaming. Various methods [2–6] for optimising bandwidth have been worked out for mobile and 802.1X networks like dynamic bandwidth allocation according to location/QoS. Some methods [7] adhere to the principle of fixed slot allocation for all cells which is not optimal in terms of bandwidth utilisation given the fact that different device types coexist in the network with different nature of bandwidth requirement. Also, the traffic pattern varies over time and location area [7]. Due to the high proliferation of the M2M devices, there may be significant impacts like network congestions, which will in turn affect the H2H services. The M2M scenarios are characterized by colossal amount of devices that interact frequently or infrequently by small amounts of data. These devices may be clustered in a small zones leading to competition amongst the network nodes. A majority of the devices can initiate connection leading to peaks in signaling and data. This penalizes the non M2M devices.

The architecture of the IMS/LTE networks [8–10] is expected to keep large use of IP and use a large spectrum of core and access technologies, stemming from the necessity to cater various types of mobile applications depending on the device type/capability. Wireless 4G networks create various challenges due to their architectural heterogeneity in terms of differences between access schemes, resource allocation techniques, and QoS requirements. To address these challenges, the existing proposals require a significant modification or new Medium Access Control (MAC).

The primary issues that may impact the state-of-the-art mobile network due to M2M communication are as follows.

Radio Network. The eNodeB in the LTE network needs to connect a large number of M2M devices. The devices will contest to use the same channel leading to collisions.

Core Network. The MME needs to attach a large number of devices. The network also needs to provide the IP and NSAPI address to those devices. A lot of devices imply sporadic and transient use of the bearer, leading to an overhead. The HSS needs to have the subscription and profile definition for the M2M devices. There can be signaling congestion towards HSS when a large number of devices try to register in the same HSS.

In M2M networks, we seldom experience congestion in the data plane. This is because devices send and receive small amounts of data. But a lot of devices transceive data simultaneously leading to congestion mainly in the Enhanced Packet Core (EPC) part between the Serving Gateway (S-GW) [10] and the Packet Gateway (P-GW) [10]. Capacity augmentations need to be done for adjuncts like MME, HSS, AAA servers, CDR processing engine.

The goal of 4G is to replace the current proliferation of core mobile networks with a single worldwide core network standard, based on IP for control, video, packet data, and voice. The IP protocol used is IPV6, which uses 128 bit addresses vis-a-vis the 32 bit addresses of IPV4 protocol. Hence the IP V6 protocol which is directly used for mobility management in 4G can offer a wider address pool for the mobile users. This is essential to accommodate all the M2M devices in the mobile network. To enhance mobility in IPv6, "micromobility" protocols (such as Hawaii, Cellular IP, and Hierarchical Mobile IPv6) have been developed for seamless handovers that is, handovers that result in minimal handover delay, minimal packet loss, and minimal loss of communication state. However, the core issues of IPV6 to be involved in mobility management still remain.

Three major issues are highlighted and the improvisations of the solutions are as below.

Issue 1: Paging Support. The base IPv6 specification does not provide any form of paging support. Hence to maintain connectivity with the backbone infrastructure, the mobile node needs to generate location updates every time it changes its point of attachment, even if it is currently in dormant or standby mode. Excessive signaling caused by mobility leads to a significant wastage of the mobile node's battery power, especially in environments with smaller cell [11]. Therefore, it is impractical to rely completely on location updates, and it is essential to define a flexible paging support in the intradomain mobility management scheme.

In order to save battery power consumption of devices, IP paging is proposed as an extension for Mobile IP [12]. Under Mobile IP paging, a device is allowed to enter a power saving idle mode when it is inactive for a period of time. During idle mode, the system approximately knows the location of the mobile node within the paging area comprised of multiple subnets. A device in idle mode does not need to register its location when moving within a paging area. It performs location update only when it changes paging areas. However, with the high penetration of the M2M devices, the size of the paging area will reduce. This will culminate in high signaling network load when the M2M device (say, implemented for a fleet management company) alters the location frequently.

Issue 2: Round Trip Delays. According to the IPV6 implementation principle for mobility management [12], the mobile node needs to send binding updates to the home agent in a given periodicity. This action implies that an authentication procedure needs to be actuated for each bind updates resulting in an increase in round trip delays.

For these reasons a new Mobile IPv6 node, called the Mobility Anchor Point (MAP), has been suggested in RFC 4140. MAP can be located at any level in a hierarchical network of routers. The MAP will limit the amount of Mobile IPv6 signaling outside the local domain. As an evolution from IPV6, Hierarchical Mobile IPv6 (or HMIPv6) proposes the implementation of the MAPs. Networks are divided into domains and subnets, with each administrative domain having a Mobility Anchor Point (MAP) at the highest level. Intradomain mobility of a mobile host is handled separately from interdomain mobility.

When the Mobile Host (MH) changes points of attachment within the same domain, the MAP of that domain is informed of the change in care of address of the MH through binding updates. Binding updates are also sent to correspondent hosts within the same domain.

This process reduces signaling traffic due to lesser binding updates. Handoff latency also decreases as "far-off home agents" and "correspondent hosts" need not be updated every time a mobile host changes point of attachment. IPV6-based mobility management ensures minimal handoff latency to achieve better QoS for real-time data conveyance.

However HMIPv6 has its drawbacks as the hierarchical, addressing model may impact the real-time applications. Due to the hierarchical nature of the protocol, each anchor point is only aware of the next anchor point down in the hierarchy. Each node stores a mapping of source destination addresses of the previous and next nodes in the hierarchical structure. These addresses are called VCoAs (Virtual Care of Address) [13]. Only the lowest anchor point in the hierarchy stores a mapping of VCoA to PCoA which is the physical care of address of the MH in the foreign environment. As the source is aware of only the VCoA of its nearest anchor point, it sets that address in the destination field header of the IP packet. At each hop, the packet is processed, depending on the source address, the new destination address (VCoA) is decided. The packet is then forwarded using the new destination address and the node's own address as the source address. This processing occurs at each hop and can effect real-time applications running at the mobile device as it creates significant delay especially if the CH (Correspondent Host) and the MH are many hops away with the MH part of a big hierarchical structure.

Issue 3: High Signaling Exchange for Mobility Management. With the evolution of the mobile networks as we find from 2G to 4G [8] we attempt to inject more symbols (data) per unit time. This has been made possible by the evolution of the modulation schemes from BPSK in case of 2G, to

64 QAM for LTE-Advanced. In the current Mobile Radio Access Network [14–16] there is a host of network processes, tightly synchronized and orchestrated by intelligent network elements, such as:

(i) handover management,

(ii) location management,

(iii) call drop off management,

(iv) interoperability and downward compatibility management,

(v) service control (like roaming control), and

(vi) feature management.

These processes involve complex signaling operations across the radio and the core network. The attempt to simplify the network in IMS [9] and LTE [10] is focused to make the core and access networks all IP. Following the evolution trail from 2G to 4G, we do not discern a significant philosophical drift towards simplification in terms of the service logic related to mobility and location management. We see a similar process for handover, frequency reuse, location updates, and cancellations in 4G [10] as compared with 3G/2G [15]. Hence the network elements of the-state-of-the-art still need to be equipped with the intelligence and processing power to handle all these complex signaling operations.

With the recent initiatives from the European Commission [1], LTE-A network will be extended to support the M2M devices. The IPV6 addressing mechanism enables a wider device reach. But there is still no evidence of significant emphasis on reducing the mobility management processes for optimised operations.

2.1. Related Work. Alot of proposals exist for optimising the access network and the nonaccess stratum (NAS) to adapt to the machine-type communications (MTC) scenario [17, 18]. Some proposals on the access layer end dwell on realisation of an adaptive RACH [19] to reduce RACH collision probability, to control network overload, and to enhance system performance. Some other solutions [17] eyeing for optimisation at access layer are time backoff classes, slotted access, and group coordination. 3GPP has a focus on adapting the network architecture and the associated parameters to be more compatible with the M2M devices. The primary goal is to provide a long-term solution for addressing mechanism and the congestion. Network improvements for MTC (NIMTC, Rel. 10) [17] and system improvements for MTC (SIMTC, Rel. 11) [18] are two mentionable developments.

In Rel 10 for, we have witnessed some improvisations namely,

(i) RAN overload control by Extended Access Barring (EAB),

(ii) APN-based prioritisation and congestion control, and

(iii) throttle control.

In Rel 11, the adaptations are

(i) avoidance of E.164 number allocation for the M2M devices,

(ii) dedicated interface for the MTC Server to ward off link congestion in the AAA and for MSISDN (E.164) less device triggering.

There has been significant work done by the research community on congestion control. These proposals [20–22] deal with scheduling algorithms in the access part and channel reservation.

In [23], a group-based mechanism was proposed. Within the group there will be a cluster head in a device area. All the signaling information passes through the cluster head towards the eNodeB (BTS). Due to aggregation mechanism, the traffic distribution is more optimised resulting in reduction of the signaling overhead.

Most of the solutions worked out so far are poised to reshape the existing mobile network in light of the M2M scenario by implementing scheduling logic, new congestion control mechanism, defining domains and cluster-heads, proposing new device addresses, and so forth. Many of these solutions attempt to reduce network congestion caused by the proliferation of the M2M devices. We attempt to shift from this paradigm by proposing SMNAT which is agnostic to the device type. We try to nip the problem in the bud by designing an access methodology which will eliminate the application layer messages itself for mobility and location management, rather than devising relentless network adaptations by standard bodies for M2M.

3. Introduction to SMNAT

With the SMNAT, we are able to perform specific layer 5 and 7 processes related to addressing and identification of a node in the RAN, right at the physical layer [24–26]. This relieves the application layer from the arduous processing in the context of the mobility management and addressing. It wards off the continuous interexchange of signaling data between the device and the network for reducing congestion.

A mobile node is provisioned with 3 network parameters, namely, the symbol coordinate, time slot, and the primary AFCRN, which are essential for channel acquisition and for processing the call or other network operations [26]. Figure 1 elaborates these parameters. The available bandwidth is divided in multiple subfrequency bands called AFCRN (Absolute Radio Frequency Channel Number). The AFCRNs are numbered in numerical order. Each AFCRN is linked to one Time Frame. Following completion of the addressing procedure, the data channel or a time slot for the traffic conveyance is identified. The access layer is involved with limited layer 7 functionalities [25]. A new entity called Coordination Processor (CP) [25] sits in between the core and access network domains. It facilitates addressing and essential signaling functionalities for mobile user node at the physical layer level. It also takes care of some layer 7 functionalities that are still required for AAA (authentication, authorization, and accounting), supplementary services, and

value-added services. The RAN comprises ad hoc repeaters, referred to as Access Points in the network area implementing TDMA/FDD multiple access.

In [26] we presented the modulation scheme, time frame structure, and the addressing principle of SMNAT. The modulation scheme is a blend of M1 PSK (for outer ring) and M2 PSK (for inner ring), where M1 < M2. Figure 1. shows the constellation diagram in the case: M1 = 8 and M2 = 4. The outer ring comprises of M1 symbols for user traffic, and the inner ring comprises of M2 symbols that will be used for addressing the mobile users. In the proposed multiple access scheme, a user is identified in the network with respect to the symbol coordinate of the outer ring in the complex plane. The constellation diagram in Figure 1 pertains to a specific time slot, $T1$, of the frame. The frame corresponds to a specific AFCRN. Symbols of the outer ring are conveyed at a rate of fr and the symbols of the inner ring are exchanged with a much higher rate denoted by fd. In particular, we could set fd = 8 ∗ fr. When the Mobile Station (MS) is provisioned in the network, it is allocated a specific AFCRN, a specific time slot in the TDMA frame, and a specific symbol coordinate pertaining to the M1-PSK. After the addressing part is completed by actuating a layer 1 handshake, the time slot is used for data transfer. There is no separate time slot for conveying the control signaling and the data part. A single-time slot can carry M1 symbols, used for addressing M1 users. Later when the time slot is seized, it will be used to convey the symbols for data traffic.

The network based on SMNAT consists of Access Points (APs) which transceive the data to the local multiplexer. Multiple local multiplexers converge the data towards an aggregate multiplexer, which is in turn connects to a central processor called Coordination Processor (CP) [25, 26]. The CP can be compared with the time/space switching matrix 2G/3G network, MME of LTE [10]. It bridges the access and the core network and is responsible for formulating the time frames actuating physical layer addressing in the forward and reverse channels. Figure 2, demonstrates the end-to-end handshake process for call initiation by A party and termination to B party. All the associated network elements in the call trajectory and the actions taken by each of them in the process are also furnished in Figure 2.

When a mobile node A calls mobile node B, the mobile node A initially scans for the time slot in the time frame pertaining to primary AFCRN to check whether it is free to carry any symbol for a new operation, or is already in use. If it finds that the time slot is free, then it injects a symbol for itself and sends it in the uplink channel. The mobile station includes the specific symbol value in the given time slot and forms a frame. It synchronises with the access point and synchronises with the frame before populating the symbol in the time slot. The frames pertaining to all the AFCRNS are aggregated by the Central Mux and finally fed to the Coordination Processor. So the layer 1 message with the specific symbol for node A reaches the Coordination Processor. The Coordination Processor injects the same symbol in the same time slot and AFCRN in the reverse direction (downlink), which acts as a layer 1 page response. If the AFCRN is busy and not available, the response message is not sent back. The mobile node has a timeout and retries the same process (i.e., invoke a page message), but with another AFCRN.

After the Coordination Processor generates the response in the downlink and the Node A receives the page response, subsequently the time slot between the Node A and the Coordination Processor is seized for layer 7 signaling data and traffic data transport. Subsequently, a layer 7 message is transmitted by the mobile node 1 via this "seized" time slot containing the A and B number plus the information elements with the details of supplementary services of node A and the authentication parameters. Once the CP obtains the B number, it queries the HSS/HLR over DIAMETER (S6a)/GSM MAP interface mainly for 3 objectives:

(i) authentication and authorisation of mobile node A,

(ii) to resolve the essential network parameters provisioned for B, namely, the symbol coordinated, time slot, and primary AFCRN in order to invoke the downlink message,

(iii) To check the available supplementary services/teleservices for mobile node B, only in case mobile node B is provisioned in the same network. This can be analysed by the Coordination Processor by examining the dialing plan.

After the signaling exchange with the HSS/HLR, the Coordination Processor generates a layer 1 message to B node. If the primary AFCRN provisioned for Node B has been seized before by another user, the Coordination Processor tries with another AFCRN according to the AFCRN scanning procedure. Node B listens to the time slots in all the relevant AFCRNs, whether a symbol (for itself) has arrived. When it finds one, it generates a layer 1 message with the same symbol and in the time slot towards the CP. After the CP receives the information A, it comprehends that it is a response message from the B party. It distinguishes between a response message and a new call request, as it maintains the call context immediately after generating a page signal in the downlink and starts a timer to await exactly the mirror message as generated in a different time frame. After this process is completed, it is acknowledged both by Node B and the CP the time slot is available for data transfer between the two entities and is now sampled at a greater rate by sub dividing in more time slots. At this stage, a layer 7 message is invoked by the CP towards the Node B to let it know all the details of the call, like A party number, B party number, supplementary services, and so forth.

Note that the procedure explained above is actuated only when the mobile node or the network commences a network operation, like mobile originating/terminating calls, data service, SMS, or other value-added services. However, the mobile node does not get attached to the network and hence there is no periodic location updates, location update during network acquisition, or location update due to change in cell. The network does not keep a track of the location or presence of the mobile node in the network area. This helps in substantial reduction of signaling messages due

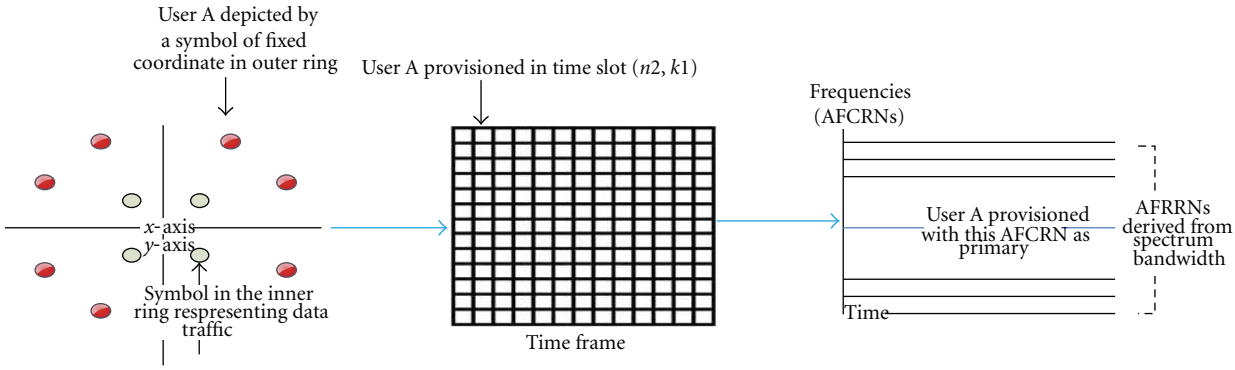


Figure 1: Parameters at the physical plane allocated for the mobile users.

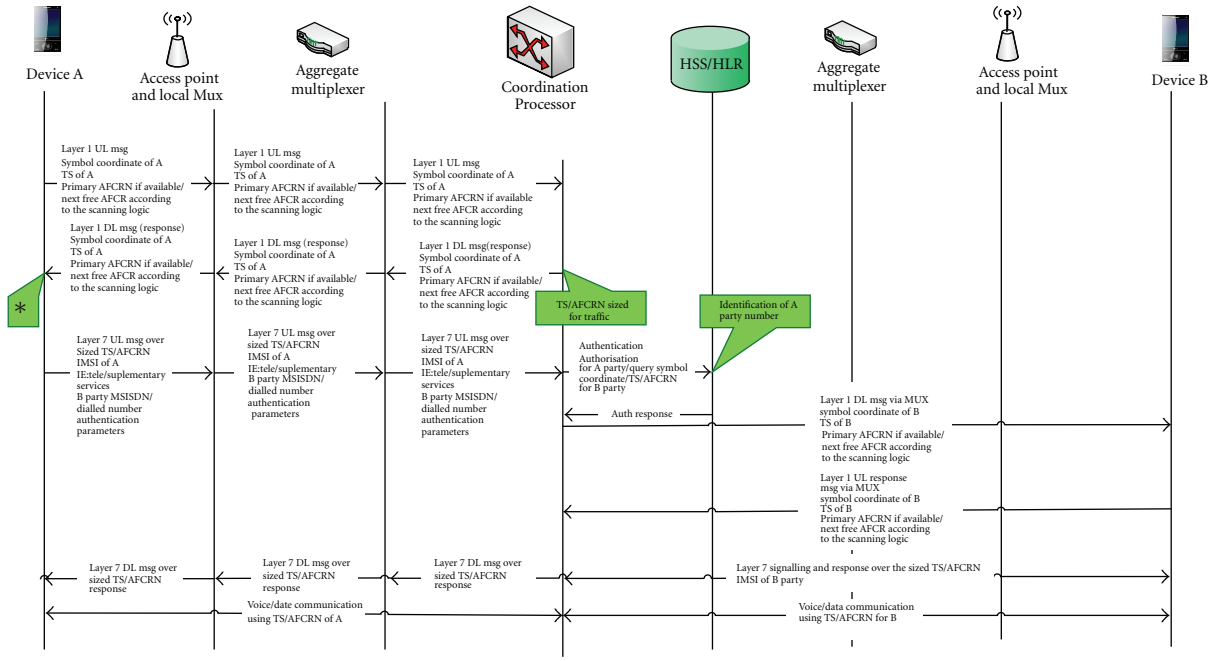


Figure 2: End-to-end call flow between A and B parties.

to mobility management. Section 3.1 covers this topic in details.

3.1. Mobility and Location Management. The cellular network has a checker board design as shown in Figure 3. Each block represents a square-shaped cell which is not identified by a Cell Global ID. Instead they are assigned Geographical Information System (GIS) ID. This parameter is not parsed in any signaling messages related to mobility and location management. So the network is agnostic to the exact location of the mobile station. The white blocks are allocated AFCRNs from 1 to n, while the black ones are allocated AFCRNs from $n+1$ to k.

Say we consider $n = 64$, $k = 128$ in this paper to keep parity with allocation methodology in traditional GSM. In the mobile station, the values of n and k are known, All the AFCRNs in white shade are mapped sequentially to the AFCRNs in the black shade in the mobile station. When a Mobile user is provisioned in the network, a primary AFCRN is allocated. The primary AFCRN pertains to a given shade of the cell, either white or black. If the mobile station attempts to invoke a message or respond to a page in a cell with opposite shade, the AFCRN mapping process (between white and the black shades) as above is invoked.

Each mobile station is fitted with 2 transceivers [25, 26]. There is an active section which takes part proactively in the network operations within the cell where the mobile user is currently located. The dormant part keeps on scanning the adjacent cells and is mainly responsible for actuating the handover procedure as in Figure 3.

A central processor in the mobile station monitors and compares the signal strength as perceived by the 2 transceivers, the active one which is already in a call and the passive one which compares the pilot. The MS also has an inbuilt GPS unit, and if in open air (with LOS with

the satellite) can also determine the direction and speed of motion. It also has an electronic compass to analyse the direction and interpret the direction of motion if the line of sight (LOS) with the satellite is not available. Analysing the signal strength of the 2 received signals (and also gathering intelligence from the GPS/compass unit if available), the handset decides to initiate the handover.

In such a case, the passive transceiver as in Figure 3, conveys the time slot and the frequency which is in use by the active transceiver. The base station serving the cell establishes a channel (TS), the same value which was intimated by the passive tranceiver. This happens exactly when the active transceiver releases the channel so that the base stations of the adjacent white and the black cells can establish channel on the same TS which is currently in use between the base station of the white cell and the MS.

Hence the physical trajectory of the call is:

passive transceiver of the MS to → *base station/access point of black cell* → *base station/access point of the white cell* (*which was already in conversation phase directly with the MS before the handover*) → *central Coordination Processor*.

After the handover, the passive transceiver in the MS becomes the primary one. The transceiver which was acting as primary before the handover process now becomes dormant which commences scanning the AFCRNs of the adjacent cells for evaluating the signal strength.

The handover never happens between the cells with the same AFCRN band that is with the same colour. The lateral handover is actuated directly between the white cell to black cell or vice versa. However, the diagonal handover happens through a process of 2 lateral handovers [25, 26].

4. Comparative Analysis of the Signaling Processes and Overhead for Mobility Management in the-State-of-the-Art and the Proposed Smart Network

The unique mobility management procedures in SMNAT are not designed specifically for one specific application scenario but for supporting coexistence of M2M, M2H, H2M, and H2H applications. The motivation is to reduce the signaling interaction between the network and the device, actuate a better congestion control, and thereby alleviate the impact due to existence of heterogeneous devices in the mobile network. The following section to scrutinise the major differences between the state-of-art and the SMNAT in terms of signaling and mobility management is applicable for all types of User Equipments including the M2M terminals.

4.1. Synopsis of Mobility and Location Management in LTE. The signaling overhead due to mobility management is related to mainly 2 subprocesses, location update and paging. During TA (Tracking area) update procedure [27], the Mobility Management Entity (MME) [10] records the TA in which the User Equipment (UE) is located. When a UE, that is, the Mobile device moves to a new TA, tracking area update is initiated by the device. Paging comes in play when the UE is being called for voice, SMS, network-initiated location messages (for location services), and network-initiated USSD. In order to place the call to the UE, MME broadcasts paging message in all cells of the UE's registered TA. In LTE, the Mobility Management Entity (MME) is responsible for the mobility management function. The MME is connected to a large number of evolved Node Bs (cells) that are grouped into the Tracking Areas (TAs). The TAs are further grouped into TA Lists (TALs). When a UE moves out of the current TAL, it reports its new location to the MME. If the LTE network attempts to connect to the UE, the MME asks the cells in the TAL to page the UE. In LTE paging, the MME may sequentially page a cell, the TA of the cell, and/or TAL of the cell.

In the following sections, we investigate the 3 main processes of a mobile network related to mobility and location management which imparts signaling overhead in the network. A comparative analysis is made between the-state-of-the-art and the SMNAT.

4.1.1. Location Update State-of-the-Art (LTE). Location update in LTE can be categorised as following.

Static Scenarios.

(i) Always-update: in this scheme, the user updates its location whenever it moves into a new cell.

(ii) Never-update: in this scheme, the user never updates its location, which means that the location update overhead is zero, but may lead to excessive paging.

(iii) Reporting cells: in this scheme, the user updates its location only when visiting one of the predefined reporting cells.

(iv) Forming LA: in this scheme, the user updates its location whenever it changes an LA.

Dynamic Update Schemes.

(i) Selective LA update: in this scheme, the LAU is not performed every time the user crosses an LA border.

(ii) Time-based: in this scheme, the user updates its location at constant time intervals.

(iii) Profile-based: in this scheme, the network maintains a list of "Location Areas" where the user was located in specific time periods.

(iv) Movement-based: in this scheme, the user updates its location after a given number of boundary crossings to other cells in the network.

(v) Distance-based: in this scheme, the user updates its location when it has moved away a certain distance from the cell where it has last updated its location. This process is primarily based on reference signal received power (RSRP) calculations by the network.

(vi) Predictive distance-based: in this scheme, the network determines the probability density function of the user's location based on location and speed reports.

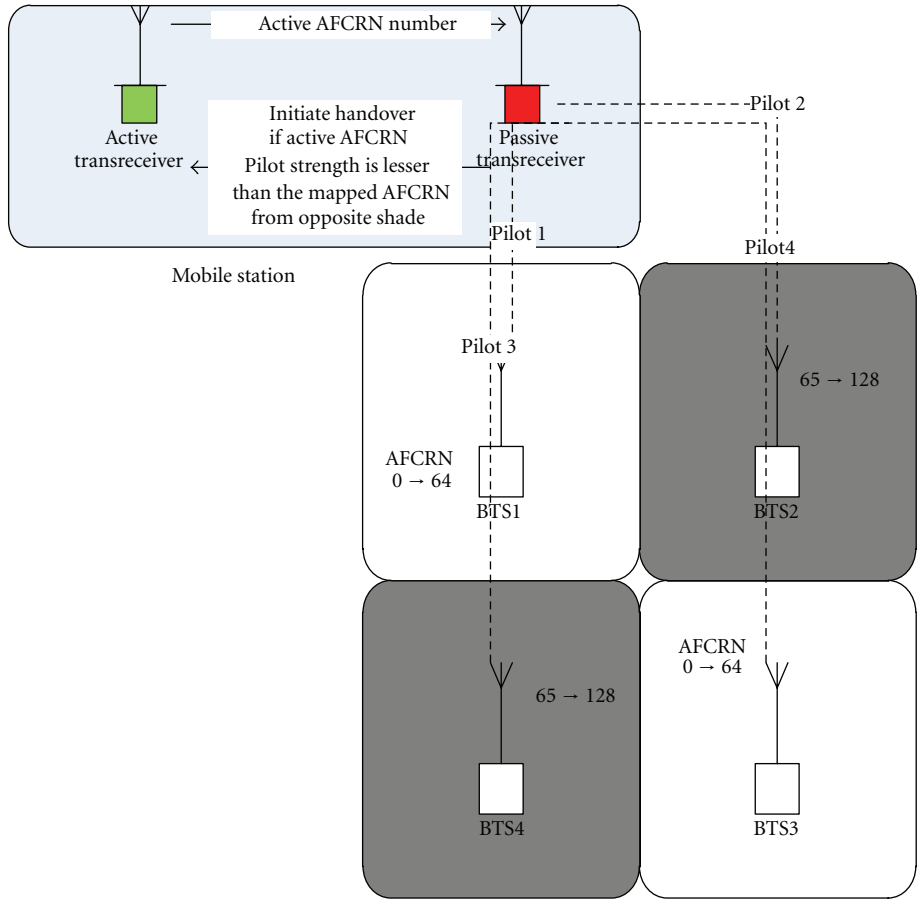


Figure 3: Active and passive transceiver and its network interaction.

The MME houses the location update algorithm to cater to any of the above-mentioned scenarios. The signaling overhead for a location update message as seen in the core network is around 378 Bytes. Depending upon the scheme chosen, the total overhead for exchanging the location update message can be determined. Figure 4 shows a location update message showing layer 3 to layer 7.

Before the application layer message for location update is sent, the LTE handset initiates the synchronisation and random access process to acquire a channel. Figure 5 demonstrates the synchronisation and the random access process for the LTE access layer.

The processing of MIB/SIBs reception by the handset (40 ms in case of MIB and 80 ms in case of SIB 1) consumes battery power in the handset.

The bandwidth required for location update can be calculated as follows.

The set of cells in a network is denoted by $N = \{1,\ldots,N\}$, and the set of TAs currently in use is denoted by $T = \{1,\ldots,T\}$. The vector $\mathbf{t} = [t1,\ldots,tN]$ is used as a general notation of cell-to-TA assignment, where t_i is the TA of cell i. TA design $\mathbf{t}$ can be alternatively represented by an $N \times N$ symmetric and binary matrix $\mathbf{S}(\mathbf{t})$; in which element $s_{ij}(\mathbf{t})$ represents whether or not two cells are in the same TA, that is,

$$s_{ij}(\mathbf{t}) = \begin{cases} 1 & \text{if } t_i = t_j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

h_{ij} is the number of UEs moving from cell i to cell j.

The overhead (B_{up}) of one location update

$$B_{\text{up}} = \sum_{i \in N} \sum_{j \in N : j+1} \left(c^u h_{ij} \left(1 - s_{ij}(\mathbf{t}) \right) \right). \tag{2}$$

4.1.2. Location Update in SMNAT. There is no location update process as there is no network acquisition required in SMNAT. More importantly, there is no location update required prior to the execution of the other network operations/processes, for example while receiving calls/SMS or during handovers. For these operations the necessary precondition is successful paging of the UE at physical layer but the UE does not need to attach to the given network. Thus by avoiding location updates, we save substantial battery power of the UE as well as spectrum bandwidth.

4.1.3. Paging in LTE. In most cases, this paging process happens while UE is in idle mode. This means that UE has

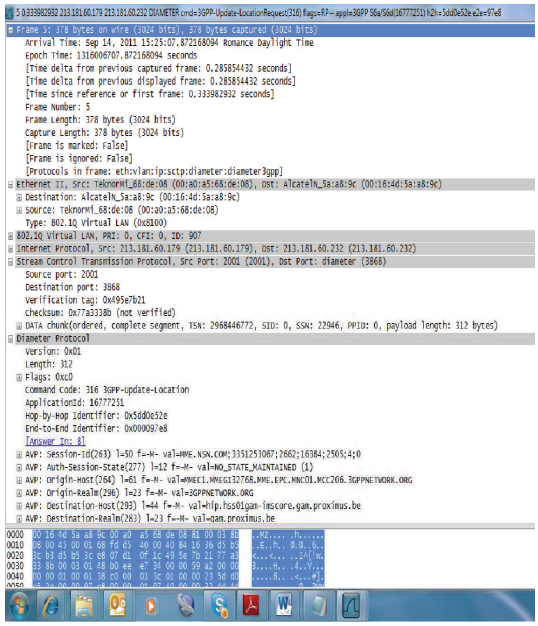


Figure 4: Update Location Request on Diameter S6 Interface.

to monitor whether the networking is invoking any paging message to it and it has to spend some energy (battery) to run this "Monitorin" process.

The paging procedure is used by the network to request the establishment of a NAS signaling connection to the UE. The network shall initiate the paging procedure for EPS services using S-TMSI with CN domain indicator set to "PS" when NAS signaling messages are sent to the UE.

Paging in LTE happens via PDCCH channel. The UE listens to the PDDCH channel in given intervals, but not continuously. This is done by DRX (Discontinuous Reception) in order to save power.

The DRX LTE has two timing units as many of other technologies. Timing Unit in Frame scale (SFN: System Frame Number) is one unit and the timing unit in subframe level (Subframe Number). It means that the network needs to know both SFN and Subframe Number to locate exact position in LTE time domain. Regarding the paging cycle, PF (Paging Frame) + PO (Paging Occasion) let you know the exact timing when UE has to wake up to catch the paging message being sent to it:

$$\text{PF} = \text{SFN} \bmod T = (T \operatorname{div} N) \times (\text{UE\_ID} \bmod N). \quad (3)$$

According to 3GPP 34.104, T is defined as follows.

T is DRX cycle of the UE. T is determined by the shortest of the UE specific DRX value, if allocated by upper layers, and a default DRX value broadcast in system information. If UE-specific DRX is not configured by upper layers, the default value is applied.

It means UE can get the T from two different sources, one from the system information (SIB2, IE default paging cycle) [28] and the other from upper layer. If upper layer sends the value, it use the T value from the upper layer, otherwise UE has to use the value from SIB2. If T is 5 milliseconds, then the handset will not be able to receive a page message unless the DRX sleep cycle is completed.

Paging in LTE is done at PDCCH using the S-IMSI (SAE-Temporary IMSI) for addressing which is a 32 bit value.

4.1.4. Paging in SMNAT. The address for the page message is the symbol coordinate at a specific time slot in the radio frame (in the Predefined AFCRN) assigned to the UE. The UE is always in the sleep mode, except the fact that it has to wake up to listen to the specific time slot assigned to itself in the available AFCRNs available within the spectrum. The UE does not need to listen to the entre radio frame. This is not possible in the-state-of-the-art as the paging information can arrive dynamically from any time slot in the frame.

The time slot carries only a single symbol with a repeated transmission of 4 times. If each symbol measures a time span of 1 microsecond, the total time of the radio frame is 10 ms (say), then the UE has to remain live for 4 microseconds. Lower time to be live implies better battery power conservation.

Secondly, no higher layer addressing parameters are used for paging which also conserves the bandwidth. The layer 7 parameters are exchanged only when the paging process is completed and the time slot is seized between the UE and the CP (Coordination Processor) for the conveyance of the signaling and data information. This is the primary advantage of the SMNAT over the-state-of-the-art where it is not possible to confine the paging process to a layer 1 activity.

The time period during which the UE needs to wake up can be calculated as follows (refer Figure 6) Let each frame pertaining to a single AFCRN contain n to $n + k$ time slots.

Say the coordinate symbol (of the complex plane) assigned to the specific UE arrives in Kth time slot.

Let there be AFCRNs ranging from i to j.

Hence the time period for which a UE needs to be live within the radio frame T live

$$T \text{ live} = 4\sum_{i}^{j} t(n), \quad (4)$$

where n can range from 0 to K.

As we have a repetitive code (4 repetitions) for error correction, so the total time is multiplied by 4.

5. A Summary of the Key Issues with the Available Network Topologies for M2M Devices to Interoperate, and How They Are Addressed/Resolved by the SMNAT

5.1. The Radio Network

Issue with Present Generation Networks. it has a complex process of mobility management and radio resource management. Basic philosophy remains the same across all the generations (2G/3G/4G). Complexity exists in radio network design. Each cell in the network area is allocated a range of frequencies. Each cell has its own cell id. Network efficiency depends much on the frequency allocation and the frequency

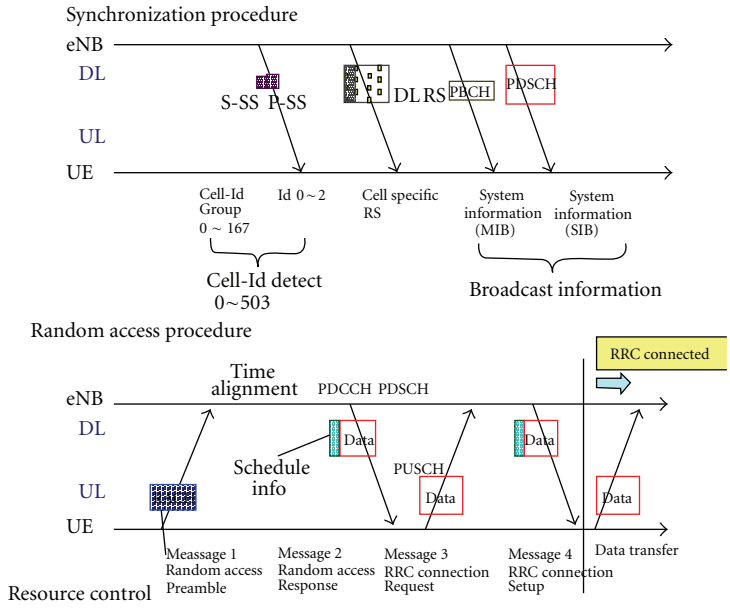


Figure 5: Synchronisation and the random access process for the LTE access layer.

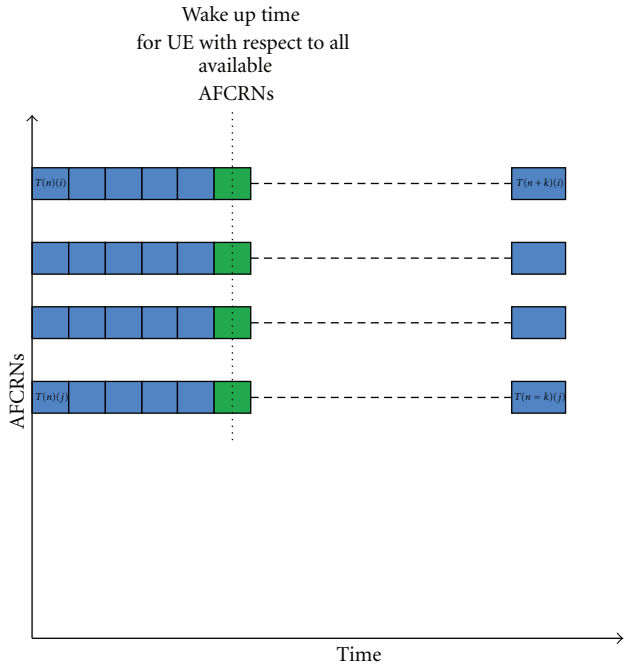


Figure 6: Wake up time for the UE.

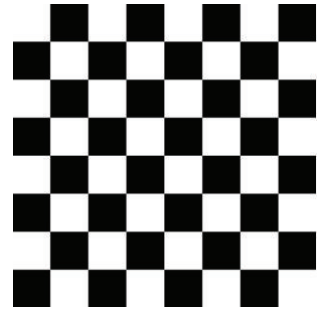


Figure 7: Cell pattern and AFCRN block assignment with proposed topology.

reuse pattern defined by the operator. Handover of a call when the subscriber is moving from one cell to the other is a process that requires lot of signaling interaction between the handset and the network. As the penetration of the M2M devices will be substantial, hence we will see a sharp increase in the signaling messages in the RAN and CN primarily due to handover, periodic location updates, location updates triggered by change in location. This poses a challenge for the M2M device itself because of the impending needs to power conservation. The more the signaling messages transceived, the more the power is consumed. Moreover, the network needs to be reengineered from to augment the capacity of the network elements. Requirements for additional spectrum requirements also need to be addressed.

With the proposed technology: we have a simple checker board design. The available bandwidth is divided in just 2 subbands (against 125 of them as in GSM). The cells are square shaped. Figure 7 defines the cell pattern. Handover is driven by the handset. The signaling interaction required for handover is lesser [26], compared to the hard and soft handoff schemes followed in the-state-of-the-art.

There are no periodic location updates invoked by the mobile station. There are no location updates traversing the RAN and reaching the core network for location changes.

5.2. Addressing. In state-of-the-art, when a mobile station acquires a network, an IP address is assigned. IP V4 addresses are 8 bits, while IPV6 IP addresses are 32 bits. So the

capability to address the mobile devices is limited by the number of devices present in the network.

With the new approach, the addressing is done by physical layer parameters. The subscriber coordinate per time slot is directly used for node addressing. So the constraint to reach the devices is limited by the time slot availability which again depends on the spectrum availability. So the physical limitation due to the IP addressing mechanism to reach a vast expanse of the M2M devices is resolved.

5.3. Dependence on Layer 7 (Application Layer according to OSI Model) Processes. Network tracks the handset continuously. Handset initiates a location update process in a given time periodicity. Network and handset manage the handover to another cell when location is changed. All these network activities are managed by application layer signaling processes at the radio network. So the network equipments and the handset need to be intelligent enough to compute and process continuously the signaling messages for mobility and location management. The network equipments should be able to actuate these processor intensive activities, thus increases the cost of overall network. For the handset as well, it is the same issue as above. Moreover, these processes drain alot of battery power, just to maintain thier "presence" in the network. This makes the mobile device more expensive. When a large number of devices are required (e.g., each device per sensor in a sensor network) the overall cost to set up the application takes an upward trail.

With the proposed technology, we render more intelligence to the lower layers so that it can simplify and actuate some processes (like addressing a mobile node), which are today layer 7 functionalities. We realise a smart modulation scheme which does not just aim to inject more symbols per unit time. It rather (also) attempts to identify the users at layer 1 (physical layer). This in turn enables us to realise the devices cheap, and when we have a swarm of the M2M devices, it makes a substantial difference from the perspective of capital expenditure.

6. Conclusion

The QoS offered by the-state-of-the-art for the Human to Human applications may be grossly impacted once the swarm of M2M devices commence to interoperate in the mobile network framework in full swing. In this paper, we identified the key issues related to this context. We attempted to redesign the mobility and the radio resource management part of the mobile networks to simplify the associated processes and hence reduce the network interaction between the device and the handset. This will help in reducing congestion and boosting the performance of both the M2M and the H2H devices by optimising the battery power, processing power and the bandwidth. As a result, the overall cost to deploy, maintain, and operate the devices and the network will reduce. This is aligned with our endeavour to make the mobile networks cleaner, greener, and leaner.

References

[1] EXALTED: Expanding LTE for Devices:FP7 Contract Number: 258512. European Commission Information Society and media.

[2] R. Gunasekaran, V. R. Uthariaraj, R. Rajesh, S. Kaarthikeyan, and S. Aravind, "Priority scheduling in mobile ad hoc networks with improved bandwidth utilization," in *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE '08)*, pp. 1955–1958, May 2008.

[3] T. Arora, P. Singh, and S. S. Kang, "Performance evaluation and improving bandwidth ulilization of AODV protocol by finding hidden terminals in wireless networks," *International Journal of Computer Science and Telecommunications*, vol. 2, no. 6, 2011.

[4] T. Gowri, K. Raja, and G. T. Arasu, "Improving the bandwidth utilization by recycling the unused bandwidth in IEEE 802. 16 networks," *International Journal of Electronics and Computer Science Engineering*, vol. 1, pp. 623–631, 2012.

[5] P. Omiyi, H. Haas, and G. Auer, "Analysis of intercellular timeslot allocation in self-organising wireless networks," in *Proceedings of the17th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '06)*, September 2006.

[6] M.-C. Chang and B.-H. Liao, "Improving bandwidth utilization of NG-SDH by dynamic bandwidth management," in *Proceedings of the 13th Asia-Pacific Network Operations and Management Symposium (APNOMS '11)*, 2011.

[7] D. G. Jeong and W. S. Jeon, "Time slot allocation in CDMA/TDD systems for mobile multimedia services," *IEEE Communications Letters*, vol. 4, no. 2, pp. 59–61, 2000.

[8] A. Kumar, Y. Liu, J. Sengupta, and Divya, "Evolution of mobile wireless communication networks: 1G to 4G," *The International Journal on Electronics & Communication Technology*, vol. 1, no. 1, 2010.

[9] 3GPP 23. 228 Architecture and main flows for a IMS system.

[10] 3GPP TS 24. 301: Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS), Stage 3.

[11] V. Simon, *Mobility management algorithms for reducing signaling overhead in next generation mobile networks [Ph.D. thesis]*, Department of Telecommunications Budapest University of Technology and Economics, 2008.

[12] F. Paint, P. Engelstad, E. Vanem et al., "Mobility aspects in 4G Networks- White Paper".

[13] C. Castelluccia, "A Hierarchical Mobile IPv6 Proposal. Theme 1—Reseaux et systemes," Rapport Technique 0226, Projet SIRAC.

[14] 3GPP TS 36. 104 V8.4.0 (2008-12) Base Station Radio Transmission and Reception.

[15] 3GPP TS 36. 331 V8.4.0 (2008-12) Radio Resource Control (RRC) Protocol Specification.

[16] 3GPP TS 36. 211 V8.5.0 (2008-12) Physical Channels and Modulation.

[17] Network Improvements for Machine Type Communications (NIMTC), Triggered by Rel-8 UID_7027 Study on Facilitating Machine to Machine Communication in GSM and UMTS (FS_M2M) TR 22. 868, created 2008-09-15, last updated 2011-02-28.

[18] System Improvements to Machine-Type Communications (SIMTC), Enhancements to Rel-10 Feature UID_410031 (NIMTC) based on Rel-11 Study on enhancements to MTC, created 2010-05-28, last updated 2011-02-28.

[19] G.-Y. Lin, H.-Y. Wei, Y.-S. Chen, and C.-C. Hsu, "Enhanced RACH design for machine-type communications," US Patent application 20120033613, August 2011.

[20] F. Yu and V. Leung, "Mobility-based predictive call admission control and bandwidth reservation in wireless cellular networks," *Computer Networks*, vol. 38, no. 5, pp. 577–589, 2002.

[21] S. Choi and K. G. Shin, "Predictive and adaptive bandwidth reservation for hand-offs in qos-sensitive cellular networks," in *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM '98)*, vol. 28, pp. 155–166, October 1998.

[22] G. Cao, "Integrating distributed channel allocation and adaptive handoff management for QoS-sensitive cellular networks," *Wireless Networks*, vol. 9, no. 2, pp. 131–142, 2003.

[23] K.-R. Jung, A. Park, and S. Lee, "Machine-type-communication (mtc) device grouping algorithm for congestion avoidance of mtc oriented lte network," in *Security-Enriched Urban Computing and Smart Grid*, T.-H. Kim, A. Stoica, and R.-S. Chang, Eds., vol. 78 of *Communications in Computer and Information Science*, pp. 167–178, Springer, Berlin, Germany, 2010.

[24] R. Sanyal, E. Cianca, and R. Prasad, "Rendering intelligence at physical layer for smart addressing and multiple access," in *Proceedings of the 9th International Conference on Networks*, Menuires, France, 2010, http://www.computer.org/portal/web/csdl/doi/10.

[25] R. Sanyal, E. Cianca, and R. Prasad, "Beyond LTE: next generation multiple access technology with intelligent lower layers," in *Proceedings of the World Congress Computer Science, Computer Engineering and Applied Computing and International Conference on Wireless Networks (ICWN '11)*, Las Vegas, Nev, USA, 2011, http://www.world-academy-of-science.org/.

[26] S. Rajarshi, E. Cianca, and R. Prasad, "Smart Mobile Networks with intelligent physical layer actuating Mobility Management," *International Journal of Mobile Network Design and Innovation*, vol. 4, no. 2, pp. 91–108, 2012.

[27] S. M. Razavi, "Tracking area planning in cellular networks-optimization and performance evaluation linkoping studies in science and technology," Licentiate Thesis 1473.

[28] 3GPP TS 36. 331 V8.5.0 (2009-03) Evolved Universal Terrestrial Radio Access (E-UTRA) Radio Resource Control (RRC) Protocol specification.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 714971, 11 pages
doi:10.1155/2012/714971

Research Article

On the Achievable User Number of the Downlinks in Cellular-Based Machine-to-Machine Communications

Yingbo Li,[1] Chen Chen,[1] Lin Bai,[2] and Ye Jin[1]

[1] *State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China*
[2] *School of Electronic and Information Engineering, Beihang University, Beijing 100191, China*

Correspondence should be addressed to Lin Bai, l.bai@buaa.edu.cn

Received 20 April 2012; Revised 5 July 2012; Accepted 9 July 2012

Academic Editor: Jianhua He

In cellular-based machine-to-machine (M2M) networks, supporting large number of machine-type communications (MTC) users (devices) has become a critical challenge in both the uplink and downlink channels. In this paper, we focus on the downlink beamforming using zero-forcing dirty paper coding (DPC). In order to characterize the system's ability of user admission, we consider the achievable user number, which is defined as the number of users whose signal-to-interference plus-noise ratios exceed a target threshold. Due to the complexity of the optimal scheme, we propose two algorithms with random user scheduling and greedy user scheduling in maximizing the achievable user number by dynamical power assignment. Using the joint distribution of effective channel gains, we derive the achievable user number of both the scheduling schemes. An upper bound on the achievable user number of the greedy scheme is then derived which is shown to be tight when there are a large number of users. From numerical results, we show that both of the schemes enjoy a linear increase in the achievable user number as the number of transmitter antennas increases. The performance of the greedy scheduling scheme is close to the optimal scheduling scheme.

1. Introduction

Wireless machine-to-machine (M2M) networks are emerging as new types of communication, where cellular systems are expected to play an important role in wireless M2M networks [1]. In cellular-based M2M networks, supporting large number of machine-type communications (MTC) users (devices) has become a critical challenge in both the uplink [2, 3] and downlink channels. The rate required by each MTC user may be low, however, their number may be large. Thus, the study on efficient user admission policies becomes one of the critical issues in M2M communications. In this paper, we focus on improving the number of active users in the cellular-based M2M communications networks.

In the cellular communications networks, beamforming techniques with multiple antennas at the base station (BS) are used for simultaneous transmission of multiple users. The multi-antenna techniques have received a lot of attention since the heuristic research work by Telatar [4] due to its diversity and multiplexing gain. When perfect channel state information is available at the BS, the optimal downlink beamforming scheme is dirty paper coding (DPC) [5] where the interference known at the transmitter is pre-subtracted at the BS. The optimal beamforming weight vectors and encoding order of DPC can be found using a duality between the downlink and the corresponding uplink channel. In [6, 7], numerical algorithms are proposed based on this duality. However, these algorithms exhibit an undesired iterative feature.

In order to reduce the complexity, a zero-forcing structure is imposed in DPC [5], that is, zero-forcing DPC. Only random user scheduling is considered in [5]. Subsequent works [8, 9] extend zero-forcing DPC in [5] by considering the user selection and propose a greedy user scheduling scheme.

In this paper, we focus on the performance evaluation and algorithms of multi-user systems with zero-forcing DPC. We are interested in the achievable user number, which is

defined as the average number of users that are admissible in the system. We say that a user is admissible if the effective signal-to-interference-plus-noise ratio (SINR) achieves a given target SINR threshold. We propose algorithms to maximize the achievable user number. The difference of the zero-forcing DPC considered in this paper from the conventional zero-forcing DPC is that the proposed algorithms should maximize the achievable user number instead of the spectral efficiency (rate). Extensive literatures have been devoted to the spectral efficiency with few exceptions, for example, [10, 11]. The achievable user number is considered in [10] for characterizing the performance of a circuit scenario.

We consider the beamforming vector design and user scheduling of zero-forcing DPC for maximizing the achievable user number. The optimal beamforming vectors for maximizing the achievable user number can be calculated by the Gram-Schmidt process which is a method for orthogonalizing a set of vectors in a successive manner. When there are more users than the number of transmitter antennas, a user scheduling problem exists. The optimal scheduling scheme can be found by solving a combinatorial optimization problem which is not easy to solve. Thus, we propose two suboptimal scheduling algorithms for maximizing the achievable user number, that is, random scheduling and greedy scheduling.

We derive the achievable user number performance of both the random and greedy user scheduling, using the joint distribution of effective channel gains. An upper bound on the achievable user number of the greedy scheme is derived. Various numerical results are present to justify our theoretical analysis. The performance of the greedy scheduling scheme is shown to be close to the optimal scheduling scheme.

The rest of the paper is organized as follows. The system model is presented in Section 2. Some preliminaries of zero-forcing DPC are given in Section 3. The algorithms for maximizing the achievable user number are proposed in Section 4. The achievable user numbers of both random scheduling scheme and greedy scheduling scheme are derived in Section 5. A comprehensive evaluation of the achievable user number of zero-forcing DPC is given in Section 6 with numerical results. We conclude the paper with some remarks in Section 8.

Notations. The superscripts T and H stand for the transpose and Hermitian transpose, respectively. Upper and lower boldfaced letters are used for matrices and column vectors, respectively. Denote by $|x|$ the absolute value of a scalar x, $||\mathbf{x}||$ the 2-norm of vector $\mathbf{x}$, and $\mathbf{I}$ the identity matrix of a certain size implicitly given by the context. We denote by $[\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N]$ the concatenation of N column vectors, and by $[\mathbf{W}, \mathbf{c}]$ the concatenation of a matrix $\mathbf{W}$ and a vector $\mathbf{c}$ of the same column size. $\mathcal{CN}(\mathbf{a}, \mathbf{R})$ represents the distribution of circularly symmetric complex Gaussian (CSCG) random vectors with mean vector $\mathbf{a}$ and covariance matrix $\mathbf{R}$. We denote the probability of an event X by $\Pr(X)$.

2. System Model

Consider the downlink transmission of a single cell system with a base station (BS) and K users in M2M communications. The users may be the group heads of group-based MTC devices. Suppose that the BS is equipped with an antenna array of L elements, while each user is with single antenna. Multiuser downlink beamforming is used for simultaneous transmission for N ($\leq K$) active users in the system. As the number of antennas at the BS is L, a typical value of N is L, that is, $N = L$, when there are enough candidate users ($K \geq L$). Generally, we assume that $K \geq L$ and $N \leq L$ in this paper. Denote the set of indices of all K candidates as $\mathcal{K} = \{1, 2, \ldots, K\}$, and the set of indices of the active N users $\mathcal{A} = \{k_1, k_2, \ldots, k_N\} \subseteq \mathcal{K}$.

The transmitted signal vector from the antenna array to N active users, denoted by $\mathbf{x}$, is given by

$$\mathbf{x} = \sum_{n=1}^{N} \mathbf{w}_{k_n} u_{k_n}, \tag{1}$$

where u_{k_n}'s are the independent data symbols for the active users, and $\mathbf{w}_{k_n}$'s are the corresponding beamforming vectors. The beamforming weight vectors $\mathbf{w}_{k_n}$'s are normalized $L \times 1$ column vectors, that is, $\|\mathbf{w}_{k_n}\| = 1$. The transmitted powers of each symbol are denoted as P_{k_n}'s, and $P_{k_n} = \mathbb{E}[|u_{k_n}|^2]$. It is assumed that the total transmitted power is limited, that is, $\sum_n^N P_{k_n} \leq P$.

The signal received at the active user k_i in the cell becomes

$$y_{k_i} = \mathbf{h}_{k_i}^H \mathbf{x} + n_{k_i}, \tag{2}$$

where $\mathbf{h}_{k_i}$ is the $L \times 1$ downlink channel vector from the transmit antenna array to the user k_i, and n_{k_i} is the independent additive white Gaussian noise (AWGN). It is assumed that the noise is CSCG distributed with zero mean and unit variance, that is, $n_{k_i} \sim \mathcal{CN}(0, 1)$.

Substituting (1) into (2) yields

$$y_{k_i} = \mathbf{h}_{k_i}^H \sum_{n=1}^{N} \mathbf{w}_{k_n} u_{k_n} + n_{k_i} = \mathbf{h}_{k_i}^H \mathbf{w}_{k_i} u_{k_i} + \mathbf{h}_{k_i}^H \sum_{n \neq i} \mathbf{w}_{k_n} u_{k_n} + n_{k_i}. \tag{3}$$

The signal-to-interference-plus-noise ratio (SINR) of user k_i, denoted by γ_{k_i} is given as follows:

$$\gamma_{k_i} = \frac{\left| \mathbf{h}_{k_i}^H \mathbf{w}_{k_i} \right|^2 P_{k_i}}{\sum_{n \neq i} \left| \mathbf{h}_{k_i}^H \mathbf{w}_{k_n} \right|^2 P_{k_n} + 1}. \tag{4}$$

The target SINR threshold for user k is denoted as Γ_k. We are interested in the achievable user number which is defined as the number of users with SINR higher than the target threshold. Thus, the achievable user number becomes

$$M = \sum_{i}^{N} \mathcal{I}(\gamma_{k_i} \geq \Gamma_{k_i}), \tag{5}$$

where $\mathcal{I}(x)$ is the indicator function, and

$$\mathcal{I}(x) = \begin{cases} 1, & x \text{ is true;} \\ 0, & x \text{ is false.} \end{cases} \tag{6}$$

The aim of this study is to support as much MTC users as possible in a cell, that is, to achieve a higher M.

Throughout the paper, we assume that the channel vectors $\mathbf{h}_k$, $k \in \mathcal{K}$ are available at the BS for both beamforming and scheduling.

3. Preliminaries of Zero-Forcing DPC

In this section, some preliminaries of downlink beamforming with zero-forcing DPC are given.

3.1. Dirty Paper Coding. Assume a specific DPC encoding order [12] π, which means the encoding order of the user n is $\pi(n)$. The signal received by user k_i can be expanded as

$$y_{k_i} = \mathbf{h}_{k_i}^H \mathbf{w}_{k_i} u_{k_i} + \mathbf{h}_{k_i}^H \sum_{\pi(n)<\pi(i)} \mathbf{w}_{k_n} u_{k_n} + \mathbf{h}_{k_i}^H \sum_{\pi(n)>\pi(i)} \mathbf{w}_{k_n} u_{k_n} + n_{k_i}. \tag{7}$$

The signals intended for the users with encoding order ahead of the user k_i cause no interference to user k_i. That is, the term $\mathbf{h}_{k_i}^H \sum_{\pi(n)<\pi(i)} \mathbf{w}_{k_n} u_{k_n}$ in (7) is pre-subtracted by DPC at the transmitter. The resulting signal received by user k_i becomes

$$y_{k_i} = \mathbf{h}_{k_i}^H \mathbf{w}_{k_i} u_{k_i} + \mathbf{h}_{k_i}^H \sum_{\pi(n)>\pi(i)} \mathbf{w}_{k_n} u_{k_n} + n_{k_i}, \tag{8}$$

and the SINR

$$\gamma_{k_i} = \frac{\left| \mathbf{h}_{k_i}^H \mathbf{w}_{k_i} \right|^2 P_{k_i}}{\sum_{\pi(n)>\pi(i)} \left| \mathbf{h}_{k_i}^H \mathbf{w}_{k_n} \right|^2 P_{k_n} + 1}. \tag{9}$$

For the index specific order, that is, $\pi(n) = n$, for all $n = 1 \cdots N$, the SINR becomes

$$\gamma_{k_i} = \frac{\left| \mathbf{h}_{k_i}^H \mathbf{w}_{k_i} \right|^2 P_{k_i}}{\sum_{n>i} \left| \mathbf{h}_{k_i}^H \mathbf{w}_{k_n} \right|^2 P_{k_n} + 1}. \tag{10}$$

It has been proven that DPC is the capacity-achieving scheme for the multiple-input multiple-output (MIMO) broadcast channel (BC). The optimal beamforming weight vectors and encoding order can be found using a duality between the BC and the corresponding multiple access channel (MAC). Several numerical algorithms have been proposed, for example, [6, 7]. However it has been recognized that these algorithms exhibit an undesired iterative feature.

3.2. Zero-Forcing DPC. In order to reduce the complexity of the calculation, a zero-forcing structure is imposed to the beamforming weight vectors, resulting in a simplified scheme, zero-forcing DPC scheme. The term $\mathbf{h}_{k_i}^H \sum_{\pi(n)>\pi(i)} \mathbf{w}_{k_n} u_{k_n}$ in (7) is nulled by the beamforming vector design, which is

$$\mathbf{h}_{k_i}^H \mathbf{w}_{k_n} = 0, \quad \forall \pi(n) > \pi(i). \tag{11}$$

The resulting signal received by user k_i becomes

$$y_{k_i} = \mathbf{h}_{k_i}^H \mathbf{w}_{k_i} u_{k_i} + n_{k_i}, \tag{12}$$

and the SINR

$$\gamma_{k_i} = \left| \mathbf{h}_{k_i}^H \mathbf{w}_{k_i} \right|^2 P_{k_i} = g_i P_{k_i}, \tag{13}$$

where $g_i = |\mathbf{h}_{k_i}^H \mathbf{w}_{k_i}|^2$ denotes the effective channel gain of the ith user.

4. Algorithm for Maximizing User Number

For a specific DPC encoding order π and N scheduled users $\mathcal{A} = \{k_1, k_2, \ldots, k_N\}$, the achieved user number becomes

$$M_{\pi,\mathcal{A}} = \sum_{i}^{N} \mathcal{I}(\gamma_{k_i} \geq \Gamma_{k_i}), \tag{14}$$

where γ_{k_i} is given by (13) and $\mathbf{w}_{k_i}$ is designed subject to the constraints (11). The optimal beamforming vectors can be found by solving the following optimization problem:

$$\begin{aligned} \underset{\mathbf{W},\mathcal{P}}{\text{maximize}} \quad & M_{\pi,\mathcal{A}} \\ \text{subject to} \quad & \mathbf{h}_{k_i}^H \mathbf{w}_{k_n} = 0, \quad \forall \pi(n) > \pi(i), \\ & \sum_{i}^{N} P_{k_i} \leq P, \end{aligned} \tag{15}$$

where $\mathbf{W} = [\mathbf{w}_{\tilde{k}_1}, \mathbf{w}_{\tilde{k}_2}, \ldots, \mathbf{w}_{\tilde{k}_N}]$, $\tilde{k}_i = k_j, \pi(j) = i$, and $\mathcal{P} = \{P_{k_1}, \ldots, P_{k_N}\}$ is the power assignment.

From Theorem 1 of [13], the optimal solution to the optimization problem in (15) is

$$\mathbf{W}_{\text{opt}} = \mathbf{Q}, \tag{16}$$

where $\mathbf{Q}$ is obtained from $\mathbf{QR}$ decomposition of the channel matrix $\mathbf{H} = [\mathbf{h}_{\tilde{k}_1}, \mathbf{h}_{\tilde{k}_2}, \ldots, \mathbf{h}_{\tilde{k}_N}]$,

$$\mathbf{H} = \mathbf{QR}, \tag{17}$$

and $\mathbf{R}$ is an upper triangular matrix.

The $\mathbf{QR}$ decomposition can be implemented using the Gram-Schmidt process which is a method for orthogonalizing a set of vectors in a successive manner. More efficient algorithm for QR decomposition can be carried out by the Householder transformation [14].

In a system where $K > N$, the optimal achievable user number can be found by

$$M = \max_{\pi,\mathcal{A}} M_{\pi,\mathcal{A}}. \tag{18}$$

Due to the successive nature of $\mathbf{QR}$ decomposition, the user scheduling process can be easily incorporated into the Gram-Schmidt process as in Algorithm 1.

Input: h_k, P.
Output: $\pi, \mathcal{A}, \mathcal{P}, \mathbf{W}$.
1: Initialize $\mathcal{A} = \emptyset$, $\mathcal{K} = \{1, 2, \ldots, K\}$, $n = 1$, $\mathbf{W} = [], \mathbf{H} = []$.
2: **while** $n \leq N$ **do**
3: Choose a user k_n from $\mathcal{K}$
4: $\mathbf{w}_{k_n} = \frac{\widetilde{\mathbf{w}}_{k_n}}{\|\widetilde{\mathbf{w}}_{k_n}\|}$, where $\widetilde{\mathbf{w}}_{k_n} = (\mathbf{I} - \mathbf{W}\mathbf{W}^H)\mathbf{h}_{k_n}$.
5: $P_{k_n} = \frac{\Gamma_{k_n}}{|\mathbf{h}_{k_n}^H \mathbf{w}_{k_n}|^2}$.
6: **if** $\sum_{i=1}^{n} P_{k_i} \leq P$ **then**
7: $\mathbf{H} \leftarrow [\mathbf{H}, \mathbf{h}_{k_n}]$.
8: $\mathbf{W} \leftarrow [\mathbf{W}, \mathbf{w}_{k_n}]$.
9: $\mathcal{A} \leftarrow \mathcal{A} \cup \{k_n\}$, $\mathcal{K} \leftarrow \mathcal{K} \setminus \{k_n\}$.
10: $\pi(n) = n$.
11: $n \leftarrow n + 1$.
12: **else**
13: break.
14: **end if**
15: **end while**
16: **return** $\pi, \mathcal{A}, \mathcal{P}, \mathbf{W}$.

Algorithm 1: Algorithm of zero-forcing dirty paper coding with user scheduling for maximizing user number.

In Algorithm 1, the powers are assigned dynamically to each active user according to their channel conditions.

The step 4 of Algorithm 1 can be expressed equivalently as

$$\widetilde{\mathbf{w}}_{k_n} = \left(\mathbf{I} - \mathbf{H}\left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H\right)\mathbf{h}_{k_n}. \tag{19}$$

The optimal user scheduling, step 3 of Algorithm 1, can be implemented by an exhaustive search of all possible active user set $\mathcal{A}$ and encoding order π. The scope of search are $K!/(K-N)!$ possible combinations, which means Algorithm 1 with specified $\mathcal{A}$ and π should be running $K!/(K-N)!$ times before the optimal user scheduling can be found.

The optimal user scheduling is a combinatorial optimization problem which is obviously not easy to solve. Therefore, two suboptimal user scheduling methods are considered in this study. The first heuristic method is random scheduling, in which a user is randomly scheduled for downlink beamforming. In random scheduling, the step 3 of Algorithm 1 is

$$\text{Randomly choose a user } k_n \text{ from } \mathcal{K}. \tag{20}$$

The second method is greedy scheduling. In greedy scheduling, the user requires the least power to achieve SINR threshold is to be selected. Thus, the step 4 of Algorithm 1 becomes

$$\text{Choose a user } k_n \text{ from } \mathcal{K} \text{ with the least } P_{k_n}. \tag{21}$$

5. Achievable User Number Performance

In this section, we derive the achievable user number of zero-forcing DPC with both random and greedy scheduling for maximizing the achievable user number.

The achievable user number M is a discrete random variable with possible values $\{0, 1, 2, \ldots, N\}$, and the probability distribution

$$\Pr(\xi = m), \quad m = 0, 1, 2, \ldots, N. \tag{22}$$

The average achieved user number, denoted by $\overline{M}$, is given by

$$\overline{M} = \sum_{m=1}^{N} m\Pr(\xi = m). \tag{23}$$

The probability

$$\Pr(\xi = m) = \begin{cases} \Pr\left(\sum_{i=1}^{m} P_i \leq P, \sum_{i=1}^{m+1} P_i > P\right), & 0 \leq m \leq N-1; \\ \Pr\left(\sum_{i=1}^{m} P_i \leq P\right), & m = N, \end{cases} \tag{24}$$

where

$$P_n = \frac{\Gamma_n}{|\mathbf{h}_n^H \mathbf{w}_n|^2} = \frac{\Gamma_n}{g_n}. \tag{25}$$

We denote the domain of (24) as $\mathcal{D}_m$. When $0 \le m \le N-1$, $\mathcal{D}_m$ is

$$\begin{cases} P_1 \le P, \\ P_2 \le P - P_1, \\ \quad \vdots \\ P_m \le P - \sum_{i=1}^{m-1} P_i, \\ P_{m+1} > P - \sum_{i=1}^{m} P_i. \end{cases} \iff \begin{cases} g_1 \ge \dfrac{\Gamma_1}{P}, \\ g_2 \ge \dfrac{\Gamma_2}{P - (\Gamma_1/g_1)}, \\ \quad \vdots \\ g_m \ge \dfrac{\Gamma_m}{P - \sum_{i=1}^{m-1}(\Gamma_i/g_i)} \\ g_{m+1} < \dfrac{\Gamma_{m+1}}{P - \sum_{i=1}^{m-1}(\Gamma_i/g_i)}. \end{cases} \tag{26}$$

When $m = N$, $\mathcal{D}_m$ is

$$\begin{cases} P_1 \le P, \\ P_2 \le P - P_1, \\ \quad \vdots \\ P_m \le P - \sum_{i=1}^{m-1} P_i. \end{cases} \iff \begin{cases} g_1 \ge \dfrac{\Gamma_1}{P}, \\ g_2 \ge \dfrac{\Gamma_2}{P - (\Gamma_1/g_1)}, \\ \quad \vdots \\ g_m \ge \dfrac{\Gamma_m}{P - \sum_{i=1}^{m-1}(\Gamma_i/g_i)}. \end{cases} \tag{27}$$

The probability $\Pr(\xi = m)$ can be calculated using the joint distribution of the effective channel gains,

$$\Pr(\xi = m) = \int \cdots \int_{\mathcal{D}_m} f_{g_1,g_2,\ldots,g_{m+1}}(x_1, x_2, \ldots, x_{m+1}) \times dx_{m+1} \cdots dx_2 dx_1, \tag{28}$$

where the integration domain $\mathcal{D}_m$ is specified by the right sides of (26) and (27).

For the purpose of analysis, we make the following CSCG assumption as those in many literatures, for example, [15, 16].

Assumption 1. The elements of each user's channel vector $\mathbf{h}_k$ are independent and identically (i.i.d.) distributed and CSCG random variables, $\mathcal{CN}(0,1)$.

Assumption 1 is reasonable in rich scattered environment. The path loss encountered by each user is assumed to be the same, and their channel directions are uniformly distributed. Assumption 1 simplifies the following performance analysis. It is noted that the algorithm in the Section 4 is not based on Assumption 1.

Besides, we assume that all the users have equal target SINR threshold, that is,

$$\Gamma_1 = \Gamma_2 = \cdots = \Gamma_K = \Gamma. \tag{29}$$

The superscripts R, G and U in this section denote the random scheduling, greedy scheduling and upper bound, respectively.

5.1. Achievable User Number with Random Scheduling. In the random scheduling scheme, we have the following lemma [5].

Lemma 1. *The channel gains g_i's are independently distributed as $g_i \sim \chi^2_{2(L-i+1)}$, where $\chi^2_{2(L-i+1)}$ denotes the central Chi-squared distribution with $2(L-i+1)$ degrees of freedom.*

Thus, the probability density function (PDF) of each channel gain g_i is

$$f^R_{g_i}(x_i) = \frac{x_i^{L-i} e^{-x_i}}{(L-i)!}, \quad i = 1, 2, \ldots, N, \tag{30}$$

and the cumulative distribution function (CDF)

$$F^R_{g_i}(x_i) = Q(x_i, L-i+1), \quad i = 1, 2, \ldots, N, \tag{31}$$

where $Q(\cdot)$ denotes the regularized incomplete gamma function (The regularized incomplete gamma function $Q(x,a) = (1/\Gamma(a)) \int_0^x e^{-t} t^{a-1} dt$, and $\Gamma(\cdot)$ is the gamma function. We use the MATLAB function gammainc to calculate the value).

Thus, the probability becomes

$$\Pr(\xi = m) = \int \cdots \int_{\mathcal{D}_m} f^R_{g_1}(x_1) f^R_{g_2}(x_2) \cdots f^R_{g_{m+1}}(x_{m+1}) \times dx_{m+1} \cdots dx_2 dx_1, \tag{32}$$

From (32), (30) and (23), we can calculate the average achievable user number of the random scheduling scheme.

5.2. Achievable User Number with Greedy Scheduling. In the greedy scheduling scheme, it can be shown that

$$g_1 \ge g_2 \ge \cdots \ge g_N. \tag{33}$$

Therefore, the integration domain with ordered effective channel gains, denoted by $\mathcal{D}^o_m$, becomes

$$\begin{cases} g_1 \ge \dfrac{\Gamma_1}{P} \\ g_1 \ge g_2 \ge \dfrac{\Gamma_2}{P - (\Gamma_1/g_1)}, \\ \quad \vdots \\ g_{m-1} \ge g_m \ge \dfrac{\Gamma_m}{P - \sum_{i=1}^{m-1}(\Gamma_i/g_i)}, \\ g_{m+1} < \dfrac{\Gamma_{m+1}}{P - \sum_{i=1}^{m}(\Gamma_i/g_i)}, \end{cases} \quad 0 \le m \le N-1, \tag{34}$$

$$\begin{cases} g_1 \ge \dfrac{\Gamma_1}{P}, \\ g_1 \ge g_2 \ge \dfrac{\Gamma_2}{P - (\Gamma_1/g_1)}, \\ \quad \vdots \\ g_{m-1} \ge g_m \ge \dfrac{\Gamma_m}{P - \sum_{i=1}^{m-1}(\Gamma_i/g_i)}, \end{cases} \quad m = N. \tag{35}$$

The probability $\Pr(\xi = m)$ becomes

$$\Pr(\xi = m) = \int \cdots \int_{\mathcal{D}^o_m} f^G_{g_1,g_2,\ldots,g_{m+1}}(x_1, x_2, \ldots, x_{m+1}) \times dx_{m+1} \cdots dx_2 dx_1, \tag{36}$$

where the joint distribution of the ordered effective channel gains in the greedy scheduling scheme is given in the Lemma 2.

Following the derivation in [9], we can have the lemma.

Lemma 2. *The channel gains g_i are jointly distributed as*

$$f^G_{g_1,g_2,\ldots,g_{m+1}}(x_1,x_2,\ldots,x_{m+1}) = \frac{K!}{(K-m-1)}\left[\int_{y=0}^{x_{m+1}}\phi_{m+1}(y,x_m,\ldots,x_1)dy\right]^{K-m-1} \times \prod_{i=1}^{m+1}\phi_i(x_i,\ldots,x_1), \tag{37}$$

where

$$\phi_i(x_i,x_{i-1},\ldots,x_1) = \frac{1}{\Gamma(L-i+1)x_i^{L-i}} \times \int_{v_{i-1}=x_i}^{x_{i-1}}\int_{v_{i-2}=v_{i-1}}^{x_{i-2}}\cdots\int_{v_1=v_2}^{x_1} dv_1\cdots dv_{i-1}, \tag{38}$$

$$\phi_1(x_1) = \frac{1}{\Gamma(L)}x_1^{L-1}e^{-x_1}. \tag{39}$$

Proof. The derivation follows from that of Proposition 1 in [9] with a notable exception that we proof the important Claim 1 in [9]. We give an equivalent of Claim 1 in [9] in Lemma 3. □

Lemma 3. *Let* $\mathbf{h}$ *and* $\mathbf{q}$ *denote independent L-dimensional random column vectors with i.i.d., CSCG entries with zero mean and unit variance. Let* $Y := \mathbf{h}^H\mathbf{h}$ *and* $X := \mathbf{h}^H\mathbf{Q}\mathbf{h}$, *where* $\mathbf{Q} = \mathbf{I} - \mathbf{q}(\mathbf{q}^H\mathbf{q})^{-1}\mathbf{q}^H$ *(cf.* (19)*). Then, the CDF of X, given Y, is given by*

$$F_{X|Y}(x \mid y) = \begin{cases}\left(\dfrac{x}{y}\right)^{L-1}, & \text{for } x\in[0,y],\\ 0, & \text{elsewhere.}\end{cases} \tag{40}$$

Proof. See Appendix A. □

From (37), (36) and (23), we can calculate the average achievable user number of the greedy scheduling scheme.

Inspired by the distribution in Lemma 1, an upper bound of the achievable user number can be found using the following lemma.

Lemma 4. *The best performance of the greedy user scheduling can be achieved when the channel gains g_i's are independently distributed as the maximum of $K-i+1$ random variables with distribution $\chi^2_{2(L-i+1)}$.*

Thus, using order statistics, the CDF of each channel gain g_i is

$$F^U_{g_i}(x_i) = \left[F^R_{g_i}(x_i)\right]^{K-i+1}, \tag{41}$$

and the PDF

$$f^U_{g_i}(x_i) = (K-i+1)\left[F^R_{g_i}(x_i)\right]^{K-i}f^R_{g_i}(x_i). \tag{42}$$

The probability becomes

$$\Pr(\xi=m) = \int\cdots\int_{\mathcal{D}_m} f^U_{g_1}(x_1)f^U_{g_2}(x_2)\cdots f^U_{g_{m+1}}(x_{m+1}) \times dx_{m+1}\cdots dx_2dx_1. \tag{43}$$

From (43), (42) and (23), we can calculate the upper bound on the average achievable user number of the greedy scheduling scheme.

6. Simulations

In this section, various numerical results are presented. In Section 6.1, the theoretic results derived from Section 5 are shown. In Section 6.2, we compare the performance of the proposed schemes with other beamforming schemes. In Section 6.3, the performances of the random and greedy scheduling schemes are compared with the optimal scheme. In Section 6.4, the improvements of the scheme with dynamic power assignment over that with equal power assignment are presented. The performance characterizations of the algorithms with dynamic power assignment are given in Section 6.5.

In the following simulations, samples of each user's channel vector $\mathbf{h}_k$ are drawn from CSCG distribution as specified in Assumption 1. The noise power spectral density is set at 1. 10^5 channel realizations are used to calculate the mean value of achievable user number.

6.1. Theoretic Results. In Figure 1 the achievable user number performance of the schemes with dynamic power assignment is plotted for different SINR thresholds Γ, with $K = 4$, $L = 2$, $N = 2$, and $P = 3$ dB. The analytical result of the random scheduling is obtained from (32), and that of the greedy scheduling from (36). The analytical lines in Figure 1 agree with those from simulations.

The upper bound of the greedy scheduling scheme with dynamic power assignment is shown in Figure 2, with $L = 4$, $N = 4$, and $P = 6$ dB. The upper bound line is obtained from (43) in Section 5. Different number of users $K = 8$ and $K = 24$ are considered. The upper bound becomes tighter when there are larger number of users.

These results justify our analysis in Section 5.

6.2. Comparison with Other Beamforming Schems. In Figure 3, the achievable user numbers of different beamforming schemes are shown. Four active users are considered with $K = 6$, $L = 4$, and $P = 6$ dB. The performance of zero-forcing DPC (ZF DPC) scheme is compared with orthogonal beamforming (OBF) [17, 18] and zero-forcing beamforming (ZFBF) schemes with dynamic power assignment. In ZFBF scheme with greedy scheduling, we select the first four users with largest channel gains. As expected, the zero-forcing DPC scheme outperforms both OBF and ZFBF schemes.

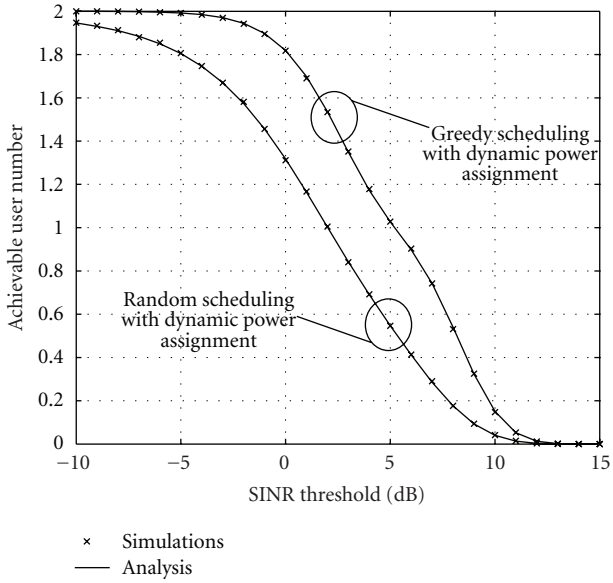


Figure 1: Achievable user number versus SINR threshold, with $K = 4$, $L = 2$, $N = 2$, and $P = 3$ dB.

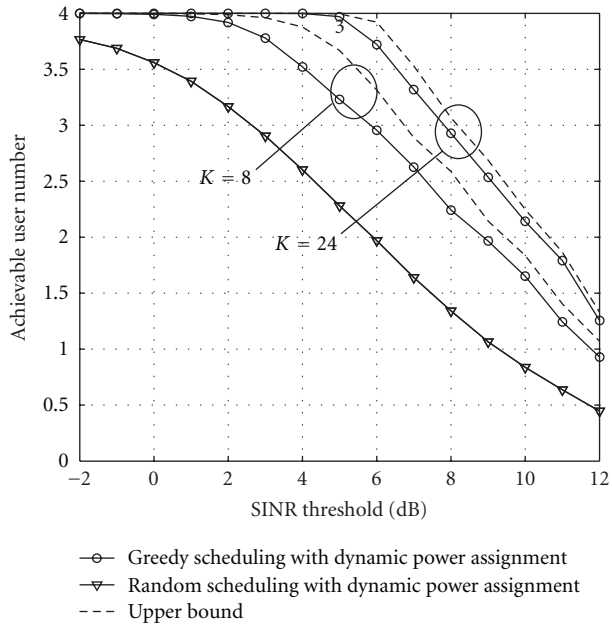


Figure 2: Achievable user number versus SINR threshold, with $L = 4$, $N = 4$, and $P = 6$ dB.

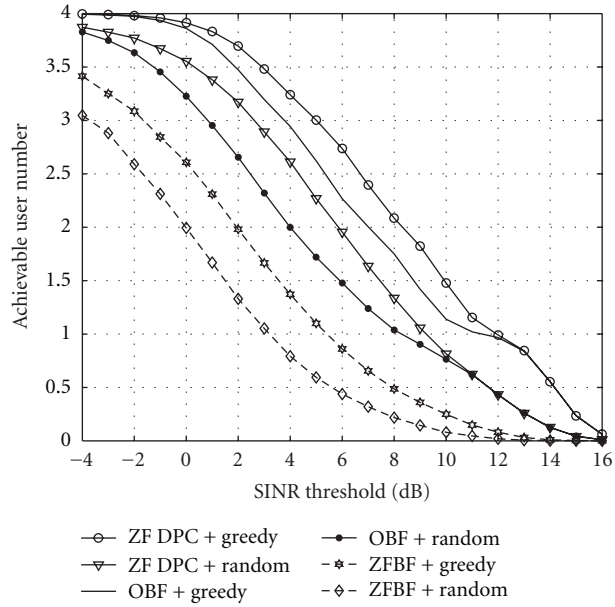


Figure 3: Achievable user number of different beamforming schemes, with $K = 6$, $L = 4$, $N = 4$, and $P = 6$ dB.

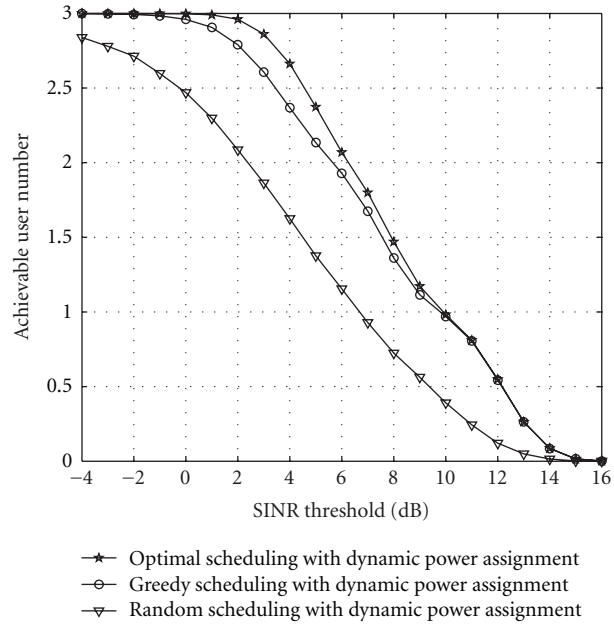


Figure 4: Achievable user number versus SINR threshold, with $K = 6$, $L = 3$, $N = 3$, and $P = 5$ dB.

6.3. Comparison with the Optimal Scheduling. The achievable user numbers of both random and greedy scheduling schemes are presented in Figures 4 and 5 compared with the optimal scheduling scheme. The active user set and DPC encoding order of the optimal scheduling scheme are found by an exhaustive search.

In Figure 4 three active users are considered with $K = 6$, $L = 3$, and $P = 5$ dB. The greedy scheme achieves a significant portion of the user number compared with the optimal scheme. In Figure 5 four active users are considered with $K = 6$, $L = 4$, and $P = 6$ dB. It is shown that the greedy scheme achieve performance close to the optimal scheme.

6.4. Improvement over Equal Power Assignment Schemes. In the equal power assignment schemes, $P_1 = P_2 = \cdots = P_N = (P/N)$.

In Figure 6 the performance of the dynamic power assignment schemes is compared with that of the equal power assignment schemes for an increasing SINR threshold Γ, with $K = 12$, $L = 8$, $N = 8$, and $P = 9$ dB. As expected,

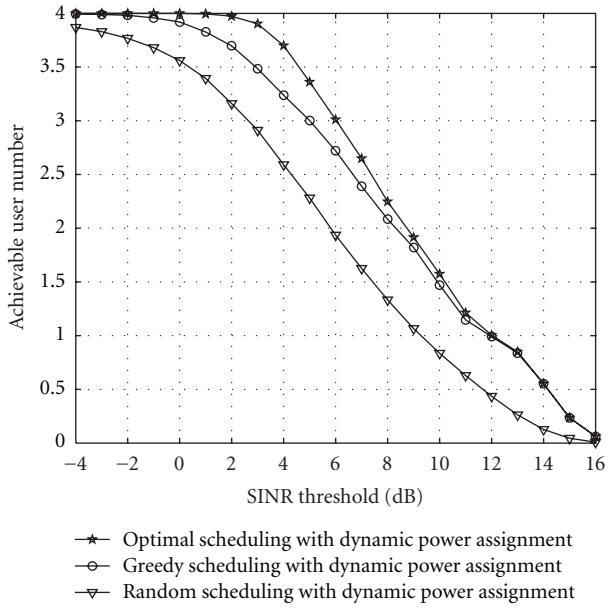


Figure 5: Achievable user number versus SINR threshold, with $K = 6$, $L = 4$, $N = 4$, and $P = 6$ dB.

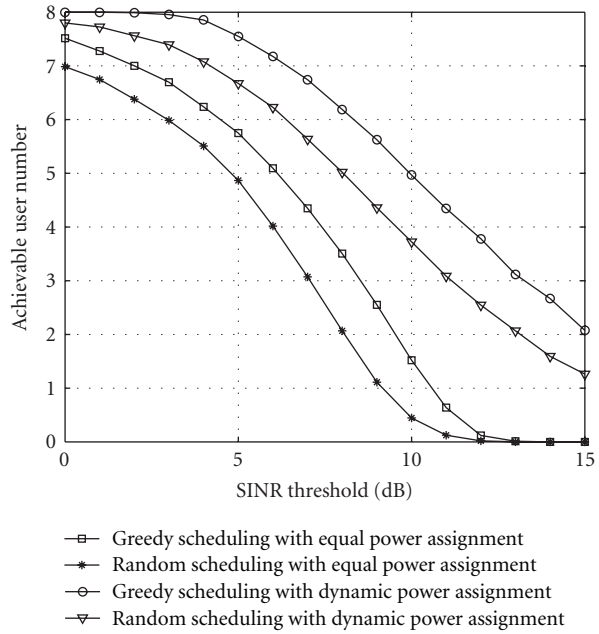


Figure 6: Achievable user number versus SINR threshold, with $K = 12$, $L = 8$, $N = 8$, and $P = 9$ dB.

the dynamic power assignment schemes outperform the equal power assignment schemes. In Figure 6, the achievable user number of greedy scheduling with dynamic power assignment is more than twice of that of greedy scheduling with equal power assignment at SINR threshold 10 dB.

In Figure 7 the performance is compared for different number of transmitter antennas, with $K = 16$, $N = 4$, $P = 6$ dB and $\Gamma = 10$ dB. It is shown that the dynamic power assignment schemes of both greedy and random scheduling need less transmitter antennas than the equal power assignment schemes to achieve the best user number, that is, $N = 4$. The number of antennas for the dynamic power assignment scheme with greedy scheduling to achieve a user number of 4 is $L = 10$, while that for the equal power assignment scheme is $L = 14$.

6.5. Performance Characterizations. The performances of the dynamic power assignment schemes with both greedy and random scheduling are investigated.

In Figure 8, the performance is shown for an increasing number of antennas L, with $K = 16$, $P = 6$ dB and $\Gamma = 10$ dB. The number of scheduled users is not restricted. Both of the schemes enjoy a linear increase as the number of transmitter antennas increases.

The achievable user number is shown in Figure 9 versus number of scheduled users, with $K = 16$, $L = 16$, $P = 10$ dB and $\Gamma = 10$ dB. The achievable user number of both schemes is saturated when $N \geq 11$ due to total power constraint, while that of the greedy scheme is higher than the random scheme.

The achievable user number is plotted in Figure 10 versus total number of users, with $L = 16$, $P = 12$ dB and $\Gamma = 10$ dB. As expected, the random scheme can not benefit from multiuser diversity of the system. It can be seen that the greedy scheduling scheme achieves a diversity gain of about 25% when $K = 60$ over the random scheme.

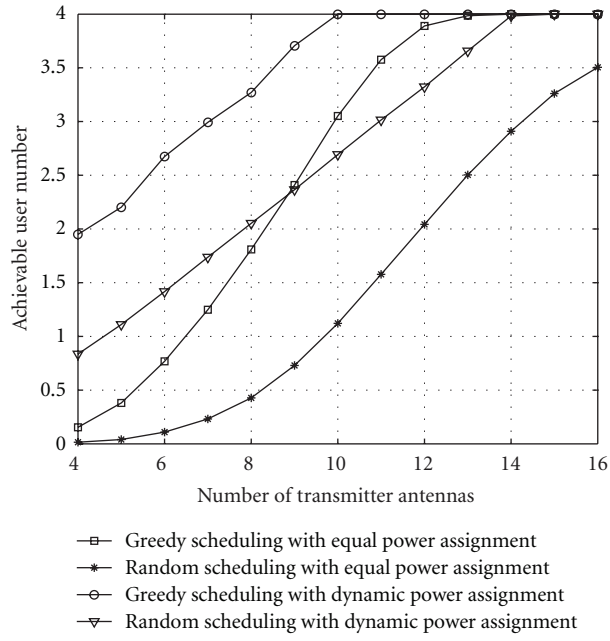


Figure 7: Achievable user number versus number of transmitter antennas, with $K = 16$, $N = 4$, $P = 6$ dB and $\Gamma = 10$ dB.

7. Discussion

In an M2M network, large number of MTC users challenge the admission capability of the communication system. However, the data rate required by each user may be small.

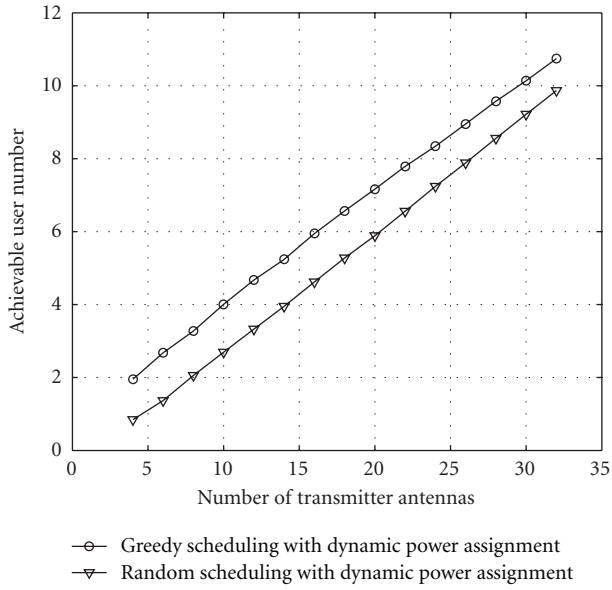


Figure 8: Achievable user number versus number of transmitter antennas, with $K = 16$, $P = 6$ dB and $\Gamma = 10$ dB. The number of scheduled users is not restricted.

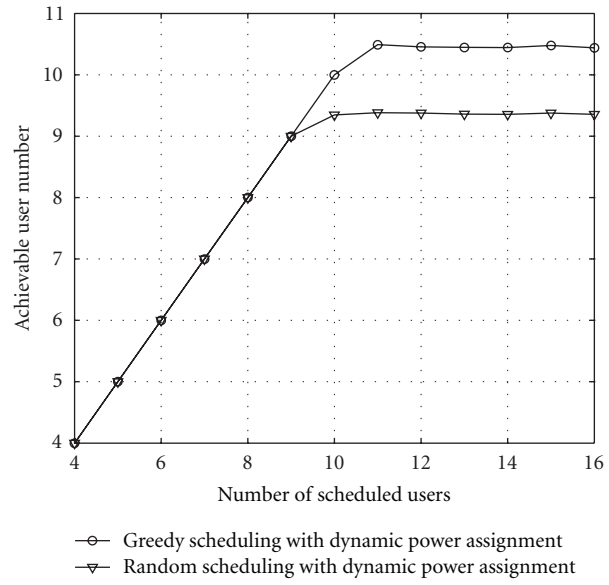


Figure 9: Achievable user number versus number of scheduled users, with $K = 16$, $L = 16$, $P = 10$ dB and $\Gamma = 10$ dB.

Thus, small portion of the system's bandwidth will be allocated to each user, and different users can be allocated with different carriers. In a multi-carrier scenario, the algorithm proposed in this paper can be incorporated with the subcarrier allocation process to maximize the achievable user number. During the subcarrier allocation process, each user is assigned with one subcarrier. After that, the beamforming weight vectors can be calculated by the proposed algorithm in Section 4.

Denote the number of carriers as Q. After the subcarrier allocation process, subcarrier q is allocated with a candidate

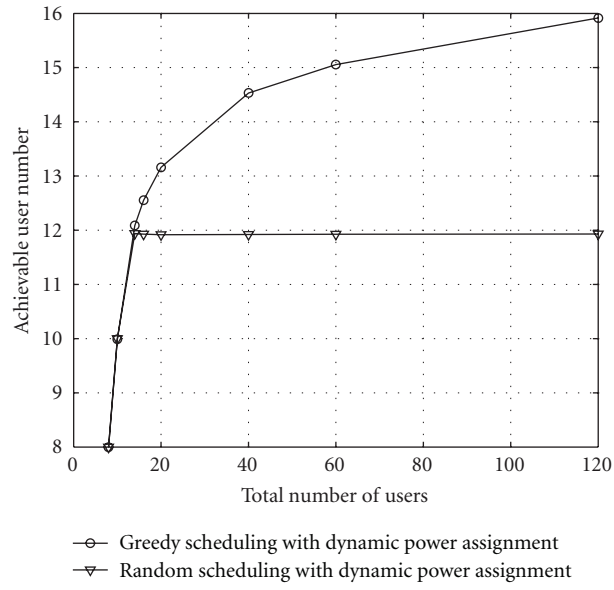


Figure 10: Achievable user number versus total number of users, with $L = 16$, $P = 12$ dB and $\Gamma = 10$ dB.

set of users $\mathcal{K}_q$. The achievable user number of subcarrier q with candidate user set $\mathcal{K}_q$ becomes

$$M_q\left(\mathcal{K}_q\right). \tag{44}$$

The maximum achievable user number can be found by solving the following problem

$$\begin{aligned} &\underset{\mathcal{K}_1,\ldots,\mathcal{K}_Q}{\text{maximize}} \quad \sum_{q=1}^{Q} M_q\left(\mathcal{K}_q\right), \\ &\text{subject to} \quad \bigcup_q \mathcal{K}_q = \mathcal{K}. \end{aligned} \tag{45}$$

The optimal subcarrier allocation remains an open question. However some other suboptimal schemes can be considered. It is noted that the greedy method adopted in the user scheduling can also be used in the subcarrier allocation process. The achievable user number of each subcarrier can be calculated one after another.

8. Conclusion

In order to characterize a system's ability of user admission, we considered the achievable user number, that is, the number of users whose SINRs exceed a threshold. The downlink beamforming using zero-forcing DPC was considered. The algorithms for maximizing the achievable user number were proposed, and the achievable user number of both random and greedy scheduling schemes were derived using the joint distribution of effective channel gains. An upper bound on the achievable user number of the greedy user scheduling scheme was derived. Various numerical results were presented. It was shown that the upper bound becomes tighter when there are larger number of users. The performance

of the greedy scheduling scheme is close to the optimal scheduling scheme. Achievable user number was shown to be one useful metric in understanding the performance of a system, especially in M2M communications, where large number of users challenge the user admission.

Appendices

A.

Proof of Lemma 3.

$$x = \mathbf{h}^H \mathbf{Q}\mathbf{h} = \mathbf{h}^H \left(\mathbf{I} - \mathbf{q}\left(\mathbf{q}^H\mathbf{q}\right)^{-1}\mathbf{q}^H\right)\mathbf{h} = \mathbf{h}^H\left(\mathbf{I} - \frac{\mathbf{q}\mathbf{q}^H}{||\mathbf{q}||^2}\right)\mathbf{h}. \tag{A.1}$$

It is easy to show that $x \le y$. Define $z = x/y$.

$$z = \frac{\mathbf{h}^H\left(\mathbf{I} - \left(\mathbf{q}\mathbf{q}^H/||\mathbf{q}||^2\right)\right)\mathbf{h}}{\mathbf{h}^H\mathbf{h}} = 1 - \frac{|\mathbf{h}^H\mathbf{q}|^2}{||\mathbf{h}||^2||\mathbf{q}||^2}. \tag{A.2}$$

Denote $\eta = |\mathbf{h}^H\mathbf{q}|^2/(\|\mathbf{h}\|^2\|\mathbf{q}\|^2)$. From Lemma B.1 given in Appendix B, it can be shown that η is beta-distributed with the parameters 1 and $L-1$:

$$\eta \sim \text{ Beta }(1,\ L-1) = (L-1)(1-\eta)^{L-2}. \tag{A.3}$$

Thus, the PDF of z is given by

$$f(z) = (L-1)z^{L-2}, \tag{A.4}$$

and the CDF

$$F(z) = z^{L-1}. \tag{A.5}$$

Then, the CDF of X, given Y, is given by

$$F_{X|Y}(\mathrm{x} \mid \mathrm{y}) = \begin{cases} \left(\frac{x}{y}\right)^{L-1}, & \text{for } x \in [0, y], \\ 0, & \text{elsewhere.} \end{cases} \tag{A.6}$$

□

B.

Lemma B.1. *Let* $\mathbf{h}$ *and* $\mathbf{q}$ *denote independent L-dimensional random column vectors with i.i.d., CSCG entries with zero mean and unit variance. Then, we have*

$$\cos^2\theta = \frac{|\mathbf{h}^H\mathbf{q}|^2}{||\mathbf{h}||^2||\mathbf{q}||^2} \sim \text{ Beta }(1, L-1), \tag{B.1}$$

where θ *stands for the angle between* $\mathbf{h}$ *and* $\mathbf{q}$*, and* Beta (α, β) *represents the Beta distribution.*

Proof. The Lemma is obtained by generalizing [19, Theorem 1.5.7(ii)] from $\mathcal{R}^L$ to $\mathcal{C}^L$. □

Acknowledgment

This work has been supported by the National Basic Research Program of China (973 Program, No. 2009CB320403).

References

[1] J. Zhang, L. Shan, H. Hu, and Y. Yang, "Mobile cellular networks and wireless sensor networks: toward convergence," *IEEE Communications Magazine*, vol. 50, no. 3, pp. 164–169, 2012.

[2] S. Y. Lien, K. C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 66–74, 2011.

[3] S.-Y. Lien, T.-H. Liau, C.-Y. Kao, and K.-C. Chen, "Cooperative access class barring for machine-to-machine communications," *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 27–32, 2012.

[4] E. Telatar, "Capacity of multi-antenna gaussian channels," Tech. Rep., Bell Laboratories, 1995.

[5] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[6] N. Jindal, W. Rhee, S. Vishwanath, S. A. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1570–1580, 2005.

[7] W. Yu, "Sum-capacity computation for the Gaussian vector broadcast channel via dual decomposition," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 754–759, 2006.

[8] Z. Tu and R. S. Blum, "Multiuser diversity for a dirty paper approach," *IEEE Communications Letters*, vol. 7, no. 8, pp. 370–372, 2003.

[9] G. Dimić and N. D. Sidiropoulos, "On downlink beamforming with greedy user selection: performance analysis and a simple new algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3857–3868, 2005.

[10] H. Viswanathan, S. Venkatesan, and H. Huang, "Downlink capacity evaluation of cellular networks with known-interference cancellation," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 802–811, 2003.

[11] S. Boppana and J. M. Shea, "Downlink user capacity of cellular systems: TDMA vs dirty paper coding," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT '06)*, pp. 754–758, July 2006.

[12] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2658–2668, 2003.

[13] L.-N. Tran, M. Juntti, M. Bengtsson, and B. Ottersten, "Beamformer designs for zero-forcing dirty paper coding," in *Proceedings of the International Conference on Wireless Communications and Signal Processing (WCSP '11)*, pp. 1–5, November 2011, 2011.

[14] J. Dai, C. Chang, Z. Ye, and Y. S. Hung, "An efficient greedy scheduler for zero-forcing dirty-paper coding," *IEEE Transactions on Communications*, vol. 57, no. 7, pp. 1939–1943, 2009.

[15] K. Huang, J. G. Andrews, and R. W. Heath, "Performance of orthogonal beamforming for SDMA with limited feedback," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 1, pp. 152–164, 2009.

[16] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, 2006.

[17] R. De Francisco, M. Kountouris, D. T. M. Slock, and D. Gesbert, "Orthogonal linear beamforming in MIMO broadcast channels," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '07)*, pp. 1210–1215, March 2007.

[18] S. Ozyurt and M. Torlak, "Performance analysis of orthogonal beamforming with user selection in MIMO broadcast channels," in *Proceedings of the 54th Annual IEEE Global Telecommunications Conference Energizing Global Communications (GLOBECOM '11)*, pp. 1–5, Houston, Tex, USA, 2011.

[19] R. Muirhead, *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, 1982.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 387192, 8 pages
doi:10.1155/2012/387192

Research Article

Distributed Energy-Efficient Topology Control Algorithm in Home M2M Networks

Chao-Yang Lee and Chu-Sing Yang

Institute of Computer and Communication Engineering, Department of Electrical Engineering, National Cheng Kung University, Tainan 701, Taiwan

Correspondence should be addressed to Chao-Yang Lee, d9072220@hotmail.com

Received 19 April 2012; Revised 7 August 2012; Accepted 7 August 2012

Academic Editor: Zhong Fan

Because machine-to-machine (M2M) technology enables machines to communicate with each other without human intervention, it could play a big role in sensor network systems. Through wireless sensor network (WSN) gateways, various information can be collected by sensors for M2M systems. For home M2M networks, this study proposes a distributed energy-efficient topology control algorithm for both topology construction and topology maintenance. Topology control is an effective method of enhancing energy efficiency and prolonging network lifetime. During the topology construction phase, each node builds a new topology based on the best parent node. Nodes can extend the network lifetime by adjusting transmission power to control the topology. In the topology maintenance phase, each node monitors the energy status of neighbors and triggers topology construction as needed. By balancing energy consumption, topology maintenance also avoids the hot spot problem. Simulation results confirm the superior performance of the proposed algorithm for distributed energy-efficient topology control (DETC) in terms of energy efficiency and network lifetime in home M2M networks.

1. Introduction

Home networks are rapidly developing to include diverse embedded devices, including mobile phones, laptops, digital TV, electronic appliances, and so forth. Recently, home networks have begun to shift from machine-to-human communications to machine-to-machine communications in home environments. The M2M network is now the dominant communication technology in home networks [1]. The M2M environments can be conceptualized as networks that provide data communications between embedded devices without human intervention. Wireless sensors are now the main drivers of M2M communication and are being deployed in various systems, including smart grid, healthcare monitoring, smart lighting control, home automation, and surveillance systems [2]. A typical application of an M2M network is the WSN, which enables the direct exchange of information between machines [3]. By interfacing with WSNs, various kinds of information can be collected by M2M network sensors. M2M networks inherit resource-limited, unguarded and mass-deployed sensor networks but are also characterized by embedded intelligence and self-organization capabilities [4]. Different applications may require different sensor capabilities. Thus, the home M2M network may need different sensors for different application requirements. Due to different sensor capabilities in M2M networks, frequent and abrupt topology change is a common problem. Moreover, the natural limitations of WSNs [5], such as their resource constraints, are unsolved issues in M2M communication. Given the above considerations, M2M network protocols and algorithms require energy-efficient and self-organizing capabilities. A potential solution for addressing these challenges is a topology control algorithm that prolongs the M2M lifetime and transforms the network topology in real time.

Effective topology control is essential for prolonged network lifetime and can optimize the tradeoff between power consumption and real-time performance [6]. In this study, we proposed a DETC algorithm in M2M networks. The DETC algorithm for M2M networks proposed in this study

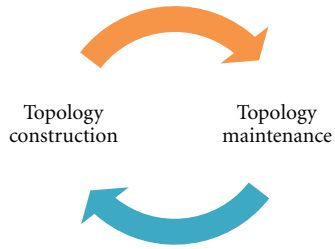


Figure 1: The topology control cycle.

is designed to control the graph representing communication links between nodes in order maximize network lifetime and minimize energy consumption. According to [7], Figure 1 shows the iterative process used for topology control. It consists of two phases: topology construction and topology maintenance. The main motivation behind the topology construction phase is building an energy-efficient topology. Transmission power is known to be the main factor in the energy consumption of each node. By adjusting its own transmission power, each node changes the topology to save energy by choosing the parent node with which it directly communicates. As soon as the topology construction phase establishes the network, the topology maintenance phase is initiated in each node. The M2M network nodes may unexpectedly move or turn power on/off. Moreover, energy consumption should be evenly distributed. The parent of each node must change at the appropriate time. Thus, a topology-maintenance algorithm monitors the topology status and triggers a topology construction phase as needed. This cycle is repeated many times over the network lifetime until all network energy is depleted.

The rest of this paper is organized as follows. Section 2 introduces relevant works. Next, Section 3 elucidates the proposed topology control algorithm. Section 4 summarizes the simulation results. Section 5 draws conclusions.

2. Related Work

Machine-to-machine technology supports wireless communications from machine to machine. This emerging technology for next-generation communications has many potential applications for improving efficiency and reducing network operating costs. Tekbiyik and Uysal-Biyikolglu [8] comprehensively studied shortest path-based energy-efficient routing alternatives for M2M networks. They also developed a classification scheme for current and future applications and the design issues of M2M networks. The evolution of M2M has recently included development of a smart grid framework. Fadlullah et al. [9] surveyed several existing communication technologies with potential smart grid applications for M2M communication. They also presented a reliable technology for improving the performance of conventional ZigBee-based M2M communications in smart grids by incorporating intelligence in a smart meter and M2M devices. Niyato et al. [10] investigated the network design issue of M2M communications for a home-energy management system (HEMS) in the smart grid. A dynamic programming algorithm is applied for optimizing the HEMS traffic concentration while minimizing cost. Based on the findings of the above studies, several researchers have designed schemes for M2M communication technologies in several applications. Energy efficiency is still a major design objective in the machine-to-machine network scenario. However, few studies have discussed the important issue of improving power efficiency in M2M networks.

Topology control is a key technique for improving energy efficiency and is applicable in M2M communication. Ding et al. [11] investigated an adaptive partitioning scheme called connectivity-based partition approach (CPA) for node scheduling and topology control in WSNs. The CPA classifies nodes according to their measured connectivity instead of estimating connectivity based on their positions. It maintains K-vertex connectivity of the backbone, which balances the trade-off between energy efficiency and communication quality. Sethu and Gerety [12] developed step topology control for reducing energy consumption while preserving connectivity in heterogeneous sensor networks without location-based assumptions. Their proposed control is also applicable for nonuniform path loss propagation. Rizvi et al. [13] presented a connected dominating set-based topology control algorithm, A1, which constructs an energy-efficient virtual backbone. By using an efficient algorithm, the topology construction phase of A1 substantially reduces the messages required for an efficient algorithm. Connectivity during topology maintenance and sensing coverage are also enhanced. Chiwewe and Hancke [14] developed the Smart Boundary Yao Gabriel Graph (SBYaoGG) topology control algorithm for increasing energy efficiency and decreasing interference in wireless sensor networks. Use of the SBYaoGG algorithm for optimization ensures symmetric and energy efficient links throughout the network. Üster and Lin [15] devised a hierarchical topology and routing structure with multiple sinks and proposed a general data aggregation approach. An integrated topology control and routing problem was used to improve energy efficiency and prolong lifetimes in data-gathering WSNs. For reduced power consumption, Ma et al. [16] constructed network topologies that minimize the number of coordinators without affecting] network connectivity. Reducing the number of coordinators reduces the average duty cycle and prolongs battery life.

Based on the above observations, most topology control algorithms focus only on topology construction. Numerous researchers have attempted to optimize topology based on a specific property such as reliability, energy efficiency, or connectivity. However, topology maintenance should also be considered when designing topology control algorithms. Because home M2M network are deployed in numerous applications, many different sensor types are needed to satisfy application requirements. To overcome this problem, this study proposes a novel topology control algorithm that prolongs system lifetime by considering both topology construction and topology management. The topology construction phase builds a topology by identifying the OR builds a topology based on the parent node with which it directly communicates. By adjusting the transmission power level and choosing the parent node based on a fair

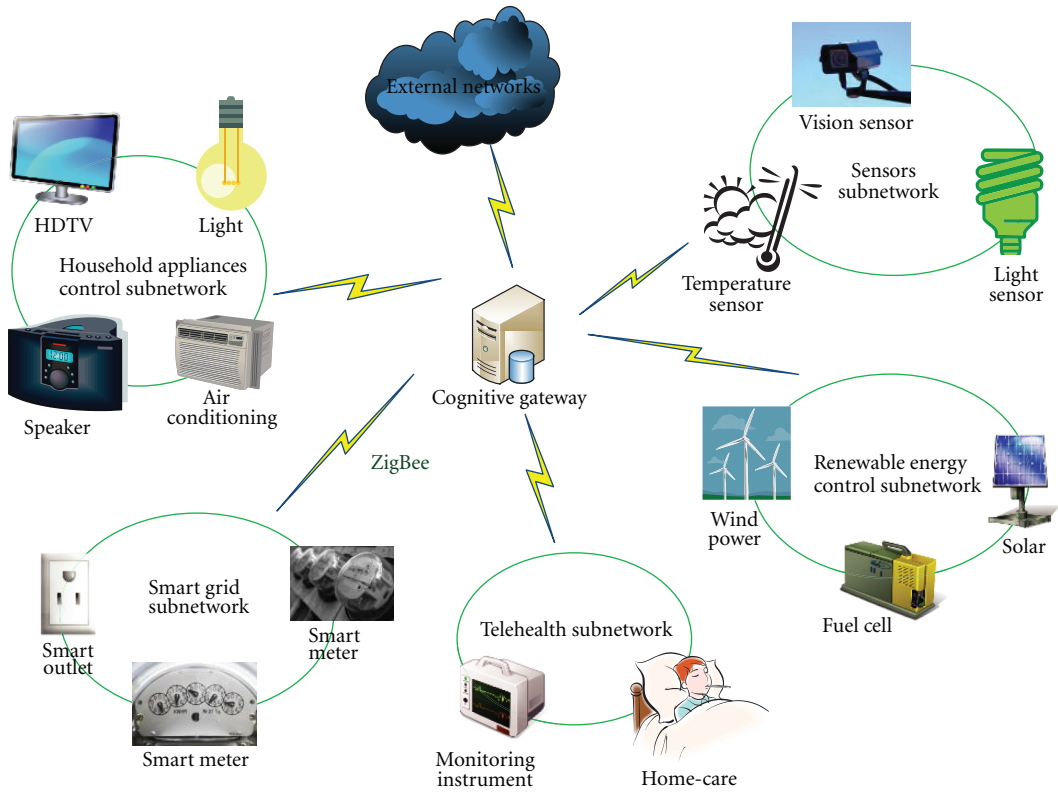


Figure 2: Home M2M sensor network architecture.

energy consumption rule, each node can change the topology to increase energy savings. Each node performs the topology maintenance algorithm to monitor the status of the neighbor and to trigger a topology construction phase that enables a rapid response when needed. By enabling nodes to create and maintain energy-efficient links autonomously, the proposed distributed algorithm scheme achieves the objective of prolonged lifetime.

3. Distributed Energy-Efficient Topology Control Algorithm

This section first defines some assumptions before giving details of the topology control algorithm and its energy efficiency.

3.1. Network Model. The M2M promises diverse services for home environments. Some exemplary applications include smart grid, mobile healthcare, home automation, security, appliance monitoring, lighting, and sensors. The home M2M network includes an external network and multiple subnetworks. The external network includes a central machine home gateway for managing the overall network, collecting data from the overall network, and connecting the home network to the outside world (e.g., the Internet). Each subnetwork is self-organized and is designed for a specific application. However, each network communicates by ZigBee [17] or 802.15.4 [18] standard. Since the IEEE 802.15.4 protocol is designed to minimize energy-efficiency and communication overheads, it is a promising standard for WSN applications. Based on the physical and medium-access control layers of IEEE 802.15.4, the upper layers are defined by the ZigBee protocol stack.

Notably, the sensors in several subnetworks have varying capabilities, as shown in Figure 2. For instance, sensors in the smart grid subnetwork have powerful computing capability and an unlimited power supply. In contrast, some sensors in the household appliances control subnetwork have unlimited energy but may suddenly turn off when the user turns off the household appliance. Sensors in the subnetwork are resource constrained in terms of power supply, computation capacity, and bandwidth. Thus, energy-efficient algorithms could help prolong lifetime in home M2M sensor networks.

This study assumes N identical sensor nodes are distributed randomly in the monitor area. The graph of the network is denoted as $G(V, E)$. Let V be the set of vertices representing the nodes and let E be the set of edges representing the communication links. The distance between node u and v is denoted as $d(u, v)$. Each sensor node can adjust the transmission power to communicate with any other nodes. Here, the widely used transmission power model is applied as follows:

$$\varepsilon(u, v) = \xi \times d(u, v)^{\alpha}, \tag{1}$$

where $\varepsilon(u, v)$ is the required transmission power from u to v, α is the path loss factor between 2 and 4, and a power threshold ξ is considered a constant.

Each sensor can send and receive hello messages with maximum power to find their neighbor nodes $N(v)$. Here, $\varepsilon_{\min}(u, v)$ denotes minimum power needed to deliver a packet from the node u to v. According to the $N(v)$ and $\varepsilon_{\min}(u, v)$, node v can construct a local subgraphic. The $\varepsilon_{\mathrm{rm}}(v)$ is the remaining energy of node v.

3.2. Energy-Efficient Topology Construction Scheme. Topology control adapts to network changes in order to enhance power savings and to prolong network lifetime. In the topology construction phase, an energy-efficient topological property that constructs topology dynamically is established in the network while maintaining connectivity. The monitor area has several sensors with different capabilities. An energy-efficient topology that corresponds with the sensor state must be built. A backbone is formed to enable node pairs in the topology to communicate. A distributed energy-efficient algorithm is proposed for building the topology. Each sensor constructs a local subgraphic $G_v \leftarrow (N(v), L(v))$ and determinates the best parent node $P(v)$. The $P(v)$ is the parent node of node v. The $L(v)$ is a set of links in the subgraphic. Thus, a tree-based backbone for data transmission can be built. Next, the topology control for home M2M networks is described.

According to the above discussion, the sensors in the home M2M network can be classified into four types: (1) The sensors have limited battery power and are always active. (2) The sensors have limited battery power and appear/disappear suddenly. (3) The sensors have unlimited power supply and always alive. (4) The sensors have unlimited power supply and appear/disappear suddenly. Some nodes may that have mobile capability are classified as type 2 or 4. Based on the different sensor characteristics, a new energy-efficient topology must be built to prolong the system lifetime and to satisfy the application requirements in H2M network. Algorithm 1 shows the proposed energy-efficient topology construction algorithm. Each node first constructs a local subgraphic (line 1). The initialization then sets the parent node of node v to NULL (line 2). Candidate set $C(v)$ must then be identified so that the child node can be removed from the neighbor node to avoid the routing loop (line 3). The sink(x) denotes the sink state, that is, whether or not the sink is connected. The objective of the process is to build an energy-efficient topology. Thus, if node u near node $v$has unlimited energy, node v can immediately apply node u at parent node $P(v)$. In lines 4 to 7, node v must find the neighbor with unlimited energy to set the parent node $P(v)$. Transmission power $\varepsilon_{\mathrm{tx}}(v)$ of node v is assumed to adjustable. If so, transmission power $\varepsilon_{\mathrm{tx}}(v)$ is set to the minimum power cost $\varepsilon_{\min}(P(v), v)$ between $P(v)$ and v. If not, the node v must try to find $P(v)$ according to transmission energy cost. The transmission power level $\varepsilon_{\mathrm{pl}}(v)$ for each node in the overall network could be the maximum or between the minimum and maximum. A connected tree topology of the overall network is formed among different nodes using an effective transmission power level $\varepsilon_{\mathrm{tx}}(v)$ such as $0 \le \varepsilon_{\mathrm{pl}}(v) \le n$. Based on analysis of sensor characteristics above, two classifications are possible based on energy considerations. First, if sensor v has unlimited energy supply, the transmission power level in a decrease order (lines 9-10). If possible, a sensor with unlimited energy can deliver packets with maximum power consumption. Second, sensor v has limited battery power, transmission power level in an increase order (lines 14-15). The shorter distance decreases transmission power cost and prolongs lifetime. For a given transmission power level, node v finds the node u with the maximum remaining energy and the shortest path (lines 11–16). The shorter path refers to the hop count HC(u), which is larger than or equal to HC(v). The HC(v) refers to the hop count from sink to v. Transmission power $\varepsilon_{\mathrm{tx}}(v)$ is then set to the minimum power cost $\varepsilon_{\min}(P(v), v)$ (lines 12–17). If no parent nodes have a shorter path from candidate set $C(v)$, node u, which has the maximum remaining energy, is set as the parent node $P(v)$ (lines 21-22). Finally, the *topology maintenance* algorithm is performed (line 24).

3.3. Energy-Efficient Topology Maintenance Scheme. Given a certain number of nodes distributed in a monitor area, the goal of topology construction is building a topology that enhances energy efficiency. Once the topology has been created, the network starts performing the topology maintenance task. Depending on its duration, maintenance can change the topology. According to [7], topology maintenance is defined as the process that periodically restores, rotates, or recreates the network topology when the current reduced topology is no longer optimal. Thus, the topology maintenance mechanism in this study is triggered to build a new topology, to increase the energy efficiency of the nodes, and to increase network lifetime. Because, many applications require a many-to-one traffic pattern in the M2M network, packet delivery may cause an energy imbalance in the nodes in the hot spot near the sink [19]. All packets must be routed through the hot spots around the sink. Because nodes in this hot spot must forward a disproportionately large amount of traffic, they often die at a very early stage. The hot-spot problem is the major limitation on lifetime. Further, numerous unlimited energy supplies and sets of sensors may appear/disappear unexpectedly in the M2M. Therefore, balanced and efficient energy consumption in each node is an important issue.

The dynamic topology maintenance techniques applied in this study create a new topology "on the fly" by triggering the topology construction mechanism as needed [7]. Dynamic topology maintenance mechanisms are advantageous because they use more network information to optimize the energy efficiency of the topology. The topology maintenance process has been described, as shown in Algorithm 2.

During the topology maintenance phase, node v periodically broadcasts a beacon message to its neighbor set $N(v)$. Similarly, each node can receive beacon messages from neighbors (lines 1-2). The beacon packet carries the following information in its header: <*Src*, *EnergyState*, *Parent_Node*, *Sink_State*, *Hop_Count*>. The *Src* is used to identify each packet; *EnergyState* refers to the remaining energy of node; *Parent_Node* records the parent node; *Sink_State* is the index

Function **Topology Construction Algorithm** at node (v) in G:
(1) $G_v \leftarrow (N(v), L(v))$; where $L(v) \in E$
(2) $P(v) \leftarrow \varnothing$
(3) $C(v) = \{xx \in N(v), P(x) \neq v, \text{sink}(x) = \text{true}\}$
(4) If $\{\varepsilon_{\text{rm}}(u) | u \in C(v)\} == \infty$
(5) $P(v) \leftarrow u$
(6) $\varepsilon_{\text{tx}}(v) = \varepsilon_{\min}(P(v), v)$
(7) Break;
(8) else
(9) If $\varepsilon_{\text{rm}}(v) == \infty$
(10) For $\varepsilon_{\text{pl}}(v)$ in decrease order
(11) $P(v) = \text{Max}_u\{\varepsilon_{\text{rm}}(u) | u \in C(v), \text{HC}(u) \leq \text{HC}(v)\}$
(12) $\varepsilon_{\text{tx}}(v) = \varepsilon_{\min}(P(v), v)$
(13) End for
(14) else
(15) For $\varepsilon_{\text{tx}}(v)$ in increase order
(16) $P(v) = \arg\max\{\varepsilon_{\text{rm}}(u) | u \in C(v), \text{HC}(u) \leq \text{HC}(v)\}$
(17) $\varepsilon_{\text{tx}}(v) = \varepsilon_{\min}(P(v), v)$
(18) End For
(19) End If
(20) End if
(21) If $P(v) == \varnothing$
(22) $P(v) = \text{Max}_u\{\varepsilon_{\text{rm}}(u) | u \in C(v)\}$
(23) End If
(24) Trigger *topology maintenance Algorithm*

Algorithm 1: Topology contraction algorithm running on a node v in G.

Function **Topology Maintenance Algorithm** at node (v):
(1) Broadcast beacon message to neighbor $N(v)$
(2) Receive beacon message from each neighbor $N(v)$
(3) $C(v) = \{xx \in N(v), P(x) \neq v, \text{sink}(x) = \text{true}\}$
(4) If $P(v)$ is miss
(5) Trigger *Topology Construction Algorithm*
(6) End If
(7) If $\varepsilon_{\text{rm}}(C(v)) == \infty$
(8) Trigger *Topology Construction Algorithm*
(9) End If
(10) If $\varepsilon_{\text{rm}}(P(v)) < \varepsilon_{\text{rm}}(C(v))$
(11) Trigger *Topology Construction Algorithm*
(12) End If

Algorithm 2: Topology maintenance algorithm running on a node v in G.

which indicating whether or not the node is connected to the sink; *Hop_Count* is the number of wireless links in a path from the sink to the node. Each can collect information about neighbors via beacon packets. The objective is to obtain the candidate set $C(v)$ to avoid the routing loop (line 3). The sensor characteristics observed in the home M2M sensor network indicate that some sensors may appear/disappear suddenly. Thus, three decision rules are used to execute the topology construction algorithm needed to build a new topology in each node. The first decision rule shows that $P(v)$ is miss (line 4). A missing parent node $P(v)$ indicates that v cannot connect to the sink, so a new topology must be built. Next, if an unlimited energy node is activated and joins the network, it should be utilized efficiently (line 7). Node v should try to find a new parent node. Finally, the parent node of v should change when v finds a better node that has more remaining energy. Each node must consume energy equally. When the remaining energy of the parent node is lower than that of the node in the candidate set, the parent node must change. In summary, the topology can vary depending on the time via topology maintenance. During the topology maintenance process, each node uses these three decision rules to trigger the topology construction algorithm. By equally distributing energy consumption, it can prolong the system lifetime.

4. Performance Evaluation

This section presents some simulation results that demonstrate the effectiveness of the proposed DETC algorithm. We evaluate the performance of DETC and compare it with EELTC [16] algorithm and A1 algorithm [13]. A random-node deployment in a 100 m × 100 m area is simulated. The sink is set at the lower left corner. The measurement unit is the time required to send 250 k bits of information to the sink over a multihop link. The communication radius R_c and initial energy level of each node are set to 10 m and 10 K J, respectively. For a mica mote with RF frequency 866 MHz [20], Table 1 [21] specifies the four transmission power levels in the interval [−13 dBm, 5 dBm] where 0 is the minimum and 3 is the maximum transmission power level for communicating among nodes. The energy consumption

Table 1: Energy consumption for different power levels.

Transmission power level ε_{tx}	Emitted power [dBm]
0	−13
1	−7
2	−1
3	5

model was based on the Berkeley mote, that is, energy costs of 0.075 W transmitting at maximum power, 0.03 W when receiving, 0.025 W when listening, and 0.001 W when sleeping. To obtain the most probable result, the average values for 50 rounds were recorded.

The first experiment measured network lifetime during topology control. Network lifetime was defined as the period from the start of network operation to the moment when the first sensor in the network runs out of energy [22]. As expected, the proposed DETC algorithm achieved a longer system lifetime (rising curve in Figure 3) compared to EELTC and A1. Notably, the curve for the EELTC algorithm was actually a declining curve. In terms of network size, the network lifetime for DETC obtained by the proposed approach is longer than that of EELTC and A1, especially for extremely large networks. Due to the hot-spot problem, packet forwarding may cause energy imbalance. Traffic routed through the nodes near the data sink can create a hot spot around the data sink. Thus, the nodes in this hot spot must forward a disproportionately high number of packets and die at a very early stage. However, the topology maintenance phase in this approach can monitor the energy state of sensors and dynamic changes in topology as needed.

In the M2M network, packet transmission is the main energy consumer. A high packet delivery rate increases the energy consumption of nodes. Figure 4 shows that the proposed algorithm performs well in terms of network lifetime under various packet delivery rates. The DETC algorithm can reconstruct a new topology to increase energy efficiency as needed. Additionally, since the node status may appear or disappear suddenly, the DETC can rapidly build a new topology in part of network. The parent node has selected via various power level control. Thus, this approach can provide the good performance of system lifetime. Figure 5 shows the performance of the network lifetime under various power levels. The total number of power levels can affect the network. A large number of power level increases network lifetime. The simulation results suggest that the selected number of power levels is appropriate since $0 \leq \varepsilon_{pl}(v) \leq 3$.

The topology control approach has two phases, topology construction and topology maintenance. The topology construction phase establishes an energy-efficient topological property in the network while ensuring connectivity. When the new topology is built, nodes trigger the topology maintenance phase. Thus, the DETC is more energy efficient compared to EELTC and A1. Figure 6 displays the results of the simulation of the average remaining energy of nodes after the end of the system. The proposed control balances the energy consumption of each node by updating the parent during the topology maintenance. Thus, for network load,

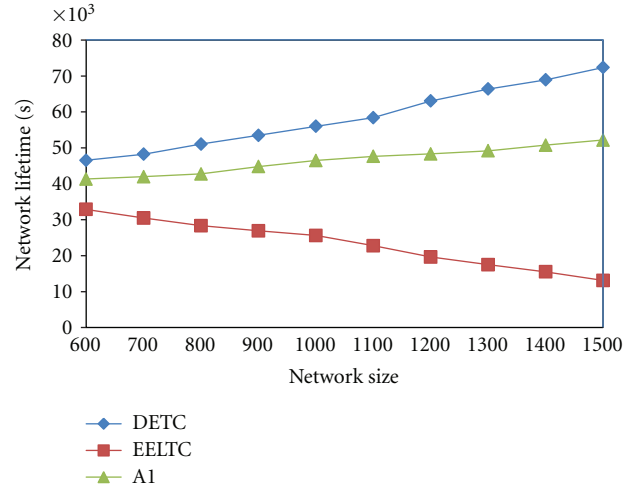


Figure 3: Network lifetime comparison in different network size.

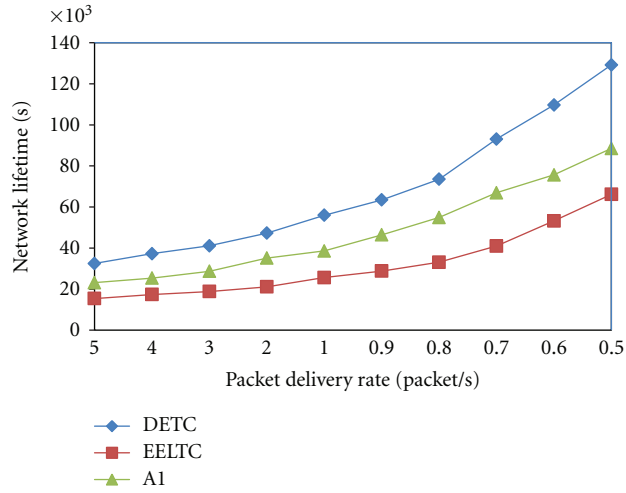


Figure 4: Network lifetime comparison for different packet delivery rate.

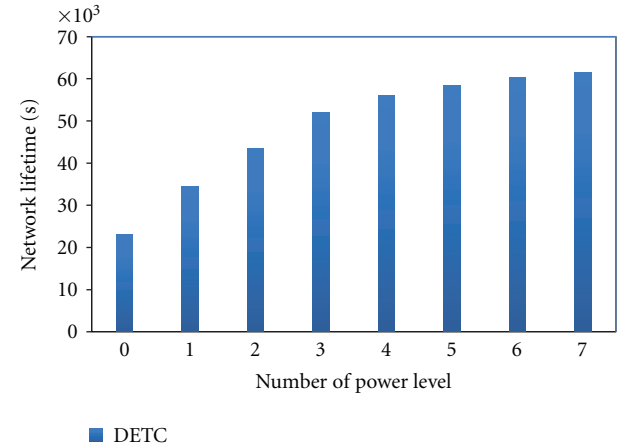


Figure 5: Network lifetime under various number of power level.

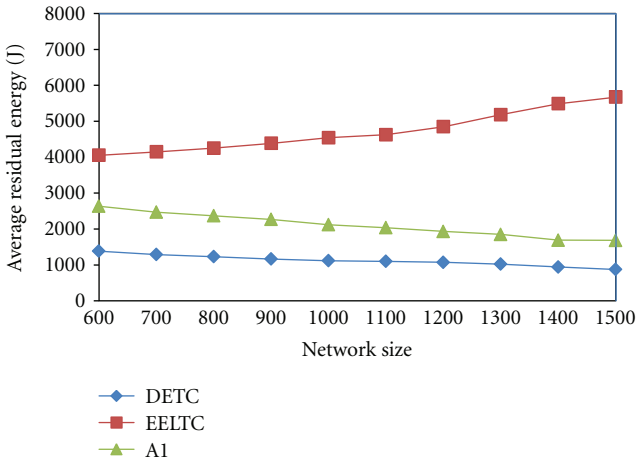


Figure 6: Average remaining energy comparison.

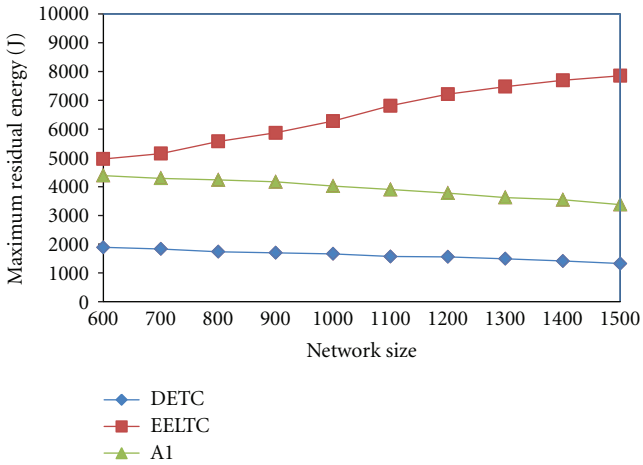


Figure 7: Maximum remaining energy comparison.

the improvement in the average remaining energy of DETC is larger than that of EELTC and A1. This proves that the scheme effectively improves network lifetime. Because the scheme balances the energy consumption of each node in the M2M network, its maximum remaining energy after the end of the system lifetime is low (Figure 7). Balanced power consumption in each node and increased network lifetime are two main benefits of DETC when applying topology maintenance to a topology control. Therefore, topology maintenance is the most important phase in topology control design. However, the topology maintenance phase may require repeated topology reconstruction, which increases energy overhead. Energy overhead is defined as the fraction of the network energy expended on topology construction [23]. In the case of topology maintenance, this metric calculates the overhead during topology reconstruction under appropriate conditions. Figure 8 shows the changes in energy overhead observed in different network sizes. The proposed algorithm obtains a higher energy overhead compared to EELTC because the EELTC is a centralized algorithm and only one topology is constructed. The DETC includes the topology maintenance algorithm, which repeatedly performs topology reconstruction until the end of the system lifetime. The energy overhead of A1 is higher than that of DETC since the topology maintenance process of A1 is triggered when the network energy falls by 10%. Energy overhead increases when the sensors reconstruct a new topology. The proposed distributed algorithm executes the topology construction phase in only one part of the overall network. Therefore, it significantly reduces energy overhead.

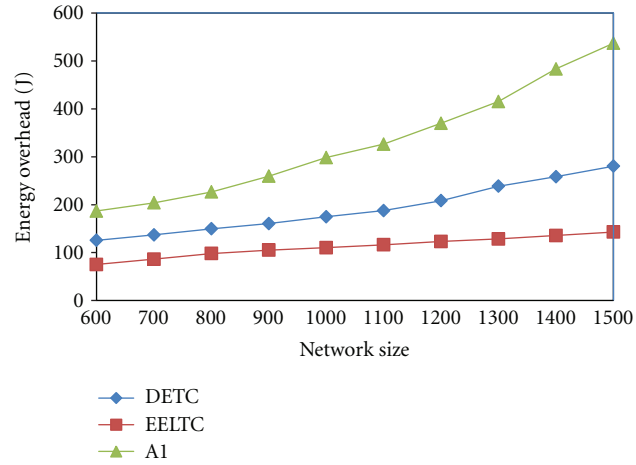


Figure 8: Energy overhead comparison.

5. Conclusions

The M2M network is expected to be among the fastest growing networks in the next several years, and numerous applications are rapidly accumulating. Generally, the wireless sensors are the primary drivers of M2M networks. Therefore, the M2M network needs various sensors to meet the application requirements. This study presented a distributed energy-efficient topology control algorithm for an M2M network. The topology control contains two phases, topology control, and topology maintenance. Unlike conventional algorithms, the control does not focus on topology construction. Once the topology constructed, each node executes a topology maintenance algorithm in place to monitor the status of the neighbor and to trigger a topology construction phase that can rapidly respond as needed. The nodes near the sink node forward traffic between the sink, which causes an energy imbalance. The topology maintenance also avoids the hot spot problem by controlling energy consumption and by rapidly responding when nodes appear/disappear. The topology construction phase builds a topology by identifying the parent node with which it directly communicates. Each node can adjust the transmission power level and choose the parent node based on a fair energy consumption rule, which ensures an energy-efficient topology. The proposed scheme uses a distributed algorithm that enables nodes to create and maintain energy-efficient links autonomously

and to achieve the objective of prolonged lifetime with reduced energy overhead. Simulation results confirm that the proposed DETC algorithm extends network lifetime, efficiently allocates energy consumption and reduces energy overhead.

Acknowledgments

The authors would like to acknowledge the valuable comments and suggestions of the reviewers, which have improved the quality of this paper. This work is supported in part by NCKU Research Center for Energy Technology and Strategy under Grant B101-MD10, Taiwan.

References

[1] Y. Zhang, R. Yu, S. Xie, W. Yao, Y. Xiao, and M. Guizani, "Home M2M networks: architectures, standards, and QoS improvement," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 44–52, 2011.

[2] S. Agarwal, C. Peylo, R. Borgaonkar, and J. P. Seifert, "Operator-based over-the-air M2M wireless sensor network security," in *Proceedings of the 14th International Conference on Intelligence in Next Generation Networks*, pp. 1–5, October 2010.

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

[4] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. D. Johnson, "M2M: from mobile to embedded internet," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 36–43, 2011.

[5] F. Wang and J. Liu, "Networked wireless sensor data collection: issues, challenges, and approaches," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 4, pp. 673–687, 2011.

[6] P. Santi, *Topology Control in Wireless Ad Hoc and Sensor Networks*, John Wiley & Sons, West Sussex, UK, 2005.

[7] M. A. Labrador and P. M. Wightman, *Topology Control in Wireless Sensor Networks with a Companion Simulation Tool for Teaching and Research*, Springer, Berlin, Germany, 2009.

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

[9] Z. M. Fadlullah, M. M. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, and Y. Nozaki, "Toward intelligent machine-to-machine communications in smart grid," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 60–65, 2011.

[10] D. Niyato, L. Xiao, and P. Wang, "Machine-to-machine communications for home energy management system in smart grid," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 53–59, 2011.

[11] Y. Ding, C. Wang, and L. Xiao, "An adaptive partitioning scheme for sleep scheduling and topology control in wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 20, no. 9, pp. 1352–1365, 2009.

[12] H. Sethu and T. Gerety, "A new distributed topology control algorithm for wireless environments with non-uniform path loss and multipath propagation," *Ad Hoc Networks*, vol. 8, no. 3, pp. 280–294, 2010.

[13] S. Rizvi, H. K. Qureshi, S. A. Khayam, V. Rakocevic b, and M. Rajarajan, "A1: an energy efficient topology control algorithm for connected area coverage in wireless sensor networks," *Journal of Network and Computer Applications*, vol. 35, no. 2, pp. 597–605, 2012.

[14] T. M. Chiwewe and G. P. Hancke, "A distributed topology control technique for low interference and energy efficiency in wireless sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 11–19, 2012.

[15] H. Üster and H. Lin, "Integrated topology control and routing in wireless sensor networks for prolonged network lifetime," *Ad Hoc Networks*, vol. 9, no. 5, pp. 835–851, 2011.

[16] J. Ma, M. Gao, Q. Zhang, and L. M. Ni, "Energy-efficient localized topology control algorithms in IEEE 802.15.4-based sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no. 5, pp. 711–720, 2007.

[17] Zigbee Alliance, "Zigbee Specification: Zigbee Document 053474r13 Version 1. 1," December 2006, http://www.zigbee.org/.

[18] IEEE Standard 802. 15.4, *Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications For Low-Rate Wireless Personal Area Networks (LR-WPANs)*, 2006.

[19] Z. Cheng, M. Perillo, and W. B. Heinzelman, "General network lifetime and cost models for evaluating sensor network deployment strategies," *IEEE Transactions on Mobile Computing*, vol. 7, no. 4, pp. 484–497, 2008.

[20] Xbow, "MPR/MIB Mote Hardware User's Manual," December 2003, http://www.xbow.com/.

[21] P. K. Sahoo, J. P. Sheu, and K. Y. Hsieh, "Power control based topology construction for the distributed wireless sensor networks," *Computer Communications*, vol. 30, no. 14-15, pp. 2774–2785, 2007.

[22] H. Wang, N. Agoulmine, M. Ma, and Y. Jin, "Network lifetime optimization in wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 7, pp. 1127–1137, 2010.

[23] H. K. Qureshi, S. Rizvi, M. Saleem, S. A. Khayam, V. Rakocevic, and M. Rajarajan, "Poly: a reliable and energy efficient topology control protocol for wireless sensor networks," *Computer Communications*, vol. 34, no. 10, pp. 1235–1242, 2011.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 843413, 9 pages
doi:10.1155/2012/843413

Research Article

Energy-Efficient Chain Formation Algorithm for Data Gathering in Wireless Sensor Networks

Se-Jung Lim and Myong-Soon Park

Department of Computer and Radio Communication Engineering, Korea University, Seoul 136-701, Republic of Korea

Correspondence should be addressed to Myong-Soon Park, myongsp@korea.ac.kr

Received 20 April 2012; Revised 27 July 2012; Accepted 29 July 2012

Academic Editor: Zhong Fan

In wireless sensor networks, since sensor nodes are distributed in inaccessible regions for data gathering, they need to be operated during an assigned time without battery recharging and relocation. For this reason, there has been abundant research on improving energy efficiency. PEGASIS, one of the well-known chain-based routing protocols for improving energy efficiency, builds a chain based on the greedy algorithm. However, due to long communication distance of some sensor nodes in a chain formed by the greedy algorithm, unbalanced energy consumption of sensor nodes occurs. Eventually, the network lifetime from this cause decreases. We propose energy efficient chain formation (EECF) algorithm to resolve the unbalanced energy consumption problem caused by long-distance data transmission of some nodes in a chain formed by the greedy algorithm. The simulation results are used to verify the energy consumption balance of sensor nodes and the whole network lifetime. In simulation, it is shown that EECF produces better results than the greedy algorithm.

1. Introduction

In wireless sensor networks (WSNs), sensor nodes enabled with sensing, computing, and communication functions gather data (e.g., temperature, humidity, infrared light, sound, shock and pressure, etc.). The data is then transmitted to the sink connected to external network [1]. However, sensor nodes are commonly distributed in inaccessible regions depending on the type of application, and the sink is located far away from sensor nodes. For this reason, sensor nodes with the limited battery resource need to be operated during the assigned time without battery recharge and relocation [2]. If each node transmits its data directly to the sink, some nodes that are far away from the sink will die much earlier than the other sensor nodes. This is as a result of rapid energy depletion due to long distance data transmission. Consequently, this problem limits the use of WSN to gather data in certain regions. This becomes a cause that cannot be done to gather data in certain regions. Thus, a more effective use of energy becomes the major challenge in WSNs [3]. To improve energy efficiency, many researchers have suggested various routing algorithms [4–12].

In these routing algorithms, power-efficient gathering in sensor information systems (PEGASIS) [12] is one of the well-known chain-based routing protocols. In PEGASIS, a chain is formed using the greedy algorithm. One among sensor nodes will be randomly selected as a leader at each round, and sensor nodes transmit data to their neighbor node along a chain toward the leader node. During data gathering, every sensor node except the end nodes in the chain fuses own data and data received from its neighbor node. The leader node typically receives data from both its neighbors and transmits fused data to the sink. Sensor nodes only communicate with their neighbors in the chain, and take turns selecting the leader node, thereby reducing their energy consumption as well as balancing of energy consumption per round. However, PEGASIS using the greedy algorithm still has weakness. The weakness is that since sensor nodes already in the chain cannot be revisited to prevent looping, distance between neighbors increases gradually. As a result, due to long distance data transmission of some sensor nodes, they consume much more energy to transmit fused data to neighbor. Besides, in WSNs, sensor nodes have limited battery resources. Thus, this unbalanced

energy consumption will lead to the lifetime decrease of a network. Hence, sensor nodes need to extend the network lifetime via balanced energy consumption.

In order to achieve this, instead of using the greedy algorithm in PEGASIS, we propose an energy efficient chain formation (EECF) algorithm based on in-order tree traversal algorithm in hierarchical tree (using Buttenfield's [13, 14] strip tree geometry algorithm). In EECF, it is possible to extend the network lifetime in terms of two things. First is computation of hierarchical tree by strip tree geometry algorithm. In this computation, binary subtrees are computed via recursive subdivision of the network field. Therefore, three nodes (i.e., parent node, left and right nodes) in each subtree will have laid the foundation for communicating within their small region. Second is chain formation using in-order tree traversal algorithm. During the chain formation, each subtree will connect with nearest subtree as well as each other within its own in the following manner: left child → parent node → right node. In summary, since each subtree means a small region of the network field, sensor nodes can communicate with their near neighbors within small regions without the need to communicate over a long distance in order to gather data.

We used OMNeT++ simulator to evaluate performance. The simulation results are used to verify the balanced energy consumption of sensor nodes and the whole network lifetime. The remaining sections of this paper are organized as follows. Section 2 presents PEGASIS protocol and PEGASIS-related research, and strip tree geometry algorithm is briefly described in Section 3. In Section 4, we describe EECF algorithm, which extends the network lifetime of PEGASIS via balanced energy consumption among sensor nodes. The simulation results for performance verification are described in Section 5. Finally, we conclude the paper with the discussion on our future work in Section 6.

2. PEGASIS Protocol and Related Research

PEGASIS [12] is the well-known chain-based routing protocol for data gathering. In PEGASIS, immobile sensor nodes are randomly distributed in the network field as shown in Figure 1. The sink is located far from sensor nodes and is fixed. Each node using global positioning system (GPS) knows its own location and that its neighbor nodes. All sensors are homogeneous and energy restricted with the same initial energy. These sensor nodes can control their power and communicate with the other nodes or the sink directly.

2.1. Chain Formation. A chain using a greedy algorithm starting from farthest node away from the sink is formed before the first round. During chain formation process by the greedy algorithm, sensor nodes select the nearest node as the next node. For example, when sensor nodes are randomly distributed as shown in Figure 2, chain index of the sensor nodes is different from identification number of each node. Each node has a unique index in the chain. As stated above, chain formation of PEGASIS start from farthest node away from the sink and use the greedy algorithm. In Figure 2, since a sensor node farthest away from the sink is node-1, chain formation is started from node-1. Therefore, the chain index of node-1 becomes 0 (i.e., C0). According to chain formation process by the greedy algorithm, the next node of node-1 (C0) connects to the nearest node-3. The index of node-3 is increased 1 for the chain index of node-1. In PEGASIS, sensor nodes already contained in the chain cannot be selected as the next node to prevent a loop. For this reason, they (i.e., node-1 and node-3) are excluded from the list of the next candidates on node-3. Until all the sensor nodes are included in the chain, the formation process of a chain is repeated.

2.2. Data Gathering. If chain formation is completed, one node is randomly selected as the leader to transmit fused data to the sink in each round. The leader node sends a small token to the end nodes of the chain for data gathering. By token passing, the data held by each node is transmitted toward the leader node along the chain. Sensor nodes except end nodes fuse own data and data received from their neighbor and then transmit the fused data to the other neighbor node along the chain as shown in Figure 3. The leader node receives data from both neighbors commonly and transmits fused data to the sink which is located far away from sensor nodes.

2.3. PEGASIS-Related Research. In PEGASIS, various algorithms [15–19] have been proposed in different contexts to increase the network lifetime in WSNs. The algorithms proposed in [15–17] basically focused on the efficient use of energy for gathering data. The topic of an investigation has gradually been changed to the importance awareness of a leader node for a data transmission to the BS [18, 19]. In [15], the authors formed the multiple chains using greedy algorithm to gather data efficiently in a chain-oriented sensor network. In each chain of different levels, a leader node is selected based on the remaining energy. The leader in higher level chain receives data from the other leaders, and then transmits to the BS. The authors of [16] proposed the concentric clustering scheme which considers the network density and location of the BS to enhance its performance and prolong lifetime. In each circle of different levels, sensor nodes form a chain using a greedy algorithm. One node is selected as a header of each chain, and one among these headers transmits aggregated data to the BS. diamond-shaped (DS) PEGASIS [17] has expanded the concentric clustering scheme [16] to transmit reliable data. In a chain of each circle of different levels, one or two nodes are selected as the header node according to levels. Aggregated data is transmitted to the BS along the selected header nodes as diamond-shaped structures. The authors proposed diverse strategies of leader selection in [18]. In a chain (using the greedy algorithm), a leader node is selected according to energy aware in the randomly chosen block among 2 or 4 blocks as well as random, shuffle, and high-energy. Lim et al. [19] proposed how to deal with the transmission failure of a leader node to raising energy efficiency. When the leader

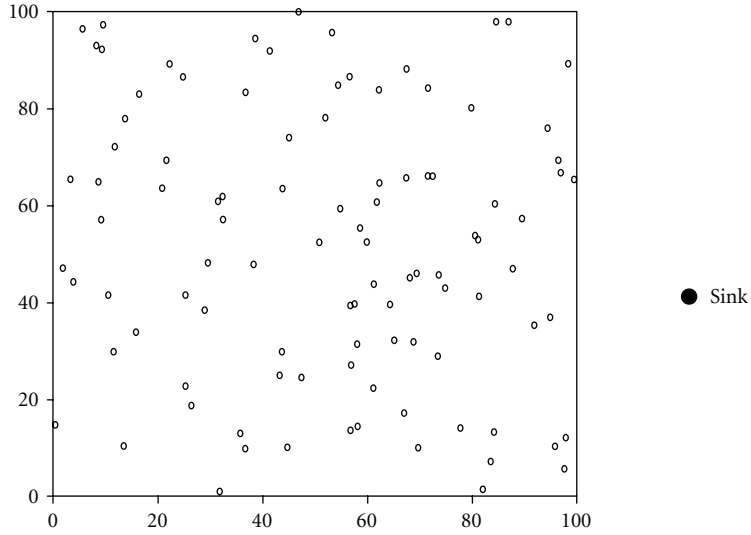


Figure 1: Sensor network with 100 sensor nodes and a sink (100 m × 100 m).

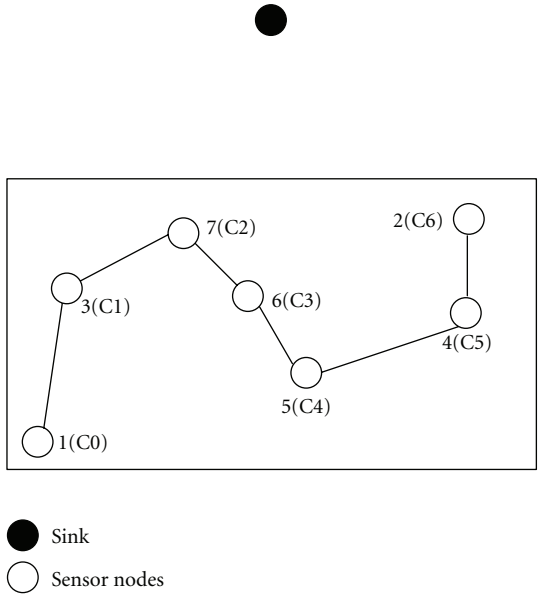


Figure 2: Example of PEGASIS's chain formation.

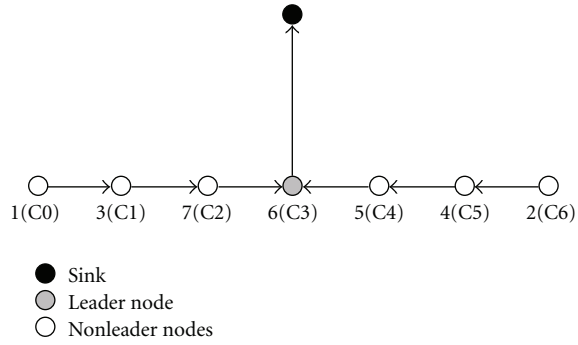


Figure 3: PEGASIS's data gathering.

node for each round cannot transmit aggregated data to the BS, data loss will occur. One node among leader node's neighbors is selected based on residual energy and transmits the aggregated data to the BS without loss of data.

As described above, most PEGASIS-related algorithms including PEGASIS used the greedy algorithm. In this case, neighbor distance will increase gradually since sensor nodes already in the chain cannot be revisited to prevent loop. Consequently, some sensor nodes need long distance data transmission, thereby consuming much more energy to transmit fused data to their neighbors. This unbalanced energy consumption will lead to the lifetime decrease of a network. In this paper, we propose an energy efficient chain formation (EECF) algorithm for the network lifetime extension via balanced energy consumption. To our knowledge, this paper is the first study based on both algorithms (strip tree geometry algorithm and in-order tree traversal algorithm). The simulation results are used to verify the balanced energy consumption of sensor nodes and the whole network lifetime.

3. Buttenfield's Strip Tree Geometry Algorithm

Buttenfield's strip tree geometry algorithm [13] is one among vector-based algorithms used to generalize spatial data in geographic information systems (GIS) environment. In [14], algorithm for transmission of vector geospatial data is developed. In EECF algorithm, we use a hierarchical tree computed based on strip tree geometry algorithm. Network field using strip tree geometry algorithm will be recursively subdivided. Subtree will be added in hierarchical tree via the recursive subdivision. Each node communicates with near neighbors in subtree. Hierarchical subdivision by strip tree geometry algorithm is explained below. In Figure 4, initial vector point of strip tree geometry algorithm is point 1 and 27. The first anchor line length is measured as

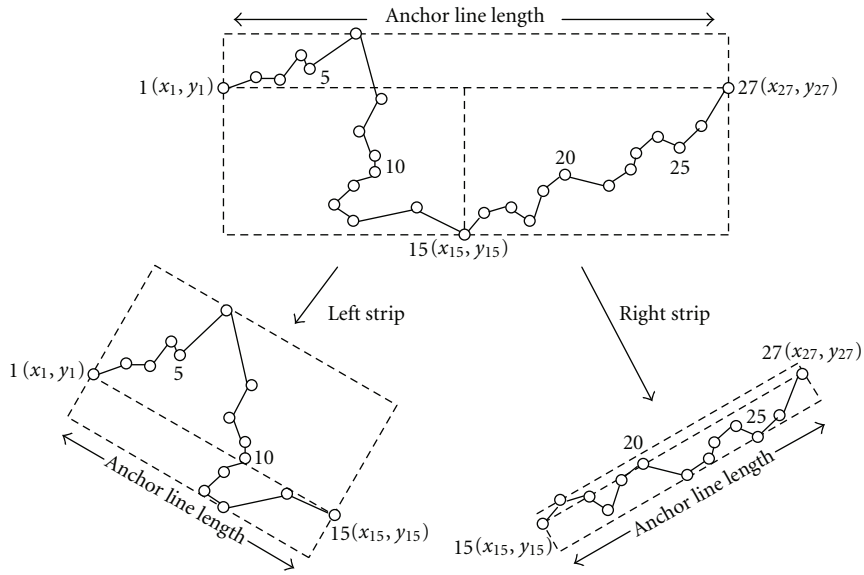


Figure 4: Hierarchical subdivision by Buttenfield's strip tree geometry algorithm.

distance between two endpoints (i.e., coordinates (x_1, y_1) and (x_{27}, y_{27})) by a mathematical formula the following:

$$\text{Length} = \sqrt{\left[(x_1 - x_1)^2 + (y_{27} - y_{27})^2\right]}. \quad (1)$$

Anchor line length is used to standardize the minimum bounding rectangle (MBR). The MBR is the rectangle bounding the line segment. A coordinate (x_{15}, y_{15}) of maximum perpendicular distance from the line connecting between point 1 and 27 will become the first vector point. This vector point 15 is stored in hierarchical strip tree. When the value of perpendicular distance is less than the threshold, that point is removed. By maximum perpendicular distance of point 15, two fields ("left-strip" and "right-strip") are indicated as shown in Figure 4. This process will continue until the preset recursive count.

4. Our EECF Algorithm

In order to balance energy consumption in PEGASIS, we propose EECF algorithm for chain formation. In EECF, if sensor nodes received a message for chain formation, each node computes a hierarchical tree using strip tree geometry algorithm and then transmits this message to the next node selected based on in-order tree traversal algorithm until all the sensor nodes are included in the chain. After chain formation, one node-like in PEGASIS using the greedy algorithm is randomly selected as the leader to transmit fused data to the sink in each round. By token passing, each node transmits fused data toward the leader node along the chain. The leader node transmits data received from its neighbors to the sink.

4.1. Basic Assumptions. In our sensor network model for simulation, the network field is a 2-dimension area with 100×100 (m) size. In the network field, immobile sensor nodes with the same capability are randomly distributed. The sink is fixed at a far location (50, 300) from sensor nodes. We assume the following to form a chain using EECF algorithm in PEGASIS. We have used the same parameters in the same environments as PEGASIS for a fair evaluation. Each node knows global knowledge of the network field (i.e., sensor nodes location, and location of sink), and can directly communicate with the other nodes or the sink. An initial energy for sensor nodes is identical and has energy restriction.

4.2. Hierarchical Tree Computation. First of all, we divide the network field using a straight line that connects a start and end node. The start node becomes furthest node from the sink, and a node with longest communication distance from this start node becomes the end node. Thus, in Figure 5, when the location of the sink is (50, 300), nodes 7 and 8 are the start and end node, respectively. The network field is divided into two regions by a straight line ($L_{7\text{-}8}$) that connects nodes 7 (x_7, y_7) and 8 (x_8, y_8). The straight line ($L_{7\text{-}8}$) equation is computed as follows:

$$y - y_7 = \left\{\frac{y_8 - y_7}{x_8 - x_7}\right\} * (x - x_7), \quad (x_7 \neq x_8). \quad (2)$$

So if two regions are R_7 and R_8, regions R_7 and R_8 have to include both start and end node. Therefore, member nodes of regions R_7 and R_8 are the following:

$$R_7 = \{\text{node}_1, \text{node}_3, \text{node}_5, \text{node}_7, \text{node}_8\},$$
$$R_8 = \{\text{node}_2, \text{node}_4, \text{node}_6, \text{node}_7, \text{node}_8, \text{node}_9\}. \quad (3)$$

Second is to decide root node for hierarchical tree of each region. In the first step, the network field was divided into

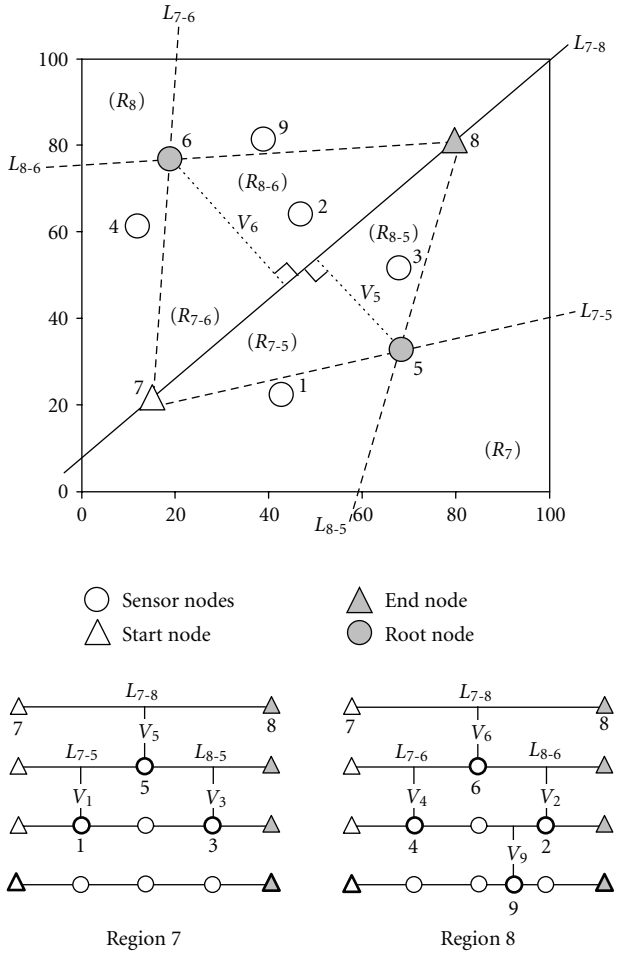


Figure 5: How the network field is divided (100 m × 100 m).

region R_7 and R_8. In R_7 and R_8, node 5 and 6 are located at longest vertical distance from straight line ($L_{7\text{-}8}$), respectively. Thus, two nodes become root node of each hierarchical tree as shown in Figure 6. If straight line ($L_{7\text{-}8}$) equation is $ax + by + c = 0$, vertical distance (V_5) of node 5 can calculate from the following:

$$V_5 = \frac{|ax_5 + by_5 + c|}{\sqrt{(a^2 + b^2)}}. \tag{4}$$

Next step is subdivision process of network field by root nodes. In Figure 5, region R_7 divides into two subregions by vertical line V_5, and region R_8 also divides by V_6. Ultimately, the network field divides into four regions (i.e., $R_{7\text{-}5}$, $R_{7\text{-}6}$, $R_{8\text{-}5}$, and $R_{8\text{-}6}$). And nodes 5, 6, 7, and 8 can connect new four straight lines ($L_{7\text{-}5}$, $L_{7\text{-}6}$, $L_{8\text{-}5}$, and $L_{8\text{-}6}$) as shown in Figure 5.

Fourth step performs child nodes addition to hierarchical tree. In four regions ($R_{7\text{-}5}$, $R_{7\text{-}6}$, $R_{8\text{-}5}$, and $R_{8\text{-}6}$), sensor nodes within each region calculate their own vertical distance from four straight lines ($L_{7\text{-}5}$, $L_{7\text{-}6}$, $L_{8\text{-}5}$, and $L_{8\text{-}6}$). As a result, node 1, 2, 3, and 4 located at longest vertical distance are added as child node of root node 5 and 6 as shown in Figure 6. Regions $R_{7\text{-}5}$, $R_{7\text{-}6}$, $R_{8\text{-}5}$, and $R_{8\text{-}6}$ are once more divided each by vertical lines of nodes 1, 2, 3, and 4, and then longest

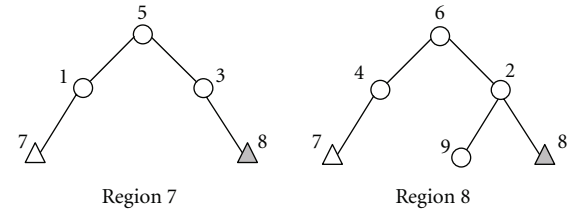


Figure 6: Hierarchical tree structures.

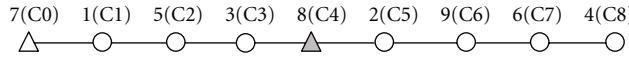


Figure 7: Chain formation by in-order tree traversal algorithm.

vertical distance. That is to says the network field divides into 8 regions. In 8 regions, each node calculates its own vertical distance, and sensor node located longest vertical distance is added to hierarchical tree. If all the sensor nodes except start (node 7) or end node (node 8) were added in hierarchical tree, two nodes (node 7 and 8) are added lastly to hierarchical tree, respectively.

Through the above computation process, we can see two hierarchical tree structures from the second figure in Figure 5 as shown in Figure 6. In this hierarchical tree, subtrees are added via recursive subdivision of the network field. Sensor nodes will communicate with each other in each subtree, and this has laid the foundation for communicating in small region.

4.3. Chain Formation. We use in-order tree traversal algorithm to form an energy efficient chain on the hierarchical tree. As first, in-order traversal visits left subtree from node of lower level, then visits on root node (i.e., parent node of left subtree). Lastly, it visits the right subtree. In our algorithm, chain formation from start node is started because a start node is located left subtree on low-level as shown in Figure 6. The start node select next node (node 1) located at a more short distance between node 1 and 4. From node 1, sensor nodes (node 5, 3, and 8) form a chain in the order by in-order traversal algorithm, and in hierarchical tree of region 8, chain formation continues in the same manner from left subtree (node 8) on low-level as shown in Figure 7.

Figure 8 shows a definite comparison between two chains using the greedy algorithm and EECF algorithm under the identical conditions for data gathering. In both algorithms, chain formation is started from the furthest node 7 from the sink (50, 300). In case of the greedy algorithm, the next node for node-7 becomes the nearest node-1, and the index of node-1 is increased by 1 for the chain index of node-7. Until all the sensor nodes are included in the chain, this process is repeated. In this chain, we can see long communication distance between nodes 4 and 8. This is because node-2 is selected as the next node for node-3 because it is the nearest. Moreover, sensor nodes already in the chain cannot be revisited to prevent loop. Thus, chain using the greedy

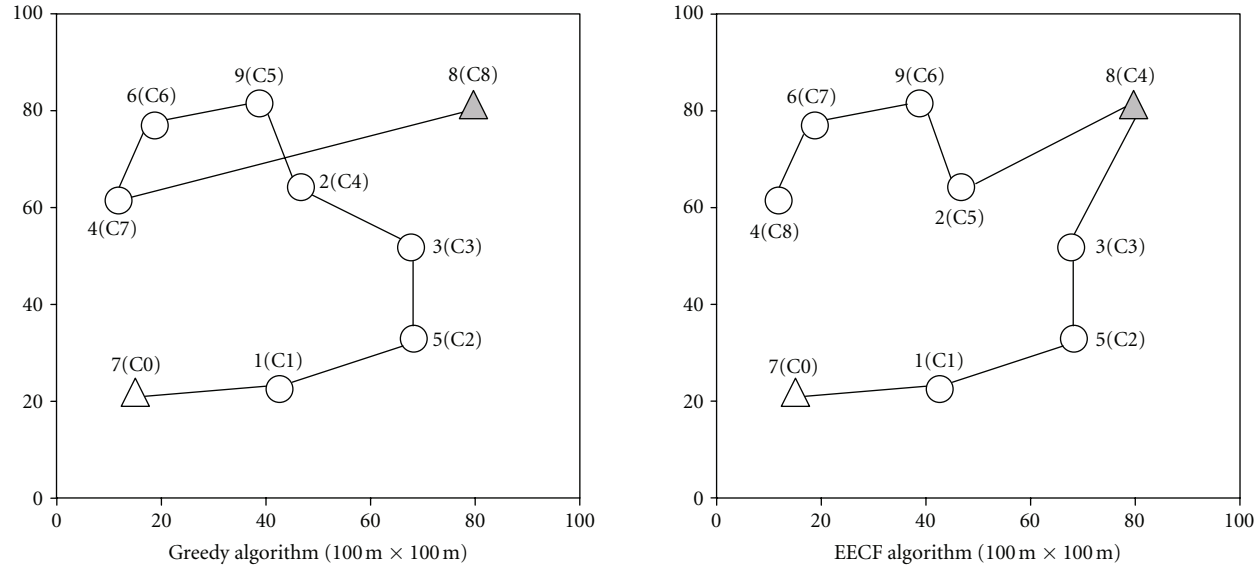


Figure 8: Chain comparison between greedy algorithm and EECF algorithm: the location of the sink is (50, 300).

algorithm is formed as follows: 7 → 1 → 5 → 3 → 2 → 9 → 6 → 4 → 8. In order to compare two algorithms, we illustrate the chain using EECF algorithm on Figure 7 at network field (100 m × 100 m) as shown below.

5. Simulation

5.1. Simulation Environment. We used OMNeT++ simulation tool to evaluate performance for EECF algorithm and two algorithms [12, 16]. In this section, our EECF algorithm is called "EECF-PEGASIS," and the algorithms in [12, 16] are called "original-PEGASIS" and "enhanced-PEGASIS," respectively. In original-PEGASIS and enhanced-PEGASIS, data held by each node is gathered along the chain using the greedy algorithm. In original-PEGASIS, after single chain formation, the leader node gathers data via token passing and transmits fused data to the sink. When the sink data is received data, this means one round. By contrast, enhanced PEGASIS forms a chain in each different level. In our simulation, we divide the network field into five levels for chain formation. According to the total number of sensor nodes, the number of sensor nodes in each level set the same. One node is selected as a header of each chain, and low level one among these headers transmits aggregated data to the BS. In case of EECF-PEGASIS, sensor nodes transmit a message for chain formation to their next node based on in-order tree traversal path in computed hierarchical tree. If this chain formation is completed, like original-PEGASIS, one leader node transmits gathered data to the sink.

We measured the performance in three different scenarios on the location of the sink: (50, 200), (50, 300), and (50, 400). In all scenarios, the total number of sensor nodes will increase from 50 up to 250 (50, 100, 150, 200, and 250). In case the total number of sensor nodes is the same, sensor nodes are distributed randomly at same location for fair evaluation. Using these scenarios we will be able to compare the performance of three algorithms through increase in number of sensor nodes in each simulation, as well as comparing results of performance based on the location of the sink. In our simulation, there is only one sink with unlimited energy. Sensor nodes randomly distributed in a 100 m × 100 m field. After distribution, the location is fixed. All the sensor nodes are homogeneous and have the same capability. Each of them can adjust its transmission range, and Initial energy is 1.0 J (energy restriction). The size of data that moves from node to node in the chain is the same by using data fusion. The energy consumption for data fusion is 5 nJ/bit/message. We do not consider token passing energy, computation energy, and delay time.

Table 1: Radio model.

Radio model	Formulas
Transmitting	$E_{TX}(k,d) = E_{TX\text{-}elec}(k) + E_{TX\text{-}amp}(k,d)$
	$E_{TX}(k,d) = E_{elec} * k + \epsilon_{amp} * k * d^2$
Receiving	$E_{RX}(k) = E_{RX\text{-}elec}(k)$
	$E_{RX}(k) = E_{elec} * k$

5.2. Simulation Measurements. In order to analyze an accurate performance, we define three measurement results: longest communication distance, average communication distance, network lifetime, and average remaining energy. Our purpose in this paper is to extend the network lifetime via balance energy consumption among sensor nodes. We measured until the energy of the first sensor node is exhausted. Firstly, the longest communication distance indicates longest distance between any two nodes in the chain. According to formulas in Table 1, sensor nodes will consume much more energy to transmit data.

Table 2: Parameters.

Symbol	Value	Description
E_{elec}	50 nJ/bit	The transmitter or receiver circuitry
ϵ_{amp}	100 pJ/bit/m^2	The transmitter amplifier
k	2000 bit	Bits message
d	meters	Distance between two nodes

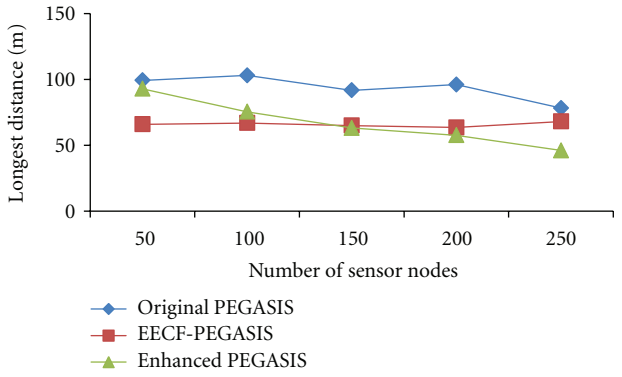


Figure 9: Longest communication distance for different number of sensor nodes.

Besides, since communication distance has a vital effect on energy consumption, the measurement of long-distance data transmission of certain nodes is critical to balance energy consumption among sensor nodes. Secondly, the network lifetime means the total number of rounds, and the need for maximization. The reason is that in WSNs, all the sensor nodes need to collect data and transmit their data to the sink during the assigned time without battery recharge. Therefore, it is important to maintain the lifetime of all the sensor nodes for a long time as possible. Lastly, the average remaining energy is used to analyze balance of energy consumption. We define this measurement as the sum of remaining energy values of all the sensor nodes divided by total number of sensor nodes in each simulation. In case of a large value, this means unbalance of energy consumption. This is because the remaining energy of one node among sensor nodes is almost zero, and other nodes except this one node maintain high energy. Therefore, the average remaining energy needs to be minimized by contrast to network lifetime for balancing of energy consumption.

5.3. Radio Model. We, as in [15–19], apply the same radio model used in PEGASIS [12] for message transmitting and receiving. In the first two formulas in Table 1, $E_{TX}(k, d)$ is energy consumption for transmitting a k-bit message and a distance d, assuming that d is d^2 energy loss caused by channel transmission. The second $E_{RX}(k)$ is the formula to compute energy consumption for receiving a k-bit message.

In order to compare performance under the identical simulation conditions as in PEGASIS, we define parameters in Table 2. E_{elec} means an energy that the transmitter or receiver circuitry needs for running. The necessary energy for transmitter amplifier is 100 pJ/bit/m^2, k is the message size in bit, and d is the distance in meters between the transmitter and receiver node.

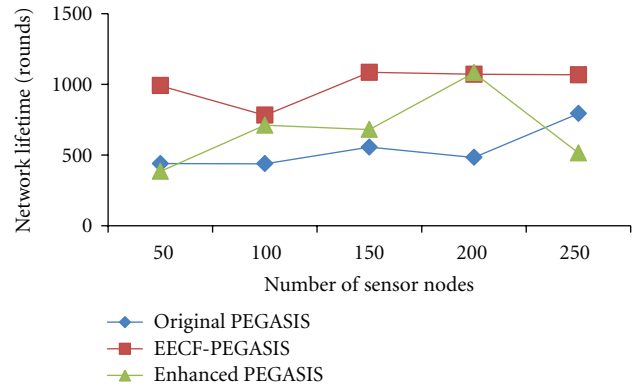


Figure 10: Network lifetime for different number of sensor nodes, sink location (50, 200).

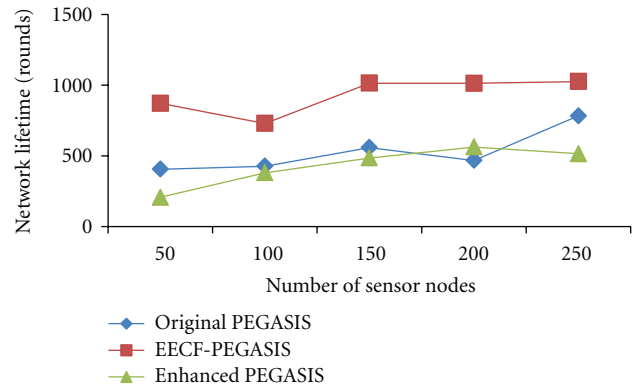


Figure 11: Network lifetime for different number of sensor nodes, sink location (50, 300).

5.4. Simulation Results. We evaluated the network lifetime of three algorithms (i.e., original-PEGASIS, Enhanced PEGASIS, and EECF PEGASIS) under different sink location environments until the energy of the first sensor node is exhausted. In each simulation, the total number of sensor nodes is increased from 50 up to 250.

Figure 9 shows longest distance between any two nodes in the chain. In EECF-PEGASISS, this longest distance was reduced by around 40% than the greedy algorithm results. In Table 1, distance d^2 energy loss will occur for transmitting. In other words, this means that communication distance is closely related with energy consumption of each node. Therefore, since higher energy consumption of some nodes that caused by long distance data transmission is very important factor to effect in the whole network lifetime, the longest communication distance needs to be minimized.

From Figures 10, 11, and 12, EECF-PEGASIS shows better performance than original PEGASIS and enhanced-PEGASIS. In EECF algorithm, since each subtree in hierarchical tree means smallest unit region in the network

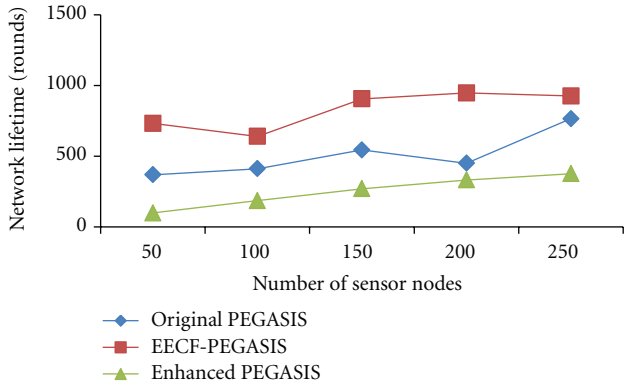


Figure 12: Network lifetime for different number of sensor nodes, sink location (50, 400).

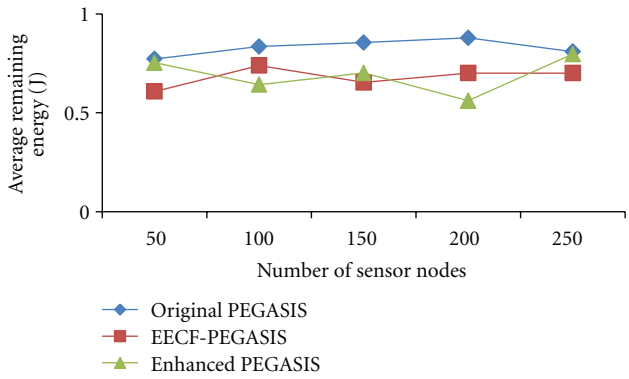


Figure 13: Average remaining energy for different number of sensor nodes, sink location (50, 200).

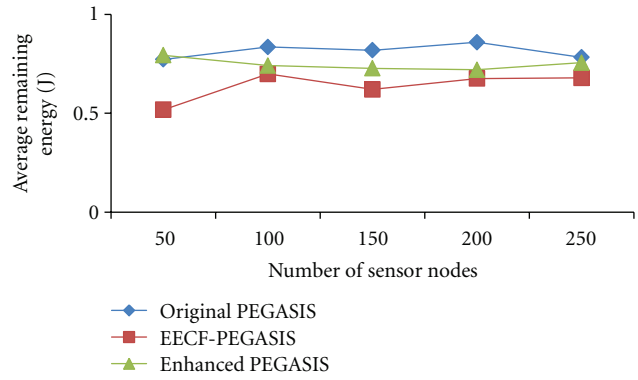


Figure 14: Average remaining energy for different number of sensor nodes, sink location (50, 300).

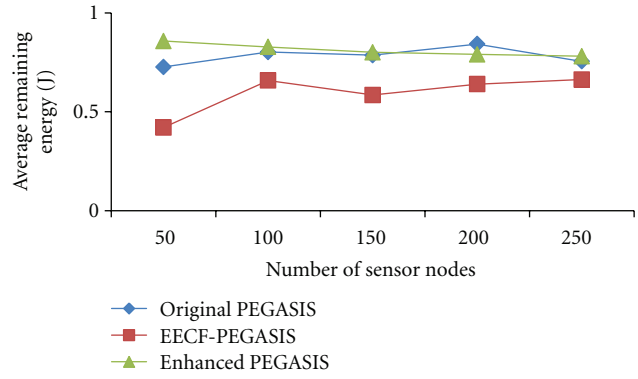


Figure 15: Average remaining energy for different number of sensor nodes, sink location (50, 400).

field, sensor nodes can communicate with their nearest nodes within each subtree. Furthermore, each subtree can be connected with its nearest subtree from the viewpoint of the whole hierarchical tree through in-order tree traversal algorithm. As a result, each node communicates with its neighbors within small regions without long-distance data transmission in order to gather data. By these reasons, the network lifetime increased under different sink location environments as shown in Figures 10, 11, and 12. In addition, because all the simulations run in the same network size (100 m $\times$ 100 m), each algorithm shows aspect analogous in Figures 10, 11, and 12.

We also evaluated an average remaining energy for different number of sensor nodes under different sink location environments until the energy of the first sensor node is exhausted. When we compare the three algorithms from Figures 13, 14, and 15, EECF-PEGASIS shows low values than original PEGASIS and enhanced PEGASIS. The average remaining energy means average values for remaining energy values of alive nodes except one node. Thus, in case of a large value, this means unbalance of energy consumption. Therefore, the average remaining energy needs to be minimized for energy consumption balance.

6. Conclusion

In this paper, we have proposed an energy efficient chain formation (EECF) algorithm for resolving unbalanced energy consumption problem caused by long communication distance of some sensor nodes in PEGASIS using the greedy algorithm. In EECF, we use two algorithms, Buttenfield's strip tree geometry algorithm and in-order tree traversal algorithm. To evaluate performance, we measured three metrics until the first sensor node fails. First measurement values are the longest communication distance among communication distance between any two nodes in the chain. Second is the network lifetime, and lastly, we measured the average remaining energy. Through simulation results, we have proved good performance of EECF algorithm in both the balance of energy consumption and whole network lifetime. In our future work, we have a plan to combine our algorithm and the greedy algorithm for reduction of total communication distance and average distance in a chain.

Acknowledgments

This work was supported by the Second Brain Korea 21 Project. Professor M.-S. Park is acknowledged for his supervision to this research.

References

[1] H. Karl and A. Willig, *Protocols and Architectures for Wireless Sensor Networks*, John Wiley & Sons, New York, NY, USA, 2005.

[2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, 2002.

[3] R. Rajagopalan and P. Varshney, "Data-aggregation techniques in sensor networks: a survey," *IEEE Communications Surveys & Tutorials*, vol. 8, no. 4, pp. 48–63, 2006.

[4] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings of the 33rd Annual Hawaii International Conference on System Siences (HICSS '00)*, January 2000.

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

[6] S. Lindsey, C. Raghavendra, and K. M. Sivalingam, "Data gathering algorithms in sensor networks using energy metrics," *IEEE Transactions on Parallel and Distributed Systems*, vol. 13, no. 9, pp. 924–935, 2002.

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

[8] O. Younis and S. Fahmy, "HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks," *IEEE Transactions on Mobile Computing*, vol. 3, no. 4, pp. 366–379, 2004.

[9] A. Manjhi, S. Nath, and P. B. Gibbons, "Tributaries and deltas: efficient and robust aggregation in sensor network streams," in *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '05)*, Baltimore, Md, USA, June 2005.

[10] Y. Xue, Y. Cui, and K. Nahrstedt, "Maximizing lifetime for data aggregation in wireless sensor networks," *Mobile Networks and Applications*, vol. 10, no. 6, pp. 853–864, 2005.

[11] Y. Lee, K. Lee, H. Lee, and A. Kusdaryono, "CBERP: cluster based energy efficient routing protocol for wireless sensor network," in *Proceedings of the 12th International Conference on Networking, VLSI and Signal Processing (ICNVS '10)*, vol. 3, pp. 24–28, 2010.

[12] S. Lindsey and C. S. Raghavendra, "PEGASIS: power-efficient gathering in sensor information systems," in *Proceedings of the IEEE Aerospace Conference*, vol. 3, pp. 1125–1130, 2002.

[13] B. P. Buttenfield, "A rule for describing line feature geometry," in *Map Generalization: Making Rules For Knowledge Representation*, B. Buttenfield and R. McMaster, Eds., pp. 150–171, Longman, London, UK, 1991.

[14] B. P. Buttenfield, "Transmitting vector geospatial data across the internet," *Lecture Notes in Computer Science*, vol. 2478, pp. 51–64, 2002.

[15] N. Tabassum, Q. E. K. M. Mamun, and Y. Urano, "COSEN: a chain oriented sensor network for efficient data collection," in *Proceedings of the 3rd International Conference on Information Technology: New Generations (ITNG '06)*, pp. 262–267, Las Vegas, Nev, USA, April 2006.

[16] S. M. Jung, Y. J. Han, and T. M. Chung, "The concentric clustering scheme for efficient energy consumption in the PEGASIS," in *Proceedings of the 9th International Conference on Advanced Communication Technology (ICACT '07)*, pp. 260–265, Gangwon-do, Korea, February 2007.

[17] J. E. Lee and K. Kim, "Diamond-shaped routing method for reliable data transmission in wireless sensor networks," in *Proceedings of the International Symposium on Parallel and Distributed Processing with Applications (ISPA '08)*, pp. 799–801, December 2008.

[18] I. Shukla and N. Meghanathan, "Impact of leader selection strategies on the PEGASIS data gathering protocol for wireless sensor networks," *Ubiquitous Computing and Communication Journal*, vol. 4, no. 5, 2009.

[19] S. J. Lim, A. K. Bashir, S. Y. Rhee, and M. S. Park:, "Energy-based re-transmission algorithm of a leader node's neighbor node for reliable transmission in the PEGASIS," in *Ubiquitous Computing and Multimedia Applications Part I*, vol. 150 of *Communications in Computer and Information Science*, pp. 120–128, Springer, Heidelberg, Germany, 2011.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 429719, 8 pages
doi:10.1155/2012/429719

Research Article

An Energy-Efficient and Fault-Tolerant Convergecast Protocol in Wireless Sensor Networks

Ting Yang,[1] ChunJian Kang,[1] and Guofang Nan[2]

[1] *School of Electrical Engineering and Automation, Tianjin University, Tianjin 300072, China*
[2] *Institute of Systems Engineering, Tianjin University, Tianjin 300072, China*

Correspondence should be addressed to Ting Yang, yangting@tju.edu.cn and Guofang Nan, gfnan@tju.edu.cn

Received 19 April 2012; Revised 9 July 2012; Accepted 9 July 2012

Academic Editor: Jianhua He

The simple graph theory is commonly employed in wireless sensor networks topology control. An inherent problem of small-granularity algorithms is the high computing complexity and large solution space when managing large-scale WSNs. Computed transmission paths are of low fault tolerance because of unattended sensor nodes and frail wireless transmitting channels. This paper uses hyper-graph theory to solve these practical problems and proposes a spanning hyper-tree algorithm (SHTa) to compute the minimum transmitting power delivery paths set for WSNs convergecast. There are three main contributions of this paper: (1) we present a novel hyper-graph model to abstract large-scale and high connectivity WSNs into a robust hyper-tree infrastructure; (2) we present a precise mathematical derivation that solves the "hyper-tree existence" problem; (3) SHTa is proposed to compute the delivery paths set, which is the minimum power transmitting convergecast hyper-tree. Variable scale hyper-edges represented as computing units limit solution space and reduce computing complexity. Mutual backup delivery paths in one hyper-edge improve the capability of fault tolerance. With experiment results, SHTa computes short latency paths with low energy consumption, compared with previous algorithms. Furthermore, in dynamic experiments scenes, SHTa retains its robust transmitting quality and presents high fault tolerance.

1. Introduction

Self-organized wireless sensor networks can be used to cooperatively detect and perceive real objects. Sensors can communicate and exchange information among themselves without human intervention. This is achieved by integrating technologies, including sensors, embedded calculations, distributed information processing, and wireless communication. Wireless sensor networks have huge potential in civil and military applications, such as smart grid, smart home, healthcare monitoring, and intelligent transport.

Self-organized wireless sensor networks are made up of highly distributed systems of small-size, wireless unattended sensors. Each sensor is capable of detecting devices' current operating conditions, such as temperature, noise, vibration, or output signals. This data is preprocessed, transmitted, and exchanged in a machine-to-machine (M2M) network [1, 2]. There is a need for reliable, scalable, and smart protocols and algorithms for self-organized M2M networks or sensor networks.

In traditional communication networks, simple graph theory is always used [3, 4]. But a large-scale self-organized wireless sensor network consists of hundreds or thousands of nodes with a complex topology. Hence, a large number of the control massages are required to establish transmission paths. On the other hand, because of the low reliability of the sensor nodes and wireless communication links, many real-time control messages have to be used to maintain an established path. These tasks use significant amount of bandwidth and consume the extra energy.

To solve this problem, in this paper, we used the hyper-graph theory and proposed Spanning Hyper-Tree algorithm (SHTa) to create a concise and robust hyper-graph infrastructure for large-scale and high connectivity self-organized wireless networks.. Based on the best of our knowledge, it is the first hyper-graph model for self-organized wireless

Table 1: Total sum cost with two transmitting methods.

	Total weight of SPT = 3.82 Routing based on SPT			Total weight of MST = 3 Routing based on MST	
Seq and time	Routing path	Consumption	Seq and time	Routing path	Consumption
1	*a*–*S*	1.41	1	*a*-*b*	1
2	*b*–*S*	1.41	1	*c*-*b*	1
3	*c*–*S*	1	2	*b*–*S*	1

networks architecture. Because a dynamic hyperedge is the minimum computing unit during routing in this type of hyper-graph architecture, fewer packets are used, which saves energy and prolongs the network's lifetime. More than one connected pairs in a hyperedge provides high bandwidth and low loss rate during transmission. This effectively improves network fault tolerance. Moreover, SHTa solves the "hyper-tree existence uncertainty problem," which is a new problem that differs from the simple graph model. An axiom "any graph has its spanning tree" is invalidated in a hyper-graph, that is, not each hyper-graph exists spanning hyper-tree with loop-free. SHTa presents an effective spanning hyper-tree method and we proposed the strict mathematical proof to prove the certainty theorem.

The remainder of the paper is organized as follows. Section 2 introduces some background material on wireless communication network architecture and optimal routing problems. Wireless self-organized sensor networks' hyper-graph model is presented in Section 3. Section 4 describes the SHTa in detail, followed by validity proof. Section 5 proposes the computer simulation and evaluation; finally, Section 6 is the conclusion and outline of future research.

2. Convergecast with Data Aggregation in Wireless Sensor Networks

For peer-to-peer (P2P) communication model, Dijkstra and Bellman-Ford algorithms are often employed to build a shortest path tree (SPT), such as OSPF used in IP backbone networks. Each router with OSPF stores an SPT in which the root is itself. Packets are transmitted following SPT's branches to arrive the minimum cost.

Different from P2P, self-organized wireless sensor network collects data from each sensor nodes to "Sink," called convergecast. During transmission, data aggregation is used to eliminate the redundancy in collection data. Many algorithms are presented to establish data aggregation tree, such as EADAT [5], E-Span [6], and HEED [7]. These algorithms set transmitting energy consumption as link weight and build SPT as an aggregation tree. Reference [8] presented the DCTC algorithm to detect and track a mobile target. DCTC used Dijkstra to establish collection tree, which is also an SPT.

Not every data packet will be transmitted from source to destination, due to data aggregation, in the intermediate nodes. A wireless sensor network is generalized as data center, and the optimum number of transmissions required per datum in the DC (Data Centre) is equal to the number of edges in the Minimum Steiner Tree (MST). Therefore, MST, not SPT, is the truly minimized sum cost tree in convergecast protocols with data aggregation.

Figure 1 shows an example to explain this optimization problem. Three nodes transmit information to Sink. SPT and MST are shown, respectively, in Figures 1(a) and 1(b). Table 1 presents the two routing configurations: the total cost of SPT method is 3.82, larger than MST. Therefore, MST is better than SPT. If the weight of edges is defined as energy consumption, MST is just the optimal energy consumption tree in the wireless sensor network.

3. Hypergraph Model for Wireless Self-Organized Sensor Networks

Whether for IP backbone network, cellar mobile network or Ad-hoc network, the simple graph theory is the main tool for research on architecture control [8–11] and counting routing protocols [3, 4, 12–14]. But in large-scale wireless self-organized sensor networks, the number of sensor nodes can be hundreds or thousands of times of that of backbone network or mobile network. Each node can connect with any one neighbour by omnidirectional antenna, which creates high node connectivity and complication in topology controlling. A simple graph algorithm with tiny granularity often has high computing complexity and uses a large amount of memory. On the other hand, in wireless sensor networks, a single transmitting path has a low fault tolerance level because of the low reliability of sensor nodes and wireless links. During data transmission, lots of control messages need to be transmitted frequently to maintain the connectivity of a delivery path, which may use lots of links' bandwidth and consume significant amount of energy.

To solve this problem, Hyper-graph theory is used as a novel mathematical tool to generalize high connectivity wireless self-organized networks into concise and robust hyper-graph infrastructure. As far as we know, it is the first hyper-graph architecture model in wireless sensor network. In the model, special nodes and connected edges among them are generalized as hyper-edges. With the growth of hyper-edges, as the minimum computing unit, fewer extra packets are used and the energy consumption is effectively reduced.

Proposition 1. *Let $X = \{x_1, x_2, \ldots, x_n\}$ be a finite set, and let $\varepsilon = \{E_i \mid i \in I\}$ be a family of subsets of X. If the following two conditions are satisfied: (1) $E_i \neq \emptyset$ $(i \in I)$; (2) $\bigcup_{i \in I} E_i = X$, the couple $H = (X, \varepsilon)$ is called a hyper-graph. $|X| = n$ is called the*

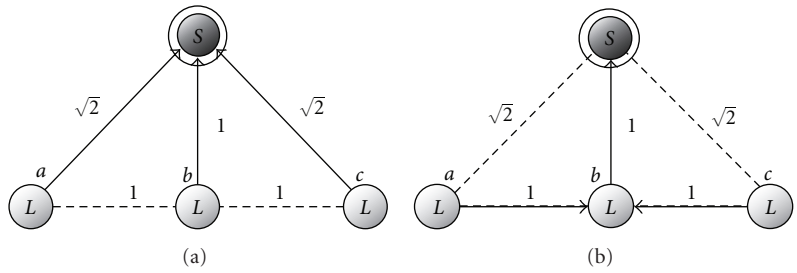


Figure 1: (a) Routing based on SPT. (b) Routing based on MST.

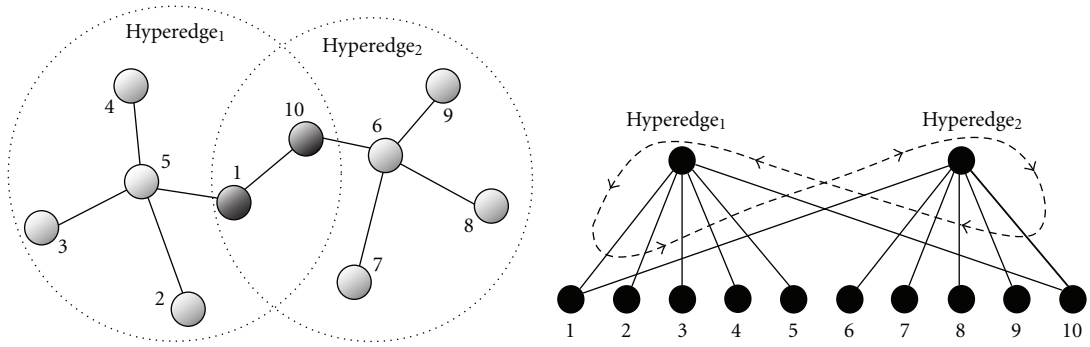


Figure 2: Hyper-graph model.

order of this hyper-graph. The elements $x_1, x_2, \ldots, x_n$ are called vertices and the $E_1, E_2, \ldots, E_m$ are called hyper-edges [15].

We describe a wireless self-organized sensor network as a hyper-graph $H = (X, \varepsilon)$, in which $X = \{x_1, x_2, \ldots, x_n\}$ is the sensor nodes set. Special characteristics of nodes are represented as a hyper-edge, that is, $E_i = \{N_1, N_2, \ldots, N_j, e_1, e_2, \ldots, e_k\}$, and hyper-edge set is $\varepsilon = \{E_1, E_2, \ldots, E_m\}$. It is obvious that cluster in simple graph is a special type of hyper-edge, and hyper-graph is the extended cluster.

In the hyper-graph model, we should also establish MST for optimal convegercast. But the binary relation of hyper-edge and vertices in hyper-graph is not the one-to-one mapping relation of vertices and edges as it is in a simple graph, which is more complex. Therefore, the axiom "any graph has its spanning tree" is invalid in hyper-graph, that is, not each hyper-graph exists spanning hyper-tree with loop-free. A hyper-graph example with no hyper-tree is shown in Figure 2. Two hyper-edges are split and Theorem 2 proposed hyper-tree does not exist, because of existing the loop (1-Hyperedge$_1$-10-Hyperedge$_2$-1) in the bipartite graph $G\langle H\rangle$, shown in Figure 2(b).

Theorem 2. *Hyper-graph H is a hyper-tree, if and only if the bipartite graph $G\langle H\rangle$ is a tree.*

To ensure one hyper-graph certainly has hyper-tree, we proved two conditions must be satisfied:

(i) *if $E_i \cap E_j \neq \emptyset$, called relative hyper-edges, then $|E_i \cap E_j| = 1$;*

(ii) *if condition one is satisfied and if $|E_i| \neq 2$, it must have $|E_j| = 2$.*

The precise mathematical proof is shown here. Firstly, if there is no hyper-cycle in $H = (X, \varepsilon)$, the proposition is true. Otherwise, if there are hyper-cycles, a break-cycle method is used. In a hyper-clcye C, three connecting hyper-edgs E_i, E_j, E_k always can be found easily, then there must be a 2-degree chained hyper-edge among them, assuming $|E_i| = 2, x_1, x_2 \in E_i, E_i \cap E_j = \{x_1\}, E_i \cap E_k = \{x_2\}$. When E_i is broken, C is not a closed cycle. But x_1 and x_2 still belong to the E_j, E_k, which is $x_1, x_2 \in \varepsilon', \varepsilon' = \varepsilon - E_i$. Then a new hyper-graph $H' = (X', \varepsilon')$ is spanned, where $X' = X$ and H' is also connective. Repeating this process till there is no hyper-cycle, the result is a hyper-graph $T = (X', \varepsilon'), X' = X$ and $\varepsilon' \subseteq \varepsilon$, which is the final spanning hyper-tree.

In Section 4, we presented a novel topology controlling algorithm to split hyper-edge, establish hyper-graph with satisfying the above two conditions, and span the minimum hyper-tree for minimum energy consumption convergecast.

4. Minimum Spanning Hyper-Tree Algorithm

In the implement of SHTa, a type of generalized synchronization mechanism with "synchronous round" was used. Firstly, we describe this synchronization mechanism.

Time synchronization is an important feature of distributed systems including wired and wireless communication systems. Many time synchronization schemes were designed including GPS [16] and Network Time Protocol

(NTP) [17] used in IP networks applications. In M2M and sensor networks, time synchronization is also used frequently for various purposes including sensor data fusion, coordinated actuation, and power-efficient duty cycling: for example integrating a time series of proximity detections into a velocity estimate; measuring the time of flight of sound for localizing its source; distributing a beam forming array; suppressing redundant messages by recognizing that they describe duplicate detections of the same event by different sensors; or supporting energy efficient scheduling and power management. Now, many good time synchronization algorithms, such as Reference Broadcast (RBS [18]), TINY/MINI-SYNC [19], and Level Synchronization [20], are presented to provide time accuracy in wireless self-organized sensor networks.

Compared with accurate time slots synchronization, generalized synchronization mechanisms with "synchronous round" can save a large number of timescale check packets, which ensures the accurate time synchronisation, and reduce the complexity of designing communication protocols, therefore reducing the transmission energy consumption. In a generalized synchronous mechanism, each processor unit should complete two steps during one synchronous round: in the first step, the processor transmits event driven messages to its neighbor; in the second step, processor switches its current state with a state transition function, once it has received any valid messages.

When synchronous network is a deterministic system, a state transition function with the same valid input must achieve the same output in each time. Mapping the two steps onto a sensor node processor, the following two operations would be implemented.

(i) Each main hyper-edge ε_m^k exchanges information with its neighbours and builds a subminimal power chained hyper-edge set $\{\varepsilon_{j1}^k, \varepsilon_{j2}^k, \ldots\}$.

(ii) A minimal power chained hyper-edge $\varepsilon_{m,n}^k$ is elected in set $\{\varepsilon_{j1}^k, \varepsilon_{j2}^k, \ldots\}$. Then ε_m^k, ε_n^k, and $\varepsilon_{m,n}^k$ merge into new main hyper-edge $\varepsilon_{m,n}^{k+1}$.

Two types of messages are employed in the above operations: (i) *Op_HC*: excite *HC* operation packet; (ii) *R_mPCHe*: request minimal power chain hyper-edge packet. The structures of the two types of packets are shown in Algorithm 1.

Without loss of generality, we suppose SHTa implements at k_{th} "synchronous round." Each main hyper-tree ε_m^k initiate to search for a minimal power chain hyper-edge connected with its neighbours. Nodes v_i, at the edge of ε_m^k, broadcasts *R_mPCHe* message. Any node receiving this message will implement SHTa as in Algorithm 2.

As soon as two main hyper-edges ε_m^k, ε_n^k confirm their conjunct minimal power chain hyper-edge, *Op_HC* messages are broadcasted in these three units. Any one node received this message will implement the operations described in Algorithm 3.

Whenever SHTa cannot find any new chained hyper-edge ε_j^k or implement consolidation operation, the algorithm stops and the spanning forest gathers into a hyper-tree. In the following section, we prove that this hyper-tree is just a minimum spanning hyper-tree.

We rewrote the conclusion of the spanning hyper-tree from SHTa algorithm: In hyper-graph $H = (X, \varepsilon), \cup_j E_j = \{(X, \varepsilon) : 1 \le j \le k\}$ is one of the hyper-graph's spanning forests. If e is the minimum weight chain hyper-edge in the set $\cup_j E_j$, there must be a hyper-tree, which is made up of $\cup_j E_j$ and e. Moreover, this hyper-tree is the minimum hyper-tree in all of the spanning hyper-trees which include $\cup_j E_j$.

Proof by Contradiction. Suppose that the conclusion is erroneous, that is, there is a hyper-tree T, which includes $\cup_j E_j$, but does not include the e. And T is strictly less than any other hyper-tree, which includes the $\cup_j E_j$ and the e. Now put e into T, and then obtain the graph T'. Obviously, there is a cycle in T', which includes another chain hyper-edge e', $e' \neq e$, and $e' \in \cup_j E_j$.

Based on the definition, weight(e') $\geq$ weight(e) is obtained. e' can be safely deleted from T'. And another hyper-tree T'' is made, including $\cup_j E_j$ and e. the power of T'' is not larger than the power of T. There is a contradiction for T. The supposition is in error, and the original proposition is true.

5. Computer Simulation

This section evaluates the performance of the novel algorithm using simulation. Firstly, seven different sensor scenes are studied, in a $200 \times 200\,\text{m}^2$ square area, and a number of sensor nodes are uniformly dispersed, ranging from 50 to 350 nodes with increment step of 50 nodes. Each node has a radio range of 40 m. We used this environment to simulate how different network density affects the energy consumption during the processing of spanning tree or hyper-tree. Then, transmission performance metrics, average latency and loss packets ratio, are evaluated when data packets are delivered following SHTa, compared with Directed Diffusion (DD) [21] and its improved algorithm EADD [22]. We use the same parameter as [21]: (1) using the 802.11 MAC protocol to ensure the data link connected; (2) setting the idle time power dissipation about 35 mW, receiving data power dissipation 395 mW, and transmiting data power dissipation 660 mW; (3) setting events modelled as 64 bytes and information control packet 36 bytes. Finally, in the simulations, we use a fixed events generated model—after every ten-second interval, ten nodes were randomly selected as sensor sources and generated constant bit rate (CBR) data streams with packet intervals of 0.1 seconds. The duration of each data streams is 5 seconds.

We first compute the maximum and average nodes' degree and the standard deviation of nodes' degree in seven different scenes to analyze the network density, shown in Table 2.

Figure 3 shows the average dissipated energy per packet as a function of networks size. DD and EADD have almost the same energy consumption and a half less than flooding. SHTa consumes less energy than DD and EADD. With the increase of the network size, SHTa can save 23.7% energy

```
Packet Op_HC struct:
typedef struct Op_HC_st          // the struct of Op_HC packet
{
   ULONG seq;                    // the serial number of the Op_HC packet
   UNIT new_mHe_ID;              // new consolidation main hyper-edge ID
   UINT Og_mHe_ID_1;             // original main hyper-edge ID
   UINT Og_mHe_ID_2;             // original main hyper-edge ID
   UINT Og_cHe_ID;               // chain hyper-edge ID between the above two main hyper-edges
   ULONG ttl;                    // the packet's lifetime
} Op_HC;
Packet R_mPCHe struct:
typedef struct R_mPCHe_st        // the struct of R_mPCHe packet
{
   ULONG seq;                    // the serial number of the R_mPCHe packet
   ULONG Source_E_ID;            // the R_mPCHe packet's source main hyper-edge ID
   UINT Source_ID;               // the R_mPCHe packet's source node ID
   UINT Dest_ID;                 // the R_mPCHe packet's destination node ID
   ULONG Tran_Pw                 // transmission power in the chain Hyper-edge
   ULONG ttl;                    // the packet's lifetime
} R_mPCHe;
```

Algorithm 1: Structures of the *Op_HC* and *R_mPCHe* packets.

$v_j \in V$ receive *R_mPCHe message and* implement the following operation in SHTa

```
(1)  if (v_j ← R_mPCHe)
(2)  {   if (R_mPCHe.ttl <= 0)
(3)          { ignore this R_mPCHe packet;
(4)            return UNSUCCESS;                  // can not recover the new delivery path
(5)          }
(6)       if (MPCHe.Source_E_ID == v_j.He_ID)     // v_j and R_mPCHe belong to the same main hyper-edge
(7)          { ignore this R_mPCHe;
(8)            R_mPCHe.ttl= R_mPCHe.ttl-1;
(9)          }
(10)      else                                    // v_j receive this R_mPCHe at first time
(11)         if (v_j ∈ ε_n^k)
(12)            R_mPCHe.Tran_Pw = Pw (v_i, v_j);  // sub-minimal power is transmission power in ε_j^k = {v_i, v_j}
(13)      if (Min_Pw(*ε_{j,i}^k) == Pw (v_i, v_j))  // ε_{m,n}^k is the minimal power in set {ε_{j1}^k, ε_{j2}^k, ...}.
(14)         ε_m^k Broadcast Op_HC;               // initiate Hyper-edges' Consolidation operation
(15) }
```

Algorithm 2: Implement flow while node receiving *R_mPCHe* message.

$v_j \in V$ receive *Op_HC message and* implement the following operation in SHTa

```
(1)  if ( v_j ← Op_HC)                           // implement Hyper-edges' Consolidation operation
(2)     if (v_i ∈ ε_m^k || v_i ∈ ε_n^k)          // only the nodes in ε_m^k and ε_n^k implement the HC operation
(3)        {  v_i. E_ID = Op_HC.new_mHe_ID       // new consolidation main hyper-edge ID: ε_m^{k+1};
(4)           v_i. syn_round = v_i. syn_round+1; // put into the next synchronous round
(5)           return SUCCESS;                   // HC operation success and generate new main hyper-edge
(6)        }
```

Algorithm 3: Implement flow while node receiving *Op_HC* message.

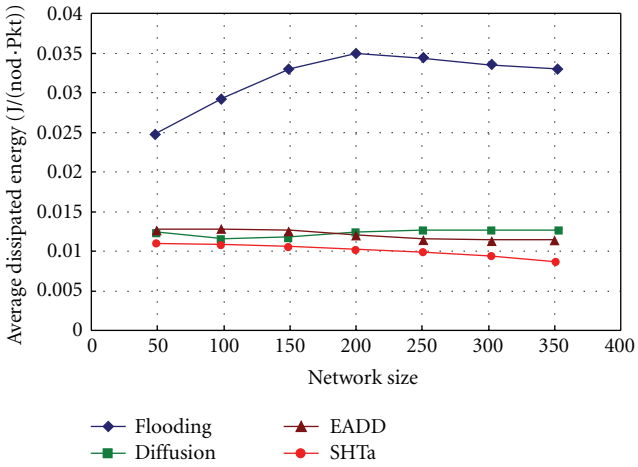


Figure 3: Average dissipated energy in different network densities.

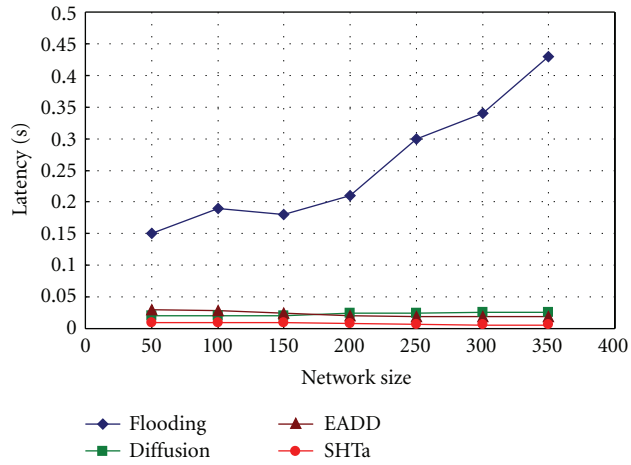


Figure 4: Average latency in different network densities.

Table 2: Maximum node degree, average nodes' degree, and the standard deviation of nodes' degree in different network size.

N	50	100	150	200	250	300	350
$\max\{d_i\}$	12	21	33	39	50	56	61
$\text{avg}\{d_i\}$	5.23	10.50	15.84	20.89	26.25	31.37	36.85
$\sigma(d)$	2.18	3.90	5.20	6.06	7.60	8.66	9.45

that of directed diffusion. It is because SHTa used hyper-edge as computing granularity, with consolidation operation, that hyper-edges become larger and less in the networks during the SHTa processing, which is completely different from the trivial nodes operation. Therefore it effectively reduces the number of overhead packets and reduces the size of the solution space, which results in the reduction of energy consumption and a prolonged network's lifetime.

Figure 4 plots the average latency observed as a function of network size. Using the shortest path, DD and EADD algorithms have lower delays than flooding algorithm. Because more available energy can give nodes a faster response time, EADD just selects these vigorous nodes as relay stations and

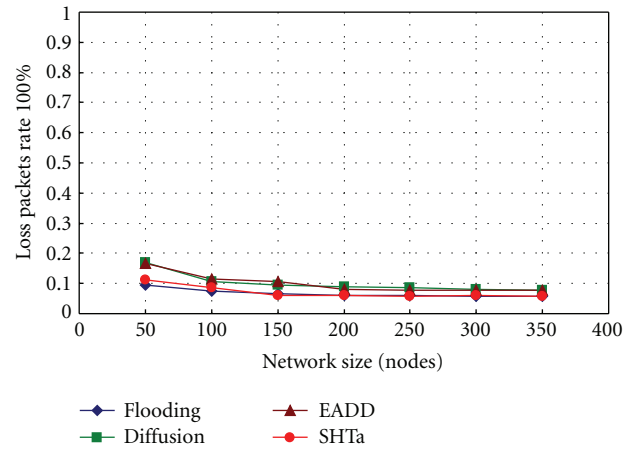


Figure 5: Packets loss rates in different densities of networks.

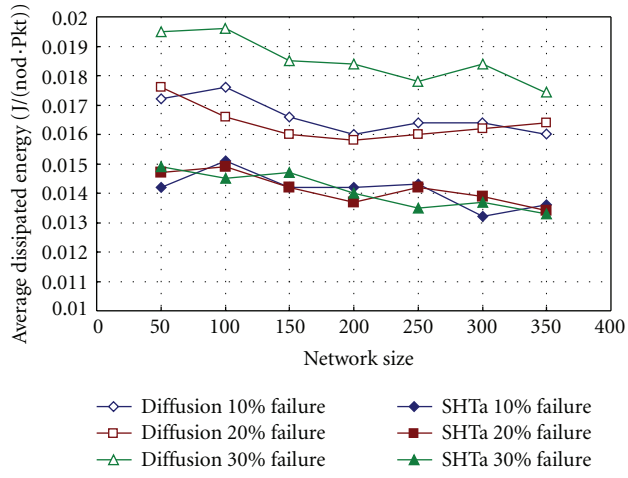


Figure 6: Average dissipated energy in dynamic experiments.

achieves lower delay than DD. Differing from the shortest edge path algorithm, SHTa uses hyper-edge, in which a set of identified nodes and edges composing multiple paths transmit information at the same time; therefore the lowest delays can be reached in four algorithms.

Figure 5 presents another performance metric-loss packets rate. SHTa and flooding have lower value than DD and EADD. Compared with signal path in DD or EADD, multiple delivery paths in one hyper-edge in SHTa or duplication flooding in the Flooding algorithm improves the transmission reliability.

We also study the impact of dynamics in wireless sensor networks with 10%, 20%, and 30% random failure nodes. Figure 6 presented the average dissipated energy per packet as a function of network size. By increasing the failure percentage from 10% to 30%, both of DD and SHTa algorithms significantly consume more energy in transmitting per packet, the increase rate of DD is 33.40%, more than 28.48% of SHTa.

Figure 7 presents the average latency measurement. The results show that SHTa also provides the lower average

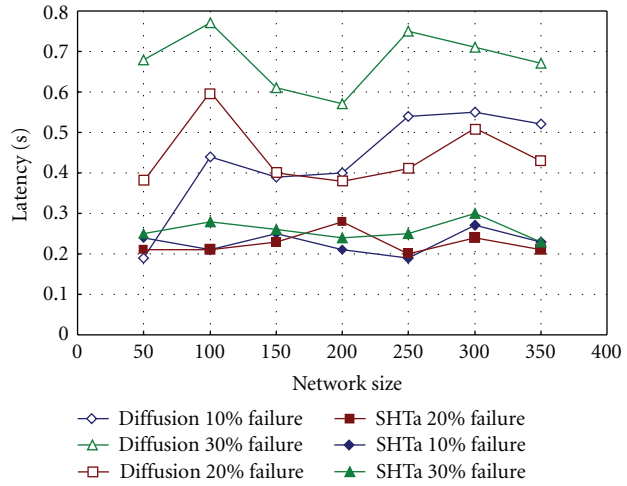


Figure 7: Average delay in dynamic experiments.

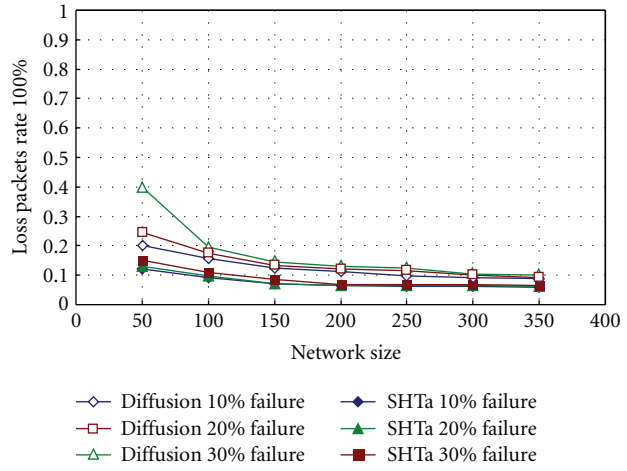


Figure 8: Packets loss rate in dynamic experiments.

delay for various fault percentages, that is, average 0.259 s when 30% of nodes fail. This is mainly because SHTa presents mutual backup delivery paths in one hyper-edge, which improves the capability of fault tolerance. In the final experiment, we evaluated the loss packets rate when the fault percentage of faulty nodes is increased. Figure 8 clearly shows that SHTa drops lower number of data packets compared with DD protocols, that is, 8.82% when 30% of nodes fail, and DD performs slightly worse, 17.12% for the same situation. All of results fairly present that SHTa is great robustness and can offers significant performance gain in networks with high fault percentage.

6. Conclusion

To consistently provide reliable communication services for machine to machine applications, scalable and smart network architecture control algorithms are needed for wireless self-organized communication networks. This paper generalizes large-scale wireless self-organized sensor networks into concise and robust hyper-graph infrastructure and proposes an algorithm called SHTa to achieve minimum spanning hyper-tree. We proved algorithm's validity with mathematical deduction and computer simulation. Based on experimental results, the SHTa algorithm can save more energy and have lower latency and packets loss rates than previous algorithms, and the algorithm is more robust in the dynamic experiments. All of these results show that SHTa is an effective technique for wireless sensor networks and M2M applications.

Acknowledgments

This work was sponsored by the National Natural Science Foundation of China No.61172014 and no. 60702037; Natural Science Foundation of Tianjin no. 12JCZDJC21300 and 09JCYBJC00800; National Basic Research Program of China (973 Program) no. 2009CB219700. The authors thank their colleagues in the laboratory.

References

[1] Y. Zhang, R. Yu, S. Xie, W. Yao, Y. Xiao, and M. Guizani, "Home M2M networks: architectures, standards, and QoS improvement," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 44–52, 2011.

[2] B. S. Choi and J. J. Lee, "Sensor network based localization algorithm using fusion sensor-agent for indoor service robot," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 3, pp. 1457–1465, 2010.

[3] I. C. Paschalidis and B. Li, "Energy optimized topologies for distributed averaging in wireless sensor networks," *IEEE Transactions on Automatic Control*, vol. 56, no. 10, pp. 2290–2304, 2011.

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

[5] M. Ding, X. Cheng, and G. Xue, "Aggregation tree construction in sensor networks," in *Proceedings of the 58th IEEE Vehicular Technology Society (VTC'03)*, vol. 4, Orlando, Fla, USA, October 2003.

[6] M. Lee and V. W. S. Wong, "An energy-aware spanning tree algorithm for data aggregation in wireless sensor networks," in *Proceedings of the IEEE Pacific RIM Conference on Communications, Computers, and Signal Processing (PACRIM'05)*, pp. 300–303, August 2005.

[7] O. Younis and S. Fahmy, "HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks," *IEEE Transactions on Mobile Computing*, vol. 3, no. 4, pp. 366–379, 2004.

[8] C. C. Tseng, H. H. Chen, K. C. Chen, S. C. Lo, and M. Y. Liu, "Quality of service-guaranteed cluster-based multihop wireless ad hoc sensor networks," *IET Communications*, vol. 5, no. 12, pp. 1698–1710, 2011.

[9] J. Baek, S. K. An, and P. Fisher, "Dynamic cluster header selection and conditional re-clustering for wireless sensor networks," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2249–2257, 2010.

[10] M. T. Hajiaghayi, N. Immorlica, and V. S. Mirrokni, "Power optimization in fault-tolerant topology control algorithms for wireless multi-hop networks," *IEEE/ACM Transactions on Networking*, vol. 15, no. 6, pp. 1345–1358, 2007.

[11] Y. Liu, Q. Zhang, and L. Ni, "Opportunity-based topology control in wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 21, no. 3, pp. 405–416, 2010.

[12] T. He, J. A. Stankovic, C. Lu, and T. F. Abdelzaher, "A spatiotemporal communication protocol for wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 16, no. 10, pp. 995–1006, 2005.

[13] E. Lee, S. Park, F. Yu, and S. H. Kim, "Data gathering mechanism with local sink in geographic routing for wireless sensor networks," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 3, pp. 1433–1441, 2010.

[14] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed diffusion:a scalable and robust communication paradigm for sensor networks," in *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking (MOBICOM'00)*, pp. 56–67, August 2000.

[15] C. Rerge, *Graph and Hypergraph*, North-Holland, Amsterdam, The Netherlands, 1973.

[16] A. Seuret, F. Michaut, J. P. Richard, and T. Divoux, "Networked control using GPS synchronization," in *Proceedings of the American Control Conference*, pp. 4195–4200, June 2006.

[17] T. Neagoe, V. Cristea, and L. Banica, "NTP versus PTP in computer networks clock synchronization," in *Proceedings of the International Symposium on Industrial Electronics 2006 (ISIE'06)*, pp. 317–362, Montreal, Canada, July 2006.

[18] J. Elson, L. Girod, and D. Estrin, "Fine-grained network time synchronization using reference broadcasts," in *Proceedings of the 5th Symposium on Operating systems Design and Implementation*, Boston, Mass, USA, December 2002.

[19] M. L. Sichitiu, C. Veerarittiphan, and Simple, "Accurate Time Synchronization for Wireless Sensor Networks," in *IEEE Wireless Communications and Networking Conference (WCNC'03)*, 2003.

[20] S. Ganeriwal, R. Kumar, S. Adlakha, and M. Srivastava, "Network-wide time synchronization in sensor networks," Technical Report UCLA, 2002.

[21] J. Choe and K. Kim, "EADD: energy aware directed diffusion for wireless sensor networks," in *Proceedings of the International Symposium on Parallel and Distributed Processing with Applications (ISPA'08)*, pp. 779–783, Sydney, Australia, December 2008.

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

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 525640, 10 pages
doi:10.1155/2012/525640

Research Article

Collaborative Relay Beamforming Strategies for Multiple Destinations with Guaranteed QoS in Wireless Machine-to-Machine Networks

Da Wang,[1] Lin Bai,[2] Xiaoning Zhang,[1] Wenyang Guan,[3] and Chen Chen[1]

[1] *State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China*
[2] *School of Electronic and Information Engineering, Beihang University, Beijing 100191, China*
[3] *College of Engineering, Swansea University, Swansea SA2 8PP, UK*

Correspondence should be addressed to Lin Bai, l.bai@buaa.edu.cn and Xiaoning Zhang, zhangxn@pku.edu.cn

Received 20 April 2012; Accepted 30 June 2012

Academic Editor: Jianhua He

The Machine-to-Machine (M2M) communications allow information exchange between machine devices, which can be carried out without any human interaction. However, there are a large number of small and low-power machine devices in the wireless M2M networks. To guarantee the quality of service (QoS) requirements of the destination devices, we study the amplify-and-forward (AF) relay beamforming, where multiple relay M2M devices can transmit signals from the source M2M device to multiple destination M2M devices. In this paper, we propose two iterative strategies to jointly optimize the source antenna selection and the collaborative relay beamforming weights with the aid of perfect channel state information (CSI). The aim of the proposed strategies is to maximize the worst-case received signal-to-interference-and-noise ratio (SINR) under two different types of relay power constraints, which are the total relay power constraint and individual relay power constraints, respectively. Using the semidefinite relaxation (SDR) technique, the optimization problem of collaborative relay beamforming can be formulated as a semidefinite programming (SDP) problem, which can be optimally solved. Simulation results validate our theoretical analysis and demonstrate that after several iterations, the performance of the proposed iterative strategies can obtain near-optimal performance.

1. Introduction

Machine devices, such as sensors, communicate with each other in machine-to-machine (M2M) networks, which do not need intervention from human beings [1]. A wireless M2M device may act as a source point and communicate data packets on behalf of the other M2M destination devices [2]. However, the wireless M2M networks are composed of a large number of small and power-limited machine devices, which can only transmit low-power signals. Due to the propagation loss between the source devices and the destination devices, it is very challenging to guarantee the quality of service (QoS) requirements of the destination devices. Recently, many increasing researches showed interest in the relay devices to deal with this problem [3]. The sources transmit signals through one or more relay devices towards the destination devices in wireless M2M networks. Relay devices can be designed to work collaboratively to guarantee the QoS requirements by forming a virtual beamforming system [4, 5].

In general, there are three types of relay schemes, including the amplify-and-forward (AF) [6], decode-and-forward (DF) [7], and compress-and-forward (CF) [8] schemes. In the AF scheme, relay nodes receive the signals transmitted from the source and forward the amplified signals to the destinations. In the DF scheme, relay nodes decode and reencod the received signals to the destinations. In the CF scheme, relay nodes send quantized and compressed signals to the destinations. Among them, the AF scheme is the most attractive for the collaborative relay beamforming due to its low implementation complexity. With the aid of channel state information (CSI), the relay nodes can work collaboratively in a way similar to a multiple antenna systems to construct virtual beams towards the destinations [9].

In this paper, we study the collaborative relay beamforming for multiple destinations in the multiple relay M2M networks, where multiple relay M2M devices form virtual beams to transmit signals from a source M2M device with multiple antennas towards multiple destination M2M devices. Firstly, we present two methods to select the source antenna to transmit the destination's signals. In Method 1, we use the greedy algorithm to select the best available channel for each destination until all the destinations are allocated properly with the source antennas; in Method 2, we balance the channel condition for each destination and its interference to the other destinations based on Method 1. Secondly, we optimize the relay beamforming weights with the aid of perfect CSI to maximize the worst-case received SINR under two different types of relay power constraints, which are the total relay power constraint and individual relay power constraints, respectively. Using the semidefinite relaxation (SDR) technique [10], we show that the optimization problem can be formulated as a semidefinite programming (SDP) problem, which can be optimally solved using interior-point methods [11]. Finally, according to the source antenna selection methods and relay beamforming schemes, we propose two iterative strategies to jointly optimize the source antenna selection and the relay beamforming weights. We show that after several iterations, the performance of the proposed iterative strategies can obtain near optimal performance.

According to our proposed iterative strategies, the source M2M device must know all the CSI of the wireless M2M networks to jointly optimize the source antenna selection and the relay beamforming weights. However in practice, the feedback channels are always limited and could be overloaded if the number of relays and destinations is large. To deal with this practical problem, Choi [12, 13] studied the distributed AF relay beamforming schemes for the single destination where the relays can work cooperatively to form a virtual beam with local CSI (The local CSI means that each relay node knows its instantaneous incoming and outgoing channel coefficients (from relay nodes' point of view). However, the CSI of the other relay nodes is assumed to be not known.) when a consensus algorithm was employed in [12], and different MMSE criteria were proposed in [13]. D. H. N. Nguyen and H. H. Nguyen [14] proposed two relay power allocation schemes in a distributed manner to minimize the sum relay power with guaranteed SNR at the destinations and maximize the SNR margin subject to individual relay power constraints, respectively, under the orthogonal transmissions of each source-destination pair. In the end of this paper, we discuss the distributed multirelay beamforming strategy to maximize the worst-case received SINR for multiple destinations, where each relay may independently calculate its beamforming weight with the local CSI.

The rest of this paper is organized as follows. In Section 2, we present a brief discussion of related works on relay beamforming which motivated our work. Section 3 describes the multiple relay M2M networks with multiple destinations. Section 4 presents two methods to select the source antenna for each destination. In Section 5, we optimize the multi-relay weights that achieve the maximum worst-case SINR under two different types of relay power constraints. Then, two iterative strategies to jointly select the source antennas and optimize the multirelay weights are proposed, respectively. Section 6 presents and analyzes the simulation results. The distributed multi-relay beamforming strategy is discussed in Section 7. Conclusions are drawn in Section 8.

Notation. Vectors are written in boldface lowercase letters, for example, $\mathbf{x}$, while matrices are denoted by boldface uppercase letters for example, $\mathbf{X}$. The superscripts $*$, T, and H stand for the conjugate, transposition, and hermitian of a complex vector or matrix, respectively. $\mathrm{Tr}(\mathbf{S})$ represents the trace of a matrix $\mathbf{S}$, and $\mathrm{diag}(s_1, s_2, \ldots, s_k)$ denotes a diagonal square matrix with $s_1, s_2, \ldots, s_k$ as the diagonal elements. $\mathbb{E}[z]$ calculates the expectation of a random entity z, while $|z|$ takes the modulus of a complex number z. The real and complex number fields are denoted by $\mathbb{R}$ and $\mathbb{C}$, respectively. In addition, $\mathbf{A} \succeq 0$ means that $\mathbf{A}$ is a hermitian semidefinite matrix, and O means the upper bound of the computational complexity.

2. Related Works

Andreev et al. [2] studied a typical smart metering M2M application and proposed a client relay scheme to improve the performance of M2M devices with poor communication link. Khajehnouri and Sayed [3] designed relay strategies in sensor networks with minimum mean square error (MMSE) performance subject to local and global relay power constraints, respectively. Elkheir et al. [15] presented an M2M communication scheme, where the source communicates with the corresponding destination node with the help of one or more relay nodes. The objective in [15] was to minimize the total relay power under specific performance constraints and the individual power of each relay. Zheng et al. [16, 17] proposed the relay beamforming schemes that achieve the maximum received signal-to-noise ratio (SNR) for a single destination under both total and individual power constraints with the aid of perfect CSI [16] and imperfect CSI [17]. In addition, distributed beamforming for AF relay nodes was investigated when a consensus algorithm was employed for cooperative beamforming in [12], and the minimum mean square error (MMSE) criteria were used in [13].

However, existing approaches focused on relay strategies with the single or multiple relay beamforming for only one destination in wireless M2M networks or sensor networks. The relay M2M devices can be used to support multiple destinations for the M2M communications. Nguyen et al. [18, 19] studied the optimal relay beamforming schemes in a multiple AF relay cooperating network with multiple source-destination pairs. The aim in [18] was to minimize the total relay power under the signal-to-interference-and-noise ratio (SINR) requirement at each destination. In [19], Nguyen et al. minimized the sum relay power with the guaranteed QoS in terms of SNR at the destinations in orthogonal channels. Liu and Petropulu [20] studied the relay beamforming to minimize the sum transmit power at the relays with multiple pairs of the source and destination, assisted by multiple relays. Zheng and Blostein [21] considered the collaborative relay beamforming system to minimize the sum power at the

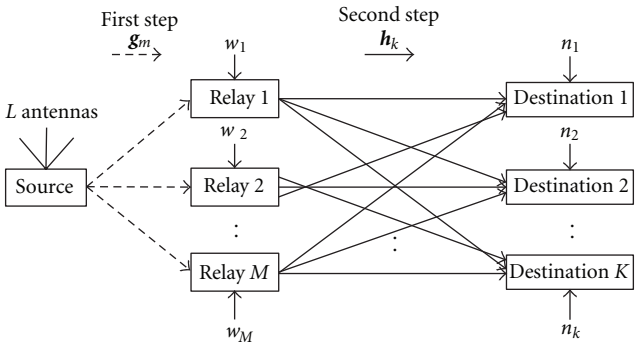


Figure 1: The multiple relay M2M networks with multiple destinations.

relays and the transmit power at the source, where a single source with multiple antennas transmits signals to multiple users through multiple single-antenna relays. Phan et al. [22] studied the joint power allocation and AF relay selection problem to maximize the destination rates and minimize the total transmit power at the relays, respectively, for multiple destinations in multiple orthogonal source-destination pair networks.

The main difference between our work and the works in [18–22] is that we present two methods to select the source antenna and optimize the relay beamforming weights to maximize the worst-case received SINR under two different types of relay power constraints, which are the total relay power constraint and individual relay power constraints, respectively, while the proposals in [18–21] were to minimize the total relay transmit power under the QoS of destinations, and Phan et al. [22] maximized the destination rates and minimized the total transmit power at the relays, respectively.

3. System Model

We consider a multiple relay M2M network consists of a source with L antennas, M relay nodes, and K destinations each with a single antenna as shown in Figure 1, where $K \leq L$. (If $K > L$, the source can select destinations to transmit the signals to exploit selection diversity among destinations [23].) Since we consider the small and power-limited M2M devices, there is no direct link between the source and destinations. The AF relay nodes form virtual beams to transmit the source signals towards multiple destinations on the same frequency band. Each destination may receive interference signals of other destinations from relay nodes. We assume that the coefficients of the channels from the source to the mth relay and the mth relay to the destination k denote $g_{l,m} \in \mathbb{C}$ and $h_{k,m} \in \mathbb{C}$, respectively. We consider a flat fading channel model, where all the channel coefficients, including $\mathbf{g}_m = [g_{1,m}, g_{2,m}, \ldots, g_{L,m}]$, $m = 1, 2, \ldots, M$ and $\mathbf{h}_k = [h_{k,1}, h_{k,2}, \ldots, h_{k,M}]$, $k = 1, 2, \ldots, K$ are independent in our multiple relay M2M networks. For convenience, the propagation losses are also contained in $\mathbf{g}_m$ and $\mathbf{h}_k$ to reflect the relative locations of the relays and the destinations.

It is assumed that the AF relay protocol has two steps. During the first step, the source selects the best available antenna to broadcast the kth destination's signal s_k to the relay nodes. Two methods to select the best antenna for the destination k will be presented in Section 4. We assume that the destinations' signals are independent with $\mathbb{E}[s_k] = 0$ and $\mathbb{E}[|s_k|^2] = P_k$. The received signal at the relay node m is given by

$$r_m = \mathbf{g}_m \sum_{k=1}^{K} \mathbf{d}_k s_k + w_m, \tag{1}$$

where w_m is independent white Gaussian noise at relay node m with zero mean and variance of N_m. The column vector $\mathbf{d}_k$ is the vector that represents the source antenna selection for the kth destination.

Then, during the second step, the relay nodes forward the received signals to the destinations using complex weights v_m, $m = 1, 2, \ldots, M$ for all destinations for the collaborative relay beamforming. Thus, the transmitted signal vector from the mth relay node is

$$x_m = v_m r_m. \tag{2}$$

The destination k receives the additive signals from all the relay nodes, which is given by

$$\begin{aligned} y_k &= \sum_{m=1}^{M} h_{k,m} v_m r_m + n_k \\ &= \sum_{m=1}^{M} h_{k,m} v_m \mathbf{g}_m \mathbf{d}_k s_k + \sum_{m=1}^{M} h_{k,m} v_m \mathbf{g}_m \sum_{j=1, j\neq k}^{K} \mathbf{d}_j s_j \\ &\quad + \sum_{m=1}^{M} h_{k,m} v_m w_m + n_k, \end{aligned} \tag{3}$$

where n_k is independent white Gaussian noise at the kth destination with zero mean and variance of N_k. Thus, the received SINR of the destination k is given by

$$\begin{aligned} \mathrm{SINR}_k &= \frac{P_k \left| \sum_{m=1}^{M} h_{k,m} v_m \mathbf{g}_m \mathbf{d}_k \right|^2}{\sum_{j=1, j\neq k}^{K} P_j \left| \sum_{m=1}^{M} h_{k,m} v_m \mathbf{g}_m \mathbf{d}_j \right|^2 + P_n} \\ &= \frac{P_k \left| \mathbf{b}_k^T \mathbf{d}_k \right|^2}{\sum_{j=1, j\neq k}^{K} P_j \left| \mathbf{b}_k^T \mathbf{d}_j \right|^2 + P_n}, \end{aligned} \tag{4}$$

where

$$\begin{aligned} P_n &= \sum_{m=1}^{M} |h_{k,m}|^2 |v_m|^2 N_m + N_k, \\ \mathbf{b}_k &= \left[\sum_{m=1}^{M} h_{k,m} v_m g_{1,m}, \ldots, \sum_{m=1}^{M} h_{k,m} v_m g_{L,m} \right]^T. \end{aligned} \tag{5}$$

4. Source Antenna Selection for Each Destination

In this section, we present two methods to optimize the source-antenna selecting vector $\mathbf{d}_k$ for the kth destination with fixed relay weights: (1) a greedy algorithm, that is, select the best available channel for each destination until all the destinations are allocated properly with the source antennas; (2) balance the channel condition for each destination k and its interference to the other destinations.

In Method 1, we first choose the element with the maximum magnitude denoted as $G_{l_k,k}$ from the $L \times K$ composite channel matrix $\mathbf{G} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K]$. Then the antenna l_k is selected to transmit signals to the destination k. The channel coefficients between the rest antennas and destinations form a new $(L-1)\times(K-1)$ channel matrix. We can repeat the above processes K times until each destination is allocated with one antenna.

In Method 2, when we select the l_kth antenna to transmit signals towards the kth destination using Method 1, we consider its interference to the other destinations, that is, $\sum_{m=1}^{M} h_{j,m} v_m g_{l_k,m}$, $j \neq k$, which are the l_kth row elements of the composite channel matrix $\mathbf{G}$. If all the magnitudes of the interference are less than or equal to the interference threshold ξ, we keep the channel and allocate the antenna for the corresponding destination on this channel. When one of the magnitudes of the interference is larger than the threshold ξ, we drop the selected channel and choose the second largest magnitude element from the composite channel matrix. We repeat the above processes until all the magnitudes of the interference are less than or equal to the threshold ξ. If the feasible solution can not be obtained under the interference threshold ξ, then we may double the threshold value and repeat the above processes to select another antenna for the destination.

In our methods, the computational complexity of selecting one antenna scales linearly with the dimension of channel matrix $\mathbf{G}$, that is, KL. Since we need to select K antennas for K destinations from the channel matrixes whose dimensions decrease step by step, the computational complexity of our methods is at most $O(K!(L!/(L-K)!))$.

5. Iterative Strategies for Max-Min SINR with Relay Power Constraints

In this section, we assume that the source-antenna selection vector $\mathbf{d}_k$ for each destination k has been already chosen by the methods in Section 4. Thus, we assume that the l_kth antenna transmits s_k to the kth destination. We can rewrite the expression of the received SINR at the destination k in (4) as

$$\begin{aligned}\mathrm{SINR}_k &= \frac{P_k \left| \sum_{m=1}^{M} h_{k,m} g_{l_k,m} v_m \right|^2}{\sum_{j=1, j\neq k}^{K} P_j \left| \sum_{m=1}^{M} h_{k,m} g_{l_j,m} v_m \right|^2 + P_n} \\ &= \frac{P_k \left| \mathbf{f}_{l_k}^T \mathbf{v} \right|^2}{\sum_{j=1, j\neq k}^{K} P_j \left| \mathbf{f}_{l_j}^T \mathbf{v} \right|^2 + \mathbf{v}^H \mathbf{H}_k \mathbf{v} + N_k},\end{aligned} \tag{6}$$

where

$$\begin{aligned}\mathbf{v} &= [v_1, v_2, \ldots, v_M]^T, \\ \mathbf{f}_{l_k} &= [h_{k,1} g_{l_k,1}, h_{k,2} g_{l_k,2}, \ldots, h_{k,M} g_{l_k,M}]^T, \\ \mathbf{H}_k &= \mathrm{diag}\left(|h_{k,1}|^2 N_1, |h_{k,2}|^2 N_2, \ldots, |h_{k,M}|^2 N_M \right).\end{aligned} \tag{7}$$

In the following, we optimize the multi-relay beamforming weights to maximize the worst-case received SINR for multiple destinations and propose the iterative strategies under two different types of relay power constrained in the multiple relay M2M network. We first investigate the case where the total relay power is limited, and then study the scenario where the individual relay transmit power is constrained.

5.1. Total Relay Power Constraint. In this subsection, we optimize the multi-relay beamforming weights for multiple destinations under a total relay power constraint P_{total}. Then we present the iterative strategy to jointly optimize the source-antenna selection and the relay beamforming weights under the total relay power constraint in the multiple relay M2M networks.

For the optimum, we can formulate the problem as

$$\begin{aligned}&\max_{\mathbf{v}} \min_{k} \quad \mathrm{SINR}_k \\ &\text{s.t.} \qquad\quad P_t \leq P_{\text{total}},\end{aligned} \tag{8}$$

where P_t is the total transmission power of relay nodes, it can be presented as

$$\begin{aligned}P_t &= \sum_{m=1}^{M} |v_m|^2 \mathbb{E}\left[|r_m|^2 \right] \\ &= \sum_{m=1}^{M} |v_m|^2 \left(\sum_{k=1}^{K} |g_{l_k,m}|^2 P_k + N_m \right) \\ &= \mathbf{v}^H \mathbf{E} \mathbf{v},\end{aligned} \tag{9}$$

$$\mathbf{E} = \mathrm{diag}\left(\sum_{k=1}^{K} |g_{l_k,1}|^2 P_k + N_1, \ldots, \sum_{k=1}^{K} |g_{l_k,M}|^2 P_k + N_M \right). \tag{10}$$

In order to guarantee the QoS for multiple destinations, we consider the maximization of the worst SINR among all destinations in (8). However, the object function of the optimization problem (8) is quasi-convex. Directly solving Problem (8) is usually difficult due to its nonconvexity. We could adopt the bisection search algorithm to efficiently find the optimal solution of Problem (8). Specifically, we could solve a sequence of the following feasibility problems for each given Γ [11].

$$\begin{aligned}&\min_{\mathbf{v}} \quad P_t \\ &\text{s.t.} \quad \begin{cases} P_t \leq P_{\text{total}}, \\ \mathrm{SINR}_k \geq \Gamma, \qquad k = 1, 2, \ldots, K. \end{cases}\end{aligned} \tag{11}$$

Denote the optimal solution of (8) by γ_{opt}. If $\Gamma = \gamma_{\text{opt}}$ in (11), then the optimal solution $\mathbf{v}$ of (11) is also optimal for (8). Thus we can repeatedly search for the optimal Γ using the bisection search method [11]. In the following, we focus on solving the problem in (11) for a given Γ.

With the received SINR in (6) and the total relay transmit power in (9), we can express Problem (11) as

$$\min_{\mathbf{v}} \quad \mathbf{v}^H \mathbf{E} \mathbf{v}$$

$$\text{s.t.} \quad \begin{cases} \mathbf{v}^H \mathbf{E} \mathbf{v} \le P_{\text{total}} \\ \dfrac{P_k}{\Gamma} \left| \mathbf{f}_{l_k}^T \mathbf{v} \right|^2 - \displaystyle\sum_{j=1, j \ne k}^{K} P_j \left| \mathbf{f}_{l_j}^T \mathbf{v} \right|^2 \\ \quad -\mathbf{v}^H \mathbf{H}_k \mathbf{v} \ge N_k, \qquad k = 1, 2, \ldots, K. \end{cases} \tag{12}$$

Using the definition $\mathbf{V} \triangleq \mathbf{v}\mathbf{v}^H$, we can reformulate Problem (12) as

$$\min_{\mathbf{V}} \quad \text{Tr}(\mathbf{E}\mathbf{V})$$

$$\text{s.t.} \quad \begin{cases} \mathbf{V} \succeq \mathbf{0}, \\ \text{rank}(\mathbf{V}) = 1, \\ \text{Tr}(\mathbf{E}\mathbf{V}) \le P_{\text{total}}, \\ \dfrac{1}{\Gamma} \text{Tr}(\mathbf{F}_k \mathbf{V}) - \displaystyle\sum_{j=1, j \ne k}^{K} \text{Tr}\left(\mathbf{F}_j \mathbf{V}\right) \\ \quad - \text{Tr}(\mathbf{H}_k \mathbf{V}) \ge N_k, \qquad k = 1, 2, \ldots, K, \end{cases} \tag{13}$$

where $\mathbf{F}_k = P_k(\mathbf{f}_{l_k}^* \mathbf{f}_{l_k}^T)$ and $\mathbf{V} \succeq 0$ means that $\mathbf{V}$ is a hermitian semidefinite matrix. Note that the rank constraint $\text{rank}(\mathbf{V}) = 1$ comes from $\mathbf{V} \triangleq \mathbf{v}\mathbf{v}^H$, which is not convex. Thus, the optimization problem (13) is not convex. We can use the SDR technique to relax Problem (13). Here, we drop the rank-one constraint and obtain the following problem.

$$\min_{\mathbf{V}} \quad \text{Tr}(\mathbf{E}\mathbf{V})$$

$$\text{s.t.} \quad \begin{cases} \mathbf{V} \succeq \mathbf{0} \\ \text{Tr}(\mathbf{E}\mathbf{V}) \le P_{\text{total}}, \\ \dfrac{1}{\Gamma} \text{Tr}(\mathbf{F}_k \mathbf{V}) - \displaystyle\sum_{j=1, j \ne k}^{K} \text{Tr}\left(\mathbf{F}_j \mathbf{V}\right) \\ \quad - \text{Tr}(\mathbf{H}_k \mathbf{V}) \ge N_k, \qquad k = 1, 2, \ldots, K. \end{cases} \tag{14}$$

Note that Problem (14) becomes an SDP problem, which can be optimally solved using interior-point methods, whose computational complexity is polynomial [11]. As shown in [24], for a given solution accuracy ϵ, the interior-point methods need at most $O(\beta(\alpha^3 + \beta\alpha^2 + \beta^2)\sqrt{\alpha}\log(1/\epsilon))$ arithmetic operations, where α is the dimension of the matrix variable, and β is the number of constraints. In our SDP problem in (14), the dimension of the matrix variable is $\alpha = M$, and the number of constraints is $\beta = K$. Therefore, the computational complexity of Problem (14) is $O(K^3 M^{3.5} \log(1/\epsilon))$.

The solution of the SDP problem in (14) provides a lower bound of the objective function in the original problem in (13) due to excluding the rank-one constraint. However, in our extensive simulations, we have never observed that the optimal solution of Problem (14) has a rank higher than one. The similar observation was also reported in [18] and [25] to design the optimal beamforming schemes. Thus, the original problem in (13) can be optimally solved. For the cases where Problem (14) has a rank higher than one, several randomization techniques [26] can be used to provide a good approximation solution $\tilde{\mathbf{v}}$ from the optimal solution $\mathbf{V}$ of Problem (14).

Based on the above results, we propose an iterative strategy (see, Algorithm 1) to jointly select the source antennas and optimize the multi-relay beamforming weights with the total relay power constraint for multiple destinations.

According to [11], the bisection search method requires $\lceil \log_2((\gamma_{UB} - \gamma_{LB})/\varepsilon) \rceil$ iterations. Thus, the complexity of Algorithm 1 is $O(K!(L!/(L-K)!)K^3 M^{3.5} \log(1/\epsilon) \lceil \log_2((\gamma_{UB} - \gamma_{LB})/\varepsilon) \rceil)$ for each iteration. Since the best available channel is selected for each destination in Step 4, and the worst-case SINR is maximized from Step 5 to Step 18, after several iterations, Algorithm 1 will converge to a fixed point. We can observe this convergence in the simulations.

5.2. Individual Relay Power Constraints. In this subsection, we study the scenario where each relay power is constrained by P_m^i, $m = 1, 2, \ldots, M$. Note that the individual relay power constraints are important when the battery lifetime of the relay nodes is limited. Then we present the iterative strategy under individual relay power constraints in the multiple relay M2M networks.

We can formulate the problem as

$$\begin{aligned} &\max_{\mathbf{v}} \min_{k} \quad \text{SINR}_k \\ &\text{s.t.} \qquad P_m \le P_m^i, \quad m = 1, 2, \ldots, M, \end{aligned} \tag{15}$$

where P_m is the transmission power of relay node m, it can be given by

$$P_m = |v_m|^2 \mathbb{E}\left[|r_m|^2\right] = |v_m|^2 \left(\sum_{k=1}^{K} \left| g_{l_k,m} \right|^2 P_k + N_m \right). \tag{16}$$

Since the optimization problem in (15) is quasi-convex, we can also obtain the optimal solution of (15) by repeatedly solving the following problem using the bisection search method to search for the Γ [11].

$$\begin{aligned} &\min_{\mathbf{v}} \quad P_t \\ &\text{s.t.} \quad \begin{cases} P_m \le P_m^i, & m = 1, 2, \ldots, M, \\ \text{SINR}_k \ge \Gamma, & k = 1, 2, \ldots, K. \end{cases} \end{aligned} \tag{17}$$

1: Initialize γ_{LB}, γ_{UB} and P_{total}. For some given $\varepsilon > 0$
2: **while** $\gamma_{UB} - \gamma_{LB} > \varepsilon$ **do**
3: Initialize the multirelay beamforming weights $v_m = c_m e^{j\theta_m}$, $m = 1, 2, \ldots, M$, where $c_m = P_{\text{total}}/M$. θ_m is a uniform random variable chosen from the interval $[0, 2\pi]$.
4: Apply Method 1 or Method 2 in Section 4 to select the source antenna for each destination with the current fixed relay weights.
5: Optimally solve Problem (14) with $\Gamma = (\gamma_{LB} + \gamma_{UB})/2$.
6: If Problem (14) is feasible, then obtain the optimal solution $\mathbf{V}_{\text{opt}}$.
7: **if** $\text{rank}(\mathbf{V}_{\text{opt}}) = 1$ **then**
8: obtain the multirelay beamforming weights vector $\mathbf{v}$, which is the optimal solution of the original problem (11).
9: **else**
10: obtain the multirelay beamforming weights vector $\mathbf{v}$ using the randomization techniques.
11: **end if**
12: If the total relay power is sufficiently close to a fixed point, or the iteration number exceeds a predetermined number, then stop the iteration and go to Step 13. Otherwise go back to Step 4.
13: **if** Problem (14) is feasible **then**
14: $\gamma_{LB} = \Gamma$
15: **else**
16: $\gamma_{UB} = \Gamma$
17: **end if**
18: **end while**
19: **return** the maximization of the worst-case SINR Γ and the multirelay beamforming weights vector $\mathbf{v}$.

Algorithm 1: Iterative Strategy under the total relay power constraint.

With the received SINR in (6) and the mth relay transmit power in (16), we can express Problem (17) as

$$\min_{\mathbf{v}} \quad \mathbf{v}^H \mathbf{E} \mathbf{v}$$

$$\text{s.t.} \quad \begin{cases} |v_m|^2 \\ \quad \times \left(\sum_{k=1}^{K} \left| g_{l_k,m} \right|^2 P_k + N_m \right) \le P_m^i, \\ \dfrac{P_k}{\Gamma} \left| \mathbf{f}_{l_k}^T \mathbf{v} \right|^2 - \sum_{j=1, j \ne k}^{K} P_j \left| \mathbf{f}_{l_j}^T \mathbf{v} \right|^2 \\ \quad -\mathbf{v}^H \mathbf{H}_k \mathbf{v} \ge N_k, \qquad m = 1, 2, \ldots, M, \\ \qquad\qquad\qquad\qquad k = 1, 2, \ldots, K. \end{cases} \tag{18}$$

Using the definition $\mathbf{V} \triangleq \mathbf{v}\mathbf{v}^H$, $\mathbf{F}_k = P_k(\mathbf{f}_{l_k}^* \mathbf{f}_{l_k}^T)$ and $A = (\sum_{k=1}^{K} |g_{l_k,m}|^2 P_k + N_m)$, we can reformulate Problem (18) as

$$\min_{\mathbf{V}} \quad \text{Tr}(\mathbf{E}\mathbf{V})$$

$$\text{s.t.} \quad \begin{cases} \mathbf{V} \succeq 0, \\ \text{rank}(\mathbf{V}) = 1, \\ [\mathbf{V}]_{m,m} \le \dfrac{P_m^i}{A}, \qquad m = 1, 2, \ldots, M, \\ \dfrac{1}{\Gamma} \text{Tr}(\mathbf{F}_k \mathbf{V}) - \sum_{j=1, j \ne k}^{K} \text{Tr}\left(\mathbf{F}_j \mathbf{V}\right) \\ \quad - \text{Tr}(\mathbf{H}_k \mathbf{V}) \ge N_k, \qquad k = 1, 2, \ldots, K, \end{cases} \tag{19}$$

where $[\mathbf{V}]_{m,m}$ denotes the (m, m)th element of $\mathbf{V}$ and $\mathbf{V} \succeq 0$ means that $\mathbf{V}$ is a hermitian semidefinite matrix. Note that the constraint $\text{rank}(\mathbf{V}) = 1$ is not convex. As mentioned in Section 5.1, we also use the SDR technique to relax Problem (19) and obtain the following SDP problem.

1: Initialize γ_{LB}, γ_{UB} and P_m^i. For some given $\varepsilon > 0$
2: **while** $\gamma_{UB} - \gamma_{LB} > \varepsilon$ **do**
3: Initialize the multirelay beamforming weights $v_m = c_m e^{j\theta_m}$, $m = 1, 2, \ldots, M$, where $c_m = P_m^i$. θ_m is a uniform random variable chosen from the interval $[0, 2\pi]$.
4: Apply Method 1 or Method 2 in Section 4 to select the source antenna for each destination with the current fixed relay weights.
5: Optimally solve Problem (20) with $\Gamma = (\gamma_{LB} + \gamma_{UB})/2$.
6: If Problem (20) is feasible, then obtain the optimal solution $\mathbf{V}_{\text{opt}}$.
7: As shown in Algorithm 1, obtain the multirelay beamforming weights vector $\mathbf{v}$ from $\mathbf{V}_{\text{opt}}$.
8: If the total relay power is sufficiently close to a fixed point, or the iteration number exceeds a predetermined number, then stop the iteration and go to Step 9. Otherwise go back to Step 4.
9: **if** Problem (20) is feasible **then**
10: $\gamma_{LB} = \Gamma$
11: **else**
12: $\gamma_{UB} = \Gamma$
13: **end if**
14: **end while**
15: **return** the maximization of the worst-case SINR Γ and the multirelay beamforming weights vector $\mathbf{v}$.

Algorithm 2: Iterative Strategy under the individual relay power constraints.

$$\min_{\mathbf{V}} \quad \text{Tr}(\mathbf{EV})$$

$$\text{s.t.} \quad \begin{cases} \mathbf{V} \succeq 0, \\ [\mathbf{V}]_{m,m} \le \dfrac{P_m^i}{A}, & m = 1, 2, \ldots, M, \\ \dfrac{1}{\Gamma}\text{Tr}(\mathbf{F}_k\mathbf{V}) - \displaystyle\sum_{j=1, j\neq k}^{K} \text{Tr}\left(\mathbf{F}_j\mathbf{V}\right) \\ \quad - \text{Tr}(\mathbf{H}_k\mathbf{V}) \ge N_k, & k = 1, 2, \ldots, K. \end{cases} \tag{20}$$

Note that Problem (20) can be optimally solved with a polynomial time complexity [11]. As shown in Section 5.1, the computational complexity of Problem (20) is the same as that of Problem (14), which is $O(K^3M^{3.5}\log(1/\epsilon))$. In our extensive simulations, we have never observed that the optimal solution of SDP relaxation problem in (20) has a rank higher than one. Thus, the original problem in (19) can be optimally solved. For the cases where Problem (20) has a rank higher than one, we can also use randomization techniques [26] to provide a good approximation solution $\tilde{\mathbf{v}}$.

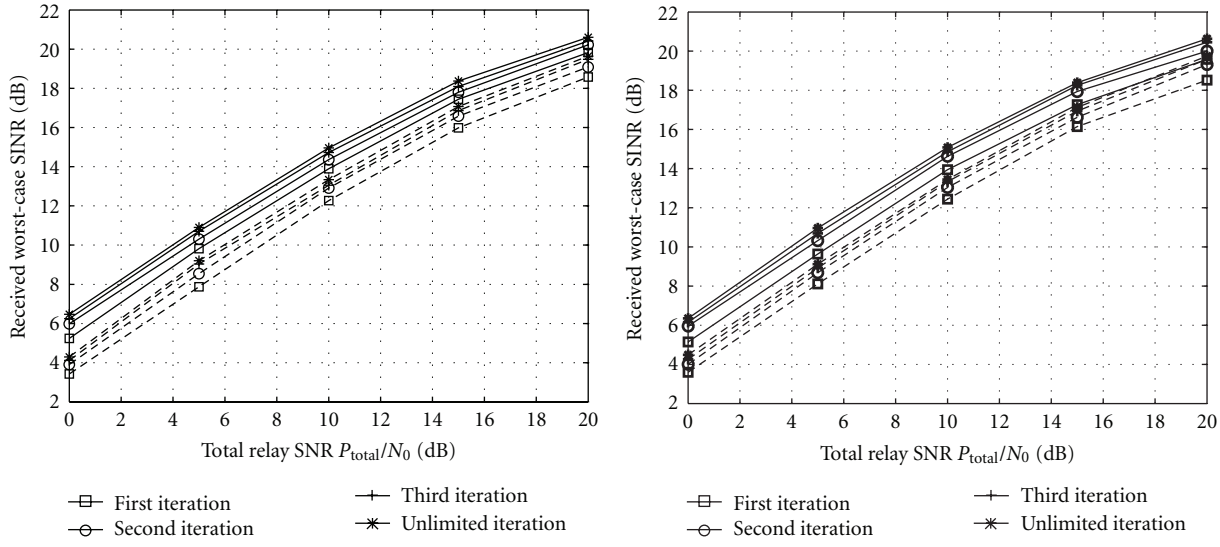
In the following, we propose another iterative strategy to jointly optimize the source-antenna selection and the multirelay beamforming weights with the individual relay power constraints for multiple destinations.

According to [11], the complexity of Algorithm 2 is $O(K!(L!/(L-K)!)K^3M^{3.5}\log(1/\epsilon)\lceil\log_2((\gamma_{UB}-\gamma_{LB})/\varepsilon)\rceil)$ for each iteration. We can observe that Algorithm 2 will converge to a fixed point after several iterations in the simulations.

6. Simulation Results

In this section, we present simulation results to validate the performance of our iteration strategies for multiple destinations in the multiple relay M2M networks. It is assumed that the flat channel coefficients $\mathbf{g}_m = [g_{1,m}, \ldots, g_{L,m}]$, $m = 1, 2, \ldots, M$ and $\mathbf{h}_k = [h_{k,1}, \ldots, h_{k,M}]$, $k = 1, 2, \ldots, K$ are independent Rayleigh random variables with zero mean and unit variance. For convenience, it is assumed that the noise at all nodes has the same power density, that is, $N_m = N_k = N_0$. In the scenario of individual relay power constraints, we assume that each relay has the same power constraint, that is, $P_m^i = P_{\text{total}}/M$. The source transmission SNR is set to $P_k/N_0 = 10\,(\text{dB})$, $k = 1, 2, \ldots, K$, which is the same for all destinations. We use CVX [27] to numerically solve the SDP problems.

In Figure 2, we compare the performance of the proposed iteration strategies for different values of the total relay SNR of P_{total}/N_0. The multiple relay M2M network has one source with 8 antennas, 20 relays, and 5 destinations. We run simulations for our proposed strategies using Method 1 and Method 2 mentioned in Section 4 to select the antenna for each destination in Figures 2(a) and 2(b), respectively. The results in Figures 2(a) and 2(b) show that the performance increases as the number of iterations increases and after 3 iterations, the proposed iteration strategies can obtain



(a) Method 1 for selecting source antennas

(b) Method 2 for selecting source antennas

FIGURE 2: Expected worst-case received SINR of the proposed iteration strategies with the total relay power constraint (solid lines) and the individual relay power constraints (dashed lines) for different numbers of iterations.

near-optimal performance. Thus, 3 iterations could be enough for the proposed iteration strategies in various multiple destination scenarios. The results indicate that the performance of the proposed iteration strategy with the total power constraint (solid lines) achieves 1∼2 (dB) higher SINR than that of the iteration strategy with the individual power constraints (dashed lines). In addition, comparing between Figures 2(a) and 2(b), the performance of our proposed strategies using Method 2 is slightly superior to that of our proposed strategies using Method 1, as Method 2 considers the interference to the other destinations from a selected channel based on Method 1.

Figure 3 demonstrates the worst-case received SINR results for the proposed iteration strategies with different values of the total relay SNR of P_{total}/N_0. In Figure 3, the curves with the square symbols represent that the multiple relay M2M network has 8 antennas, 20 relays, and 5 destinations; the curves with the circle symbols represent that the multiple relay M2M network has 6 antennas, 10 relays, and 5 destinations; the curves with the star symbols represent that the multiple relay M2M network has 6 antennas, 5 relays, and 5 destinations. We run simulations for the proposed iteration strategies using Method 1 to select the source antenna for each destination with 3 iterations in Figure 3. As shown in the figure, in order to improve the performance of the multiple relay M2M network, we can increase the number of antennas or relays with the same relay power constraint. It is also confirmed that the performance of the proposed iteration strategy with the total power constraint (solid lines) is superior than that of the iteration strategy with the individual power constraints (dashed lines). The reason is that in the case of total power constraints, more power can be allocated to the relay nodes with good channel gains by optimization.

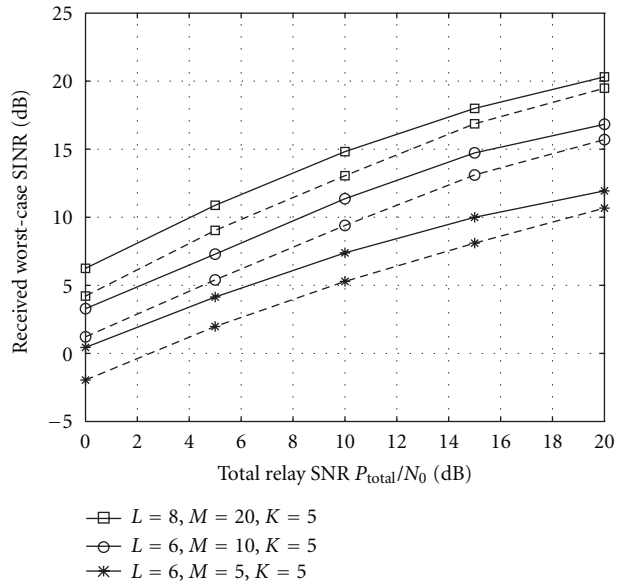


FIGURE 3: Expected worst-case received SINR of the proposed iteration strategies with the total relay power constraint (solid lines) and the individual relay power constraints (dashed lines).

7. Discussion

In the above results, the source M2M device must calculate the collaborative relay beamforming weights with the aid of perfect CSI and then broadcast them to the relays, respectively. Thus, the perfect CSI from the source to relays and from relays to the destinations needs to provide feedback to the source M2M device. The source can also know the

noise level of relays and destinations from feedback channels. However in practice, wireless M2M networks, the feedback channels are always limited and could be overloaded if the number of relays and destinations is large. To deal with this issue, we can develop the distributed relay beamforming strategies to reduce the CSI exchange and allow each relay to learn its beamforming weight based on the local CSI. It is assumed that each relay node can learn the local CSI from the source by training and from destinations by feedback, respectively, and measure its noise level. In [16], a distributed strategy with local CSI is proposed for maximizing the received SNR at the single destination based on the Karush-Kuhn-Tucker (KKT) analysis. For the optimization problem of the received SINR at destinations, most previous works have focused on reducing the CSI exchange overhead. Instead of the instant CSI feedback, statistical channel information and partial CSI feedback are used in [28]. In [29], the CSI exchange overhead is drastically reduced by partitioning relays into two clusters which have different needs for CSI for calculating the relay beamforming weights. However, distributed relay beamforming strategies with the local CSI for maximizing the worst-case received SINR at multiple destinations have not been intensively discussed. In order to derive distributed beamforming strategies for the received SINR problem in the nonorthogonal systems, we can consider another criterion, that is, MMSE criterion to construct distributed relay beamforming strategies in [13] and [30]. It has been proved to be adopted for the single destination in the multiple relay systems to develop distributed beamforming strategies with the local CSI [13] and can be expected to be suitable for multiple destinations [30]. It is demonstrated that the MMSE minimization problems can be converted into the maximization of the worst-case SINR problems. Thus, in the future work, we can develop the maximization of the worst-case SINR distributed relay beamforming strategies by using the MMSE distributed beamforming strategies.

8. Conclusion

In this paper, we proposed two iterative strategies to jointly optimize the source-antenna selection and the AF relay beamforming weights with the guaranteed QoS in terms of SINR for multiple destinations in wireless M2M networks. We presented two methods to select the source antenna for each destination: (1) use the greedy algorithm to select the best available channel for each destination; (2) balance the channel condition for each destination and its interference to the other destinations based on Method 1. The maximization of the worst-case received SINR for collaborative relay beamforming was studied under two different types of power constraints, which are the total relay power constraint and the individual relay power constraints, respectively. To obtain the solution of the optimal relay beamforming weights, an SDR technique was employed to formulate the optimization problem as an SDP problem, which can be optimally solved. Simulation results demonstrated that after 3 iterations, the performance of the proposed iterative strategies can obtain near-optimal performance. In the end of this paper, we discussed the distributed relay beamforming strategy to maximize the worst-case received SINR for multiple destinations in the practical M2M networks.

Acknowledgment

This work has been supported by the National Basic Research Program of China (973 Program, no. 2009CB320403).

References

[1] J. Zhang, L. Shan, H. Hu, and Y. Yang, "Mobile cellular networks and wireless sensor networks: toward convergence," *IEEE Communications Magazine*, vol. 50, no. 3, pp. 164–169, 2012.

[2] S. Andreev, O. Galinina, and Y. Koucheryavy, "Energy-efficient client relay scheme for machine-to-machine communication," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '11)*, pp. 1–5, Houston, Tex, USA, December 2011.

[3] N. Khajehnouri and A. H. Sayed, "Distributed MMSE relay strategies for wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 55, no. 7 I, pp. 3336–3348, 2007.

[4] A. K. Sadek, W. Su, and K. J. R. Liu, "Multinode cooperative communications in wireless networks," *IEEE Transactions on Signal Processing*, vol. 55, no. 1, pp. 341–355, 2007.

[5] M. Alatossava, A. Taparugssanagorn, V. M. Holappa, and J. Ylitalo, "Measurement based capacity of distributed MIMO antenna system in Urban microcellular environment at 5.25 GHz," in *Proceedings of the IEEE 67th Vehicular Technology Conference-Spring (VTC '08)*, pp. 430–434, Marina Bay, Singapore, May 2008.

[6] Y. W. Liang and R. Schober, "Amplify-and-forward multi-antenna beamforming with joint source-relay power constraint," in *Proceedings of the IEEE 72nd Vehicular Technology Conference Fall (VTC '10)*, pp. 1–5, Ottawa, Canada, September 2010.

[7] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, 2004.

[8] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037–3063, 2005.

[9] X. He, T. Luo, and G. Yue, "Optimized distributed MIMO for cooperative relay networks," *IEEE Communications Letters*, vol. 14, no. 1, pp. 9–11, 2010.

[10] Z. Q. Luo, W. K. Ma, A. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 20–34, 2010.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[12] J. Choi, "Distributed beamforming using a consensus algorithm for cooperative relay networks," *IEEE Communications Letters*, vol. 15, no. 4, pp. 368–370, 2011.

[13] J. Choi, "MMSE-based distributed beamforming in cooperative relay networks," *IEEE Transactions on Communications*, vol. 59, no. 5, pp. 1346–1356, 2011.

[14] D. H. N. Nguyen and H. H. Nguyen, "Power allocation in wireless multiuser multi-relay networks with distributed beamforming," *IET Communications*, vol. 5, no. 14, pp. 2040–2051, 2011.

[15] G. A. Elkheir, A. S. Lioumpas, and A. Alexiou, "Energy efficient AF relaying under error performance constraints with application to M2M networks," in *Proceedings of the IEEE 22nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC '11)*, pp. 56–60, Toronto, Canada, September 2011.

[16] G. Zheng, K. K. Wong, A. Paulraj, and B. Ottersten, "Collaborative-relay beamforming with perfect CSI: optimum and distributed implementation," *IEEE Signal Processing Letters*, vol. 16, no. 4, pp. 257–260, 2009.

[17] G. Zheng, K. K. Wong, A. Paulraj, and B. Ottersten, "Robust collaborative-relay beamforming," *IEEE Transactions on Signal Processing*, vol. 57, no. 8, pp. 3130–3143, 2009.

[18] D. H. N. Nguyen, H. H. Nguyen, and H. D. Tuan, "Distributed beamforming in relay-assisted multiuser communications," in *Proceedings of the IEEE International Conference on Communications (ICC '09)*, pp. 1–5, Dresden, Germany, June 2009.

[19] D. H. N. Nguyen, H. H. Nguyen, and T. T. Pham, "Distributed beamforming in multiuser multi-relay networks with guaranteed QoS," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '09)*, pp. 1–6, Honolulu, Hawaii, USA, December 2009.

[20] Y. Liu and A. P. Petropulu, "Cooperative beamforming in multi-source multi-destination relay systems with SINR constraints," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10)*, pp. 2870–2873, Dallas, Tex, USA, March 2010.

[21] Y. Zheng and S. Blostein, "Downlink distributed beamforming through relay networks," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '09)*, pp. 1–6, Honolulu, Hawaii, USA, December 2009.

[22] K. T. Phan, D. H. N. Nguyen, and T. Le-Ngoc, "Joint power allocation and relay selection in cooperative networks," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '09)*, pp. 1–5, Honolulu, Hawaii, USA, December 2009.

[23] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and user selection," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1478–1491, 2007.

[24] A. Mobasher, M. Taherzadeh, R. Sotirov, and A. K. Khandani, "A Near-maximum-likelihood decoding algorithm for MIMO systems based on semi-definite programming," *IEEE Transactions on Information Theory*, vol. 53, no. 11, pp. 3869–3886, 2007.

[25] M. Bengtsson and B. Ottersten, "Optimal and suboptimal transmit beamforming," in *Handbook of Antennas in Wireless Communications*, CRC Press, 2001.

[26] N. D. Sidiropoulos, T. N. Davidson, and Z. Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Transactions on Signal Processing*, vol. 54, no. 6 I, pp. 2239–2251, 2006.

[27] M. Grant and S. Boyd, "CVX users guide for CVX version 1. 21 (build 808)," 2011, http://cvxr.com/.

[28] D. Hammarwall, M. Bengtsson, and B. Ottersten, "Utilizing the spatial information provided by channel norm feedback in SDMA systems," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3278–3293, 2008.

[29] C. Esli and A. Wittneben, "Distributed multiuser cooperative network with heterogenous relay clusters," in *Proceedings of the 41st Asilomar Conference on Signals, Systems and Computers (ACSSC '07)*, pp. 829–834, Pacific Grove, Calif, USA, November 2007.

[30] S. Verdu, *Multiuser Detection*, Cambridge University Press, Cambridge, UK, 1998.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 860864, 11 pages
doi:10.1155/2012/860864

Research Article

Resource Description Language: A Unified Description Language for Network Embedded Resources

André C. Santos,[1,2] Luís D. Pedrosa,[1,3] Martijn Kuipers,[1,2] and Rui M. Rocha[1,3]

[1] *IST, Technical University of Lisbon, Avenida Professor Dr. Aníbal Cavaco Silva, 2744-016 Porto Salvo, Portugal*
[2] *INESC-ID, Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa, Rua Alves Redol 9, 1000-029 Lisboa, Portugal*
[3] *Instituto de Telecomunicações, Avenida Rovisco Pais 1, 1049-011 Lisboa, Portugal*

Correspondence should be addressed to André C. Santos, acoelhosantos@ist.utl.pt

Received 20 April 2012; Revised 10 July 2012; Accepted 10 July 2012

Academic Editor: Jianhua He

As machine-to-machine networks become larger and more pervasive, manual configuration and discovery of resources will become intractable. It is in this context that we propose the RDL, a Resource Description Language that represents a uniform way of describing embedded resources, allowing them to be shared and enabling a new class of resource-aware applications. The RDL can describe a wide range of resources, characterizing individual nodes or entire networks. It can contribute to overcome performance issues in dense networks or mobility-driven problems in highly dynamic machine-to-machine topologies by providing the means for self-adaptability and manageability, as well as opportunistic resource sharing in context-aware embedded applications. The main goal for the RDL is to define a reusable and extensible resource description specification, which can only be reached if the resources are described in a standardized format. To illustrate the feasibility of our approach, we have also developed a Java implementation of the RDL framework, as well as a TinyOS implementation targeting resource constrained platforms. Furthermore, we have developed Modulus, a modular middleware for the development of resource-aware distributed applications.

1. Introduction

Over the past few years, the increasing demand for "always on" communications has been imposing a big pressure on network architectures that not only have to provide efficient and sustainable support to an increasing range of services but also have to cope with scenarios where heterogeneity and mobility are of key importance. In parallel, the emerging smart objects concept brought along new challenges in architectures, protocols, and services where scalability, heterogeneity, and mobility are also issues to consider, perhaps even more important as they are supposed to interact with the physical world. Commonly, these systems are not well prepared for changes in the surrounding environment as with changes in the evolving resources, resulting in problems and malfunctions, which are especially critical in such resource-constrained devices that support resource-aware applications.

Machine-to-machine (M2M) systems are also experiencing a decisive evolution in this last decade. Indeed, as the first years of the new millennium went by, one could notice how solutions used in typical control applications for reaching remote embedded devices from central control hubs, using a one-to-one interconnection approach, were progressively turning into rich network-based architectures encompassing a multitude of network resources and converging towards a global mobile Internet paradigm. This Internet of embedded devices—the so called Internet of Things—relies on a multi-tier architecture mixing already operating wired and wireless networks with emerging embedded networks specially developed and deployed for M2M applications.

Beyond the traditional challenges associated with wide, local, or personal area communication networks, there are a number of challenges specifically related with M2M settings. Challenges like coping with device heterogeneity, "zero-touch" manageability, standardized plug-n-play capability,

augmented sensing support based on resource-constrained devices are among those that should be tackled efficiently in future M2M networks [1].

A number of usage models have been identified as promising for M2M applications. Examples more often cited are smart homes that use embedded systems integrated in home appliances for automation, security, entertainment, and energy management; vehicular networks used for safety, traffic information, navigation, and entertainment; healthcare in a variety of scenarios based on interconnected body area networks capable of gathering vital information from biomedical sensor networks and triggering appropriate responses. These examples are just the "tip of the iceberg" in a world of interconnected smart objects. However, as the complexity of this world increases, in terms of scale and popularity, the system's overall reliance on human intervention to determine network roles and policies in the presence of complex dynamics is no longer sustainable.

Adapting to new environments is indeed an important task for true M2M networks, involving the collection of information about neighboring nodes and the negotiation of available resources to communicate on, as well as available services [2].

Highly dynamic topologies, where new nodes can appear or existing ones can move around and shut down, are also envisaged scenarios in the networked embedded system's world. Thus, a set of mechanisms providing autoconfiguration and self-organization of network nodes are mandatory to handle topology changes. Also, given the heterogeneous nature of modern networks, interoperability issues are of extreme importance as they define how adaptable a system can be. Furthermore, due to the hierarchical nature of M2M systems architecture, the existence of network nodes where information should be aggregated and processed is practically mandatory for efficiency reasons, leading to the need to determine and configure optimal aggregation points.

Besides several tentative approaches (e.g., IEEE 1451 [3], UPnP [4]), a suitable standardized solution to describe embedded resources in a network is still lacking. In fact, solutions for network nodes autoconfiguration tend to be extremely complex given the heavy resource constraints of small embedded nodes.

In this paper, we define a unified approach for the description of resources, called the Resource Description Language (RDL), which can be used by various entities, such as resource management algorithms. In short, the RDL concept is about describing objects that do not exist as content, such as temperature sensors and queues. The RDL, for instance, can be used for cross-layering information from the physical layer, so that actions can be triggered from events happening at the lowest layer, such as a fading of the wireless channel in case of a wireless link. Another example is the use of the RDL to describe the resources in a Wireless Sensor Network (WSN) or in M2M networks. As the RDL is a neutral specification, it allows M2M communication between different devices that adhere to the RDL. The specification is robust in a way that machines can easily extract known and necessary information from the RDL and ignore or (if needed) forward the unknown parts, without the need to implement all possible uses of the specification in each participating device.

The main goal for the RDL is to define a reusable resource description specification, which can only be reached if the resources are described in a standard, interchangeable form. Since the RDL can be input to management algorithms, there is a need to separate the definition of the resources from the functionality of the algorithms that use them. This way, the algorithms can support the heterogeneity of resources, which is beneficial for opportunistic networking, where each node might try to use whatever resources that are available at any given time.

To better illustrate the benefits of using the RDL in M2M environments, we consider two use cases where the RDL may represent a significant improvement on their functionality allowing an efficient and flexible network operation.

The first use case relates to cost-effective solutions in highly dense structures comprising a large number of embedded devices providing information that should be uploaded to monitoring centers. In such situations, network capacity and devices' energy resources are at a premium, demanding for mechanisms capable of saving precious network resources. One efficient mechanism is the aggregation of information produced by the sensor devices right at the lower levels of the sensor subnetwork tree as depicted in Figure 1.

A common situation in sensing systems is often characterized by sensor devices that operate in overlapping sensing regions. In these cases, their sensor readings will have practically the same values, resulting in a waste of resources since many sensor messages are just duplicated copies of previously sent ones. For optimization purposes, a data aggregation mechanism is used.

In order to help the system to achieve optimal efficiency, namely, in the case of a large sensor field, the aggregator nodes should be configured automatically, without user intervention, whenever two streams of the same resource type pass through these nodes. This is where the RDL comes into play conveying information about resource types and requirements, and allowing aggregator nodes to configure themselves to operate over data streams upon requests (also adhering to the RDL semantics) received from upper level nodes. What is not at all obvious is whether aggregated values should be aggregated with raw data or if aggregated values should be aggregated with other aggregated values when both regard the same data type. This multilevel aggregation limits message forwarding even further, increasing energy efficiency. It is precisely the situation illustrated in Figure 1. In it, the straight red lines are raw data readings, the black dotted lines are aggregated readings, and the final small dotted line regards are the hypothetic aggregation of previously aggregated values.

The second application scenario concerns a mobility setting, commonly found in vehicular networks or in mobile healthcare applications, where M2M subnetworks may become in contact occasionally with a fixed or mobile infrastructure. We chose a road system scenario to demonstrate the usefulness of the RDL in opportunistic situations. Maintenance and safety assurance on extensive road

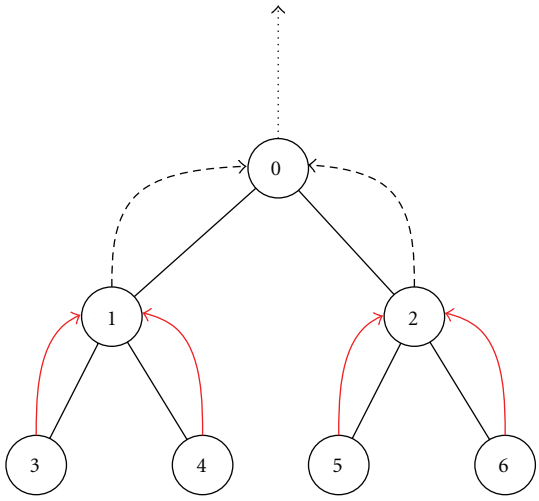


Figure 1: RDL used in multilevel autoaggregation.

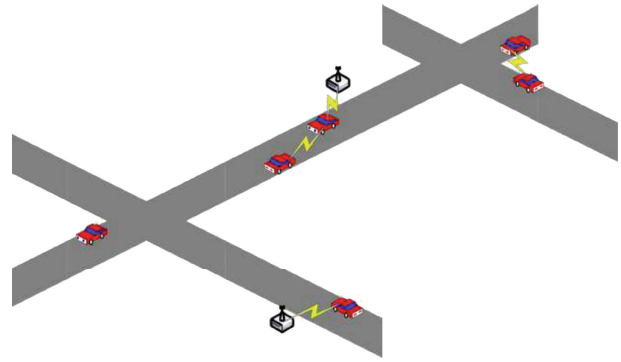


Figure 2: A road system scenario.

networks present relevant challenges for the organizations responsible for them. Highways are usually equipped with expensive monitoring equipment. Secondary and regional roads, on the other hand, are of too great extent for such constant monitoring to be cost-effective or even sustainable.

A M2M system could be the right answer for these scenarios as small, low-cost embedded sensor devices can be freely deployed wherever needed, allowing information collection with an acceptable investment strategy. There is, however, the problem of how to transmit the captured information to the monitoring center. Using public cellular communication infrastructures is certainly a possibility but carries significant operational costs. An opportunistic M2M network, as the one depicted in Figure 2, using cheap, plug-n-play On Board Units (OBUs) on regular vehicles would present a less expensive alternative. Such OBUs are already commonly deployed by governments and transit companies for vehicle identification for billing purposes (e.g., for electronic toll collection) and in many places are even mandatory. These systems could be easily repurposed to allow some degree of vehicle-to-vehicle and vehicle-to-infrastructure communications, while also increasing their processing and storage capacity. By using such enhanced OBUs as mobile carriers, sensor nodes would only have to perform local message transfers using built-in short-range radios.

As drivers do not all follow the same routes, there would be opportunities for vehicle-to-vehicle communications which could positively impact the network performance. However, there are several problems associated with such opportunistic strategies. On the one hand, when the density of vehicles becomes too high, the impact on the network performance is affected by the multiple secondary nodes within interference range of each other, all trying to gain access to the same spectrum. On the other hand, as users expect the OBUs batteries to last for several years, even with the aid of an unobtrusive energy harvesting scheme (e.g., a solar panel), the minimization of energy consumption, through a series of mechanisms (e.g., carefully choosing the next nodes, based on limited available information), is of prime importance.

This is where the RDL can play a significant role. By allowing vehicles and road-side sensor nodes to negotiate available resources in a standard way, the underlying application or middleware systems (as discussed in Section 6) can easily ensure that data is only passed onto a node that has sufficient storage and power resources to carry it. Furthermore, some OBUs (e.g., those on public transit or road maintenance vehicles) could be equipped with sensors of their own and could perform more enhanced sensor data fusion and processing. The ability to allow such systems to be incrementally developed and deployed is the RDL's main strength. Indeed, all three of the above described systems (road-side sensors, common and enhanced vehicle OBUs) could be developed independently and in a decoupled fashion.

As for the interference mitigation, common channel sensing policies for tracking white space in the spectrum are seen as an important part of opportunistic spectrum access [5]. Interfering nodes must collaborate in order to reduce the collisions, which would affect the efficiency of the spectrum utilization. In such cases, the RDL can be used as an effective method for the nodes to convey channel information to others, as the RDL can give a common and very compact description of their resources. Although the RDL itself is not an algorithm to solve the spectrum access among secondary nodes, it can be of great benefit as a sharing mechanism of information for the negotiation game in opportunistic channel access.

The RDL can be specified in two supported formats, that is, XML or KLV, and we further contribute with the development of a Java implementation for the RDL framework which embodies the translation of the RDL concepts for platforms which support the Java environment. For user interaction and development of the RDL language, a Java applet tool has been developed (RDL applet tool available at http://web.ist.utl.pt/acoelhosantos/rdl/). Within the tool, using a graphical user interface, it is possible to build an RDL graphically using a tree interface, having a clear perception of the hierarchical structure of the resource definitions and respecting the RDL element structure; import external RDL

files to be viewed as a resource tree which allows further development of already defined RDL's; export the RDL created by generating it to XML and/or KLV formats. For more computational limited devices, a TinyOS implementation of the RDL framework is currently being developed. For a stable establishment as a language, the RDL specification will not be closed, always aiming at addressing and tackling new and arising resource description necessities. Moreover, we have developed Modulus, a modular middleware for the development of resource-aware distributed applications.

The structure of the remainder of the article is as follows. In Section 2, an overview of related work is presented. Section 3 describes the RDL concept and framework. Next, Section 4 presents the RDL specification, presenting the XML and KLV format representations; followed by the resource matching algorithm in Section 5. An application case study for the RDL language, using the Modulus middleware, is shown in Section 6. Conclusions are drawn in Section 7, being complemented with future work remarks.

2. Background and Related Work

Relevant background and related work focuses on research for the description of specific resources, mainly sensors, actors, and services. These research efforts mostly concentrate on concrete areas or applications and are used to define only a subset of items regarding resource description. The RDL differentiates itself by defining a broader resource scope, enabling therefore the production of multiple focused resource-based applications. In addition to the main related works mentioned, it is also important to refer research work accomplished on the topic of resource awareness (e.g., [6]) and network policy languages (e.g., [7, 8]). Both topics focus on the need for monitoring the level and quality of resources in the operating environment and on the need of being able to control them and to react upon changes, aiming for better network resource management and security.

Relatively to sensors, attention has been given to the description of their characteristics for interoperability, sharing, and discovery. Within this topic, we point out the Semantic Specification of Sensors [9], Sensor Web Enablement [10], and IEEE 1451 [3], whose summary follows.

The work by Compton et al. [9] reviews the state of the art of semantic specification of sensors. Semantic sensor networks use declarative descriptions of sensors to promote reuse and integration and to help solve the difficulties of installing, querying, and maintaining complex, heterogeneous sensor networks.

The Sensor Web Enablement (SWE) is the definition of web service interfaces and data encodings to make sensors discoverable, taskable, and accessible through the Internet, thus enabling a standardized communication and interaction with arbitrary types of sensors and sensor systems [10]. SWE makes use of Sensor Model Language (SensorML) [11] which provides models and encodings to describe any kind of process in sensor or postprocessing systems and metadata descriptions. SensorML has been applied in research, such as in the work of Aloisio et al. [12], to build an information structure to integrate sensor networks in grid environments.

The IEEE 1451 [3] is a set of standards that describe open, common, network-independent communication interfaces for connecting both sensors and actuators to other systems or entities. The IEEE 1451 depends on the definition of the Transducer Electronic Data Sheet (TEDS) for each transducer (sensor or actuator) which contains identification, calibration, correction data, and manufacturer-related information. This allows systems to automatically identify sensors and obtain their calibration and operating parameters in an opportunistic manner [13]. The IEEE 1451 standard is already being used for several sensor-related applications, such as for networking systems of intelligent vehicles [14].

Considering services, great importance goes to the Service Location Protocol (SLP), which is a scalable framework for the discovery and selection of network services formalized as an IETF standard [15, 16]. This protocol provides a dynamic configuration mechanism that allows applications to find and use services based on required attributes and not by name or address. The objective is that internet-enabled devices become less dependent on static and manual configuration of network services. This is extremely important as computers become more portable and networks larger and more pervasive, the need to automate the location and client configuration for network services also increases [17]. Considering mobile users, who frequently change network service environments, the SLP is essential in supporting user mobility [18].

3. RDL Framework Overview

By defining the Resource Description Language as a framework, we provide an abstraction that provides the general concepts, relationships, and functionalities that can be translated and specialized for specific systems. The RDL is designed to provide a flexible resource description framework, capable of not only describing the resources per se (e.g., printers, printing services, sensors), but also of contextualizing them as capabilities or requirements. Capabilities define which resources incorporate certain characteristics, whilst requirements define which necessary characteristics the resources need to possess. Requirements may also consider alternatives to the resources it is requesting, allowing a sort of requirement priority requests. A visual overview of the framework is presented in Figure 3.

The resources in the RDL are either scalar or service resources. Scalar resources represent the most fundamental aspects of the described objects, that is, describing what the objects are and characterizing them both qualitatively and quantitatively. These scalar resources describe such things as node hardware capabilities, as well as any specific metrics or state information provided by the network layers. Service resources describe coherent packages of scalar resources that export some sort of functionality. For example, whereas a network router can describe its queues and protocol capabilities as scalar resources, it can export its gateway capabilities as a service for its peers to route packets through it. In addition, it is also possible to specify additional constraints to characterize resource functionality.

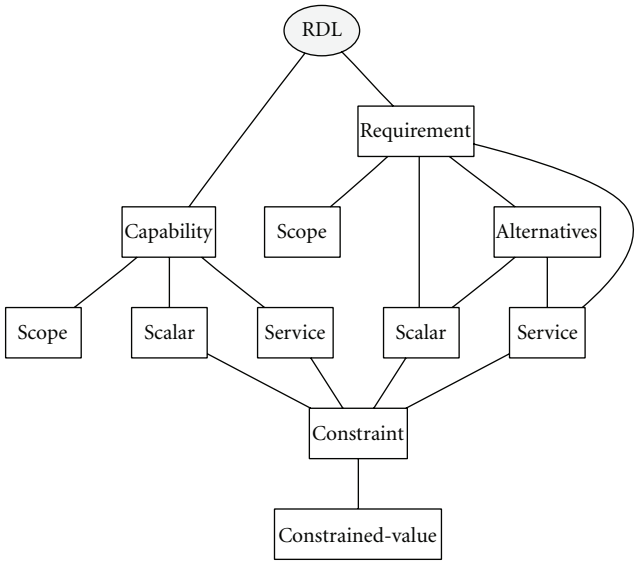


Figure 3: RDL framework diagram depicting main concepts and their relations.

Resources, whether they are scalar or service ones, belong to nodes in the network. Scalar resources that are, most likely, found in nodes in the network, describing what resources the node has. Service resources can be available either in a single node, a group of nodes, or an entire network. The RDL defines two scopes in order to define where the resources exist, or where they can be used. These scopes are the node-scope, where the scope indicates a single node, or a network-scope, where it indicates a group of nodes or a network.

4. RDL Format Specification

Given the heterogeneous nature of modern networks, interoperability issues come into play. As such, the RDL must be built upon standard formats with clear and unambiguous specifications, assuring that different nodes can successfully manage their resources cooperatively.

For the RDL specification, two formats are proposed, an official Extensible Markup Language (XML) [19] format and an alternative, compact binary representation, using a Key-Length-Value (KLV) [20] format. Whereas the XML format is suited for most applications, enabling the use of a vast library of XML-based tools and APIs that have already been developed, the KLV representation can be used in networked embedded systems and wireless sensor networks where the limited network characteristics and processing power as well as the need for energy efficiency advise against the use of a more computational demanding XML format.

The format specification concretely specifies the element structure of the concepts defined in the framework. The RDL describes resources as either capabilities or requirements. These resources, in turn, can be specified as one of two types: scalars or services. Additionally, scopes are used to define the context within which these resources are defined, as well as to limit the reach of the queries that look them up (e.g., a resource may be described as belonging to "node 1," and one may search for all of the resources that belong to "node 1" to find it). Constraints, on the other hand, are used to further define any limitations that apply to these resources by describing intrinsic properties (e.g., the printer is only capable of printing in black and white).

Having the intention of not limiting the expressiveness of the language, we acknowledge the need for it to accept extensions that further augment its reach and adaptability, while at the same time maintaining its logical structure.

4.1. XML Format. Following the general RDL element structure model, the official format describes capabilities, requirements, resources, and their scopes and constraints using XML elements and attributes. An example of a sensor node RDL specified in this format is illustrated on Algorithm 1.

The first lines need to show the XML declaration (version and encoding) and the schema which it obeys to. At the root, a top level "`<rdl>`" element is defined. Within this element, multiple capabilities and requirements can be described through the elements "`<capability>`" and "`<requirement>`", respectively. The capability or requirement element, in turn, contains two main sections, the scope description and the resource description. The scope description, specified within a "`<scope>`" subelement, describes the context within which the capability/requirement is needed or can be used. This element has a mandatory "`type`" attribute, indicating whether the scope describes a "`node`" or a "`network`," as well as any number of optional descriptive subelements.

The resource description section, within a capability or requirement element, is specified through the subelements "`<scalar>`" or "`<service>`". Each of these subelements is uniquely identified by an "`id`" attribute and described by a mandatory "`type`" attribute, that specifies what kind of resource they represent. Optional descriptive subelements can be used to further better describe these resources, while the "`used-resource`" subelements may be used to identify any additional resources used.

Finally, constraints are specified through "`constraint`" and "`constrained-value`" subelements. The "`constrained-value`" subelements indicate a limitation on the values a particular variable may take, through the use of "key"-"`operator`"-"`value`" attributes, specifying an inequality that applies to a property (e.g., pages < 20), according to the semantics specified in the optional "`operator`" attribute (i.e., the default values "`equals`," "`different`," "`greater`," "`less`," "`greater-equal`," "`less-equal`," "`contains`," "`does-not-contain`," "`is-contained`," "`is-not-contained`"). The constraint subelement clusters and groups multiple "constrained-values", performing a composition. The constraint composition follows different semantics whether it is being used in a capability or a requirement which will be visible when matching resources.

Using a mechanism similar to the one used to describe composite constraints, simple composite requirements can also be built. Using the additional "`<alternatives>`"

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE s1 PUBLIC "<location>/rdl.dtd" "rdl.dtd">
<RDL>
  <capability>
    <scope type="node">
      <nodeAddress>123456789</nodeAddress>
    </scope>
    <scalar id="5" type="module-sensor-light">
      <constraint>
        <constrained-value key="sensor-max-value" operator="equals" value="100"/>
        <constrained-value key="sensor-min-value" operator="equals" value="0"/>
      </constraint>
    </scalar>
    <scalar id="6" type="module-sensor-temperature">
      <constrained-value key="sensor-precision-ppm" operator="equals" value="1"/>
    </scalar>
  </capability>
</RDL>
```

Algorithm 1: RDL KLV description of a sensor node with two sensor modules (light and temperature).

subelement, further "<scalar>" or "<service>" subelements can be specified which consist of alternative requirements to the ones stated outside the "<alternatives>" element.

4.2. KLV Format. The XML-based format proposed gives a complete and human readable form of the description of resources. However, in some scenarios, such as WSNs, there is the need for a more compact representation, due to node limitations with respect to processing power and storage. As such, a KLV-based format is proposed. KLVs are tuples that associate a variable length value with a fixed size integer key identifier, while also explicitly specifying the size of the value. The new format can be built using a structure similar to the one used in the XML format, as illustrated through the example on Algorithm 2. The example is presented in a more understandable human-readable format, since the KLV format represents information in a binary form, not directly human readable, but optimized for machine interpretation.

The KLV format allows the creation of a hierarchal system by recursively encapsulating entire children KLV tuples within their parents value. The main advantage of this system is the fact that if a node is not familiar with a given key, while it will not be able to use its value, it will still be capable of skipping over it and using the remaining, otherwise usable, KLV tuples. Additionally, for a small set of mandatory key identifiers with fixed size values, the format can be further compacted by omitting the length value, as the nodes preexisting knowledge can be used to calculate it.

4.3. Performance Metrics. In this section, we compare XML and KLV descriptions in terms of memory to justify the existence of the two available formats. The XML format is human-readable and easily understandable but comes at a higher memory cost in comparison to the more lightweight and simple KLV format.

Comparing the RDL specification examples presented in Algorithms 1 and 2, the memory ratio between the XML and the KLV version is tenfold. For the same description, the XML version used around 546 bytes of memory whilst the KLV version used only around 55 bytes. This difference can be understood by the fact that the binary representation of the KLV is more compact, which although not directly human-readable has a greater memory advantage for machine processing and RDL message interchange.

Further experimentation on the memory difference between the two formats is depicted in Figure 4. The figure, presents four example cases (A, B, C, and D) which specify different RDL descriptions that increase in complexity. Case A describes a single capability with scope and a single scalar resource. Case B describes a single capability with two scalar resources with constraints associated. Case C describes three capabilities and three requirements, each with one scalar resource and constraints. Case D describes four capabilities and four requirements, with scalar resources, constraints, and requirement alternatives.

In Figure 4, the differences for the individual cases can be seen, as well as the general memory trend when increasing the complexity of RDL specifications. It is interesting to notice that, as the RDL specification increases in complexity, the XML format has a more marked slope increase in comparison with the KLV format, making XML growth surpass KLV's as more and more RDL resource descriptors are defined.

5. Resource Matching Algorithm

The RDL format is intended to ease the verification of the fulfillment of requirements with the available capabilities.

```
<capability> = {
  <scope> = {
    <type> = <node>;
    <node-address> = 123456789;
  }
  <scalar> = {
    <type> = <modulus-module-sensor-light>;
    <constraint> = {
      <constrained-value> = {
        <key> = <modulus-sensor-max-value>;
        <operator> = <equals>;
        <value> = '0 0 0 100';
      }
      <constrained-value> = {
        <key> = <modulus-sensor-min-value>;
        <operator> = <equals>;
        <value> = '0 0 0 0';
      }
    }
  }
  <scalar> = {
    <type> = <modulus-module-sensor-temperature>;
    <constrained-value> = {
      <key> = <modulus-sensor-precision-ppm>;
      <operator> = <equals>;
      <value> = '0 0 0 1';
    }
  }
}
```

Algorithm 2: RDL KLV description of a sensor node with two sensor modules (light and temperature).

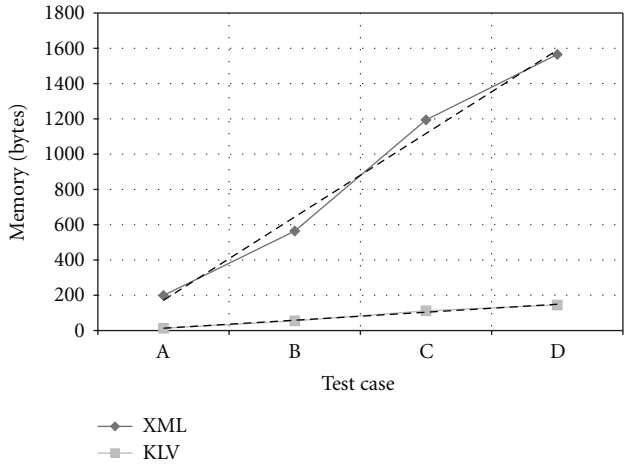


Figure 4: Memory comparison for the XML and KLV formats as RDL descriptions increase in complexity.

To verify such fulfillment, a resource matching algorithm was developed which matches the known requirements with capabilities, or vice versa. Pseudocode for the algorithm can be seen in Algorithm 3.

When matching resources, a match occurs when at least one requirement has one capability that satisfies it. In order to check if a requirement is provided by a capability (or if a capability satisfies a requirement), both the scope and the scalar/service resources need to be evaluated. Scopes must be compatible, and for all scalar and service resources, constraints must be checked. Considering the logic behind the resource matching and defining n as the number of requirements to match against m capabilities, complexity of the algorithm can be generally defined as $O(n \times m)$.

The algorithm executes and is implemented as follows. Within capabilities, all constraints and constrained-values are logically OR-ed at the resource description root-level and logically AND-ed within the constraint subelements, in other words, if any of the root conditions is met, the capability is considered to satisfy the requirement. On the other hand, all of the constrained values within a constraint must be met to satisfy the requirement. Requirements, in contrast, follow the opposite logic: constraints and constrained-values are AND-ed at the root level and OR-ed within subconstraints, that is, all root-level constraints must be met and subconstraints represent alternative choices.

Whereas capabilities can only be logically OR-ed at the root-level, that is, a requirement need only matches one capability to be satisfied, a requirement can describe a richer relationship. The idea is that, like in capabilities, any of the root-level requirements can be met to achieve a match between two RDL descriptors but, within each requirement

```
Input : Resource Descriptor (RDL-A), Resource Descriptor (RDL-B)
Output: True (match) or False (mismatch)
foreach Requirement in RequirementsList of RDL-1 Ado
   foreach Capability in CapabilitiesList of RDL-Bdo
      if Requirement.Scope matches Capability.Scopethen
         foreach ResourceR in ResourcesList of Requirementdo
            foreach ResourceC in ResourcesList of Capabilitydo
               if (ResourceR.Type equals ResourceC.Type) and (ResourceR.Constraints
               matches ResourceC.Constraints)then
                  return true;
               end
            end
         end
      end
      return false;
   end
end
```

Algorithm 3: High-level pseudocode for the match between resources.

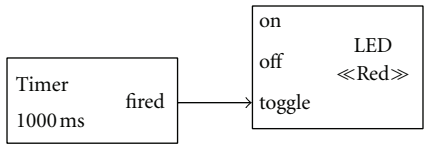


Figure 5: LED blinking modulus application diagram.

element, all scalar and service resources must be met (i.e., they are logically AND-ed). This description can be further enhanced through alternatives, containing additional scalar or service resources, allowing one to have a third level of depth, now following an OR-ed logic.

6. Application Scenario: Modulus-Modular Middleware Solution

In this section, to better exemplify how the RDL can be used in a real context and to show that the definition of the resources in a interchangeable format assumes a very important role, we present a modular middleware solution that we have been developing, called Modulus, that enables rapid application adaptation to available resources.

Modulus is a modular middleware solution for M2M systems and wireless sensor networks. The concept is loosely borrowed from the way systems are designed in electrical engineering. Electrical engineers do not go about reinventing the transistor every time they project a new amplifier. An assortment of prefabricated reusable components can be looked up in a catalogue and complex systems are built up by interconnecting these components in the right way. Modulus brings this sort of approach to the design and development of embedded software, focusing on rapid application development through intensive component level reusage. As before, large and complex applications are built from smaller building blocks, and the concept of application is redefined as being an interconnected set of simple reusable components.

More concretely, in Modulus, modules are instantiated, each instance sharing the same implementation, but keeping its own state and configuration. These modules communicate with each other in an abstract way using interfaces, allowing one component instance that uses a given interface to be wired to another that provides the same interface. With such a system in place, and with applications designed according to this paradigm, the runtime system is built upon a message passing framework that pushes data from module to module, according to the specified wiring.

One of the key advantages of Modulus lies in how it leverages the RDL to describe its applications. In a sense, Modulus applications do not connect predefined modules to each other, but rather connect their RDL descriptions. This small difference is a key factor for runtime adaptability and makes even more sense from the developer's point of view. Indeed, in an abstract sense, a module's implementation is not as important as the functionality that it exports. As long as the interface by which such functionality is provided is known to all parties involved, each module can be seen as a little black box. Furthermore, through the use of RDL constraints, a higher level of matching can be performed, by which wiring is not only constrained to interfaces of the same type, but also to other application level measures of compatibility or quality.

Figure 5 illustrates this concept in action. The LED blinking application is one of the simplest applications that can be implemented on an embedded system and is frequently used to quickly test tool-chains and build automation systems, much like Hello World applications are used on traditional computing systems. The application is simply built by connecting two components: a Timer and an LED. The LED component uses three interfaces of type Event; one to turn it on, one to turn it off and one to toggle

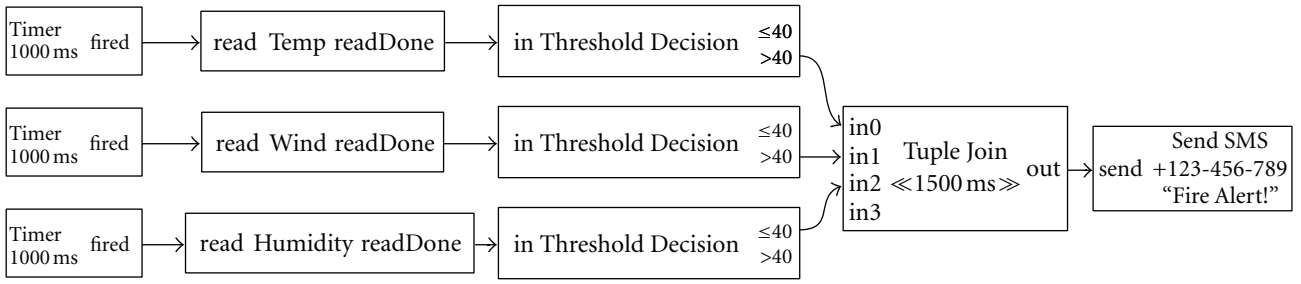


Figure 6: Fire alert modulus application diagram.

its current state. The Event interface is a generic interface used to signal that something happened. The Timer module provides this interface to periodically generate this signal at the configured rate and, by wiring these two components together, the periodic LED toggling is achieved, as expected.

Following this basic concept, the Modulus application is simply encoded as the component configurations (only the timer periodicity, in this case), and the interface wiring based on RDL descriptors. The entire application is described in KLV as shown in Algorithm 4. Furthermore, the inherent flexibility of the RDL encodes both static constraints such as the module specifications shown here (Timer and LED), as well as any additional application specific constraints, such as the LED color.

The power of this approach is even more clear in networked distributed systems. By transparently allowing wiring to traverse the network, Modulus allows functionality to be exported between nodes, enabling the development of complex distributed systems, using the same building blocks as before. Under these circumstances, application developers might not even have direct access to knowledge of which resources will be available in the deployed system. So long as an adequate description of the resource is given, the application will be adapted accordingly, at run-time.

A good example of this is the hypothetical application illustrated in Figure 6. In this example, it is determined that there is a higher risk of forest fires when all of the following conditions are met: the temperature is higher than 40°C, the wind speed is faster than 40 km/h, and the relative humidity is less than 40%. As such, park rangers want an application to automatically send an SMS to a predetermined phone number when such conditions are met.

To this end, each of the three sensors must be sampled every second and their values compared against the given threshold. If all of the preconditions are met within a certain time window (1500 ms in this case), then the SMS module is signaled and the message sent. What is interesting, however, is that such an application would work equally well whether all three sensors and the GSM module were all on the same system or in four networked nodes. Modulus abstracts away such details and transparently marshals the data to where it is needed.

From the user's perspective, the RDL can be used to provide a powerful insight into the networks built-in capabilities, allowing them to create better informed policies. From within Modulus, the additional information provided by the RDL can be used to enable automated application optimizations, or even to decide when it would be wise to migrate a particular application or functionality to another, more capable, node. Under these circumstances, the RDL can be a key element in enabling opportunistic resource sharing and creating more powerful, context-aware, embedded applications.

7. Conclusions

This paper describes the initial steps in the specification of the RDL. The main goal for the RDL is to define a reusable resource description specification for different algorithms, which can only be reached if the resources are described in a standard form. By separating the actual resources from the functionalities behind the algorithms that use them, new systems can take advantage of a pool of available resources. The RDL language is explicitly specified in the official XML format or the alternative KLV format. The KLV is a more memory compact representation as evaluation shows.

The RDL defines two layers of resources, that is, scalar resources and services. Scalar resources describe the objects and characterize them both qualitatively and quantitatively. Scalar resources are, for example, node hardware capabilities or packet error rate of a wireless link, and so forth. Services, on the other hand, describe packages of scalar resources that export some sort of functionality. For example, whereas a network router can describe its queues and protocol capabilities as scalar resources, it can export its gateway capabilities as a service, allowing its peers to route packets through it. Resources are bound to a scope, and there have been two scopes defined, that is, node and network scope. Resource matching is accomplished through an algorithm that analyzes resource descriptors and checks if their requirements and capabilities match.

In order to further develop the RDL framework, we also proposed a Java translation, which respects the language concepts and provides a core implementation for language usage. A Java applet tool was also developed to provide a means of interacting with the RDL language, allowing easy graphical creation of RDL descriptors as well as their generation to XML or KLV code.

To further illustrate the power of the RDL, we developed Modulus, a modular middleware solution for networked

```
<module-instance> = {
  <rdl-description> = {
    <requirement> = {
      <scalar> = {
        <type> = <modulus-module-timer-periodic-milli>;
      }
    }
  }
  <configuration> = 1000;
}
<module-instance> = {
  <rdl-description> = {
    <requirement> = {
      <scalar> = {
        <type> = <modulus-module-actuator-led>;
          <constrained-value> = {
            <key> = <modulus-module-actuator-led-color>;
            <operator> = <equals>;
            <value> = <red>;
          }
        }
      }
    }
  }
}
<wiring-data> = '0 0 0 0 1 2';
```

Algorithm 4: LED Blinking Modulus Application KLV Encoding.

embedded sensing systems. Modulus uses the RDL to describe software components and resources available on sensor nodes, allowing the development of applications to be decoupled from the actual resources they will use. This allows applications to automatically adapt, at runtime, to available resources, thus increasing the systems overall flexibility.

In short, this initial specification represents an initial effort to aid resource-based application development. Further improvements are planned for continuous development of the language. Main future work additions include

(i) further development of the specification to include more necessary properties and values that need to be defined to better describe resource nodes;

(ii) implementing additional functionality to the RDL Applet tool to fully respect the specification;

(iii) proposing the RDL as an IETF RFC.

Acknowledgments

Initial work on the description and specification of the RDL was supported by the Instituto de Telecomunicações, Lisbon, in the context of the EU-FP7 NEWCOM++ NOE [21] by M. Kuipers, L. D. Pedrosa, and R. M. Rocha. Further development took place at GEMS as part of L. D. Pedrosa and A. C. Santos' doctoral research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Instituto de Telecomunicações. This work was partially supported by national funds through *Fundação para a Ciência e a Tecnologia* (FCT), under project PEst-OE/EEI/LA0021/2011 and Doctoral Grant no. SFRH/BD/47409/2008.

References

[1] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. D. Johnson, "M2M: from mobile to embedded internet," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 36–43, 2011.

[2] J. Zander and O. Queseth, *Radio Resource Management for Wireless Networks*, Artech House, Norwood, Mass, USA, 2001.

[3] D. Wobschall, "IEEE 1451—a universal transducer protocol standard," in *Proceedings of the 42nd Annual IEEE AUTOTESTCON Conference*, pp. 359–363, September 2007.

[4] M. P. Bodlaender, "UPnP 1.1—designing for performance & compatibility," *IEEE Transactions on Consumer Electronics*, vol. 51, no. 1, pp. 69–75, 2005.

[5] H. Liu, B. Krishnamachari, and Q. Zhao, "Negotiating multichannel sensing and access in cognitive radio wireless networks," in *Proceedings of the 6th IEEE Annual Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks Workshops (SECON '09)*, pp. 1–6, June 2009.

[6] A. Acharya, M. Ranganathan, and J. H. Saltz, "Sumatra: a language for resource-aware mobile programs," in *Proceedings of the 2nd International Workshop on Mobile Object Systems—Towards the Programmable Internet (MOS '96)*, pp. 111–130, Springer, London, UK, 1997.

[7] N. Damianou, N. Dulay, E. Lupu, and M. Sloman, "The ponder policy specification language," in *Proceedings of the 9th*

IEEE Workshop on Policies for Distributed Systems and Networks (POLICY '01), pp. 18–38, London, UK, 2001.

[8] G. N. Stone, B. Lundy, and G. G. Xie, "Network policy languages: a survey and a new approach," *IEEE Network*, vol. 15, no. 1, pp. 10–21, 2001.

[9] M. Compton, C. Henson, H. Neuhaus, L. Lefort, and A. Sheth, "A survey of the semantic specification of sensors," in *Proceedings of the 2nd International Workshop on Semantic Sensor Networks at the 8th International Semantic Web Conference*, vol. 522, pp. 17–32, October 2009.

[10] A. Bröring, K. Janowicz, C. Stasch, and W. Kuhn, "Semantic challenges for sensor plug and play," in *Proceedings of the 9th International Symposium on Web and Wireless Geographical Information Systems (W2GIS '09)*, pp. 72–86, Springer, Berlin, 2009.

[11] M. Botts and A. Robin, "OpenGIS R Sensor Model Language (SensorML) Implementation Specification," OpenGIS Implementation Specification OGC 07-000, Open Geospatial Consortium Inc., Version: 1.0.0, July 2007.

[12] G. Aloisio, D. Conte, C. Elefante, G. P. Marra, G. Mastrantonio, and G. Quarta, "Globus monitoring and discovery service and sensorML for grid sensor networks," in *Proceedings of the 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE '06)*, pp. 201–206, June 2006.

[13] D. Wobschall, "Networked sensor monitoring using the universal IEEE 1451 standard," *IEEE Instrumentation and Measurement Magazine*, vol. 11, no. 2, pp. 18–22, 2008.

[14] K. C. Lee, M. H. Kim, S. Lee, and H. H. Lee, "IEEE 1451 based smart module for in-vehicle networking systems of intelligent vehicles," in *Proceedings of the The 29th Annual Conference of the IEEE Industrial Electronics Society*, pp. 1796–1801, November 2003.

[15] J. Veizades, E. Guttman, C. Perkins, and S. Kaplan, "Service Location Protocol," RFC 2165 (Proposed Standard). Updated by RFCs 2608, 2609, June 1997.

[16] E. Guttman, C. Perkins, J. Veizades, and M. Day, "Service Location Protocol, Version 2," RFC 2608 (Proposed Standard). Updated by RFC 3224, June 1999.

[17] E. Guttman, "Service location protocol: automatic discovery of IP network services," *IEEE Internet Computing*, vol. 3, no. 4, pp. 71–80, 1999.

[18] C. E. Perkins, "Service location protocol for mobile users," in *Proceedings of the 9th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '98)*, pp. 141–146, September 1998.

[19] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, and F. Yergeau, "Extensible Markup Language (XML) 1.0 (5th Edition)," W3C Recommendation, November 2008.

[20] "The KLV Standard—Data Encoding Protocol Using Key-Length Value," 2007.

[21] L. Galluccio, A. Leonardi, G. Morabito et al., "EU-FP7 NEWCOM++– WPR11: Opportunistic Networks—Intermediate Report on Resource Management Issues and Routing/Forwarding Schemes for Opportunistic Networks," Tech. Rep. Bilkent/KHAS, CNIT-CT, CNITBO, CNIT-PD, CNIT-TO, CNRS-LAAS, ISMB, IST-TUL, KAU, NKUA/IASA, PUT, UPC, RWTH, 2009.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 762953, 8 pages
doi:10.1155/2012/762953

Research Article

Service-Oriented Radio Architecture: A Novel M2M Network Architecture for Cognitive Radio Systems

Xu Dong,[1] Shengqun Wei,[2] Ying Li,[2] Lifeng Wang,[2] and Lin Bai[3]

[1] *Institute of Communications Engineering, PLA University of Science and Technology, Nanjing 210007, China*
[2] *Institute of China Electronic System Engineering Corporation, Beijing 100141, China*
[3] *School of Electronic and Information Engineering, Beihang University, Beijing 100191, China*

Correspondence should be addressed to Lin Bai, l.bai@buaa.edu.cn

Received 20 April 2012; Revised 4 July 2012; Accepted 4 July 2012

Academic Editor: Jianhua He

In future cognitive radio networks, a number of spectrum sensors can be distributedly deployed to monitor the surrounding wireless environment, where the machine-to-machine (M2M) technology is considered to provide the interactions among sensors, cognitive engines, and other system modules. Thus, a flexible M2M network architecture is desired to develop cognitive radio networks. As a distributed system framework, service-oriented architecture (SOA) has been well studied to provide the loose coupling, reusability, and scalability, where various system function modules are encapsulated into web services to build a virtual system. In this paper, based on SOA, we propose a flexible M2M architecture, namely, the service-oriented radio architecture (SORA), to develop the cognitive radio systems. It shows that our proposed architecture provides coordinated implementation of cognitive radio systems based on different development platforms. Furthermore, an SORA-based cognitive radio testbed is implemented, where the standard web service technologies and open source software tools are used to support the characteristics of SORA.

1. Introduction

In cognitive radio (CR) systems, the sensors are employed to monitor the surrounding spectrum environment, where the cognitive engines (CE) and software defined radio (SDR) techniques are needed to generate and carry out different communication strategies according to the actual environment. As an efficient method to develop CR systems, the machine-to-machine (M2 M) networks among sensors, CEs, databases, and SDR systems are drawing wide attentions. However, large-scale CR systems have posed a big challenge to develop the M2 M networks with respect to the flexibility and compatibility. For example, since different CR systems, testbeds, and prototypes could be implemented on different software and hardware development platforms, the heterogeneity of those platforms would lead to high incompatibility for cooperative development [1]. Then, the interfaces of a certain machine may not be reused by others due to the unique design of data interfaces. Moreover, considering a tightly-coupled architecture of SDR systems, for example, software communication architecture (SCA) [2], the interaction and integration among SDR system and other machines become even more difficult. Therefore, to deal with the above issues in future CR networks design, a flexible and compatible M2 M architecture needs to be considered.

Recently, building a large-scale and multinode testbed over wide areas becomes an urgent need for the research on future CR networks and has already been concerned by the European cooperation in science and technology (COST) IC0902 project [3]. The main objective of IC0902 project is to develop cooperative research activities in the field of CR networks. For the comprehensive cooperative research, different mechanisms, for instance, a larger-scale CR network testbed integrated from different CR testbeds or prototypes, are established, where the challenges to develop the M2 M networks mentioned in above need to be carefully considered. To this end, different complex inter-operations and interfaces are implemented among various platforms, especially when components of the system are developed

with different programming languages and hardware platforms.

Since heterogeneous system integration and cooperative development have posed a big challenge over the past decade, as a solution, the service-oriented architecture (SOA) [4, 5] has been used to integrate and develop different interenterprise systems. An SOA-based system consists of the loose-coupled and platform-independent services with well-defined interoperable interfaces. Through web-based publish/subscribe mechanisms, the services can be flexibly orchestrated to develop appropriate applications. Thus, the SOA allows cooperative design and development, where reusability and scalability are considered.

In this respect, SOA has already been used to design the M2 M platforms [6]. In addition, Virginia Tech has built a cognitive radio open source system (CROSS) to integrate cognitive engines based on different intelligent algorithms [7], where CROSS utilizes socket connections and SOA approaches for inter-component communications to support portability and interoperability between components developed in different programming languages.

Using the principles of SOA, in this paper, we propose an approach to collaboratively design and develop the CR system, namely, the service-oriented radio architecture (SORA). To be different from the traditional-distributed radio architectures (e.g., SCA), SORA provides a distributed and modular radio communication system framework, which assembles a radio system over networks or even Internet (i.e., based on web service technologies). Moreover, as building a radio system consists of some blocks, the services in SORA are loosecoupled. Comparing to traditional radio system architectures, it shows that our proposed SORA is able to support the underlying heterogeneity, remote system invocation and integration, and meanwhile provide higher scalability, reusability, and flexibility.

The rest of this paper is organized as follows. In Section 2, we introduce the SORA in details. The work on developing an SORA-based CR testbed is shown in Section 3. Finally, the paper is concluded in Section 4.

2. Service-Oriented Radio Architecture

The principles and methodologies of SOA have been widely employed in different wireless communication systems with various applications [8–11]. Specifically, in CROSS, the SOA has been used to design and construct the CR systems. Comparing to the CROSS, a number of web service standards are considered in SORA-based systems to provide higher flexibility and compatibility. The CROSS developers utilize the SOA approaches to integrate their cognitive engines developed in different languages with an usual SDR platform, but the program and protocols of SOA are designed and developed by themselves, where the open standards of SOA are not employed. Thus, when a new system component is added into CROSS, its dedicated SOA framework needs a corresponding change. In contrast, the SORA is based on a series of international standards. The new component can be encapsulated according to the standards so that it is compatible with the original system without any change. Additionally, in SORA-based system, the components with different functions can be flexibly chosen and integrated to complete the corresponding communication task. Then, the characteristics of SORA will be detailed as follows.

2.1. An Overview. Our proposed SORA is an open architecture, where the principles and methodologies of service orientation are considered to design and integrate radio communication systems. Using SORA, different radio function modules are encapsulated into the network-base components (viz., services). Since radio services are published and discovered over networks, users are able to invoke and orchestrate different required services to implement various radio communication applications.

Two key issues are considered to define the SORA. Firstly, SORA is a distributed and modular system architecture, where the radio function modules can be deployed as services over networks. Secondly, SORA provides an open architecture, where the radio services adopt open and standardized interface descriptions and interoperation contracts. Different characteristics of SORA-based radio systems are summarized as follow.

2.1.1. Loose-Coupling. As the building blocks of radio systems, radio services are abstracted from various radio functions of actual employed hardware platforms, software programs, and operating systems. Thus, different radio services are able to interoperate in a neutral format.

2.1.2. Reusability. Services with different or same radio functions can be used in multiple separate radio communication systems to reduce the cost of development.

2.1.3. Scalability. Radio systems are expediently upgraded as either the software programs of the original services are updated or new services are integrated. Then, the legacy systems can be well compatible with new systems.

2.1.4. Reconfigurability. Reconfiguration is implemented to re-organize the newly involved services or re-define operational parameters of the original services according to the surrounding radio frequency (RF) environment.

2.1.5. Cooperative Development. Academic and commercial communities are able to independently design and develop the radio function modules and systems, which are then provided in a service manner. Therefore, different services are cooperatively integrated to develop a large-scale testbed or demonstration system over networks.

All these characteristics of SORA lead to the significant improvement of flexibility and compatibility during the processes of radio system design, development, and implementation, especially for cooperative research, test, and exploitation.

2.2. Framework. The framework of our proposed SORA is summarized in Figure 1, where three terms are considered, namely, radio infrastructures, services, and applications. Specifically, radio infrastructures provide the basic function of radios (e.g., encoder, modulator-demodulator (MODEM), RF module, and antenna, etc.), services that as act the infrastructure functionalities (i.e., blocks of the radio system that can be assembled), and radio communication applications are implemented through design and integration of various services.

The three terms of the framework are specified as follows.

2.2.1. Radio Infrastructures. It is known that different wireless networks (e.g., 2G/3G, Wi-Fi, and WiMAX) exist in the current wireless environment, simultaneously. In such environment, the radio infrastructures can be characterized by the heterogeneity, as they could be based on different development platforms and may belong to different operators and networks. In Figure 1, three components are considered as the infrastructures, including antennas, RF modules, and multiple programmable processors. With the development of digital telecommunications, software-based methods are widely used at the back of antennas and RF modules, where different functions, including encoding, modulation, and equalization, can be implemented in application-specific integrated circuits (ASIC), field programmable gate arrays (FPGA), or digital signal processors (DSP). Therefore, in our proposed structure, these processors are included as the infrastructures of radio systems, while different radio infrastructures are encapsulated as services using the standard web service interfaces.

2.2.2. Services. Using functionality description, different radio infrastructures are encapsulated and abstracted from the corresponding software/hardware platforms. The abstracted infrastructure is regarded as a service to be published, discovered, and invoked over networks. In our proposed architecture, the services are classified into four different categories, namely, waveform services, information security services, sensing services, and cognition services. The waveform services include functions of encoding, modulation, and protocols for interaction; the sensing services provide the RF environment to radios; and the cognition services enable reasoning and learning from the experiences. According to the requirements of different users, these services are employed to form various applications, where the service broker is considered to perform the management for service registration, publication, and discovery.

2.2.3. Applications. Through the service integration, a virtual radio is constructed to execute the different applications of radio communication. The applications, for example, voices, short messages, and multimedia, and so forth, are then converted into digital signals using the source encoder. Note that the source encoder can be regarded as a service.

According to the framework of our proposed SORA systems, we can show that the service layer plays a crucial role, since different radio infrastructures are encapsulated and abstracted as services to support the heterogeneity of underlying development platforms. Comparing to the traditional radio architectures, the services are deployed distributedly over different networks with our proposed SORA configuration, while a virtual radio system is constructed through integration of appropriate services to meet the requirements of different users. From this perspective, SORA is regarded as an agile virtualization method to assemble radio function modules in a network-based concept.

2.3. Technical Standards. The implementation of SORA is based on a series of web service standards, including the web service description language (WSDL) for service description, the simple object access protocol (SOAP) for service interaction, and the universal description, discovery, and integration (UDDI) for service registration, publishing, discovery, and integration, respectively.

It is noteworthy that these web service standards are open access and can be employed on multiple underlying protocols. For example, multiple physical and link layer protocols for different physical environments are usually considered with different transport and network protocols, including HTTP, TCP, UDP, and IP, for different network environments. Using these web service technologies, SORA can provide various protocols, interfaces, and platforms to support the underlying complexity. Thus, the SORA-based radio systems can be flexibly deployed, developed, implemented, and managed.

The web service standards used in SORA are briefly introduced as follows.

2.3.1. WSDL. IT is a description language, which is based on extensible markup language (XML) to define the functionality of a web service. A service provider creates a WSDL document to describe the interface, message format, and network address. Since the information of a WSDL file is published in a service broker, the users are able to obtain the information of the available operations for the service. The XML Schema is used to define the data types in the WSDL document.

2.3.2. SOAP. IT is a protocol specification to exchange messages for the implementation of web services. SOAP is based on XML with the characteristics of extensibility, neutrality, and independence. Hence, SOAP provides a high-level interoperation capability over different hardware platforms, operating systems, and programming languages.

2.3.3. UDDI. IT is a set of XML-based protocols of the service broker (viz., the registry), which are considered for web service registration, publishing, and discovery. The UDDI registry is interrogated by the SOAP messages to provide an access to the WSDL documents that are required to interact with the web services listed in the registry.

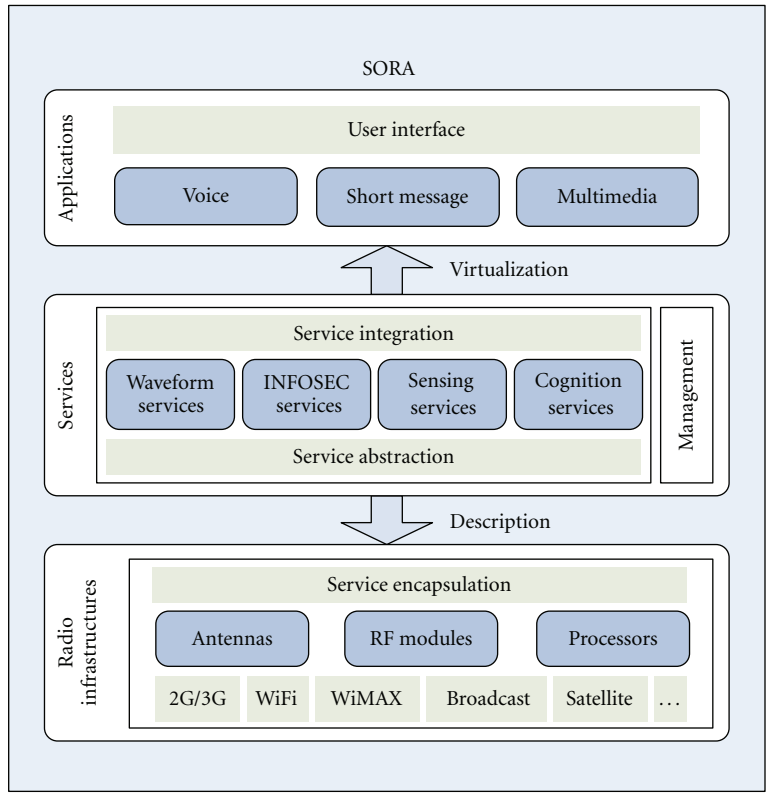


Figure 1: Framework of SORA.

According to the above introduction, different web service standards are defined and expressed in XML format. Therefore, the cross-platform sharing and reusing capabilities of XML are inherited to enjoy the flexibility and compatibility of SORA-based radio systems.

2.4. Use Case. A use case to illustrate the cooperation in developing an SORA-based CR access network is shown in Figure 2. In the proposed scenario, there are two operators and two research institutes, which are deployed over wide areas. Each one is considered with an individual network connected to an IP core network. It shows that although Operator A takes charge of a radio access network (RAN) with no cognitive capabilities, there have been some high-performance baseband processing and modem modules. On the contrary, Operator B possesses an expert system. Moreover, it is known that Institute A concentrates on the design and development of various sensors, while the artificial intelligence technologies and learning theory are developed at the Institute B.

Suppose that Operator B would like to build a CR access network for testing, however, due to the limited funds, the cooperation with Operator A is considered, where the research institutes A and B could work together to develop the SORA-based testbed. The process of the testbed development can be divided into four steps, including the service encapsulation, publication, discovery, and orchestration, which are defined as follows.

2.4.1. Service Encapsulation. All the participants, including Operator A, Operator B, Institute A, and Institute B, encapsulate their individual CR function modules as the invokable web services. Thus, these participants act as the service providers that generate the WSDL documents to define different service functionalities.

2.4.2. Service Publication. The WSDL documents are published in a UDDI service registry over the IP core network, where the registry is developed by a third party for public service publication and discovery.

2.4.3. Service Discovery. Operator B accesses to the service registry by utilizing the SOAP messages to find out the required services in the service list. Correspondingly, the locations, interfaces, and operations of different services are discovered using the WSDL documents.

2.4.4. Service Orchestration. According to the order of the signal processing in the CR system, different discovered services are orchestrated by the Operator B. Then, the services are invoked through the SOAP messages, where a virtual SORA-based CR access network is finally constructed.

The process mentioned above can be particularly used to develop the access point (AP) for the CR access networks. The AP is developed as a service client together with antenna and RF modules to provide the functionalities of service

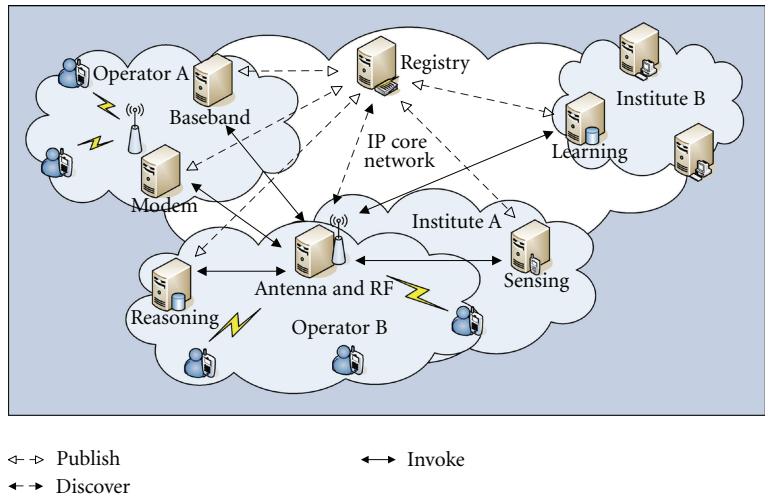


Figure 2: Use case of CR access network based on SORA.

discovery, orchestration, and invocation. Thus, as a cognitive user accesses to the AP, a communication mission would be triggered. The process of the mission is summarized in 5 steps as follows.

Step 1. As the service client, the AP invokes the baseband and modem services from Operator A for signal processing;

Step 2. At Institute A, the sensing service observes the real-time changes of the RF environment and then sent the results to the service client;

Step 3. The client invokes the local reasoning service to choose and decide the new transmission channels and operational parameters according to the observation results;

Step 4. The client invokes the signal processing services to be reconfigured according to the reasoning results;

Step 5. The whole process of the mission is memorized as one experience, while the experiences are periodically analyzed by the learning service at Institute B.

From this case, we can show that the SORA-based CR system provides a solution for cooperative development and implementation, where the reduced time and cost of the development are considered.

Additionally, our proposed SORA can deal with the more complicated situations, for example, when multiple users with the same or different radio interfaces communicate with one SORA-based CR AP at the same time. In such a situation, the AP should invoke and integrate some new services to satisfy the new communication requirement, such as admission control service and access protocol processing services. Meanwhile, AP needs to organize multigroup services to process different radio access interfaces and protocols. Each group of services can be orchestrated to process a certain kind of radio access interface and protocol, for example, 3G, 802.11, and so forth. Thus, when devices with different radio interfaces communicate with the AP, it can individually and parallel process these requirements of devices by invocating of different groups of services. Note that since one AP has limited resources (groups of services), it should first invoke the admission control service to check out the user's access model. If AP allows the user to access, the needed resources for this communication will be allocated to the user and the corresponding waveform services (including the protocol processing services) are invoked to process this communication.

3. Cognitive Radio Testbed Based on SORA

To verify the feasibility of SORA, a CR testbed is developed based on the approaches of SORA while a dynamic spectrum access (DSA) scenario is demonstrated on the testbed. As shown in Figure 3, three components are considered for the SORA-based CR testbed, including a service registry, a service client, and several services. All the components are connected to an IP-based local area network (LAN). Then, using the service encapsulation, publication, discovery, orchestration, and invocation, a virtual CR system can be constructed.

3.1. Development Platforms. Different software and hardware tools are used to develop the SORA-based CR testbed. The software development environment is considered on the basis of Linux operating system. The hardware development platforms include various types of hosts and the universal software radio peripherals (USRPs) for the RF frontends.

3.1.1. Service Registry. The service registry is constructed on a web server which is developed by Tomcat. The registry is implemented by jUDDI for service publication and discovery in a UDDI standard format. A database management system developed by MySQL is employed to store different types of data, for example, WSDL documents.

3.1.2. Service Client. The Python-based client program is used to orchestrate and invoke the services. A software tool of the client, namely SUDS, is carried out to perform the SOAP invocations and analyze the WSDL documents.

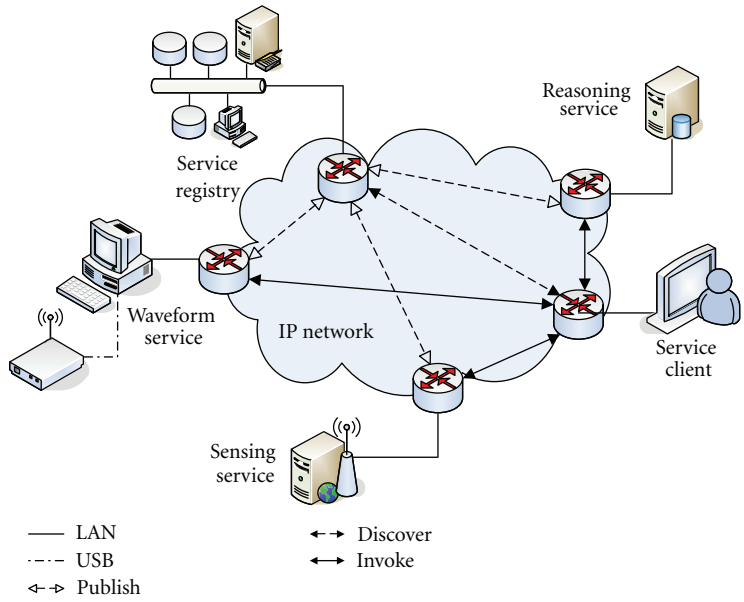


Figure 3: CR testbed based on SORA.

3.1.3. Services. The service programs and the registry are developed on the Tomcat web servers. Since the Tomcat is based on java, a java runtime environment is built by eclipse to provide the services. Axis2 is used to generate the WSDL documents and manage the SOAP messages for the services.

The hardware and software tools are summarized in Table 1, where the development tools of different service functionalities (including sensing, waveform, and reasoning) and the simulation of the RF environment are introduced.

3.2. Service Encapsulation. From Table 1, it is noticed that the sensing, waveform, and reasoning function modules are developed in different software tools and programming languages. Since the Tomcat is a java-based web server tool, the software modules are encapsulated into a java-oriented service.

The service encapsulation plays a key role in the process of development, as it is crucial to develop the function modules of the SORA-based CR testbed. Thus, two steps are considered for the encapsulation as follows. The SOAP interaction is first enabled based on java, and then the WSDL documents is generated for service description.

As a function module interacts with SOAP messages based on java, according to web service deployment standard, the original program module, which can be developed in C, C++, or Python, is additionally equipped with a java interface. This work is implemented on the platform of Eclipse. As shown in Figure 4, the java interface program module communicates with the original function module through a socket connection. As an SOAP request arrives at the java interface, the function module will be informed and return the data processing result to the java interface, then the result will be sent back to the requester through an SOAP message. Thus, the java interface enables the original function module to be encapsulated as a service and communicate in SOAP contracts.

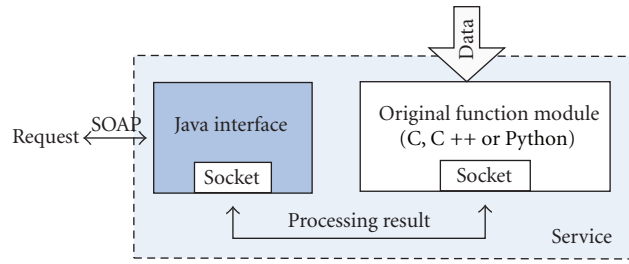


Figure 4: Service encapsulation.

To make the encapsulated service independent from the original programming language, the additional java interface should be defined in a neutral manner. This work can be achieved by using Axis2 that generates a WSDL document for the service description. When the java interface is developed in eclipse, a ".java" file is generated. The related WSDL document can then be obtained automatically by loading the java file to Axis2. Note that Axis2 can be used to analyze and generate the SOAP messages.

3.3. DSA Demonstration. The process of service publication, discovery, orchestration, and invocation is demonstrated using a DSA scenario. In this scenario, the vector signal generator (VSG) randomly generates different waveforms in different frequency bands to simulate the appearances of primary users. As for the secondary user, the SORA-based CR system dynamically chooses and accesses the spectrum holes to communicate while avoiding interference with the primary users.

Table 1: Development tools for The CR test base on SORA.

Element	Hardware	Function	Software Solution	Software Language	Developer
Service registry	Host	Web server	Tomcat	Java	Apache Software foundation
		UDDI	jUDDI	Java	Apache Software foundation
		Database	MySQL	C and C++	MySQL AB
Service client	Host	Orchestration	Center controller	Python	Self-development
		SOAP client	SUDS	Python	Jeff Ortel
Service	Build on each service host	Web server	Tomcat	Java	Apache Software foundation
		Java environment	Eclipse	Java	Eclipse foundation
		SOAP & WSDL enabler	Axis2	Java	Apache Software foundation
	USRP and host	Spectrum sensing	Energy detection	Python	Self-development
	USRP and host	Reconfigurable waveform	GNU radio	C++	Massachusetts Institute of Technology
	Host	Reasoning	CLIPS	C	NASA-Johnson Space Center
RF environment	Vector signal generator	Dynamic spectrum	Matlab	Matlab	Mathworks

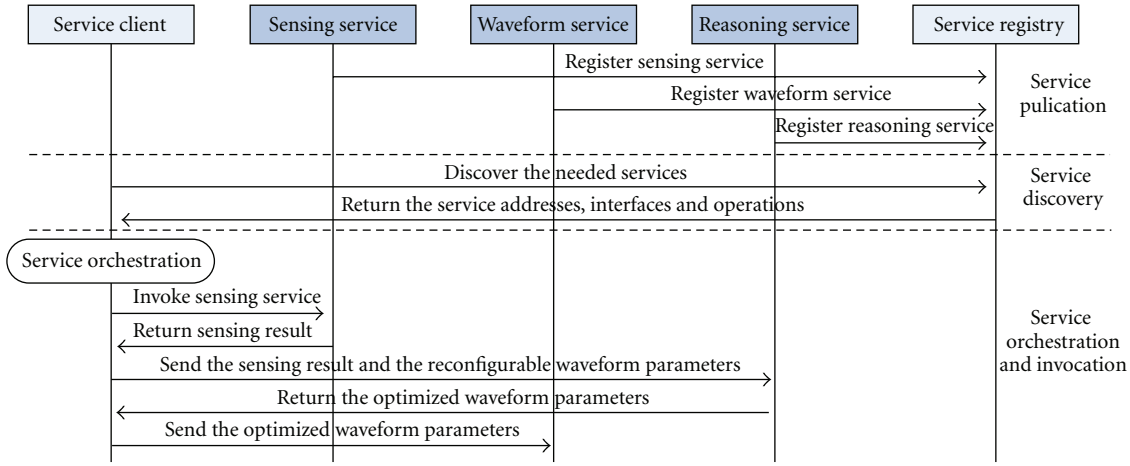


Figure 5: The information flow of service publication, discovery, orchestration, and invocation.

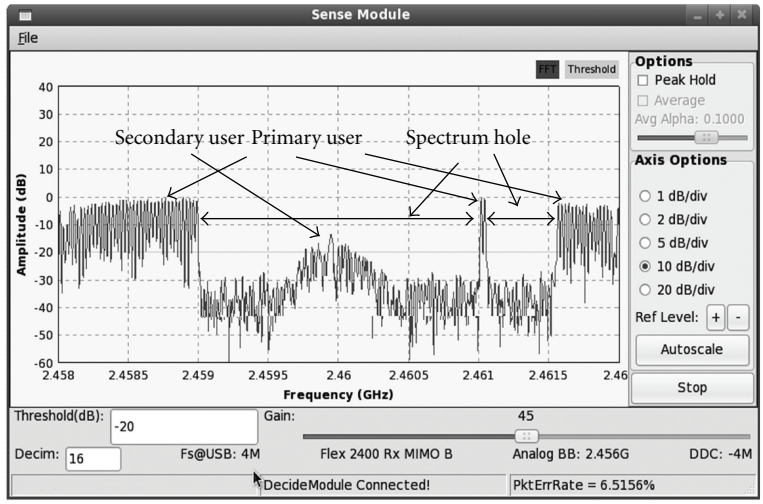


Figure 6: DSA demonstration.

The information flow of service publication, discovery, orchestration, and invocation is shown in Figure 5. The WSDL documents are published in the registry by following the process of service registration. On the other hand, the service client discovers the services and obtains their WSDL documents, and then orchestrates the services.

According to the orchestration, the sensing service is invoked by the client to start and keep sensing the spectrum, while the current sensing results are returned to the client. Then, the client sends the sensing results and reconfigurable waveform parameters to the reasoning service. After that, the reasoning service decides the optimized waveform parameters in terms of the relating radio knowledge and returns the results to the client. Finally, the client sends the optimized waveform parameters to the waveform service. Waveforms at the transmitter and the receiver are configured, respectively, and then the communication starts.

When a primary user appears at the frequency band that is being used by the secondary user, the process of service invocation according to the orchestration is implemented again. Thus, the DSA is implemented.

Figure 6 shows a screenshot of the DSA demonstration on the SORA-based CR testbed. It is illustrated that the testbed can dynamically choose and access the spectrum hole with the maximum bandwidth to avoid interference with the primary users and satisfy the spectrum requirement of the secondary user. Then, the spectrum usage efficiency can be increased. Although SORA may not improve the performance of CR systems, it certainly facilitates the flexibility of system development and implementation.

4. Conclusion

In this paper, we have proposed a novel M2 M architecture for future CR networks, namely, the SORA. We have first introduced the difficulties faced by the M2 M network in future large-scale CR networks and recently cooperative research activities, where the SOA provides a solution. Then, we have proposed our SORA and the components to construct CR networks. Comparing to the CROSS, it has been shown that our proposed SORA provides a higher flexibility and compatibility in cooperative development of CR network. Moreover, we have presented our SORA-based CR testbed to demonstrate the feasibility of the architecture.

In conclusion, the proposed SORA shows the characteristics of loose coupling, reusability, scalability, platform-neutrality, and cost reduction. Therefore, SORA is able to facilitate the M2 M communications in future CR networks.

Acknowledgments

This paper was supported by the National Basic Research Program of China (973 Program, no. 2009CB320403), National Natural Science Foundation of China (nos. 60832006, 60832008), and National Science and Technology Major Project (no. 2009 ZX03007-004).

References

[1] P. Pawelczak, K. Nolan, L. Doyle, S. Oh, and D. Cabric, "Cognitive radio: ten years of experimentation and development," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 90–100, 2011.

[2] Software Communications Architecture Specification (v2.2), http://www.eng.auburn.edu/users/hamilton/security/SDR/SRD_Release_V2.2.pdf.

[3] COST Action IC0902, http://newyork.ing.uniroma1.it/IC0902/.

[4] Web Service Architecture, http://www.w3.org/TR/ws-arch/.

[5] Service Oriented Architecture, http://en.wikipedia.org/wiki/Service-oriented_architecture.

[6] S. K. Zhang, J. W. Zhang, and W. Li, "Design of M2M platform based on J2EE and SOA," in *Proceedings of the 1st International Conference on E-Business and E-Government (ICEE '10)*, pp. 2029–2032, Guangzhou, China, May 2010.

[7] Virginia Tech Cognitive Radio Open Source System, http://ossie.wireless.vt.edu/trac/wiki/Cross.

[8] Q. Duan, "Applying the service-oriented architecture for network discovery and selection in the next generationwireless mobile networks," in *Proceedings of the 12th International Conference on Network-Based Information Systems (NBiS '09)*, pp. 380–385, Indianapolis, Ind, USA, August 2009.

[9] C. L. Wu, C. F. Liao, and L. C. Fu, "Service-oriented smart-home architecture based on OSGi and mobile-agent technology," *IEEE Transactions on Systems, Man and Cybernetics C*, vol. 37, no. 2, pp. 193–205, 2007.

[10] K. Y. Yee, A. W. Tiong, F. S. Tsai, and R. Kanagasabai, "OntoMobiLe: a generic ontology-centric service-oriented architecture for mobile learning," in *Proceedings of the 10th International Conference on Mobile Data Management (MDM '09)*, pp. 631–636, Taipei, Taiwan, May 2009.

[11] H. Neema, A. Kashyap, R. Kereskenyi, Y. Xue, and G. Karsai, "SOAMANET: a tool for evaluating service-oriented architectures on mobile ad-hoc networks," in *Proceedings of the IEEE/ACM Symposium on Distributed Simulation and Real-Time Applications*, pp. 179–188, Virginia, Va, USA, October 2010.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 840391, 9 pages
doi:10.1155/2012/840391

Research Article

A Dynamic Pricing Scheme for Congestion Game in Wireless Machine-to-Machine Networks

Zhifei Mao,[1] Guofang Nan,[1] and Minqiang Li[2]

[1] *Institute of System Engineering, Tianjin University, Tianjin 300072, China*
[2] *Department of Information Management and Management Science, Tianjin University, Tianjin 300072, China*

Correspondence should be addressed to Guofang Nan, guofangnan@gmail.com

Received 8 April 2012; Revised 9 July 2012; Accepted 9 July 2012

Academic Editor: Jianhua He

The problem of assigning a set of source nodes to a set of routes in wireless machine-to-machine (M2M) networks is addressed using a game theoretic approach. The objective is to minimize the maximum latency over all source nodes as far as possible while the game achieves a pure Nash Equilibrium (NE). To compute such an NE efficiently, we present a distributed dynamic pricing (DP) scheme, where each source node is assumed to pay for using any route so that the route has incentive to relay data for the source node. A loose upper bound is given for the convergence time of DP, and simulation results show that it performs much faster in practice. The price of anarchy in this game is also investigated by comparing DP with a cost-reducing path method; the results show that DP produces optimum assignment in more than 90% of the simulation runs.

1. Introduction

Features such as self-organization, ease of deployment, low cost, and infrastructurelessness bring to wireless M2M networks the advantages of robustness, easy maintenance, economy, and flexibility. Therefore, M2M networks are envisioned to be the key technology for next generation wireless networks [1]. An M2M network consists of a large number of self-organized and self-configured nodes, the behavior of which cannot be controlled easily without any available centralized scheduling command. Thus, the network can easily be congested when there are a great many nodes that want to transmit data via a common set of routes.

Extensive work has addressed the congestion problem by adjusting the source rate based on a feedback method [2–4], but prevention strategies for the congestion have been ignored. Nevertheless, by joining the congestion control in the routing layer, this problem can be solved via the coordination among nodes [5, 6]. However, in most wireless M2M network applications, each user is its own authority, and unconditional forwarding of packets to other users cannot be assumed directly. Therefore, a scheme that stimulates routes to relay the packets of other nodes is required. Motivated by the general network congestion game [7–9] and pricing scheme [10], our study jointly considers the problems of routing congestion and incentive provision.

In this paper, a distributed dynamic pricing (DP) scheme is developed for wireless M2M networks to minimize the maximum latency over all source nodes. First, routing congestion is formulated as a dynamic game in which pricing is employed in the definition of players' utility function. Each route is encouraged to charge source nodes for delivering their information, and the amount is decided by its latency. Second, dynamics is introduced in the game. The price of each route is changeable rather than fixed so that its load can be adjusted dynamically. We assume that the number of source nodes a route can take has an upper limit, and while the number of source nodes assigned to this route equals its upper limit, we say that this route is supply-demand balanced. Once all routes achieve supply-demand balance, the game terminates and converges to a pure Nash Equilibrium (NE). Finally, simulations are conducted to evaluate the performance of the proposed scheme.

The rest of this paper is organized as follows. Section 2 provides related work in recent years. Section 3 illustrates the system model and Section 4 formulates the network

congestion game. A DP scheme is proposed in Section 5. In Section 6, the simulation results are presented. And we conclude in Section 7.

2. Related Work

In [2], an algorithm jointly addressing congestion control and scheduling in wireless M2M networks is developed, the basis of which is rate allocation among flows in the network. By adjusting backoff min-slot and window size of each flow, the network can achieve high overall throughput and low per-flow delay. However, this algorithm cannot be applied for scenarios with dynamic routing, since route of each flow is assumed to be fixed. EWCCP, a protocol that can be added as a thin layer between IP and TCP, is proposed in [3]. It coordinates flows that compete for common channel according to explicit multibit congestion feedback from routers. Authors in [4] found that the main reason of congestion collapse in wireless M2M networks for streaming services is the severe contention among individual nodes at MAC layer, and thus presented a TCP-friendly congestion control scheme where a contention state estimator is defined as the difference between the number of arriving packets and that of leaving packets during each control interval. Similar to the aforementioned literature, congestion control schemes proposed in [11–13] are all feedback-based. Although these existing feedback-based approaches have achieved some success in congestion control for M2M networks, most of them ignored prevention/avoidance strategies for congestion control.

Congestion avoidance can be achieved by jointly considering congestion and routing. The work in [11] provides a multipath routing algorithm, I2MR, which discovers zone disjoint paths using the concept of path correlation so as to avoid interference in multiple nodes and increase system throughput. Cross-layer design of joint congestion control, routing, and scheduling was presented in [12], which enable each source node to adjust its routes for data transmission in each period according to a local congestion price from its neighbors. Other works such as [14–17] have also jointly considered the problems of congestion control and routing in multihop wireless networks or wireless sensor networks. The major distinction between our work and aforementioned work lies in the formulation of the congestion routing. Specifically, we propose a congestion game to model the route-selection behavior of each node who wants to transmit its data via a route with the lowest price. Moreover, we design a distributed algorithm with low complexity by making use of pricing to implement the congestion game framework.

3. Network Model

Consider a wireless M2M network with a set N of n parallel routes from a set M of m source nodes to a destination node, $M = \{1, 2, \ldots, m\}$ and $N = \{1, 2, \ldots, n\}$. Figure 1 shows an illustration of the network. For each source node i, let the strategy set $S_i \subseteq N$ denote the set of routes to which source node i can possibly be assigned. For each route j, let $B_j \subseteq M$ denote the set of source nodes that can possibly be served by this route. Each source node i intends to send a particular amount of traffic along a fixed route to the destination; thus, their choices form an assignment $\vec{s} = (s_1, s_2, \ldots, s_m)$, where $s_i \in S_i$ is the chosen route of source i. Given an assignment, let q_j be the number of the source nodes choosing route j. Let h_j be the length of route j, which equals the total of nodes contained in this route.

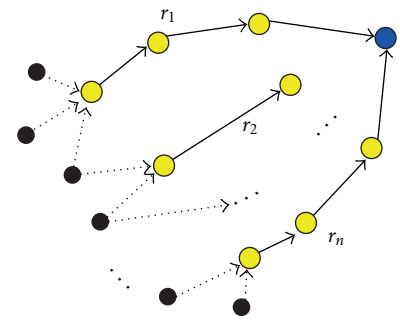


Figure 1: An example of wireless M2M network.

In the network, each route j has a latency d_j. Since the packet of each source node has been generated and processed before its transmission toward one route, the process time will not be included in routes' latency. Furthermore, comparing to queuing time and transmission time, the propagation time can be neglected in most M2M networks except some like satellite networks. Thus, the latency d_j of route j is dependent on the number of its length and current load, which can be expressed as $f(q_j) \oplus g(h_j)$. We assume that all source nodes transmit their packets via a common wireless channel, which means that routes will receive the packets one by one when multiple transmissions arrive simultaneously. Thus, both $f(q_j)$ and $g(h_j)$ can be set as a linear function. Based on the above assumptions, we compute the latency d_j as

$$d_j = q_j \times \frac{L}{R_s} + h_j \times \frac{L}{R_r}, \tag{1}$$

where L is the size of packet generated by source nodes, and R_s is the rate of transmitting one packet from a source node to a route while R_r is the rate of transmitting the same packet between two nodes within the same route. Moreover, we assume that R_s equals R_r. Then, d_j can be rewritten as

$$d_j = \left(q_j + h_j\right) \times D, \tag{2}$$

where $D = L/R_s = L/R_r$.

Our study aims to minimize the maximum latency over all source nodes. Hence, the congestion control problem can be modeled as

$$\min \max_{j \in M} \left(q_j + h_j\right) D. \tag{3}$$

4. Network Congestion Game

The network congestion game deals with the problem of congestion which is fundamental in networks and distributed

systems [7]. Whenever a set of users intend to utilize a much smaller set of resources in a common period, one needs to schedule these users orderly so as to reduce the overall cost and exploit available resources efficiently.

In wireless M2M networks, users and resources are referred to as source nodes and routes, respectively. The network congestion game can be formally defined as a tuple $\Gamma = \langle N, (S_i)_{i\in M}, (u_i)_{i\in M}\rangle$, where M is the set of players, that is, the source nodes. For the remainder of this paper, the terms "source node" and "player" are used interchangeably. S_i is the strategy set of player i. Let $\vec{s} = (s_1, s_2, \ldots, s_m)$ be the strategy profile when each player i chooses s_i, which is equivalent to the assignment in Section 2. $u_i : \Pi = S_1 \times S_2 \times \cdots \times S_m \rightarrow R$ is the utility function of player i. Since each player intends to connect to a route with the lowest latency, the utility function of player i can be formulated as

$$u_i = \{-d_{s_i} \mid s_i \in S_i\}. \tag{4}$$

The problem of incentive provision is managed here by transforming the latency to price which is charged by a route. The source node(s) that chooses this route is the price taker(s). This transformation provides the following benefits: (1) payment can effectively motivate a route to relay the data of a source node, and (2) the route can exploit its price as a lever to adjust its load. The instrument of pricing adopted here is not for maximizing the routes' revenues, but for achieving a socially beneficial objective. The price of route j is set as a function of its latency and denoted as $p_j(d_j)$. The utility function of each player i can then be rewritten as

$$u_i = \{-p_{s_i}(d_{s_i}) \mid s_i \in S_i\}. \tag{5}$$

Suppose the game is played in a one-shot style, which indicates the game experiences only one iteration and then terminates. Since each player has to make a decision without the decisions of other players, the result may be extreme; that is, routes with low prices may share the entire load, whereas those with high prices may share no load at all. Furthermore, a single iteration does not allow routes to adjust their load. Therefore, we treat this game as a dynamic game where players interact with each other by playing the one-shot game numerous times until it converges to an NE. The next section proposes specific rules for the dynamic game.

5. Algorithm

In this section, a distributed DP scheme is presented to implement the dynamic game and enable it to converge in polynomial time. If a source node wants to deliver information via some routes, it first sends a request to all available routes in the strategy set. The game goes by the following specific rules.

(1) Each route j broadcasts an initial price to all source nodes that can communicate with j by wireless connection, and then waits for their responses. The initial price is set as $p_j = \alpha(|B_j| + h_j)$, where α is a price factor.

(2) After receiving the prices from all available routes, each source node i sends a response to the one(s) with minimum price.

(3) After obtaining the response from each source node in B_j, route j decides which node should be accepted. The result is sent to all source nodes in B_j. Whether to accept a source node or not depends on the supply-demand degree of this route, which is formally defined as follows.

Definition 1. The supply-demand degree of a route j is

$$\eta_j^t = \frac{\left|B_j^t\right| + h_j}{p_j^t/\alpha} = \frac{\alpha\left(\left|B_j^t\right| + h_j\right)}{p_j^t}. \tag{6}$$

The above equation means that at iteration t, the maximum number of connections (old and new) that each route can take in is $p_j^t/\alpha - h_j$. Thus, route j balances its supply-demand when the actual number of taken-in source nodes equals the maximum. If newly requested connections plus old connections exceed the maximum, j needs to randomly reject some of the newly requested connections to obtain its supply-demand balance. j then broadcasts the decision.

(4) After receiving the decisions of all routes in S_i, each source node i responds with an acknowledgement to the chosen route. Node i must choose one route randomly if multiple routes intend to accept its request.

(5) Thus far, the interaction between routes and source nodes in a single iteration is finished. Afterwards, each route j needs to determine the price of the next iteration. If the supply-demand is not balanced at the current iteration and there is no new connection accepted or no old connection broken, the price of the next iteration is set to $p_j \leftarrow p_j - \alpha$. Otherwise, the price remains unchanged. If the supply-demand is balanced at the current iteration, the price also remains unchanged and adds a "balanced" mark with the announced price.

The above five steps illustrate the process of each iteration in the dynamic game. The following step decides when the game would be over. First, a definition of the market level supply-demand degree is provided.

Definition 2. The market level supply-demand degree is the average of all the degrees of the routes' supply-demand, that is,

$$N_t = \frac{\sum_{j\in N} \eta_j^t}{n}. \tag{7}$$

Once $N_t = 1$, the whole market is supply-demand balanced, and the game is considered finished (Algorithms 1 and 2).

Two basic metrics for evaluating this algorithm are (1) its ability to converge to a pure NE and (2) the speed it needs to do so. Some significant features of DP are proven from these two aspects.

DP, part 1: executed at each route $j \in N$
Initialization
Set $B_j^0 = \phi$, $\eta_j^0 < 1$, $p_j^1 = \alpha(|B_j| + h_j)$.
repeat
if $\eta_j^{t-1} < 1$ (i.e., unbalanced)
Broadcast its price p_j^t to each source node $i \in B_j$.
else
Broadcast its price p_j^t with a "balanced" signal to each source node $i \in B_j$.
End if
Wait for responses from all source nodes in B_j.
if $|C_j^t| > p_j^t/\alpha - h_j - |B_j^{t-1}| + |L_j^t|$
Randomly choose a subset A_j^t of C_j^t, such that supply-demand is balanced.
Set $B_j^t = B_j^{t-1} \bigcup A_j^t \setminus L_j^t$.
else
Set $B_j^t = B_j^{t-1} \bigcup C_j^t \setminus L_j^t$.
End if
Broadcast its decision to each source node $i \in B_j$.
if $\eta_j^t < 1$ and $|C_j^t| + |L_j^t| == 0$
Set $p_j^{t+1} = p_j^t - \alpha$.
else
Set $p_j^{t+1} = p_j^t$.
$t = t + 1$.
End if
until $N_t = 1$ or $\eta_j^t == 1$ for each $j \in N$(i.e., balance its Supply-demand).
L_j^t: set of source nodes that leave route j during iteration t.
C_j^t: set of source nodes that turn to route j during iteration t.

Algorithm 1

DP, part 2: executed at each source node $i \in M$
repeat
Wait until receiving price from each route $j \in S_i$, then Send a request to route(s) k maximizing u_i^t, then wait for the decision of requested route(s).
if there is more than one route saying "yes" to its request
Choose one randomly to build a formal wireless connection.
End if
$t = t + 1$.
until each route balances its supply-demand.

Algorithm 2

Theorem 3. *The proposed algorithm converges to a pure NE.*

Proof. Let t^* be the number of iterations that the proposed algorithm needs to converge, and let $p_j^{t^*}$ be the final price of any route j.

First, by contradiction, we prove that for every source node $i \in B_j^{t^*}$, $p_j^{t^*} - \alpha \le p_k^{t^*} \le p_j^{t^*} + \alpha$ holds, where $k \in S_i$ is any route available to source node i. Assume $p_k^{t^*} \le p_j^{t^*} - 2\alpha$, and without losing generality let $p_k^{t^*} = p_j^{t^*} - 2\alpha$. Then

(1) suppose the price of route k at iteration $t = 1$ already equals $p_j^{t^*} - 2\alpha$, which means k has balanced its supply-demand from the beginning. Clealy, $p_j^{t^*} - 2\alpha < p_j^{t^*} \le p_j^1$. Thus, the chosen route of player i would be k rather than j, which contradicts the preassumption that i decides to connect to route j;

(2) suppose the price of route k at iteration $t = 1$ is not $p_j^{t^*} - 2\alpha$. There must be some iteration t' between $t = 1$ and $t = t^*$, at which the price of k equals $p_j^{t^*} - \alpha$ and at iteration $t' + 1$, the price equals $p_j^{t^*} - 2\alpha$. Since $p_j^{t^*} - \alpha < p_j^{t^*}$, the chosen route of player i at iteration t' would be k rather than j, which also contradicts the preassumption that i decides to connect to route j.

Considering (1) and (2) jointly, $p_k^{t^*} \le p_j^{t^*} - 2\alpha$ is invalid. Therefore, $p_k^{t^*} \ge p_j^{t^*} - \alpha$. Similarly, we can prove that $p_k^{t^*} \le p_j^{t^*} + \alpha$. Thus, $p_j^{t^*} - \alpha \le p_k^{t^*} \le p_j^{t^*} + \alpha$ holds for every player $i \in B_j^{t^*}$.

Second, we prove that each player i has no incentive to change its choice unilaterally. Without loss of generality, let $p_k^{t^*} = p_j^{t^*} - \alpha$. If i changes its choice from j to k, then k and j need to adjust their prices to $p_j^{t^*}$ and $p_j^{t^*} - \alpha$, respectively, to keep their supply-demand balanced. Note that, the price i has to take is still unchanged. Hence, its choice will not change. □

Lemma 4. *If there is a source node that changes its choice or a route that lowers its price at an iteration t, then we have $N_t > N_{t-1}$ and $\Delta = N_t - N_{t-1} \ge \alpha/n\Lambda(\Lambda - 1)$ where $\Lambda = \max_j(|B_j| + h_j)$.*

Proof. $\Lambda = \max_j(|B_j| + h_j)$ implies that $p_j^t \le \alpha\Lambda$ for each $j \in N$.

(1) Suppose player i changes its choice from k to j, then $p_k^t < p_j^t$. Considering that there may be no alternation in the sets of taken-in source nodes of j and k at iteration $t-1$, their prices may be reduced at iteration t accordingly. Thus, we have $p_j^t \le p_j^{t-1}$ and $p_k^t \le p_k^{t-1}$. Then, the following can be obtained:

$$\begin{aligned}
\Delta &= \frac{1}{n}\left[\left(\eta_j^t - \eta_j^{t-1}\right) + \left(\eta_k^t - \eta_k^{t-1}\right)\right] \\
&= \frac{1}{n}\left[\frac{\alpha\left(\left|B_j^t\right| + h_j\right)}{p_j^t} - \frac{\alpha\left(\left|B_j^{t-1}\right| + h_j\right)}{p_j^{t-1}} + \frac{\alpha\left(\left|B_k^t\right| + h_k\right)}{p_k^t} - \frac{\alpha\left(\left|B_k^{t-1}\right| + h_k\right)}{p_k^{t-1}}\right] \\
&= \frac{\alpha}{n}\left[\frac{\left|B_j^t\right| - 1 + h_j}{p_j^t} - \frac{\left|B_j^{t-1}\right| + h_j}{p_j^{t-1}} + \frac{\left|B_k^t\right| + 1 + h_k}{p_k^t} - \frac{\left|B_k^{t-1}\right| + h_k}{p_k^{t-1}}\right] \\
&\ge \frac{\alpha}{n}\left[\frac{\left|B_j^t\right| - 1 + h_j}{p_j^t} - \frac{\left|B_j^{t-1}\right| + h_j}{p_j^t} + \frac{\left|B_k^t\right| + 1 + h_k}{p_k^t} - \frac{\left|B_k^{t-1}\right| + h_k}{p_k^t}\right] \\
&= \frac{\alpha}{n}\left[\frac{1}{p_k^t} - \frac{1}{p_j^t}\right] \ge \frac{\alpha}{n}\left[\frac{1}{p_j^t - \alpha} - \frac{1}{p_j^t}\right] = \frac{\alpha^2}{np_j^t\left(p_j^t - \alpha\right)} \\
&\ge \frac{\alpha^2}{n\alpha\Lambda(\alpha\Lambda - \alpha)} = \frac{\alpha}{n\Lambda(\Lambda - 1)} > 0.
\end{aligned} \tag{8}$$

This means that $N_{t+1} > N_t$ holds if a source node changes its choice at iteration t.

(2) Suppose route j lowers its price at some iteration t, then $p_j^t = p_j^{t-1} - \alpha$. Furthermore, a lower price may attract extra source nodes to choose j, thus $|B_j^t| \ge |B_j^{t-1}|$. Hence, the following can be obtained:

$$\begin{aligned}
\Delta &= \frac{1}{n}\left(\eta_j^t - \eta_j^{t-1}\right) = \frac{1}{n}\left[\frac{\alpha\left(\left|B_j^t\right| + h_j\right)}{p_j^t} - \frac{\alpha\left(\left|B_j^{t-1}\right| + h_j\right)}{p_j^{t-1}}\right] \\
&\ge \frac{\alpha}{n}\left[\frac{\left|B_j^{t-1}\right| + h_j}{p_j^t} - \frac{\left|B_j^{t-1}\right| + h_j}{p_j^{t-1}}\right] \\
&= \frac{\alpha}{n}\left(\left|B_j^{t-1}\right| + h_j\right)\left[\frac{1}{p_j^{t-1} - \alpha} - \frac{1}{p_j^{t-1}}\right] \\
&\ge \frac{\alpha}{n}\left[\frac{1}{p_j^{t-1} - \alpha} - \frac{1}{p_j^{t-1}}\right] = \frac{\alpha^2}{np_j^{t-1}\left(p_j^{t-1} - \alpha\right)} \\
&\ge \frac{\alpha^2}{n\alpha\Lambda(\alpha\Lambda - \alpha)} = \frac{\alpha}{n\Lambda(\Lambda - 1)} > 0,
\end{aligned} \tag{9}$$

which implies that $N_{t+1} > N_t$ holds if a route changes its price at iteration t.

□

Lemma 5. *$N_t \ge N_{t-1}$ holds and N_t can remain unchanged during two successive iterations at most.*

Proof. Lemma 4 states that $N_{t+1} > N_t$ if there is a source node that changes its choice or if a route lowers its price at an iteration t. Consider the following case: there is neither any source node changes its choice nor any route lowers its price at an iteration t, and some routes are not supply-demand balanced. It can be computed that $\eta_j^t = \eta_j^{t-1}$ according to Definition 1. Thus we have $N_t - N_{t-1} = (1/n)\sum_{j\in N}(\eta_j^t - \eta_j^{t-1}) = 0$. By jointly considering Lemma 4, $N_t \ge N_{t-1}$ can be obtained. N_t can remain unchanged during two successive iterations. Since some routes are still in an unbalanced state and no source node deviates from its initial decision, these routes will reduce their prices at iteration $t+1$ (according to Rule 3). $N_{t+1} > N_t$ can then be obtained according to Lemma 4. This completes the proof. □

Theorem 6. *The number of iterations required to complete the proposed algorithm is $O(n\Lambda^2)$, where $\Lambda = \max_j(|B_j| + h_j)$.*

Proof. Evidently, $N_{t^*} = 1$ and $N_0 \ge 0$. Furthermore, Lemma 5 states that N_t can remain unchanged during two successive iterations at most. The following can then be obtained:

$$\begin{aligned}
t^* &\le 2 \cdot \frac{N_{t^*} - N_0}{\alpha/n\Lambda(\Lambda - 1)} \le 2 \cdot \frac{1}{\alpha/n\Lambda(\Lambda - 1)} \\
&= \frac{2n\Lambda(\Lambda - 1)}{\alpha} = \frac{2}{\alpha}n(\Lambda^2 - \Lambda).
\end{aligned} \tag{10}$$

Table 1: Simulation parameters.

Parameter	Value
Number of source nodes	$20, 40, \ldots, 1000$
Ratio of source nodes to routes	2, 3, 4
Degree of source nodes	$1, 2, \ldots, 10$

Equation (10) indicates that the proposed algorithm terminates in polynomial time, and the upper bound is $O(n\Lambda^2)$. □

6. Simulation Results

We conduct simulations to assess the performance of the proposed algorithm in terms of executed time and to ascertain the optimality of the NE and provide the results in this section. To obtain reliable results, several parameters are investigated, including the number of source nodes, the ratio of source nodes to routes denoted by r, and the degree of source nodes, which is defined as the number of routes that each source node can connect to. The values of these parameters are given in detail in Table 1. Since having assumed that the transmission rate and packet length are the same for all nodes, these two measurements are not included in the variable table. And note that our simulations are dedicated for quite generalized wireless M2M networks and thereby do not rely on specific protocols in MAC layer; it is unnecessary to specify which wireless channel is in use and how to set the window size. A large number of runs are conducted for each simulation and the results are averaged.

6.1. Executed Time. Section 4 shows that the proposed algorithm can be completed in polynomial time and gives an upper bound of the executed time. However, the real speed is essential in practice. Here, we evaluate the practical speed by comparing it with the theoretical time bound T given by (10). $\lambda = P/T$ is denoted as the ratio of practical time to the theoretical time bound (P-T ratio), where P is the practical time that the proposed algorithm takes.

First, we study the effect of the number of source nodes on the proposed algorithm's speed, and the other parameters are set as constants. Figure 2 shows how many iterations the proposed algorithm needs to be completed versus the different number of source nodes when the degree of source nodes is set to an average of 5, and Figure 3 illustrates the value of λ under each corresponding case. It can be seen that with the increase of the number of source nodes, the proposed algorithm requires more iterations to be completed, but the increasing trend is not rapid and it needs only 200 iterations with 1000 source nodes and 250 routes ($r = 4$). Almost in contrast to the trend of absolute iterations, the P-T ratio λ decreases sharply from the beginning with the number of source nodes increasing until around 200, where $\lambda < 0.01$, followed by a gradual decline. Moreover, the value of λ is fairly close to 0 with 1000 source nodes. The reason for this phenomenon may be that the given upper time bound is loose.

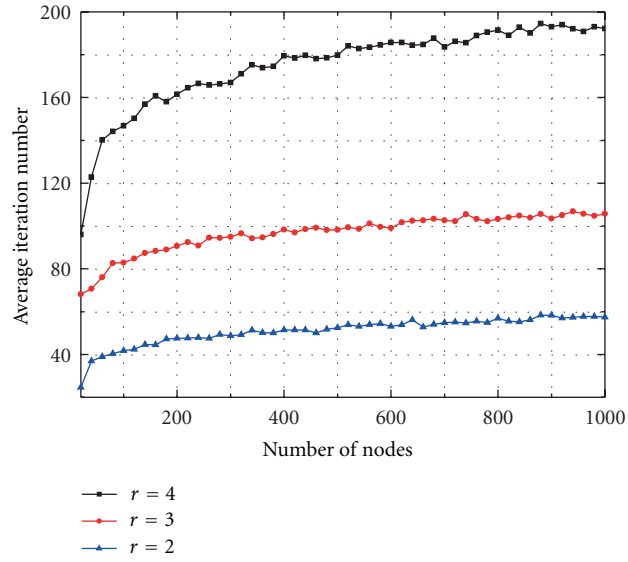


Figure 2: Number of iterations the proposed algorithm takes versus the number of source nodes.

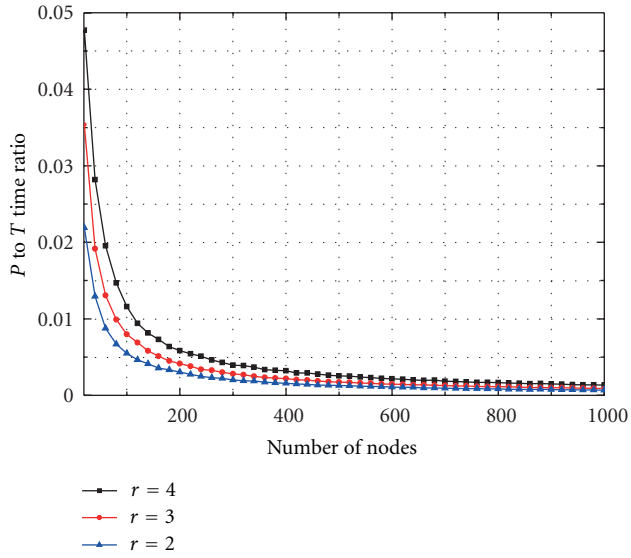


Figure 3: P-T ratio λ versus the number of source nodes.

Second, the degree of source nodes is considered as variable, and 10 groups of the maximum degree of the source nodes are set to $1, 2, \ldots, 10$, respectively. Figure 4 shows that the required average iterations by the proposed algorithm increases almost linearly with the increase in the number of source nodes. In particular, when the maximum degree of each source node is 1, only one iteration is needed because each has only one choice. When there are 400 source nodes with the maximum degree of 10, the proposed algorithm takes nearly 400 iterations to be completed. Figure 5 shows that λ reaches its maximum (slightly below 0.02) when each source node has two available routes on average. Except in a special case where each source node has only one choice,

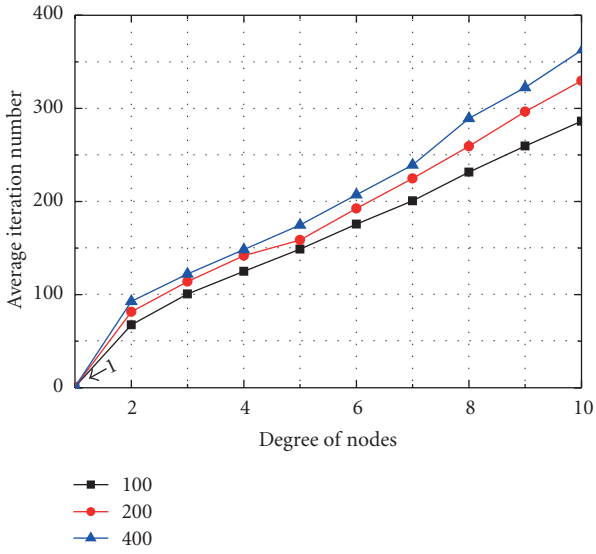


Figure 4: Number of iterations the proposed algorithm takes versus the degree of source nodes.

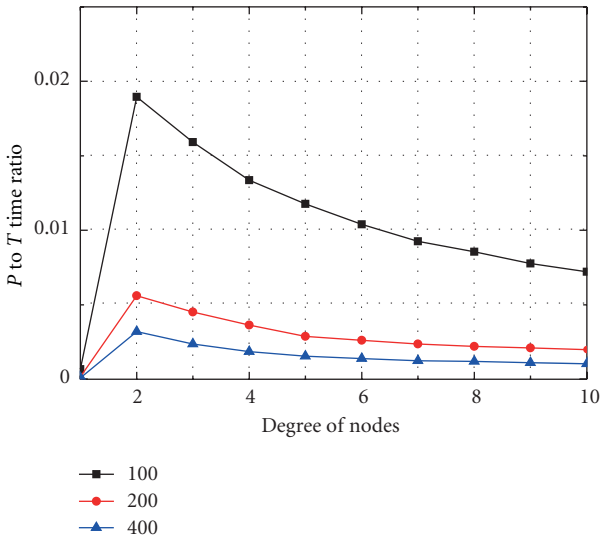


Figure 5: P-T ratio λ versus the degree of source nodes.

the P-T ratio λ decreases slowly as each source node has more choices because each has multiple routes to choose from.

On one hand, the given upper time bound is rather loose. On the other hand, the time bound is polynomial and the P-T ratio λ is extremely small in the given simulation. Therefore, the proposed algorithm has good speed performance.

6.2. Optimality. Section 4 indicates that a one-shot style may render the NE of the game undesirable. This part shows how the dynamic game can improve NE by simulation. Figure 6 provides the comparison results when the degree of source nodes is set to an average of 5 and r to 4. Evidently, the latency under OneShotHop is nearly triple of DynamicPricing. Moreover, OneShotPricing is merely better by one point compared with OneShotHop.

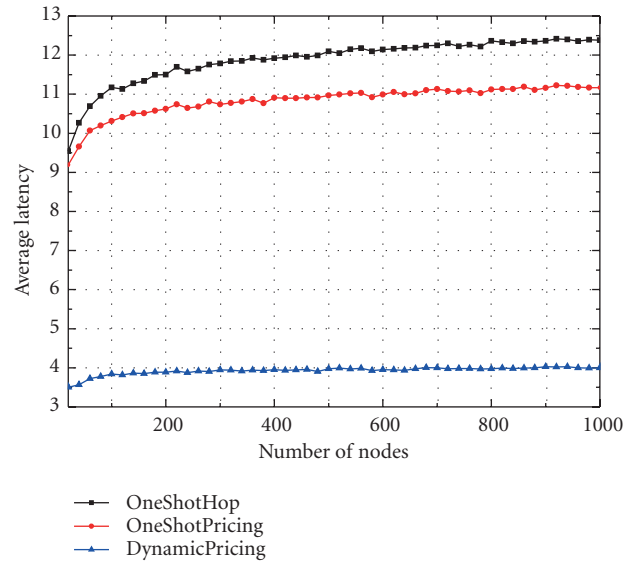


Figure 6: Latency under DynamicPricing, OneShotHop, and OneShotPricing versus the number of source nodes.

Since our proposed algorithm is distributed, there is no guarantee that it can compute the optimal NE every time as a centralized approach can. Therefore, learning the optimality of the NE generated by DP is significant. Here, several experiments are conducted to examine the optimality of the proposed algorithm by comparing it with a well-known centralized approach based on the cost-reducing path and to determine whether it can terminate in polynomial time [18]. Before showing the results, price of anarchy (PoA) [19], which measures the optimality of a game, is introduced as a metric and formally given as follows.

Definition 7 (PoA). For game G, let $\cos t(G)$ denote the social cost of G's NE, and let $\text{opt}(G)$ denote the minimum social cost over all assignments. The price of anarchy is then defined by

$$\text{PoA}(G) = \frac{\cos t(G)}{\text{opt}(G)}. \tag{11}$$

PoA is constantly not smaller than 1 (1 means optimal). The smaller the PoA is, the better the NE becomes.

First, the effect of the number of source nodes on the proposed algorithm's optimality is studied. The degree of source nodes is set to an average of 5. Figure 7 shows that as the number of source nodes increases, the NE's PoA tends to become larger despite small fluctuations. The maximum value of PoA is below 1.40 with 1000 source nodes and 500 routes. The corresponding ratio of the number of runs that generates an optimal NE to that of all runs (optimal-to-all for short) is given in Figure 8. Most of the runs can generate an optimal assignment.

Second, the degree of source nodes is considered to be variable. Ten groups of the maximum degree of the source nodes are set to $1, 2, \ldots, 10$, respectively. When each source

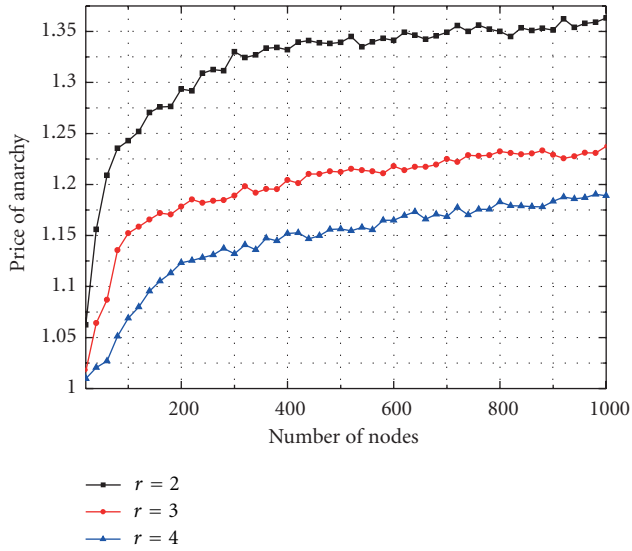


Figure 7: PoA versus the number of source nodes.

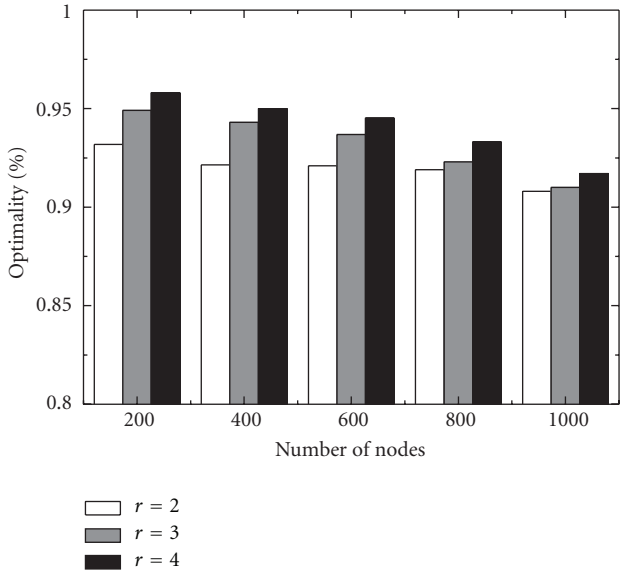


Figure 8: Ratio of optimal-to-all versus the number of source nodes.

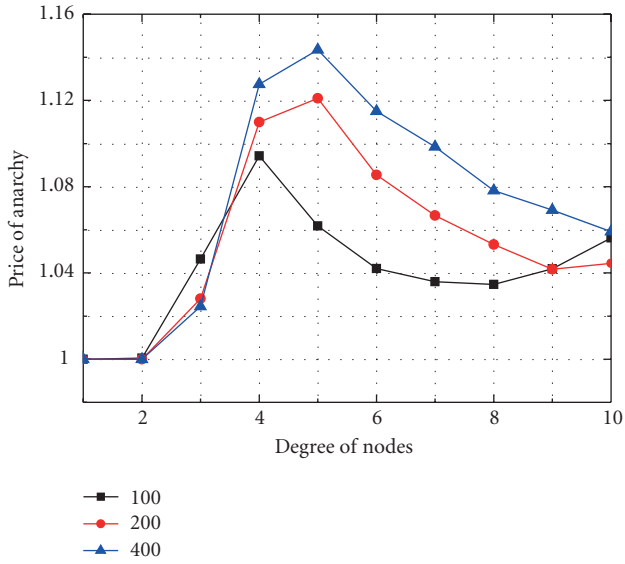


Figure 9: PoA versus the degree of source nodes.

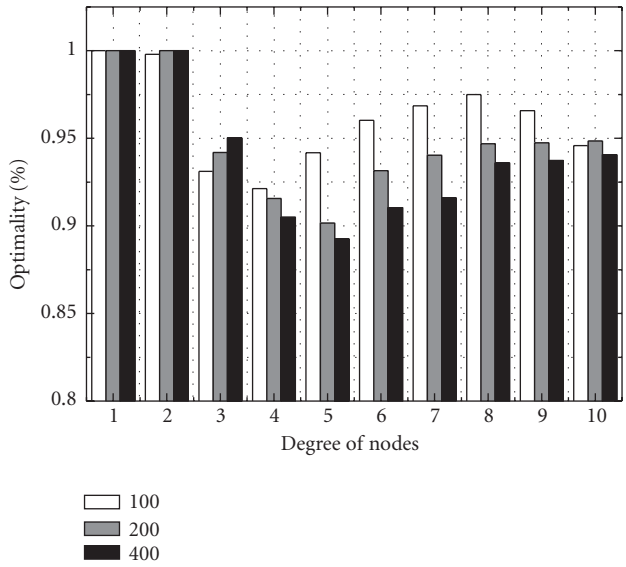


Figure 10: Ratio of optimal-to-all versus the degree of source nodes.

node has only one route to connect to, the assignment is unique, and thus PoA = 1, which is in accordance with that shown in Figure 9. If each source node has multiple choices, the optimality of the proposed algorithm becomes slightly poor. However, the maximum PoA does not exceed 1.16, which is close to 1. Figure 10 provides the information on the ratio of optimal-to-all in each case corresponding to Figure 9. Over 90% of runs can generate an optimal assignment.

7. Conclusion

The routing congestion problem in wireless M2M networks has been investigated using a game theoretic approach. A network congestion game has been formulated, in which many selfish source nodes compete for connecting to a smaller set of routes. Since full cooperation among nodes cannot be assumed in wireless M2M networks, pricing has been applied in the game to motivate cooperative data relay. By setting the pricing scheme dynamic, routes altering their prices and source nodes accordingly altering their connected routes, we have proved that the network congestion game converges to a pure NE in polynomial time. The optimality of the proposed algorithm was evaluated by simulation. The results indicate good practical performance.

Acknowledgments

This research is partially supported by research grants from the National Science Foundation of China under Grant-nos.70701025 and 71071105, the Program for New Century Excellent Talents in Universities of China under Grant no. NCET-08-0396, and a National Science Fund for Distinguished Young Scholars of China under Grant no. 70925005, and the Program for Changjiang Scholars and Innovative Research Team in the university (IRT1028).

References

[1] Q. Zhang and Y. Q. Zheng, "Cross-layer design for qos support in multihop wireless networks," *Proceedings of the IEEE*, vol. 96, no. 1, pp. 64–76, 2008.

[2] P. K. Huang, X. Lin, and C. C. Wang, "A low-complexity congestion control and scheduling algorithm for multihop wireless networks with order-optimal per-flow delay," in *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM '11)*, pp. 2588–2596, April 2011.

[3] K. Tan, F. Jiang, Q. Zhang, and X. Shen, "Congestion control in multihop wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 2, pp. 863–873, 2007.

[4] H. J. Lee and J. T. Lim, "Congestion control for streaming service in IEEE 802.11 multihop networks," *IET Communications*, vol. 4, no. 12, pp. 1415–1422, 2010.

[5] J. Y. Teo, Y. Ha, and C. K. Tham, "Interference-minimized multipath routing with congestion control in wireless sensor network for high-rate streaming," *IEEE Transactions on Mobile Computing*, vol. 7, no. 9, pp. 1124–1137, 2008.

[6] L. Chen, S. H. Low, M. Chiangs, and J. C. Doyle, "Cross-layer congestion control, routing and scheduling design in ad hoc wireless networks," in *Proceedings of the 25th IEEE International Conference on Computer Communications (INFOCOM '06)*, April 2006.

[7] B. Vöcking and R. Aachen, "Congestion games: optimization in competition," in *Proceedings of the Proceedings of the 2nd Algorithms and Complexity in Durham Workshop*, H. Broersma, S. Dantchev, M. Johnson, and S. Szeider, Eds., Kings College Publications, London, UK, 2006.

[8] H. Sperber, "How to find Nash equilibria with extreme total latency in network congestion games?" in *Proceedings of the International Conference on Game Theory for Networks (GameNets '09)*, pp. 158–163, May 2009.

[9] D. Fotakis, S. Kontogiannis, and P. Spirakis, "Symmetry in network congestion games: pure equilibria and anarchy cost," *Approximation and Online Algorithms*, vol. 3879, pp. 161–175, 2006.

[10] C. A. Gizelis and D. D. Vergados, "A survey of pricing schemes in wireless networks," *IEEE Communications Surveys and Tutorials*, vol. 13, no. 1, pp. 126–145, 2011.

[11] X. Zhu, S. Han, and B. Girod, "Congestion-aware rate allocation for multipath video streaming over ad hoc wireless networks," in *Proceedings of the International Conference on Image Processing (ICIP '04)*, pp. 2547–2550, October 2004.

[12] L. Chen, S. H. Low, and J. C. Doyle, "Joint congestion control and media access control design for ad hoc wireless networks," in *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM '05)*, pp. 2212–2222, March 2005.

[13] M. Chiang, "Balancing transport and physical layers in wireless multihop networks: jointly optimal congestion control and power control," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 104–116, 2005.

[14] C. Wang, B. Li, K. Sohraby, M. Daneshmand, and Y. Hu, "Upstream congestion control in wireless sensor networks through cross-layer optimization," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 786–795, 2007.

[15] P. Wang and S. Bohacek, "Practical computation of optimal schedules in multihop wireless networks," *IEEE/ACM Transactions on Networking*, vol. 19, no. 2, pp. 305–318, 2011.

[16] C. Jiang, Y. Shi, Y. T. Hou, and S. Kompella, "On optimal throughput-energy curve for multi-hop wireless networks," in *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM '11)*, pp. 1341–1349, April 2011.

[17] Y. P. Hsu and K. T. Feng, "Cross-layer routing for congestion control in wireless sensor networks," in *Proceedings of the IEEE Radio and Wireless Symposium (RWS '08)*, pp. 783–786, January 2008.

[18] N. J. A. Harvey, R. E. Ladner, L. Lovász, and T. Tamir, "Semi-matchings for bipartite graphs and load balancing," *Journal of Algorithms*, vol. 59, no. 1, pp. 53–78, 2006.

[19] C. H. Papadimitriou, "Algorithms, games, and the internet," in *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing*, pp. 749–753, July 2001.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 963758, 11 pages
doi:10.1155/2012/963758

Research Article

Optimal Routing Control in Disconnected Machine-to-Machine Networks

Yahui Wu, Su Deng, and Hongbin Huang

Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China

Correspondence should be addressed to Yahui Wu, wuyahui@nudt.edu.cn

Received 18 April 2012; Revised 9 June 2012; Accepted 25 June 2012

Academic Editor: Lin Bai

Machine-to-machine (M2M) networks could be connected by a wide range of wireless technologies (e.g., Bluetooth, WiFi, RFID). Due to some factors (e.g., mobility of machines, limited communication range), it is hard to maintain the connectivity of the network, that is, the network is disconnected and it is a specific application of delay tolerant networks (DTN). Communication in such network often needs nodes working in a cooperative way. However, due to selfishness, nodes have no incentive to help others. Therefore, if the source requests help from other nodes, it needs to pay certain fees to them. The main goal of this paper is to explore efficient ways for the source to maximize the average delivery ratio when the total fees are limited. First, we mathematically characterize the average delivery ratio under different policies. Then we get the optimal policy through Pontryagin's maximum principle, and prove that the optimal policy conforms to the *threshold* form when the fees that other nodes require satisfy certain conditions. Simulations based on both synthetic and real motion traces show the accuracy of our model. Through extensive numerical results, we demonstrate that the optimal policy obtained by our model is the best.

1. Introduction

At present, Machine-to-machine (M2M) networks have emerged as an innovative topic and are undergoing rapid development [1]. In particular, M2M networks could be connected by a wide range of wireless technologies (e.g., Bluetooth, WiFi). Due to the limited communication range, mobility of machines or other factors, it is hard to maintain the connectivity of the M2M networks. On the other hand, the concept of delay tolerant networks (DTN) has been proposed to support many new networking applications, where the end-to-end connectivity cannot be assumed [2]. Therefore, M2M networks can be seen as a specific application of DTN. In order to provide communication services in such disconnected networks, nodes in DTN communicate through a *store-carry-forward* way. In particular, when the next hop is not immediately available for the current node, some relay nodes will *store* the message in their buffer, *carry* the message along their movements, and *forward* the message to other nodes when a new communication opportunity is occurring [2].

By abuse of language, we will use node and machine alternately in this paper. It is easily to see that the *store-carry-forward* communication mode needs nodes working in a cooperative way. In particular, the source (one specific machine) may need other nodes serving as *relays* to forward message to the destination (another machine). In the real world, machines may be vehicles, smart phones, and so on. Though nodes in M2M networks can communicate with each other without human intervention, they can be manipulated by human, too. Therefore, peoples' social behavior (e.g., selfishness) can have certain impact on communication. Due to the selfish nature, nodes have no incentive to help others [3, 4]. To make these nodes be cooperative, some incentive policies are necessary. This paper adopts the fees-based incentive policy, that is, the source has to pay certain fees to other nodes if it requests help from them. The fees may be real or virtual currency, by which nodes can buy certain services (e.g., downloading files) from others. Moreover, these fees may be varying with time. In fact, the buffer space or the forwarding ability can be seen as goods. The event that the source requests help from other

nodes can be seen as that the source buys certain goods from them, so the message transmission process can be seen as the commodities trading process, and nodes want to maximize its income in this process. Therefore, nodes may adjust the price of their goods according to the market state. For example, if the remaining lifetime of message is shorter, nodes may guess that the source is eager to transmit the message as soon as possible, so they may guess that their goods (e.g., buffer space, the forwarding ability) are important for the source. In this case, they may increase the price. In addition, if the remaining lifetime of the message is long, nodes may guess that the source may be not eager to transmit the message quickly and is not willing to pay more fees. In this case, they may help the source with only a fewer fees. Therefore, the fees that the ordinary nodes require may be varying with time. If the source has enough fees and can make every node be cooperative, the message will be transmitted according to the classic epidemic routing (ER) algorithm [5], but it has to pay more fees. As the buyers, the source may be not willing to pay more fees, for example, it may think that it is not economical [6]. Therefore, when the total fees are limited, how to maximize the probability that the destination gets message before the message deadline is a very important problem.

Note that our work is similar to the packet purse model (PPM) of credit-based incentive policy, in which nodes can get certain *nuggets* (denoted by fees in our work) from the source if they help the source [6]. Therefore, the fees in our work can be seen as *nuggets* in the PPM model. However, state of the art works about the PPM model have some differences with our work. First, most of them do not consider the case that the fees that nodes require may be varying with time. In addition, they do not study how to use the limited *nuggets* efficiently. For example, supposing that node j is willing to help the source by charging m*nuggets* at time $t1$, but it only requires n *nuggets* at time $t2$. If $m > n$ and $t1 < t2$, the source can pay less *nuggets* when it requires help from node j at time $t2$, but this may decrease the probability that node j forwards message to the destination, so it may be not good for the source. On the other hand, because the source uses fewer *nuggets* to make node j be cooperative, the remaining *nuggets* are more, so the source may have enough *nuggets* to make more nodes be cooperative and this is good for the source. Therefore, how to use the limited *nuggets* of the source is not a simple problem, and this will be our main contribution. At last, most of existing works assume that once a node obtains certain *nuggets*, they will be cooperative all the time. However, the cost (e.g., energy) of nodes in the routing process is closely related with the *forwarding times* [7], so nodes may ask for *nuggets* according to the *forwarding times*. In other words, once they forward message to one node, the source has to pay certain fees to them.

On the other hand, our work is similar to the optimal control problem of ER algorithm in DTN, which is a popular topic recently, and there are some good works in this field, such as [7, 8]. However, they do not consider the selfishness nature of nodes. For the selfishness behavior, there are some works, which evaluate its impact on the routing performance [3, 4]. However, none of these works considers the optimal control problem. In addition, the selfishness behavior in those works is not varying with time and is denoted by a simple way. In particular, they use the probabilistic way to denote the selfishness behavior. For example, node i is willing to help node j with probability q.

The main contribution of this paper is to study the optimal routing problem in such complex application. In particular, we propose a unifying framework through a continuous time Markov process, which can be used to evaluate the total fees that the source pays. Then based on the framework, we formulate an optimization problem. Through Pontryagin's maximum principle, we explore the stochastic control problem, and we prove that the optimal policy conforms to the *threshold* form in some cases. By comparing the simulation results with the theoretical results, we show that our theoretical framework is very accurate. We compare the performance of the optimal policy with other policies through extensive numerical results and find that the optimal policy obtained by Pontryagin's maximum principle is the best.

The rest of this paper is organized as follows. In Section 2, we briefly present some related works, and we describe the network model in Section 3. In Section 4, we first present the theoretical framework, and then formulate the optimal control problem. In Section 5, we introduce the simulation and numerical results. Finally, we conclude our main work in Section 6.

2. Related Work

In the past few years, many routing protocols have been proposed in DTN, but most of them need certain prior knowledge of the network, or they can obtain such knowledge through some online learning processes, such as the works in [9–11]. In some applications, there may be no enough time to learn, so these algorithms have certain shortage. On the other hand, ER algorithm does not need any prior knowledge about the network and can be used in many environments. For this reason, this algorithm is still a very popular topic. However, it works in a flooding way, so it wastes much energy and suffers from poor scalability in large networks. At present, many policies have been proposed to reduce its overhead. Among them, there are probability forwarding policy [12], hop-based forwarding policy [13], and so forth. These algorithms try to achieve big delivery ratio and relatively low transmission cost. Generally speaking, big delivery ratio is obtained at the expense of more cost. Therefore, how to accurately evaluate both strengths and limitations of these algorithms is very important. Some works use the simulation manner [14], but recently, theoretical manner is more popular. The performance of ER algorithm based on the sparsely exponential graph is studied in [15] and then the problem is explored again with heterogeneous nodes in [16]. The authors in [17] get the generic theoretical upper bounds for the information propagation speed in large-scale mobile and intermittently connected network, and then the work in [18] explores the information propagation speed in bidirectional vehicular delay tolerant network. The work in [19] studies

the performance of two-hop relay routing under limited packet lifetime. Performance of the ER routing when the contention exists is explored in [20].

Most routing algorithms in DTN need nodes working in a cooperative way, so the selfishness behavior can have important impact on the performance. Panagakis et al. evaluate the impact of selfishness through simulation in [21]. There are also some works, which study the impact of nodes' selfishness behavior by theoretical method, such as [3, 4]. The work in [22] is the first to propose the *social selfishness* behavior in DTN and proposes a user-centered routing policy, which is adaptive to the selfishness nature. Specially, the *social selfishness* behavior means that nodes are more willing to help the one with whom they have certain social ties (e.g., citation relation). The work in [23] studies the impact of the *social selfishness* behavior on ER algorithm and finds that ER algorithm is very robust to the *social selfishness* behavior. Then they study its impact on multicasting in DTN [24].

Because ER algorithm has certain shortage, there are some works, which begin to consider the optimal control problem. The optimal control problem of two-hop routing algorithm based on discrete time model is studied in [7], and this work proves that the optimal forwarding policy conforms to the *threshold* form. Then the work in [8] explores the problem again with a continuous time Markov process. The work in [25] proposes an optimal activation and transmission policy, and then the work in [26] proposes an energy-efficient optimal beaconing policy. Above works try to maximize certain objective function under some constraints, but the work in [27] study the trade-off between the delivery delay and energy consumption in DTN that uses two-hop relaying method. This work is different from our work in certain aspects. First, the work in [27] is an optimization problem without constraint. Second, the energy consumption within given time is fixed. In particular, it is a special case of our work when PR(t) (see the definition in next section) is a constant. Third, that work uses the two-hop routing method, which is too simple. There are also some works, which consider the optimal control of ER algorithm with *SIRD* model, in which nodes carrying message (e.g., virus) may exist from the network [28, 29].

3. Network Model

In this paper, we assume that there are one source node S, N ordinary nodes and one destination node D, so the network totally has $N + 2$ nodes. At time 0, only S is carrying message and it tries to make the destination D obtain message before the deadline T.

To make the destination get message quickly, S requests help from the ordinary nodes. If one ordinary node (e.g., j) gets message, it can forward message to other node. Due to the selfishness nature, S has to pay certain fees to j every time j forward message to other node. Therefore, considering the cost constraint, S needs to control the behavior of itself and the ordinary nodes. In this paper, we assume that both S and the ordinary nodes have the same forwarding policy. In particular, any node carrying message forwards message

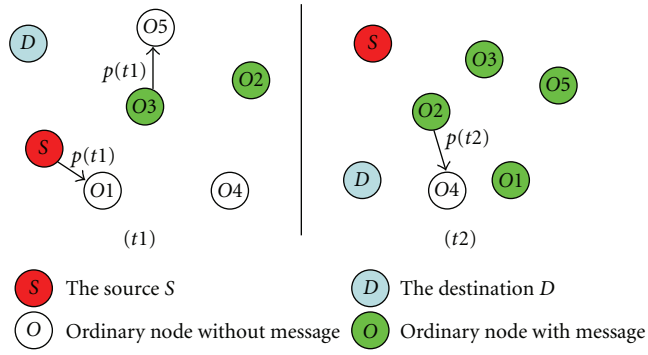


Figure 1: The snapshot of the network at time $t1$ and $t2$.

to other ordinary nodes with probability $p(t)$ at time t. However, they forward message to the destination D with probability 1. That is to say, once a node carrying message encounters with D, D can get message immediately. This may need the third organization to make the ordinary nodes abide by the rules. On the other hand, if an ordinary node (e.g., j) which is carrying message forwards message to another node at time t, the fees that S pays to j is PR(t). The goal of this paper is to explore the optimal value of $p(t)$ at any time t to maximize the average delivery ratio when the total fees that the source can pay is limited. The snapshot at time $t1$ and $t2$ can be seen as in Figure 1. At time $t1$, ordinary node $O1$ encounters with the source, and the source will forward message to $O1$ with probability $p(t1)$. In addition, node $O5$ encounters with the ordinary node $O3$, which is carrying message at time $t1$. Because ordinary nodes have the same forwarding policy as the source, node $O3$ will forward message to $O5$ with probability $p(t1)$, too. If the ordinary $O3$ successfully forwards message to $O5$, the source has to pay PR($t1$) fees to $O3$. Similarly, at time $t2$, node $O4$ encounters with the ordinary node $O2$, and node $O2$ will forward message to $O2$ with probability $p(t2)$.

Nodes in the network can communicate with each other only when they come into the transmission range of each other, which means a communication contact, so the mobility rule of nodes is critical. In this paper, we assume that the occurrence of contacts between two nodes follows a Poisson distribution. This assumption has been used in wireless communications many years. At present, some works show that this assumption is only an approximation to the message propagation process, and they reveal that nodes encounter with each other according to the power law distribution [30]. However, they also find that if you consider long traces, the tail of the distribution is exponential. In addition, the work in [31] shows that individual inter-meeting time can be shaped to be exponential by choosing an approximate domain size with respect to given time scale. Moreover, there are also some works, which describe the intermeeting time of human or vehicles by Poisson process and validate their model experimentally on real motion traces [32, 33]. According to these descriptions, the exponential intermeeting time is rational in some applications, and we assume that the intermeeting time between two nodes follows an exponential distribution with parameter λ. Simulations based on both

Table 1: The list of commonly used variables.

N	Number of ordinary nodes
λ	Exponential parameter of the intermeeting time
T	The maximal lifetime of the message
$\mathrm{PR}(t)$	The fees that the ordinary node require at time t
$p(t)$	Forwarding probability at time t
$X(t)$	The number of ordinary nodes carrying message at time t
$F(t)$	The delivery ratio at time t
$U(t)$	The total fees that the source pays till time t

synthetic and real motion traces show that our theoretical framework based on such assumption is very accurate. The list of commonly used variables can be seen as in Table 1.

4. Optimization Formulation

4.1. Theoretical Framework. Let $X(t)$ denote the number of ordinary nodes carrying message at time t. Only the source has message at time 0, so we have $X(0) = 0$. In this paper, we assume that nodes carrying message do not receive the same message anymore. Given a small time interval Δt, we can obtain

$$X(t+\Delta t) = X(t) + \sum_{j\in\{Y(t)\}} \varphi_j(t, t+\Delta t). \tag{1}$$

Symbol $\{Y(t)\}$ denotes the set of ordinary nodes without message at time t, so the set has $N - X(t)$ elements. $\varphi_j(t, t+\Delta t)$ denotes the event whether the ordinary node j gets message in time interval $[t, t+\Delta t]$. If $\varphi_j(t, t+\Delta t) = 1$, we can say that node j gets message, but if $\varphi_j(t, t+\Delta t) = 0$, node j does not get message. As shown above, two nodes encounter with each other according to an exponential distribution with parameter λ. Therefore, node j encounters with a specific node (e.g., k) in interval $[t, t+\Delta t]$ with probability $1 - e^{-\lambda\Delta t}$. Moreover, if node k has message and j encounters with k at time $t+t1$, node j receives message with probability $p(t+t1)(1-e^{-\lambda\Delta t})$, $0 \le t1 \le \Delta t$. Because time interval Δt is very small, we can say that such probability equals to $p(t)(1-e^{-\lambda\Delta t})$. In addition, there are $1+X(t)$ nodes which have message at time t, so we have

$$p\left(\varphi_j(t, t+\Delta t) = 1\right) = 1 - \left(1 - p(t)\left(1 - e^{-\lambda\Delta t}\right)\right)^{1+X(t)}. \tag{2}$$

Therefore, we can get the expectation of $\varphi_j(t, t+\Delta t)$ as follows:

$$E\left(\varphi_j(t, t+\Delta t)\right) = E\left(1 - \left(1 - p(t)\left(1 - e^{-\lambda\Delta t}\right)\right)^{1+X(t)}\right). \tag{3}$$

Base on (1) and (3), we have

$$\begin{aligned} E(X(t+\Delta t)) &= E(X(t)) + (N - E(X(t)))E\left(\varphi_j(t, t+\Delta t)\right) \\ &\Longrightarrow \lim_{\Delta t\to 0} \frac{E(X(t+\Delta t)) - E(X(t))}{\Delta t} \\ &= \lim_{\Delta t\to 0} \frac{(N - E(X(t)))E\left(\varphi_j(t, t+\Delta t)\right)}{\Delta t} \\ &\Longrightarrow E(\dot{X}(t)) = \lambda p(t)(N - E(X(t)))(1 + E(X(t))). \end{aligned} \tag{4}$$

Note that $E(X(t))$ denotes the expectation of $X(t)$. One main metric of routing algorithm in DTN is the delivery ratio, which denotes the probability that the destination D obtains message within given time. Let $F(t)$ denote the delivery ratio when the given time is t. Before getting its value, we first give another symbol $W(t) = 1 - F(t)$, which denotes the probability that D does not obtain message before time t. Moreover, let $W(t, t+\Delta t)$ denote the probability that D does not get message in time interval $[t, t+\Delta t]$. Therefore, we have

$$W(t+\Delta t) = W(t)W(t, t+\Delta t). \tag{5}$$

There are $1 + X(t)$ nodes which may forward message to D at time t, so we can get the following equation easily:

$$W(t, t+\Delta t) = e^{-\lambda\Delta t(1+X(t))}. \tag{6}$$

Further, we can obtain

$$\begin{aligned} E(\dot{W}(t)) &= -\lambda E(W(t))(1 + E(X(t))), \\ E(\dot{F}(t)) &= \lambda(1 - E(F(t)))(1 + E(X(t))). \end{aligned} \tag{7}$$

Let $U(t)$ denote the total fees that the source pays till time t, we have

$$U(t+\Delta t) = U(t) + \sum_{j\in\{Y(t)\}} \mathrm{PR}\left(t+\rho_j\right)\sigma_j(t, t+\Delta t). \tag{8}$$

Let $\sigma_j(t, t+\Delta t)$ denotes that whether the ordinary node j gets message from another ordinary node in interval $[t, t+\Delta t]$. If this event happens (e.g., at time $t+\rho_j$, $0 \le \rho_j \le \Delta t$), one ordinary node carrying message forward message to another node, so S has to pay certain fees denoted by $\mathrm{PR}(t+\rho_j)$. In fact, if the destination D gets message from an ordinary node, the source has to pay certain fees to this ordinary node, too. However, the network only has one destination. Therefore, we can ignore the cost in this paper. Note that $\varphi_j(t, t+\Delta t)$ denotes the event whether the ordinary node j gets message in time interval $[t, t+\Delta t]$, so we have

$$p\left(\sigma_j(t, t+\Delta t) = 1\right) = p\left(\varphi_j(t, t+\Delta t) = 1\right)\frac{X(t)}{1+X(t)}. \tag{9}$$

Based on (8) and (9), we have

$$\lim_{\Delta t\to 0}\frac{E(U(t+\Delta t)-U(t))}{\Delta t} = \mathrm{PR}(t)\lim_{\Delta t\to 0}\frac{E\left(\sum_{j\in\{Y(t)\}}\sigma_j(t,t+\Delta t)\right)}{\Delta t} \tag{10}$$

$$\Longrightarrow E(\dot{U}(t)) = \frac{E(X(t))}{1+E(X(t))}\mathrm{PR}(t)E(\dot{X}(t)).$$

Then, we have

$$E(U(T)) = \int_0^T \frac{E(X(t))}{1+E(X(t))}\mathrm{PR}(t)E(\dot{X}(t))dt. \tag{11}$$

4.2. Optimal Control. Our object is to solve the following optimization problem:

$$\begin{aligned}&\max\ E(F(T)) = \int_0^T E(\dot{F}(t))dt,\\ &\text{subject } E(U(T)) \le C.\end{aligned} \tag{12}$$

Symbol C denotes the maximal fees that the source can pay, and T is the maximal lifetime of the message. The controlling variable is $p(t)$. Let $((X, F), p)$ be an optimal solution. In particular, at time t, X denotes the value of $E(X(t))$, and F denotes the value of $E(F(t))$. Similarly, p denotes the value of $p(t)$. In addition, U denotes the value of $E(U(t))$ at time t. Consider the *Hamiltonian* H, and *costate* or *adjoint* functions λ_X, λ_F and λ_U defined as follows:

$$\begin{aligned}H &= \dot{F} + \lambda_X\dot{X} + \lambda_F\dot{F} + \lambda_U\dot{U}\\ &= (1+\lambda_F)\dot{F} + \left(\lambda_X + \lambda_U\mathrm{PR}\frac{X}{1+X}\right)\dot{X}\\ &= \lambda(1+\lambda_F)(1+X)(1-F)\\ &\quad + \lambda(\lambda_X(1+X)+\lambda_U\mathrm{PR}X)(N-X)p,\end{aligned} \tag{13}$$

$$\begin{aligned}\dot{\lambda}_X &= -\frac{\partial H}{\partial X} = -(1+\lambda_F)\lambda(1-F)\\ &\quad +\lambda(\lambda_X(1+X)+\lambda_U\mathrm{PR}X)p\\ &\quad -\lambda(\lambda_X+\lambda_U\mathrm{PR})(N-X)p\\ \dot{\lambda}_F &= -\frac{\partial H}{\partial F} = \lambda(1+\lambda_F)(1+X)\\ \dot{\lambda}_U &= -\frac{\partial H}{\partial U} = 0.\end{aligned} \tag{14}$$

Note that at time t, PR is a simple expression of $\mathrm{PR}(t)$.

The *transversality* conditions are shown as follows:

$$\lambda_X(T) = \lambda_F(T) = \lambda_U(T)(U(T)-C) = 0, \quad \lambda_U(T) \le 0. \tag{15}$$

Then according to Pontryagin's maximum principle ([34, Page 109, Theorem 3.14]), there exist continuous or piecewise continuously differentiable state and *costate* functions, which satisfy

$$p \in \arg\max_{0\le p^*\le 1} H(\lambda_X,\lambda_F,\lambda_U,(X,F,U),p^*). \tag{16}$$

This equation between the optimal control parameter p and the *Hamiltonian* H allows us to express p as a function of the state (X, F, U) and *costate* $(\lambda_X, \lambda_F, \lambda_U)$, resulting in a system of differential equations involving only the state and *costate* functions, and not the control function. In fact, this equation means that maximize the value of $E(F(T))$ equals to maximize the corresponding *Hamiltonian* H. In particular, at given time t, the state (X, F, U) and *costate* $(\lambda_X, \lambda_F, \lambda_U)$ can be seen as constants, and $p(t)$ can maximize H at this time. Therefore, according to (14), we can obtain the optimal policy as follows:

$$p = \begin{cases}1, & (\lambda_X(1+X)+\lambda_U\mathrm{PR}X)(N-X) > 0,\\ 0, & (\lambda_X(1+X)+\lambda_U\mathrm{PR}X)(N-X) < 0.\end{cases} \tag{17}$$

The total number of ordinary nodes is N, so we have $X \le N$. When $X = N$, every ordinary node is carrying message, and the nodes cannot forward message to the ordinary node any more, so the value of p cannot have any impact and it may be any value. If $X < N$, we have $N - X > 0$, so if $\lambda_X(1+X)+\lambda_U\mathrm{PR}X > 0$, we can obtain $(N-X)(\lambda_X(1+X)+\lambda_U\mathrm{PR}X) > 0$. Therefore, H is increasing with p at present, and the optimal value of p is 1. The optimal value of p can be obtained easily in other cases in similar way, and we have (17). For simplicity, we only consider the case $X < N$ next in this paper, that is, at least one ordinary node is not carrying message.

Below, we will prove that when the function of $\mathrm{PR}(t)$ satisfies certain conditions, the optimal policy has a simple structure. The conditions are: $\mathrm{PR}(t)$ is nondecreasing with time t; it is differentiable; it is nonnegative. In fact, the maximal lifetime (T) of the message is fixed, so if the value of t is bigger, the remaining lifetime $(T - t)$ is shorter. In this case, the ordinary nodes may guess that the source may be eager to transmit message to D quickly, so they may ask for more fees from the source. That is to say, if the value of t is bigger, the value of $\mathrm{PR}(t)$ may be bigger. Therefore, the condition that $\mathrm{PR}(t)$ is increasing is rational in some environments. In this case, the optimal policy has at most one jump, and it conforms to the *threshold* form. Then we have the following Theorem.

Theorem 1. *If the* $\mathrm{PR}(t)$ *satisfies the above conditions, the optimal policy satisfies:* $p(t) = 1$, $t < h$, *and* $p(t) = 0$, $t > h$, $0 \le h \le T$.

Proof. This theorem means that the source requests help with probability 1 before time h, but it stops requesting help after time h. We simply use $X(t)$ or X to denote $E(X(t))$ in the proving process, and we only consider the case that $X < N$. Now, we define a new function as follow:

$$f(t) = \lambda_X(t)(1+X(t)) + \lambda_U(t)\mathrm{PR}(t)X(t). \tag{18}$$

Therefore, we have

$$\begin{aligned}\dot{f} &= \dot{\lambda}_X(1+X) + \lambda_X\dot{X} + \dot{\lambda}_U\mathrm{PR}X + \lambda_U\dot{\mathrm{PR}}X + \lambda_U\mathrm{PR}\dot{X}\\ &= \dot{\lambda}_X(1+X) + \lambda_X\dot{X} + \lambda_U\dot{\mathrm{PR}}X + \lambda_U\mathrm{PR}\dot{X}.\end{aligned} \tag{19}$$

If $f(t) < 0$, we know that $p(t) = 0$ according to (17), so we have

$$\dot{f}(t) = \dot{\lambda}_X(t)(1 + X(t)) + \lambda_U(t)\dot{\mathrm{PR}}(t)X(t). \tag{20}$$

Then based on (14) and $p(t) = 0$, we know that

$$\dot{\lambda}_X(t) = -(1 + \lambda_F(t))\lambda(1 - F(t)). \tag{21}$$

In fact, we have $1 + \lambda_F > 0$ at any time. Otherwise, if $1 + \lambda_F \leq 0$, we have

$$\begin{aligned} &\lambda_F \leq -1 < 0, \\ &\dot{\lambda}_F = \lambda(1 + \lambda_F)(1 + X) \leq 0. \end{aligned} \tag{22}$$

From (22), we can see that λ_F cannot be increasing, and it cannot be 0 at time T. This is contradiction with (15). Therefore, we have $1 + \lambda_F > 0$ at any time. Based on (20) and (21) and the fact that PR is nondecreasing with t and λ_U is nonpositive, we can see that f is not increasing at time t. Because $f(t) < 0$, we have $f(s) < 0$, $s > t$. Further, we have $p(s) = 0$, $s > t$. Therefore, once the forwarding probability equals to 0, it remains unchanged.

Now, we assume that $f(t) = \lambda_X(t)(1 + X(t)) + \lambda_U(t)\mathrm{PR}(t)X(t) = 0$, and then we have

$$\begin{aligned} \dot{\lambda}_X(t) = &-(1 + \lambda_F(t))\lambda(1 - F(t)) \\ &- \lambda(\lambda_X(t) + \lambda_U(t)\mathrm{PR}(t))(N - X(t))p(t). \end{aligned} \tag{23}$$

Combined with (19), we have,

$$\begin{aligned} \dot{f}(t) &= \dot{\lambda}_X(t)(1 + X(t)) + \lambda_X(t)\dot{X}(t) + \lambda_U(t)\dot{\mathrm{PR}}(t)X(t) \\ &\quad + \lambda_U(t)\mathrm{PR}(t)\dot{X}(t) \\ &= -(1 + \lambda_F(t))\dot{F}(t) - (\lambda_X(t) + \lambda_U(t)\mathrm{PR}(t))\dot{X}(t) \\ &\quad + \lambda_X(t)\dot{X}(t) + \lambda_U(t)\dot{\mathrm{PR}}(t)X(t) + \lambda_U(t)\mathrm{PR}(t)\dot{X}(t) \\ &= -(1 + \lambda_F(t))\dot{F}(t) + \lambda_U(t)\dot{\mathrm{PR}}(t)X(t). \end{aligned} \tag{24}$$

Note that we have proved that $1 + \lambda_F > 0$ at any time. Based on (14) and (15), we know that λ_U is a nonpositive constant. If it is smaller than 0, we can easily know that $f(t)$ is decreasing at time t. If λ_U equals to 0, we have

$$\dot{f}(t) = -(1 + \lambda_F(t))\dot{F}(t). \tag{25}$$

If $E(F(t)) = 1$, the objective function reaches to the maximal value, and it is not necessary to forward anymore, so we only consider the case $E(F(t)) < 1$. In this case, based on (7), we have

$$\begin{aligned} &\dot{F}(t) > 0, \\ &\dot{f}(t) = -(1 + \lambda_F(t))\dot{F}(t) < 0. \end{aligned} \tag{26}$$

Because $f(t) = 0$, we have $f(s) < 0$, $s > t$. Further, we have $p(s) = 0$, $s > t$. This means that there is at most one time t at which p can be any value, and then p equals to 0 after time t.

In summary, for time h, if $f(h) \leq 0$, then we have $f(t) < 0$, $t > h$. Therefore, according to (17), the optimal policy satisfies: $p(t) = 1$, $t < h$, and $p(t) = 0$, $t > h$, $0 \leq h \leq T$. This proves that Theorem 1 is correct. □

In fact, Theorem 1 means that if $\mathrm{PR}(t)$ satisfies certain conditions, the optimal policy has a bang-bang structure. In particular, the source will require help from others with probability 1 before certain *threshold*, and then stop. In addition, the value of the *threshold* is denoted by h.

5. Simulation and Numerical Results

5.1. Model Validation. In this section, we will check the accuracy of our framework by comparing the theoretical results obtained by our model with the simulation results. We run several simulations using the opportunistic network environment (ONE) [35] based on both synthetic mobility model and realworld-based scenarios. For the synthetic mobility trace, we use the famous random waypoint (RWP) mobility model [36], which is commonly used in many mobile wireless networks. There are totally 500 nodes, and all nodes move according to the RWP model within a 10000 m × 1000 m terrain with a scale speed chosen from a uniform distribution from 4 m/s to 10 m/s. The communication range is 10 m. Moreover, the source and destination nodes are randomly selected among these nodes. For the realworld-based scenarios, we use a real motion traces from about 2100 operational taxis for about one month in Shanghai city collected by global position system (GPS) [37]. The location information of the taxis is recorded at every 40 seconds with the area of 102 km^2. We randomly pick 500 nodes from this trace, and the source and destination nodes are randomly selected among these nodes, too.

The first metric is the total fees that the source pays to the ordinary nodes under different forwarding policies. The source may forward message with any probability, that is, the value of $p(t)$ may be any value between 0 and 1 at time t. Because our main goal is to check the accuracy of our theoretical framework, we only consider two special cases. Case 1: $p(t) = 1$, $t \geq 0$; Case 2: $p(t) = 0.5$, $t \geq 0$. The first case means that the source requests help all the time and message is propagated according to ER algorithm, but in the second case, nodes forward message to other ordinary nodes with probability 0.5. The function of $\mathrm{PR}(t)$ may be any form. For simplicity, we define: $\mathrm{PR}(t) = 3(1 - e^{-t/10000})$. In this section, the total fees is not limited. Let the maximal message lifetime increase from 0 s to 10000 s, we can get the results in Figures 2 and 3, for the RWP model and Shanghai city motion trace, respectively.

From these results, we can see that the average deviation between the theoretical and the simulation results is very small. For example, the average deviation is about 2.92% for the RWP mobility model, and 4.01% for the Shanghai city motion trace. In addition, Figure 3 shows that the source has to pay more fees if it adopts the policy of Case 2. In fact, when it adopts the policy of Case 1, nodes can get message

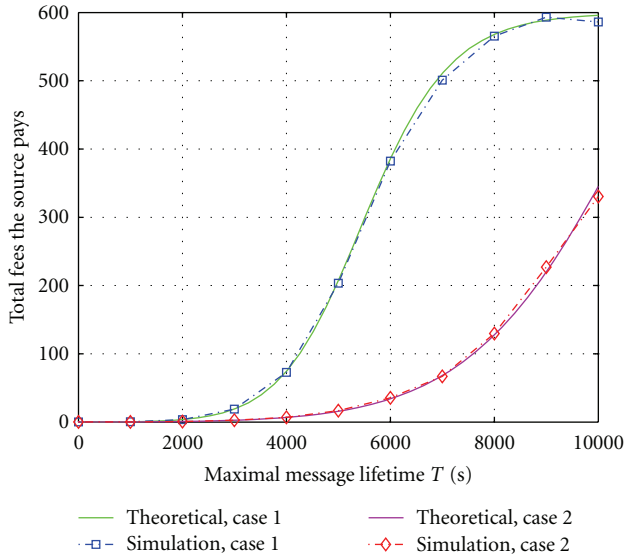


Figure 2: Simulation and numerical results comparison of total fees with RWP mobility model.

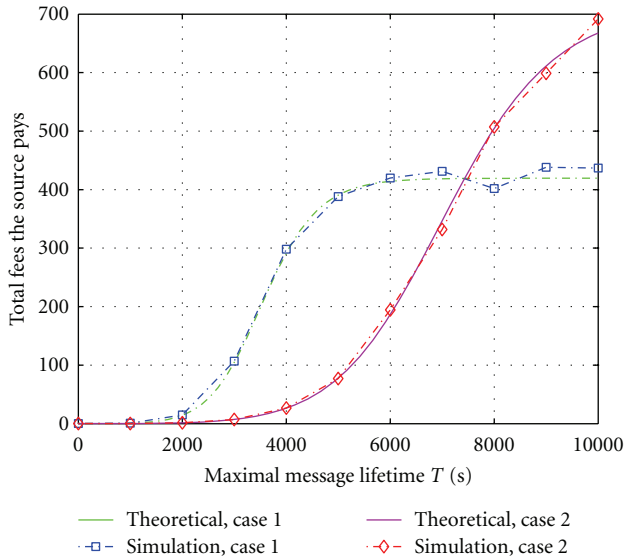


Figure 3: Simulation and numerical results comparison of total fees with Shanghai city mobility model.

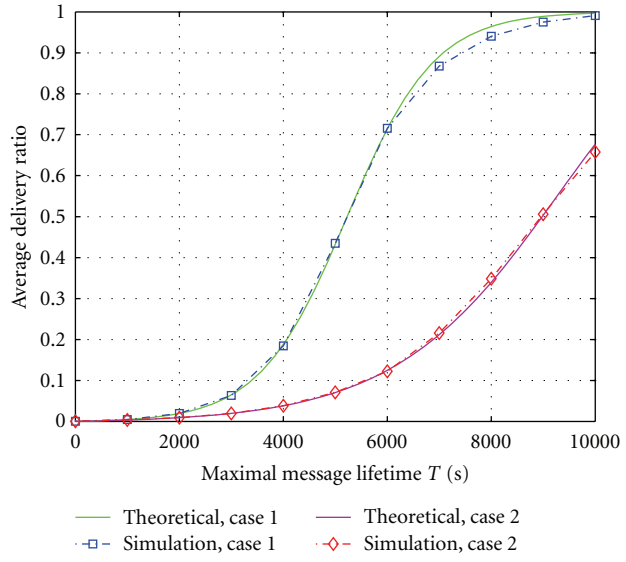


Figure 4: Simulation and numerical results comparison of average delivery ratio with RWP mobility model.

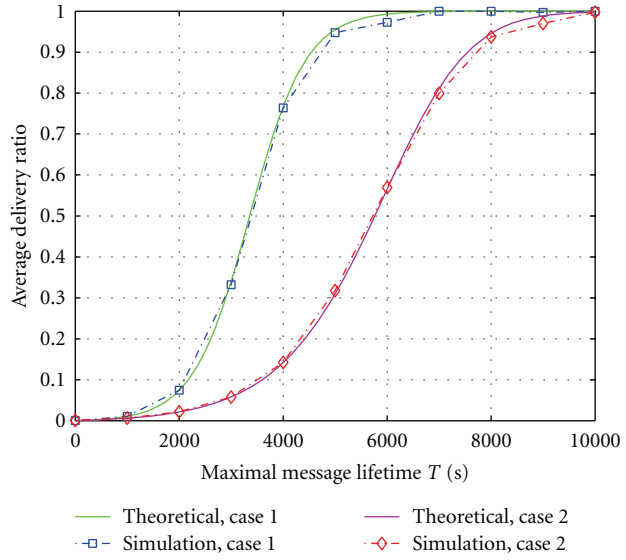


Figure 5: Performance comparison of different policies with Shanghai city mobility model.

timely, when the fees that the ordinary nodes required is less. However, if it adopts the policy of Case 2, many nodes can get message when the time t is bigger, so the source has to pay more fees to them.

Then based on the same settings, we explore the average delivery ratio in Figures 4 and 5, respectively. These results also show that the average deviation between the theoretical and the simulation results is very small.

To further check the accuracy of the model, we want to explore the performance when the number of nodes is different. In particular, we assume that there are 300 and 600 nodes, respectively. For simplicity, we only consider Case 1, that is, the value of $p(t)$ is 1 all the time. Other settings remain unchanged, and we can obtain Figures 6 and 7, respectively. Both results show that our theoretical framework is very accurate.

All of the above results demonstrate the accuracy of our theoretical framework. For this reason, we can use the numerical results obtained by our theoretical framework to evaluate the performance of different policies.

5.2. Performance Analysis with Numerical Results. In this section, all of the numerical results are obtained by our theoretical framework based on the best fitting for the Shanghai city motion trace.

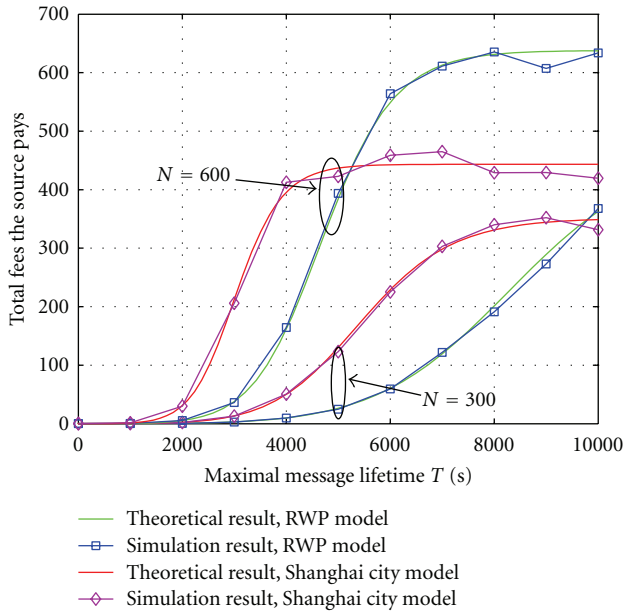


Figure 6: Simulation and numerical results comparison when the number of nodes is different.

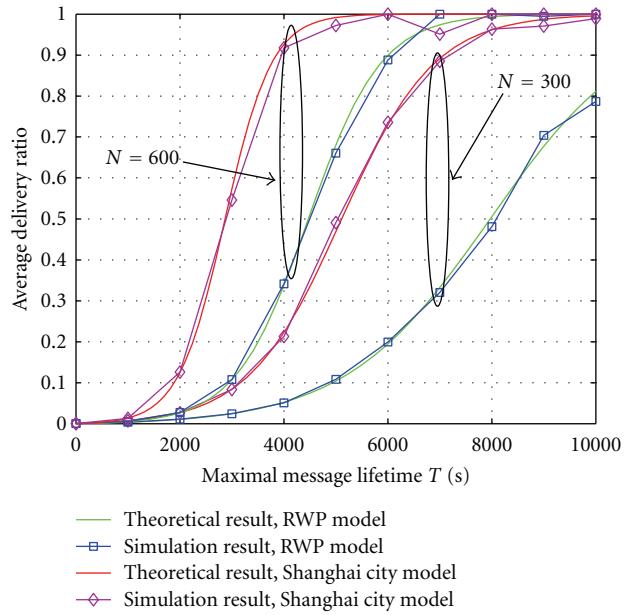


Figure 7: Performance comparison when the number of nodes is different.

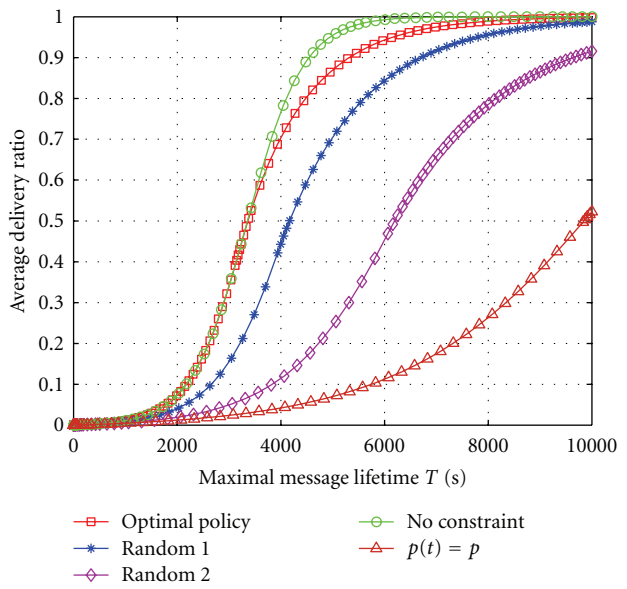


Figure 8: Performance comparison of different policies when the maximal message lifetime is different.

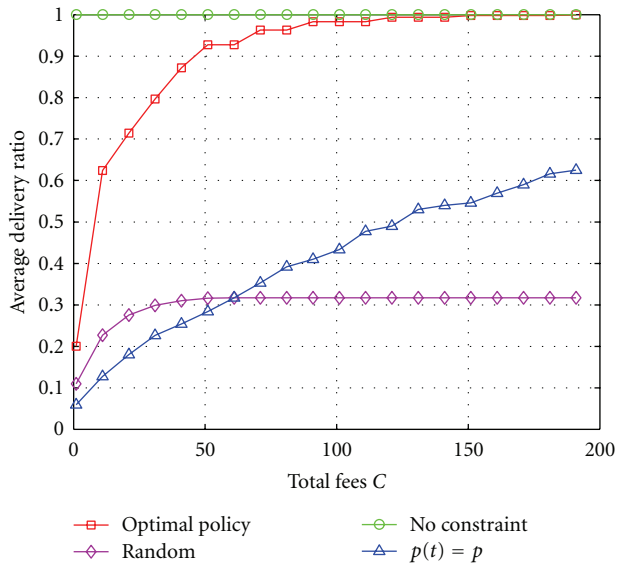


Figure 9: Performance comparison of different policies when the total fees are different.

First, we evaluate the performance of the optimal policy, which is the *threshold* form. For comparison, we consider three other cases. Case 1: $p(t) = p$, $t \geq 0$; Case 2: *No Constraint*; Case 3: *random*. Case 1 means that nodes forward message with the same probability all the time. The policy *No constraint* means that the total fees C is not limited, and $p(t)$ equals to 1 all the time. The *random* policy means that the value of $p(t)$ is randomly selected from the interval $[0, 1]$ at time t. In addition, we assume that total fees C equals to 150. Other settings are the same as that in simulation, and then we can obtain Figure 8. The result shows that the optimal policy is better than other policies when the total fees is limited. On the other hand, when total fees C equals to 150, the average delivery ratio is only a little smaller than that when the total fees is not limited.

Now, we let the maximal message lifetime T be 10000 s, and the value of C increase from 1 to 200. Other settings remain unchanged, and we can obtain Figure 9. This result shows that the optimal policy is best when the total fees are limited, too. On the other hand, when the total fees reach about 100, the average delivery ratio is about 1. This means that the source only needs to pay limited fees.

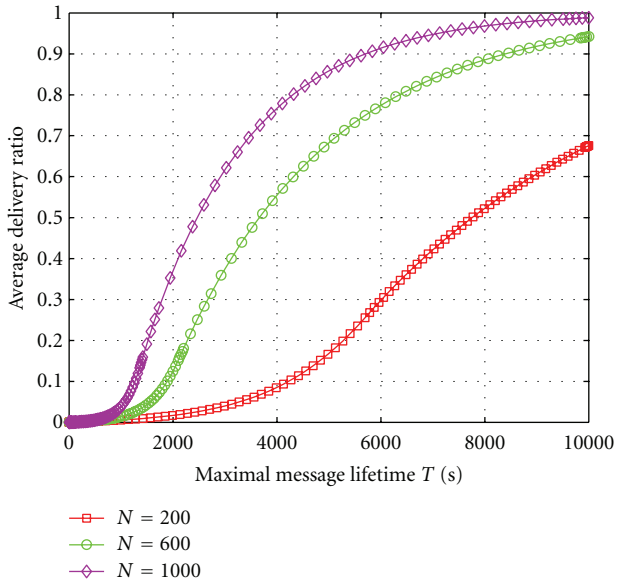


Figure 10: Average delivery ratio of the optimal policy when the value of N is different.

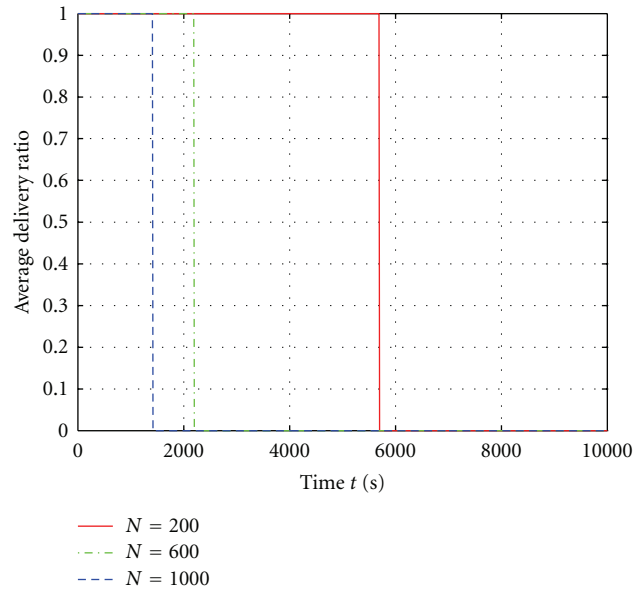


Figure 11: Optimal policy when the value of N is different.

The number of ordinary nodes may have certain impact on the routing performance. At present, we only consider the optimal policy obtained from (17). We let the value of C be 50, and the maximal message lifetime increase from 0 s to 10000 s. Other settings remain unchanged. Based on these settings, we can obtain Figure 10 when N equals to 200, 600, and 1000, respectively.

The result in Figure 10 shows that if the number of ordinary nodes is bigger, the performance is better. As shown above, the message transmission process can be seen as the commodities trading process. The event that the source requests help from other nodes can be seen as that the source buys certain goods from them. The price of the goods is increasing with time. Therefore, it is good for the source to buy more goods early. However, when the number of ordinary nodes is smaller, the goods are limited, so the source cannot buy many goods early. Therefore, if the number of ordinary nodes is bigger, the performance is better.

On the other hand, as shown in Theorem 1, the source will stop requesting help at certain time. In particular, the optimal policy satisfies: $p(t) = 1$, $t < h$, and $p(t) = 0$, $t > h$, $0 \le h \le T$. Therefore, the forwarding probability equals to 0 after time h. When the number of nodes is bigger, the value of h is smaller, so the source stops requesting help earlier. We can see the result in Figure 11 more clearly, when there are 200, 600, and 1000 nodes, respectively. For example, the value of h equals to about 2160s when there are 600 nodes.

In the above simulation and numerical results, we define: $\mathrm{PR}(t) = 3(1-e^{-t/10000})$. However, there may be many different forms for the function. Here, we study another special case, that is, we define: $\mathrm{PR}(t) = b^*(t+1)^y$, b, $y > 0$. Other settings are the same as that in above numerical result. However, we let the value of N be 600. The performance of the optimal policy is shown in Figure 12. This result shows that the performance is different when the parameters are different.

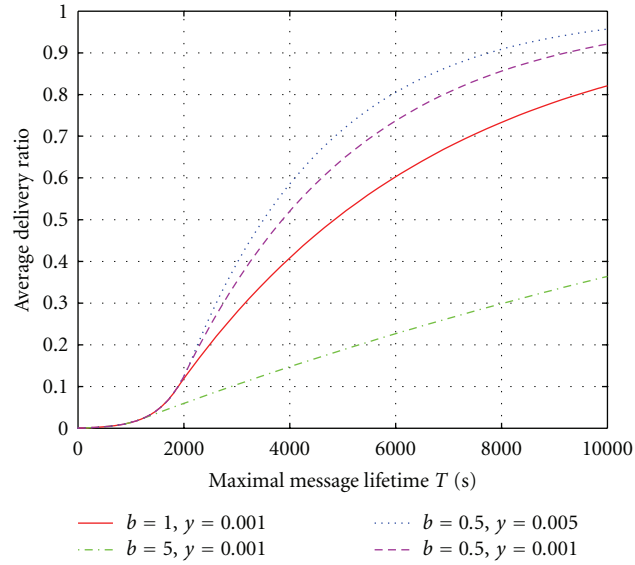


Figure 12: Optimal policy with different function.

6. Conclusions

Due to the mobility of nodes and many other factors, it is hard to maintain the connectivity in M2M networks. Therefore, the *store-carry-forward* communication mode is an important message propagation way. Such communication mode needs nodes working in a cooperative way. Due to the selfishness nature, nodes may ask for some fees (denoted by $\mathrm{PR}(t)$) from the source, which may be varying with time. For example, when the message stays in the network a long time, the ordinary nodes may think that its remaining lifetime is shorter, so they may ask for more fees from the source. In this paper, we propose a unifying framework to

evaluate the total fees that the source pays under different forwarding policies. Then based on the framework, we study the optimal control problem. In particular, we prove that the optimal forwarding policy conforms to the *threshold* form when $\mathrm{PR}(t)$ satisfies certain conditions. Simulations based on both synthetic and real motion traces show the accuracy of our theoretical framework. Numerical results show that the optimal forwarding policy obtained by (17) is the best one.

References

[1] Z. M. Fadlullah, M. M. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, and Y. Nozaki, "Toward intelligent machine-to-machine communications in smart grid," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 60–65, 2011.

[2] K. Fall, "A delay-tolerant network architecture for challenged internets," in *Proceedings of ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM '03)*, pp. 27–34, August 2003.

[3] G. Resta and P. Santi, "The effects of node cooperation level on routing performance in delay tolerant networks," in *Proceedings of the 6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON '09)*, June 2009.

[4] M. Karaliopoulos, "Assessing the vulnerability of DTN data relaying schemes to node selfishness," *IEEE Communications Letters*, vol. 13, no. 12, pp. 923–925, 2009.

[5] A. Vahdat and D. Becker, "Epidecmic routing for partially connected ad hoc networks," Tech. Rep. CS-200006, Duke University, Durham, NC, USA, 2000.

[6] L. Buttyan and J. P. Hubaux, "Enforcing service availability in Mobile Ad-Hoc WANs," in *Proceedings of the 1st ACM International Symposium on Mobile Ad Hoc Networking & Computing (MobiHoc '00)*, pp. 87–96, August 2000.

[7] E. Altman, G. Neglia, F. De Pellegrini, and D. Miorandi, "Decentralized stochastic control of delay tolerant networks," in *Proceedings of the 28th IEEE Conference on Computer Communications (INFOCOM '09)*, pp. 1134–1142, Rio de Janeiro, Brazil, April 2009.

[8] Y. Li, Y. Jiang, D. Jin, L. Su, L. Zeng, and D. Wu, "Energy-efficient optimal opportunistic forwarding for delay-tolerant networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 9, pp. 4500–4512, 2010.

[9] Q. Yuan, I. Cardei, and J. Wu, "An efficient prediction-based routing in disruption-tolerant networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 1, pp. 19–31, 2012.

[10] H. Samuel, W. Zhuang, and B. Preiss, "Improving the dominating-set routing over delay-tolerant mobile ad-hoc networks via estimating node intermeeting times," *Eurasip Journal on Wireless Communications and Networking*, vol. 2011, Article ID 402989, 2011.

[11] J. Wu and Y. Wang, "Social feature-based multi-path routing in delay tolerant networks," in *Proceedings of IEEE International Conference on Computer Communications (INFOCOM '12)*, pp. 1368–1376, March 2012.

[12] C. Liu and J. Wu, "An optimal probabilistic forwarding protocol in delay tolerant networks," in *Proceedings of the 10th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc '09)*, pp. 105–114, May 2009.

[13] M. Vojnović and A. Proutiere, "Hop limited flooding over dynamic networks," in *Proceedings of IEEE International Conference on Computer Communications (INFOCOM '11)*, pp. 685–693, Shanghai, China, April 2011.

[14] R. Ramanathan, R. Hansen, P. Basu, R. Rosales-Hain, and R. Krishnan, "Prioritized epidemic routing for opportunistic networks," in *Proceedings of the 1st International MobiSys Workshop on Mobile Opportunistic Networking*, pp. 62–66, June 2007.

[15] X. Zhang, G. Neglia, J. Kurose, and D. Towsley, "Performance modeling of epidemic routing," *Computer Networks*, vol. 51, no. 10, pp. 2867–2891, 2007.

[16] Y. K. Ip, W. C. Lau, and O. C. Yue, "Performance modeling of epidemic routing with heterogeneous node types," in *Proceedings of IEEE International Conference on Communications (ICC '08)*, pp. 219–224, May 2008.

[17] P. Jacquet, B. Mans, and G. Rodolakis, "Information propagation speed in mobile and delay tolerant networks," *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5001–5015, 2010.

[18] E. Baccelli, P. Jacquet, B. Mans, and G. Rodolakis, "Information propagation speed in bidirectional vehicular delay tolerant networks," in *Proceedings of IEEE International Conference on Computer Communications (INFOCOM '11)*, pp. 436–440, Shanghai, China, April 2011.

[19] A. A. Hanbali, P. Nain, and E. Altman, "Performance of ad hoc networks with two-hop relay routing and limited packet lifetime," in *Proceedings of the 1st International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS '06)*, October 2006.

[20] A. Jindal and K. Psounis, "Contention-aware performance analysis of mobility-assisted routing," *IEEE Transactions on Mobile Computing*, vol. 8, no. 2, pp. 145–161, 2009.

[21] A. Panagakis, A. Vaios, and I. Stavrakakis, "On the effects of cooperation in DTNs," in *Proceedings of the 2nd International Conference on Communication System Software and Middleware and Workshops (COMSWARE '07)*, January 2007.

[22] Q. Li, S. Zhu, and G. Cao, "Routing in socially selfish delay tolerant networks," in *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM '10)*, March 2010.

[23] Y. Li, P. Hui, D. Jin, L. Su, and L. Zeng, "Evaluating the impact of social selfishness on the epidemic routing in delay tolerant networks," *IEEE Communications Letters*, vol. 14, no. 11, pp. 1026–1028, 2010.

[24] Y. Li, G. Su, D. O. Wu, D. Jin, L. Su, and L. Zeng, "The impact of node selfishness on multicasting in delay tolerant networks," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 5, pp. 2224–2238, 2011.

[25] E. Altman, A. P. Azad, T. Başar, and F. De Pellegrini, "Optimal activation and transmission control in delay tolerant networks," in *Proceedings of IEEE International Conference on Computer Communications (INFOCOM '10)*, March 2010.

[26] Y. Li, Z. Wang, D. Jin, L. Su, L. Zeng, and S. Chen, "Optimal beaconing control for epidemic routing in delay-tolerant networks," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 1, pp. 311–320, 2012.

[27] C. Singh, A. Kumar, and R. Sundaresan, "Delay and energy optimal two-hop relaying in delay tolerant networks," in *Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt '10)*, pp. 256–265, June 2010.

[28] M. H. R. Khouzani, S. Sarkar, and E. Altman, "Optimal control of epidemic evolution," in *Proceedings of IEEE International Conference on Computer Communications (INFOCOM '11)*, pp. 1683–1691, April 2011.

[29] M. H. R. Khouzani, S. S. Venkatesh, and S. Sarkar, "Market based control of epidemics," in *Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing*, pp. 314–320, Monticello, Ill, USA, September 2011.

[30] T. Karagiannis, J. Y. Le Boudec, and M. Vojnovic, "Power law and exponential decay of inter contact times between mobile devices," in *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking (MobiCom '07)*, pp. 183–194, September 2007.

[31] H. Cai and D. Y. Eun, "Crossing over the bounded domain: from exponential to power-law intermeeting time in mobile ad hoc networks," *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, pp. 1578–1591, 2009.

[32] W. Gao, Q. Li, B. Zhao, and G. Cao, "Multicasting in delay tolerant networks: a social network perspective," in *Proceedings of the 10th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc '09)*, pp. 299–308, May 2009.

[33] H. Zhu, L. Fu, G. Xue, Y. Zhu, M. Li, and L. M. Ni, "Recognizing exponential inter-contact time in VANETs," in *Proceedings of IEEE International Conference on Computer Communications (INFOCOM '10)*, pp. 1–5, March 2010.

[34] D. Grass, J. Caulkins, G. Feichtinger, G. Tragler, and D. Behrens, *Optimal Control of Nonlinear Processes: With Applications in Drugs, Corruption, and Terror*, Springer, 2008.

[35] A. Keranen, J. Ott, and T. Karkkainen, "The ONE simulator for DTN protocol evaluation," in *Proceedings of the 2nd International Conference on Simulation Tools and Techniques (Simutools '09)*, vol. 55, March 2009.

[36] C. Bettstetter and C. Wagner, "The spatial node distribution of the random waypoint mobility model," in *Proceedings of the Deutscher Workshop über Mobile Ad-Hoc Netzwerke (WMAN '02)*, pp. 41–58, March 2002.

[37] Shanghai Taxi Trace Data, http://wirelesslab.sjtu.edu.cn/.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 134238, 8 pages
doi:10.1155/2012/134238

Research Article

Adaptive Message Rate Control of Infrastructured DSRC Vehicle Networks for Coexisting Road Safety and Non-Safety Applications

Wenyang Guan,[1] Jianhua He,[2] Chao Ma,[1] Zuoyin Tang,[2] and Yue Li[1]

[1] *College of Engineering, Swansea University, Swansea SA2 8PP, UK*
[2] *School of Engineering and Applied Science, Aston University, Birmingham B4 7ET, UK*

Correspondence should be addressed to Chao Ma, mac1@aston.ac.uk

Received 22 April 2012; Revised 20 June 2012; Accepted 25 June 2012

Academic Editor: Lin Bai

Intelligent transport system (ITS) has large potentials on road safety applications as well as nonsafety applications. One of the big challenges for ITS is on the reliable and cost-effective vehicle communications due to the large quantity of vehicles, high mobility, and bursty traffic from the safety and non-safety applications. In this paper, we investigate the use of dedicated short-range communications (DSRC) for coexisting safety and non-safety applications over infrastructured vehicle networks. The main objective of this work is to improve the scalability of communications for vehicles networks, ensure QoS for safety applications, and leave as much as possible bandwidth for non-safety applications. A two-level adaptive control scheme is proposed to find appropriate message rate and control channel interval for safety applications. Simulation results demonstrated that this adaptive method outperforms the fixed control method under varying number of vehicles.

1. Introduction

Intelligent transport system (ITS) has received wide interests since the last decade due to its huge potentials on traffic safety applications, business logistics, route planning, entertainment, and many other applications. However, one of the big challenges for ITS is on the vehicle machine to machine communications. Due to the large quantity of vehicles, high mobility, and bursty traffic from the safety and nonsafety applications, the traditional cellular networks can not provide cost effective and real-time communications for large-scale ITS applications, especially for safety applications. On the other hand, among the broad ITS applications, road traffic safety applications have been a subject of worldwide concern. It has been extensively studied to actively prevent accidents or passively minimize the consequences of accidents. Driven by advances in wireless communications and mobile networking, collaborative safety applications (CSAs) enabled by vehicular communications are widely considered to be key for the success of future road safety [1]. Vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) communications can enable exchange of vehicle information and proactive warning of potential hazards for collaborative safe application, such as emergency stops, merging traffic, vehicles in a driver's blind spot, imminent collision, and driving assistant messages that related to safety driving.

Among direct V2V communication technologies, dedicated short-range communications (DSRC) is a strong candidate for CSA. Compared to cellular networks, it can provide very high data transfer rates at low cost in circumstances where minimal communication latency and isolated relatively small communication zones are important. DSRC technology is robust and can be built into large-scale vehicles [2–4]. The US Federal Communications Commission (FCC) has allocated 75 MHz of spectrum in the 5.9 GHz band for DSRC [5]. DSRC standards are currently being developed by organizations including the IEEE [6] and the Society of Automobile Engineers (SAE). IEEE is specifying a wireless access in vehicular environment (WAVE) for DSRC to provide seamless, interoperable V2V, and vehicle to roadside unit (RSU) (V2R) communication services [6]. SAE is defining a standard message set and data dictionary for DSRC-based vehicle safety applications. The US National Highway Traffic Safety Administration (NHTSA) has undertaken several projects to test vehicle safety applications performance by simulation and field experiments.

To reduce system costs, both road safety applications and non-safety applications are likely to be deployed over multiple DSRC channels. According to the specified multichannel operations, time-division multiplexing is used for the DSRC devices to monitor the control channel (CCH) for safety information and service channels (SCH) for non-safety applications. All the DSRC devices need to monitor CCH at the CCH intervals. One big challenge for the coexisting safety and non-safety application is how to effectively ensure QoS for safety applications while leave as much as bandwidth for non-safety applications. Safety applications have higher priority and they have stringent requirements for reliable real-time message delivery, as excessive message delays or message loss hinders the effectiveness of CSA and can even cause unexpected negative consequences. However, the QoS requirements are hard to be met by the random channel access specified in the IEEE 802.11 DCF [7–10]. On the other hand, the non-safety applications should not get as much bandwidth as possible to provide efficient non-safety services. For the safety applications, their QoS perceived are affected by a wide range of factors, such as resource provisioning and congestion control. In this paper, we consider two major types of safety applications: event-driven safety applications (ESA) and periodic safety applications (PSA). Their QoS can be differentiated by channel access schemes and message rate control schemes. ESA is designed to be used for emergency scenarios. It creates and broadcast messages if accidents happened or are emerging. PSA is designed for announcing existence of a vehicle and broadcast non-emergent messages. PSA messages are periodically generated and broadcasted to help build mutual awareness and implement some simple CSAs [1]. Compared to PSA messages, ESA messages have higher priority to inform or make a caution to the following vehicles with global positioning system (GPS) information included.

With the challenges on the development of coexisting safety and non-safety applications over DSRC-based vehicle networks, it is important to improve the utilization of the limited spectrum resources for DSRC networks, while meeting the QoS requirements for the road safety applications. In this paper, we propose an adaptive control scheme to avoid network congestion and provide good QoS for safety applications. The objectives are to provide high-availability and low-latency channel for high-priority, ESA messages and maximize channel utilization for low priority PSA messages and non-safety applications. To facilitate the adaptive control of the DSRC networks, we use an off-line simulation based approach to find out the best possible configurations of CCH interval, safety message rate, and channel access parameters for given combinations of safety QoS requirements and the number of vehicles. Here we assume each vehicle in the network is equipped with a DSRC radio. A utility function is proposed to take the QoS requirements of safety applications into account and solve the multiple objectives optimization problem for the coexisting safety and non-safety applications. The identified configurations are then adaptively used online by a roadside access point (AP) for both CCH interval control and channel access control. We focus on the broadcast-based safety applications in this paper.

In the literature, Wang and Hassan [11] investigated the impact of CCH interval on the QoS of single safety application and channel availability for non-safety applications. However, the service differentiated channel access and congestion control are not considered in [11]. The authors have studied adaptive message rate control for DSRC vehicle networks, in which safety message rate is controlled in a distributed manner by the vehicles in freeways [10]. In this paper, the focus is on a road intersection where an access pointer (AP) is deployed for centralized network control. A distributed message rate control method for ad hoc vehicle networks is proposed for single safety application in [12]. A centralized message rate control in road intersections is studied for two differentiated safety application in [13]. However, it is noted that all the above works have not considered the impact and adaptive configuration of CCH interval.

The remaining of this paper is organized as follows. We briefly introduce the background knowledge on DSRC and 802.11p Standard in Section 2. Section 3 presents the design of the adaptive congestion control method. Numerical results are presented in Section 4. Finally, we make a conclusion in Section 5.

2. Background

2.1. DSRC Standard Activities. For economic concerns, DSRC is expected to provide both road safety and commercial services. The overall WAVE architecture developed by IEEE for DSRC includes IEEE 802.11p (MAC and PHY standards) and IEEE Std 1609.1 to 1609.4. At the MAC layer, IEEE 802.11p is based on IEEE 802.11e, which has been augmented with QoS support. IEEE 802.11e can provide multiple priorities to different applications by differentiating DCF-based channel access parameters [7]. At the physical layer, 802.11p is the same as 802.11a except that 802.11p is operated with 10 MHz bandwidth instead of 20 MHz for 802.11a. More details on the 802.11 channel access schemes is referred to in [7].

Multichannel operation is specified in IEEE Std 1609.4. In the multichannel framework, a control channel (CCH) is to be used exclusively for road safety messages and service announcements, while the other channels are service channels (SCH). It is required that all WAVE devices need to monitor CCH at regular intervals. To account for the devices that can not simultaneously monitor CCH and SCH, synchronization procedure has been proposed to coordinate the channel using time-division multiplexing [14]. A synchronization interval comprises a CCH interval, a SCH interval, and two guard intervals.

2.2. Channel Access in 802.11p Standard. For each Access Category (AC), an enhanced distributed coordination access (EDCA) process will be started to contend for transmission opportunities (TXOPs) using a set of distinct EDCA parameters, including arbitration interframe space (AIFS) instead of DIFS in DCF. AIFS(AC) is determined by AIFS(AC) = SIFS + AIFSN(AC), where AIFSN(AC) is an integer indication of the number of slots that a station belonging to AC should defer before either invoking a backoff or starting a

Table 1: Default EDCA parameter set.

AC	Example	CWmin	AIFSN
AC0	BK	a CWmin	9
AC1	Best effort (BE)	a CWmin	6
AC2	Video	(a CWmin + 1)/2 − 1	3
AC3	Voice	(a CWmin + 1)/4 − 1	2

transmission after a SIFS duration. AC values of 0, 1, 2, and 3 present background, best effort, video, and voice, as shown in Table 1, respectively.

3. Adaptive Control Scheme

In this section, we present a two-level adaptive control scheme for the coexisting safety and non-safety applications. For the safety applications, we consider both emergency and routine safety applications. We take a road intersection as an example network scenario where a fixed roadside AP has the full control of setting for the CCH interval and other system parameters. In the first level, the time allocated to the CCH and the SCH is controlled and adapted according to traffic loads in a relatively long-time scale. In the second level, adaptive congestion control is applied to the CCH in a relatively short-time scale. The objectives of the design are to ensure QoS of high priority ESA messages while maximizing channel utilization for low priority PSA messages and non-safety applications. The reason to maximize channel utilization for low priority PSA messages is that low priority PSA applications which coexist with ESA application over the DSRC control channel are also important for CSA. For example, periodically broadcasted PSA messages which include vehicle positions enable mutual awareness.

There are two major parts included in the adaptive control scheme. The first part is an offline procedure to find out the optimal configurations of CCH interval, message rate, and backoff exponent (BE) for a set of QoS requirements and given number of vehicles. These optimal configurations are then applied in the second part where the roadside AP requests the vehicles to update the configurations according to the QoS requirements and an estimated number of vehicles in the road intersection.

It is noted that in addition to the control of CCH interval and message rate, a MAC layer blocking mechanism is used by all the vehicles for safety applications [10]. The MAC layer blocking mechanism is used to immediately block low priority PSA messages by a vehicle if it detects that the channel is busy for longer than a channel busy threshold in any CCH interval. The proposed adaptive control scheme is implemented in a centralized manner. This is different from traditional network congestion control protocols such as TCP and TFRC protocols, which control only the packet transmission rate and are implemented in a distributed approach at the transport layer. The proposed method is also different from distributed rate adaptation method proposed in [10] as the AP can fully control the system configurations for the vehicles in the road intersection.

3.1. Offline Determination of Optimal Configurations. To facilitate the adaptive control, we use an offline simulation-based approach to find out the best possible configurations of CCH interval, safety message rate, and channel access parameters for given combinations of and safety QoS requirements (e.g., message successful probability and message delivery delay) and the number of vehicles.

Here offline simulation approach means determination of optimal configurations by simulation of a system that is not in operation, which is contrast to the approach that may be used to adaptively find the proper configurations from the real system in operation. A simulator is developed for this purpose. Although it is possible to use analytical models to determine the optimal configurations, we believe the analytical models may not be efficient to take into account the complex system operations and parameters, such as unsaturated traffic load, MAC layer backoff, and blocking.

A challenge on the determination of optimal configuration is the multiple objectives optimization for the whole vehicle network, namely, provisioning of high available channel for ESA messages and high channel utilization for PSA messages, and leaving more channel time to non-safety applications. For example, a low PSA message rate will present higher channel availability to ESA messages but at the cost of less transmitted PSA messages. To tackle the multiple objective optimization problem, we use a utility function to find out the combination of BE, message rate, and the minimal CCH interval which can satisfied the specific QoS. In the proposed utility function, the performance metrics of message success probability, average transmit delay, and transmit rate are taken into account. It is noted that there could be alternative utility functions defined for the multiple objective optimization problem. Investigation of alternative utility functions is left for our future work.

Let P_e and P_p denote message success probability for ESA messages and PSA messages, respectively. Let D_e and D_p denote message delivery delay for ESA messages and PSA messages, respectively. Let R_e and R_p denote the average number of successfully received ESA messages and PSA messages by one vehicle in one second, respectively. We can have the following proposed utility function (denoted by Θ):

$$\begin{aligned}\Theta = {} & R_e + \left(P_{s,e} - P_{\text{thr},e}\right)^+ + \left(D_{\text{thr},e} - D_e\right)^+ \\ & + R_p + \left(P_{s,p} - P_{\text{thr},p}\right)^+ + \left(D_{\text{thr},p} - D_p\right)^+,\end{aligned} \tag{1}$$

where $P_{\text{thr},e}$ and $P_{\text{thr},p}$ are preset thresholds for the message success probability of ESA and PSA messages, respectively. $D_{\text{thr},e}$ and $D_{\text{thr},p}$ are preset thresholds for delivery delay of ESA and PSA messages, respectively.

The threshold function $(x)^+$ used in the utility function is expressed by

$$\begin{aligned}(x)^+ &= 0, \quad \text{if } x \le 0, \\ (x)^+ &= 1, \quad \text{if } x > 0.\end{aligned} \tag{2}$$

The reason that the threshold function is used is that for some given QoS requirements on the ESA and PSA applications, we want to maintain a message success

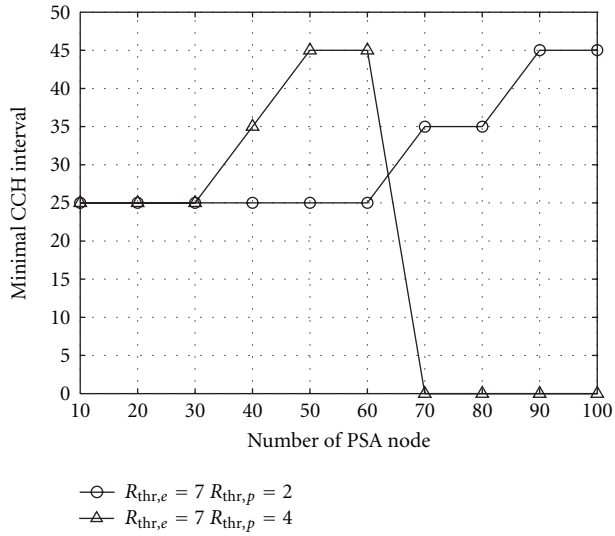


Figure 1: Minimal CCH interval satisfying the preset QoS requirements with $R_{thr,p} = 2$ and 4.

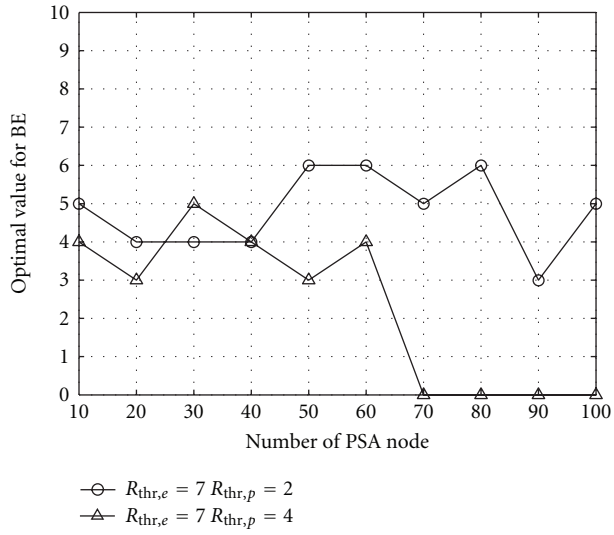


Figure 2: Optimal BE against the number of PSA vehicles with $R_{thr,p} = 2$ and 4.

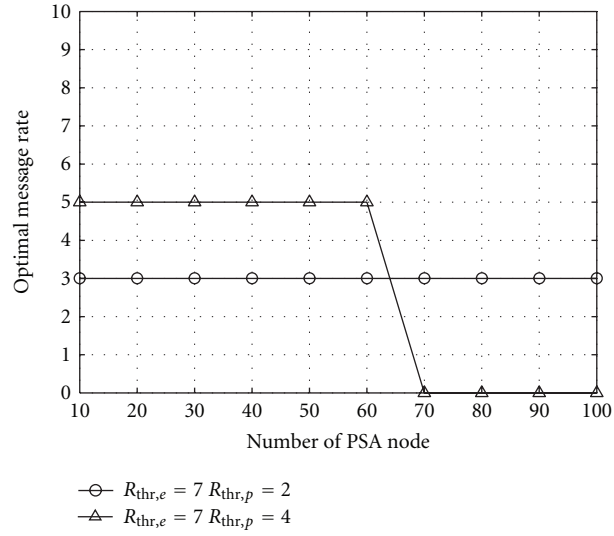


Figure 3: Optimal message rate against the number of PSA vehicles with $R_{thr,p} = 2$ and 4.

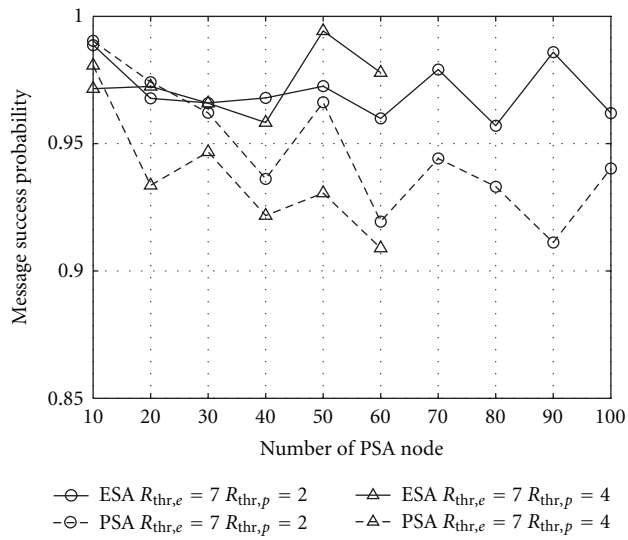


Figure 4: Message success probability against the number of PSA vehicles with $R_{thr,p} = 2$ and 4.

probability larger than the message success threshold, a message delivery delay smaller than the delay thresholds and an average transmit rate larger than the rate threshold. If the requirement on one service metric (message success probability or delivery delay) is not satisfied, value Θ will be set to 0 which means that rate of R_p is unusable from that specific service metric.

With the preset parameters, we can determine a configuration table which gives the minimal CCH interval and the optimal configuration of message rate, and BE which meet given QoS requirements with various number of vehicles in the network. Note that the determination of the optimal configurations is only needed at the roadside AP.

3.2. Online Adaptation of Configurations. In this procedure, the AP applies the findings from the offline procedure on the minimal CCH interval and the optimal configurations of message rate for PSA applications. The procedure operates as follows. Firstly, the AP estimates the number of vehicles (N_{est}) at the road intersection for every T_{est} seconds through the received PSA messages, which are broadcasted by the vehicles at the road intersection. In this paper, we set $T_{est} = 60$. According to the estimated number of vehicles and the preset QoS requirements, the AP looks up the configuration table to get the minimal CCH interval and the optimal configurations of message rate and BE. The value of CCH interval and the optimal configuration is then broadcasted

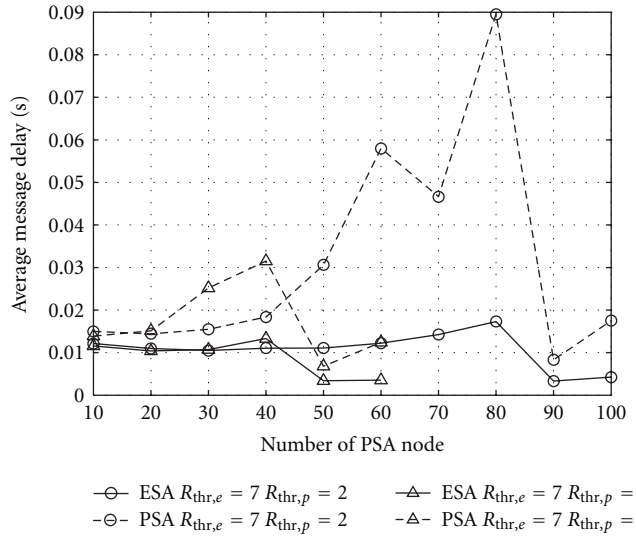


Figure 5: Message delivery delay against the number of PSA vehicles with $R_{thr,p} = 2$ and 4.

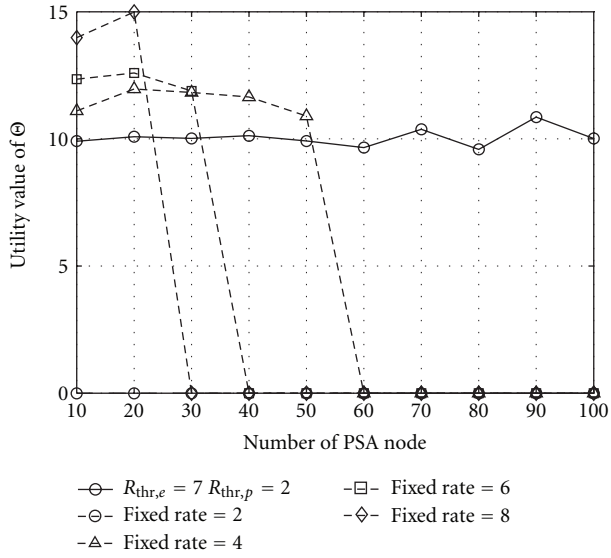


Figure 6: Utility Θ of adaptive control and Type I FCS for $R_{thr,p} = 2$.

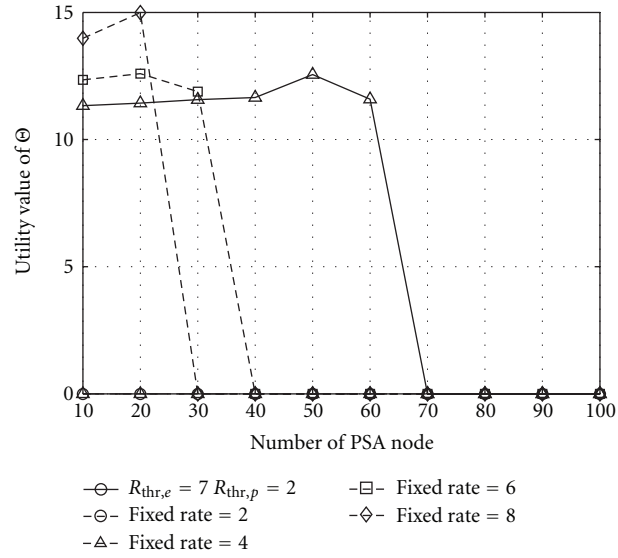


Figure 7: Utility Θ of adaptive control and Type I FCS for $R_{thr,p} = 4$.

in ESA messages by the AP at different time scales (every TS_{CCH} seconds for CCH interval broadcast and every TS_{rate} for message rate broadcast) to the vehicles at the intersection. In default, TS_{CCH} is set to 300 seconds and TS_{rate} is set to 30 seconds. Vehicles received the AP configuration instructions update their configurations accordingly. In addition, the AP keeps monitoring the safety applications QoS performances during the system operations. If the perceived QoS performances are better than the required, the CCH interval is increased with a step of 10 ms to improve the QoS performances for safety applications. In reverse the CCH interval is reduced with the same step if the QoS performances are poor than the QoS requirements.

4. Numerical Result

We have built a discrete event-driven simulator to evaluate the performance of the adaptive control scheme for DSRC vehicle networks. All vehicles are located with uniform distribution along the roads at a junction and a roadside; AP is located at the center of the junction. We assume a single hop ad hoc network in which each vehicle can hear transmissions from other vehicles. For simplicity, we assume there are two classes of vehicles in the network. The first class of vehicles transmit only ESA messages while the second class of vehicles transmit only PSA messages. Message block event at MAC layer is triggered to provide high available bandwidth for ESA messages with a MAC blocking threshold of 70%

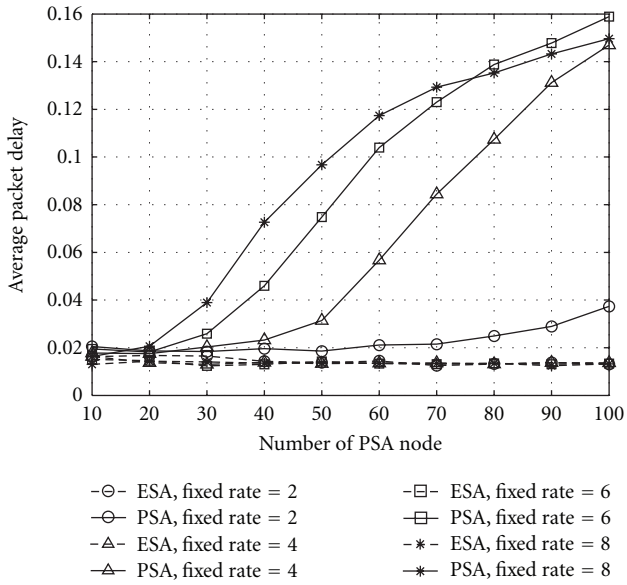


Figure 8: Message delay of Type I FCS.

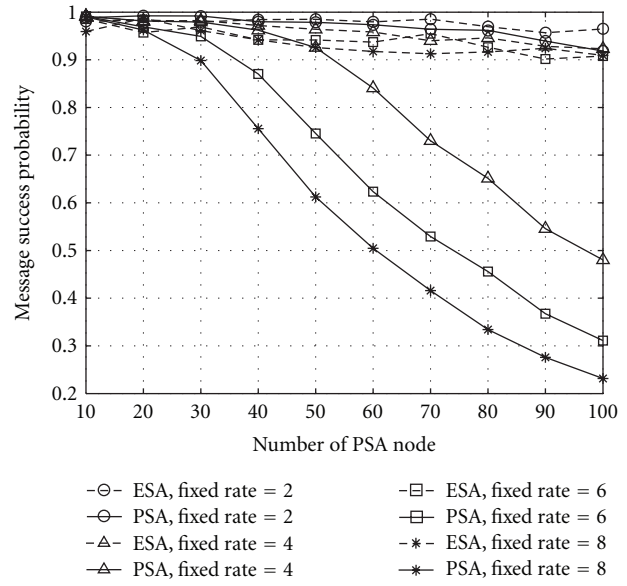


Figure 9: Message success probability of Type I FCS.

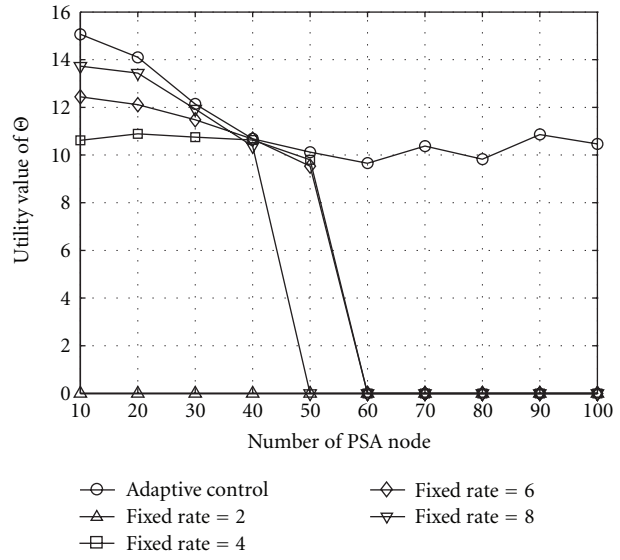


Figure 10: Utility Θ of adaptive control and Type II FCS for $R_{\text{thr},p} = 2$.

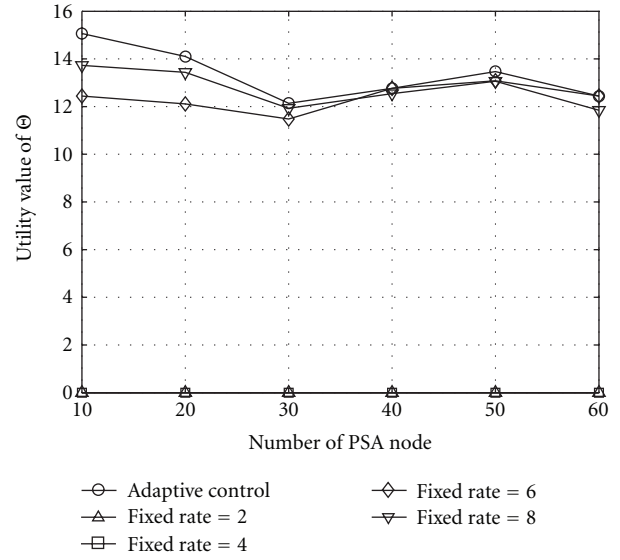


Figure 11: Utility Θ of adaptive control and Type II FCS for $R_{\text{thr},p} = 4$.

in a single SI. Performance for MAC blocking thresholds of 50% and 90% is also investigated in the simulations. For simplicity, we assume that there are three first-class vehicles, which periodically send eight ESA messages per second. All safety messages have the same length of 250 bytes and are broadcasted at the rate of 3 Mbps. An ideal channel is assumed where a message can be successfully received if no collision happens.

We have used the following configurations for the thresholds in the proposed utility function: $R_e = 7$, R_p ranging from 2 to 8, $P_{\text{thr},e} = 0.9$, $P_{\text{thr},p} = 0.9$, $D_{\text{thr},e} = 0.02$, and $D_{\text{thr},p} = 0.1$. The CCH interval is configurable in the set [15, 25, 35, 45] ms, which corresponds to 30%, 50%, 70%, and 90% of a 50 ms synchronization interval (SI), respectively.

With the above parameter configurations, we obtained the optimal configuration of message rate, BE, and CCH interval length. The minimal CCH interval length satisfying the preset QoS requirements and the corresponding optimal configurations of BE and message rate is plotted against the number of vehicles in Figures 1, 2, and 3, respectively. The preset QoS requirements are with message success probability $P_{\text{thr},p} = 0.9$ and message rate $R_{\text{thr},e} = 7$, $R_{\text{thr},p} = 2$ and 4.

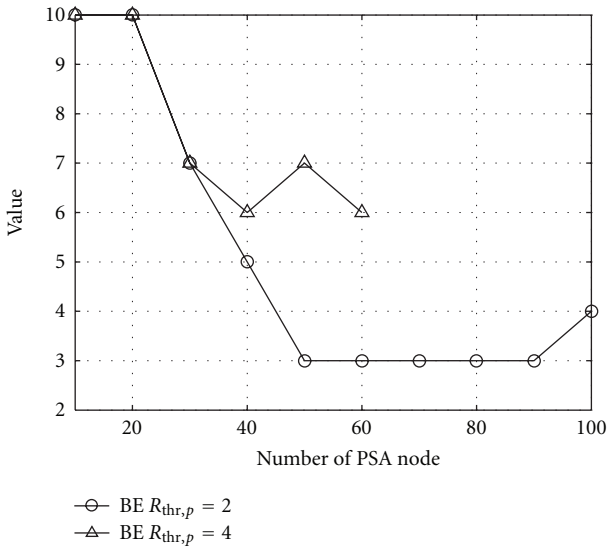


Figure 12: Optimal message rate of adaptive control scheme.

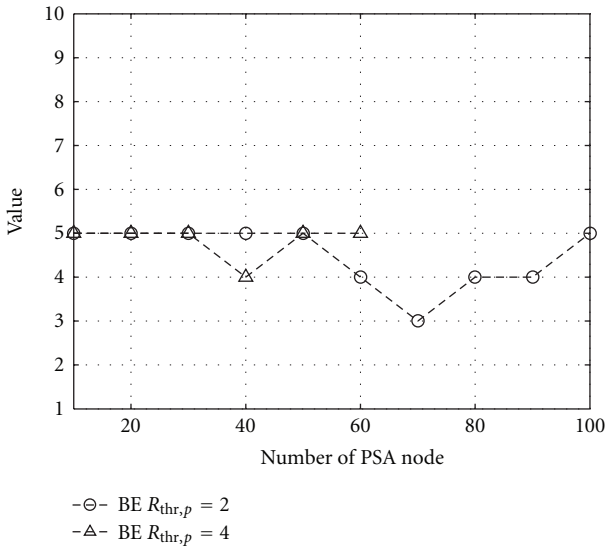


Figure 13: Optimal BE of adaptive control scheme.

It can be observed from Figure 3 that message rate must be at least 5 messages per second to satisfy the QoS requirement $R_{\text{thr},p} = 4$. However, the minimal CCH interval length is increased to 45 ms with 60 vehicles and $R_{\text{thr},p} = 4$. With more than 60 vehicles in the network, none combination of CCH interval and message rate can satisfy the QoS requirement. In these cases, message rate, BE, and CCH interval length are plotted as 0 in the figures.

Figures 4 and 5 present the performances of message success probability and delivery delay under the QoS requirement of $R_{\text{thr},p} = 2$, and 4. It can be seen that the combination of BE, rate, and CCH interval length selected by the offline procedure performs well in QoS provisioning, that is, ESA messages delay is lower than 20 ms, message success probability for both ESA and PSA is over than 90% as required.

We plot the utility value Θ for QoS requirement of $R_{\text{thr},p} = 2$ and 4 in Figures 6 and 7. For performance comparison, we plot the results obtained for a type fixed control scheme (called Type I FCS) in which the CCH interval and message rate are fixed irrespective of the dynamic traffic loads. For the Type I FCS we set BE to 4 and CCH interval length to 35 ms. Several fixed message rates are selected for comparison. It is observed from the figures that the adaptive control scheme can achieve larger utility than the Type I FCS in most of the cases. And more importantly the adaptive control scheme can use much smaller CCH interval length to satisfy safety applications QoS, which means non-safety applications are left with more channel time. In Figures 8 and 9, we plot the corresponding average message delay and success probability of ESA and PSA messages for Type I FCS.

Next we compare the performance of the adaptive control scheme with that of another type of fixed control scheme (called Type II FCS), which uses the CCH interval length identified in the adaptive control scheme but uses fixed message rate for PSA. The utility values of the adaptive control scheme and the Type II FCS are plotted in Figures 10 and 11 for the QoS requirements of $R_{\text{thr},p} = 2$ and 4, respectively. We can see from Figures 10 and 11 that although Type II FCS can leave the same amount of channel time to non-safety applications as the adaptive control scheme, it has much smaller utility values than the adaptive control scheme due to the use of fixed message rates in the FCS. The optimal message rate and BE for PSA are plotted in Figures 12 and 13, respectively.

It can be observed from Figure 12 that with less vehicles in the network, message rate can reach as high as 10, and still meet the QoS requirement, and BE maintains stable relatively. With the number of vehicle increasing, message rate is as low as 3, and BE becomes unstable when $R_{\text{thr},p} = 2$, while none of any combinations can meet the QoS requirement when $R_{\text{thr},p} = 4$.

5. Conclusion

In this paper, we investigated a system control issue faced by the coexisting safety and non-safety application deployed over DSRC vehicle networks. A two-levels adaptive control scheme was proposed with one level on CCH interval control and the other level on message rate and channel access control. The objective is to ensure QoS requirements for safety applications while leaving as much bandwidth as possible for non-safety applications. An offline procedure is used to determine the optimal configurations of CCH interval, safety message rate, and channel access parameters. A utility function is proposed to solve the multiobjectives optimization problem and take the safety application QoS into account. The identified configurations are applied online by the roadside AP according to the estimated number of vehicles. Results demonstrate that the adaptive control scheme significantly improves system performances over the fixed control scheme with changing number of vehicles and QoS requirements from the road safety applications.

Acknowledgments

The work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) with Grant Reference no. *EP/1010157/1* and the National Natural Science Foundation of China (NSFC) under the Grant no. 61103177.

References

[1] Vehicle Safety Communications Project, "Vehicle safety communications project," Final Report, CAMP IVI Light Vehicle Enabling Research Program DOT HS 810 591, 2006.

[2] J. Zhu and S. Roy, "MAC for dedicated short range communications in intelligent transport system," *IEEE Communications Magazine*, vol. 41, no. 12, pp. 60–67, 2003.

[3] S. Biswas, R. Tatchikou, and F. Dion, "Vehicle-to-vehicle wireless communication protocols for enhancing highway traffic safety," *IEEE Communications Magazine*, vol. 44, no. 1, pp. 74–82, 2006.

[4] D. Jiang, V. Taliwal, A. Meier, W. Holfelder, and R. Herrtwich, "Design of 5.9 GHz DSRC-based vehicular safety communication," *IEEE Wireless Communications*, vol. 13, no. 5, pp. 36–43, 2006.

[5] Federal Communications Commission, FCC Report and Order FCC 03-324, 2004.

[6] IEEE 802, *Amendment For Wireless Access For the Vehicular Environment (WAVE)*, 2010.

[7] IEEE Std 802, *Wireless LAN Medium Access Control (MAC) enhancements for Quality of Service (QoS)*, 2005.

[8] J. Yin, T. Elbatt, G. Yeung et al., "Performance evaluation of safety applications over DSRC vehicular ad hoc networks," in *Proceedings of the 1st ACM VANET International Workshop on Vehicular Ad Hoc Networks*, pp. 1–9, October 2004.

[9] J. He, Z. Tang, T. O'Farrell, and T. M. Chen, "Performance analysis of DSRC priority mechanism for road safety applications in vehicular networks," *Wireless Communications and Mobile Computing*, vol. 11, no. 7, pp. 980–990, 2011.

[10] J. He, H. H. Chen, T. M. Chen, and W. Cheng, "Adaptive congestion control for DSRC vehicle networks," *IEEE Communications Letters*, vol. 14, no. 2, pp. 127–129, 2010.

[11] Z. Wang and M. Hassan, "How much of dsrc is available for non-safety use?" in *Proceedings of the 5th ACM VANET International Workshop on VehiculAr Inter-NETworking*, pp. 23–29, September 2008.

[12] M. Torrent-Moreno, J. Mittag, P. Santi, and H. Hartenstein, "Vehicle-to-vehicle communication: fair transmit power control for safety-critical information," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 7, pp. 3684–3703, 2009.

[13] W. Guan and J. He, "Adaptive congestion control of DSRC vehicle networks for collaborative road safety applications," in *IEEE Workshop on Wireless Local Area Networks (WLN '11)*, 2011.

[14] IEEE Std 1609, *IEEE Trial-Use Standard For Wireless Access in Vehicular Environments (WAVE) Multi-Channel Operation*, 2006.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 213727, 9 pages
doi:10.1155/2012/213727

Research Article

Distributed Beamforming for Relay Assisted Multiuser Machine-to-Machine Networks

Chen Chen,[1] Lin Bai,[2] Meiping Feng,[1] Minhua Huang,[1] and Tian Tian[3]

[1] *State Key Laboratory of Advanced Optical Communication Systems and Networks, Peking University, Beijing 100871, China*
[2] *School of Electronic and Information Engineering, Beihang University, Beijing 100191, China*
[3] *Strategic Investment Department, China Unicom, Beijing 100033, China*

Correspondence should be addressed to Lin Bai, l.bai@buaa.edu.cn and Meiping Feng, feng.meiping@pku.edu.cn

Received 20 April 2012; Revised 5 July 2012; Accepted 9 July 2012

Academic Editor: Jianhua He

We consider signal transmission with the aid of multiple half-duplex single-antenna relay nodes using the amplify-and-forward (AF) strategy for a multiuser wireless machine-to-machine (M2M) communication system. In such a scenario, relay beamforming has been proved to be an effective way to improve system performance by employing spatial diversity gains. Early works have mostly focused on various centralized relay beamforming schemes based on global channel state information (CSI). Since global CSI is often unavailable due to geometric locations, power limitation, or other constraints of the relay nodes in M2M networks, our work aims to develop distributed algorithms that each relay node individually learns its own beamforming weights with local CSI. We propose two suboptimal relay beamforming schemes that only require local CSI to minimize mean square error (MSE) for all the users with nonorthogonal channels. For multiuser systems with orthogonal channels, we divide the optimization problem into multiple single user problems which then can be solved by each relay independently. Numerical simulations for the proposed algorithms are presented showing the performance of the proposed schemes is close to that of the optimal scheme with global CSI in terms of bit error rate (BER) criterion.

1. Introduction

With the ability to provide reliable transmission and improve the coverage and capacity of wireless networks, cooperative communication using relay beamforming has attracted considerable attention [1]. Multiple cooperative relays can act as a virtual array of transmit antennas to provide spatial multiplexing and hence gain diversity for users. Such kind of relay techniques can be employed in many different communication scenarios. For instance, relays can be placed at the edge of a cell to extend the coverage of the network or at the intersection of adjacent cells for interference coordination [2]. Wireless machine-to-machine (M2M) sensor networks are another possible applications. In these networks, signals can be relayed by one or more relay sensor nodes while they might be attenuated beyond detection by propagation loss if directly targeting the destination sensor nodes, due to the low-power machine devices [3]. Also, the spatial diversity of distributed relays can help to mitigate the effect of fading channels.

Typically, communications with the aid of relay beamforming are implemented in two phases where the relays work in a half-duplex mode. In the first phase, signals are broadcast from source nodes to relay nodes. In the second phase, the relay nodes process the received signals according to relay beamforming strategies and then forward the processed messages to destination nodes. Orthogonal or nonorthogonal transmissions can be used in cooperative communication. In this paper, transmission systems over orthogonal channels are referred to as orthogonal systems, where each source node transmits messages to its destination node through a dedicated orthogonal channel with the help of relays. A practical example of such kind of system is the M2M networks that based on the long-term evolution or long-term evolution-advanced (LTE/LTE-A) system, in which each user receives its signals through an orthogonal frequency-division multiplexing (OFDM) subchannel. Transmission systems over nonorthogonal channels are referred to as nonorthogonal systems, where all users transmit signals simultaneously in a multiple access channel

and cause interference to each other. Such kind of multiple access channel system can be realized with time division multiple access (TDMA). Since M2M networks could be connected by various wireless network technologies, we consider both orthogonal and nonorthogonal systems in this paper.

In general, there are broadly three kinds of relay beamforming strategies: amplify-and-forward (AF), compress-and-forward (CF), and decode-and-forward (DF) [4–6]. In the AF strategy, relay nodes receive and amplify the signals transmitted from the source nodes and forward them to the destination nodes. Among the three strategies, the AF strategy is the most attractive one due to its low implementation complexity, which means low cost for the relay node devices and is essentially important for M2M sensor networks with large amounts of distributed relay sensor nodes. In our paper, the AF beamforming strategy is used in relay schemes for M2M sensor networks.

Various research work has been presented to propose AF relay beamforming schemes. To guarantee certain quality of service (QoS), relay beamforming approaches are mostly focused on signal-to-noise ratio (SNR) at the destinations for single user systems or orthogonal multiuser systems and signal-to-interference-plus-noise ratio (SINR) for nonorthogonal multiuser systems. Optimal relay beamforming schemes with relay power constraints are studied in [7, 8], in which algorithms are designed to achieve maximum SNR (MSNR) for a single user system and an orthogonal multiuser system, respectively. In [9], optimal relay beamforming weights for a nonorthogonal multiuser system are found to meet a given set of target SINR at the destinations while minimizing transmission power at the relays. These relay schemes aiming to MSNR or target SINR assume that all the relay nodes can communicate with each other and therefore each relay has global channel state information (CSI), that is, channel coefficients of all the relay nodes. The requirement of global CSI makes the MSNR or target SINR relay beamforming schemes not applicable for M2M sensor networks with a large number of relay nodes, where either the overhead of CSI exchange could be exceedingly high or even CSI exchange among relay nodes might be impossible due to the geographical locations of the relay nodes. It is appealing to derive an algorithm that each relay can learn its own beamforming weights independently based on local CSI without knowing other relays' channel information. To avoid the need of global CSI, minimum mean square error (MMSE) criterion instead of MSNR or target SINR is considered in [10–12]. In [10], MMSE-based approaches subject to individual power constraints with local CSI are derived for a single user system, which yield suboptimal but effective results. It is also demonstrated in [12] that the MMSE-based distributed implementation of signal detection can guarantee the system performance close to that of joint implementation with global CSI. However, the previous works do not consider the relay networks, such as the M2M networks. As a result, it is expected that the MSE cost function can be used to design the relay beamforming schemes with local CSI for multiuser M2M sensor networks. Thus, in our work, we study the distributed beamforming scheme with multiple relay nodes. The MSE cost function instead of the SNR or SINR cost function is considered for relay beamforming schemes. We aim at developing efficient distributed relay beamforming algorithms with local CSI that minimize the sum MSE of all users for a multiuser wireless M2M network using the AF strategy. We define MMSE as the minimum of sum MSE of all users in this paper, and, therefore, the relay beamforming schemes studied in our work are also referred to as the MMSE-based relay beamforming schemes.

The rest of the paper is organized as follows. In Section 2, the system model for relay assisted multiuser M2M networks is presented. In Section 3, we study the relay beamforming schemes with local CSI that minimizes the sum MSE of all users in nonorthogonal systems. Relay beamforming schemes for orthogonal systems are discussed in Section 4. In Section 5, simulation results under various conditions are presented. Conclusion of the paper can be found in Section 6.

Notations. Bold lower letters are used for vectors, for example, $\mathbf{x}$, and bold capital letters for matrices, for example, $\mathbf{X}$. $[\mathbf{X}]_{ij}$ stands for the (i, j)th element of $\mathbf{X}$. $\mathbf{diag}(\mathbf{x})$ denotes a diagonal square matrix with vector $\mathbf{x}$ as the diagonal elements. Superscripts $\mathbf{X}^{\mathcal{T}}$, $\mathbf{X}^*$, and $\mathbf{X}^\dagger$ stand for transpose, complex conjugate, and Hermitian transpose of $\mathbf{X}$, respectively. $|x|$ takes the modulus of a complex number and $\|\mathbf{X}\|$ gives the Frobenius norm of a vector or matrix. In this paper, the operator $\cdot$ stands for element-wise multiplication. The real and complex number fields are denoted by $\mathbb{R}$ and $\mathbb{C}$. $\mathbb{E}$ is the expectation operator. The notation $x \sim \mathcal{CN}(m, \sigma^2)$ means that x is a complex Gaussian random variable with mean m and variance σ^2.

2. System Model

Consider an AF relay assisted wireless M2M network, consisting of L pairs of source-destination user nodes and K relay nodes. All the users and the relays are equipped with only one antenna. As discussed in Section 1, relays operate in a half-duplex mode, that is, they cannot receive and transmit at the same time. The communication is implemented in two phases. Assume there is no direct link between any source node and destination node. In this paper, we consider both orthogonal and nonorthogonal systems. In orthogonal system, each user-destination pair takes up one orthogonal subchannel and the relay nodes design different sets of beamforming weights for different users. In nonorthogonal systems, each user causes interference to the other users in the same system, and only one set of beamforming weights are designed for all the users to cooperatively transmit all the signals to the destination nodes on the same frequency band. In addition, both cases face the relay power transmission constraints. The system model is depicted in Figure 1.

Let $\mathbf{s} = [s_1, \ldots, s_L]^{\mathcal{T}}$, an $\mathbf{L} \times 1$ column vector, denote the transmitted signal vector from the L source nodes. It is assumed that each source node transmits at the same power level, that is, $P_s = \mathbb{E}[|s_i|^2]$ with $\mathbb{E}(s_i) = 0$ for $i = 1, \ldots, L$. $\mathbf{h}_i = [h_{i1}, h_{i2}, \ldots, h_{iK}]^{\mathcal{T}} \in \mathbb{C}^{K\times 1}$ is

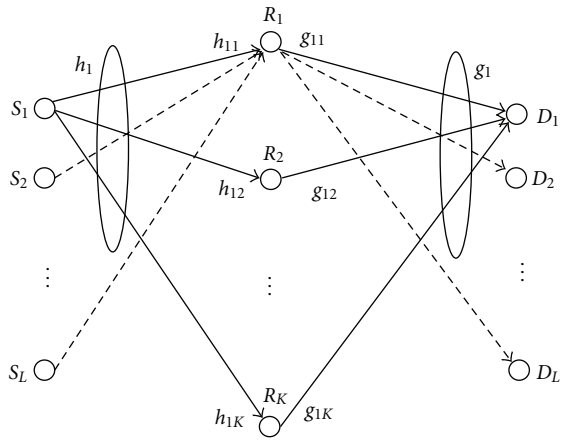


Figure 1: AF multiuser multirelay M2M communication system model.

the channel coefficient vector from source node i to the K relay nodes. Column vector $\mathbf{g}_i = [g_{i1}, g_{i2}, \ldots, g_{iK}]^{\mathcal{T}} \in \mathbb{C}^{K\times 1}$ is the channel coefficient vector from the K relay nodes to destination node i. In nonorthogonal systems, $\mathbf{n}_r = [n_{r1}, n_{r2}, \ldots, n_{rK}]^{\mathcal{T}}$ represents the relay background noise over the shared channel, with independent variable $n_{ri} \sim \mathcal{CN}(0, \sigma_{ri}^2)$ denoting the background noise at relay i, for $i = 1, \ldots, K$. In orthogonal systems, $\mathbf{n}_{lr} = [n_{lr1}, n_{lr2}, \ldots, n_{lrK}]^{\mathcal{T}}$ represents the relay background noise over the subchannel of user l, with independent variable $n_{lri} \sim \mathcal{CN}(0, \sigma_{lri}^2)$ denoting the background noise at relay i, for $i = 1, \ldots, K$. $n_i \sim \mathcal{CN}(0, \sigma_{di}^2)$ is the background noise at destination i.

As in most research work, in this paper, all the channels are assumed to be independent, with a Rayleigh flat fading distribution expressed as $h_{ij} \sim \mathcal{CN}(0, \sigma_h^2)$ and $g_{ij} \sim \mathcal{CN}(0, \sigma_g^2)$ for all i and j. For convenience, we assume that each relay faces a same level of added white Gaussian noise (AWGN), that is, $\sigma_{ri}^2 = \sigma_{rli}^2 = \sigma_r^2$, for $i = 1, \ldots, K$ and $l = 1, \ldots, L$. The power of AWGN at each destination node is assumed to be the same, that is, $\sigma_{di}^2 = \sigma_0^2$, for $i = 1, \ldots, L$.

2.1. Nonorthogonal Channel Scenario. If the channels are nonorthogonal, the received signals at the relays in the first phase can be written as

$$\mathbf{r} = \sum_{l=1}^{L} \mathbf{h}_l s_l + \mathbf{n}_r, \tag{1}$$

where $\mathbf{r}$ is a $\mathbb{C}^{K\times 1}$ column vector $(r_1, \ldots, r_K)^{\mathcal{T}}$ and r_i is the signal received by the ith relay.

The relays then process the received signals with complex beamforming weights. Define a row vector $\mathbf{w} = [w_1, \ldots, w_K] \in \mathbb{C}^{1\times K}$ as the beamforming weight vector. The signals transmitted by the relays can be expressed as

$$\mathbf{x} = \mathbf{w}^{\mathcal{T}} \cdot \mathbf{r}. \tag{2}$$

In the second phase, the relays broadcast the processed signals to all the destination nodes simultaneously. The channels between relay nodes and destination nodes are also nonorthogonal. As a result, the received signal at the ith destination node is given by

$$\begin{aligned} y_i &= \mathbf{g}_i^{\mathcal{T}} \mathbf{x} + n_i \\ &= \mathbf{g}_i^{\mathcal{T}} \left[\mathbf{w}^{\mathcal{T}} \cdot \left(\sum_{l=1}^{L} \mathbf{h}_l s_l + \mathbf{n}_r \right) \right] + n_i. \end{aligned} \tag{3}$$

2.2. Orthogonal Channel Scenario. In the case of orthogonal channels, relays can separate the messages sent from different users. For user l, the received signal at the relays can be written as

$$\mathbf{r}_l = \mathbf{h}_l s_l + \mathbf{n}_{rl}, \tag{4}$$

where $\mathbf{r}_l$ is a $\mathbb{C}^{K\times 1}$ column vector and r_{li} is the signal received by the ith relay from user node l. $\mathbf{n}_{rl}$ is the relay AWGN for user l over its subchannel.

The relays design different sets of complex beamforming weights for different users. Define a row vector $\mathbf{w}_l = [w_{l1}, \ldots, w_{lK}] \in \mathbb{C}^{1\times K}$ as the beamforming weight vector for user l. The lth user's signal transmitted by the relays can be written as

$$\mathbf{x}_l = \mathbf{w}_l^{\mathcal{T}} \cdot \mathbf{r}_l. \tag{5}$$

In the second phase, the relay nodes simultaneously broadcast the processed signals to all the destination nodes. The transmission to each destination node is carried out over orthogonal channels to avoid interuser interference. Thus, the received signal at the lth destination node is given by

$$\begin{aligned} y_l &= \mathbf{g}_l^{\mathcal{T}} \mathbf{x}_l + n_l \\ &= \mathbf{g}_l^{\mathcal{T}} \left[\mathbf{w}_l^{\mathcal{T}} \cdot (\mathbf{h}_l s_l + \mathbf{n}_{lr}) \right] + n_l. \end{aligned} \tag{6}$$

It can be seen from the above that the goal of relay beamforming schemes for nonorthogonal systems is to design a set of $K \times 1$ row vector $\mathbf{w}$ for all the users, while for orthogonal system L sets of $K \times 1$ row vectors, that is, $\mathbf{w}_1, \ldots, \mathbf{w}_L$, each one corresponding to one source-destination pair.

As mentioned in Section 1, the aim of this paper is to develop distributed algorithms using local CSI instead of full global CSI. In the discussed system, local CSI for relay i is referred to as incoming channel coefficients h_{li} from the source nodes to relay node i, outgoing channel coefficients g_{li} from relay node i to the destination nodes for $l = 1, \ldots, L$ and its noise variance σ_r^2. We can denote $\overline{h}_i = [h_{1i}, h_{2i}, \ldots, h_{Li}]$ and $\overline{g}_i = [g_{1i}, g_{2i}, \ldots, g_{Li}]$ as the local CSI vectors for relay node i. Besides, the signal power P_s and the destination noise variance σ_0^2 are assumed to be known at the relays. In the following sections, we will derive suboptimal beamforming schemes based on MMSE that only require local CSI for nonorthogonal and orthogonal systems.

3. MMSE-Based Distributed Beamforming for Nonorthogonal Systems

In this section, we are interested in choosing a beamforming weight vector that minimizes the sum of the mean square

error between the uncorrupted received signals at the destination nodes and the transmitted signals at the source nodes for nonorthogonal systems. The optimization problem is first formulated and then two suboptimal ways to obtain beamforming weights from the original problem using local CSI are put forward.

3.1. MMSE Optimization Problem Formulation. The uncorrupted received signal at the ith destination node is $\mathbf{g}_i^{\mathcal{T}}\mathbf{x}$, so the MSE of the ith user is expressed as

$$\begin{aligned}\mathrm{MSE}_i(\mathbf{w}) &= \mathbb{E}\left[\left|\mathbf{g}_i^{\mathcal{T}}\mathbf{x} - s_i\right|^2\right] \\ &= P_s\sum_{l=1}^{L}\mathbf{w}(\mathbf{h}_l\cdot\mathbf{g}_i)(\mathbf{h}_l\cdot\mathbf{g}_i)^{\dagger}\mathbf{w}^{\dagger} - P_s\mathbf{w}(\mathbf{h}_i\cdot\mathbf{g}_i) \\ &\quad - P_s(\mathbf{h}_i\cdot\mathbf{g}_i)^{\dagger}\mathbf{w}^{\dagger} + \sigma_r^2\mathbf{w}\{\mathbf{diag}(\mathbf{g}_i\cdot\mathbf{g}_i^*)\}\mathbf{w}^{\dagger} \\ &\quad + P_s.\end{aligned} \tag{7}$$

The sum of the MSE of all the source and destination pairs is given by

$$\begin{aligned}\mathrm{SUM\_MSE}(\mathbf{w}) &= \sum_{i=1}^{L}\mathrm{MSE}_i(\mathbf{w}) \\ &= P_s\sum_{i=1}^{L}\sum_{l=1}^{L}\mathbf{w}(\mathbf{h}_l\cdot\mathbf{g}_i)(\mathbf{h}_l\cdot\mathbf{g}_i)^{\dagger}\mathbf{w}^{\dagger} \\ &\quad - P_s\mathbf{w}\sum_{i=1}^{L}(\mathbf{h}_i\cdot\mathbf{g}_i) - \sum_{i=1}^{L}P_s(\mathbf{h}_i\cdot\mathbf{g}_i)^{\dagger}\mathbf{w}^{\dagger} \\ &\quad + \sigma_r^2\mathbf{w}\left\{\sum_{i=1}^{L}\mathbf{diag}(\mathbf{g}_i\cdot\mathbf{g}_i^*)\right\}\mathbf{w}^{\dagger} + LP_s.\end{aligned} \tag{8}$$

Transmission power limitation of the relay sensor nodes needs to be taken into consideration in wireless M2M networks. The transmission power of each relay node is constrained within the maximum power capacity denoted by P_{rel}. Then we have the relay power constraints as

$$\begin{aligned}P_i &= \mathbb{E}\left[\|x_i\|^2\right] = w_i w_i^*\left[P_s\sum_{l=1}^{L}h_{li}h_{li}^* + \sigma_r^2\right] \\ &= \mathbf{w}\mathbf{D}_i\mathbf{w}^* \le P_{\mathrm{rel}}, \quad i = 1,\ldots,K,\end{aligned} \tag{9}$$

where $\mathbf{D}_i$ has only one nonzero element with $[\mathbf{D}_i]_{ii} = P_s\sum_{l=1}^{L}h_{li}h_{li}^* + \sigma_r^2$.

To obtain the best performance for the system, we design a beamforming scheme that generates the smallest sum of MSE. Thus, the optimization problem can be formulated as an MMSE problem subject to individual relay transmission power constraints and is expressed as

$$\begin{aligned}&\underset{\mathbf{w}}{\text{minimize}}\ \mathrm{SUM\_MSE} \\ &\text{subject to } P_{\mathrm{rel}} \ge P_i, \quad i = 1,\ldots,K.\end{aligned} \tag{10}$$

To solve the optimization problem, we first introduce the Lagrangian of the problem

$$\begin{aligned}\mathcal{L}(\mathbf{w},\boldsymbol{\lambda}) &= \mathrm{SUM\_MSE} + \sum_{k=1}^{K}\lambda_k(P_k - P_{\mathrm{rel}}) \\ &= P_s\sum_{i=1}^{L}\sum_{l=1}^{L}\mathbf{w}(\mathbf{h}_l\cdot\mathbf{g}_i)(\mathbf{h}_l\cdot\mathbf{g}_i)^{\dagger}\mathbf{w}^{\dagger} \\ &\quad - P_s\mathbf{w}\sum_{i=1}^{L}(\mathbf{h}_i\cdot\mathbf{g}_i) - \sum_{i=1}^{L}P_s(\mathbf{h}_i\cdot\mathbf{g}_i)^{\dagger}\mathbf{w}^{\dagger} \\ &\quad + \sigma_r^2\mathbf{w}\left\{\sum_{i=1}^{L}\mathbf{diag}(\mathbf{g}_i\cdot\mathbf{g}_i^*)\right\}\mathbf{w}^{\dagger} + LP_s \\ &\quad + \sum_{i=1}^{K}\lambda_i(\mathbf{w}\mathbf{D}_i\mathbf{w}^* - P_{\mathrm{rel}}),\end{aligned} \tag{11}$$

where $\boldsymbol{\lambda} = [\lambda_1,\ldots,\lambda_K]$ are the Lagrangian multipliers and $\lambda_i \ge 0$, for $i = 1,\ldots,K$.

The Lagrangian dual function is defined as

$$g(\boldsymbol{\lambda}) = \min_{\mathbf{w}}\mathcal{L}(\mathbf{w},\boldsymbol{\lambda}). \tag{12}$$

The dual problem is to maximize $g(\boldsymbol{\lambda})$. According to Karush-Kuhn-Tucker (KKT) conditions, the optimal solution should satisfy

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = 0, \tag{13}$$

$$\lambda_i(P_{\mathrm{rel}} - P_i) = 0, \tag{14}$$

$$\lambda_i \ge 0, \quad i = 1,\ldots,K. \tag{15}$$

Due to the symmetry of $\mathbf{w}$ and $\mathbf{w}^{\dagger}$ in (13), (13) is equivalent to $\partial\mathcal{L}/\partial\mathbf{w}^{\dagger} = 0$. Expand the differentiation equation, we have

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}^{\dagger}} = P_s\mathbf{w}\mathbf{A} + \sigma_r^2\mathbf{w}\mathbf{B} - P_s\mathbf{c} + \mathbf{w}\mathbf{D}, \tag{16}$$

where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{D}$ are $K\times K$ matrices while $\mathbf{c}$ is a $1\times K$ row vector. For $\mathbf{A}$, $[\mathbf{A}]_{mk} = \sum_{i=1}^{L}\sum_{j=1}^{L}(g_{im}h_{jm})(g_{ik}h_{jk})^*$. $\mathbf{B}$ and $\mathbf{D}$ are both diagonal with $[\mathbf{B}]_{kk} = \sum_{i=1}^{L}g_{ik}g_{ik}^*$ and $[\mathbf{D}]_{kk} = \lambda_k(P_s\sum_{j=1}^{L}h_{jk}h_{jk}^* + \sigma_r^2)$ for $k = 1,\ldots,K$, respectively. As for row vector $\mathbf{c}$, $c_m = \sum_{i=1}^{L}(g_{im}h_{im})^*$.

Note that $\mathbf{w}$ is a $1\times K$ row vector, and (16) is consisted of K equations. The jth equation of (16) can be reformulated as

$$\begin{aligned}w_j &= \frac{P_s c_j - P_s\sum_{i=1,i\ne j}^{K}w_i[\mathbf{A}]_{ij}}{P_s[\mathbf{A}]_{jj} + \sigma_r^2[\mathbf{B}]_{jj} + \lambda_j\left[\mathbf{D}_j\right]_{jj}} \\ &= \frac{P_s\sum_{i=1}^{L}\left(g_{ij}h_{ij}\right)^* - P_s\sum_{i=1,i\ne j}^{K}w_i[\mathbf{A}]_{ij}}{P_s[\mathbf{A}]_{jj} + \sigma_r^2\sum_{i=1}^{L}g_{ij}g_{ij}^* + \lambda_j\left(P_s\sum_{m=1}^{L}h_{mj}h_{mj}^* + \sigma_r^2\right)},\end{aligned} \tag{17}$$

where $[\mathbf{A}]_{jj}$, $[\mathbf{B}]_{jj}$, c_j, and $[\mathbf{D}_j]_{jj}$ only contain local CSI at relay j while $[\mathbf{A}]_{ij}$ for $i \neq j$ has other relays' CSI. It can be seen clearly that to calculate the optimal beamforming weight of relay j requires the complete knowledge of all channels, that is, the global CSI. This is not desirable for wireless M2M networks, where there are a large number of relay nodes limited in memory, battery, and processing capability. As a result, we proposed two suboptimal algorithms that each relay can learn its own beamforming weight with local CSI.

3.2. MSE Minimization Suboptimal Algorithm. In this part, we are dedicated to find the suboptimal beamforming weights assuming that each relay only has local CSI and there is no information exchange between the relays. Below, we propose two suboptimal approaches to deal with the lack of global CSI.

3.2.1. Suboptimal Algorithm 1. First, in (17), we can see that it does not involve channel information from other relays except for the second term of the numerator. It can also be seen from (14) that λ_j corresponds to relay j's transmission power constraint which is determined by local beamforming weights and local CSI. In this case, we can ignore the other relays' contribution to the sum MSE in (8). As a result, we can set all the elements $[A]_{ij}$ where $i \neq j$ to zero for relay j. Then we have the following expression:

$$\begin{aligned} w_j &= \frac{P_s c_j}{P_s[\mathbf{A}]_{jj} + \sigma_r^2[\mathbf{B}]_{jj} + [\mathbf{D}]_{jj}} \\ &= \frac{P_s c_j}{P_s[\mathbf{A}]_{jj} + \sigma_r^2[\mathbf{B}]_{jj} + \lambda_j\left[\mathbf{D}_j\right]_{jj}} \\ &= \frac{P_s \sum_{i=1}^{L}\left(g_{ij}h_{ij}\right)^*}{P_s[\mathbf{A}]_{jj} + \sigma_r^2 \sum_{i=1}^{L} g_{ij}g_{ij}^* + \lambda_j\left(P_s \sum_{m=1}^{L} h_{mj}h_{mj}^* + \sigma_r^2\right)}. \end{aligned} \tag{18}$$

Substitute (18) into (14), and λ_j is given by $\lambda_j = \max\{0, z_{j1}\}$, where z_{j1} is

$$z_{j1} = \frac{1}{\left[\mathbf{D}_j\right]_{jj}}\left\{\frac{P_s\left|c_j\right|\sqrt{\left[\mathbf{D}_j\right]_{jj}}}{\sqrt{P_{\text{rel}}}} - P_s[\mathbf{A}]_{jj} - \sigma_r^2[\mathbf{B}]_{jj}\right\}. \tag{19}$$

When $z_{j1} < 0$, λ_j is 0, and w_j is

$$\begin{aligned} w_j &= \frac{P_s c_j}{P_s[\mathbf{A}]_{jj} + \sigma_r^2[\mathbf{B}]_{jj}} \\ &= \frac{P_s \sum_{i=1}^{L}\left(g_{ij}h_{ij}\right)^*}{P_s \sum_{i=1}^{L}\sum_{l=1}^{L}\left(g_{ij}h_{lj}\right)\left(g_{ij}h_{lj}\right)^* + \sigma_r^2 \sum_{i=1}^{L} g_{ij}g_{ij}^*}. \end{aligned} \tag{20}$$

In this case, the relay node does not need to turn to the maximum transmission power. To let $z_{j1} < 0$, we have

$$\frac{P_s^2\left|\sum_{i=1}^{L}\left(g_{ij}h_{ij}\right)^*\right|^2}{P_{\text{rel}}\left(\sum_{i=1}^{L}\left|g_{ij}\right|^2\right)^2\left(P_s\sum_{l=1}^{L}\left|h_{lj}\right|^2 + \sigma_r^2\right)} < 1. \tag{21}$$

With the inequality $|\sum_{i=1}^{L}(g_{ij}h_{ij})^*|^2 \leq \sum_{i=1}^{L}|g_{ij}|^2\sum_{l=1}^{L}|h_{lj}|^2$, the left term of (21) satisfies the inequality below,

$$\begin{aligned} &\frac{P_s^2\left|\sum_{i=1}^{L}\left(g_{ij}h_{ij}\right)^*\right|^2}{P_{\text{rel}}\left(\sum_{i=1}^{L}\left|g_{ij}\right|^2\right)^2\left(P_s\sum_{l=1}^{L}\left|h_{lj}\right|^2 + \sigma_r^2\right)} \\ &\leq \frac{P_s^2\sum_{i=1}^{L}\left|h_{ij}\right|^2}{P_{\text{rel}}\left(\sum_{i=1}^{L}\left|g_{ij}\right|^2\right)\left(P_s\sum_{l=1}^{L}\left|h_{lj}\right|^2 + \sigma_r^2\right)}. \end{aligned} \tag{22}$$

When the right term of (22) is less than 1, (21) is satisfied. It means that if the channel gains between the relay nodes and the destination nodes are strong enough, that is,

$$\frac{\sum_{i=1}^{L}\left|g_{ij}\right|^2}{\sum_{i=1}^{L}\left|h_{ij}\right|^2} > \frac{P_s^2}{P_{\text{rel}}\left(P_s\sum_{l=1}^{L}\left|h_{lj}\right|^2 + \sigma_r^2\right)}, \tag{23}$$

the relays do not need to transmit the maximum power. This result is similar to that in single user systems [10]. However, if $z_{j1} > 0$, then $P_i = P_{\text{rel}}$, and the relay transmits at the maximum relay power level.

By ignoring other relays' CSI, the above algorithm allows each relay to calculate its own beamforming weight independently. Simulation results show that this suboptimal approach can give satisfactory performance which is very close to the optimal approach.

3.2.2. Suboptimal Algorithm 2. Second, instead of ignoring other relays' channel coefficients, we may substitute them with appropriate approximations. In [10], it assumes symmetric conditions for all the channels in a single user system, that is, it uses the corresponding local CSI statistics as approximations of the unknown global CSI. Here, we extent this idea to our multiuser system. As shown in (17), relay node j still needs to know $w_i[\mathbf{A}]_{ij}$ where $i \neq j$ to learn its own weight. Under the assumption of symmetric channel condition, we have $w_j[\mathbf{A}]_{jj} = w_i[\mathbf{A}]_{ij}$. Replace this into (17), and we can get

$$\begin{aligned} w_j &= \frac{P_s c_j - P_s\sum_{i=1,\, i\neq j}^{K} w_i[\mathbf{A}]_{ij}}{P_s[\mathbf{A}]_{jj} + \sigma_r^2[\mathbf{B}]_{jj} + [\mathbf{D}]_{jj}} \\ &\approx \frac{P_s c_j - (K-1)P_s w_j[\mathbf{A}]_{jj}}{P_s[\mathbf{A}]_{jj} + \sigma_r^2[\mathbf{B}]_{jj} + [\mathbf{D}]_{jj}}. \end{aligned} \tag{24}$$

Reformulate the above equation, and the suboptimal solution of relay j's beamforming weight is given by

$$w_j \approx \frac{P_s c_j}{KP_s[\mathbf{A}]_{jj} + \sigma_r^2[\mathbf{B}]_{jj} + [\mathbf{D}]_{jj}} = \frac{P_s \sum_{i=1}^{L} \left(g_{ij}h_{ij}\right)^*}{KP_s[\mathbf{A}]_{jj} + \sigma_r^2 \sum_{i=1}^{L} g_{ij}g_{ij}^* + \lambda_j \left[\mathbf{D}_j\right]_{jj}}. \tag{25}$$

As discussed in the Suboptimal Algorithm 1, substitute (25) into the power constraint of (14). The Lagrangian multiplier λ_j is given by $\lambda_j = \max\{0, z_{j2}\}$, where z_{j2} is

$$z_{j2} = \frac{1}{\left[\mathbf{D}_j\right]_{jj}} \left\{ \frac{P_s \left| c_j \right| \sqrt{\left[\mathbf{D}_j\right]_{jj}}}{\sqrt{P_{\text{rel}}}} - KP_s[\mathbf{A}]_{jj} - \sigma_r^2[\mathbf{B}]_{jj} \right\}. \tag{26}$$

Similar to Suboptimal Algorithm 1, if

$$\frac{\sum_{i=1}^{L} \left| g_{ij} \right|^2}{\sum_{i=1}^{L} \left| h_{ij} \right|^2} > \frac{P_s^2}{P_{\text{rel}} \left(KP_s \sum_{l=1}^{L} \left| h_{lj} \right|^2 + \sigma_r^2 \right)}, \tag{27}$$

then $z_{j2} < 0$, and the relay transmission power is less than the maximum allowable power. Otherwise the relay will transmit signals at the maximum power.

As can be seen, both of the proposed suboptimal algorithms only require local CSI.

4. MMSE Beamforming for Orthogonal Systems

In this section, we develop distributed relay beamforming algorithms for a different multiuser system where each user transmits signals through orthogonal channels. In this scenario, the sum of MSE criterion is also considered to design the beamforming weights. The optimization problem is first established and then we will show the optimization problem can be divided into several single user problems.

4.1. Optimization Problem Formulation. The uncorrupted received signal at the ith destination node is $\mathbf{g}_i^{\mathcal{T}} \mathbf{x}_i$, so the MSE of the ith user is expressed as

$$\begin{aligned} \text{MSE}_i(\mathbf{w}_i) &= \mathbb{E}\left[\left| \mathbf{g}_i^{\mathcal{T}} \mathbf{x}_i - s_i \right|^2 \right] \\ &= P_s \mathbf{w}_i (\mathbf{h}_i \cdot \mathbf{g}_i)(\mathbf{h}_i \cdot \mathbf{g}_i)^{\dagger} \mathbf{w}_i^{\dagger} - P_s \mathbf{w}_i (\mathbf{h}_i \cdot \mathbf{g}_i) \\ &\quad - P_s (\mathbf{h}_i \cdot \mathbf{g}_i)^{\dagger} \mathbf{w}_i^{\dagger} + \sigma_r^2 \mathbf{w}_i \{\mathbf{diag}(\mathbf{g}_i \cdot \mathbf{g}_i^*)\} \mathbf{w}_i^{\dagger} \\ &\quad + P_s. \end{aligned} \tag{28}$$

The sum of MSE of all the source nodes and destination nodes is given by

$$\begin{aligned} &\text{SUM\_MSE}(\mathbf{w}_1, \ldots, \mathbf{w}_L) \\ &= \sum_{i=1}^{L} \text{MSE}_i(\mathbf{w}_i) \\ &= P_s \sum_{i=1}^{L} \mathbf{w}_i (\mathbf{h}_i \cdot \mathbf{g}_i)(\mathbf{h}_i \cdot \mathbf{g}_i)^{\dagger} \mathbf{w}_i^{\dagger} - P_s \sum_{i=1}^{L} \mathbf{w}_i (\mathbf{h}_i \cdot \mathbf{g}_i) \\ &\quad - P_s \sum_{i=1}^{L} (\mathbf{h}_i \cdot \mathbf{g}_i)^{\dagger} \mathbf{w}_i^{\dagger} + \sum_{i=1}^{L} \sigma_r^2 \mathbf{w}_i \{\mathbf{diag}(\mathbf{g}_i \cdot \mathbf{g}_i^*)\} \mathbf{w}_i^{\dagger} \\ &\quad + LP_s. \end{aligned} \tag{29}$$

Similar to that in nonorthogonal channel systems, power constraints of the relay sensor nodes also need to be taken into consideration. The transmission power of each relay is constrained within the maximum power denoted by P_{rel}, and then we have the relay power constraints as

$$\begin{aligned} P_i &= \sum_{l=1}^{L} \mathbb{E}\left[\|x_{li}\|^2 \right] = \sum_{l=1}^{L} w_{li} w_{li}^* \left[P_s h_{li} h_{li}^* + \sigma_r^2 \right] \\ &= \sum_{l=1}^{L} \mathbf{w}_l \mathbf{D}_{li} \mathbf{w}_l^* \le P_{\text{rel}}, \quad i = 1, \ldots, K, \end{aligned} \tag{30}$$

where $\mathbf{D}_{li}$ has only one nonzero element with $[\mathbf{D}_{li}]_{ii} = P_s h_{li} h_{li}^* + \sigma_r^2$.

Then the optimization problem can be formulated as

$$\begin{aligned} &\underset{\mathbf{w}_1, \ldots, \mathbf{w}_L}{\text{minimize}} \ \text{SUM\_MSE} \\ &\text{subject to } P_{\text{rel}} \ge P_i, \quad i = 1, \ldots, K. \end{aligned} \tag{31}$$

4.2. Suboptimal Algorithms for Orthogonal Scenarios. In orthogonal systems, each relay needs to learn L weights for L users. To learn the beamforming weights independently by relay nodes, we can establish the Lagrangian function as in the nonorthogonal cases,

$$\mathcal{L}(\mathbf{w}_1, \ldots, \mathbf{w}_L, \boldsymbol{\lambda}) = \text{SUM\_MSE} + \sum_{k=1}^{K} \lambda_k (P_k - P_{\text{rel}}). \tag{32}$$

The optimal solution $\mathbf{w}_1, \ldots, \mathbf{w}_L, \boldsymbol{\lambda}$ should satisfy the KKT conditions. We can have

$$w_{li} = \frac{P_s (h_{li} g_{li})^* - P_s \sum_{k=1,\, k \ne i}^{K} w_{lk} [\mathbf{A}_l]_{ki}}{P_s [\mathbf{A}_l]_{ii} + \sigma_r^2 g_{li} g_{li}^* + \lambda_i \left(P_s h_{li} h_{li}^* + \sigma_r^2 \right)}, \tag{33}$$

where w_{li} is the beamforming weight at relay i for user l, and $\mathbf{A}_l$ is a $\mathbb{C}^{K \times K}$ matrix with $[\mathbf{A}_l]_{ij} = h_{li} g_{li} h_{lj}^* g_{lj}^*$.

For relay i, it needs to decide the value of $w_{1i}, w_{2i}, \ldots, w_{Li}$. The expression of w_{li} exhibits similar features to those in nonorthogonal systems with only the second term on the numerator involving global CSI, which may be ignored or replaced with approximations.

Substituting the suboptimal w_{li} for $l = 1,\ldots,L$ into the power constraints of (30), we can get a polynomial equation concerning the Lagrange multiplier λ_i. It decides the allocation of the relay transmission power to different users. We can use iterative methods to solve the polynomial equations.

Also, note that the transmissions through orthogonal channels make the minimization of sum MSE without power constraints equivalent to each user minimizing its own MSE with its beamforming vector. By preallocating the total relay transmission power capacity to each user, the original problem can be decoupled into L single user beamforming problems expressed as

$$\begin{aligned} &\underset{\mathbf{w}_l}{\text{minimize}}\ \text{MSE}_l \\ &\text{subject to } P_{rli} \geq P_{li}, \quad i = 1,\ldots,K. \\ &\text{for} \quad l = 1,\ldots,L, \end{aligned} \tag{34}$$

where P_{rli} is the maximum transmission power allocated for user l at relay node i and should satisfy $\sum_{l=1}^{L} P_{rli} = P_{\text{rel}}$. P_{li} is the power that relay node i uses to transmit user l's signal. At this point, the original problem becomes L single user subproblems. The approaches introduced in the nonorthogonal section can be used for the relays to learn the weights based on local CSI. A simple way is to equally allocate the available transmission power to each user, that is, to let $P_{rli} = P_{\text{rel}}/L$. By decoupling the original problem into single user cases, relays are able to calculate the weights on local CSI avoiding any information exchange with low complexity. However, the cost is the sacrifice of resource efficiency which means performance degradation from another point of view.

5. Simulation Results

In this section, we present numerical results for the proposed algorithms. It is assumed that all the channel coefficients are generated as zero-mean and unit-variance independent complex Gaussian random variables. Noise variances at the relays and the destinations are set to be the same as $\sigma_r^2 = \sigma_0^2 = 1$. BPSK is employed in our simulation system for simplicity. Define SNR $= P_s/\sigma_r^2$. We study the bit error rate (BER) performance under different circumstances.

Figures 2, 3, and 4 show the performance of Suboptimal Algorithm 1 in nonorthogonal systems. Figure 2 shows the BER performance over the level of SNR with different number of relays deployed in a three-user system. It is clearly that increasing the number of relays improves the system performance as expected because of the benefit of diversity gain. In Figure 3, BER performance is depicted when the number of users increases under different source transmission power. It shows that the number of users can have a great impact on the system performance. Although the source transmission power increases, the BER performance still degrades a lot with six or more users in the system. This is reasonable for that in nonorthogonal systems each user causes interference to the others and it needs sufficient relay nodes to provide a stronger beamforming to cancel out the interference for all the users at the same time. Thus, the same performance could be maintained for a larger number of users by increasing the number of relay nodes. Wireless M2M networks can deploy large numbers of low-cost relay nodes or allow some user nodes to act as relays to guarantee communication quality.

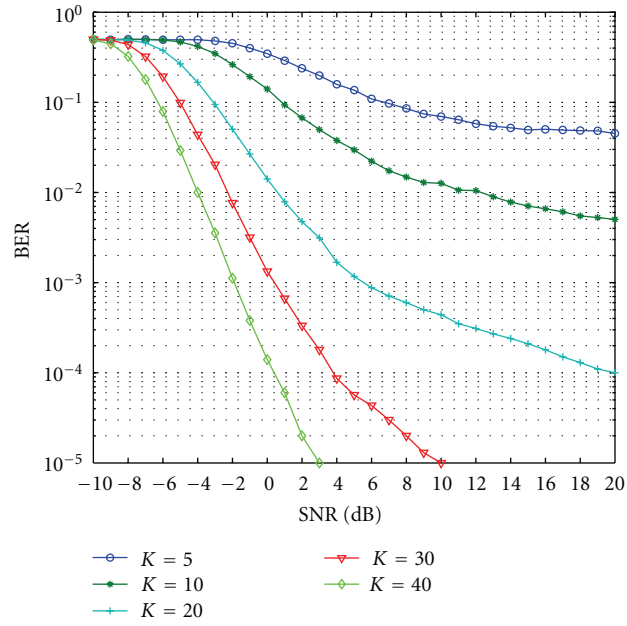


Figure 2: BER performance versus source transmission power with different number of relays in a 3 user nonorthogonal system.

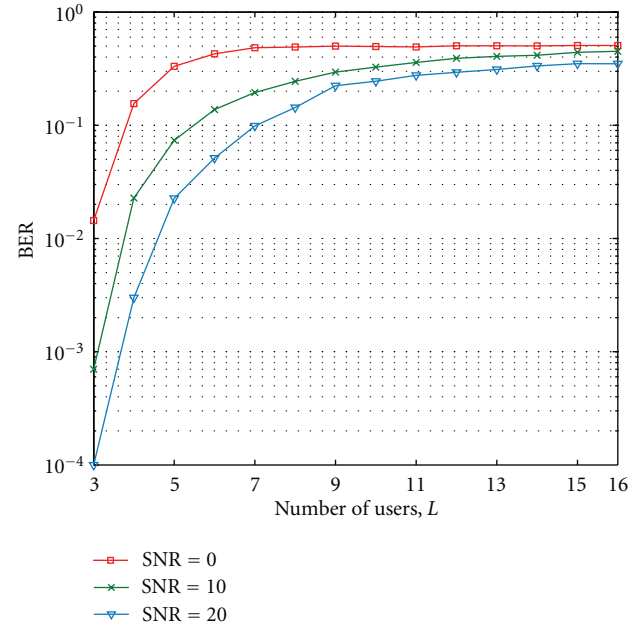


Figure 3: BER performance versus the number of users for different values of SNR with 20 relays in nonorthogonal system.

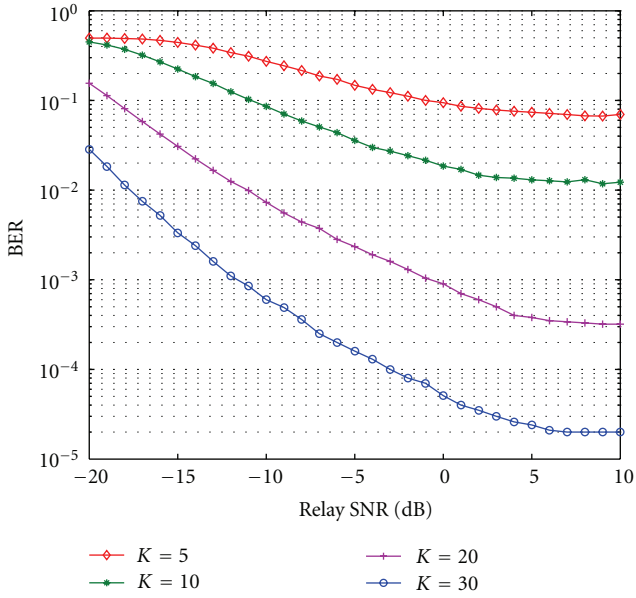


Figure 4: BER performance versus relay power constraint for different relays in nonorthogonal system.

Figure 4 shows the BER performance versus different relay power constraints when different number of relays are deployed. Relay SNR is defined as the relay maximum power with relay noise σ_r^2 set to be 0 dB. It can be seen that, at the beginning, with the maximum relay transmission power increasing, the BER drops greatly but then all the curves exhibit error floor even though the power value keeps rising. This result is consistent with the analysis in Section 3 that the best transmission power depends on (22). When the inequality satisfies, the relays do not need to transmit signals at the maximum power.

Figure 5 shows the performance in orthogonal channel systems. In Figure 5, we consider both cases where the relay power is equally allocated to all the users and optimized based on lagrange multipliers, with lines labeled as Eq and Ineq, respectively. Also, we compare the proposed schemes with the one based on MSNR criterion using global CSI proposed in [8] labeled as MSNR in Figure 5. It can be seen that the scheme optimizing the relay power allocation to different users have better performance than that with equal allocation. It is also clear that the performance of both schemes is close to that of [8]. Since users are not interfered with each other, BER performance is slightly affected by the number of users. However, as the number of users grows, the relay transmission capacity for each user decreases and ends in performance degradation. Compare Figure 5 with Figure 2, and it can be seen that the performance of orthogonal system is better than the nonorthogonal system at the cost of more frequency resource.

Figure 6 compares the optimal results using global CSI with that of the Suboptimal Algorithm 1 and Suboptimal Algorithm 2 in nonorthogonal systems. The schemes using global CSI are proposed in [9, 13], based on MSNR and MMSE criterion, and labeled as joint MSNR and joint

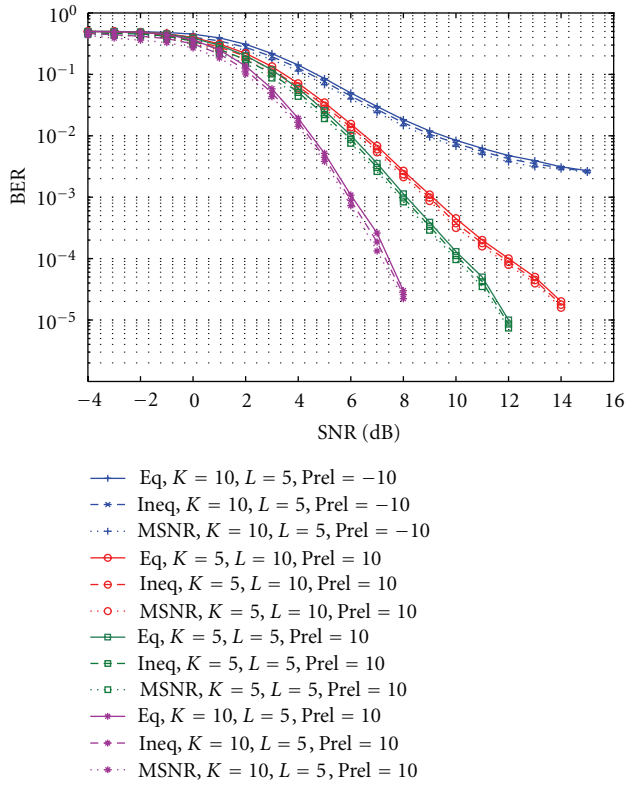


Figure 5: BER performance versus source transmission power for orthogonal system.

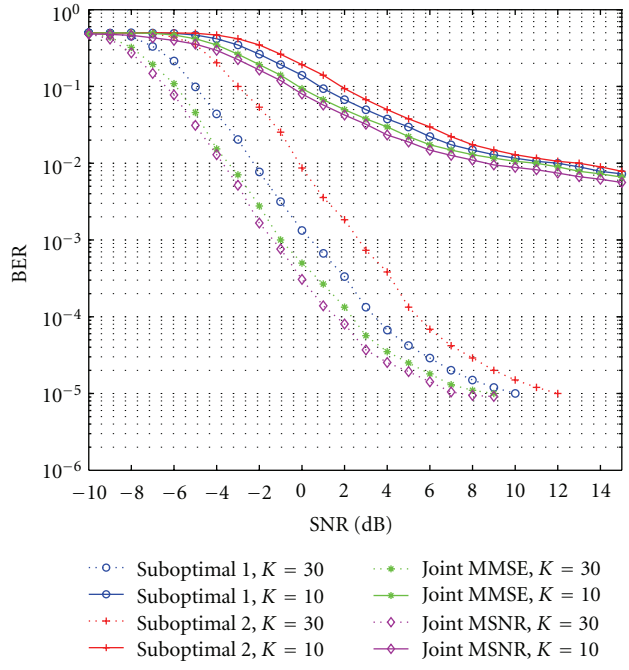


Figure 6: BER performance comparison for nonorthogonal system.

MMSE, respectively. It can be seen that the performance of Suboptimal Algorithm 1 is very close to the optimal performance. As the number of relays increases, the difference becomes larger but remains acceptable. Suboptimal Algorithm 2 gives quite good approximates when the number of relays is small. The instantaneous global CSI estimated with local CSI gets less accurate as the number grows, and hence degrades the performance of Suboptimal Algorithm 2.

6. Conclusion

In this paper, we considered distributed relay beamforming schemes with local CSI for AF multiuser multirelay M2M networks with orthogonal channels and nonorthogonal channels based on MMSE criterion. In nonorthogonal systems, two approximate approaches were provided where each relay could learn its own weight independently and respectively with local CSI. Simulation results showed that the performance of the proposed algorithm was close to the optimal one using global CSI. In orthogonal systems, the optimization problem was separated into single user problems which then could be solved with distributed algorithms. Simulation results were presented in the end of the paper to show the BER performance of the proposed relay beamforming schemes.

Acknowledgment

This work has been supported by the National Basic Research Program of China (973 Program no. 2009CB320403).

References

[1] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bölcskei, "An overview of MIMO communications—a key to gigabit wireless," *Proceedings of the IEEE*, vol. 92, no. 2, pp. 198–217, 2004.

[2] M. Fadel, A. El-Keyi, and A. Sultan, "QOS-constrained multiuser peer-to-peer amplify-and-forward relay beamforming," *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1397–1408, 2012.

[3] G. J. Pottie and W. J. Kaiser, "Wireless sensor networks," in *Proceedings of the IEEE Information Theory Workshop (ITW '98)*, pp. 139–140, June 1998.

[4] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037–3063, 2005.

[5] T. M. Cover and A. A. E. Gamal, "Capacity theorems for the relay channel," *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–584, 1979.

[6] P. Gupta and P. R. Kumar, "Towards an information theory of large networks: an achievable rate region," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1877–1894, 2003.

[7] G. Zheng, K. K. Wong, A. Paulraj, and B. Ottersten, "Collaborative-relay beamforming with perfect CSI: optimum and distributed implementation," *IEEE Signal Processing Letters*, vol. 16, no. 4, pp. 257–260.

[8] D. H. N. Nguyen and H. H. Nguyen, "SNR maximization and distributed beamforming in multiuser multi-relay networks," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '09)*, Honolulu, Hawai, USA, December 2009.

[9] D. H. N. Nguyen, H. H. Nguyen, and H. D. Tuan, "Distributed beamforming in relay-assisted multiuser communications," in *Proceedings of the IEEE International Conference on Communications (ICC '09)*, Dresden, Germany, June 2009.

[10] J. Choi, "MMSE-based distributed beamforming in cooperative relay networks," *IEEE Transactions on Communications*, vol. 59, no. 5, pp. 1346–1356, 2011.

[11] N. Khajehnouri and A. H. Sayed, "Distributed MMSE relay strategies for wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3336–3348, 2007.

[12] S. Verdu, *Multiuser Detection*, Cambridge University Press, 1998.

[13] S. Berger and A. Wittneben, "Cooperative distributed multiuser MSE relaying in wireless ad-hoc networks," in *Proceedings of the 39th Asilomar Conference on Signals, Systems and Computers*, pp. 1072–1076, November 2005.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 307467, 8 pages
doi:10.1155/2012/307467

Research Article

Multilayer Orthogonal Beamforming for Priority-Guaranteed Wireless Communications

Jindong Xie, Jun Zhang, and Lin Bai

School of Electronic and Information Engineering, Beihang University, Beijing 100191, China

Correspondence should be addressed to Lin Bai, l.bai@buaa.edu.cn

Received 20 April 2012; Revised 5 July 2012; Accepted 9 July 2012

Academic Editor: Jianhua He

To utilize the benefits of cellular systems, wireless machine-to-machine (M2M) communications over cellular systems are being widely considered. In order to support efficient spectrum sharing between M2M devices and normal mobile users, in the paper, we propose a multilayer orthogonal beamforming (MOBF) scheme for M2M communications over orthogonal frequency division multiple access (OFDMA-) based cellular systems. Using MOBF, each subcarrier in OFDMA systems could be efficiently reused by both normal mobile users and machine-type devices which are organized into multiple virtual layers. The users located in higher layers (e.g., mobile users) are not to be interfered by those in lower layers (e.g., machine devices). To improve the performance, the orthogonal deficiency (OD-) based user selection is carried out, where the intralayer fairness and quasimaximal performance can be guaranteed, simultaneously. Moreover, the signal-to-interference plus noise ratio (SINR) is investigated to measure the performance lower bound of different layers. It is demonstrated by both theoretical and numerical results that the proposed approach provides a stable SINR performance for each layer, that is, the interference free ability from lower level layers.

1. Introduction

Wireless machine-to-machine (M2M) communications over cellular systems are being widely considered as an increasingly important interaction mechanism for M2M applications [1, 2]. In order to utilize the benefits of cellular systems, for example, ubiquitous coverage, the beyond 3rd generation (B3G) and the 4th generation (4G) cellular systems are expected to play crucial roles in the future wireless M2M networks [3, 4]. However, due to different services with quality of service (QoS) are required for different users in cellular systems, that is, human users with mobile terminals and a large number of machine devices, it may not be straightforward to consider M2M communications over cellular systems with limited spectrum resources. More specifically, a higher QoS is needed to support instant communications for mobile users, while the delay tolerance of M2M communications could be measured in a wide range (i.e., from a few milliseconds to several hours) [5]. Thus, it is desired to develop a communication strategy that provides different QoS requirements of humans and machines under limited spectrum resources.

Recent studies have shown that the space division multiple access (SDMA) could be an efficient means to exploit the spectrum efficiency over cellular systems. The SDMA approaches are capable of achieving a much higher capacity and wireless resource usage efficiency [6], which have been proposed for the 3rd generation partnership project long-term evolution (3GPP-LTE) standard [7]. As the optimal SDMA strategy, the dirty paper coding (DPC) is proposed in [8, 9], yet it is difficult to be implemented in practical systems due to the prohibitively high complexity. Thus, various suboptimal beamforming (BF) approaches are proposed to reduce the complexity. It is shown in [10] that the zero-forcing BF (ZFBF) achieves a large fraction of DPC capacity as the number of users approaches infinity. However, good performance of ZFBF cannot be guaranteed with limited number of users and low signal-to-noise ratio (SNR) regime. Thus, orthogonal BF (OBF) is proposed in [11] and further developed in [12] to improve the performance, while a hybrid scheme is studied in [13] to achieve promising performance in various scenarios. In [14], the probability density function (PDF) expressions for the scheduled users and closed-form expression of the first and second scheduled

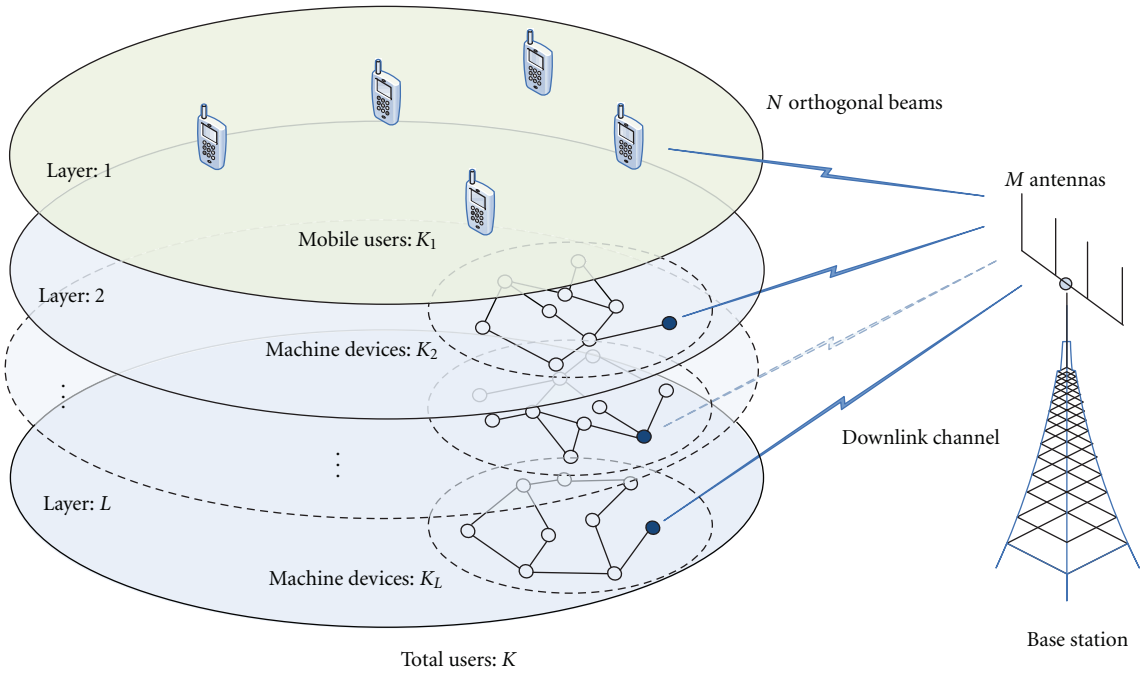


Figure 1: The multilayer SDMA model for wireless M2M communications over cellular systems. Note that users in higher layers should not be interfered by those located in lower layers.

users are derived to theoretically analyze the performance of OBF. Note that conventional methods of BF have been developed to maximize the total sum rate. If different performance requirements are considered for different users, for example, mobile users require a higher QoS to support instant communications which cannot be interfered by machine users, existing BF approaches may not provide a proper solution.

In this paper, to provide flexible performance for different users with diverse priorities, we propose a multilayer orthogonal beamforming (MOBF) for wireless M2M communications over cellular systems under the orthogonal frequency division multiple access (OFDMA) framework. (As the key technology of the 3GPP-LTE mobile broadband standard [15], OFDMA is considered as a background in this paper.) Although different metrics can be carried out to define the physical layer QoS, we consider the signal-to-interference plus noise ratio (SINR) outage probability as the one used in [16]. Therefore, candidate users with various QoS demands could be organized into multiple layers based on required SINR outage probability. By using the proposed MOBF, users in higher layers (i.e., mobile users) enjoy a sufficiently high SINR and will not be interfered from those in lower layers (i.e., machine users). Although users in lower layers may suffer the interference from higher layers, a user selection strategy is carried out to maximize the performance under the orthogonal constraint. It is shown that our proposed approach supports a stable SINR performance for each layer, which is then illustrated by theoretical and numerical results.

The rest of the paper is organized as follows. Section 2 provides the system model. Our proposed MOBF is introduced in Section 3, while its performance is then analyzed in Section 4. After that, numerical results are presented in Section 5. Finally, we conclude this paper in Section 6.

Notation. The superscripts T and H stand for the transpose and Hermitian transpose, respectively. $\mathbf{I}_M$ represents the $M \times M$ identity matrix. Denote by $\|\cdot\|$ and $|\cdot|$ the 2-norm and amplitude of the enclosed complex valued quantity, respectively. $\langle\cdot\rangle$ is the inner product. The statistical expectation is represented by $\mathbb{E}[\cdot]$. The statistical distribution of a circularly symmetric complex Gaussian (CSCG) random variable with mean a and covariance b is denoted by $\mathcal{CN}(a, b)$.

2. System Model

Consider the downlink channel of cellular systems (as shown in Figure 1) with K users requiring L priorities separately (i.e., L virtual layers needed), where each user is equipped with single receive antenna, and let $\mathcal{A} = \{1, 2, \ldots, K\}$ be the index set of all K users. For a given subcarrier (MOBF for one certain subcarrier is emphasized throughout this paper, and the extension method for the whole OFDMA system is straightforward,) the base station (BS) equipped with M transmit antennas simultaneously transmits data to N users selected from $\mathcal{A}$, where $N \leq M$ and denote by $\mathcal{S}$ the index set of N selected users. Since in a typical system, the number of users is larger than the number of transmit antennas, we assume that $M \leq K$ throughout the paper. Thus, the

transmitted signal through *one subcarrier* from BS can be expressed as

$$\mathbf{s} = \sum_{n=1}^{N} \mathbf{w}_n x_n = \mathbf{W}_N \mathbf{x}, \tag{1}$$

where $\mathbf{w}_n \in \mathcal{C}^{M\times 1}$ is the nth beamforming vector (which is also regarded as the nth beam) for the selected user. Let $\mathbf{W}_N = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N]$ be the unitary beamforming matrix with $\mathbf{W}_N^{\mathrm{H}}\mathbf{W}_N = \mathbf{I}_N$ and denote by $\mathbf{x} = [x_1, x_2, \ldots, x_N]^{\mathrm{T}}$ the transmitted symbol vector in one time slot, where (Equal power allocation over scheduled users is considered, which is equivalent to the optimal water-filling method as the users have high SNR.) $\mathbb{E}[|x_n|^2] = P_s$. Assume that the channel gain from the BS to the kth user, denoted by $\mathbf{h}_k \in \mathcal{C}^{M\times 1}$, is independent and identically distributed (i.i.d.) flat Rayleigh fading. Then, the signal received by the nth selected user k_n is given by

$$y_{k_n} = \mathbf{h}_{k_n}^{\mathrm{H}} \mathbf{s} + z_{k_n}, \quad k_n \in \mathcal{S}, \tag{2}$$

where z_{k_n} denotes the additive white Gaussian noise with zero mean and σ^2 variance. According to (2), the SINR of the nth beam observed by the k_nth user becomes

$$\mathrm{SINR}_{k_n,n} = \frac{\left|\mathbf{h}_{k_n}^{\mathrm{H}}\mathbf{w}_n\right|^2 P_s}{\sum_{i\neq n}^{N}\left|\mathbf{h}_{k_n}^{\mathrm{H}}\mathbf{w}_i\right|^2 P_s + \sigma^2}, \quad k_n \in \mathcal{S}. \tag{3}$$

Throughout this paper, the perfect channel state information (CSI) at the base station is assumed. For the sake of simplification, we assume $L = N$ in the paper, which leads to that a single user is considered in each virtual layer at a time. The cases of multiple users selected in one layer are beyond the scope of the paper and can be treated as an extension for future works.

3. Multilayer Orthogonal Beamforming with ODS User Selection

In order to avoid the interference from lower layers to higher layers, in this selection, the MOBF is carried out to provide interference-free ability for users with high priorities. Based on that, an orthogonal deficiency (OD) based user selection strategy is considered to improve the performance.

3.1. Multilayer Orthogonal Beamforming. Letting $\mathcal{S} = \{k_1, \ldots, k_n, \ldots, k_N\}$ and $\mathbf{h}_{k_n}$ be the channel gain of the selected user k_n for the nth layer, then the beamforming vector is given by

$$\mathbf{w}_n = \frac{\widetilde{\mathbf{w}}_n}{||\widetilde{\mathbf{w}}_n||}, \tag{4}$$

where

$$\widetilde{\mathbf{w}}_n = \left[\mathbf{I}_M - \sum_{i=1}^{n-1}\frac{\widetilde{\mathbf{w}}_i\widetilde{\mathbf{w}}_i^{\mathrm{H}}}{||\widetilde{\mathbf{w}}_i||^2}\right]\mathbf{h}_{k_n} \tag{5}$$

$$= \left[\mathbf{I}_M - \mathbf{W}_{n-1}\mathbf{W}_{n-1}^{\mathrm{H}}\right]\mathbf{h}_{k_n}, \tag{6}$$

for $(n = 1, 2, \ldots, N)$. From (5), we can show that $\mathbf{W}_N$ is a matrix whose column vectors are orthogonal to each other. Thus, it can be easily derived that $\mathbf{w}_n$ would be orthogonal to $\mathbf{h}_{k_i}$ when $i < n$, that is, $\langle \mathbf{h}_{k_i}, \mathbf{w}_n\rangle = \mathbf{h}_{k_i}^{\mathrm{H}}\mathbf{w}_n = 0$. In this case, according to the SINR expression in (3), the ith selected user does not receive any interference from the nth one, that is, the user k_n located in the lower layers.

However, $\mathbf{w}_i$ may not be orthogonal to $\mathbf{h}_{k_n}$ as the $\mathbf{w}_i$ is generated without taking $\mathbf{h}_{k_n}$ into consideration when $i < n$. The well-known orthogonal deficiency (OD) [17, 18] can be used as a metric to measure the orthogonality as

$$\xi(k_n, i) = \frac{\left|\mathbf{h}_{k_n}^{\mathrm{H}}\widetilde{\mathbf{w}}_i\right|^2}{||\mathbf{h}_{k_n}||^2||\widetilde{\mathbf{w}}_i||^2} \quad (i < n). \tag{7}$$

Note that two vectors are orthogonal when $\xi(k_n, i) = 0$.

Theorem 1. *Denote by $\xi(k_n, i)$ the* OD *of $\mathbf{h}_{k_n}$ and $\mathbf{w}_i$ for the nth selected user with index k_n, the* SINR *of the user is given by*

$$\mathrm{SINR}_{k_n} = \frac{1 - \sum_{i=1}^{n-1}\xi(k_n, i)}{\sum_{i=1}^{n-1}\xi(k_n, i) + \left(1/\gamma_{k_n}\right)}, \quad n \le N, \tag{8}$$

where $\gamma_{k_n} = (\|\mathbf{h}_{k_n}\|^2 P_s/\sigma^2)$ denotes the corresponding SNR.

Proof. Assuming $n < j \le N$, it can be easily obtained that $\langle \mathbf{h}_{k_n}, \mathbf{w}_j\rangle = 0$. Therefore, we reshape (3) as

$$\begin{aligned}\mathrm{SINR}_{k_n} &= \frac{\left|\mathbf{h}_{k_n}^{\mathrm{H}}\mathbf{w}_n\right|^2 P_s}{\sum_{i=1}^{n-1}\left|\mathbf{h}_{k_n}^{\mathrm{H}}\mathbf{w}_i\right|^2 P_s + \sigma^2} \\ &= \frac{\left(\left|\mathbf{h}_{k_n}^{\mathrm{H}}\widetilde{\mathbf{w}}_n\right|^2/||\widetilde{\mathbf{w}}_n||^2\right)}{\sum_{i=1}^{n-1}\left(\left|\mathbf{h}_{k_n}^{\mathrm{H}}\widetilde{\mathbf{w}}_i\right|^2/||\widetilde{\mathbf{w}}_i||^2\right) + (\sigma^2/P_s)}, \quad n \le N.\end{aligned} \tag{9}$$

According to (5), we can show

$$\begin{aligned}\left|\mathbf{h}_{k_n}^{\mathrm{H}}\widetilde{\mathbf{w}}_n\right|^2 &= \left|||\mathbf{h}_{k_n}||^2 - \sum_{i=1}^{n-1}\frac{\mathbf{h}_{k_n}^{\mathrm{H}}\widetilde{\mathbf{w}}_i\widetilde{\mathbf{w}}_i^{\mathrm{H}}\mathbf{h}_{k_n}}{||\widetilde{\mathbf{w}}_i||^2}\right|^2 \\ &= \left|||\mathbf{h}_{k_n}||^2 - \sum_{i=1}^{n-1}\frac{\left|\mathbf{h}_{k_n}^{\mathrm{H}}\widetilde{\mathbf{w}}_i\right|^2}{||\widetilde{\mathbf{w}}_i||^2}\right|^2, \\ ||\widetilde{\mathbf{w}}_n||^2 &= \mathbf{h}_{k_n}^{\mathrm{H}}\left(\mathbf{I}_M - \sum_{i=1}^{n-1}\frac{\widetilde{\mathbf{w}}_i^{\mathrm{H}}\widetilde{\mathbf{w}}_i}{||\widetilde{\mathbf{w}}_i||^2}\right)^2\mathbf{h}_{k_n} \\ &= ||\mathbf{h}_{k_n}||^2 - \sum_{i=1}^{n-1}\frac{\left|\mathbf{h}_{k_n}^{\mathrm{H}}\widetilde{\mathbf{w}}_i\right|^2}{||\widetilde{\mathbf{w}}_i||^2}.\end{aligned} \tag{10}$$

Since $\|\widetilde{\mathbf{w}}_n\|^2 \ge 0$, we have

$$\begin{aligned}\mathrm{SINR}_{k_n} &= \frac{||\mathbf{h}_{k_n}||^2 - \sum_{i=1}^{n-1}\left(\left|\mathbf{h}_{k_n}^{\mathrm{H}}\widetilde{\mathbf{w}}_i\right|^2/||\widetilde{\mathbf{w}}_i||^2\right)}{\sum_{i=1}^{n-1}\left(\left|\mathbf{h}_{k_n}^{\mathrm{H}}\widetilde{\mathbf{w}}_i\right|^2/||\widetilde{\mathbf{w}}_i||^2\right) + (\sigma^2/P_s)} \\ &= \frac{1 - \sum_{i=1}^{n-1}\xi(k_n, i)}{\sum_{i=1}^{n-1}\xi(k_n, i) + \left(1/\gamma_{k_n}\right)}, \quad n \le M.\end{aligned} \tag{11}$$

This completes the proof. □

It is noteworthy that as $\gamma_{k_n} \gg 1$, the SINR of user k_n at the nth orthogonal beam can be approximated by its signal-to-interference ratio (SIR), that is,

$$\mathrm{SINR}_{k_n} \approx \mathrm{SIR}_{k_n} = \frac{1 - \sum_{i=1}^{n-1} \xi(k_n, i)}{\sum_{i=1}^{n-1} \xi(k_n, i)}, \tag{12}$$

and upper bounded by

$$\begin{aligned} \mathrm{SINR}_{k_n} &\le \min\{\mathrm{SIR}_{k_n}, \mathrm{SNR}_{k_n}\} \\ &= \min\left\{\frac{1 - \sum_{i=1}^{n-1} \xi(k_n, i)}{\sum_{i=1}^{n-1} \xi(k_n, i)}, \gamma_{k_n}\right\}. \end{aligned} \tag{13}$$

In general, $\xi(k_n, i) \neq 0$, as $\mathbf{h}_{k_n}$ and the elements of $\mathbf{W}_{n-1}$ are independent. As shown in (12), SINR_{k_n} is more likely to decrease with n increases. Thus, the user with a small n (at high layer) enjoys a higher SINR thanks to the MOBF, where a high QoS requirement could be guaranteed.

However, the user with a large n (at low layer) suffers the interference from those with small n because of OD existing. Note that different channel conditions of users result in different ODs, which may lead to quite different performances of the MOBF. Thus, OD-based user selection plays a crucial role to minimize the interference and meanwhile maximize the system performance.

3.2. OD-Based User Selection Criterion. Let $\mathcal{A}_n$ be the subset of candidate users with the same SINR requirement for layer n, where the user number of subset is K_n, $n \in \{1, 2, \ldots, N\}$. Note that $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_N$ and $K = K_1 + K_2 + \cdots + K_N$, where "$\cup$" denotes the set union. In order to maximize the system performance, the user who has the maximum achievable rate can be chosen as

$$\begin{aligned} k_n &= \underset{k \in \mathcal{A}_n}{\operatorname{argmax}} \log_2(1 + \mathrm{SINR}_k) \\ &= \underset{k \in \mathcal{A}_n}{\operatorname{argmax}} \mathrm{SINR}_k. \end{aligned} \tag{14}$$

According to (12) and (14), when γ_k is sufficiently large, the SINR-based user selection criterion is approximately identical to choose the user who has the smallest sum of ODs. Hence, the resulting selection scheme is regarded as the orthogonal deficiency sequential (ODS) user selection, which is summarized as follows.

(1) Let $n = 1$. According to Theorem 1, the user to be selected in the first layer would not be interfered by other users. Thus, the first user can be selected by

$$k_1 = \underset{k \in \mathcal{A}_1}{\operatorname{argmax}} \gamma_k. \tag{15}$$

Once the index of the first user is determined, we update $\mathcal{S} = \{k_1\}$. In addition, we let $\mathbf{w}_1 = \mathbf{h}_{k_1}/\|\mathbf{h}_{k_1}\|$.

(2) Let $n = n + 1$. The nth user is chosen as

$$\begin{aligned} k_n &= \underset{k \in \mathcal{A}_n}{\operatorname{argmax}} \frac{1 - \sum_{i=1}^{n-1} \xi(k_n, i)}{\sum_{i=1}^{n-1} \xi(k_n, i) + (1/\gamma_{k_n})} \\ &\approx \underset{k \in \mathcal{A}_n}{\operatorname{argmin}} \sum_{i=1}^{n-1} \xi(k, i). \end{aligned} \tag{16}$$

Once the nth user is determined, we update the user subset $\mathcal{S} = \mathcal{S} \cup \{k_n\}$. After that, we generate $\mathbf{w}_n$ using the proposed MOBF in (4).

(3) If $n = N$, $\mathcal{S} = \{k_1, \ldots, k_N\}$ and stop. Otherwise, go back to step (2).

As N users are selected, the achievable rate for the nth layer is given by

$$R_n = \log_2(1 + \mathrm{SINR}_{k_n}), \quad k_n \in \mathcal{S}. \tag{17}$$

And the sum rate for the given subcarrier is carried out as

$$R_{\mathrm{sum}} = \sum_{n=1}^{N} \log_2(1 + \mathrm{SINR}_{k_n}), \quad k_n \in \mathcal{S}. \tag{18}$$

In comparison with SINR-based selection criteria, whose performances are highly dependent on SNR (which may have fairness problem due to the SNR is close related to the distance between BS and users [6]), our proposed ODS user selection could provide an equal chance for each candidate user with the same priority even in the near-far environment. Note that, the ODS user selection neglects the effect of channel gain for the fairness requirement, which means only quasimaximal performance can be guaranteed. However, the performance loss is negligible, when the system is interference limited and the SNR is sufficient high.

4. Performance Analysis

In this section, the theoretical analysis of our proposed MOBF is carried out. The outage probability of the threshold SINR for each layer is derived to measure the lower bound of the performance. It is assumed that the elements of channel vectors $\mathbf{h}_k, k \in \mathcal{A}$, are i.i.d. CSCG random variables distributed as $\mathcal{CN}(0, 1/M)$.

4.1. The Distribution of OD. Suppose that $\mathbf{a}$ and $\mathbf{b}$ are two independent $m \times 1$ complex-valued Gaussian random vectors. Then, we have

$$\cos^2\theta = \frac{|\mathbf{a}^{\mathrm{H}}\mathbf{b}|^2}{\|\mathbf{a}\|^2 \|\mathbf{b}\|^2} \sim \mathrm{Beta}(1, m-1), \tag{19}$$

where θ is the angle between $\mathbf{a}$ and $\mathbf{b}$, and Beta (α, β) denotes the Beta distribution [19].

Lemma 2. *$\xi(k_n, 1), \ldots, \xi(k_n, n-1)$ are i.i.d. random variables and follow the Beta distribution.*

Proof. According to (5), all the column vectors of $\mathbf{W}_{n-1}$, that is, $\mathbf{w}_i$ $(i = 1, 2, \ldots, n-1)$, are unitary Gaussian random vectors, as the orthogonalization operation is a linear transformation and $\mathbf{h}_{k_n}$ is a Gaussian random vector.

Besides, it can be derived from (5) that all $\mathbf{w}_i$ of $\mathbf{W}_{n-1}$ are uncorrelated with each other. Therefore, they are independent due to equivalence between the uncorrelated and the independent for Gaussian random variables.

Using the above results, different $\xi(k_n, i)$ in (7) are independent of each other, since (7) is the function of $\mathbf{h}_{k_n}$ and $\mathbf{w}_i$, when $i < n$. Thus, according to (19), $\xi(k_n, 1), \ldots, \xi(k_n, n-1)$ are i.i.d. random variables and follow the Beta distribution, where

$$\xi(k_n, i) \sim \text{Beta}\,(1, M-1). \tag{20}$$

This completes the proof. □

Using Lemma 2, the PDF and the cumulative distribution function (CDF) of $\xi(k_n, i)$ are given by

$$\xi(k_n, i) \sim f_\xi(x) = (M-1)(1-x)^{M-2}, \quad 0 \le x \le 1, \tag{21}$$

$$F_\xi(x) = 1 - (1-x)^{M-1}, \quad 0 \le x \le 1, \tag{22}$$

respectively.

4.2. SINR Outage Probability. In this section, the SINR performance of the MOBF with ODS is analyzed by using the distribution of OD. Since it is not easy to find the PDF and CDF of SINR_{k_n} in (8), we consider SIR as the approximation of SINR in (12) and the upper bound in (13).

Letting Γ be the threshold SINR, the outage probability of SINR in (12) for a random user $k \in \mathcal{A}_n$ is approximately obtained as

$$\begin{aligned}\Pr(\text{SINR} \le \Gamma) &\approx \Pr\left(\frac{1 - \sum_{i=1}^{n-1} \xi(k,i)}{\sum_{i=1}^{n-1} \xi(k,i)} \le \Gamma\right) \\ &= 1 - \Pr\left(\sum_{i=1}^{n-1} \xi(k,i) \le \frac{1}{1+\Gamma}\right),\end{aligned} \tag{23}$$

where $0 < \Gamma < +\infty$.

In order to deal with the sum of ODs, we can prove the following lemma.

Lemma 3. *Assume $\zeta_1, \zeta_2, \ldots, \zeta_n$ are i.i.d random variables. Let $f_{\zeta_i}(x_i)$ and $F_{\zeta_i}(x_i)$ be the PDF and CDF of ζ_i, respectively, $i = 1, \ldots, n$. The CDF of $\eta_n = \sum_{i=1}^{n} \zeta_i$ is given by*

$$\begin{aligned}F_{\eta_n}(x) &= F_{\zeta_1} * F_{\zeta_2} * \cdots * F_{\zeta_n} \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\zeta_1}(x_1) f_{\zeta_2}(x_2) \cdots f_{\zeta_{n-1}}(x_{n-1}) \\ &\quad \cdot F_{\zeta_n}(x - x_1 - x_2 \cdots - x_{n-1}) \\ &\quad \cdot dx_{n-1} \cdots dx_2 dx_1,\end{aligned} \tag{24}$$

where "$*$" represents the convolution operation.

Proof. See Appendix. □

Using Lemma 3 together with (21) and (22), the CDF of SIR for a random user in the nth layer can be rewritten as

$$\begin{aligned}F(\Gamma) = 1 - \int_0^x \int_0^{x-t_1} \cdots \int_0^{x-t_{n-3}} & f_\xi(x_1) f_\xi(x_2) \cdots f_\xi(x_{n-2}) \\ & \cdot F_\xi(x - t_{n-2}) dx_{n-2} \cdots dx_2 dx_1,\end{aligned} \tag{25}$$

where $x = 1/(1+\Gamma)$ and t_n denotes the sum of x_i, that is, $t_n = \sum_{i=1}^{n} x_i$.

Since (25) is complicated, it is desired to have a concise expression for analysis. Supposing that $\delta = 1/((n-1)(1+\Gamma))$, and $n > 1$, we have

$$\Pr(\xi_1 \le \delta, \xi_2 \le \delta, \ldots, \xi_{n-1} \le \delta) \le \Pr\left(\sum_{i=1}^{n-1} \xi_i \le \frac{1}{1+\Gamma}\right), \tag{26}$$

where $\xi_i = \xi(k, i)$, $k \in \mathcal{A}_n$ for notational simplicity. Since $\xi_1, \xi_2, \ldots, \xi_{n-1}$ are i.i.d. random variables, we have

$$\begin{aligned}&\Pr(\xi_1 \le \delta, \xi_2 \le \delta, \ldots, \xi_{n-1} \le \delta) \\ &\quad = \Pr(\xi_1 \le \delta) \cdot \Pr(\xi_2 \le \delta) \cdots \Pr(\xi_{n-1} \le \delta) \\ &\quad = \Pr^{(n-1)}(\xi_i \le \delta).\end{aligned} \tag{27}$$

Then, the CDF of SIR in (25) for a random user in the nth layer is approximated by

$$\begin{aligned}F(\Gamma) &\approx \widetilde{F}(\Gamma) \\ &= 1 - \Pr^{(n-1)}\left(\xi_i \le \frac{1}{(n-1)(1+\Gamma)}\right) \\ &= 1 - \left(1 - \left(1 - \frac{1}{(n-1)(1+\Gamma)}\right)^{M-1}\right)^{n-1},\end{aligned} \tag{28}$$

where $0 < \Gamma < \infty$ and $n > 1$.

Using order statistics [20], the selected user has the maximum SIR among user subset $\mathcal{A}_n$, which can be derived by

$$\begin{aligned}F_{\max}(\Gamma) &= \Pr\left(\max_{k_n \in \mathcal{A}_n} \text{SIR}_{k_n} \le \Gamma\right) \\ &= F^{K_n}(\Gamma) \approx \widetilde{F}^{K_n}(\Gamma).\end{aligned} \tag{29}$$

By taking into account of the SNR, from (13), the SINR of the selected user k_n based on ODS is upper bounded by

$$\text{SINR}_{k_n} \le \min\left\{\max_{k \in \mathcal{A}_n} \text{SIR}_k, \text{SNR}_k\right\}. \tag{30}$$

Thus, the lower bound of SINR outage probability for the selected user in the nth layer is given by

$$\begin{aligned}&\Pr(\text{SINR}_{k_n} \le \Gamma) \\ &\quad \ge P_{\text{LB}}(\Gamma) \\ &\quad = 1 - (1 - F_{\max}(\Gamma))\left(1 - F_{\chi(2M)}\left(\frac{2M\sigma^2\Gamma}{Ps}\right)\right),\end{aligned} \tag{31}$$

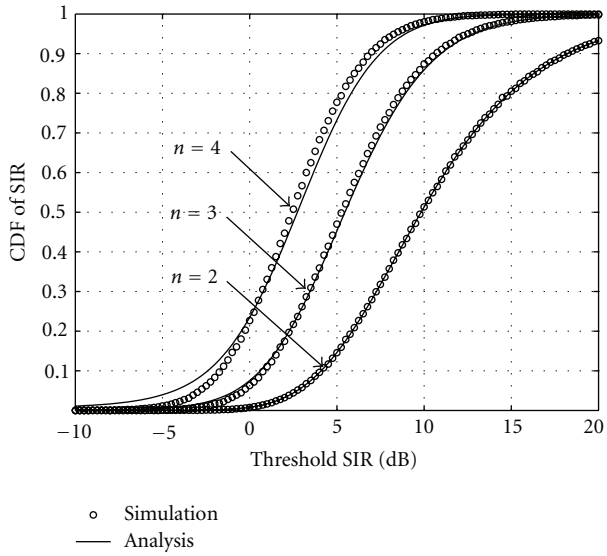


Figure 2: The CDF of SIR for a random user in the nth layer in a noise-free scenario with $M = 8$ antennas and $N = 4$ layers.

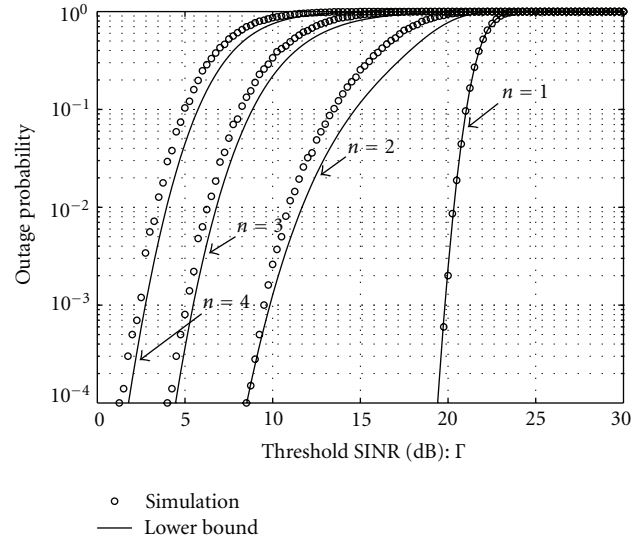


Figure 3: Comparison of the SINR outage probability simulation and lower bound with subset user number $K_n = 10$, $M = 8$ antennas, $N = 4$ layers, and total transmit power $P_s = 20$ dB.

and the approximation is

$$P_{\mathrm{LB}}(\Gamma) \approx \widetilde{P}_{\mathrm{LB}}(\Gamma) = 1 - \left(1 - \widetilde{F}^{K_n}(\Gamma)\right)\left(1 - F_{\chi(2M)}\left(\frac{2M\sigma^2\Gamma}{Ps}\right)\right), \tag{32}$$

where $F_{\chi(n)}$ denotes the chi-square CDF with n degrees of freedom.

4.3. Achievable Rate Analysis. With the lower bound of SINR outage probability, the achievable rate can be expressed with a given Γ. Then, the lower bound of average achievable rate using the outage probability in (31) is shown as

$$\begin{aligned} R_n &\geq \Pr(\mathrm{SINR}_{k_n} \leq \Gamma) \times 0 \\ &\quad + \Pr(\mathrm{SINR}_{k_n} > \Gamma)\log_2(1+\Gamma) \\ &= (1 - P_{\mathrm{LB}}(\Gamma))\log_2(1+\Gamma). \end{aligned} \tag{33}$$

A tight lower bound on R_n can be achieved using integration method with respect to Γ as

$$R_n = \int_0^{+\infty} \log_2(1+\Gamma)\, d(P_{\mathrm{LB}}(\Gamma)) \tag{34}$$

$$\geq \sum_{\gamma \in \hat{\Gamma}} \left(\widetilde{P}_{\mathrm{LB}}\left(\gamma + \delta_\gamma\right) - \widetilde{P}_{\mathrm{LB}}(\gamma)\right)\log_2(1+\gamma), \tag{35}$$

where $\hat{\Gamma}$ denotes the practical cutoff limit and δ_γ is the quantification step. Note that a smaller δ_γ provides a more accurate approximation.

5. Numerical Results

In this section, numerical results are presented where we evaluate the priority-guaranteed ability of the proposed

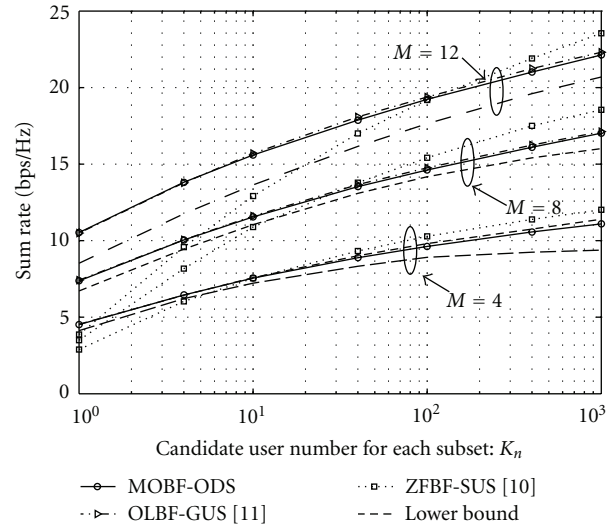


Figure 4: Sum rate performance comparison of MOBF with $M = 4, 8, 12$ antennas, $N = M$ layers, and total transmit power $P_s = 5$ dB. Equal number of candidate users K_n is used ranging from 1 to 1000.

MOBF scheme. We consider homogeneous users in a cell with an M-antenna base station, that is, user channels are independent unit-variance complex Gaussian vectors. Then, the squared channel magnitude $\|\mathbf{h}_k\|^2$ follows a chi-square distribution with 2 M degrees of freedom. Additionally, in the simulation of Figures 2–5, we assume that all users experience the same pass-loss condition, while shadow fading is not considered. Moreover, without loss of generality, the number of candidate users K_n in each subset $\mathcal{A}_n$ is set to be the same. Thus, the total number of candidate users can be calculated by $K = K_n \times N$. Furthermore, the noise variance $\sigma^2 = 1$ is assumed for simplicity.

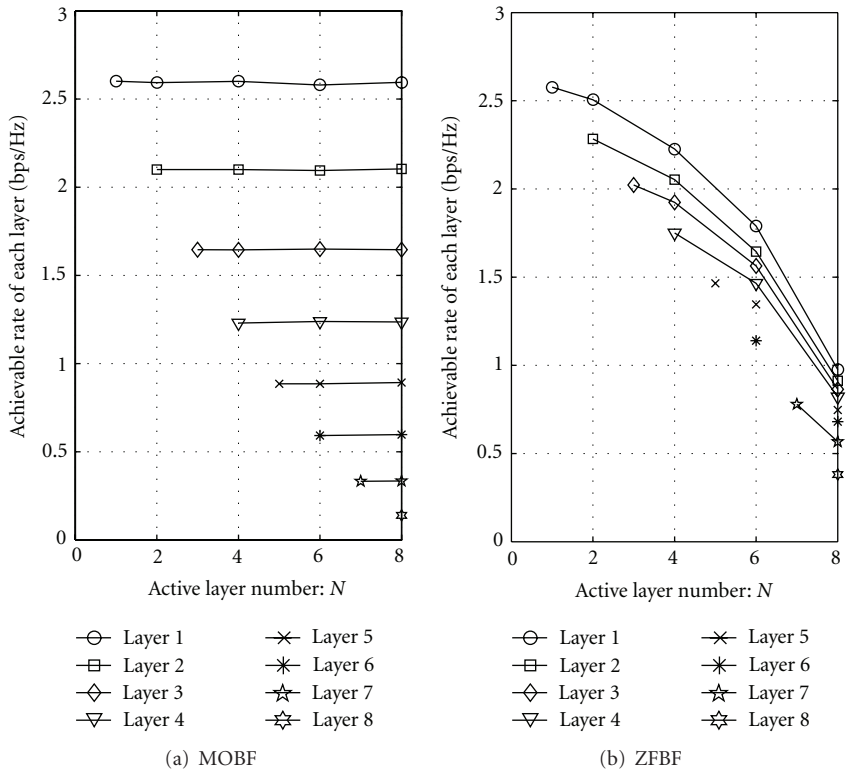


Figure 5: Achievable layer rates of MOBF and ZFBF as a function of active layers number N, that is, the total number of simultaneous beams, where $M = 8$ antennas, $K_n = 10$ candidate users for each subset $\mathcal{A}_n$, and transmit power $P_s = 5$ dB, are considered.

In Figure 2, the OD-based analytical CDF of SIR in (25) and the simulation results obtained in a noise-free scenario are compared. Since the first layer is an interference-free layer according to the MOBF, the SIR outage probability comparison is obtained with the parameters $n = 2, 3$, and 4. It is shown from Figure 2 that the analytical and simulated results exhibit the same behavior.

Figure 3 shows the outage probabilities versus threshold SINR among different layers of MOBF (i.e., $n = 1, 2, 3$, and 4), where numerical results and the lower bound derived in (31) are compared. Through simulations, $N = 4$, $K_n = 10$, and $M = 8$ are considered. It shows that the user with a smaller n (in a higher layer) has lower outage probability since the impact of interference becomes smaller. In particularly, the user in the first layer won't be interfered by those in lower layers, therefore the first layer SINR is reduced to SNR which could be described as the chi-square distribution. Furthermore, from Figure 3, we can also show that the derived lower bound is able to provide an accurate approximation. Note that the total number of candidate users is $K_n \times N$.

Simulation results of sum rate versus number K_n of each user subset are presented in Figure 4, where $N = M$ for the cases of $M = 4$, 8, and 12 antennas are considered, respectively. It shows that the sum rate of the system increases with more candidate users (i.e., K_n) and transmitting antennas (i.e., M). Besides, we can confirm that the appropriate theoretical lower bound in (35) with $\delta_\gamma = 0.1$ dB and $\widetilde{\Gamma} = [-10\,\text{dB}, 20\,\text{dB}]$ is reasonable accurate, even though a large K_n is considered. In particular, as K_n increases, the sum rates of all schemes increase because of more multi-user diversity.

Additionally, ZFBF with semiorthogonal user selection (SUS) [10] and OBF with greedy user selection algorithm-B (GUS-B) [11] are also considered for performance comparison in Figure 4. It can be observed that both MOBF and OBF outperform ZFBF at the regime with fewer K_n and relatively low SNR, which is more practical for the power-limited M2M networks. Besides, though marginal sum rate gain can be obtained by GUS-B over OBF, a complexity order of $\mathcal{O}(K^2)$ is required, which is higher compared to the complexity order of $\mathcal{O}(K)$ required by the proposed ODS.

In Figure 5, we compare the performance of our proposed MOBF and the ZFBF-SUS in [10] in terms of the achievable rate versus N, where $M = 8$ and total power allocation $P_s = 5$ dB are considered. It shows that the ZFBF provides descending performances as N increases, since the effective channel gain decreases dramatically because of the interlayer interference [10]. On the contrary, our proposed MOBF provides stable performances of all layers even when $N = M$, because the MOBF layer rates are only affected by higher layers. Moreover, it is confirmed that the user with a smaller n (in a higher layer) enjoys a higher performance by using our proposed approach.

6. Conclusion

In this paper, the MOBF strategy for wireless M2M communications over cellular systems is proposed together with

an ODS user selection criterion. It has been shown that users with high priorities enjoy high SINR due to the use of MOBF, since users in higher layers won't be interfered by those in lower layers. On the other hand, the SINR of users with low priorities can also be maximized using ODS user selection. The performance of different layers has been then analyzed using theoretical tools. Through analysis and numerical results, it has been shown that our proposed scheme provides more stable performance for different users compared to the existing approach. Thus, for wireless M2M communications over cellular systems, the MOBF can be regarded as a stable mechanism for both mobile and machine users.

Appendix

Proof. The proof of lemma 3 is shown as follows. Let $f_{\zeta_i}(x_i)$ and $F_{\zeta_i}(x_i)$ be the PDF and CDF of the random variable ζ_i, respectively.

First, if $n = 2$, ζ_1 and ζ_2 are i.i.d. random variables, the CDF of $\eta_2 = \zeta_1 + \zeta_2$ is given by

$$\begin{aligned} F_{\eta_2}(x) &= F_{\zeta_1}(x_1) * F_{\zeta_2}(x_2) \\ &= \int_{-\infty}^{+\infty} f_{\zeta_1}(x_1) F_{\zeta_2}(x - x_1) dx_1, \end{aligned} \tag{A.1}$$

which is the precise expression of (24) with $n = 2$. Provided that (24) is established for $n = k$ $(k \geq 2)$, we have

$$\begin{aligned} F_{\eta_k}(x) &= F_{\zeta_1} * F_{\zeta_2} * \cdots * F_{\zeta_k} \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\zeta_1}(x_1) f_{\zeta_2}(x_2) \cdots f_{\zeta_{k-1}}(x_{k-1}) \\ &\quad \cdot F_{\zeta_k}(x - x_1 - \cdots - x_{k-1}) \\ &\quad \cdot dx_{k-1} \cdots dx_2 dx_1. \end{aligned} \tag{A.2}$$

When $n = k + 1$, we can show that

$$\begin{aligned} F_{\eta_{k+1}}(x) &= F_{\zeta_1} * F_{\zeta_2} * \cdots * F_{\zeta_k} * F_{\zeta_{k+1}} \\ &= F_{\zeta_1} * \left(F_{\zeta_2} * \cdots * F_{\zeta_k} * F_{\zeta_{k+1}}\right) \\ &= \int_{-\infty}^{+\infty} f_{\zeta_1}(x_1) \cdot F_{\eta_{k \leftarrow k+1}}(x - x_1) dx_1 \\ &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_{\zeta_1}(x_1) f_{\zeta_2}(x_2) \cdots f_{\zeta_k}(x_k) \\ &\quad \cdot F_{\zeta_{k+1}}(x - x_1 - x_2 \cdots - x_k) \\ &\quad \cdot dx_k \cdots dx_2 dx_1. \end{aligned} \tag{A.3}$$

This completes the proof of lemma 3. □

References

[1] G. Lawton, "Machine-to-machine technology gears up for growth," *Computer*, vol. 37, no. 9, pp. 12–15, 2004.

[2] A. Alheraish, "Design and implementation of home automation system," *IEEE Transactions on Consumer Electronics*, vol. 50, no. 4, pp. 1087–1092, 2004.

[3] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. D. Johnson, "M2M: from mobile to embedded internet," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 36–43, 2011.

[4] S. Y. Lien and K. C. Chen, "Massive access management for QoS guarantees in 3GPP machine-to-machine communications," *IEEE Communications Letters*, vol. 15, no. 3, pp. 311–313, 2011.

[5] A. S. Lioumpas and A. Alexiou, "Uplink scheduling for machine-to-machine communications in LTE-based cellular systems," in *Proceedings of IEEE Global Telecommunications Conference Workshops*, pp. 353–357, December 2011.

[6] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1912–1921, 2003.

[7] 3GPP TR 25.814, *Physical Layers Aspects for Evolved Universal Terrestrial Radio Access (Rel. 7)*, June 2006.

[8] M. H. M. Costa, "Writing on Dirty Paper," *IEEE Transactions on Information Theory*, vol. IT-29, no. 3, pp. 439–441, 1983.

[9] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[10] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, 2006.

[11] R. De Francisco, M. Kountouris, D. T. M. Slock, and D. Gesbert, "Orthogonal linear beamforming in MIMO broadcast channels," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC '07)*, pp. 1211–1216, March 2007.

[12] J. Duplicy, D. P. Palomar, and L. Vandendorpe, "Adaptive orthogonal beamforming for the MIMO broadcast channel," in *Proceedings of the 2nd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMPSAP '07)*, pp. 77–80, December 2007.

[13] C. Zhang, W. Xu, and M. Chen, "Hybrid zero-forcing beamforming/orthogonal beamforming with user selection for MIMO broadcast channels," *IEEE Communications Letters*, vol. 13, no. 1, pp. 10–12, 2009.

[14] S. Ozyurt and M. Torlak, "Performance analysis of orthogonal beamforming with user selection in MIMO broadcast channels," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '11)*, pp. 1–5, Houston, Tex, USA, 2011.

[15] 3GPP TR 36.211, *Evolved Universal Terrestrial Radio Access (EUTRA); Physical Channels and Modulation*, March 2010.

[16] R. Zhang, C. C. Chai, and Y. C. Liang, "Joint beamforming and power control for multiantenna relay broadcast channel with QoS constraints," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 726–737, 2009.

[17] X. Ma and W. Zhang, "Fundamental limits of linear equalizers: diversity, capacity, and complexity," *IEEE Transactions on Information Theory*, vol. 54, no. 8, pp. 3442–3456, 2008.

[18] J. Choi, "Opportunistic beamforming with single beamforming matrix for virtual antenna arrays," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 3, pp. 872–881, 2011.

[19] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, New York, NY ,USA, 1982.

[20] H. A. David and H. N. Nagaraja, *Order Statistics*, John Wiley & Sons, New York, NY, USA, 3rd edition, 2003.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 940302, 11 pages
doi:10.1155/2012/940302

Research Article

Investigation of Uncoordinated Coexisting IEEE 802.15.4 Networks with Sleep Mode for Machine-to-Machine Communications

Chao Ma,[1] Jianhua He,[1] Zuoyin Tang,[1] Wenyang Guan,[2] and Yue Li[2]

[1] *School of Engineering and Applied Science, Aston University, Birmingham B47ET, UK*
[2] *College of Engineering, Swansea University, Wales SA28PP, UK*

Correspondence should be addressed to Wenyang Guan, 440501@swansea.ac.uk

Received 20 April 2012; Revised 15 June 2012; Accepted 25 June 2012

Academic Editor: Lin Bai

The low-energy consumption of IEEE 802.15.4 networks makes it a strong candidate for machine-to-machine (M2M) communications. As multiple M2M applications with 802.15.4 networks may be deployed closely and independently in residential or enterprise areas, supporting reliable and timely M2M communications can be a big challenge especially when potential hidden terminals appear. In this paper, we investigate two scenarios of 802.15.4 network-based M2M communication. An analytic model is proposed to understand the performance of uncoordinated coexisting 802.15.4 networks. Sleep mode operations of the networks are taken into account. Simulations verified the analytic model. It is observed that reducing sleep time and overlap ratio can increase the performance of M2M communications. When the networks are uncoordinated, reducing the overlap ratio can effectively improve the network performance.

1. Introduction

The increasingly popular M2M technology can enable machines to communicate directly with one another through wireless and/or wired system [1, 2]. With the increasing volume of wireless networks in recent years, the information can be exchanged between machine devices much easier and faster at low cost, which makes M2M technology more attractive to business as well as customers due to its huge potential on cost reduction and services improvement [3, 4]. Effective and reliable support from wireless networks for communications between the number of M2M devices is pivotal to the success of M2M technology. This paper aims to study the effectiveness of IEEE 802.15.4 network technology on support of M2M communications [5]. With increasing number of M2M devices which may be connected by IEEE 802.15.4 network, it is very likely to witness that multiple 802.15.4 networks are closely and independently deployed for M2M appellations. The interference of each other and hidden terminal problem may arise among these M2M devices, which can become severe problems for 802.15.4 networks-based M2M applications.

In this paper we investigate the issues of multiple uncoordinated coexisting 802.15.4 networks when they are closely deployed with sleep mode. Two representative network scenarios where two 802.15.4 networks are working in the sleep mode with different overlap ratios in the channel access periods are presented. We propose an analytic model to predict the system throughput and energy consumption with different media access control (MAC) parameters, frame length, the number of network devices for each network, and overlap ratios. The results show that the impact of uncoordinated multiple networks operations can lead to significant system performance drop when their channel access periods are most overlapped. The approaches of reducing sleep time and overlap ratio to increase the system performance are also studied. The analytic model is verified by simulations.

In the literatures, simulation-based evaluation of single 802.15.4 network has been widely reported including [6, 7]. Additionally, many analytic models have been proposed to capture the throughput and energy consumption performance of single 802.15.4 network with either saturated or

unsaturated traffic. The limited scalability of the 802.15.4 MAC was pointed out by Yedavalli and Krishnamachari [8] where the performance in terms of throughput and energy consumption are studied. They showed that 802.15.4 MAC performed poorly when the number of contending devices was high. Mišić et al. proposed a Markov model to evaluate the throughput of 802.15.4 networks with unsaturated downlink and uplink traffic [9]. However, their analytical models do not have high accuracy. A simplified Markov model was proposed in [10], in which a geometric distribution was used to approximate the uniform distribution for the random backoff counter. But the approximation results in large inaccuracy in throughput prediction. A three-dimensional Markov model was proposed in [11] to evaluate the throughput of slotted carrier sense multiple access (CSMA). However, the state transitions in [11] were not correctly modelled. The model was revised with improved accuracy in [12]. Both Pollin et al. [13] and Singh et al. [14] also considered a star network topology and analysed the MAC protocol performance under assumption that saturated traffic conditions. They found out that a large fraction of packets was dropped during the channel access, and the dropping probability increases with the number of sensor devices. Park et al. [15] and He et al. [16] developed accurate analytical modes for 802.15.4 MAC protocol in the beacon-enabled mode for star network topology. In addition, the performance analysis was mainly targeted at validating the accuracy of the proposed model. Energy consumption and throughput performance of 802.15.4 MAC was analysed in [17]. However, none of these works has studied the uncoordinated operations of multiple coexisting 802.15.4 networks with sleep mode.

The remainder of this paper is organised as follows. In Section 2, we introduce the superframe structure and channel access algorithm of IEEE 802.15.4 standard briefly. Two scenarios and model assumption of the analytic model are introduced in Section 2 and the analytic model of two uncoordinated networks is presented in Section 3. Section 4 discusses the numerical results and performance analysis. The conclusions and future works are in Section 5.

2. Channel Access Algorithm of IEEE 802.15.4

An IEEE 802.15.4 network can work either in nonbeacon-enabled or in beacon-enabled mode [5]. In the nonbeacon-enabled mode, there are no regular beacons, and devices communicate with each other using unslotted CSMA-CA algorithm. In the beacon-enabled mode, the coordinator transmits regular beacons for synchronisation and association procedures to control communication. A superframe structure is imposed in the beacon-enabled mode as shown in Figure 1, whose format is defined by the coordinator. The superframe is bounded by network regular beacons, and can have an active portion and an optional inactive portion. All communications take place in the active period while devices are allowed to enter a low-power (sleep) mode during the inactive period. The structure of this superframe is described by the values of *macBeaconOrder* (*BO*) and *macSuperframeOrder* (*SO*). The BO describes the beacon interval (BI), the SO describes the length of superframe duration (SD), and they are related, respectively, as follow:

$$\begin{aligned} \mathrm{BI} &= \mathrm{aBaseSuperframeDuration} \times 2^{\mathrm{BO}}, \\ \mathrm{SD} &= \mathrm{aBaseSuperframeDuration} \times 2^{\mathrm{SO}}, \end{aligned} \tag{1}$$

where aBaseSuperframeDuration = 960 symbols and $0 \leq \mathrm{SO} \leq \mathrm{BO} \leq 14$.

The active portion of each superframe shall be composed of two parts: a contention access period (CAP) and an optional contention-free period (CFP). The CAP shall start immediately following the beacon and complete before the beginning of CFP on a superframe slot boundary. If the CFP is zero length, the CAP shall complete at the end of the active portion of the superframe. We neglect every CFP in this study, because it is designed for low-latency applications requiring specific data bandwidth which means only CAP in each active portion of superframe structure. In CAP, communication among devices uses slotted CSMA-CA algorithm for contention access. The 802.15.4 slotted CSMA-CA algorithm operates in unit of backoff slot. One backoff slot has the length of 20 symbols. In the rest of the paper backoff slot is simply called slot unless otherwise specified.

According to the acknowledgement (ACK) of successful reception of a data frame, the slotted CSMA-CA algorithm can be operated in two modes: ACK mode, if an ACK frame is to be sent and non-ACK mode, if an ACK frame is not expected to be sent. In this paper we will work on the non-ACK mode. In the non-ACK mode, every device in the network maintains three variables for each transmission attempt: NB, W, and CW. NB denotes the backoff stage, representing the backoff times that have been retried in the slotted CSMA-CA process while one device is trying to transmit a data frame in each transmission. W denotes the backoff window, representing the number of slots that one device needs to back off before clear channel assessment (CCA). CW represents the contention window length and is used to determine how many slots for the CCA before transmissions. CW will be set to two before each transmission and reset to two when the channel is sensed busy in CCAs.

Before each device starts a new transmission attempt, NB sets to zero and W sets to W_0. The backoff counter chooses a random number from $[0, W_0 - 1]$ and it decreases every slot without sensing channel until it reaches zero. If the number of backoff slots is greater than the remaining number of slots in the CAP, the MAC sublayer shall pause the backoff countdown at the end of the CAP and resume it at the start of the CAP in the next superframe [5]. If the number of backoff slots is less than or equal to the remaining number of slots in the CAP, the MAC sublayer shall check whether the number of remaining slots in current CAP is enough to complete the two CCAs and the frame transmission when backoff counter reaches zero. If the two CCAs and the frame transmission can be completed, the MAC sublayer shall request that the PHY perform the first CCA (denoted by CCA1) when backoff counter reaches zero. If channel is idle at CCA1, CW decreases one and the second CCA (denoted

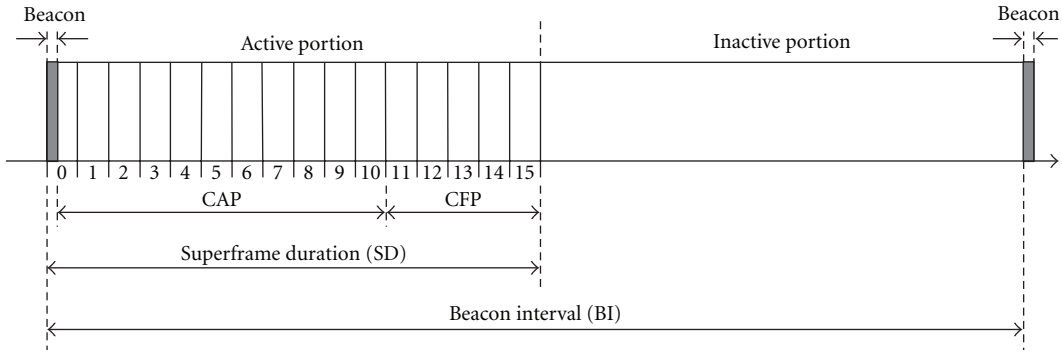


Figure 1: An example of the superframe structure. In this case, the beacon interval, BI, is twice as long as the active superframe duration, SD, and it contains CFP.

by CCA2) will be performed after CCA1. If channel is idle for both CCA1 and CCA2, the frame will be transmitted in next slots. If channel is busy in either CCA1 or CCA2, CW resets to two, NB increases by one, and W is doubled but not exceed W_x. If NB is smaller or equal to the allowed number of backoff retries macMaxCSMABackoffs (denoted by m), the above backoff and CCA processes are repeated. If NB exceeds m, the CSMA-CA algorithm ends. If two CCAs and the frame transmission cannot be completed in the remaining CAP, the MAC sublayer shall wait until the start of the CAP in the next superframe and apply a further random backoff delay before evaluating whether it can proceed again.

3. Model Assumption

There can be many scenarios with which the 802.15.4 networks may or may not interfere with each other if their operations are not coordinated when multiple 802.15.4 networks are deployed independently and closely. We assume two 802.15.4 networks are deployed closely and two simple representative scenarios are considered to focus on obtaining insights to the impact of uncoordinated operations on system performance.

We consider a star network topology with a PAN coordinator. These two networks are labelled by NET1 and NET2 with N_1 and N_2 denoting the number of basic devices in addition to one coordinator in each network, respectively. All the devices in its own network are within communication ranges of each other. Only uplink traffic from the basic devices to the coordinator in each network is considered. Each data frame has a fixed length which requires L slots to transmit over the channel. The data payload in MAC layer frame is fixed L_d slots, which is transmitted as the MAC payload in the MAC protocol data unit. In our scenarios, we assume that the two networks are both transmitting equal length of data payload L_d slots using the same L slots through the channel. We assume a saturated traffic with non-ACK mode, which means that each device has always traffic to send frames to its coordinator. The BOs of NET1 and NET2 are the same, which means the length of BIs of two networks are equal. The SOs of NET1 and NET2 are also set the same, which gives them equal active portions as well. We also consider an idle superframe structure, which means the active portions only consist of CAP without CFP, and an idle channel over which a frame fails if and only if collision happens.

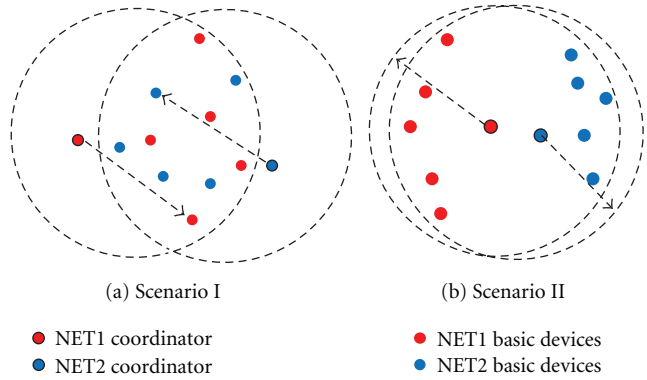


Figure 2: Communication range of each network is fully overlapped; (a) basic devices from two networks can detect each other's transmissions through CCAs; (b) basic devices from each network cannot detect transmission from other network's transmission through CCAs.

(1) Scenario I. For this scenario we assume that both considered networks are operated on the same frequency channel and the communication range of each network is fully overlapped as shown in Figure 2(a). We consider the beacon-enabled mode as mentioned in the previous section. Each network has a coordinator, which is responsible for broadcasting the beacon frames in the beginning of superframes. For simplicity, we assume that the beacons from any network can be correctly received by all the basic devices which belong to that network. The two networks share the whole channel frequencies, which means they can detect each other's transmissions through CCAs. For this scenario, the CAPs from each network can be partially overlapped or fully overlapped with different overlap ratio g as shown in Figure 3. When $g = 1$, which is the worst case that the two networks are fully overlapped in channel access periods, and when $g = 0$, which means there is no interference between these two networks.

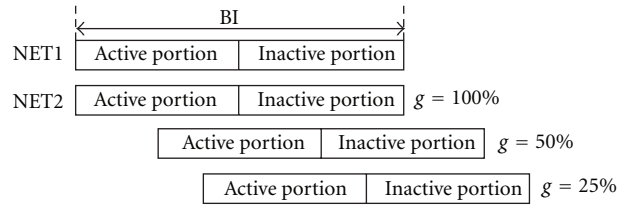


Figure 3: Illustration of channel access periods overlap in the superframe structure with different overlap ratio g, where g is the ratio of the number of slots in overlapped part to the number of slots in CAP, $0 \le g \le 1$.

(2) Scenario II. In this scenario, we assume these two networks share the channel frequencies and their communication range is fully overlapped as shown in Figure 2(b). We consider that the basic devices of each network can only hear transmissions from the other devices in its own network but cannot detect transmissions from other networks, which means that the CCA detections for each device are not affected by the channel activities from the other networks. This could happen because the distance between the basic devices and the two networks is too far to hear each other, although they operate on the same frequency channel. But the coordinators for the networks can detect transmissions from all the basic devices not only their own networks but also the other networks. With this assumption hidden terminals are present from neighbour networks. The transmissions could be collided by the data from other networks if they have overlap in the channel access portion. The same with Scenario I, when $g = 1$, which is the worst case that the CAPs of two networks are fully overlapped, and $g = 0$ means there is no interference in the CAPs for each network. We assume that the beacons from either network can be correctly received by all the devices belong to that network.

4. Analytic Model for Scenarios

4.1. Analytic Model of Scenario I. In the superframe structure, there is no activity from any device in the inactive portion for each network. The system performance of networks in CAP for Scenario I could be analysed separately with two parts according to different overlap ratio g: nonoverlapped and overlapped part.

For the nonoverlapped part, the performance of networks could be analysed with the existing analytic model for single 802.15.4 network proposed in [12]. According to the idea of performance modelling in [12], the nonoverlapped part channel states sensed by each device of either 802.15.4 network can be modelled by a renewal process, which starts with an idle period and followed by a fixed length of L slots (frame transmission), as shown in Figure 4.

The idle period depends on the random backoff slots and the transmission activities from each device. It is noted that the maximal number of idle slots is $W_x - 1$ plus two slot CCAs. On the other hand, the slotted CSMA-CA operations of each individual device could be modelled by a Markov chain with finite states. Let $p_{n,k}$ denote the probability of a transmission from devices in network n (n represents network identification, being 1 or 2) other than a tagged basic device in network n starting after exactly kth idle slots since the last transmission, where $k \in [0, W_x + 1]$ [12]. The transmission probability of a basic device in a general backoff slot can be calculated with the Markov chain constructed for each device.

Without loss of generality we consider NET1 and a tagged basic device in NET1. For the tagged basic device, its Markov chain consists of a number of finite states and each corresponds to a state of the CSMA-CA algorithm in one slot. These finite states are introduced below. Let $\overline{M}$ denote the steady-state probability of a general state M in the Markov state space. For simplicity we ignore the subscript "1" which corresponds to NET1 in the Markov states. In the following derivation we assume NET1 and NET2 use the same set of MAC parameters. It is trivial to extend to the cases with different sets of MAC parameters.

(1) Busy State. Denoted by $B_{i,j,l}$, during which at least one device other than the tagged basic device transmits the lth part of a frame of L slots, with the backoff stage and backoff counter of the tagged basic device being i and j, respectively, where $i \in [0, m]$, $j \in [0, W_i - 1]$, and $l \in [2, L]$, W_i is the minimum of $2^i W_0$ and W_m [12]:

$$\overline{B}_{0,j,2} = \sum_{k=2}^{W_0-1} p_k \overline{K}_{0,j+1,k} + \frac{1}{W_0} \sum_{k=2}^{W_m} p_k \left(\overline{K}_{m,0,k} + \overline{C}_{m,k}\right), \quad i = 0,\ j \in [0, W_0 - 1],$$

$$\overline{B}_{i,j,2} = \sum_{k=2}^{W_i-1} p_k \overline{K}_{i,j+1,k} + \frac{1}{W_i} \sum_{k=2}^{W_i-1} p_k \left(\overline{K}_{i-1,0,k} + \overline{C}_{i-1,k}\right), \quad i \in [1, m],\ j \in [0, W_i - 1],$$

$$\overline{B}_{i,j,l} = \begin{cases} \overline{B}_{0,j+1,l-1} + \dfrac{\overline{B}_{m,0,l-1}}{W_0}, & i = 0,\ j \in [0, W_i - 1], \\ \overline{B}_{i,j+1,l-1} + \dfrac{\overline{B}_{i-1,0,l-1}}{W_i}, & i \in [1, m],\ j \in [0, W_i - 1]. \end{cases} \tag{2}$$

(2) Backoff State. Denoted by $K_{i,j,k}$, during which the tagged basic device backoff with backoff counter being j at backoff stage i, after k idle slots since the last transmission, where $i \in [0, m]$, $j \in [0, W_i - 1]$, and $k \in [0, W_i - 1]$ [12]:

$$\overline{K}_{0,j,0} = \overline{B}_{0,j+1,L} + \frac{\left(\overline{B}_{m,0,L} + \overline{T}_L\right)}{W_0}, \quad i = 0,\ j \in [0, W_0 - 1],$$

$$\overline{K}_{i,j,0} = \overline{B}_{i,j+1,L} + \frac{\overline{B}_{i-1,0,L}}{W_i}, \quad i \in [1, m],\ j \in [0, W_i - 1],$$

$$\overline{K}_{i,j,k} = \begin{cases} \overline{K}_{i,j+1,k-1}, & k \in [1, 2], \\ (1 - p_{k-1}) \overline{K}_{i,j+1,k-1}, & 3 \le k \le W_i - 1. \end{cases} \tag{3}$$

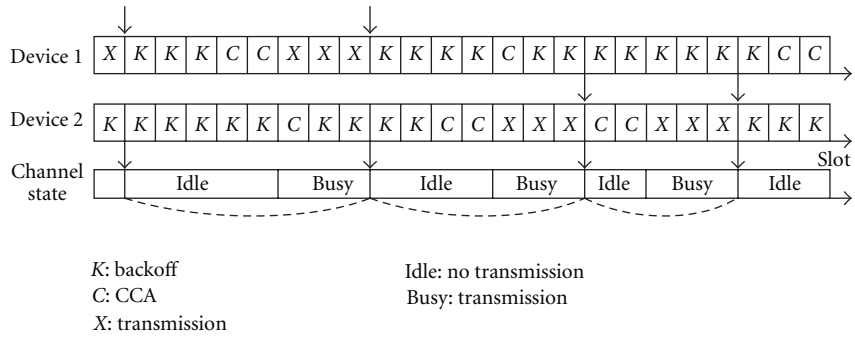


Figure 4: Example of channel renewal process for 802.15.4 networks. A random number of idle slots followed by a fixed length of transmission in the CAP.

(3) Sensing State. Denoted by $C_{i,k}$, during which the tagged basic device performs CCA2 at the ith backoff stage, after k idle slots since the last transmission, where $i \in [0, m]$ and $k \in [1, W_i]$ [12]:

$$\overline{C}_{i,k} = \begin{cases} \overline{K}_{i,0,k-1}, & k \in [1,2], \\ (1-p_{k-1})\overline{K}_{i,0,k-1}, & k \in [3, W_i]. \end{cases} \tag{4}$$

(4) Initial Transmission State. Denoted by $X_{i,k}$, during which the tagged basic device starts to transmit a frame at backoff stage $i \in [0, m]$, after $k \in [2, W_i + 1]$ idle slots since the last transmission [12]:

$$\overline{X}_{i,k} = \begin{cases} \overline{C}_{i,k-1}, & k = 2, \\ (1-p_{k-1})\overline{C}_{i,k-1}, & k \in [3, W_i + 1]. \end{cases} \tag{5}$$

(5) Transmission State. Denoted by T_l, during which the tagged basic device transmits the lth part of a frame, where $l \in [2, L]$. The first part is transmitted in the state $X_{i,k}$ [12]:

$$\overline{T}_l = \begin{cases} \sum_{i=0}^{m} \sum_{k=2}^{W_i+1} \overline{X}_{i,k}, & l = 2, \\ \overline{T}_{l-1}, & l \in [3, L]. \end{cases} \tag{6}$$

The transmission probability τ_k that the tagged basic device transmits after exactly k idle slots since the last transmission for the nonoverlapped part in CAPs can be computed by $\tau_k = 0$, for $k \in [0, 1]$, and for $k \in [2, W_x + 1]$ [12]:

$$\tau_k = \frac{\sum_{i=0}^{m} \overline{X}_{i,k}}{\sum_{i=0}^{m} \left[\overline{X}_{i,k} + \overline{C}_{i,k} + \sum_{j=0}^{W_i-1} \overline{K}_{i,j,k} \right]}. \tag{7}$$

With the above expressions derived for transmission probability τ_k ($\tau_{1,k}$ and $\tau_{2,k}$ for NET1 and NET2, resp.), we can calculate channel busy probability p_k^f ($p_{1,k}^f$ and $p_{2,k}^f$ for NET1 and NET2, resp.) for the nonoverlapped part of CAPs with the tagged basic device in Scenario I (in NET1 and NET2, resp.) with $k \in [0, W_x + 1]$:

$$\begin{aligned} p_{1,k}^f &= 1 - (1 - \tau_{1,k})^{N_1 - 1}, \\ p_{2,k}^f &= 1 - (1 - \tau_{2,k})^{N_2 - 1}. \end{aligned} \tag{8}$$

Since the balance equations for all steady-state probabilities and expressions for $p_{1,k}^f$ and $p_{2,k}^f$, $k \in [0, W_x + 1]$ have been derived, the Markov chain for the tagged basic device can be numerically solved. After the Markov chains are solved, we can calculate the throughput of nonoverlapped part $S_{I,n}^f$ for Scenario I with individualnetwork:

$$S_{\mathrm{I},n}^f = N_n L_d \sum_{i=0}^{m} \sum_{k=1}^{W_i} C_{n,i,k-1} \left(1 - p_{n,k-1}^f\right) \left(1 - p_{n,k}^f\right), \quad n = 1, 2. \tag{9}$$

To analyse energy consumption, we use normalised energy consumption, defined in [17] as the average energy consumed to transmit one slot of payload. The energy consumption of transmitting a frame in a slot (denoted by E_t) and perform a CCA (denoted by E_c) in a slot is set to 0.01 mJ and 0.01135 mJ, respectively [17]. We use the $\eta_{\mathrm{I},n}^f$ to represent the normalised energy consumptions of the nonoverlapped part for Scenario I with NET1 and NET2, respectively:

$$\eta_{\mathrm{I},n}^f = \frac{N_n}{S_{\mathrm{I},n}^f} \sum_{i=0}^{m} \left\{ \sum_{l=2}^{L} E_c B_{n,i,0,l} + \sum_{k=0}^{W_i+1} \left[E_c \left(K_{n,i,0,k} + C_{n,i,k} \right) + L E_t X_{n,i,k} \right] \right\}, \quad n = 1, 2. \tag{10}$$

For the overlapped part of CAPs in Scenario I, we have the transmission probabilities $\tau_{1,k} = \tau_{2,k}$, and we use $p_{n,k}^o$ to denote the new channel busy probabilities:

$$\begin{aligned} p_{1,k}^o &= 1 - (1 - \tau_{1,k})^{N_1 + N_2 - 1}, \\ p_{2,k}^o &= 1 - (1 - \tau_{2,k})^{N_1 + N_2 - 1}. \end{aligned} \tag{11}$$

The Markov chain for the tagged basic device can be numerically solved again with the new channel busy probabilities. The throughput S_{I}^o of the overlapped part for overall system in Scenario I is obtained as follows:

$$S_{\mathrm{I}}^o = (N_1 + N_2) L_d \sum_{i=0}^{m} \sum_{k=1}^{W_i} C_{1,i,k-1} \left(1 - p_{1,k-1}^o\right) \left(1 - p_{1,k}^o\right). \tag{12}$$

For NET1 and NET2, the throughputs of overlapped part S_n^o with Scenario I are obtained, respectively:

$$S_{\mathrm{I},n}^o = \frac{S_{\mathrm{I}}^o N_n}{(N_1 + N_2)}, \quad n = 1, 2. \tag{13}$$

The normalised energy consumptions of overlapped part for Scenario I are defined as $\eta_{\mathrm{I},n}^o$ with NET1 and NET2, respectively:

$$\eta_{\mathrm{I},n}^o = \frac{N_n}{S_{\mathrm{I},n}^o} \sum_{i=0}^{m} \left\{ \sum_{l=2}^{L} E_c B_{1,i,0,l} + \sum_{k=0}^{W_i+1} \left[E_c (K_{1,i,0,k} + C_{1,i,k}) + L E_t X_{1,i,k} \right] \right\}, \quad n = 1, 2. \tag{14}$$

Now, we can combine the nonoverlapped and overlapped part of CAPs together with different overlap ratio g. The various sleep time in BIs with different SOs is also considered. For Scenario I, we have throughputs $S_{\mathrm{I},n}$ of NET1 and NET2 are calculated by

$$S_{\mathrm{I},n} = 2^{\mathrm{SO}_n - \mathrm{BO}_n} \left[(1 - g) S_{\mathrm{I},n}^f + g S_{\mathrm{I},n}^o \right], \quad n = 1, 2. \tag{15}$$

The overall network throughput S_{I} is calculated by

$$S_{\mathrm{I}} = S_{\mathrm{I},1} + S_{\mathrm{I},2}. \tag{16}$$

The normalised energy consumption $\eta_{\mathrm{I},n}$ of Scenario I for NET1 and NET2 is, respectively

$$\eta_{\mathrm{I},n} = 2^{\mathrm{SO}_n - \mathrm{BO}_n} \frac{(1 - g) S_{\mathrm{I},n}^f \eta_{\mathrm{I},n}^f + g S_{\mathrm{I},n}^o \eta_{\mathrm{I},n}^o}{S_{\mathrm{I},n}}, \quad n = 1, 2. \tag{17}$$

4.2. Analytic Model of Scenario II. So far we have presented the performance analytic mode for Scenarios I with different sleep time and different overlap ratios. In this subsection we will give the analytic mode for Scenario II. Firstly, we assume that the two networks are fully overlapped in the CAPs ($g = 1$), and we focus on the channel access periods only. The performance with different overlap ratios can be obtained after we get the fully overlapped performance. As we discussed previously in Scenario I the channel access operation is not affected by channel activities at other networks. But for Scenario II correct reception of frame transmissions in one network can be affected by the frame transmissions in the other network. If a frame from the tagged basic device transmitted to coordinator in one network does not collide with frames from the other devices in the same network, it is still subject to collide with frames from the other network. An illustration of the uncoordinated operations for Scenario II, is shown in Figure 5. We could reuse the Markov states from [12] and calculate the new channel busy probability $p_{1,k}$ and $p_{2,k}$ for NET1 and NET2 in Scenario II, respectively. The problem that remains to be solved is on the calculation of successful frame reception probability, which depends on the probability of transmissions from both networks.

The performance of NET1 under the impact of uncoordinated operation from NET2 is considered firstly. The impact of NET1 transmissions to NET2 performance can be analysed similarly. With the Markov chain states we can compute the transmission probability $\tau_{2,k}$ of NET2 as done for nonoverlapped part in Scenario I by (7). Now the probability of exact k idle slots before one transmission in the active portion of NET2 can be derived, which is expressed by $p_{2,\mathrm{idle},k} = 0$ (identifier 2 means NET2) for $k \in [0, 1]$ and for $k \in [2, W_x + 1]$:

$$p_{2,\mathrm{idle},k} = \begin{cases} 1 - (1 - \tau_{2,k})^{N_2}, & k = 2, \\ \left(1 - (1 - \tau_{2,k})^{N_2}\right) \prod_{z=2}^{k-1} (1 - \tau_{2,z})^{N_2}, & k \in [3, W_x + 1]. \end{cases} \tag{18}$$

For each transmission from NET2 following k idle slots there is a probability $p_{2,\mathrm{suc},k}$ that an independent transmission from NET1 will not collide with the transmission from NET2. It is noted that the probability $p_{2,\mathrm{suc},k}$ is larger than zero only if idle slots k from NET2 is larger than or equal to the transmission data length L_1 in NET1. An illustration of the collision of frames from NET1 with frames from NET2 is presented in Figure 6.

We can calculate $p_{2,\mathrm{suc},k}$ for $k \in [2, W_x + 1]$ by

$$p_{2,\mathrm{suc},k} = \begin{cases} 0, & k < L_1, \\ \dfrac{k - L_1 + 1}{k}, & k \geq L_1. \end{cases} \tag{19}$$

The average probability $p_{2,\mathrm{suc,avg}}$ that a transmission from NET1 does not collide with transmissions from NET2 can be calculated by

$$p_{2,\mathrm{suc,avg}} = \frac{\sum_{k=2}^{W_x+1} k \cdot p_{2,\mathrm{idle},k} \cdot p_{2,\mathrm{suc},k}}{\sum_{k=2}^{W_x+1} (k + L_2) \cdot p_{2,\mathrm{idle},k}}, \tag{20}$$

where L_2 is the transmission data length in NET2.

For Scenario II, the throughput of nonoverlapped part of each network can be calculated by (9) as done in Scenario I, for there is no interference from each other. Then we can calculate the throughput of overlapped part $S_{\mathrm{II},1}^o$ for NET1 in this scenario:

$$S_{\mathrm{II},1}^o = p_{2,\mathrm{suc,avg}} S_{\mathrm{I},1}^f. \tag{21}$$

Similarly we can use the same way to calculate the throughput of overlapped part S_2^o for NET2:

$$S_{\mathrm{II},2}^o = p_{1,\mathrm{suc,avg}} S_{\mathrm{I},2}^f. \tag{22}$$

After the throughputs of nonoverlapped and overlapped part have been derived, we can calculate the throughputs $S_{\mathrm{II},n}$ for NET1 and NET2. For the different overlap ratio g and sleep time, the throughputs for NET1 and NET2 $S_{\mathrm{II},n}$ in Scenario II can be calculated by

$$S_{\mathrm{II},n} = 2^{\mathrm{SO}_n - \mathrm{BO}_n} \left[(1 - g) S_{\mathrm{I},n}^f + g S_{\mathrm{II},n}^o \right], \quad n = 1, 2. \tag{23}$$

The overall system throughput for Scenario II is calculated by $S_{\mathrm{II}} = S_{\mathrm{II},1} + S_{\mathrm{II},2}$.

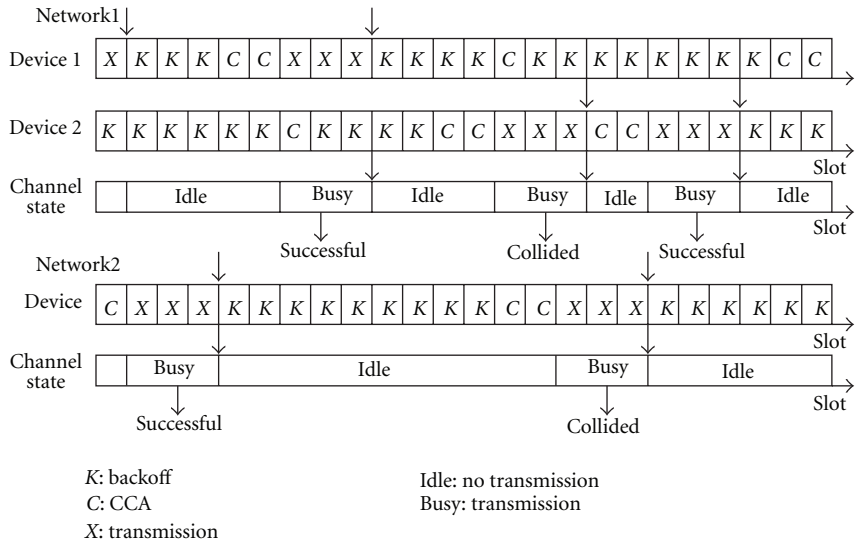


Figure 5: Example of transmission collisions for two uncoordinated 802.15.4 networks for Scenario II.

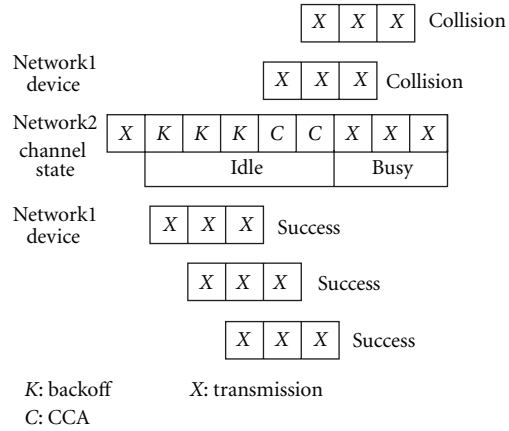


Figure 6: Illustration of transmissions from NET1 with/without collisions with frames from NET2 for Scenario II.

Because of the only impact on the transmissions in each network for Scenario II is the outcomes of frame reception. The normalised energy consumptions $\eta_{\mathrm{II},n}$ for NET1 and NET2 in Scenario II can be calculated by using $\eta^{f}_{\mathrm{I},n}$ (10) in Scenario I with the new throughputs $S_{\mathrm{II},n}$ in Scenario II:

$$\eta_{\mathrm{II},n} = 2^{\mathrm{SO}_n - \mathrm{BO}_n} \frac{\eta^{f}_{\mathrm{I},n} S^{f}_{\mathrm{I},n}}{S_{\mathrm{II},n}}, \quad n = 1, 2. \tag{24}$$

5. Numerical Results and Performance Analysis

We consider an IEEE PHY at frequency band 2400–2483.5 MHz with O-QPSK modulation and data rate of 250 kbps. A discrete event simulator is used to investigate the performance of uncoordinated problem and verify the proposed analytic model. The symbol rate is 62500 symbols per second for the PHY and at most 3000 slots of data could be transmitted in one second. We set BO for each network is fixed 6, which means each superframe length BI is fixed 3072 slots for both NET1 and NET2. Then the CAP of each superframe is only decide by SOs. For example, if SO sets to 5, which means half of the superframe 1536 slots are active portion and the rest is inactive portion. We can vary different SOs for various sleep slots for both two networks to investigate the impact of sleep mode. Typical results are presented with default MAC parameters for NET1: $W_0 = 2^3$, $W_x = 2^5$, and $m = 4$. The MAC parameter in NET2 are varied to investigate the impact of uncoordinated operations from NET2. The overhead of the header L_h in a data frame is 1.5 slots and the data length with MAC and PHY layer herder is $L = L_d + L_h$. We assume that both networks transmit frames with the same data length L. Each simulation results presented in the figures was obtained from the average of 20 simulations. In each simulation 10^5 data frames are transmitted.

Figures 7(a)–7(d) give the normalised throughput and normalised energy consumption of Scenario I. Figures 8(a)–8(d) show the normalised throughput and normalised energy consumption of Scenario II.

5.1. Analysis of Scenario I. Figure 7(a) shows the throughput S of overall system and throughput S_1 of NET1 for Scenario I. Only 5 M2M devices are in NET2 and the MAC parameters of NET2 are as same as those in NET1. For $L = 3$, we have $L_d = 1.5$ and the data length in one frame is 15 bytes. Similarly for $L = 6$, the data length is 55 bytes. Half of the slots in BI are in inactive portion with SO = 5 and CAPs of NET1 and NET2 are fully overlapped with $g = 1$. Consider the case of 20 M2M devices working in NET1. The throughput of NET1 is 0.03 for $L = 3$, which means that at most 30 data messages could be successfully delivered in one second in total NET1. Each M2M device in NET1 could deliver at most 1.5 data messages in one second with message size $L = 3$. This performance may be reasonably acceptable for M2M applications, for most of

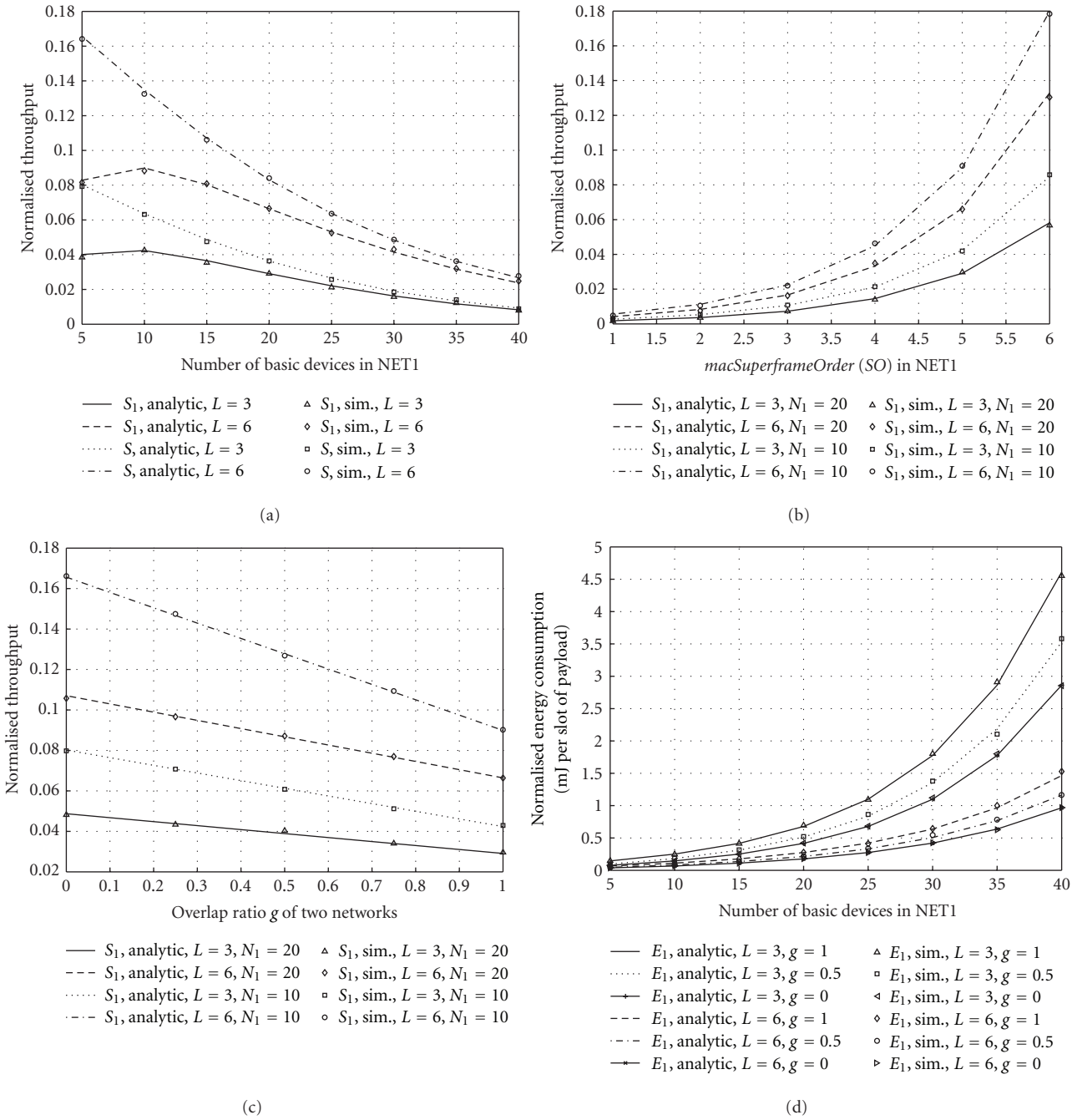


FIGURE 7: Normalised throughput S of overall system, S_1 of NET1 and energy consumption E_1 of NET1 for Scenario I. The number of basic devices in NET2 is fixed 5. (a) $L = 3$, $L = 6$ slots, SO = 5, and $g = 1$ for each network. (b) $L = 3$, $L = 6$ slots, and $g = 1$ for each network. The number of basic devices in NET1 is $N_1 = 20$ and $N_1 = 10$, respectively. (c) $L = 3$, $L = 6$ slots, and SO = 5 for each network. The number of basic devices in NET1 is $N_1 = 20$ and $N_1 = 10$, respectively. (d) $L = 3$, $L = 6$ slots, and $g = 1$; $g = 0.5$, $g = 0$, and SO = 5 for each network.

them do not need high data rate such as smart metering and environment monitoring. However, the throughput of NET1 decreases further when there are more M2M devices and the normal applications may not be effectively supported by the uncoordinated operation in Scenario I.

Figure 7(b) presents the throughput S_1 of NET1 with different SOs in Scenario I. Consider the case of 20 and 10 M2M devices in NET1 as examples. When the SO increased by one (not over maximum BO), the throughput is doubled with the condition overlap ratio $g = 1$. With the assumption of Scenario I, reducing the sleep time or using the nonsleep mode will dramatically increase the throughput S_1 of NET1, which will make it more suitable for most M2M application. For example, if we set SO = 6 for Scenario I, which

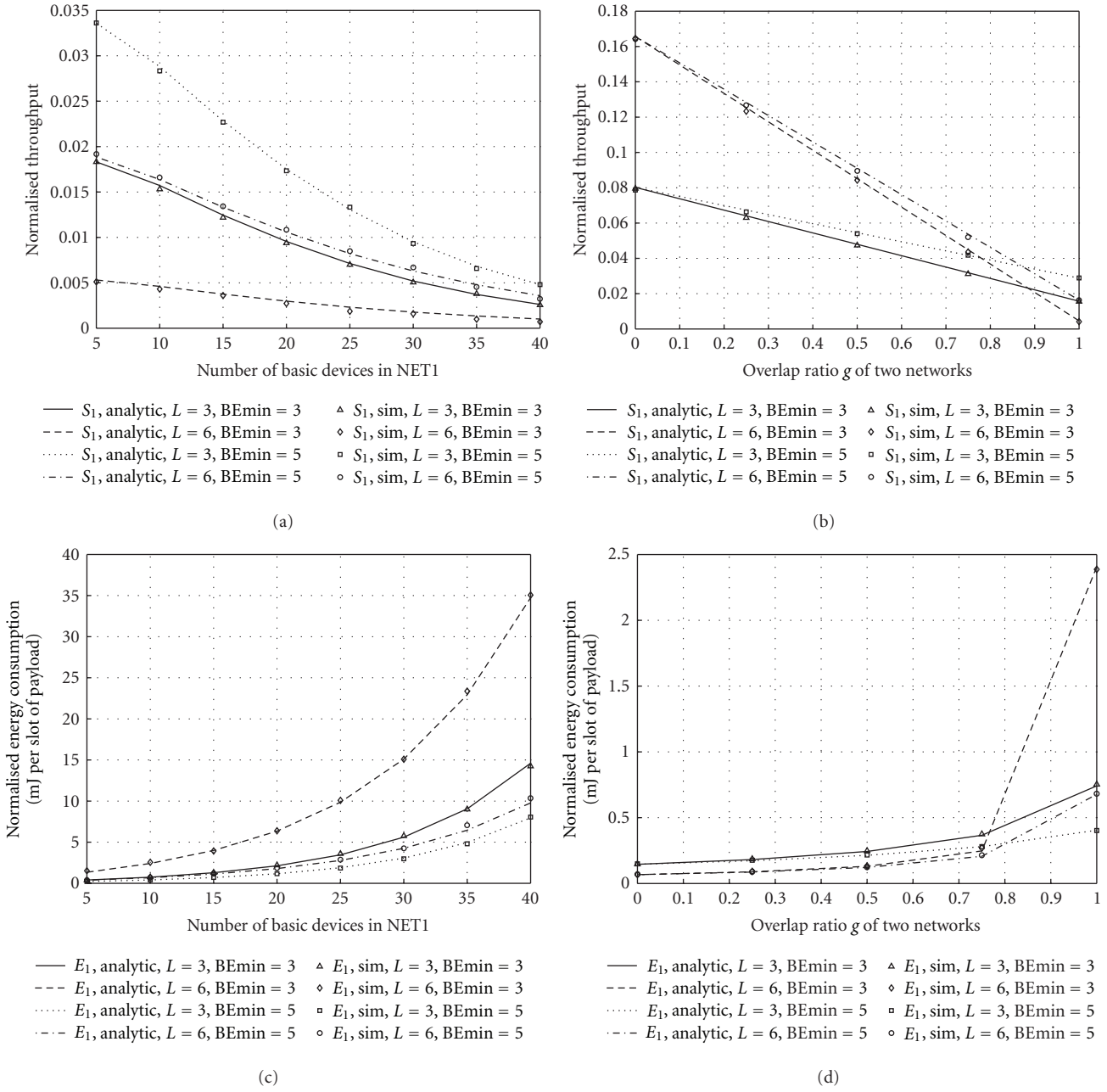


Figure 8: Normalised throughput S_1 of NET1 and energy consumption E_1 of NET1 for Scenario II. The number of basic devices in NET2 is fixed 5 and the BEmin of NET2 is set to 3 and 5 with initial backoff windows are $W_0 = 2^3$ and $W_0 = 2^5$. (a) $L = 3$, $L = 6$ slots and SO = 5, $g = 1$ for each network. (b) $L = 3$, $L = 6$ slots and SO = 5 for each network. The number of basic devices in NET1 is $N_1 = 10$. (c) $L = 3$, $L = 6$ slots, $g = 1$, and SO = 5 for each network. (d) $L = 3$, $L = 6$ slots, and SO = 5 for each network. The number of basic devices in NET1 is $N_1 = 10$.

means these two networks are working in nonsleep mode, the throughput S_1 of NET1 will be doubled than what Figure 7(a) shows, for the active slots in superframe BIs are doubled.

Figure 7(c) shows the relationship between throughput S_1 of NET1 and overlap ratio g in Scenario I. We take $N_1 = 20$ and $N_1 = 10$ as examples. When $g = 0$, which means there is no interference between two networks and when $g = 1$, which means the CAPs are fully overlapped. With the increasing of g from 0 to 1, the throughput S_1 of NET1 linearly drops. Consider the case of 10 M2M devices in NET1. The throughput S_1 of NET1 is 0.08 for $L = 3$ and $g = 0$ which means at most 80 data messages could be successfully transmitted in total in NET1. When the overlap ratio increases to $g = 0.5$ and $g = 1$, the successfully transmitted data messages drop to about 60 and 40, respectively.

Figure 7(d) represents the energy consumption E_1 of NET1 for Scenario I. We consider the cases of $N_1 = 20$ with $L = 3$ for three conditions $g = 1, g = 0.5$, and $g = 0$.

When $g = 1$, the E_1 is at most 0.7 mJ, which is the highest one compared to nearly 0.5 mJ for $g = 0.5$ and about 0.4 mJ for $g = 0$. For Scenario I, it is observed that increasing the active slots in superframe and reducing the overlap ratio can both obtain higher throughputs. Reducing sleep time will not affect the normalised energy consumption, but will increase the total energy consumption, for more slots will be active in superframes. Taking 20 M2M devices with $L = 3$ in NET1 as an example from Figure 7(d), the throughput will be doubled from 0.03 to 0.06 when SO increases from 5 to 6, which means 30 data messages will be increased to 60 data messages that could be successfully delivered in one second in total NET1 with message size $L = 3$. The normalised energy consumption is at most 0.7 mJ for this example with $g = 1$, and it will not change for different SOs. With the SO = 5 at most 10.5 mJ per second energy will be consumed in total of NET1, and when SO = 6 at most 21 mJ per second energy will be consumed in total of NET1. Compared to increase SOs, reducing the overlap ratio can also reduce the energy consumption to get higher throughputs.

5.2. Analysis of Scenario II. Figure 8(a) shows that the throughput S_1 of NET1 in Scenario II with 5 basic devices in NET2. Two sets of initial backoff window (BEmin = 3 and BEmin = 5) are used to study the impact of the slotted CSMA-CA parameters set for NET2 on the NET1 performance. It shows that for BEmin = 3 and $L = 3$, the throughput S_1 of NET1 drops below 0.04 even with only 5 devices in NET1. Results for throughput S of overall system have been obtained but not resented here due to concern on the readability of the figure. With larger frame length $L = 6$, the NET1 throughput S_1 drops further. It is also observed that the analytic results match very well with the simulation results, which demonstrates the high accuracy of the proposed analytic mode. Consider the case of 10 M2M devices in the NET1. The throughput of NET1 is 0.015 for $L = 3$ and 0.005 for $L = 6$, respectively. It means each M2M device in NET1 could successfully deliver at most 1.2 data messages in one second for $L = 3$ and 0.25 data messages for $L = 6$, respectively. When there are more M2M devices the NET1, the throughput of NET1 drops further and the normal M2M applications could not be effectively supported by the 802.15.4 networks. The above analysis shows that for Scenario II, uncoordinated operation of 802.15.4 networks can significantly affect the effectiveness of the networks on supporting M2M applications.

It is observed that with increased random backoff window in NET2, the throughput S_1 of NET1 for Scenario II is largely improved. This can be explained by the fact that with large random backoff window for devices in NET2, there will be smaller collision probabilities between frames from NET1 and NET2. Increasing random backoff window may be an effective measure to improve the system performance in case of multiple uncoordinated 802.15.4 networks and hidden terminals. But it is noted that such improvement may be achieved at cost of increased message delivery delay due to the larger backoff windows. Compared to Scenario I, the throughput S_1 of NET1 is limited even with increased random backoff window in NET2.

Figure 8(b) gives the throughput S_1 of NET1 with different overlap ratio g for Scenario II. The number of basic devices in NET2 is still 5 and in NET1 is 10 as an example. With the increasing of g from 0 to 1, the throughput S_1 of NET1 linearly drops dramatically. When $g = 0.5$, the throughputs S_1 of NET1 for $L = 3$ are about 0.045 and 0.055 with BEmin = 3 and BEmin = 5, respectively. It means each M2M device could successfully transmit 4.5 data messages and 5.5 data message in one second, respectively. When g increases to 1, they drop to 1.5 data messages and 3 data messages in one second, respectively. The throughputs S_1 of NET1 with BEmin = 3 drop quickly than with BEmin = 5 for both $L = 3$ and $L = 6$. With larger frame length ($L = 6$), the NET1 throughput S_1 drops further with increasing g and becomes lower than $L = 3$ when g is near 1.

Figures 8(c) and 8(d) represent the energy consumption E_1 of NET1 for Scenario II. It is observed that the uncoordinated operation of 802.15.4 networks could lead to significant increase of the energy consumption. With increasing number of M2M devices in NET1, the energy consumption E_1 of NET1 will increase dramatically and hardly support M2M application. For larger backoff window in NET2 (BEmin = 5), the energy consumption E_1 of NET1 largely drops, for more data messages can be successfully transmitted without collision. It is still higher compared to Scenario I. The overlap ratio can affect the energy consumption significantly, with lower overlap ratio g, the energy consumption can be largely improved. Compared to the measure of increasing random backoff window in NET2, the adaptive sleep mode which can decrease the overlap ratio g may be a more effective way to improve the system performance of both throughput and energy consumption.

6. Conclusion

Wireless M2M networks could play a critical role in the M2M technology. In this paper, we investigated the effectiveness of IEEE 802.15.4 networks in support of M2M communications. Two representative scenarios of closely deployed and uncoordinated IEEE 802.15.4 networks are studied. An analytic mode was proposed to understand the impact of uncoordinated operations and sleep mode on the overall system performance. Simulations demonstrate the high accuracy of the proposed analytical model. It was observed that the uncoordinated operations of 802.15.4 networks have significant impact on the M2M application performances. Reducing sleep time and overlap ratio can improve the overall networks performances. But by reducing sleep time approach the energy consumption may be increased due to longer active periods. On the other hand reducing sleeping overlap ratio g in the channel access periods of networks can largely improve the throughput and decrease the energy consumption.

Acknowledgments

The work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) with Grant reference no. EP/1010157/1 and the National Natural Science Foundation of China (NSFC) under the Grant no. 61103177.

References

[1] R. Q. Hu, Y. Qian, H. H. Chen, and A. Jamalipour, "Recent progress in machine-to-machine communications," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 24–26, 2011.

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

[3] Z. M. Fadlullah, M. M. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, and Y. Nozaki, "Toward intelligent machine-to-machine communications in smart grid," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 60–65, 2011.

[4] S. Y. Lien and K. C. Chen, "Massive access management for QoS guarantees in 3GPP machine-to-machine communications," *IEEE Communications Letters*, vol. 15, no. 3, pp. 311–313, 2011.

[5] "IEEE standard for information technology- local and metropolitan area networks- specific requirements- part 15. 4: wireless medium access control (mac) and physical layer (phy) specifications for low rate wireless personal area networks (wpans)," IEEE Standards 802. 15. 4-2006 (Revision of IEEE Std 802. 15. 4-2003), vol. 7, pp. 1 –320, 2006.

[6] J. Zheng and M. Lee, "A comprehensive performance study of IEEE 802. 15. 4," *Sensor Network Operations*, pp. 218–237, 2004.

[7] G. Anastasi, M. Conti, and M. Di Francesco, "A comprehensive analysis of the MAC unreliability problem in IEEE 802.15.4 wireless sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 1, pp. 52–65, 2011.

[8] K. Yedavalli and B. Krishnamachari, "Enhancement of the IEEE 802.15.4 MAC protocol for scalable data collection in dense sensor networks," in *Proceedings of the Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (Wiopt'08)*, pp. 152–161, April 2008.

[9] J. Mišić, S. Shafi, and V. B. Mišić, "Performance of a beacon enabled IEEE 802.15.4 cluster with downlink and uplink traffic," *IEEE Transactions on Parallel and Distributed Systems*, vol. 17, no. 4, pp. 361–376, 2006.

[10] I. Das and S. Roy, "Analysis of the contention access period of ieee 802. 15. 4 mac," Tech. Rep. UWEETR-2006-0003, Department of Electrical Engineering, University of Washington, 2006.

[11] Z. Tao, S. Panwar, D. Gu, and J. Zhang, "Performance analysis and a proposed improvement for the IEEE 802.15.4 contention access period," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC'06)*, pp. 1811–1818, April 2006.

[12] J. He, Z. Tang, H. H. Chen, and S. Wang, "An accurate Markov model for slotted CSMA/CA algorithm in IEEE 802.15.4 networks," *IEEE Communications Letters*, vol. 12, no. 6, pp. 420–422, 2008.

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

[14] C. K. Singh, A. Kumar, and P. M. Ameer, "Performance evaluation of an IEEE 802.15.4 sensor network with a star topology," *Wireless Networks*, vol. 14, no. 4, pp. 543–568, 2008.

[15] P. Park, P. Di Marco, P. Soldati, C. Fischione, and K. H. Johansson, "A generalized Markov chain model for effective analysis of slotted IEEE 802.15.4," in *Proceedings of the IEEE 6th International Conference on Mobile Adhoc and Sensor Systems (MASS '09)*, pp. 130–139, October 2009.

[16] J. He, Z. Tang, H. H. Chen, and Q. Zhang, "An accurate and scalable analytical model for IEEE 802.15.4 slotted CSMA/CA networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 440–448, 2009.

[17] T. R. Park, T. H. Kim, J. Y. Choi, S. Choi, and W. H. Kwon, "Throughput and energy consumption analysis of IEEE 802.15.4 slotted CSMA/CA," *Electronics Letters*, vol. 41, no. 18, pp. 1017–1019, 2005.