Performance Enhancement for Wireless Networks:
Modulation, Clock Synchronization and Resource Management

by

Zhe Yang
B.Eng., Xi'an Jiaotong University, 2005
M.Eng., Xi'an Jiaotong University, 2008

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Electrical and Computer Engineering

Performance Enhancement for Wireless Networks:
Modulation, Clock Synchronization and Resource Management

by

Zhe Yang
B.Eng., Xi'an Jiaotong University, 2005
M.Eng., Xi'an Jiaotong University, 2008

Supervisory Committee

Dr. Lin Cai, Supervisor
(Department of Electrical and Computer Engineering)

Dr. Hongchuan Yang, Departmental Member
(Department of Electrical and Computer Engineering)

Dr. Yi Shi, Outside Member
(Department of Mechanical Engineering)

Supervisory Committee

Dr. Lin Cai, Supervisor
(Department of Electrical and Computer Engineering)

Dr. Hongchuan Yang, Departmental Member
(Department of Electrical and Computer Engineering)

Dr. Yi Shi, Outside Member
(Department of Mechanical Engineering)

ABSTRACT

Wireless networks become more and more important in modern information systems as the last mile/meter solutions, thanks to the flexibility of mobile access to facilitate Internet access anytime, anywhere. Given the limited resources, e.g., spectrum and energy supplies, to meet the ever increasing demand for wireless data services, new approaches are beckoned to enhance the spectrum and energy efficiency. We investigate this problem from three important aspects, digital modulation, clock synchronization and concurrent transmission scheduling. The contributions of this dissertation are four-fold.

First, we employ the cross-layer design to explore the spatial diversity and broadcast nature of wireless links and propose a novel network modulation scheme that can superpose the information bits of different priorities into one symbol. It offers a new dimension to improve the network throughput since we can flexibly configure the transmission according to the channels among transceivers. Moreover, it is compatible with the main-stream hardware and we just need a software upgrade to implement the idea.

Second, we propose modulation schemes based on hexagonal tiling, which is known to be the most compact way of two-dimensional regular tiling. In order to fully utilize the advantage of hexagonal constellation, we employ the non-binary error control

coding since the number of constellation points of hexagonal constellation is not necessarily to be an integer power-of-two. The feasibility of these new modulation schemes is verified by the prototype system based on the software defined radio platform USRP2 and GNU Radio.

Third, to facilitate a wide range of wireless communications technologies and protocols, clock synchronization among several wireless devices is a fundamental requirement. We investigated this problem by tracing to the source of clock desynchronization, which is the clock skew. However, as shown by measurement results, the clock skew is not constant and related to the working temperature. We propose a novel clock skew estimation algorithm that can leverage the temperature information to accurately estimate the clock skew. Based on the estimation results, we propose a clock synchronization scheme that can directly remove the clock skew according to the working temperature.

Fourth, the traditional time-sharing based scheduling schemes usually schedule one transmission within certain area. The emerging broadband wireless devices can dynamically adjust the transmitted data rate according to the received signal to interference and noise ratio (SINR). Allowing concurrent transmissions may be more efficient, while optimal scheduling problem for concurrent transmissions is an NP-hard problem. We propose simple yet effective heuristic algorithms that can significantly improve the system throughput with moderate computational complexity.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

This dissertation has benefited greatly from the support of many people, some of whom I would sincerely like to thank here.

To begin with, it gives me great pleasure in expressing my appreciation to my PhD advisor Dr. Lin Cai, for all her help and support during the past five years. She helped me come up with the thesis topic and guided me to overcome the technical difficulties I encountered. It was a great experience to work with her, whose expertise, commitment and enthusiasm for research have always been an inspiration.

I would like to thank the other members of my committee, Dr. Hongchuan Yang, and Dr. Yang Shi for their help and suggestions at all levels of the research work and the dissertation. Also, I would like to thank Dr. Wei Song from the Department of Computer Science, University of New Brunswick, for taking time out from her busy schedule to serve as my external examiner.

Dozens of people, including professors, staffs and students, have helped and taught me immensely at Univeristy of Victoria. I would like to thank Dr. Jianping Pan, Dr. Wusheng Lu and Dr. Aoron Gulliver for their enlightening discussions and guidance on my research projects, which significantly enhanced the quality of my work. Furthermore, I would like to thank my colleagues and friends who supported and helped me during my study at University of Victoria especially Siyuan, Yuqian, Min, Xuan, Lei(Zheng), Kan, Yi, Haoyuan, Zhe, Bojiang, Haoling, Vivek, Ahmad, Le, Guowei, Shuai, Yangyang, Lei(Zhang), Meng etc.

My most heartfelt gratefulness goes to my parents, my parents-in-law, my wife and my brother for their selfless and consistent love and support. Words cannot express my thanks for them. Without their help, I cannot imagine I am able to make any tiny progress and achievements.

Finally, very special thanks goes out to my son, Jirong, the adorable angel who just joined us several months ago, for giving me unlimited happiness, pleasure and the special support in his own way.

Zhe YANG, Coquitlam, BC, Canada

DEDICATION

To my family for their endless love, support and encouragement.

Chapter 1

Introduction

1.1 Background

With the continuous growth of the demand for ubiquitous Internet access anywhere, anytime, wireless communication networks become an essential component of modern information systems. It has triggered the fast development of wireless communication technologies, ranging from cellular networks, wireless metropolitan area networks (WMANs), wireless local area networks (WLANs), to mobile ad hoc networks (MANETs). Although the achievable link data rate in the wireless systems keeps increasing, it is out-paced by the demand for emerging information services. According to the Cisco Visual Networking Index and numerous studies based on market status and trends, mobile data traffic will increase 18-fold from 2011 to 2016 [15]. Meanwhile, most of the mobile/handheld devices are battery powered, which indicates that they usually have limited energy supplies. This ever-growing demand for mobile data creates significant demands on the scarce wireless spectrum and limited power of mobile devices. Thus how to improve the spectral and energy efficiency of wireless communications systems is a major challenge for the research community and industry, which also motivates this dissertation.

The layered architecture has been adopted in network protocol design. The overall networking tasks are divided into different layers. The protocol for each layer is designed individually to provide services to upper layer protocols. This design concept allows the upper layer takes the services of the lower layer transparently, which simplifies the design and implementation of each layer. On the other hand, the upper layer protocols do not access the detailed information and implementation of the

lower layers. In wireless communication, just treating the lower layer as transparent components is not favorable since it may lose the chance to further improve the system performance. Therefore, cross-layer design has attracted intensive attention.

To further enhance the wireless network performance, both the properties of the physical medium and the quality of service requirements of applications should be taken into account. We can adaptively adjust the transmission configuration to meet the requirements of applications, such as adjusting transmission power, data rate and coding schemes. Knowledge should be shared among layers to achieve the highest possible adaptivity [6]. We work on three important issues in wireless communication networks, including clock synchronization, digital modulation and scheduling for concurrent transmission. In the following section, we will briefly overview each of them.

1.2 Research Objects and Solutions

1.2.1 Digital Modulation

Digital modulation is the process to convey coded messages in a physically transmittable signal, which is an indispensable building block for a digital communication system. For digital modulation, one of the most popular design concepts is the QAM scheme, where a signal can be represented in the signal space domain using an in-phase and quadrature-phase (I/Q) constellation diagram. Each constellation point is mapped to the amplitudes of the in-phased and quadrature-phased carrier waves. In this dissertation work, we investigate the digital modulation from two different perspectives.

Network Modulation

It is well know that wireless medium are usually open and shared. Besides, the channel quality of different users varies greatly due to the topology and the fading effect etc. Inspired by these facts, we investigate the problem from a new angle and propose the network modulation (NM) scheme for a multi-point wireless system, which can significantly improve network throughput with a software upgrade to wireless devices. The reason why we call it network modulation is that it allows us to combine topology control, scheduling and routing (which belong to the functions of network protocols) with modulation for multi-point wireless systems. In current wireless systems, when

sources transmit data to receivers through single-hop or multi-hop wireless paths, the physical layer modulates and demodulates the information bits hop-by-hop, and the transmission over each hop is treated the same as that in a point-to-point communication link.

The proposed NM scheme allows the sender transmit messages to multiple receivers simultaneously, using a software mapping technology to redefine the constellation of typical quadrature amplitude modulation (QAM). It can superpose different data to different layers corresponding to different channel qualities. By selecting and re-mapping some constellation points to different layers, they will have the ability to explore the spatial diversity. As the proposed NM is software-based, it does not require specialized communication hardware and can be implemented with low cost and high flexibility. NM can be used to improve network performance in a wide spectrum of scenarios, for applications with or without differentiated service requirements, anycast (broadcast, multicast and unicast) services, one-way or two-way traffic, and single-hop or multi-hop wireless paths, in infrastructure or ad hoc networks. The minimum requirement for applying network modulation is that there are no less than three nodes within each other's transmission range, so we can consider modulation, topology control, resource allocation, and routing jointly.

Hexagonal Modulation

The adaptive modulation and coding (AMC) schemes are widely adopted by modern digital wireless communication systems, which can dynamically adjust the transmission rate according to the channel state information (CSI) to efficiently utilize channel resources. One of the most popular modulation designs for wireless communications are based on the QAM concept.

The amount of information transmitted by a modulation symbol is determined by the number of constellation points of this modulation scheme. At the receiver side, the demodulation is usually based on the Voronoi diagram, which can be considered as the partition of the two-dimension signal space. With the conventional QAM, the signal space is partitioned into square decision regions; however, it is well known that hexagon instead of square is the densest two-dimension packing. With hexagonal tilting, we can pack more constellation points and thus transmit more information under the same energy constraint. However, it is non-trivial to adopt the hexagon based QAM (H-QAM) designs since the number of constellation points is

no longer the integer power-of-two, which is very inconvenient for the widely used binary representation of digital information. The straightforward solutions that just selects part of the constellation points or just converts the coded bit stream to non-binary symbols may not be able to fully explore the potential of H-QAMs. Selecting part of the constellation points might deprive the energy efficiency of H-QAM. Just directly converting coded bit stream to non-binary symbols that are compatible with H-QAM may suffer from the error propagation problem and results in an even worse bit-error-rate (BER) performance.

In this dissertation work, we not only adopt the H-QAM designs, but also use ternary digits (trits) to carry information. To avoid the error propagation problem, we employ the ternary based error correction coding/decoding to provide the forward error protection for the transmitted packets. Combing with ternary-based error coding, the proposed H-QAM can fill the signal-to-noise-ration (SNR) gap of the conventional AMC schemes and therefore can achieve much higher overall system throughput.

1.2.2 Clock Synchronization

Clock synchronization is the process to mitigate the clock offset between two or more hosts in a network and let them obtain a common notion of time in the network. It serves as a fundamental requirement to facilitate pervasive computing and communications, and spreads a wide range of network applications including wired network measurements, wireless sensor network data fusion, peer-to-peer video on demand (P2P VoD) systems, and so on. All of them require clock synchronization with a good resolution and tractable computational cost, especially for those in a resource-constrained environment.

However, how to effectively estimate and compensate clock offset among two or more hosts with low communication and computational cost is still an open issue. It is especially important for wireless/handheld devices which are resource-constrained. The inherent and dominant reason is clock skew. It is the instantaneous clock drift that is caused by the imperfect crystal and peripheral circuit, i.e., the difference between the rates of the pulses generated by periodic component of two clocks. The effect of clock skew will be accumulated and lead to an unbounded clock offset.

Therefore, the estimation and compensation of clock skew are essential to improve the performance and efficiency of clock synchronization, which can be especially ben-

eficial to resource-constrained devices. In this dissertation work, we first propose an additional information aided clock skew estimation algorithm to accurately estimate the clock skew. Then, based on the estimation results, we propose a two-phase clock synchronization scheme. The first phase is for external clock synchronization, during which nodes update their clock by exchanging timestamp messages with the reference clock. Different from the conventional solutions, we propose to directly remove the clock skew during the external synchronization phase, for a higher synchronization accuracy and lower computational complexity. The second part is the clock self-calibration phase, as the accumulated clock skew will make the synchronized clock drift away again, therefore we need to compensate the clock skew to maintain the clock synchronization accuracy. However, the compensation is non-trivial as the clock skew may not be constant. Thus we propose the temperature-assisted clock self-calibration (TACSC) to dynamically compensate the clock skew according to the working temperature. Extensive simulation results demonstrate that the proposed synchronization scheme can achieve a much lower root mean square error in the external synchronization phase. Furthermore, during the clock self-calibration phase, the TACSC scheme can improve the synchronization accuracy by more than one order of magnitude, which is verified by simulation and experimentation with a Mica2 sensor nodes testbed.

1.2.3 Scheduling for Concurrent Transmission

In traditional narrow-band rate-nonadaptive wireless systems, each flow can reliably transmit at a specified data rate so long as the received signal to SINR is greater than a certain threshold. In this case, letting more flows transmit concurrently without violating their SINR requirements implies that a higher system throughput can be achieved. The optimal scheduling problem for this case can be converted to a Knapsack problem which is known to be NP-complete [77]. We have seen a great deal of research efforts in determining the system capacity and optimizing the concurrent scheduling for such systems [29].

In emerging broadband wireless systems, however, the physical (PHY) layer uses adaptive modulation and coding(AMC) schemes, so the instantaneous link data rate is a variable determined by its received SINR. When we schedule multiple flows to share a time slot, from the perspective of an individual flow, its performance in terms of data rate or bit error probability in the slot might be degraded due to the interference from

other flows. On the other hand, with more flows sharing the time slots, each flow might be allocated with more time slots, hence the long-term throughput of each flow and the entire network throughput will be improved. Scheduling for rate-adaptive broadband wireless networks with random network topology is thus much more difficult than that in rate-nonadaptive systems: in a time slot, each flow's throughput (data rate) is not known before the scheduling decision of that slot is finalized; when we add or remove one flow from the set sharing the slot, the throughput of all other flows in the set will change. The capacity bound derived for rate-nonadaptive networks is no longer applicable either.

Given the hardness of the scheduling problem for concurrent transmissions in rate-adaptive wireless systems and the fact that the real time scheduling decision should be made within a few milliseconds, optimal scheduling algorithms are impractical to deploy due to the computational cost. In this dissertation work, we propose practical heuristic scheduling algorithms that can achieve better performance than the existing solutions and maintain long-term fairness among competing flows with tolerable computational complexity.

1.3 Dissertation Organization

This dissertation work is intended to improve wireless networks performance from different perspectives to exploit the broadcast nature of wireless channel, the environment related behaviors and the spatial diversity gain. The rest of this dissertation is organized as follows.

The detailed research work and results are presented in Chapters 2, 3, 4 and 5 discussing three different research objects, namely, digital modulation, clock synchronization and network scheduling for concurrent transmissions, respectively. In each chapter, we first introduce the research topic, related research work and the proposed solutions, followed by the simulation/experiment results.

The network modulation scheme is proposed in Chapter 2. It can superpose different information bits into different layers in one transmitted signal and hence have the ability to explore the spatial diversity. Also inspired by the fact that the wireless channel is open and shared, we propose to leverage the help of error control coding to further improve the performance of user cooperative network. The simulation results have shown that the proposed schemes can significantly improve the spectrum efficiency with very low computational cost.

In Chapter 3, we present the detailed design and performance evaluation of H-QAM. In order to explore the capacity of H-QAM, it is critical to introduce the non-binary error correction coding/decoding. Combining with different coding rates and H-QAM, we can obtain the augmented AMC set. The new AMC schemes can fill the gap between the conventional QAM based AMC schemes and lead to a finer grained AMC set. Therefore, we can substantially improve the system throughput. Meanwhile, since we fix the average symbol energy, the per-bit energy consumption is also reduced.

Clock skew is the inherent reason of clock desynchronization. In Chapter 4, based on the measurement results, we first present a new hybrid clock skew model consisting of two different models, which describes the behavior of clock skew under different working environment. The additional information aided multi-model Kalman filter (AMKF) is proposed to tackle the model uncertainty problem and we can obtain a much higher estimation accuracy. Based on the estimation results, we can obtain the value of clock skew w.r.t. the working temperature. The proposed two-phase clock synchronization scheme can directly remove the clock skew accordingly. The synchronization accuracy can be improved during the external synchronization phase. In the clock self-calibration phase, we can dynamically compensate the clock skew and keep the synchronization status for a much longer period without introducing communication overhead.

Chapter 5 presents the scheduling for concurrent transmissions, as it may not be efficient to allocate wireless channel(s) to one link exclusively. How to select the concurrent transmission subset is an NP-Complete problem. In this chapter, we propose a heuristic algorithm that can improve the system throughput with moderate computational cost.

Chapter 6 concludes the dissertation with further research issues.

1.4 Bibliographic Notes

Most of the works reported in this dissertation have appeared in research papers. The works in Chapter 2 have been published in [109, 107]. The works in Chapter 3 have been submitted as [104]. The works in Chapter 4 have been published in [110, 105] and submitted as [108], and those in Chapter 5 have appeared in [106].

Chapter 2

Network Modulation

2.1 Introduction

For the wireless communications among users at different locations, exploring the spatial diversity and the broadcast nature of wireless media is a promising direction. Previous approaches to explore the spatial diversity can be classified into two categories. One is to take advantage of the spatial diversity gain to improve the physical (PHY)-layer performance, such as improving the spectrum and energy efficiency and reducing the bit error rate [51, 66]. The other one is to arrange multi-hop relaying and routing considering the network topology to improve the network performance, such as network throughput and end-to-end delay [44]. On the other hand, it is well known that, given the open and shared medium of a wireless channel, isolating communication and networking designs may lead to performance degradation.

We investigate the cross-layer design for the performance enhancement of wireless networks in a new direction. Specifically, we propose to enhance the PHY-layer modulation and error-correction coding design considering the network topology, aimed to improve both the PHY-layer spectrum and energy efficiency, which can lead to a better network performance in terms of a higher network throughput and a lower bit-energy consumption.

We propose the NM scheme, a network topology-aware modulation scheme. Different from the traditional digital modulation schemes that are designed for point-to-point transmissions, we design and select the NM schemes considering the network topology and the three communication channels among the source node, the destination node, and the relay node that assists the transmission. It can also explore

the broadcast nature of wireless media and serve two users simultaneously without introducing any new resources. Furthermore, for relay communication, we design the error-correction coding assisted relay (EAR) scheme, where a strategically located relay can generate and transmit Reed-Solomon (RS) code instead of the original information bits to the receiver. The configuration of the RS code also depends on the topology and the channels among the three nodes.

2.2 Related Work

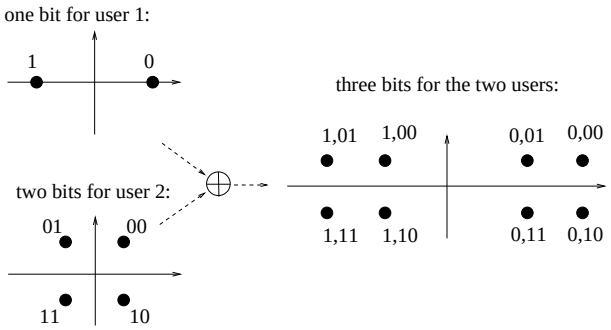


Figure 2.1: Super-positioning precoding (SPC).

There are many approaches in different layers to improve the performance of wireless communication systems. We take a new direction to configure the modulation and error-correction coding schemes to enhance the network performance via user cooperation.

The proposed NM scheme is inspired by super-positioning precoding (SPC) [16], which was designed for multicast transmissions, and recently how to use SPC for relay communications has also appeared [11, 78]. The idea can be illustrated using an example shown in Fig. 2.1. When a source needs to deliver information bits to two users with different channel quality, instead of transmitting the bits in different time slots (called the time-sharing scheme), the source can superpose the bits targeted to different users, and transmit them in one shot, such that both users can demodulate and decode its own bits successfully (w.h.p.). The potential of SPC has been realized not until recently, when the digital video broadcasting (DVB) systems have standardized and implemented the hierarchical modulation (h-mod) schemes based on the idea of SPC [96, 99]. The nice fit of SPC (or h-mod) and video broadcasting is because the video sources use scalable video coding schemes and prefer differentiated services

(DiffServ) provided by SPC. For instance, each video frame is encoded to a base layer and a number of enhancement layers, and bits of the base layer are much more important than the others. Using SPC (or h-mod), the base layer bits can be received successfully with much lower SNR than that for the enhancement layer bits, so the receiver can always decode the base layer successfully, and can decode the enhancement layers when the transmission channel quality is sufficiently good. However, SPC (or h-mod) requires specialized hardware and sophisticated signal processing, so it is not desirable for wireless hand-held devices.

Another direction is to extend the functions of the PHY layer to networking protocols, which can effectively enhance system performance. For example, error coding has been investigated for the PHY layer for several decades. Recently, network coding has been proposed and been applied in many wired and wireless scenarios to improve network performance [2, 42, 114, 43]. Similarly, we also let the upper layer carry some function that is traditionally considered to be a PHY layer issue.

In [12], a scalable modulation (s-mod) scheme has been proposed which uses software to remap the constellations of typical QAM to provide DiffServ for video multicast over wireless networks. Performance study shows that the software based solution, s-mod, can achieve performance similar to or even better than SPC (or h-mod) for supporting video multicast in dynamic wireless environment. SoftCast is another approach to support scalable video multicast in wireless networks [41].

In the proposed network modulation scheme, we also use a software based modulation solution, similar to s-mod, to improve the wireless network performance. Different from the work in [12, 41], here, we focus on how the software based modulation remapping can be combined with the link layer resource allocation and network layer relay and routing to significantly improve the throughput of wireless networks, for any types of applications (with or without DiffServ requirements), for anycast services, and for single-hop or multi-hop wireless networks.

Using error-correction coding to improve throughput is a well-studied topic [20, 40, 102, 93, 46, 98, 14, 54]. For instance, in [20], the optimal error-correcting code is selected based on time-varying wireless channel conditions. An adaptive link-level forward error correction algorithm was proposed in [40] to adjust the error-correcting code according to the predicted packet loss rate. In [39], a partial packet recovery scheme was proposed: instead of retransmitting the corrupted packet, the sender only retransmits the bits in a packet that are likely in error and therefore the retransmitted packet length can be much smaller than that of the original packet. To overcome the

shortcoming of the previous solution in [39] that requires sophisticated hardware modifications, in [59], the sender retransmits a smaller packet contains error-correcting codes. Our proposed EAR scheme is motivated by [59]. However, in a relay assisted transmission, the relay node may not have a correct copy of the original packet, which makes the configuration more complicated. In addition, the original messages and the error-correcting codes are transmitted through different channels. Therefore, the selection of error-correcting codes needs to consider the topology and channels among multiple nodes, and its performance gain thus is topology sensitive.

2.3 Network Modulation

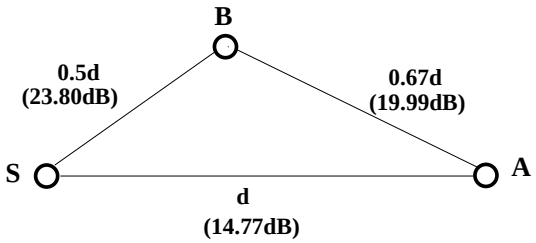


Figure 2.2: A three-node example.

2.3.1 Design Intuition

We use a three-node topology shown in Fig. 2.2 to briefly explain the idea and demonstrate the performance gain. Nodes S, A, B communicate with each other using wireless transmissions, and the transceiver distances and the corresponding signal to noise ratios (SNRs) are given in the figure assuming their transmission power (symbol energy) is a constant, E_s.

In current broadband wireless communication systems, the AMC schemes are used to improve transmission efficiency in highly dynamic environment, so the sender can adjust the modulation and coding schemes (and thus the link data rate) according to the received SNR. For example, in WiMAX systems, the transmitter can vary the data rate by using BPSK, QPSK, 16-QAM, and 64-QAM, combined with different error control codes.

We consider three typical cases: (1) S unicasts data to A; (2) S broadcasts data to both A and B; (3) S unicasts different data to A and B, respectively. For the

first case, if using the direct transmission link with AMC, S can transmit at most two bits per symbol duration (b/sym) (using QPSK); if using B as the relay, the first hop can transmit 6 b/sym (using 64-QAM) and the second hop can transmit 4 b/sym (using 16-QAM), so the total throughput is $6/(1+6/4) = 2.4$ b/sym. Using the proposed NM scheme, we can let S send two bits to A and three other bits to B simultaneously during one symbol transmission (in Section 2.3.3, we will discuss how NM can achieve it), then let B relay the three bits to A using 16-QAM, so we can achieve the throughput of $5/(1+3/4) = 2.857$ b/sym. The per-bit energy consumption using direct transmission, relay, and NM are $E_s/2$, $E_s/2.4$, and $E_s/2.857$, respectively. In other words, the proposed NM scheme can achieve both 43% and 19% throughput gain and 30% and 16% energy saving than the direct transmission and the traditional relay transmission schemes, respectively. The throughput gain for case (2), the broadcast scenario, is the same as that for case (1).

For case (3), assume that the lengths of the two messages to A and B are the same. If S transmits to A and B directly, it takes $1/2 + 1/6 = 2/3$ symbol durations to transmit one bit to A and one bit to B separately, i.e., the system throughput is 3 b/sym. If S transmits all bits to B first using 64-QAM, and let B relay the bits for A using 16-QAM, then the throughput is $2/(1/6 + 1/2.4) = 3.429$ b/sym. Using the proposed NM scheme, S can transmit two bits to A and three bits to B simultaneously in one symbol; then, for the remaining bits to A, we can use B as the relay using the same NM scheme as that for case (1). The throughput using NM is $6/(1 + 1/2.857) = 4.444$ b/sym. Thus, the throughput gains of using the proposed NM over the direct transmission and traditional relay schemes are 48% and 30%, respectively; and the energy saving is 32% and 23%, respectively.

2.3.2 System Model

Wireless signal suffers from path-loss, shadowing, multipath fading and other impairments. The transceiver distance d is assumed fixed during the period of each transmission, so the path-loss can be estimated by

$$PL(d)[\text{dB}] = PL(d_0)[\text{dB}] + 10 \cdot \alpha \log_{10}(\frac{d}{d_0}), \tag{2.1}$$

where α is the path-loss exponent, and $PL(d_0)$ is the path-loss at reference distance d_0. $PL(d_0)$ can be calculated by the Friis free-space model:

$$PL(d_0)[\text{dB}] = 10 \cdot \log_{10}(\frac{G_{tx}G_{rx}v^2}{(4\pi)^2 d_0^2 L}), \tag{2.2}$$

where v is the wavelength corresponding to the center frequency f_c, L is the system loss factor, and G_{tx} and G_{rx} are the transmitter and receiver antenna gains, respectively.

The received signal power P_r can be expressed as

$$P_r = \kappa P_t d^{-\alpha}, \tag{2.3}$$

where $\kappa = 10^{PL(d_0)/10}\beta$ equals the reference path-loss times the fading and shadowing factor β, and P_t is the transmission power. The proposed NM can be applied when different links have different values of β. For simplicity, we assume $\beta = 1$ for all links. Then, the received SNR, γ, is given by

$$\gamma = \frac{P_r}{N_0} = \frac{\kappa P_t d^{-\alpha}}{N_0} = \gamma_0(\frac{d}{d_0})^{-\alpha}, \tag{2.4}$$

where γ_0 is the SNR at the reference distance d_0.

2.3.3 Motivation

The design of network modulation is inspired by the example shown in Fig. 2.2. In this example, α is three, the received SNR of SA (link between node S and node A) is 14.77 dB which can support QPSK (2 b/sym), the received SNR of SB is 23.80 dB which can support 64-QAM (6 b/sym), and the received SNR of BA is 19.99 dB which can support 16-QAM (4 b/sym). The first observation is that, to deliver information bits from node S to node A, the path loss between node S and node A is so severe that it is preferable to use node B as a relay. This relay strategy has been widely used in relay networks and multi-hop routing protocol designs.

The second observation is that, if using node B as the relay, when node S transmits bits to node B using 64-QAM, node A can receive a copy of the information. But as the channel quality of SA is so poor, node A may not be able to demodulate any bits successfully from this transmission. In the PHY layer, user cooperative communication has been extensively studied, which tries to use sophisticated signal

processing techniques to combine the two copies of signal node A received (the one from the transmission by node S and the other from the transmission by node B) to enhance system performance. Here, we found that, even without such techniques which require hardware upgrade, we can enhance the relay performance by re-map the constellation of the 64-QAM used by node S, as shown in Fig. 2.3.

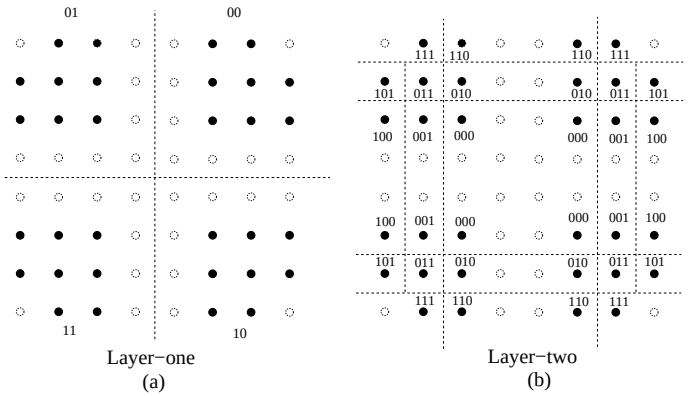


Figure 2.3: A network modulation design.

Using the re-defined constellation map in Fig. 2.3, each symbol contains five bits information represented by the location of the 32 black points. These 32 points belong to four clusters, and the index of the clusters represent two layer-one bits, as shown in Fig. 2.3 (a). Each cluster contains eight points, and the location of them represents three layer-two bits, as shown in Fig. 2.3 (b). The dotted lines in Fig. 2.3 (a) and (b) give the demodulation decision boundaries of the layer-one and layer-two bits, respectively. As the minimum distances from the constellation points to the layer-one decision boundaries are three times that to the layer-two decision boundaries, therefore the minimum SNR needed to transmit the layer-one bits successfully (which is close to the SNR needed for QPSK) is much less than that for the layer-two bits (close to the SNR needed for 64-QAM).

Therefore, when node S transmits a symbol using the NM design shown in Fig. 2.3, node A can demodulate two (layer-one) bits successfully, and node B can demodulate both the two layer-one bits and three layer-two bits successfully. Then, node B just need to relay the layer-two bits to node A using 16-QAM (with 4 b/sym). Node A can receive five bits with $1 + 3/4$ symbol duration, so the throughput using NM-assisted relay is $5/1.75 = 2.857$ b/sym, which is higher than both direct transmission (2 b/sym) and purely relying on node B to relay all bits (2.4 b/sym).

As we use the same energy to transmit each symbol, the energy consumed per

bit is inversely proportional to the throughput. Thus, in the above example, using NM-assisted relay can enjoy both throughput gain and energy saving.

2.3.4 NM Implementation

To implement NM schemes, in the sender side, we can add a mapping function before modulation. Using the above example, node S groups five bits and map them to a six-bit symbol according to the standard constellation map of 64-QAM. For example, "00000" is mapped to "000000" in Gray-mapped 64-QAM constellation, and "00111" is mapped to "001101".

The software re-mapping is so flexible that we can arrange the bit-mapping in different clusters in an arbitrary order such that the decision regions of each symbol can be maximized. For example, we can put the layer-two bits "000" in the neighboring clusters closest to each other as shown in Fig. 2.3 (b). This strategy can slightly enhance the BER performance of layer-two bits.

In the receiver side, both node A and node B can demodulate the symbols using standard 64-QAM demodulator first. Then, node A re-maps each 6-bit symbol to two (layer-one) bits, and B re-maps each 6-bit symbol to three (layer-two) bits.

Therefore, the implementation of NM (for both modulation and demodulation) can rely on software upgrade using the existing standard QAM modems.

2.3.5 BER Estimation

The BERs for layer-one and layer-two bits are determined by the probability that the received signal is outside the decision regions of them, respectively.

Using an AWGN channel as an example, we can integrate the 2-D Gaussian distribution over the area of a layer-one symbol's decision region to get the probability that the corresponding layer-one bits are successfully demodulated, p_s. If this symbol contains m layer-one bits, the BER of layer-one bits is approximately $(1 - p_s)/m$. Similarly, we can obtain the BER of layer-two bits.

Thus, when designing NM to ensure the BER performance, a key issue is to set the minimum Euclidean distance (MED) of the constellation points to the decision boundaries appropriately. On the other hand, the average bit symbol energy is proportional to the square of the Euclidean distance of the constellation points to the origin. Therefore, when we do the re-mapping, we should avoid the points farthest

away from the origin if possible. The design given in Fig. 2.3 has used these two insights to achieve performance gains.

2.3.6 NM Schemes

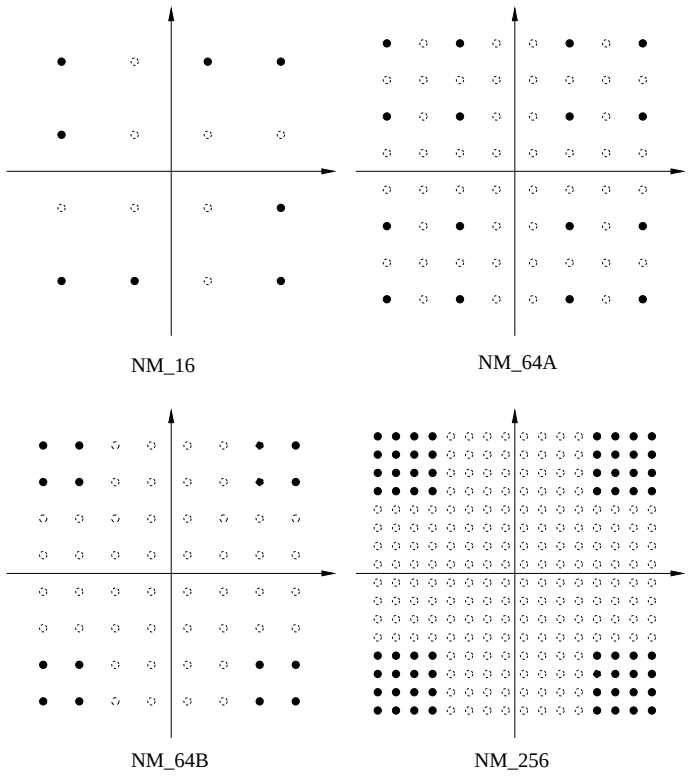


Figure 2.4: Network Modulation Schemes.

So far, in addition to the NM scheme shown in Fig. 2.3, we design four other NM schemes shown in Fig. 2.4. We use these five NM schemes to study how effectively we can improve the network performance. These NM schemes are based on 16-QAM, 64-QAM, and 256-QAM, and each symbol contains two layer-one bits (with larger MED) and one to four layer-two bits (with smaller MED).

Note that we have NOT exhaustively searched the design space of NM yet, which remains an open issue. All existing standard modulations can be viewed as special cases of NM.

2.3.7 NM Applications

In this section, we study how to apply NM in various network scenarios to achieve performance gains.

I. Relay

The first scenario is that a source node delivers information to a destination, with the help of another node which serves as a relay. In this subsection, we first give the necessary conditions that using a relay is preferable, with or without NM. The conditions can be used to identify relay candidates. Then, we propose how to optimally configure NM.

1. Necessary condition of using a relay

With traditional modulation, the direct transmission between the source and the destination can support up-to m b/sym. We have the following theorem.

Theorem 1. *It is preferable to use a relay with traditional modulation if the following condition holds:*

$$\frac{m_1 m_2}{m_1 + m_2} > m, \tag{2.5}$$

where m_1 and m_2 are the bits per symbol rates corresponding to the highest modulation schemes that can be supported by the source to the relay and the relay to the destination, respectively.

For a NM scheme, the source can transmit m_1' bits and m_2' bits to the destination and the relay in one symbol, respectively. The following theorem gives the necessary condition that the NM-assisted relay is preferable.

Theorem 2. *The condition that using the NM-assisted relay is preferable to the direct transmission is*

$$\frac{(m_1' + m_2')m_2}{m_2 + m_2'} > m. \tag{2.6}$$

The LHSs of (2.5) and (2.6) give the throughputs (in terms of bits per symbol) of the relay with traditional modulation and with NM, respectively. The proofs of the above two theorems are straight-forward and we omit them here.

2. Optimal NM configuration

Given an identified relay candidate, the next question is how to choose the best NM scheme under the constraint of BER, p_e. Denote $p_{e,sr}$, $p_{e,rd}$, and $p_{e,sd}$ the BERs of the source (S) to the relay (R), the relay to the destination (D), and S to D,

respectively, which depend on the modulation schemes used and the received SNRs. Assume that all nodes transmit using the same power P_t. We can formulate the following optimal NM configuration problem.

Problem 1. *(P1)*

$$\begin{aligned} \max \quad & \frac{(m_1{}' + m_2{}')m_2}{m_2 + m_2'}, \\ \text{s.t.} \quad & p_{e,sr}(m_2', P_t) + p_{e,rd}(m_2, P_t) \leq p_e \\ & p_{e,sd}(m_1{}', P_t) \leq p_e \end{aligned}$$

As the number of modulation schemes that we can choose is limited (five proposed NM schemes and four traditional modulation schemes), it is feasible to solve the above optimization problem by exhaustive searching.

3. Optimal power allocation

As the channels between SR and RD might be quite different, appropriate transmission power allocation might be desirable to further enhance the system performance. Denote the total power budget by P. The optimal configuration considering power allocation can be formulated as the following optimization problem.

Problem 2. *(P2)*

$$\begin{aligned} \max \quad & \frac{(m_1{}' + m_2{}')m_2}{m_2 + m_2{}'} \\ \text{s.t.} \quad & p_{e,sd}(m_1{}', P_1) \leq p_e \\ & p_{e,sr}(m_2', P_1) + p_{e,rd}(m_2, P_2) \leq p_e \\ & P_1 + P_2 \leq P_{\max} \\ & P_1 \geq 0, \quad P_2 \geq 0 \end{aligned}$$

The optimization problem (P2) is a mixed-integer problem which is difficult to solve, especially without the close-form expression of BER. Considering the power control in a practical system is discrete, P_1 and P_2 are discrete variables with limited number of values. Using exhaustive searching is still tractable as the searching space is moderate. We set the number of power levels to be 100 here, which is sufficient for a practical system.

II. Broadcast and Multicast

When a node broadcasts or multicasts to a group of receivers using wireless transmissions, it needs to adjust the modulation so the receiver with the worst channel quality can successfully receive the message. In this case, we can use the similar strategy as discussed in the previous subsection to enhance the system performance. That is, first identify relay candidates for the worst-channel receiver; then use the optimal NM configuration algorithm to select the appropriate NM scheme to maximize the throughput and save energy. Then we repeat the procedure to consider the second worst channel receiver. Due to space limitation, we do not discuss this problem in detail.

III. Downlink Unicast

Next, we consider the scenario that, in an infrastructure-based network, a base station (BS) or an access point (AP) needs to send different messages to users using the downlink. We can also use NM to save the downlink bandwidth and energy. If we let an idle node to act as a relay for an active user, the problem is again the same as the relay problem discussed in subsection 2.3.7, and we omit it here. If only active nodes can collaborate with each other, the problem is different, which is the focus of this subsection.

Let $\mathcal{F}$ be the active user set (the downlink receiver set), and each user needs to receive one unit of data per scheduling period from the BS. Using NM, the BS can transmit messages to multiple users with different channel quality simultaneously to achieve performance gain. As a starting point, we use NM to combine messages to two users only, and how to combine the messages belong to more than two users remains a further research issue.

IV. Uplink Unicast

The uplink scenario is similar to the downlink case. A user may need to rely on other nodes (who can be either active or inactive) to relay messages for it. We can also use the WLF matching algorithm to group nodes into pairs if using NM is preferable. The NM configuration problem for each pair can also be solved by global searching over the extended modulation set and we can also take power allocation into account.

2.4 Error-correction Coding Assisted Relay

2.4.1 Motivation

Still consider the three-node topology in Fig. 2.2. For the transmission from S to R, D also receives a copy. Due to the worse channel condition of SD, the received copy by D may contain severe errors. In the existing solutions of relay networks, D solely relies on the transmission from R to decode the message from S; in the traditional cooperative systems, D combines two copies of transmissions from S and R to decode the message [64, 84]. The relay approach is simpler, but it wastes the erroneous copy from S to D entirely. The second cooperative communication approach can utilize both copies to achieve the spatial diversity gain, but it requires sophisticated signal processing and, more importantly, transmitting two copies of the same information bits increases the cost.

Here, we investigate a new direction for cooperative transmissions. The transmission takes two steps. In the first step, S transmits a packet to R and D. In the second step, R uses the received packet to generate a smaller packet that contains the error-correcting code (ECC) of the original packet. Then, D can use the corrupted copy of the original packet from the transmission by S and the ECC from the transmission by R to recover the original packet.

2.4.2 Selection of ECC

Error coding is a well developed yet still very active area. Different ECCs have different applicable scenarios considering the trade-off of overhead, performance, and implementation complexity. In traditional applications, both the message and the parity bits are transmitted over the same channel and thus with equal error probability. For EAR, the transmissions from S and R to D are over independent channels with different BERs due to the spatial diversity.

Given the special scenario of EAR, the desirable ECC should satisfy three properties. First, it must be a systematic code, because the relay node needs to receive the original packet before it can encode it. Second, it should be scalable, so the code length can be adjusted according to different error correction requirements. Third, the computational cost should be acceptable for current hardware platforms. Considering these requirements, we choose the Reed-Solomon (RS) codes, as they satisfy the requirements above and there are many high performance software and/or hardware

implementations readily available.

2.4.3 RS-code Configuration

The RS-code is a systematic linear block code and it is organized into fixed-length blocks. Each block contains b bits and the encoding and decoding operations are in Galois field G(2^b) [60]. Since modern computing systems are byte oriented, it is straightforward to choose the block size of 8 bits to be compatible with the mainstream hardware and software systems. Therefore, we use GF(2^8) consisting of $2^8 - 1$ blocks to construct the RS-code.

Denote the RS-code by a two-tuple (n, k), where $n = 255$ is the length of encoded blocks and k is the length of message blocks (in bytes). Their difference, $r = n - k$, is the length of additive parity blocks appended to the k message blocks. RS-code (n, k) can correct up to $t = r/2$ block errors. RS-code with longer parity blocks (i.e., larger r) can correct more bit errors at the cost of a higher overhead, so we need to select the code length carefully.

Assume that RS-code is employed during all transmissions for fair comparison. Given the BER, the block error rate, P_E, can be calculated as $P_E = 1 - (1 - BER)^b$. Different from the conventional configuration of RS-code in point-to-point communications, here, we need to consider the BERs of both the direct transmission and the one by the relay. The corresponding block error rate are denoted as P_E^{SD} and P_E^{RD}, respectively, which can be estimated by R.

Given the RS(n, k) code, the probability that D can decode the k-block message correctly by combining the ECC from R and the original message from S is

$$P_{c,EAR}(r) = \sum_{i=0}^{\frac{r}{2}} \binom{k}{i} (P_E^{SD})^i (1 - P_E^{SD})^{k-i} \cdot \left[\sum_{j=0}^{\frac{r}{2}-i} \binom{r}{j} (P_E^{RD})^j \left(1 - P_E^{RD}\right)^{r-j} \right]. \tag{2.7}$$

When wireless channels are worse than the estimation and the chosen ECC is insufficient to to recover the original packet, a failed transmission occurs. We have two choices: (a) to transmit another stronger ECC, and (b) to retransmit the whole packet. The first scheme may require less channel time but it is difficult to determine how many bits are in error and which ECC code is sufficient. Thus, we let the relay

node retransmit the whole packet if the ECC attempt fails for simplicity.

We denote the maximum retransmission limit by R as $R_{\max}$, and the RS-code length for the retransmission by R as r_{RD}. The probability that the packet can be successively received by D by a retransmission is

$$P_{c,r}^{(1)} = \sum_{i=0}^{r_{RD}/2} \binom{k + r_{RD}}{i} (P_E^{RD})^i (1 - P_E^{RD})^{k+r_{RD}-i}. \tag{2.8}$$

If the previous retransmission fails, the message will be retransmitted again till reaching the retransmission limit. Hence, the probability that the message can be successfully received by retransmissions is

$$P_{c,r} = \sum_{i=0}^{R_{\max}-1} (1 - P_{c,r}^{(1)})^i P_{c,r}^{(1)}. \tag{2.9}$$

Thus, for each successfully transmitted $L(r) = k(P_{c,EAR} + (1 - P_{c,EAR})P_{c,r})$ bytes, the expected transmitted data length (in bytes) by R is

$$\begin{aligned} U(r) = {} & r + (1 - P_{c,EAR})(k + r_{RD}) \\ & \left(\sum_{i=0}^{R_{\max}-1} (i+1)(1 - P_{c,r}^{(1)})^i P_{c,r}^1 + R_{\max}(1 - P_{c,r}^{(1)})^{R_{\max}} \right). \end{aligned} \tag{2.10}$$

Given R using the modulation type with m_{RD} bit/sym, the corresponding channel time is $T(r) = \frac{U(r)}{m_{RD}}$ symbol duration.

Based on $T(r)$, the expected throughput of the second hop (RD) is $E_t(r) = \frac{L(r)}{T(r)}$ (bytes per symbol duration), which should be maximized.

Discussion: As there are many efficient implementations of RS codec, the extra computation cost of EAR is moderate. Experimental results in [59] showed that the computation cost of RS encoding and decoding is less than 10% of the capacity for a commodity computer. If using FPGA to implement RS codec, the efficiency is even higher. For example, the decoding rate is up to 496 Mbps on Spartan-II by Xilinx [1].

2.5 Performance Evaluation and Discussion

Extensive simulations have been conducted to evaluate the performance of NM and EAR in different scenarios with different network topologies. For comparison purpose,

we also implement the traditional relay scheme based on adaptive modulation. We use the same average symbol energy for all transmission schemes. Thus, if a scheme can achieve a higher throughput, it can also save energy (using less energy per bit).

2.5.1 Line Topology

We first use a linear topology, where three nodes S, R, and D, are located on a line, as shown in Fig. 2.5.

The distance between node S and node D is fixed at 1 unit with the corresponding reference SNR equal to 12 dB (15.85). α is set to three, so the average received SNR at distance x unit equals $15.85 \cdot (x)^{-3}$ and the AWGN channel model is considered. The tolerable BER (before error coding) is set to $p_e = 10^{-3}$. Therefore, with direct transmission, node S can transmit to node D at most 2 b/sym using QPSK. The distance between node R and node S, x, is varying between 0.05 and 0.95 to evaluate the system performance under different channel conditions.

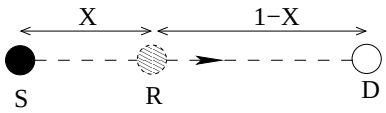
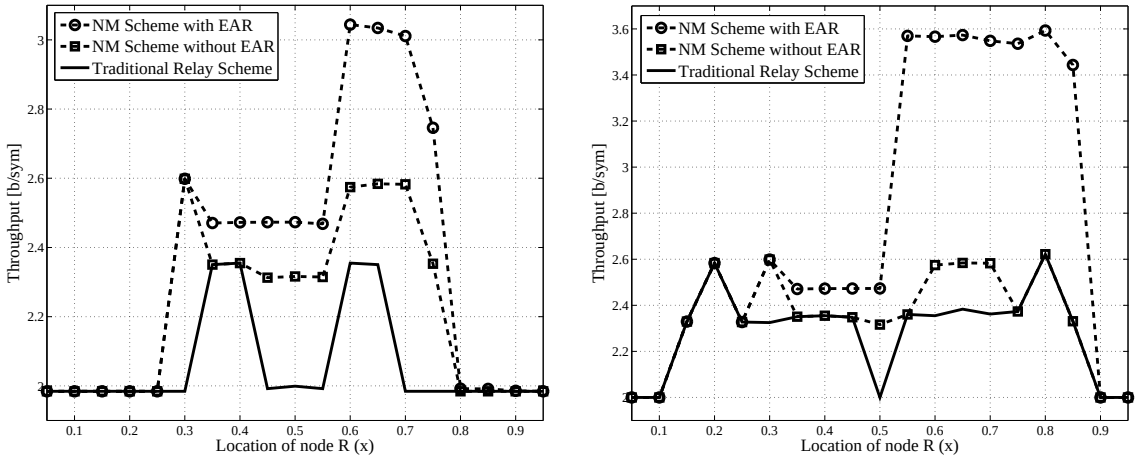


Figure 2.5: Line topology.

I. Single-receiver case



(a) Equal power allocation (b) Optimal power allocation

Figure 2.6: Throughput comparison, single-receiver.

First, we evaluate the performance of the single-receiver case, i.e., S transmits to D using R as a relay. In Fig. 2.6(a) and (b), we compare the expected throughput under the equal power allocation and optimal power allocation, respectively. In these figures, x-axis represents the location of node R, and y-axis is the expected throughput (i.e., the expected number of bits received by node D divided by the expected number of symbol durations). For the direct transmission, the throughput is 1.984 b/sym considering packet retransmissions and error-control coding.

As shown in Fig. 2.6(a) and (b), first, to outperform the direct transmission, the location of R can be in a wider range using the NM-assisted relay schemes than that using the traditional relay schemes. Second, the proposed NM with a fixed transmission power can outperform the traditional relay schemes (by up to 30% for $x = 0.3$ or 0.7); if combined with the EAR scheme, it can achieve an even higher throughput and outperform the traditional relay scheme by 50% at $x = 0.7$. The improvement comes from twofold: 1) using the NM scheme, two-layered bits are transmitted simultaneously and only the L2 bits need to be relayed; and 2) with the EAR scheme, the relayed packet contains the RS parity bits, which is much smaller than the original packet.

Furthermore, compared the results in Fig. 2.6(a) and (b), the system throughput can be further improved by the optimal power allocation between S and R, and NM/EAR can outperform the traditional relay by around 50% for $x \in (0.55, 0.85)$. The results demonstrate the importance of power allocation. In the following, all the results are obtained with the optimal power allocation.

The throughput curves in Fig. 2.6 are in zig-zag patterns. This is because both the number of modulation schemes and the RS code length are limited and cannot be adjusted continuously.

The throughput curve of the traditional relay scheme is symmetric with respect to $x = 0.5$, while those of the NM-assisted relay schemes are asymmetric. In specific, when $x < 0.5$, the performance gain is mainly from the NM; when $x > 0.5$, EAR can achieve a substantial improvement. This is because, with EAR, when R is closer to D, the SNR of link SR is closer to that of SD, we need a less strong RS code (fewer parity bits) to help D to successfully decode the packets. When R is very close to S, R has to retransmit all L2 bits similar to the traditional relay scheme, so the performance gain of EAR largely disappears.

II. Two-receiver case

Next, let node S transmit different messages to both node A and node B. The throughput results are given in Fig. 2.7. In this scenario, the NM can achieve even higher gains than the traditional schemes. From the figure, the NM schemes can achieve up to 72% and 91% throughput gains than the traditional schemes with or without relay, respectively. The preferable region of the relay (in which the NM-assisted relay schemes can outperform the direct transmission schemes) is wider than that for the single-receiver case. This is because that, for the two-receiver case, with NM, the data to R can be piggybacked with the transmission to D, which can achieve a much higher throughput gain than using R as a relay only.

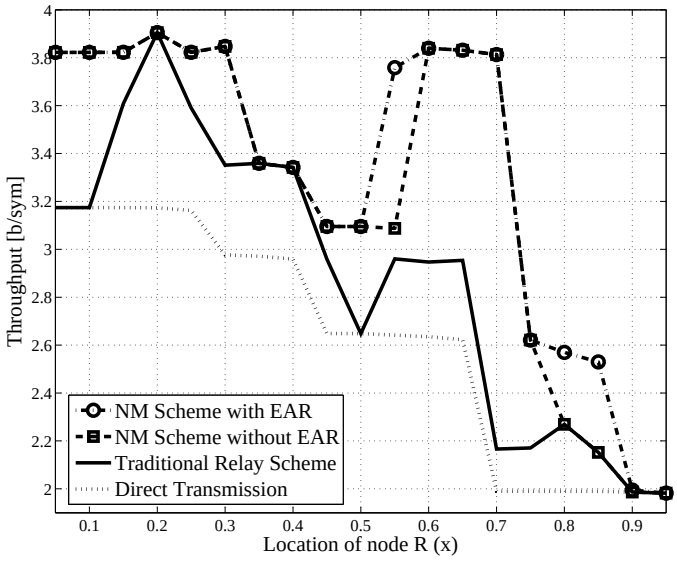


Figure 2.7: Throughput comparison, two-receiver.

Another observation is that the preferable relay region (in which relay schemes can outperform direct transmission schemes) is different for the single-receiver case and that for the two-receiver case. This is because that the throughput of single-receiver case is highly dependent on the rate of the second hop, which should support more than 2 b/sym in order to outperform the direct transmission schemes. Thus, when B is too close to S, using relay with or without NM cannot achieve higher throughput. This is not the case for the two-receiver case, where the transmissions are finished in two stages: unicast transmission and relay transmission (if transmission for node B finishes first) or direct transmission (if transmission for node A finishes first). In the unicast stage, the data for different nodes can be transmitted simultaneously,

so the overall throughput is high. Thus, when node B is close to node S, even though the second hop channel condition is not good enough to support modulation schemes higher than QPSK, the NM-assisted transmission scheme can still achieve throughput gain. Only when node B is very close to node A, we cannot obtain any gain by traditional or NM-assisted relay schemes.

However, for the two-receiver case, the EAR scheme can only achieve a marginal gain. This is because, using NM, the percentage of the data to be relayed by R to D in the two-receiver case is much lower than that in the previous one-receiver case.

Note that these results are sensitive to the received SNR of SD. If the received SNR of SD changes, the desirable relay locations and the performance gains of NM and EAR also change. Nevertheless, the overall trend will remain the same.

The results with the linear topology not only provide important insights of where the performance gain comes from and for which scenarios the proposed scheme is most useful, but also indicate where the desirable relay nodes should be located.

2.5.2 Network Scenario

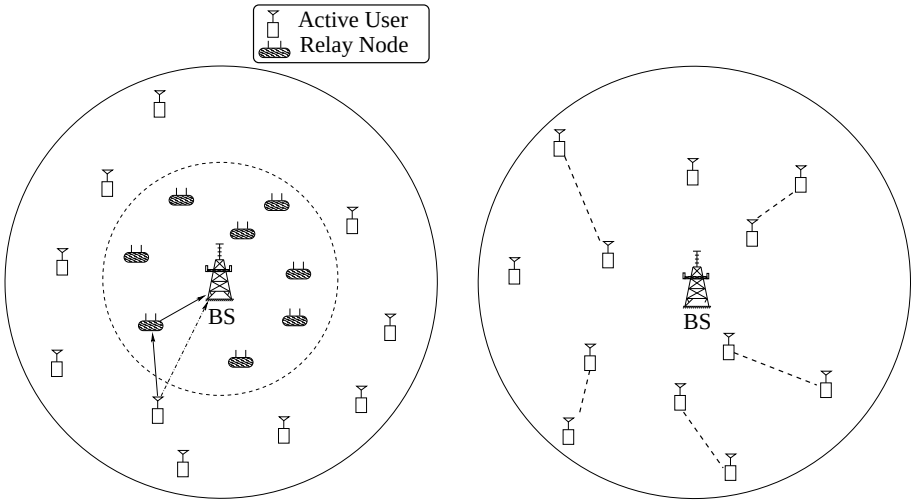


Figure 2.8: Network topology.

We further evaluate the performance in a two-dimensional network, where a base station (BS) serves all the active users uniformly distributed in the cell as shown in Fig. 2.8. For simplicity, we assume that there is no interference for the downlink and uplink transmissions, and each node has the same amount of data to transmit

for both the downlink and the uplink. Assume that the BS has the knowledge of the network topology and channel conditions that are stable during the scheduling period.

First, grouping users into pairs for cooperative transmission can be formulated as the following weighted non-bipartite matching problem, which is applicable to both the uplink and downlink scenarios. Denote the set of the mobile users in the cell by $\mathcal{F}$. Denote by $R_{i,j}$ the throughput for a pair of users i and j, and R_i and R_j their (downlink or uplink) individual throughput with traditional modulations, respectively. The gain of using NM (with or without EAR) for users i and j is $G_{i,j} = \max(1/R_i + 1/R_j - 2/R_{i,j}, 0)$. The optimal matching will maximize the sum of $G_{i,j}$ for all users.

If the $G_{i,j}$ for all users in $\mathcal{F}$ is known, the state-of-the-art optimal matching algorithm can solve the above matching problem with the time complexity of $O(N^3)$ [25]. However, obtaining $G_{i,j}$ for all pairs is nontrivial. Thus, instead of using the optimal matching algorithm, we use a simple heuristic algorithm called worst-link-first (WLF) matching [67] to group users: the BS first sorts the channel qualities of all users; then it selects an unmatched user j with the worst channel quality and finds another unmatched user i who is most desirable for j (i.e., the pair can achieve the highest performance gain); it repeats the above procedure till all users are matched. If using relay schemes (with NM and/or EAR) for a pair cannot achieve any throughput gain, the direct transmission schemes will be used. We also use the optimal power allocation for all schemes and all pairs.

Network throughput

Network throughput is used to evaluate the performance of different schemes, which is the total amount of information transmitted in the whole network over the total channel time consumed.

To consider different node densities, we vary the number of active users from 2 to 50 with the step size of 4. Set the SNR at the cell boundary as 11 dB (which can support QPSK using the traditional modulation scheme). For each node density, we repeated the simulation $1,000$ times using different randomly generated topologies, and calculated the average results. In the following figures, the network throughputs are normalized to the one with direct transmission (which adopts the adaptive modulation scheme) for easy comparison.

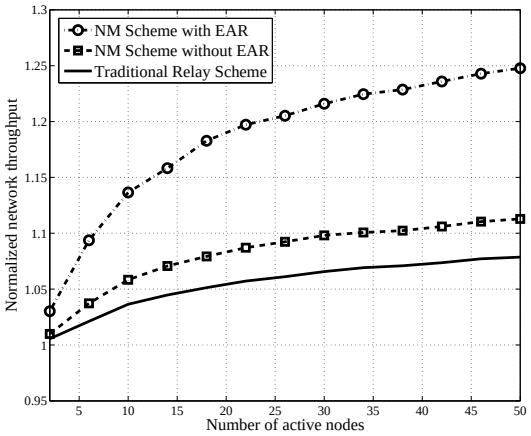


Figure 2.9: Network throughput (uplink).

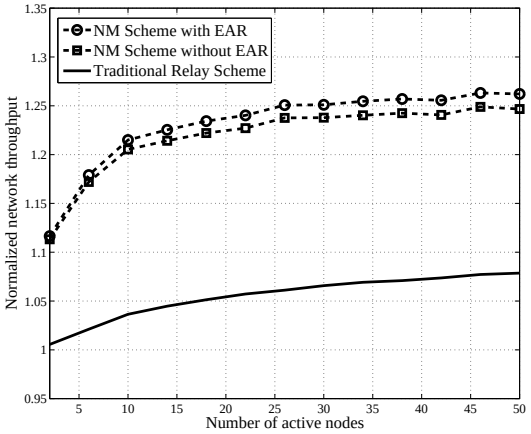


Figure 2.10: Network throughput (downlink).

The normalized network throughputs for uplink and downlink are shown in Fig. 2.9 and Fig. 2.10, respectively. We note that the downlink and uplink throughputs are the same with the traditional relay schemes. This is because, with the traditional relay schemes, the most preferable relay nodes are at the same or symmetric locations for the downlink and uplink transmissions. Thus, the same node will serve as the relay for both uplink and downlink transmissions. However, the throughput of downlink and uplink may not be the same with NM-assisted relay, since different relay nodes might be chosen for uplink and downlink transmissions. Thus, with NM, the workloads of relay nodes can be more balanced, which is desirable for energy-constrained wireless networks.

From the simulation results, the traditional relay scheme can achieve a marginal

network throughput gain, while the proposed NM with EAR scheme can achieve a 25% throughput gain.

Bit-energy

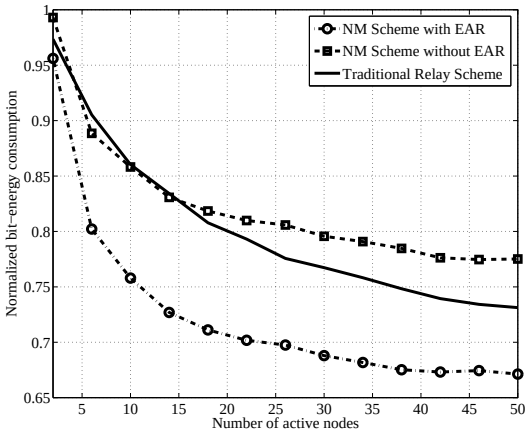


Figure 2.11: Per bit-energy consumption (uplink).

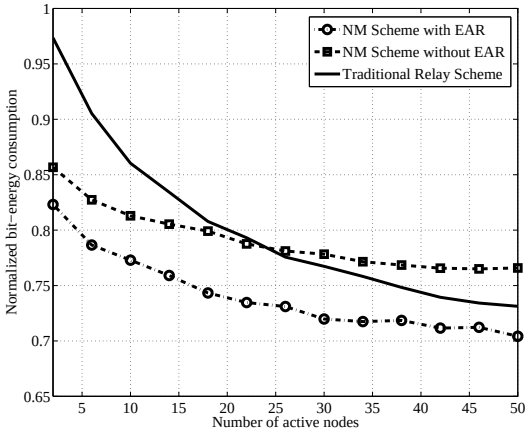


Figure 2.12: Per bit-energy consumption (downlink).

In addition to the throughput gain, the average bit-energy consumption is also reduced by using NM with EAR. Fig. 2.11 and Fig. 2.12 show the bit-energy results, which are normalized to the bit-energy of direct transmission. With more than 30 users in the cell, for the uplink case, the NM without and with EAR schemes can reduce the bit-energy by more than 20% and 31%, respectively, compared to the direct transmission. For the downlink case, the NM scheme without and with EAR

can reduce the bit-energy consumption by more than 22% and 28%, respectively. Note that there is a tradeoff between energy consumption and throughput. How to minimize bit-energy consumption while maintaining certain throughput using NM and EAR can be formulated as a different optimization problem which can be an interesting further research issue.

Further discussions

The gain by the NM is more significant for the downlink transmissions while the gain by the EAR is more significant for the uplink transmissions, because the uplink scenario is similar to the one-receiver case and the downlink scenario is similar to the two-receiver case in the linear topology.

For the uplink, a higher node density can lead to a much higher gain as the performance of EAR is more sensitive to the relay location. For the downlink, even when the network only has two active users, using NM can achieve around 11% throughput gain on average.

Note that we only consider five NM schemes and the RS(255, $*$) codes in this dissertation work. If more well-designed NM schemes and RS codes are used, it is possible to further improve the network performance.

2.6 Conclusions

In this chapter, we have presented the detailed design of the network modulation and the error-correction coding assisted relay schemes, which take the network topology and channel conditions among multiple nodes into consideration to explore the spatial diversity and broadcast nature of wireless communications. The proposed NM is based on the mainstream QAM modem, and the proposed EAR uses the widely-used RS codes. Therefore, they are easy to be adopted in practical wireless systems. Extensive simulation results have demonstrated the substantial performance gains and lower bit-energy consumption of the proposed schemes.

2.7 Symbol List

Symbol	**Explanation**
d_0, d	transceiver distance
$PL(\cdot)$	path-loss
G_{tx}, G_{rx}	transmitter and receiver antenna gains
P_t, P_r	transmitted and received signal power
α	path-loss factor
γ	received SNR
m	number of bits transmitted by a symbol
p_e	BER
P_E	block error rate
$P_{c,EAR}(\cdot)$	decode probability
(n, k)	RS-code
$U(\cdot)$	expected transmitted data length
$T(\cdot)$	channel time (in symbol duration)
$E_t(\cdot)$	expected throughput
$G_{i,j}$	performance gain by grouping user i and j
R	link throughput

Chapter 3

Hexagonal Modulation and Non-binary Error Control Coding

3.1 Introduction

As mentioned in Chapter 2, the QAM is one of the most popular design principles where two quadrature (90-degree out-of-phase) carrier waves (usually sinusoids) are changed (i.e., modulated) in their amplitudes and mixed to represent the coded information bits. The most common signal constellations are 16, 64 and 256-QAM which carry 4, 6 and 8 bits per symbol, respectively [37]. Other modulation schemes such as BPSK and QPSK share the same principle.

At the receiver end, the QAM demodulator converts the received signal, which may be corrupted by fading, noise and interference, to bits. The signal space is partitioned into decision regions for this purpose, and demodulation is equivalent to determining the region that contains the received signal. With the conventional QAM, the signal space is partitioned into rectangular decision regions, however, it is well known that a two-dimensional regular tiling with hexagons is the most efficient packing in terms of compactness [28]. Therefore, if hexagonal decision regions are employed to partition the signal space, referred to as hexagonal modulation (H-QAM), the spectrum and/or energy efficiency can be improved. H-QAM maximizes the minimum distance between symbols and thus minimizes the symbol error probability for a given signal energy, as well as the peak-to-average power ratio, which is important for OFDM systems [88, 31, 73, 22].

However, H-QAM schemes have not been widely adopted in wireless systems yet

because the number of the constellation points may not be the integer power-of-two and thus it has the inherent difficulty to be compatible with current information systems, which usually use the binary digits to represent information. To improve the compatibility, there are two existing approaches [88, 31, 73]. One is to convert the binary data stream to non-binary symbols, e.g., ternary, at the transmitter and reverse the procedure at the receiver. However, this scheme suffers from the error propagation problem, i.e., a one non-binary digit error may cause several binary digit errors and thus degrade the BER performance. The other approach is to leave some symbols unused, but this reduces the efficiency which limits the gains possible with H-QAM.

Thus hexagonal modulation alone can only provide marginal performance improvement. To fully realize the potential of H-QAM, it is necessary to go beyond binary bit-mapping and coding. Thus, in this dissertation we not only employ H-QAM , but also break the limitation of binary bit-mapping and coding, i.e., using ternary digits (trits) to carry information and conduct error correction to achieve a higher spectral and energy efficiency. Interestingly, ternary architectures have been considered in computing and storage systems due to their higher radix economy and the three usable states for certain electromagnetic materials [50, 49]. Even though ternary communication and computing systems have not yet reached commercially viability, their future use has been predicted by Knuth [47].

With this new approach, the main contributions of this chapter are three-fold. First, we propose new modulation schemes based on hexagonal tiling and evaluate the corresponding BER performance. The new H-QAM schemes proposed in this dissertation contain 3, 6, 8, and 12 constellation points to represent 1 trit, 1-bit plus 1-trit, 3-bits, and 2-bits plus 1-trit, respectively. Second, we employ ternary convolutional coding to protect the ternary digit directly. For H-QAM with hybrid bit and trit information, we consider a combination of binary and ternary coding. The BER performance is evaluated for different modulation schemes, including the conventional QAM, with code rates 1/2 and 3/4 to conform to the IEEE 802.11 standard [75]. These results show that increasing the number of modulation schemes does not lead to a large increase in AMC schemes, and will be discussed later. Third, due to the ring structure of hexagonal modulations, bits and trits can naturally have different levels of protection against transmission errors. For example, using an appropriate information mapping in H12-QAM, the trit (with roughly 1.5-bit information) can have a lower BER than the bit with the same average received SNR, due to the large

inter-symbol-cluster distance. This intrinsic feature is especially favorable for scalable video applications (e.g., base vs enhancement layers in layered video coding, and I, P and B frames in MPEG-like video coding schemes) due to the different levels of sensitivity to information loss on user-perceived video quality. The last but not the least, extensive simulation results are presented to demonstrate the efficiency of the proposed scheme, which can provide considerable performance gains compared to existing binary systems. A prototype H-QAM wireless communication system was developed with non-binary information mapping using GNU Radio and USRP2, a commonly used software-defined radio (SDR) platform [23].

3.2 Background and Related Work

3.2.1 Hexagonal Signal Constellations

Modulation is the process of converting a data stream to waveforms suitable for transmission through a communication channel. For bandwidth limited bandpass modulation, quadrature amplitude modulation (QAM) is commonly employed. Typical QAM constellations can be considered as rectangular partitions of the two-dimension signal space. It is well known that regular hexagons provide the densest two-dimension packing, and this has motivated research into the potential of H-QAM [31, 88, 73, 32]. Existing work on hexagonal modulation usually considers one of the following approaches. The first considers binary data, so that each modulation symbol represents an integer number of bits. However, this scheme requires that the number of constellation points be a power-of-two. Since the number of H-QAM constellation points is not a power-of-two [31], some of the constellation points are not used, which is a waste of available resources.

The second approach uses all points in the hexagonal constellation for transmission to maximize the per-symbol throughput. It has been shown in [88] that hexagonal-18 QAM modulation (H18-QAM) requires less energy per-bit than 16-QAM. As 18 can be factorized as $2 \times 3 \times 3$, one H18-QAM symbol can be decomposed into one bit and two trits. To accommodate a binary data stream, binary symbols can be mapped to ternary symbols using binary-input ternary-output (BITO) convolutional or turbo codes to also provide error correction [88]. However, this conversion is not flexible and suffers from poor performance since it fails to offer adequate protection for the ternary digits. According to the simulation results, the coded error performance of

H18-QAM is 0.6 dB lower than that of the coded 16-QAM, so in this case hexagonal modulation provides no improvement over rectangular modulation.

It is clear that using binary data with non-binary modulation may be suboptimal as some useful constellation points are not used, and employing BITO codes for error correction leads to the inflexibility in multiplexing bits and trits. An alternative approach is to combine H-QAM with non-binary, in particular ternary, coding.

3.2.2 Ternary Computing and Communications

Radix economy is used to measure the cost of storing or transmitting numbers in a base [33]. It is defined as the number of digits needed to represent a number N in the base of b, multiplied by the radix b. A base with a lower radix economy has a higher efficiency. It has been proven that a radix of three provides the lowest radix economy among all integer bases. Thus, the cost of storing numbers can be minimized if a ternary base is used [33]. This implies that a ternary base outperforms the widely used binary base in terms of both storage and communications. Ternary based data recording/storing systems have been investigated [49, 50]. Computers using balanced ternary logic were implemented in the late 1950's and were shown to be more efficient in ternary than binary. Although the interests in ternary computing have diminished, Knuth has predicted that the elegance and efficiency of ternary logic will bring it back in the future [47].

As most existing information and computing systems are binary based, ternary has not received much attention in wireless communication systems yet. For example, the implementation of non-binary error correction coding such as Reed-Solomon codes has been focused on binary extension fields GF(2^s), where s is an integer [19]. Thus the number of known ternary error correcting codes suitable for communication applications is limited.

As discussed in Sec. 3.2.1, the use of ternary codes with hexagonal constellations can fully realize the potential of H-QAM and improve the performance of wireless communication systems. Thus, we have conducted extensive literature scan to find out a ternary convolutional code and applied appropriate puncturing schemes to obtain different code rates, and the details will be discussed in Section 3.3.3.

3.2.3 Differentiated Services

In many multimedia applications, some types of information are more important than others in terms of how they are perceived by users, so we should treat them differently to fully utilize the limited resources. We use scalable video as an example, which can encode the video stream into a base layer and several enhancement layers to provide spatial, temporal, and quality scalability. For each video frame, if the receiver can receive its base layer successfully, a low quality video can be played back, and the quality is improved when more enhancement layers are received. To improve the coding efficiency, the correlations among the information in different layers are used to compress the higher layers. Consequently, if the lower layer information is lost during transmission, the corresponding higher layer information becomes useless. Given the different importance of different layers, it is essential to provide differentiated services for information in different layers.

Differentiated Services (DiffServ) can be provided in different layers of the network protocol stack. Providing DiffServ at the coding and modulation level has the advantage of easy synchronization and the robustness against varying wireless channel conditions: when multiplexing the information from different layers and transmitting it in the same symbol, different layers of the same frame can be received *simultaneously*. When the wireless channel quality is good, all layers can be received successfully; when the channel quality is poor, the most important layers can be received successfully and the other layers may be corrupted or lost. In the literature, hierarchical modulation has been proposed to support differentiated services, and the unequal protection of QAM constellation mapping has also been used to support data bits of different importance [83]. Different from the previous work, here we use the intrinsic performance difference of bits and trits multiplexed in each hexagonal symbol to support layered video applications.

3.2.4 Modulation vs. Coding Gain

While low density parity check (LDPC) codes have been shown to provide near capacity performance [81], complexity remains a challenge.

Capacity achieving LDPC codes are very complex to implement because long code lengths are required (millions of bits). In addition, the BER after demodulation should be sufficiently low. If this BER is below a given threshold, coding can be employed to reduce it to a negligible level. Otherwise, it is difficult for coding to

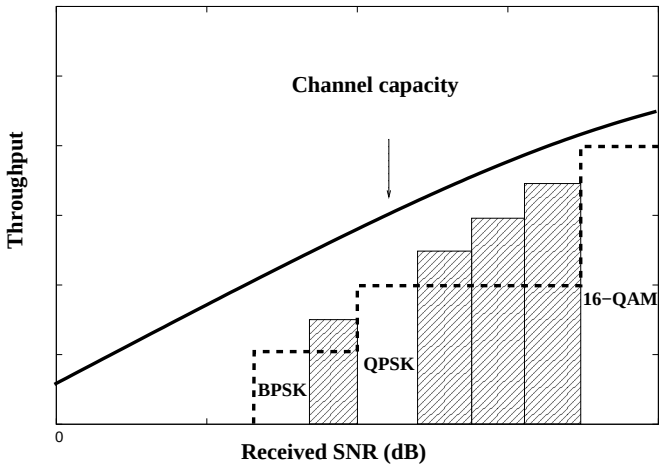


Figure 3.1: Illustration of achievable throughput with different modulation schemes.

reduce the BER, and the BERs with and without coding are similar. Thus given the SNR at the receiver, an appropriate modulation scheme can be used to ensure that the BER after demodulation is below a threshold (often 10^{-3} for wireless systems), and then error control coding is applied to reduce this BER to an acceptable region (e.g., below 10^{-6}). As error control coding is not effective unless the uncoded BER is sufficiently low, a better modulation scheme can help to improve the overall system performance. For wireless communication systems employing AMC, several combinations of modulation and coding schemes are usually adopted to fit with different channel conditions. In this chapter, we also consider different code rates for H-QAM. Introducing H-QAM modulation does not lead to much more AMC schemes as H-QAM with coding can outperform and thus replace some of the existing, conventional AMC schemes.

In addition, if the SNR gap between two AMC schemes is large, e.g., QPSK with a rate 3/4 code and 16-QAM with a rate 1/2 code, this gap can be filled with an H-QAM based scheme. As illustrated in Figure 3.1, since the SNR at the receiver is continuous, there is space (shaded areas) for new modulation schemes (using new or existing coding schemes) to improve the system performance, which motivates the work reported in this chapter.

3.3 System Design

3.3.1 Constellation Geometry

A signal can be represented in the signal space domain using an in-phase and quadrature-phase (I/Q) constellation diagram. For a constellation with N points, the information carried in each symbol equals $\log_2 N$ bits (or $\log_3 N$ trits). The distance between a constellation point to the origin, d, is proportional to the square root of the transmitted symbol energy. In the absence of noise, the received signal constellation has the same shape as the transmitted constellation, except that, at the receiver, the distance from a constellation point to the origin is proportional to the square root of the received symbol energy. In the following, *constellation* refers to the constellation at the receiver unless otherwise stated.

In an additive white Gaussian noise (AWGN) channel, a received symbol follows two-dimensional Gaussian distribution centered at the corresponding constellation point. A Voronoi diagram can be used to determine the decision boundary of each

symbol. The probability that a symbol is demodulated in error is equal to the probability that the received symbol lies outside the decision region of the intended symbol.

Given the requirement that the BER after demodulation should be low (e.g., below 10^{-3}), and the fact that the Gaussian distribution decays exponentially, we can approximate the symbol error probability for H-QAM by

$$2Q\left(\sqrt{\frac{2r^2}{N_0}}\right), \tag{3.1}$$

where $Q(\cdot)$ is the Q-function, r is the shortest distance from the constellation point to its decision boundary, and N_0 is the noise spectral density. As r is equal to half of the minimum Euclidean distance in the signal space between two constellation points, the BER is a function of the minimum Euclidean distance between constellation points.

The optimal design of a modulation constellation can be considered as the maximization of the number of constellation points under the average power (symbol energy) and BER constraints. Using (3.1), this can be converted to the following geometry problem. In a circle of radius $\sqrt{E}+r$, pack as many non-overlapping circles with radius r as possible, where E is the maximal received symbol energy, and r is determined by the BER constraint.

When the radius of the region, $\sqrt{E}$, is sufficiently large, the optimal packing is a hexagonal tiling. Comparing this tiling with the rectangular tiling widely used in existing QAM schemes, a hexagonal tiling can cover the region more efficiently, which leads to approximately a 0.6 dB gain [24].

The Proposed H-QAM Geometry

Given the commonly employed modulation schemes such as BPSK, QPSK, and 16-QAM, we propose to add ternary PSK (TPSK), H6-QAM, H8-QAM, and H12-QAM schemes to fill the gaps in Figure 3.1. Each received symbol can provide from one to four bits of information with a granularity of about 0.5 bit. We use these signal constellations to demonstrate the benefits of hexagonal modulation combined with ternary coding.

Given a fixed number of constellation points, the optimal constellation arrangement problem is slightly different from the circle packing problem. Conditioned on maximizing the minimum Euclidean distance, the *average* symbol energy should be minimized. This is equivalent to minimizing the average of the square of the distances

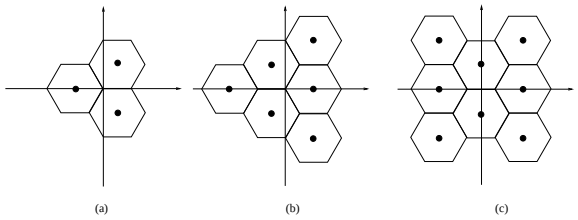


Figure 3.2: TPSK, H6-QAM, and H8-QAM constellations.

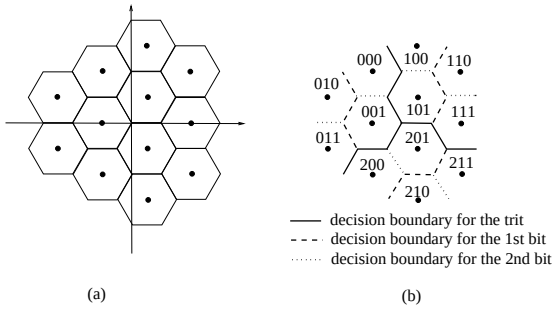


Figure 3.3: Constellation and mapping for H12-QAM.

from the constellation points to the origin under the condition that the minimum distance of each constellation point to its decision boundary is no smaller than r, i.e.

$$\min \frac{\sum_i d_i^2}{M}, \tag{3.2}$$

where d_i is the distance of the i-th constellation point to the origin and M is the number of constellation points. Thus, the points should be as close to the origin as possible, and so the location of the origin should be chosen carefully.

For the TPSK constellation shown in Figure 3.2(a), each symbol represents one trit, or $\log_2 3 \approx 1.585$ bits of information. It is straightforward to arrange the constellation points according to an equilateral triangle, and set the origin to the center of the triangle. The minimum Euclidean distance from each point to its decision boundary is $r = \sqrt{3E_s/4}$, where E_s equals the average symbol energy.

For H6-QAM, each symbol carries one trit plus one bit of information, and the constellation arrangement is shown in Figure 3.2(b). The minimum Euclidean distance from a constellation point to its decision boundary is $r = \sqrt{8E_s/15}$, which is much larger than that of 6-PSK ($\sqrt{E_s/4}$), and thus the symbol error performance of H6-QAM is better.

For H8-QAM, each symbol carries three bits of information, and the constellation

arrangement is shown in Figure 3.2(c), where the origin is located at the midpoint of an edge of a hexagon. The minimum Euclidean distance from a constellation point to its decision boundary is $r = \sqrt{2E_s/9}$, which is larger than that of rectangular 8-QAM ($\sqrt{E_s/6}$). For H12-QAM, each symbol carries one trit plus two bits of information, and the proposed constellation arrangement is shown in Figure 3.3(a), where the origin is the joint vertex of the center hexagons. The minimum Euclidean distance from a constellation point to its decision boundary is $r = \sqrt{3E_s/19}$. Note that TPSK, H6-QAM and H12-QAM are rotationally symmetric by 120°, which is an additional benefit as will be shown in Section 3.5.

3.3.2 Constellation Mapping

Given the geometry of the constellation points, the next step is to map the bits and trits to the constellation point such that the BER is minimized. As there are more neighboring points (with the smallest distance to a constellation point) using hexagonal modulation, careful mapping is required to limit the number of bits and/or trits in error due to a symbol error. It is not straightforward to obtain a Gray type of mapping (with only one bit/trit difference between neighboring points), because the number of neighbors with hexagonal tiling often exceeds the number of bits and trits represented by each symbol.

An exhaustive search for the optimal mapping is only possible when the number of constellation points is small as the computational cost increases exponentially with the number of points. Here, we use the following design principle to search for good mappings. Starting from a bit, the constellation points are divided into two clusters. Similarly, starting from a trit, the constellation points are divided into three clusters. Then, '0' and '1' (for a bit) or '0', '1' and '2' (for a trit) are arbitrarily assigned to each of the clusters. For the remaining bits or trits, binary or ternary numbers are assigned to the points in one cluster first, and in turn to the points in the other clusters. The same number is assigned to neighboring points in different clusters as much as possible. For example, with the H12-QAM constellation in Figure 3.3(b), starting with a trit, the 12 points are divided into three clusters, and the first, second and third one are assigned '0', '1' and '2', respectively. Within the first cluster, a Gray type mapping is used to assign '00', '01', '11' and '10' to the four points. For the second cluster, '00' and '01' are assigned to the points neighboring '000' and '001'. Similarly, for the third cluster, '01' and '11' are assigned to the points neighboring

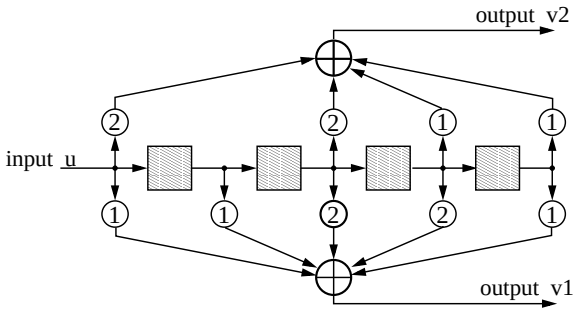


Figure 3.4: A ternary convolutional encoder [95].

'101' and '111'.

Although the above mapping is not of a Gray-type, the results presented in Section 3.4 show that it still leads to a performance improvement.

3.3.3 Ternary Error Control Coding and Interleaving

Given a noisy channel, a received constellation point may differ from that transmitted which will result in bit/trit errors. Interleaving and error control coding can be used to mitigate these errors.

Non-binary Convolutional Codes

Convolutional coding has been widely used in wireless systems such as 802.11 because of the relatively simple implementation and good performance improvements [27]. A binary convolutional encoder can be represented by the parameters (n, k, m) where k and n are the numbers of input and output bits, respectively, and m is the encoder memory size. The code rate is k/n. Three code rates, 1/2, 3/4 and 2/3, are employed in the IEEE 802.11 standard. At the receiver, the received coded bit stream is decoded to recover the original message bit stream. The Viterbi algorithm follows the maximum-likelihood decoding approach and is widely used in practice because of the low implementation complexity.

Similarly, a ternary (n, k, m) convolutional encoder maps k input trits to n output trits [95]. A ternary convolutional encoder can be simply implemented using shift registers and modulo-3 adders, as shown in Figure 3.4. In the figure, the shaded squares are memory elements and the circled numbers are the coefficients that the trits are multiplied by. The circled sum represents a modulo-3 adder. As there are two output trit streams v1 and v2, the code rate is 1/2. Similar to binary convolutional codes,

puncturing can be employed to obtain different code rates to adjust the protection level according to the modulation scheme, channel condition and required BER. To be consistent with the 802.11 standard, a rate 3/4 punctured ternary convolutional encoder is considered. The puncturing pattern is the same as that used for the binary convolutional encoder, [1 1 1 0 0 1], where 0 means the coded digit at that position is punctured.

Following the similar principle, we implemented the ternary based Viterbi decoding [86], where the difference is the operation is on GF(3) rather than GF(2). The complexity associated with a convolutional code is primarily in the decoder. A ternary decoder using the Viterbi algorithm requires $O(3^m L_T)$ memory space and $O(3^{2(m+1)} L_C)$ computation time, where L_T is the trace back length of the decoder and L_C is the code block length, respectively. The computational complexity of a ternary convolutional is comparable to that of a binary decoder with a similar number of states.

Interleaving

Many error correcting codes such as convolutional codes cannot tolerate burst errors, so it is desirable to separate these errors using an interleaver. An interleaver can also help mitigate errors when Gray mapping is not employed. Bits or trits are interleaved within a packet as the performance results show that this intra-packet interleaving provides an obvious performance gain for fading channel and it does not introduce extensive delay.

3.3.4 System Architecture

Using the IEEE 802.11a standard as an example, Figure 3.5 shows the system architecture which utilizes H-QAM. The shaded blocks indicate the new modules or the existing modules that need to be updated. The messages are divided into two queues, one of which is converted to the trit stream. The bits/trits conversion module converts the bit sequence to a trit sequence (details of this conversion will be discussed later), and this is the input to the ternary-based convolutional coding (CC) encoder. The cross-layer controller uses the CSI to determine the modulation and coding schemes to be used, which is included in the Physical Layer Convergence Procedure (PLCP) header [27].

The PLCP header includes the physical layer control information. The *rate* field of

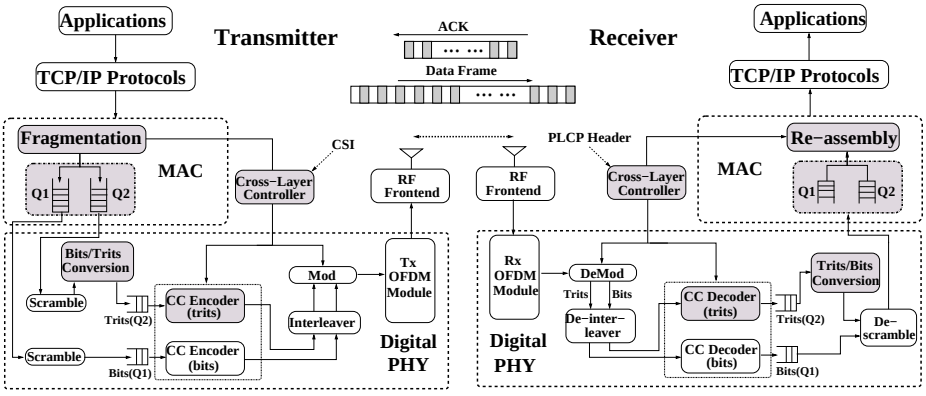


Figure 3.5: The new system architecture based on IEEE 802.11.

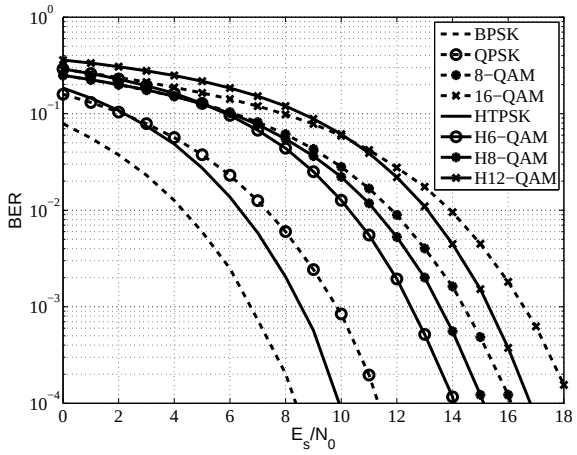


Figure 3.6: BER without error control coding.

the PLCP header is used to inform the receiver of the modulation and coding scheme to be employed. There are 8 different AMC schemes in the current IEEE 802.11a standard, which is represented by a 4-bit *rate* field. There is one reserved bit in the header which can be combined with the 4-bit rate field to represent up to 32 different modulation and coding schemes. This should be sufficient to include the new AMC schemes based on H-QAM and ternary convolutional coding proposed in this chapter, as some of these schemes can replace existing AMC schemes based on conventional QAM. The proposed non-binary H-QAM communication system is compatible with conventional systems and does not require additional communication overhead as will be shown later.

3.4 Performance Evaluation

3.4.1 Uncoded BER Performance

A key performance index is the BER given the received SNR over an AWGN channel. We consider a received SNR from 0 to 18 dB and examine the uncoded BER for different modulation schemes. The received signal to noise ratio is E_s/N_0 where E_s is the received energy per symbol and N_0 is the noise power spectral density. Monte Carlo simulation is used to obtain the BER results shown in Figure 3.6.

Although H-QAM is the most efficient modulation in terms of achieving the minimum symbol error rate, each constellation point has more neighbors than that of rectangular QAM, so the resultant non-Gray code mapping will degrade the BER

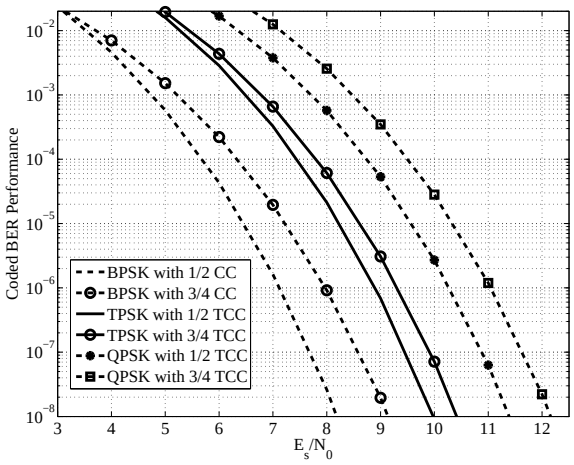


Figure 3.7: Coded BER with BPSK, TPSK and QPSK modulation.

performance of H-QAM. Comparing H8-QAM and 8-QAM, we note that even with this loss due to non-Gray mapping, H8-QAM still outperforms rectangular 8-QAM by 0.8 dB at a BER of 10^{-3}, and this gap increases for smaller BERs. These results demonstrate the advantage of using hexagonal constellations.

Considering uncoded systems and using 10^{-3} as the uncoded BER threshold, the addition of TPSK, H6-QAM, H8-QAM and H12-QAM provides more choices to adapt the modulation. For instance, TPSK can be used to replace BPSK in the SNR range of $[6.2,\ 8]$ dB to achieve a 58% throughput gain. Similarly, H6-QAM, H8-QAM, and H12-QAM can replace QPSK in the SNR range of $[9.8,\ 15.3]$ dB to achieve 29% to 79% throughput gains. H-QAM constellations such as H27-QAM and H54-QAM should provide throughput gains when the received SNR is higher, which is left for future research.

3.4.2 Coded BER Performance

For the coded BER performance, an AWGN channel is considered with the rate 1/2 ternary convolutional code shown in Figure 3.4. This code has $3^4 = 81$ states, so for a fair comparison a rate 1/2 binary convolutional code is used with $2^6 = 64$ states. This binary code is employed in IEEE 802.11 standards [27]. To evaluate the performance of the H-QAM based AMC, we also consider punctured convolutional coding with the puncture pattern discussed in Sec. 3.3.3, which provides a code rate of 3/4.

The BERs for BPSK, TPSK and QPSK with different code rates are shown in Figure 3.7, and the BERs for QPSK, H6-QAM, H8-QAM, H12-QAM and 16-QAM

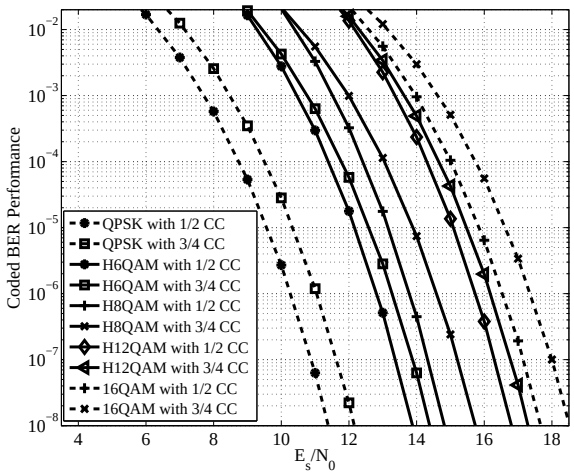


Figure 3.8: Coded BER with QPSK, H6-QM, H8-QAM, H12-QAM and 16-QAM modulation.

Table 3.1: A comparison of modulation and coding schemes.

Modulation	**Coding rate**	**Throughput (b/sym)**	**SNR (dB)**
BPSK*	1/2	0.5	> 7.12
BPSK*	3/4	0.75	> 7.97
TPSK	1/2	0.785	> 8.89
TPSK*	3/4	1.178	> 9.3
QPSK	1/2	1	> 10.2
QPSK*	3/4	1.5	> 11.04
H6-QAM	1/2	1.285	> 12.81
H6-QAM*	3/4	1.928	> 13.27
H8-QAM	1/2	1.5	> 13.78
H8-QAM*	3/4	2.25	> 14.58
H12-QAM	1/2	1.785	> 15.73
H12-QAM*	3/4	2.678	> 16.18
16-QAM	1/2	2	> 16.53
16-QAM*	3/4	3	> 17.35

with different code rates in Figure 3.8. Comparing Figures 3.6 and 3.7, when the SNR is below 5 dB, the BER for BPSK with or without coding is similar (10^{-2} or above). In addition, when the SNR is below 6 dB, the BER for TPSK with or without coding is similar (10^{-2} or above). Thus, error correction coding is effective only when the uncoded BER is sufficiently low.

Figures 3.7 and 3.8 show that for a given code rate and BER, the required SNR increases with respect to the number of constellation points. This is because the denser the constellation, the smaller the minimum Euclidean distance. However, some combinations of H-QAM and coding outperform the QAM combinations in terms of both throughput (information per symbol) and BER. For instance, TPSK with a rate 3/4 code has a throughput of 1.178 b/sym (bits per symbol), which is higher than that of QPSK with a rate 1/2 code, 1 b/sym. In addition, the BER performance of TPSK with a rate 3/4 code is better than that of QPSK with a rate 1/2 code, with more than 0.9 dB improvement at a BER of 10^{-6}. Therefore, TPSK with rate 3/4 coding can replace QPSK with rate 1/2 coding in AMC. Similarly, H12-QAM with rate 3/4 coding can replace 16-QAM with rate 1/2 coding. Using 10^{-6} as the threshold for coded BER, the required SNR and the corresponding throughput are shown in Table 3.1.

Considering AMC, we can not only use TPSK and H12-QAM with rate 3/4 coding to replace QPSK and 16-QAM with rate 1/2 coding, respectively, but also use H6-

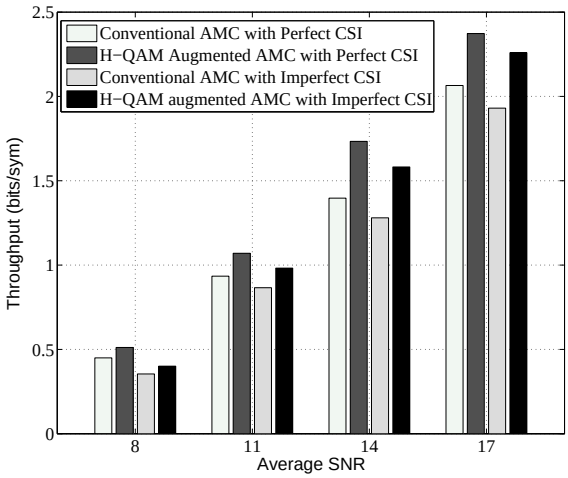


Figure 3.9: Single link throughput comparison.

QAM and H8-QAM with rate 3/4 coding to fill the SNR gap between QPSK with rate 3/4 coding and 16-QAM with rate 1/2 coding. This provides us a finer grained set of choices for AMC. These new AMC schemes can be indicated in the PLCP header for the receiver to demodulate and decode properly. Next, we study how these new modulation and coding schemes contribute to a higher system throughput and efficiency.

3.4.3 Single Link Throughput

The enriched modulation and coding combinations are now considered in an AMC system. A Rician fading channel with Rician factor $K = 6$ dB is considered for a single communications link. The conventional AMC set contains BPSK, QPSK and 16-QAM with rate 1/2 and 3/4 coding according to the IEEE 802.11 standard [27]. The augmented AMC set includes the existing QAM and the new H-QAM based transmission schemes marked by * in Table 3.1. For a fair comparison, we use the same symbol rate and energy for all the modulation and coding schemes.

In the simulation, we transmit data packets with the size of 1 kilo-byte, and $1,000$ packets are transmitted to obtain the average performance. Intra-packet interleaving is used for the bits and trits separately. The simulated transmitter and receiver follow the same structure as given in Figure 3.5.

For AMC with or without H-QAM, one practical issue is the imperfect estimation of channel conditions. If the received SNR is underestimated, the sender may select a modulation and coding scheme with a lower throughput (i.e., the number of bits per

symbol received by the receiver). Conversely, if the received SNR is overestimated, a higher BER than the required threshold may result, which is even more undesirable. The impact of channel estimation errors on the system performance is thus of critical importance. To examine this impact, channel estimation errors are modeled as a Gaussian random process with zero mean and unit variance. To reduce the probability that the received SNR is overestimated, which may severely degrade the system performance, the transmitter uses the estimated SNR minus its standard deviation to select the modulation and coding type.

Figure 3.9 compares the link throughput using AMC with and without the proposed H-QAM schemes, for an average received SNR of 8, 11, 14 and 17 dB. The average throughput was obtained for $1,000$ runs using Monte Carlo simulation to average the effects of fading. With perfect channel estimation, the proposed non-binary communication system outperforms the conventional one by 13.7%, 14.5%, 24.1% and 14.9% when the average SNR is 8, 11, 14 and 17 dB, respectively. The performance of both conventional QAM and H-QAM is degraded with imperfect channel estimation, but the proposed system still achieves throughput gains of 13.1%, 13.5%, 23.6% and 17% for an average SNR of 8, 11, 14 and 17 dB, respectively.

3.4.4 Network Throughput

The system performance is then evaluated in an infrastructure-based network, where an access point (AP) is centrally located to serve all users in the network, e.g., a WiFi network, and the mobile users are randomly distributed. The wireless channels suffer from independent Rician block fading. The path-loss exponent α is set to 3. All symbols have the same average transmitted symbol energy. We consider the downlink performance where the AP transmits packets (with the same size of 1 kilo-byte) to all mobile users in a round-robin manner. It is assumed that the AP has the channel information which is used to select the AMC scheme for each packet according to the estimated SNR. The average SNR is set to 7 dB when users are at the boundary of the network.

The system performance was evaluated using Monte Carlo simulation with different node densities. For each density, 1,000 simulation runs were executed using random topologies, and the average network throughput was determined in terms of bits per symbol. Figure 3.10 presents the results for 5 and 30 users, which correspond to sparse and dense networks, respectively. These results show that the proposed

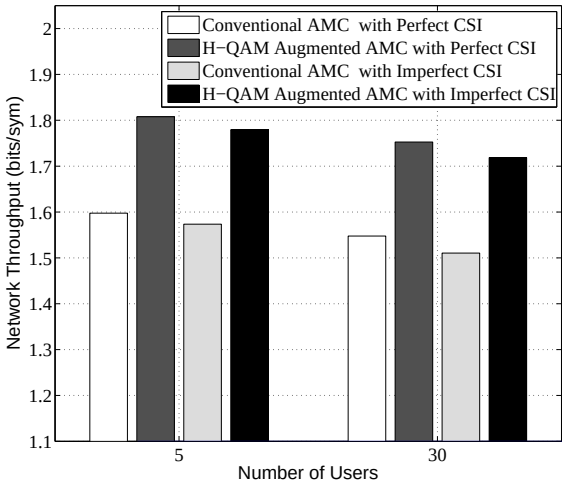


Figure 3.10: Network Throughput.

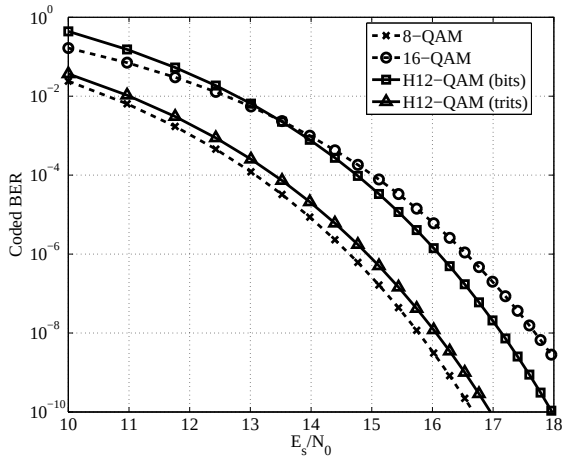
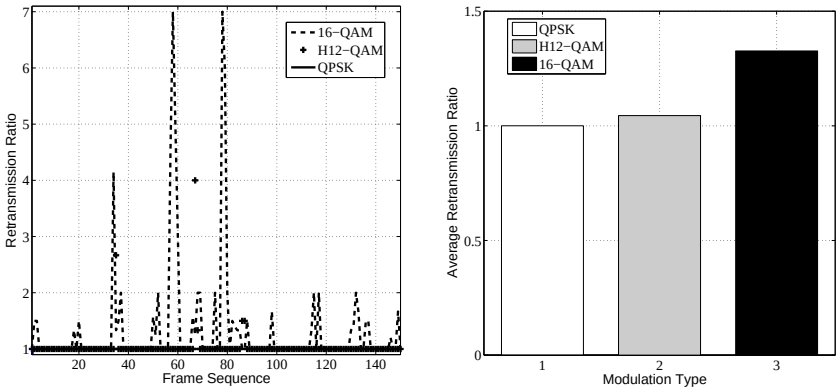


Figure 3.11: H12-QAM Trit and Bit BER Performance.

non-binary H-QAM schemes increase the average network throughput by more than 13.2% and 13.3% for the 5 and 30 user cases, with perfect or imperfect channel information. In addition to the throughput gain, for a given average transmitted symbol energy, with more bits transmitted per symbol, the per-bit energy consumption is reduced by 11.6% and 11.7%, respectively.

BER Diversity and Link-Layer Performance

We use H12-QAM which contains 1 trit and 2 bits of information per symbol as an example to transmit the layered video. For comparison, we also evaluate the performance using the traditional 16-QAM and QPSK. The coded trits/bits BER



(a) One simulation run (b) Average over 100 runs

Figure 3.12: Link-Layer Transmission Times.

performance of H12-QAM is shown in Figure 3.11. From the figure, the BER performance of trits and that of bits in H12-QAM have more than one order of difference. The BER of trits is close to that of 8-QAM while the BER of bits is slightly better than that of 16-QAM. It is anticipated to have a higher video quality if we use trits to carry the base layer information and bits to carry the enhancement layer information, as we have a better error resilience performance for the base layer due to the BER diversity.

We still consider a Rician fading channel, where the average SNR is 75 (18.75 dB) and the Rician factor is 8 dB. The wireless communication is error prone even with a strong error control coding. Therefore, the link-layer retransmission schemes are adopted to further improve the system robustness and link stability. Given the importance of the base layer, we design the retransmission scheme accordingly: if the base-layer in a packet contains errors, a retransmission will be triggered until the base-layer is received correctly or reaching the maximum retransmission limit of 7 according to the IEEE 802.11 standard [27]. For simplicity, the video data are fragmented to packets with the maximal packet size of $1,500$ bytes.

First, we use the average number of transmissions per packet to evaluate the link-layer performance. Figure 3.12(a) shows the transmission times of each video frame during one run of the simulation. There are several pulses in the curves of 16-QAM and H12-QAM, which are due to the channel variation and varying video frame size. In the figure, all packets can be transmitted successfully with QPSK and H12-QAM, but there are a few packet losses with 16-QAM after reaching the retry limit. The average transmission times per packet are shown in Figure 3.12(b), which

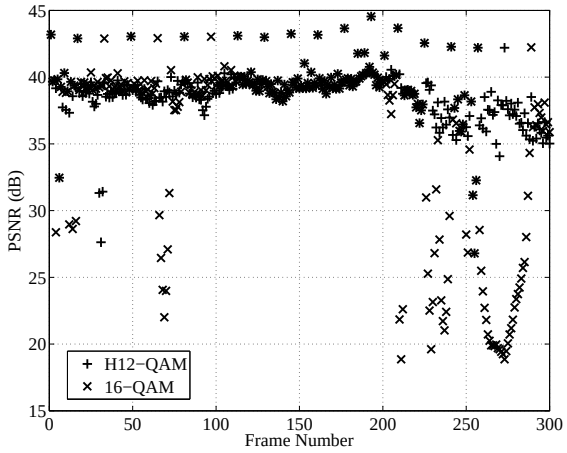


Figure 3.13: Video Quality with H12-QAM vs 16-QAM.

are 1.32, 1.04 and 1.00 for 16-QAM, H12-QAM and QPSK, respectively. There is almost no retransmission for QPSK due to the low BER, and H12-QAM has slightly larger retransmission times but its spectrum efficiency is more than 70% higher than QPSK. On the other hand, the average retransmission times with 16-QAM is very high, which not only leads to more channel time consumed but also results in a larger delay and jitter. Considering the larger number of information carried per symbol and the larger number of link layer retransmissions, the spectrum efficiency of 16-QAM is similar to that of H12-QAM. Next, we compare their video service quality.

Video Quality Comparison

Figure 3.13 compares the PSNR of the video using the 16-QAM and the H12-QAM. With 16-QAM, the video performance degrades to unacceptable regions frequently due to the losses of the base layer information even with up to 7 link-layer retransmissions.

We also conducted the simulation using QPSK. The average PSNR of the video using 16-QAM, H12-QAM and QPSK are 33.31, 38.4 and 39.16 dB, respectively. Compared with QPSK, we can see that the average video quality using H12-QAM is similar, but H12-QAM can achieve both the bandwidth and energy saving by 70%.

In [83], Apex was proposed based on the conventional QAMs, where the unequal error protection of constellation mapping codes is used to carry the I, P, and B frames in a non-layered video, as the I frames are more important than P and B frames. The trits and bits in H-QAM have an even larger performance difference than that of the

conventional QAMs given in [83]. Thus, the non-layered video can also benefit from the performance difference by using hybrid trits/ bits for I, P, and B frames.

3.5 Prototype and Measurements

A prototype for the non-binary H-QAM communication system was developed using the SDR platform USRP2 [23] and GNU Radio. SDR offers the the flexibility to develop the radio systems by moving the general signal processing modules to software end. GNU Radio [9] is is one of the most popular SDR software development tool stack.

The USRP2 platform can support different carrier frequencies according to the daughter board mounted, and we used the XCV2450 daughter board which can be operated at the license-free bands of 2.4 to 2.5 GHz and 4.9 to 5.85 GHz. One USRP2 was connected to the laptop host (DELL E5400) as the transmitter and another to the PC host (DELL OPTIPLEX 755) as the receiver. The USRP2-based OFDM implementation [76] was augmented with the proposed H-QAM. The carrier frequency, number of subcarriers and FFT length are 2.49 GHz, 80 and 512, respectively.

Three new hexagonal modulation designs, TPSK, H6-QAM and H8-QAM, have been implemented. As mentioned previously, bits/trits conversion is required to implement H-QAM. This conversion will introduce some performance loss as 2 and 3 are both prime numbers [88]. Using a long bit sequence will reduce this loss but increase the delay and complexity. The conversion efficiency is defined as

$$\eta = \frac{l_b}{l_t \cdot \log_2(3)}, \tag{3.3}$$

where l_b is the length of the input bit sequence and l_t is the length of the output trit sequence. Converting 11 bits to 7 trits provides an efficiency of $\frac{11\log_2 2}{7\log_2 3} = 99.1\%$ [48]. As the conversion block is small, this can be easily implemented using a lookup table. After the bits/trits conversion and scrambling at the transmitter, the mapping to modulation symbols is done.

$1,000$ test frames were transmitted where each frame contains 100 blocks of data, and each block contains 16 bytes of data with a 4-byte CRC. The transmitter converts all or part of the bits in each block to trits depending on the modulation employed. The receiver demodulates the received symbols to bits and/or trits according to the modulation scheme used as indicated in the PLCP header. Then the trits are con-

Table 3.2: The experimental BER results.

Modulation	**Uncoded BER**	**Coded BER**
BPSK	5.4×10^{-5}	$< 10^{-7}$
TPSK	5.1×10^{-5}	$< 10^{-7}$
QPSK	2.9×10^{-4}	2.13×10^{-7}
H6-QAM	1.2×10^{-3}	5.49×10^{-6}
H8-QAM	1.6×10^{-3}	1.04×10^{-5}
8PSK	2.3×10^{-3}	1.22×10^{-5}

verted back to bits and the CRC is checked to determine whether there are any transmission errors in the block. If a block fails the CRC check, the number of errors is obtained by comparing it with the original block. The BER after demodulation can then be calculated using the total number of errors. To obtain the coded BER, the received bits or trits are further processed using the binary and ternary convolutional codes to determine the coded BER.

The BER results are shown in Table 3.2. As expected, the performance of TPSK is better than that of QPSK, and H6-QAM and H8-QAM perform better than 8PSK. These results demonstrate the feasibility and simplicity of deploying the proposed H-QAM. Note that the uncoded BER performance of TPSK is even slightly better than that of BPSK. This is because, given the constellation of BPSK, there might be a 180° phase shift between two consecutive symbols transmitted, and these phase-shifts result in large amplitude fluctuations, which degrade the performance of the communication systems [86]. The relative phase difference of TPSK symbols, 120°, is much smaller than that of BPSK, which brings an additional benefit in the practical communication system that was usually not counted in the analysis. Despite the accuracy limitations of the USRP2 hardware and the effects of the fading wireless channel, the measured results given for an indoor environment confirm the feasibility of the approach.

3.6 Conclusions

In this chapter, the design and implementation of a non-binary communication system have been presented which employs H-QAM and ternary error control coding. Both bits and trits are employed to improve the spectral and power efficiency. The proposed H-QAM can be implemented by modifying existing QAM systems. We have also explored the inherent performance diversity of bits and trits in the same symbol to offer a differentiated information protection, which can be used to support video applications for a higher efficiency and/or better QoS.

3.7 Symbol List

Symbol	Explanation
E	maximal received symbol energy
d_i	distance of the i-th constellation point to the origin
M	the number of constellation points
E_s	average symbol energy
l_b	length of bit sequence
l_t	length of trit sequence

Chapter 4

Environment-aware Clock Skew Estimation and Clock Synchronization

4.1 Introduction

Having a synchronized clock, i.e., all nodes having a common notion of *time*, is one of the most fundamental requirements in wireless sensor networks. Clock synchronization is critical for data transmission, localization, sleep-scheduling, information fusion, and etc. However, due to the imperfect crystal cutting and clock circuit, and the dynamic working temperature, the output value of a node's local clock, namely the *local time*, is often different from others' local time, leading to the clock offset. An unregulated clock offset will degrade the network performance and even endanger the proper network functioning.

Clock synchronization, the process of removing clock offset w.r.t. the reference clock, still remains a challenging issue especially in the contexts of sensor networks. First, sensor nodes usually have limited computational capacity and power supply and thus the synchronization algorithms for sensor networks should be with moderate computational complexity and communication overhead. Besides the limited resources, in many sensor network applications, such as environment monitoring, wild animal tracking, surveillance systems, etc., the environments surrounding sensor

nodes are usually highly dynamic and/or even hostile. Thus it is clear the environment information could be utilized to improve the synchronization of sensor nodes.

In the modern electronic system, a crystal oscillator is usually used to generate the clock ticks, which are periodical signals with certain frequency. However, the output frequency of the clock tick is not always the same as the nominal frequency and changes according to the working temperature. The deviation of the output frequency results in different clock tick durations and thus the clock skew, which is the instantaneous clock drift rate between two or more clocks. It is the inherent reason for clock de-synchronization [74]. The output value of a clock is the counting number of its ticks and therefore the effect of clock skew will be accumulated and lead to an unbounded clock offset. Thus the clock needs to be re-synchronized to maintain the clock synchronization accuracy. It is obvious that if we can estimate and compensate the clock skew, we can improve the clock synchronization accuracy and therefore prolong the clock re-synchronization period [68].

However, the estimation on the clock skew is non-trivial because the output frequency of a crystal is affected by the working temperature, which is dynamic especially for contexts such as outdoor applications [115]. Previously, a constant clock skew was assumed for skew estimation, which is not accurate according to the measurement results. Kalman filters have been used to improve online skew estimation performance, which can achieve the minimum mean square error (MSE) for linear systems with white Gaussian noise [7, 5, 30, 110]. However, in WSNs, clock skew is non-stationary due to environment variations. Different from the previous approaches, in this chapter, we use the correlation of temperature and clock skew in a multi-model Kalman filter to accurately estimate the clock skew in a dynamic environment. Here, the dynamics of the environment mainly refers to the changing environment temperature.

Based on the clock skew estimation results, we propose a two-phase synchronization scheme. The first one is the *temperature-assisted external clock synchronization* phase, during which the local node exchanges timestamps to eliminate the clock offset. Different from the previous solutions, we propose to leverage the relationship between the clock skew and the temperature to estimate the current clock skew, which allows us to directly remove the clock skew and thus improves the clock offset estimation accuracy and reduces the computational complexity. During the second phase there is no timestamp exchanged and the time is updated by the local clock only. We propose the *temperature-assisted clock self-calibration* (TACSC), with which the clock skew is dynamically compensated according to the current working temperature.

The main contributions of this chapter are fourfold. First, based on our measurements, we demonstrate that the clock skew is highly correlated to temperature, so we use a hybrid two-model system to describe the clock skew. For the hybrid system, the main challenge is to determine which model should be used in the Kalman filter for a specific time period, as clock skew measurements contain severe noise. To tackle this problem, we propose an additional information aided multi-model Kalman filter (AMKF) to obtain the model likelihood for clock skew based on the local temperature model likelihood estimation. Second, based on the clock skew estimation results w.r.t temperature, we propose to directly remove the clock skew during a clock synchronization process by exploring the relationship between the crystal frequency and the working temperature. It can not only improve the estimation accuracy but also reduce the computational complexity. Third, during the period when the local clock cannot obtain the timestamps from the reference clock, we propose the TACSC scheme to dynamically compensate the clock skew. Thus we can maintain the synchronization of the local clock and therefore significantly prolong the period between two synchronization processes under certain synchronization error tolerance. The proposed TACSC scheme relies on local information only and does not require any timestamp exchange. Therefore, it can not only reduce the communication overhead in clock synchronization but also be used in the case when the reference clock is temporarily unreachable. The last but not the least, extensive simulation and testbed verification are conducted to demonstrate the efficacy of the proposed scheme. For the clock skew estimation, the proposed scheme can reduce the RMSE by more than 50% and we also obtain the estimation lower bound as benchmark. For the clock synchronization phase, the proposed scheme can reduce the mean square error (MSE) of clock offset estimation by more than 50%. For the duration when there is no timestamp message exchanged, the proposed TACSC scheme can reduce the synchronization error by more than one order of magnitude.

4.2 Related Work

In the previous research work, many clock synchronization approaches have been proposed for different scenarios. For example, the Network Time Protocol (NTP) [70] has been widely used in the Internet, which allows computers to extract the timestamp information from NTP packets to update their local clocks. However, it is not well suited for sensor networks due to the complexity. The Timing-sync Protocol for

Sensor Networks (TPSN) was proposed for sensor networks instead in [26] to correct the clock offset. However, this scheme requires frequent re-synchronizations because it does not compensate clock skew, which makes the clock drift away quickly. The Flooding Time Synchronization Protocol (FTSP) [68] has been proposed for multi-hop wireless networks. It calculates the clock skew in a moving window and uses linear-regression algorithms to mitigate both clock offset and clock skew in a hierarchical way. FTSP takes the clock skew into account. Therefore, it can achieve a relatively high synchronization accuracy. In [57], an interesting scheme called FLIGHT was proposed, which can explore the light intensity changes with a stable period that equals half of the alternating current's to perform clock synchronization in the indoor environment.

With the estimation of clock skew, it can be removed and/or compensated to improve the synchronization accuracy and/or prolong the re-synchronization period. A linear optimization problem was formulated and solved to estimate clock skew in [71]. A convex hull algorithm for both offline and online clock skew estimation was proposed in [113], and it was also used for clock skew mitigation. In [53], the joint estimation of clock offset and clock skew with unknown synchronization delay was addressed. In [45], a direct clock skew removal technique was proposed. However, the relationship between clock skew and working environment has not been considered in their work, and the skew estimation using a moving window might not be able to track the change of clock skew accurately. In [115], the authors proposed an interesting on-demand synchronization scheme that can give an uncertainty analysis of the clock skew estimation and compensation. According to this uncertainty analysis, the synchronization error can be statistically estimated and therefore the resynchronization can be triggered wisely. However, since this scheme ignored the impact of temperature and thus the analytical model can only work well in a stable environment. As demonstrated in [105, 87], clock skew changes dramatically when the working temperature changes.

In general, crystal oscillators are sensitive to the temperature and the changing temperature will deviate the output frequency of crystal oscillator. Although the temperature compensated crystal oscillator (TCXO) can compensate the temperature effect, e.g., MAXIM DS32KHz, but the cost is still too high ($2.81 each at the volume of 1K [69]) for the low cost sensor nodes. Besides, the TCXO also increases the circuit complexity, energy consumption and footprint of wireless devices. In addition, the TCXO cannot totally eliminate the clock skew instability w.r.t. temperature

(e.g., up to ±7.5ppm [69]). Thus, we are motivated to devise low-cost, flexible software solutions.

In [94], the authors validated that the clock skew of sensor node was unique and sensitive to temperature and they utilized this information to detect malicious and mal-functioning nodes. In [82], a temperature-compensated time synchronization algorithm was proposed to compensate the clock according to the environment temperature and can achieve a much longer resynchronization period. The limitation of that work is that the calibration algorithm treated the clock skew as a stationary random process, which may not be true in dynamic environments and will thus degrade the estimation accuracy.

On the other hand, it is well known that Kalman filter is optimal for linear systems with white Gaussian noise in terms of MSE. Several approaches have been proposed based on Kalman filter [7, 5, 30, 110]. The algorithms proposed in [7, 5] assumed that the clock skew is constant, which works well in a relatively stable environment, while the performance will degrade in dynamic environments. For example, the temperature may vary more than 20℃ within 12 hours in an outdoor environment. Temperature changes may be more dramatic for some applications such as wild fire detection. The circuit temperature changes can also be faster due to the heat generated by the circuit components.

In [30], an $AR(1)$ model was developed to describe the changing clock skew, which provides some adaptivity w.r.t the changing clock skew.

In [110], an interacting multi-model (IMM) Kalman filter was employed in clock skew estimation to consider the changing clock skew. However, due to the high noise level of clock skew measurement, the model uncertainty problem still exists. In addition, how to utilize the high correlation between temperature and clock skew to better estimate clock skew has not been investigated in [30, 110].

In [82], a temperature-compensated time synchronization algorithm was proposed to compensate the clock skew according to the environment temperature. Since it dynamically compensate the clock skew, it can achieve a longer re-synchronization period. The limitation of that work is that the adopted clock skew estimation algorithm treats the clock skew as a stationary random process, which may not be true in dynamic environments.

Different from the previous approaches, in this work, we first identify the correlation between temperature and clock skew from the measurement results, and propose to accurately estimate clock skew with the assistance of temperature measurements

using a low-cost Kalman filter. According to the estimation of clock skew w.r.t. temperature, we can directly remove it in the external clock synchronization phase, which results in both higher synchronization accuracy and lower computational complexity. In the clock self-calibration phase, we use the correlation between temperature and clock skew to compensate the instantaneous clock skew, so it can maintain the clock synchronization level for a much longer duration without exchanging timestamp messages. The proposed solution can be easily extended to any other wireless devices equipped with temperature sensors.

4.3 Notations and System Model

4.3.1 Notations

For presentation clarity, we first define the important terms used in this chapter, which are consistent with our published work [105].

Clock is the device to measure time. It consists of a periodic component that ticks at a given frequency and a counter that counts the number of ticks from a starting time instant. Denote by $C_A(t)$ the output value of clock A at a given reference time instant t.

Clock frequency is the frequency of clock signal generated by the clock circuit, which is usually determined by the crystal and the corresponding peripheral circuit components. Let $f(T)$ denote the actual output frequency of clock circuit at temperature T. It may not be the same as the nominal frequency f_n with perfect crystal and peripheral circuit at the targeted temperature. The difference of $f(T)$ and f_n results in the clock skew, which will be discussed later.

Clock offset is the difference between the reported time of two or more clocks. If the time reported by clock A and clock B are $C_A(t)$ and $C_B(t)$, respectively, the offset of clock A w.r.t. clock B can be written as

$$\theta_B^A(t) = C_A(t) - C_B(t). \tag{4.1}$$

Clock skew is the differential coefficient of the clock offset. The clock skew reflects the difference of tick durations of two or more clocks, and is denoted by

$$\alpha_B^A(t) = \frac{d\theta_B^A(t)}{dt} \approx \frac{\theta_B^A(t+\tau(t)) - \theta_B^A(t)}{\tau(t)}, \tag{4.2}$$

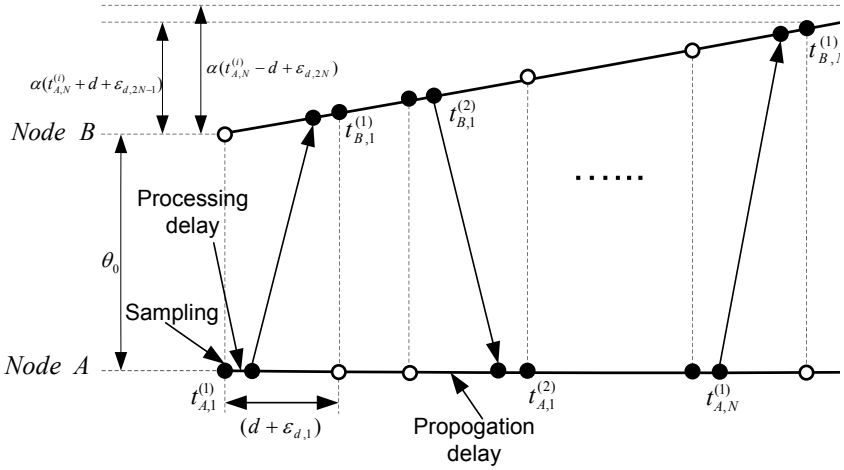


Figure 4.1: Two-way timestamp exchange

where $\tau(t)$ is the sampling interval.

Using a uniform sampling where $\tau(t) = \tau$, the clock offset and clock skew can be discretized, respectively, as

$$\theta_B^A[n] = C_A[n] - C_B[n], \tag{4.3}$$

and

$$\alpha_B^A[n] = \frac{\theta_B^A[n+1] - \theta_B^A[n]}{\tau}. \tag{4.4}$$

4.3.2 Two-way Timestamp Echange

The two-way timestamp exchange mechanism between the local node and the reference node is usually adopted to estimate the clock offset [53, 101], as shown in Fig 4.1. Several timestamps might be exchanged to eliminate the measurement noise caused by delay and clock jitters, etc., to obtain a higher clock offset estimation accuracy.

In Fig. 4.1, θ_0 is the initial clock offset to be estimated. The absolute value of clock offset might change as time elapses due to the clock skew, i.e.,

$$\theta(\Delta t) = \theta_0 + \alpha(T)\Delta_t \tag{4.5}$$

where the clock skew $\alpha(T)$ can be assumed as a constant during the timestamp exchanges, since the timestamp message exchange can be finished within a very short time. In the ith round of timestamp exchange, node A first sends the timestamp $t_{A,1}^{(i)}$ at $t_{A,1}^{(i)}+d_{p,A}$, where $d_{p,A}$ is the processing and queuing delays at node A. The message arrives at node B at $t_{B,1}^{(i)}-d_{p,B}$ and node B records the time $t_{B,1}^{(i)}$, where $d_{p,B}$ is the processing and queuing delays at node B. The reply message contains timestamp $t_{B,2}^{(i)}$ and is sent at $t_{B,2}^{(i)}+d_{p,B}$ along with the previous recorded timestamp $t_{B,1}^{i}$. Then, node A records the reception time of the reply message as $t_{A,2}^{i}$.

Denote $D_1[i]$ as the difference between the timestamps $t_{A,1}^{i}$ and $t_{B,1}^{i}$, i.e., $D_1[i]=t_{B,1}^{i}-t_{A,1}^{i}$. However, we cannot simply use $D_1[i]$ as the estimation or measurement of clock offset $\theta[i]$, since there are several delays introduced in different stages during the message transmission, which make it deviate from the actual clock offset. Instead, the clock offset can be expressed as

$$D_1[i]=\theta_1[i]+(1+\alpha(T))(t_c[i]+t_t[i]+t_p[i]), \tag{4.6}$$

where $\theta_1(i)=\theta_0+\alpha(T)(t_{A,1}^{i}-t_{A,1}^{0})$ denotes the current clock offset; $t_c[i]$ and $t_p[i]$ are the message construction delay at the sender side and the processing delay at the receiver side, respectively; $t_t[i]$ is the propagation delay. $t_c[i]$ and $t_p[i]$ may be different from system to system due to the CPU capacity and the system load. Their variations are usually at μs level or even smaller, by setting the corresponding processes at a high priority. Hence, we can approximate them as constants and their variations can be included in the system observation noise. $t_t[i]$ can also be approximated as a constant if the channel is stable, and thus its variance can be included in the system observation noise as well. Therefore, we can attribute all these delays as a constant delay d_1 plus a zero-mean random variable and rewrite $D_1[i]$ as

$$D_1[i]=\theta_1[i]+(1+\alpha(T))(d_1+\omega_{d,1}[i]), \tag{4.7}$$

where $\omega_{d,1}[i]$ is the delay jitter during the transmission from node A to node B. Similary, let $D_2[i]$ denote the difference between timestamps $t_{B,2}^{i}$ and $t_{A,2}^{i}$, i.e. $D_2[i]=t_{B,2}^{i}-t_{A,2}^{i}$, and it can be written as

$$D_2[i]=\theta_2[i]-(1+\alpha(T))(d_2+\omega_{d,2}[i]), \tag{4.8}$$

where $\omega_{d,2}[i]$ is the delay jitter during the transmission from node B to node A. Since

the aforementioned process happens within a short period, we assume that the delay is symmetric, i.e. $d_1 = d_2 = d$ as in [74, 53]. Therefore, we can model this two-way timestamp exchange as

$$\begin{cases} D_1[i] = t^i_{B,1} - t^i_{A,1} = \theta_0 + \alpha(T)(t^i_{A,1} - t^1_{A,1}) + (1 + \alpha(T))(d + \omega_{d,1}[i]), \\ D_2[i] = t^i_{B,2} - t^i_{A,2} = \theta_0 + \alpha(T)(t^i_{A,2} - t^1_{A,1}) - (1 + \alpha(T))(d + \omega_{d,2}[i]), \end{cases} \tag{4.9}$$

where $\omega_{d,1}[i]$ and $\omega_{d,2}[i]$ are assumed to be zero mean independent Gaussian distributed random variables with variance σ_d^2, which has been experimentally verified in [21]. If the two-way timestamp exchange is conducted for N rounds, we can obtain a sequence of recorded timestamps as $\{t^1_{A,i}, t^1_{B,i}, t^2_{A,i}, t^2_{B,i}\}_{i=1}^N$. Then, the clock offset θ_0 can be estimated based on these N equation sets similar to (4.9) by different algorithms [53, 74].

4.3.3 Clock Skew Model

To accurately estimate clock skew, we first investigate its property from measurement results. As all oscillators and PLL modules are sensitive to temperature, in addition to the existing measurement results [72, 97, 110, 87], we conduct further measurements over a time span of six months to verify and quantify the correlation of temperature and clock skew.

Measurement Setting

We used one laptop (DELL Latitude-D520) as the reference node and the widely used sensor node, Berkeley Mica2 Mote [89], as the client to send timestamps periodically, where the period duration is set to one second for simplicity. We used two Mica2 nodes and put them together. One Mica2 node was connected to the MIB board [90], which offers the serial interface to send timestamp messages to the laptop. To avoid the buffer jitter, when a serial port event interrupt is triggered, the laptop will read the system time right away and then process the timestamp messages. The second node was connected to a sensor board MTK300 [91], which has a temperature sensor. The reason we used two nodes here is that the sensor board MTK300 is too thick for Mica2 to connect with the MIB board.

To obtain the clock skew measurement in a dynamic environment, we used a heater to increase and a fan to reduce the environment temperature for the Mica2

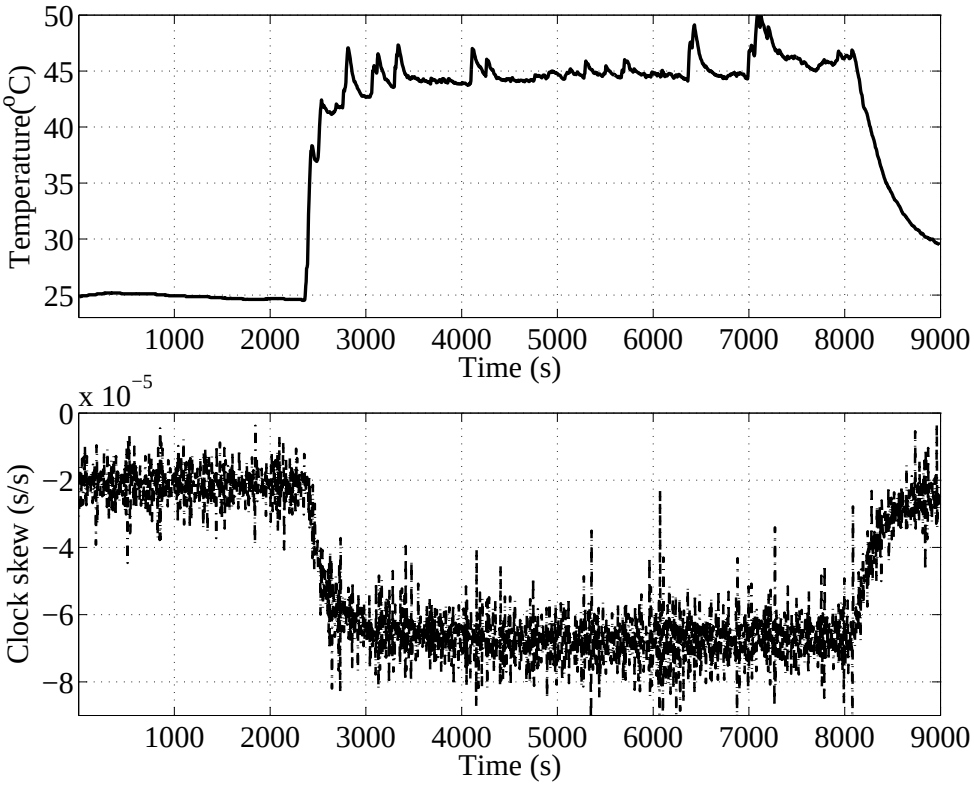


Figure 4.2: Experiments on temperature and clock skew.

sensor nodes. This can mimic the outdoor situation where the sunshine and wind can increase and reduce the temperature. The Mica2 node equipped with the temperature sensor sent the temperature measurement to the other node on demand. The Mica2 node connected to the MIB board periodically transmitted timestamp messages containing the temperature information to the laptop. Hence, we can use these messages to obtain the clock skew measurement.

Measurement Results and Analysis

The measurement results are shown in Fig. 4.2. From the figure, when the environment is stable, the clock skew can be well modeled as a constant with additive noise. On the other hand, the absolute value of clock skew increases (decreases) as the temperature increases (decreases) and hence cannot be considered as a stationary random process. Therefore, it is difficult to use a single model to describe the dynamic behavior of clock skew [110], and we use a hybrid two-model system to describe the

clock skew as

$$\begin{aligned} \alpha[n] &= \alpha + \phi[n] && \text{(a)}, \\ \alpha[n] &= \alpha[n-1] + \tau\rho + \phi[n] && \text{(b)}, \end{aligned} \tag{4.10}$$

where ρ is the changing rate of clock skew and τ is the sampling interval and $\phi[n]$ represents the processing noise, which can be considered as crystal jitter caused by electromagnetic interference and/or some other uncertain factors.

If the temperature is relatively stable, the change of clock skew is marginal and the constant clock skew model in (4.10-a) should be employed. If the temperature changes, we can use the constant velocity model shown in (4.10-b) to characterize the change of clock skew, because the temperature change is relatively smooth during the sample period and the constant velocity model has been widely used in kinematics, thus we can treat ρ as constant.

However, given the severe noise level of clock skew measurements, it is very difficult if not impossible to determine which model to use at a time instant. In addition, from the measurement results, we note that the change of clock skew is not necessarily linear. However, we avoid to introduce other models, such as constant acceleration model, as more models will further exaggerate the model uncertainty problem. In addition, the constant velocity model has the ability to keep up with the constant acceleration process if the sampling rate is sufficiently high, and in this case, the model imperfection can be treated as part of the processing noise.

A more important observation is that, in Fig. 4.2, the trend of the temperature curve matches the clock skew curve very well. We repeated the experiments, and the results all show the similar tendency. To reveal the correlation, we first used the raw measurement data to calculate the Pearson product-moment correlation coefficient ($\rho_{X,Y} = E[(X - u_X)(Y - u_X)]/(\sigma_X\sigma_Y)$) of the clock skew and temperature measurements. The coefficient result is -0.505 only, as the clock skew and temperature measurements both contain noises. We then used a moving window of size 5 to smooth the measurement results. We used the average of the measurement results within the window and calculated their correlation coefficient again, and the result is -0.973, indicating that the temperature and clock skew are indeed highly correlated (negative temperature coefficient). However, we need to use advanced estimation techniques to use this correlation, as the raw measurements, particularly for clock skew, contain severe noise.

Furthermore, from our measurements repeated over a half-year span, for the same pair of clocks, the relationship of clock skew w.r.t. temperature is stable; but different

pairs of clocks exhibit different clock skew change patterns w.r.t. temperature. With this observation, we propose to dynamically compensate the clock skew for each sensor node by first estimating the relationship between the clock skew (w.r.t. the reference clock) and temperature, which can enhance the effectiveness of the clock offset compensation process and prolong the resynchronization period.

Also, the impact of temperature on clock skew is dominant when compared with other environment factors, such as humidity, shock etc. [94, 87], which have not been studied in this chapter, and we treat them as system noise.

4.3.4 Crystal Oscillator

With its low cost, relatively good stability and accuracy, crystal oscillator is widely used in modern electronic systems especially in embedded systems, e.g., the Mica2 sensor nodes [89] are equipped with the 32KHz crystal SE2412CT-ND [34], a tuning fork crystal unit. Crystal oscillator is an electronic oscillator circuit that uses the mechanical resonance of a vibrating crystal of piezoelectric material to create a periodical electrical signal with certain frequency to provide a relatively stable clock signal. The output frequency of the clock circuit depends on the crystal, i.e., the shape or the cut of the crystal, and the peripheral devices, such as the capacitor [87]. However, due to the imperfect cutting techniques and the load capacitance of the peripheral circuits, the actual output frequency $f(T)$ may not be the same as the nominal frequency f_n even at the targeted temperature T [87] (typically between 20℃$\sim$ 25℃ [34]).

The output frequency is also highly related to the working temperature and can be modeled as [87, 35, 110, 34]

$$f(T) = f(T_0)(1 - \beta(T - T_0)^2), \tag{4.11}$$

where β is the parabolic coefficient (or temperature coefficient) and T_0 is usually called the turn-over temperature (might be different from the targeted temperature). A common parabolic coefficient for a regular 32KHz tuning fork crystal is around $\pm 0.04 ppm/℃$ [35, 34]. It indicates that the crystal will resonate close to $f(T_0)$ when the temperature is close to T_0, and it will slow down when either the temperature increases or decreases from the turn-over temperature. As $f(T)$ over T is a quadratic function, it is very sensitive to the operating temperature, e.g., if $T - T_0 = 10℃$, the frequency drift can be up to 4ppm (around 4μs per second).

The clock output is the count of clock ticks, where the duration of a clock tick

is determined by the frequency of the crystal oscillator. Therefore, the difference between $f(T)$ and f_n results in a clock skew of

$$\alpha(T) = f_n/f(T) - 1. \tag{4.12}$$

Given the fact that the temperature sensor becomes the standard element of many sensor board [91], we can substitute the measured temperature (may contain noise) into (4.11) to estimate the current clock frequency and to obtain the corresponding clock skew by (4.12). Since the measurement of temperature is noisy, it will result in the estimation noise in the frequency and clock skew estimation, which will be discussed in Section 4.5.1.

4.4 AMKF Clock Skew Estimation

According to the measurement results, clock skew is dependent on and sensitive to the temperature and is modeled as a hybrid system with two models. It cannot be considered as a stationary random process and many estimators designed for stationary random processes are not applicable here, such as Maximum-likelihood, Wiener filter, etc. Here, we employ the Kalman filter for clock skew estimation.

Considering the hybrid system of clock skew models, it is difficult to determine the suitable model for a specific time, because the switching of models is random and unpredictable. In our previous work [110], an IMM Kalman filter is used to tackle this model uncertainty in clock skew, which can combine the output of several filters based on different models and/or parameters. Instead of making explicit decisions, the IMM Kalman filter uses the weighted sum of the outputs from different filters as the system output. The output combination is based on a soft decision by assigning different probabilities to different models. The transition between these models is considered as a Markov chain with preset transition probabilities. In each iteration, every model processes the measurement data independently and the likelihood function is calculated based on the estimation of each filter. Then we can obtain the dynamic weights, which can be considered as the posterior probability of each filter based on these likelihood functions.

Although the IMM Kalman filter has the potential to deal with model uncertainty, it is still difficult to accurately estimate the model probability because the measurement noise is relatively large. The measurement noise may lead to severe errors in

selecting the clock skew models, so the previous solution in [110] is not robust and accurate enough.

On the other hand, we note that the measurement noise for temperature is much smaller than that of clock skew as there are more additive error sources for clock skew measurement. Given the high correlation of temperature and clock skew, we can use the temperature information to assist the decision of which model to use for clock skew estimation.

Therefore, we propose the additional information aided multi-model Kalman filter (AMKF), which can estimate the model probability for one process using the model probability of another related process of which the model probability is easier to obtain. This approach is different from the traditional approaches, where the decision is based on the estimated process itself only.

To use AMKF, we need to know the model probability of the related process, which is temperature in this work. Given the fact that the temperature measurements contain relatively small noise and the environment temperature can be either stable or slowly changing, temperature is also modeled as a hybrid two-model system with a constant model and a constant velocity model. We use the IMM Kalman filter [8] to process the temperature first and obtain the corresponding model probability. At any time instant, the model for the temperature is assumed identical to the model for the clock skew because these two processes are highly positive-correlated.

The structure of the IMM Kalman filter is shown in Fig. 4.3. The IMM Kalman filter takes three steps: reinitialization, model conditioned filtering, and output estimation. The first step is to reinitialize the estimation state $\tilde{x}^{(i)}_{k-1|k-1}$ and covariance matrix $\tilde{P}^{(i)}_{k-1|k-1}$ based on the estimation results $\hat{x}^{(i)}_{k-1|k-1}$ and the corresponding covariance matrix $\hat{P}^{(i)}_{k-1|k-1}$ of all filters in the system. In the second step, we maintain several standard Kalman filters based on different models independently to process the measurements z_k. At last, the output $\hat{x}_{k|k}$ is estimated based on the estimation results from different filters and the corresponding model likelihoods Λ^i_k, which is calculated based on the estimation residuals and corresponding covariance matrix.

Once we obtain the model probability from the IMM filter for temperature, we can set a threshold to determine which model to use for clock skew estimation. In our work, the threshold may not be 0.5. This is because the constant-velocity model is a higher order model when compared with the constant model. Therefore, the estimation error caused by the inaccurate model selection of using the constant velocity model for a constant process is smaller than vice versa. Hence, the constant-velocity

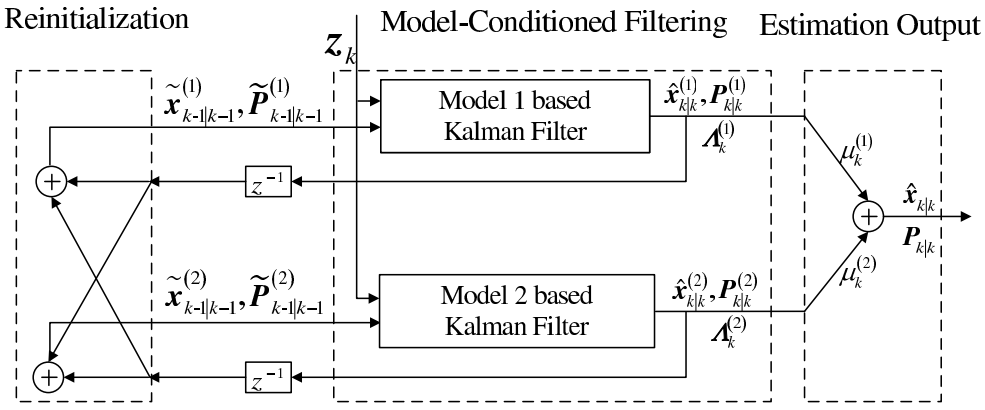


Figure 4.3: Structure of the IMM Kalman Filter.

model is slightly preferable, so it is associated with a larger likelihood. Since we largely eliminate the model uncertainty of clock skew, we can use the sequential Kalman filter here with different state transition matrix and covariance matrix of the processing noise. The mathematical details of IMM [55] and Kalman filter [100] are not presented here due to the space limitation.

The computational complexity of Kalman filter is mainly caused by the matrix inversion. As it comes to our problem, the matrices involved are all 2×2 and therefore the inversion can be calculated efficiently, so the computational complexity of the proposed solution is affordable even for sensor nodes.

4.5 Temperature-Assisted Clock Synchronization and Self-Calibration

In this section, we present the design of the two-phase temperature-assisted clock synchronization algorithm. We first explore the relationship between the crystal output frequency and the working temperature to estimate the clock skew. Then it can be directly removed in the external synchronization phase and also dynamically compensated thereafter till the following re-synchronization.

4.5.1 Online Clock Skew Estimation

As mentioned in Section 4.3.4, we can estimate the crystal output frequency based on (4.11) and the current working temperature. Then the clock skew can be estimated by (4.12). Since the temperature information from the temperature sensor

also contains noise, it will introduce errors in the crystal frequency estimation and thus the clock skew estimation. We analyze the estimation error performance first. We assume that the temperature measurement noise δ_T is Gaussian distributed with the mean of 0 and the variance of σ_T^2, i.e. $\delta_T \in \mathcal{N}(0, \sigma_T^2)$.

As presented in Section 4.3.4, the output frequency of a crystal is modeled as a quadratic function w.r.t. temperature. Therefore, an intuitive solution is to substitute the measured temperature into (4.11) to obtain the frequency estimation. However, this estimation is biased.

Lemma 1. *The frequency estimation $\hat{f}(T)$ obtained by substituting the measured temperature into (4.11) is not an unbiased estimation of the actual frequency $f(T)$ at temperature T.*

Proof. The measured temperature $\tilde{T}$ is expressed as

$$\tilde{T} = T + \delta_T, \tag{4.13}$$

where T denotes the actual temperature and δ_T is the measurement noise as mentioned before. Therefore, the estimation of frequency directly from (4.11) is

$$\tilde{f}(T) = f_0(1 - \beta(\tilde{T} - T_0)^2), \tag{4.14}$$

where f_0 is the frequency at temperature T_0. Then, we obtain the mean value of the frequency estimation at temperature T as

$$\begin{aligned} E[\tilde{f}(T)] &= E[f_0(1 - \beta(\tilde{T} - T_0)^2)] \\ &= f_0 - f_0\beta E[((T + \delta_T - T_0)^2)] \\ &= f_0 - f_0\beta E[(T - T_0)^2 + \delta_T^2 + 2\delta_T(T - t_0)] \\ &= f_0(1 - \beta(T - T_0)^2) - f_0\beta\sigma_T^2. \end{aligned} \tag{4.15}$$

As we know, the actual crystal output frequency at T should be $f_0(1-\beta(T-T_0)^2)$ [87, 35, 34], which is different from $E[\tilde{f}(T)]$ as we calculated above. This finishes the proof. □

From the proof of Lemma 1, we can see that the difference between $E[\tilde{f}(T)]$ and $f(T)$ is $-f_0\beta\sigma_T^2$, which is a known constant. Therefore, we can take it into consideration and obtain the unbiased crystal output frequency estimation.

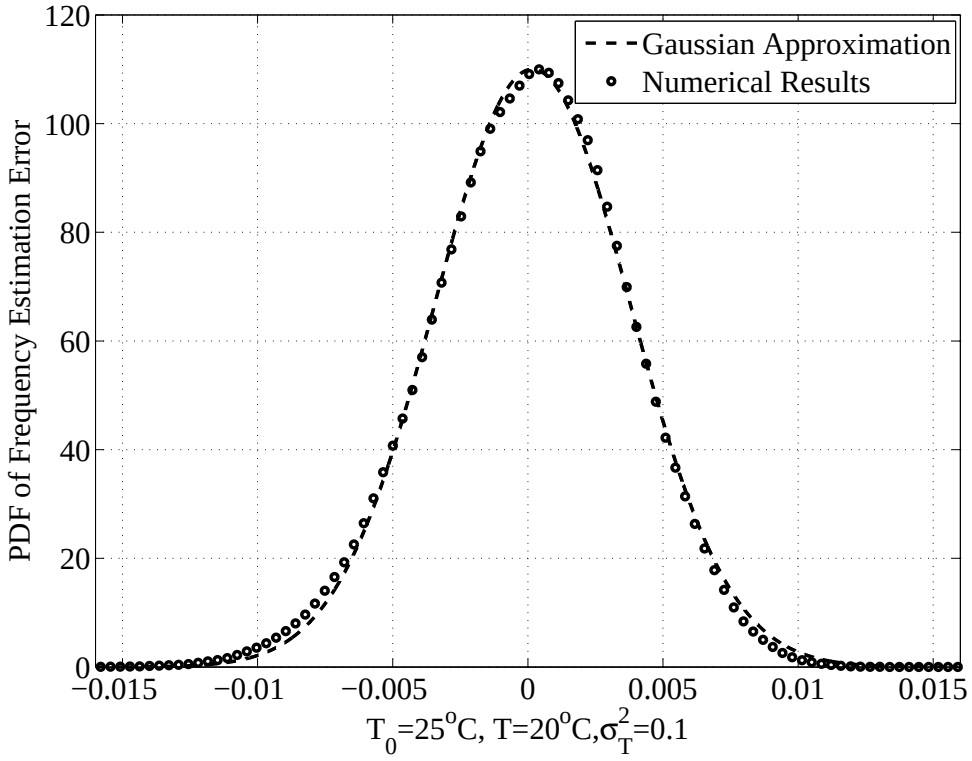


Figure 4.4: Approximated distribution of frequency estimation

Theorem 3. *The unbiased estimation of crystal output frequency based on the temperature measurement is*

$$\tilde{f} = f_0(1 - \beta(\tilde{T} - T_0)^2) + f_0\beta\sigma_T^2, \tag{4.16}$$

and the mean value of the estimated frequency is $E[\tilde{f}(T)] = f(T)$.

The proof of Theorem 3 is similar to that of Lemma 1 and therefore is omitted here due to the space limit.

The estimation error of the crystal output frequency at certain temperature is also a random variable. Numerically, we found its distribution can be well approximated by a Gaussian distribution as shown in Fig. 4.4. Then we can derive the variance of the crystal frequency estimation error as

$$\begin{aligned} \sigma_{f(T)}^2 &= E[(\tilde{f} - \bar{f})^2] \\ &= E[(f_0(1 - \beta(\tilde{T} - T_0)^2) - f_0(1 - \beta(T - T_0)^2))^2] \\ &= f_0^2\beta^2(3\sigma_T^4 + 4(T - T_0)^2\sigma_T^2). \end{aligned} \tag{4.17}$$

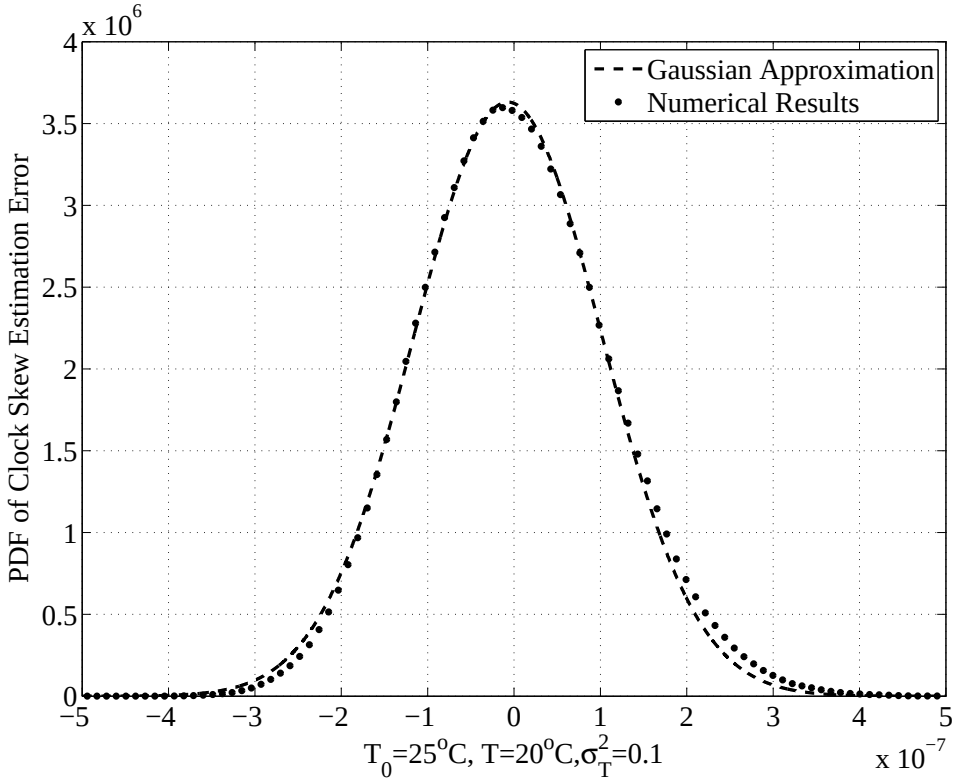


Figure 4.5: Approximated distribution of clock skew estimation

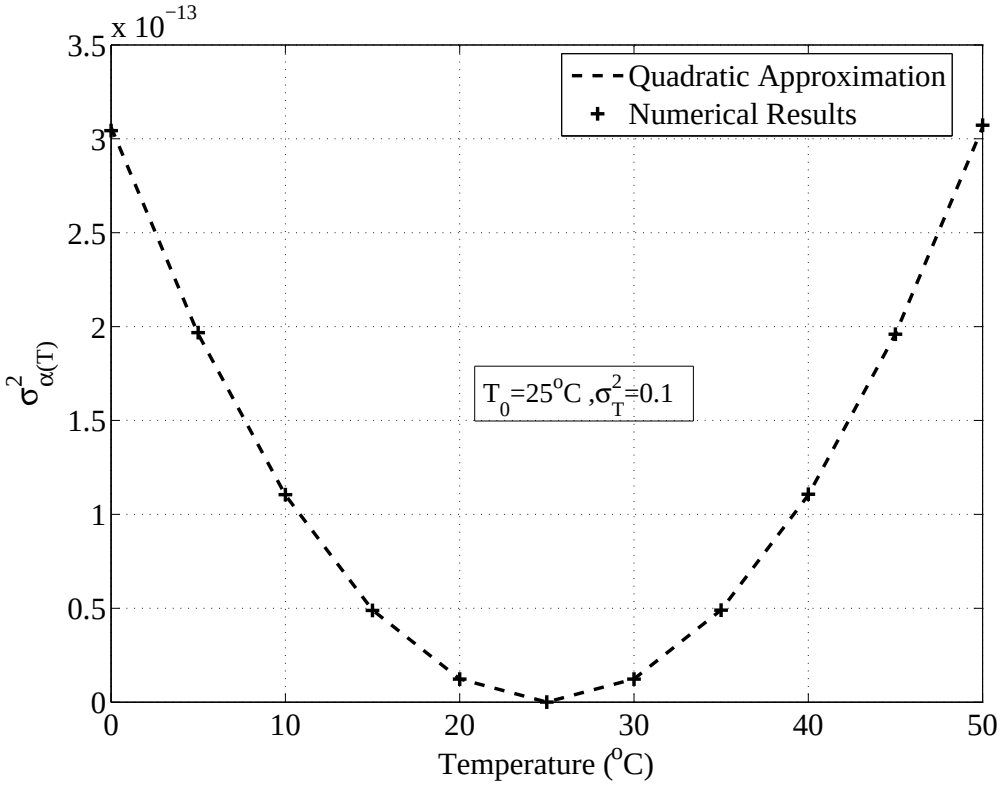


Figure 4.6: Variance of clock skew estimation error

From (4.17), the variance of the frequency estimation error is also related to temperature, and the further from the turn-over temperature T_0, the larger the variance is.

From (4.12), we can obtain the current clock skew based on the estimation of the crystal output frequency

$$\alpha(T) = \frac{f_n}{\tilde{f}(T)} - 1. \tag{4.18}$$

As we use a Gaussian distribution to approximate the crystal frequency estimation error, we can obtain the probability density function (PDF) of $(1+\alpha(T))$ accordingly. Unfortunately, even with the PDF, we still cannot obtain the closed-form expression of the mean value and variance of clock skew estimation error. We found that we can still use Gaussian distribution to well approximate the clock skew estimation error as shown in Fig. 4.5. The clock skew estimation obtained from (4.18) can be biased. However, from the numerical results, the mean value of the clock skew estimation error is below 10^{-10} level, which is negligible. The variance of the clock skew estimation error can be approximated by a quadratic function w.r.t. temperature as in Fig. 4.6.

Discussions: The clock skew changes in dynamic environments. To address this challenge, we advocate to estimate the clock skew in advance based on the relationship between the crystal output frequency and the temperature. Besides the temperature measurement error, the errors contained in the parabolic parameter β, the actual turn-over temperature T_0 and the corresponding crystal output frequency $f(T_0)$ also affect the clock skew estimation performance. These parameters can be obtained by pre-deployment system profiling [87] or by the curve fitting from the clock skew estimation results [105].

4.5.2 Two-phase Clock Synchronization Scheme

Clock synchronization is the process to mitigate the clock offset among two or multiple clocks based on the timestamps. However, the timestamp messages may not be always available due to the communication overhead or when the reference clock is temporarily unreachable, especially for WSN applications. Therefore, the sensor node needs to maintain the synchronization by its local clock. Unfortunately, as we discussed before, the uncompensated, time-varying clock skew will make the clock offset drift away quickly and unbounded. The proposed clock synchronization scheme contains two phases: (i) the external synchronization phase, during which nodes refresh their clock by exchanging the timestamp messages; (ii) the clock self-calibration, which is the interval between the two external synchronization phases and calibrate the clock based on the local information. Different from the previous work [115, 53, 87, 82], we take the environment information into account.

External Clock Synchronization Phase

During this phase, local node exchanges timestamp messages with the reference clock as presented in Section 4.3.2 and the procedure is shown in Fig. 4.1. Based on this two-way timestamp exchange, we can obtain a sequence of equation sets as (4.9). Since the timestamp exchange can be accomplished within a short period, we assume the two-way transmission between the local node and the reference node undergoes the same delay with independent delay jitters [53, 74]. We can rewrite the two-way timestamp exchange model (4.9) by adding the two equations together, and the new model is

$$D_1[i] + D_2[i] = 2\theta_0 + \alpha(T)(t^i_{A,1} + t^i_{A,2} - 2t^1_{A,1}) + (1 + \alpha(T))(\omega_{d,1}[i] - \omega_{d,2}[i]). \quad (4.19)$$

Different from the previous work which tries to estimate both clock offset θ_0 and clock skew α simultaneously, such as using maximum likelihood estimators [53, 74], we leverage the temperature information to estimate clock skew, as presented in Section 4.5.1. Then we directly remove the clock skew accordingly to improve the clock offset estimation accuracy. On the other hand, since we only need to estimate clock offset instead of two unknown variables, we can reduce the computational complexity of the estimator. Let $\tilde{\alpha}(T)$ denote the estimated clock skew, we can obtain the ith clock offset estimation $\tilde{\theta}_0[i]$ from (4.19) as

$$\tilde{\theta}_0[i] = \frac{1}{2}[D_1[i] + D_2[i] - \tilde{\alpha}(T)(t^i_{A,1} + t^i_{A,2} - 2t^1_{A,1})]. \tag{4.20}$$

As we mentioned before, the two-way timestamps exchange is conducted multiple times sequentially to smooth the noise out. Therefore, the estimation of clock offset θ_0 is

$$\tilde{\theta}_0 = \frac{1}{2N}\sum_{i=1}^{N}\tilde{\theta}_0[i], \tag{4.21}$$

where N is the number of two-way timestamps. This estimator only involves scalar operations. From (4.19) and (4.21), we can see that there are $9N$ add operations and N multiply operations involved, which is much simpler than the previous work [53, 74], where there are at least $14N$ add operations and $7N$ multiply operations.

Clock Self-calibration Phase

An effective compensation of clock skew can improve the clock synchronization accuracy and prolong the clock re-synchronization period and thus less frequent timestamp exchange is needed, which is particularly desirable for WSNs where communication energy is costly and frequent re-synchronization may not be feasible due to mobility, sleep scheduling, etc. However, as demonstrated in previous work [105, 87, 82] and in the crystal datasheet [36, 35, 34], the clock skew is not constant and varies with the surrounding environment. Therefore, the compensation of clock skew for WSNs is non-trivial, so we propose the temperature-assisted clock self-calibration (TACSC) scheme which takes the temperature information into account to dynamically compensate clock skew.

During the self-calibration phase, a local variable $\tilde{\theta}[n]$ is used to indicate the estimated clock offset in the n-th time slot, which is the accumulated clock skew since the last external synchronization phase. Due to the thermal inertia, temperature

usually does not change quickly in a small time period, such as a few seconds. We assume that clock skew does not change during one sampling period. Thus, the instantaneous temperature measurement can represent the temperature for the whole sampling period. We can use the current temperature to obtain the instantaneous clock skew $\tilde{\alpha}(T_n)$. Then we can update the estimated clock offset as

$$\tilde{\theta}[n] = \tilde{\theta}[n-1] + \tau\tilde{\alpha}(T_n) = \tilde{\theta_0} + \tau\sum_{i=1}^{n}\tilde{\alpha}(T_i), \tag{4.22}$$

where T_i is the temperature measurement in the ith slot and τ is the temperature (clock skew) sampling duration. As proved in [115], (4.22) is an unbiased estimation of clock offset $\theta_i[n]$. We are also interested in its error uncertainty.

Theorem 4. *The variance of the clock offset estimation by (4.22) at Nth slot is*

$$\sigma^2_{\theta[N]} = \sigma^2_{\theta_0} + \tau^2\sum_{i=1}^{N}\sigma^2_{\alpha(T_i)}, \tag{4.23}$$

where $\sigma^2_{\theta_0}$ is the variance of clock offset estimation in the external synchronization phase.

Proof. The true value of clock offset at Nth time slot is

$$\theta[N] = \theta_0 + \tau\sum_{i=1}^{N}\alpha(T_i). \tag{4.24}$$

Therefore, from (4.22) and (4.24), the clock offset estimation error is

$$\delta_{\theta[N]} = \tilde{\theta}[N] - \theta[N] = \delta_{\theta_0} + \tau\sum_{i=1}^{N}\delta_{\alpha(T_i)}, \tag{4.25}$$

where δ_{θ_0} is the estimation error in the external clock synchronization phase and $\delta_{\alpha(T_i)}$ is the clock skew estimation error, which is approximated as a Gaussian noise with the variance related to temperature as presented in Section 4.5.1. Therefore, from (4.25), we have the error variance as

$$\sigma^2_{\theta[N]} = \sigma^2_{\theta_0} + \tau^2\sum_{i=1}^{N}\sigma^2_{\alpha(T_i)}.$$

This finishes the proof. □

From Theorem 4, given the elapsed time of TACSC ($\Delta t = N\tau$), we can see that a smaller sampling period τ leads to a smaller error variance at the cost of a higher energy consumption. Without loss of generality, we set the sampling period τ as 1s for simplicity in simulation.

Given the estimation of clock offset $\tilde{\theta}[n]$, we can compensate it to the local clock as follows:

$$C_i[n] = C_i[n] - \tilde{\theta}[n]. \tag{4.26}$$

Given the unavoidable clock offset and clock skew estimation errors, the clock will eventually drift away but at a much slower pace, as we will show in Section 4.6.

4.6 Performance Evaluation

In this section, we first present the numerical results of the clock skew estimation. We then evaluate the performance of clock offset estimation, i.e., the results of the external clock synchronization phase. At last, the performance of the proposed TACSC scheme by simulation and by experiments using our sensor testbed is studied.

4.6.1 Performance of AMKF

We generate the simulation traces based on the measurement results in a temperature-changing environment. The temperature alternates between two states: either being constant for a while, or changing smoothly in a certain duration. Assume that the reference clock works in a stable environment or is well calibrated, and its clock jitter can be treated as part of the noise. In the simulation, the temperature varied from 20℃ to 50℃ and the corresponding clock skew varied from 20 μs to 40 μs. The measurement noises followed a Gaussian distribution. We used a set of training data to obtain the table for clock skew compensation w.r.t. temperature.

We first applied the IMM Kalman filter on temperature measurements to determine the clock skew model at each time instant, and the results are shown in Fig. 4.7. During the period that the temperature changes from 20℃ to 50℃ (from the 500 to 600 s), the probability of the constant velocity model is close to one; during the period that the temperature is stable, the probability is below 0.5. As a result, during the temperature changing period, the correct model is selected for clock skew almost

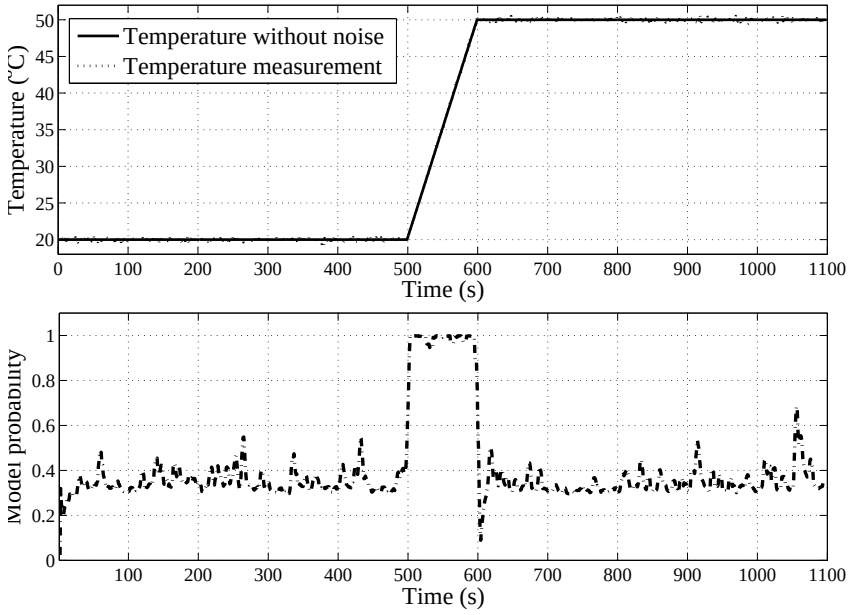


Figure 4.7: Temperature measurements and the probability of constant velocity model.

surely; while during the temperature stable period, there is a small probability that the constant velocity model is selected, which is acceptable as discussed in Sec. 4.4.

For the two clock skew models, the constant and the constant velocity model, we choose the clock state vector as $x_k = [\theta_k, \alpha_k]$. The corresponding transition matrix and processing noise covariance matrix are

$$\begin{aligned} A^{(1)} &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad Q^{(1)} = \beta \begin{bmatrix} T^2 & 0 \\ 0 & 0 \end{bmatrix}, \\ A^{(2)} &= \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}, \qquad Q^{(1)} = \beta \begin{bmatrix} T^4/4 & T^3/2 \\ T^3/2 & T^2 \end{bmatrix}, \end{aligned} \tag{4.27}$$

where β is a scalar to adjust the processing noise level, which reflects the accuracy of the model. It can be used to make a tradeoff between estimation accuracy and filter suitability. The similar setting is used for temperature except that the state vector contains temperature and temperature change.

Given the model probability obtained from the IMM filter on temperature measurements, the model in the Kalman filter for skew estimation is selected. The estimation results are shown in Fig. 4.8. From the figure, the estimation results using the proposed AMKF (the dashed line) are close to the real values (the solid line),

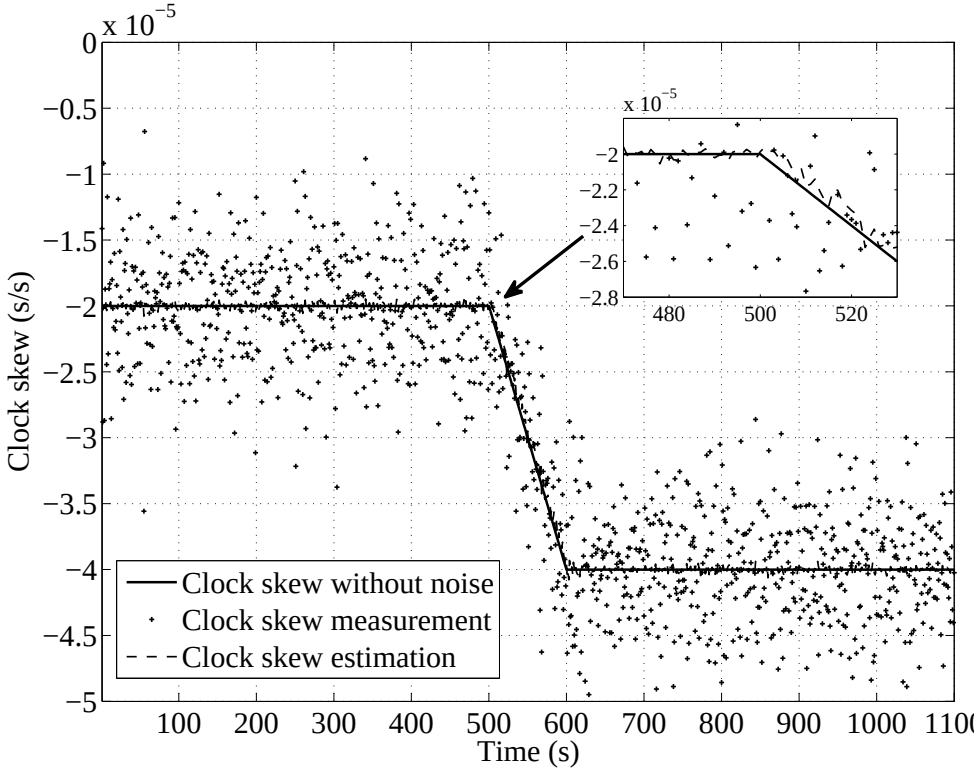


Figure 4.8: Clock skew estimation.

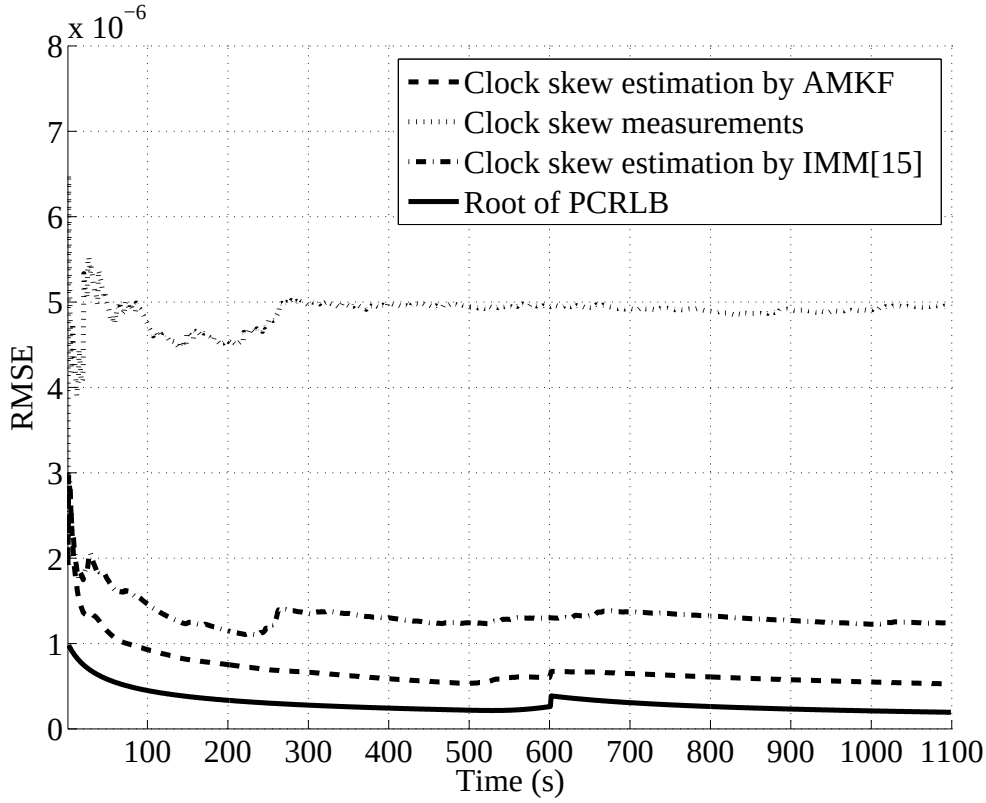


Figure 4.9: RMSE of clock skew estimation.

even though the clock measurements (the dots) contain severe noise. This is because we can detect the change of model efficiently using temperature estimation, so the negative impact of model uncertainty is minimized. From the figure, the estimation errors are most obvious during the initial stage (0 to 50 s) and the period that the temperature changes to the constant state (600 to 650 s). This is because that we intentionally set the threshold of the IMM filter to favor constant velocity model, so the probability of selecting a wrong model during these periods is higher than that of other periods. Nevertheless, the estimation error of the proposed algorithm is always much less than 1 μs/s if the filter converges.

To further evaluate the estimation error, the root mean square error (RMSE) of the proposed AMKF algorithm is shown in Fig. 4.9. We also compare the proposed AMKF with our previous solution in [110], which used the IMM Kalman filter to process the clock skew measurements without the assistance from temperature measurements. As shown in [110], the IMM used in [110] can outperform the non-adaptive estimation algorithms (such as the one in [30]), so it is the best known solution. From the figure, the RMSE of the proposed AMKF is less than 50% of the best known re-

sults in [110]. PCRLB offers a statistic lower bound of RMSE, and the results in Fig. 4.9 show that the performance of the proposed AMKF is close to the bound. Generally speaking, the Kalman filter can approach the PCRLB in a linear Gaussian system with sufficient measurements. However, in dynamic environments, the process within the same model cannot be arbitrary long due to the variations of clock skew and temperature models. Therefore, the gap to reach the PCRLB exists, which is mainly caused by the model change.

4.6.2 Performance of the External Clock Synchronization Phase

The external clock synchronization is based on the two-way timestamp exchange shown in Fig. 4.1. The transmission delay D and the standard deviation of delay jitter σ_D are set as 100ms and 10μs, respectively. The temperature measurement noise is assumed to be zero mean Gaussian distributed with the variance $\sigma_T^2 = 0.1$. Considering the tuning folk crystal oscillator used in Mica2 nodes [89], the nominal frequency is 32,768Hz and the parabolic parameter β is set to 0.04ppm. The turnover temperature and the corresponding output frequency are 25℃ and $32,768.5$Hz, respectively.

The root mean square errors (RMSE) of the clock offset estimation based on the two-way timestamp exchange are shown in Fig. 4.10. We compare the proposed temperature-assisted external clock synchronization scheme with the existing solution [53], which tries to estimate the clock offset and clock skew simultaneously. We also evaluate the Cramer-Rao Lower Bound (CRLB) as a benchmark [74]. As presented in Section 4.3.4, the estimation variance of clock skew is also related to the temperature, which will affect the synchronization accuracy. Therefore, we evaluate the synchronization performance under different temperatures (0℃ and 10℃).

From Fig. 4.10, without the parabolic parameter error, the proposed temperature-assisted synchronization can significantly outperform the existing algorithm and reduce the RMSE by more than 90% and close to the performance bound. This is because the previous solutions estimate the clock offset and clock skew simultaneously, while we only need to estimate the clock offset and therefore the information contained in the measurements is dedicated to one unknown variable. Besides, the computational complexity is reduced. We just show the results at $T = 0$℃ and omit the results at $T = 10$℃, which largely overlap with the performance bound.

In reality, in order to obtain the parameters that define the parabolic function

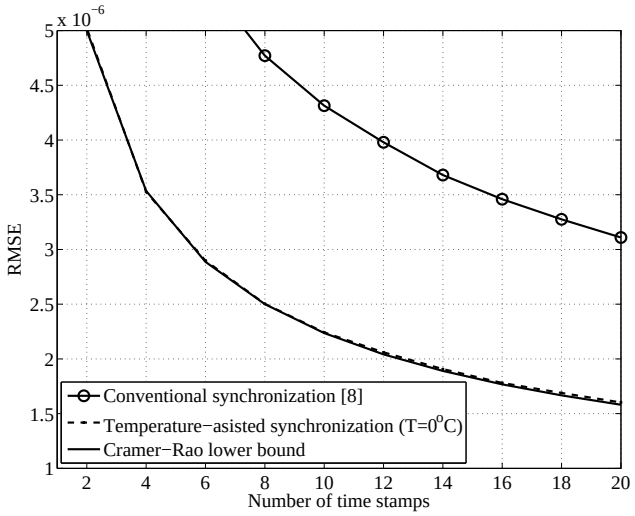


Figure 4.10: Clock offset estimation, no parabolic parameter error

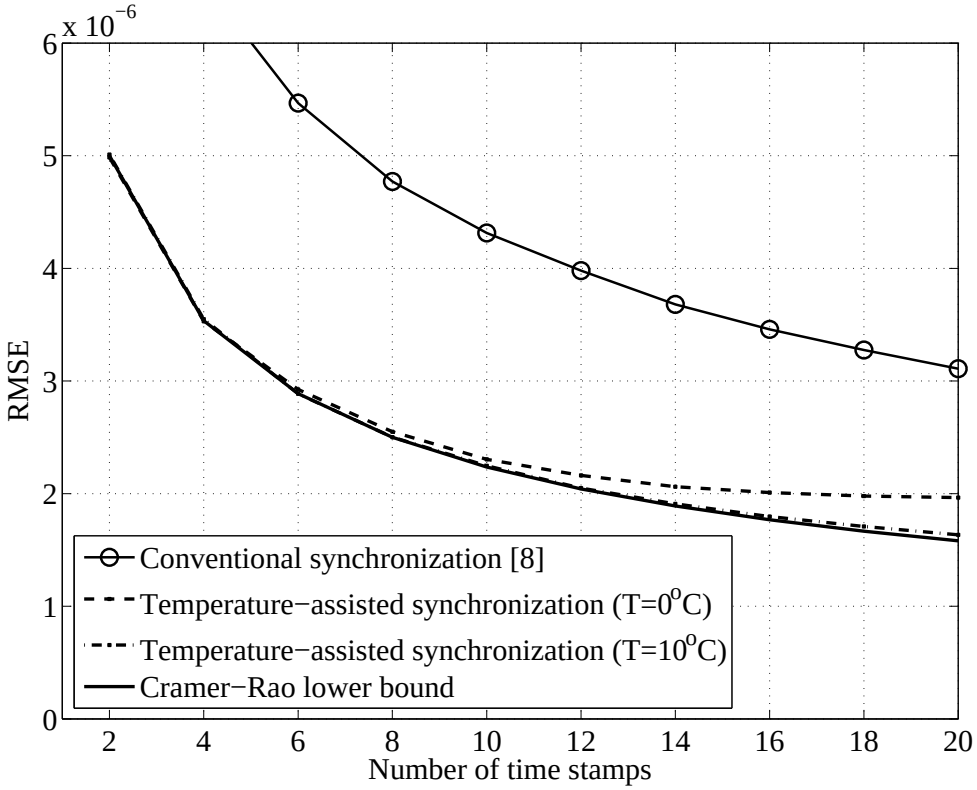


Figure 4.11: Clock offset estimation, with parabolic parameter error

of crystal output frequency v.s. temperature, i.e., f_0, T_0 and β, we need clock skew values w.r.t. different temperature values as the training data.

We can obtain the parabolic parameters with at least three clock skew v.s. temperature estimation pairs, and the estimation accuracy will be improved with more training data. Then we can use the minimum mean square error (MMSE) algorithm to estimate these parabolic parameters. However, such estimation also contains errors, which will affect the synchronization performance. The standard deviation of the clock skew estimation error can be as low as $10^{-2}\mu$s [115, 105]. Assuming that the standard deviation of the clock skew estimation error is 0.1μs and that of the temperature is 0.1℃, we used 8 clock skew v.s. temperature pairs as the training data set to estimate the parabolic parameters and then use them to dynamically estimate the crystal output frequency and thus the corresponding clock skew. From the results shown in Fig. 4.11, we can see the RMSE is lower at $T = 10$℃ than that at $T = 0$℃. This is because that the parabolic parameters errors have a larger impact at the temperature further away from the turn-over temperature (25℃). Besides, the variance of clock skew estimation error is also larger at $T = 0$℃ than that at 10℃. Nevertheless, even with estimation errors in parabolic parameters, the proposed scheme can still significantly outperform the existing algorithm.

4.6.3 Performance of the Clock Self-calibration Phase

To evaluate the performance of the TACSC scheme and compare it with the constant skew compensation solution, we generated simulation data as shown in Fig. 4.12, which simulate the temperature changing from the sunset to midnight (summer time, Victoria, BC) and the corresponding clock skew. We set the sampling period T to be 1 s.

The results of TACSC and the constant clock skew compensation are compared in Fig. 4.13. As shown in the figure, the constant clock skew compensation only works for certain circumstances and the performance degrades severely once the working environment changes because it cannot adaptively compensate the changing clock skew. On the other hand, we can see that the clock offset is well compensated using the proposed TACSC and the clock synchronization error is more than two orders of magnitude lower than that by the constant compensation algorithm.

Different from the external synchronization phase, the TACSC scheme is more sensitive to the errors of parabolic parameters. As demonstrated in Fig. 4.14, the

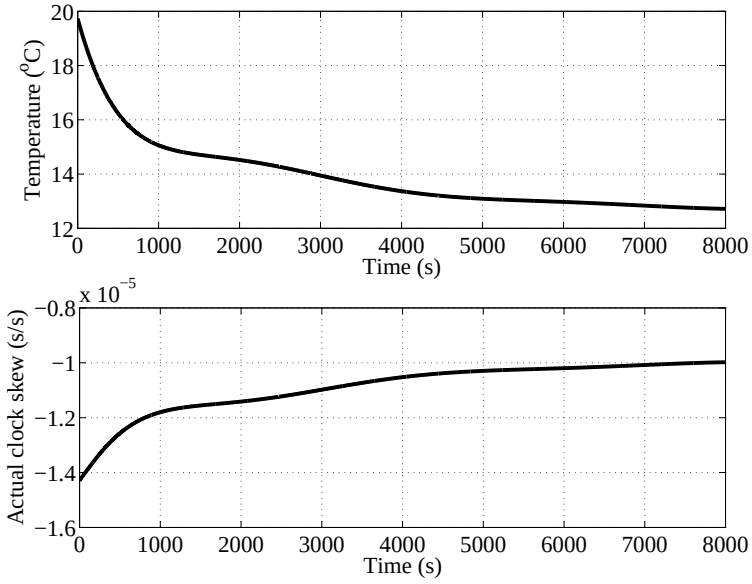


Figure 4.12: Simulation data for TACSC

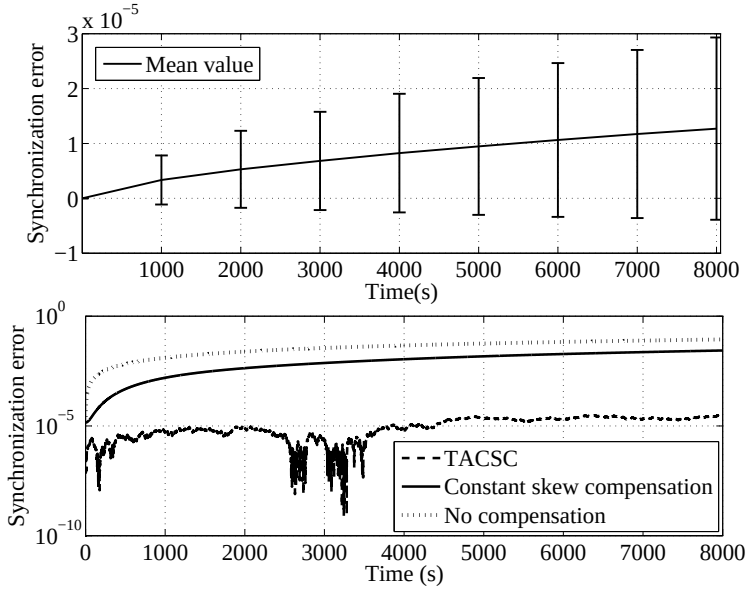


Figure 4.13: Temperature-assisted clock self-calibration, without parabolic parameter error

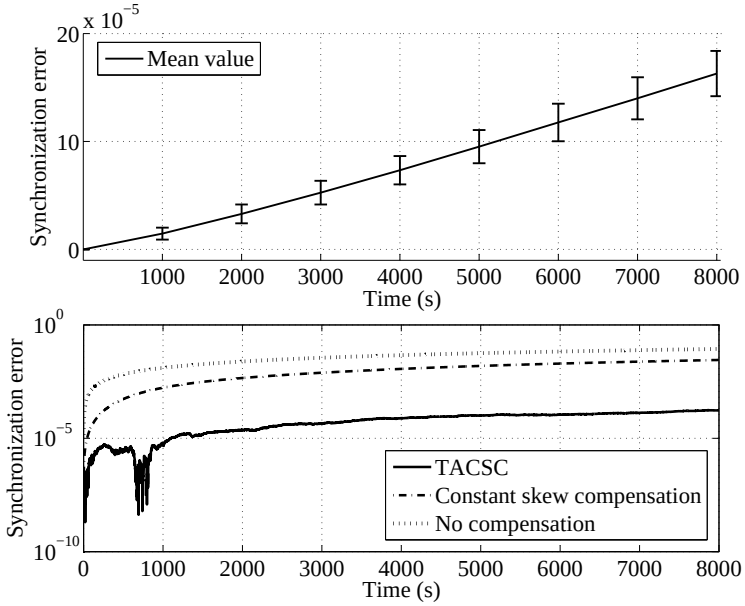


Figure 4.14: Temperature-assisted clock self-calibration, with parabolic parameter error

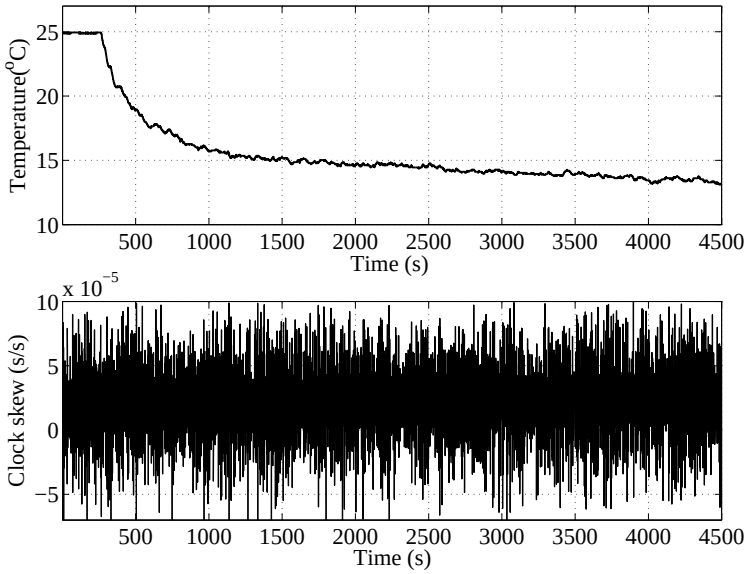


Figure 4.15: Training trace

mean synchronization error of TACSC is up to 160μs, which is much higher than that in Fig. 4.13 (less than 15μs). This is because the clock skew estimation error caused by the imperfect parabolic parameters will be accumulated over a much longer calibration phase. However, the synchronization error of the proposed TACSC scheme is still much lower than that of the non-dynamic clock skew compensation scheme, which is up to 0.115s, and the improvement is about three orders of magnitude.

We also conducted experiments using the Mica2 testbed to demonstrate the applicability and feasibility of the proposed scheme in a real world. We used the temperature and clock skew measurements shown in Fig. 4.15 as the training dataset to obtain the clock skew v.s. temperature and estimate the parabolic parameters by curve fitting. Then, several months later, we used the same sensor and the laptop to conduct the same measurement, where the sensor was working in outdoor environment in Victoria, BC, and the traces of temperature and clock skew are shown in Fig. 4.16. According to the parabolic parameters estimated by the training data, we can estimate the crystal output frequency (clock skew) and then dynamically compensate the clock skew of the sensor node.

The compensation results are shown in Fig. 4.17. From the figure, the clock offset by TACSC is much lower than those without compensation or with a constant skew compensation. The clock offset by TACSC is always below 8.5ms over the 88, 593s (more than 24h) test duration, which is an order of magnitude improvement

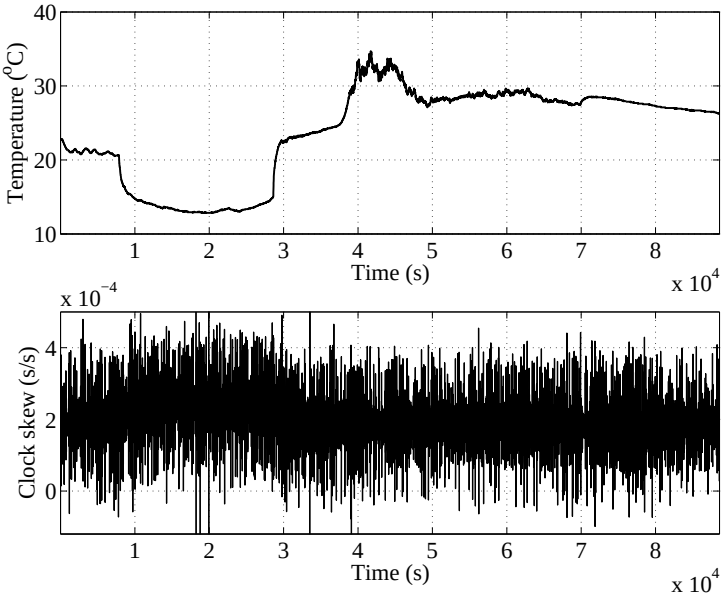


Figure 4.16: Verification trace

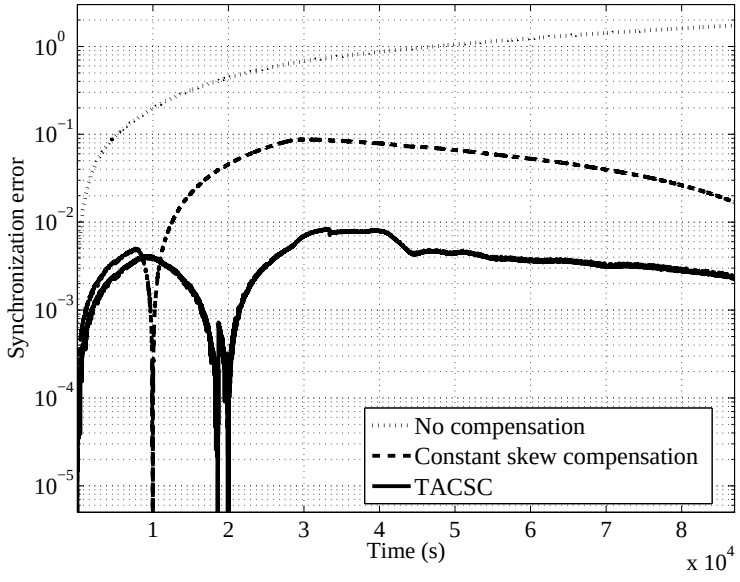


Figure 4.17: Verification results

to the constant skew compensation, where the clock offset is up to 95ms.

On the other hand, it is not surprising that the experiment results are worse than that of simulations due to the following reasons. First, some often-used assumptions may not hold in a practical system, such as the measurement noise may not be strictly Gaussian distributed, which will degrade the estimation accuracy. More importantly, we used a laptop in a lab environment as the reference node, as the clock of the laptop is much more stable than that of the sensor node. However, the laptop clock circuit also suffers from thermal noise and other uncertain factors, and the lab room temperature is not a constant as our lab has no temperature control, which also introduces errors. Furthermore, the estimation error of parabolic parameters will be accumulated. These imperfect practical conditions degrade the performance of the clock skew estimation in the real test. Even with these impairments, the test results show that TACSC can meet our design objective, i.e., reduce the clock offset and prolong the re-synchronization period by an order of magnitude, and its performance is reliable over several months (and probably more).

4.7 Conclusions

In this chapter, we have investigated the temperature-assisted clock synchronization and clock self-calibration synchronization for wireless networks, which is specially designed for the dynamic working environments. The crystal output frequency has been modeled as a quadratic function w.r.t. temperature, so we can estimate it according to the current working temperature and thus the clock skew. We have then proposed to directly remove the clock skew during the external clock synchronization process to improve the synchronization accuracy with a much lower computational cost. For the interval between synchronization actions, we have proposed a TACSC scheme to conduct the real-time dynamic clock skew compensation. Simulation and experiment results have demonstrated the effectiveness and efficiency of the proposed TACSC scheme.

4.8 Symbol List

Symbol	Explanation
$C_A(t)$	output of clock A at time t
$\theta_B^A(t)$	clock offset of clock A w.r.t. clock B
$\alpha_B^A(t)$	clock skew of clock A w.r.t. clock B
τ	sampling period
D	measured clock offset
t_c	message construction delay
t_t	propagation delay
t_p	processing delay
ω	delay jitter
$t^i_{A(B),*}$	time stamp of node A(B) at ith exchange
d	total message delay
ρ	changing rate of clock skew
ϕ	processing noise
T	temperature
$f(T)$	crystal frequency at temperature T
$\tilde{x}^{(i)}_{k-1\|k-1}$	estimation result of Kalman filter i at $k-1$th step
$\tilde{P}^{(i)}_{k-1\|k-1}$	estimation covariance matrix of Kalman filter i at $k-1$th step
z_k	measurement of AMKF at kth step
δ_T	temperature measurement noise
β	parabolic coefficient
A	state transition matrix
Q	process noise covariance matrix

Chapter 5

Scheduling for Concurrent Transmission

5.1 Introduction

Many wireless networks have a mesh topology where nodes can communicate with each other in a peer-to-peer fashion. Representatives of such systems include the wireless mesh networks [3], ultra-wide band (UWB) wireless personal networks (WPANs) and home networks [103]. In these systems, how to improve the system performance by scheduling concurrent transmissions appropriately is an important and challenging issue, due to the multi-user interference (MUI) of concurrent transmissions.

In a narrow-band rate-nonadaptive wireless system, each flow can successfully transmit at a specified data rate so long as the received SINR is above certain threshold. In this case, letting more flows transmit concurrently without violating their SINR requirements means that we can have higher system throughput. The optimal scheduling problem for this case can be converted to a Knapsack problem which is NP-complete [77]. We have seen a great deal of research efforts in determining the system capacity and optimizing the concurrent scheduling for such systems [29].

In most broadband wireless systems, the PHY-layer can adjust the modulation and coding schemes used, so the link data rate is a variable which is determined by its received SINR. When we schedule multiple flows to share a time slot, from the perspective of an individual flow, its performance in terms of data rate or bit error probability in the slot might be degraded due to the interference from other flows. On the other hand, with more flows sharing the time slots, each flow might be

allocated with more time slots, hence the long-term throughput of each flow and the entire network throughput will be improved. Scheduling for rate-adaptive broadband wireless networks with random network topology is thus much more difficult than that in rate-nonadaptive systems: a) in each time slot, each flow's throughput (data rate) is not known before the scheduling decision of that slot is finalized; and b) when we add or remove one flow from the set sharing the slot, the throughput of all other flows in the set will change.

Given the hardness of the concurrent scheduling problem for rate-adaptive wireless systems and the fact that the real-time scheduling decision should be made within a few milliseconds, optimal scheduling algorithms will be impractical to deploy due to the computational cost. In this chapter, we propose practical heuristic scheduling algorithms that can achieve significantly better performance than the existing solutions and with tolerable computational complexity.

The main contributions of this chapter are three-fold. First, we formulate an optimal scheduling problem aimed to maximize the network throughput and maintain long-term fairness among competing flows. We decompose the problem and propose a simple searching algorithm which can quickly move towards better scheduling solutions. Thus, the scheduling decision can achieve high throughput and maintain long-term fairness. Second, in the case that each flow has a short-term minimum service rate requirement, we formulate a constrained optimal scheduling problem. To satisfy the constraints, we consider the dual problem and apply Lagrangian relaxation. We incorporate a dual update procedure in the proposed searching algorithm to ensure that the searching results satisfy the constraints. Finally, extensive simulations are conducted to demonstrate the effectiveness and efficiency of the proposed scheduling algorithms which are found to achieve higher throughputs than the existing solutions and ensure better fairness.

5.2 Related Work

The joint scheduling and routing problem has been heavily investigated [4, 56, 112]. These work employed a conflict graph to bound mutual-interference when scheduling concurrent transmissions. However,this approach is not suitable for rate-adaptive wireless systems where the transceiver can adjust the link data rate to tolerate different levels of interference. References [17, 61] considered rate-adaptive wireless networks, but they focused on the minimum power allocation problem and the

application-rate assignment problem instead of the throughput maximization problem we consider here.

As most broadband wireless systems adopt adaptive modulation and coding schemes, concurrent scheduling problem for rate-adaptive wireless networks begins to attract attention. In [62, 63],based on the concept of "exclusive region" [79], several heuristic scheduling algorithms with polynomial time complexity were proposed. In their work, each receiver defines an exclusive region, and flows are scheduled to transmit concurrently only if the senders are outside the exclusive regions of other receivers. By allowing concurrent transmissions without violating the exclusive region conditions, the network throughput can be made much higher than that with the time division multiple access (TDMA) scheme. However, using exclusive region cannot guarantee the optimal searching direction. In addition, how to constrain the scheduling decisions (such as to guarantee the minimum flow throughput) has not been addressed.

In [80], a novel quadratic lower bound for the capacity of single flow was employed and such approximation is involved in the dual problem of the primal non-convex optimization problem, where certain constrained sub-problems were then solved. However, the computational complexity of the approach is $O(n^3 2^n)$ which is too high for real time schedulers.

Besides the wireless network scheduling, there are problems in other communication and networking systems with similar mathematical models. In [52], Lee et al. studied the optimization and rate control problem for multi-class services in the Internet, aimed to maximize the utility function which may not be concave (i.e., non-convex minimization). They showed that rate control algorithms based on the dual method by a pricing-based mechanism, developed for concave utility function, can be used for problems with non-concave utility function. Their problem is still simpler than the scheduling problem to be investigated here. This is because their non-concave utility functions are chosen to be sigmoid-like which are not difficult to evaluated, compared to our non-concave objective function (for flow date rate under MUI). Lagrangian dual relaxation was studied in [13, 111, 65] to solve the dynamic spectrum management problem in digital subscriber line (DSL) systems. Yu and Lui [111] first discovered the zero duality gap under the so-called time sharing condition. Subsequently, Luo and Zhang [65] rigorously proved some results for the continuous Lebesgue integral formulation. However, the optimal algorithms proposed in [111] is of very high computational cost, and they do not consider the fairness and minimum flow throughput requirements.

5.3 Practical Scheduling Algorithms for Concurrent Transmissions

5.3.1 System Model

We consider a wireless network with a set of $2N$ active nodes, $\mathcal{N}$, randomly deployed in a region. These nodes are divided into two disjoint sets, namely the transmitter set $\mathcal{S}$ and the receiver set $\mathcal{C}$. The two sets have the same cardinality $|\mathcal{S}| = |\mathcal{C}| = \frac{1}{2}|\mathcal{N}| = N$. One transmitter s_i in $\mathcal{S}$ and one receiver c_i in $\mathcal{C}$ form a communication pair (s_i, c_i) (flow i). The instantaneous data rate of flow i is determined by the received SINR.

We consider the scenarios in that wireless devices can directly communicate to or relay for each other in a peer-to-peer fashion, i.e., the network has a mesh topology. For presentation clarity, in what follows we use IEEE 802.15.3 UWB WPANs to describe the scheduling problem. Our approach can be applicable to other wireless systems with a mesh topology.

According to the IEEE 802.15.3 standard, several wireless devices can autonomously form a piconet in which one of the nodes should be selected as the piconet coordinator (PNC). The system model of a piconet is shown in Fig. 5.1. The system uses a superframe structure in the time domain, where each superframe consists of three phases: the Beacon period (BP) for network synchronization and control messages broadcast by the PNC, the contention access period (CAP) for devices sending requests to the PNC using the carrier sensing multiple access/collision avoidance (CSMA/CA) MAC protocol, and the channel time allocation period (CTAP) for data transmissions among devices in a peer-to-peer fashion. The CTAP contains several channel time slots. Currently, TDMA is used to allocate each time slot to a specific flow, i.e., each time slot is occupied by one flow exclusively. Without loss of generality, we assume that the number of time slots in a superframe equals N, and the scheduler (the PNC) makes a scheduling decision for every superframe. For the physical layer, we employ the path-loss model as presented in 2.3.2.

5.3.2 Scheduling Problem Formulation

For time slot k, the scheduling decision can be represented by a control vector $U_k = (u_{k,1}, u_{k,2}, \cdots u_{k,N})$, where variable $u_{k,i}$ controls the channel access of flow i with $u_{k,i} = 1$ if flow i is scheduled to transmit in slot k, and $u_{k,i} = 0$ if i is idle in slot k. Denote

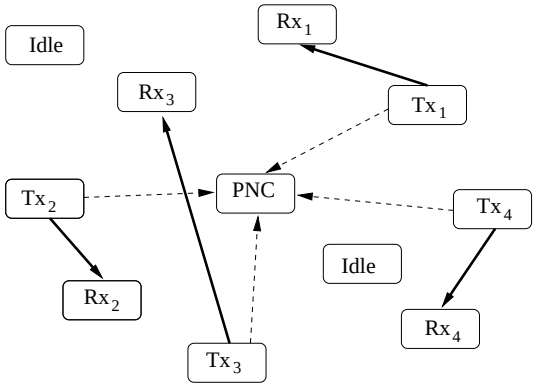


Figure 5.1: System model of a piconet.

by $R_{U_{k,i}}$ the throughput of flow i in slot k, we can write

$$R_{U_{k,i}} = \eta W \cdot \log_2(1 + \frac{u_{k,i}\kappa G_{i,i} P_t d_{i,i}^{-\gamma}}{N_0 + b\sum_{l \neq i} u_{k,l}\kappa G_{l,i} P_t d_{l,i}^{-\gamma}}), \tag{5.1}$$

where b is the multiple user interference factor, W is the signal bandwidth, η is the coefficient describing the efficiency of the transceiver design, G is the fading gain, d is the transceiver distance, κ is the reference path-loss, and γ is the path-loss factor.

Under these circumstances, the design goal of the scheduling algorithm is to determine the optimal $U_1, U_2, ..., U_N$ to maximize the system throughput.

To maximize the total throughput, we first formulate the scheduling problem as an unconstrained optimization problem:

Problem 3. *(P3)*

$$\max_{u_{k,i} \in \{0,1\}} \sum_{k=1}^{N} \sum_{i=1}^{N} R_{U_{k,i}}. \tag{5.2}$$

The optimization problem (P3) is a nonlinear integer programming problem. One possible approach is to relax the integer variables to continuous ones, and use optimization tools to solve the approximated problem. However, even the approximated problem is difficult to solve, since its objective function is not necessarily concave in U_k. This is because the first term of the Hessian matrix of the objective function can be negative. There is no existing software to effectively solve the non-convex optimization problem.

The optimization problem (P3) is more difficult than the optimal single slot scheduling problem for rate-nonadaptive networks, which can be reduced to a 0-1 Knapsack problem [77]: Consider the case where the data rate of each flow in a slot,

corresponding to the profit of items in the knapsack problem, is fixed so long as the total interference is smaller than a threshold. Items (flows) can be added into the knapsack (allocated to use the slot). The objective is to maximize the total profit (total throughput) with the weight (interference) constraint. It is well-known that the Knapsack problem is NP-complete [77]. The problem we investigate here is even harder than the Knapsack problem, since the profits of items (flow throughput) will change according to the selected subsets. It results in the hardness of this scheduling problem, and the existing approximation algorithms for Knapsack problems cannot be used for our problem.

In addition, it is obvious that the optimal algorithm (if exists) for (P3) will lead to a biased solution, i.e., all the resources (time slots) will be allocated to those flows with better channel conditions while some unlucky flows will be starving.

To ensure long-term fairness among competing flows, we employ a control parameter based on the weighted fair queuing [18] as

$$\rho_{k-1,i} = \frac{w_i}{(\sum_{m=1}^{k-1} R_{U_m,i} + \epsilon)^\alpha}, \tag{5.3}$$

where ϵ is a small positive scalar in order to prevent zero denominator, w_i is a weight for flow i to provide differentiated services, and α is a parameter to make a tradeoff between fairness and network throughput. This control variable may be considered as the dynamic weight of each flow determined by the resource allocation schemes up to $(k-1)th$ time slot, $U_1, U_2, \cdots, U_{k-1}$. The optimal scheduling problem considering the long term fairness can now be formulated as

Problem 4. *(P4)*

$$\max_{u_{k,i} \in \{0,1\}} \quad \sum_{k=1}^{N} \sum_{i=1}^{N} \rho_{k-1,i} \cdot R_{U_{k,i}}, \tag{5.4}$$

where initially $\rho_{0,i} = 1/\epsilon$.

For the cases that a minimum throughput $R^i_{\min}$ of each flow needs to be ensured, the optimal scheduling problem with the constraints is formulated as

Problem 5. *(P5)*

$$\max_{u_{k,i}\in\{0,1\}} \quad \sum_{k=1}^{N}\sum_{i=1}^{N} \rho_{k-1,i} \cdot R_{U_{k,i}} \tag{5.5}$$

$$\text{s.t.} \quad \sum_{k=1}^{N} R_{U_{k,i}} \geq R^{i}_{\min} \ \forall i. \tag{5.6}$$

To deal with the NP-hard optimal scheduling problems (P3), (P4) and (P5), we use a dynamic programming approach. We decompose each problem into several smaller decision problems to reduce the computational complexity. The technical details of our approach are described in the following part.

5.3.3 Single-flip Algorithm

It is difficult to approximate the unconstrained scheduling problems (P3) and (P4), due to the data rate adaptation. Considering the hardness of the problem and that a scheduling algorithm needs to be executed with very limited time (typically in the order of milliseconds), we develop heuristic algorithms that offer good performance with tolerable complexity.

Intuitively, the optimal scheduling result U^* can be found by evaluating the system throughput for all the scheduling decisions (the states) in the searching space $\mathcal{U}$. Such a brute force is not feasible since the induced computation load grows exponentially w.r.t. the multiplication of the number of flows and the number of slots. The inefficiency is mainly caused by the huge searching space that includes many undesirable candidates. Therefore, to develop a practical search algorithm, the key is to find a better state in each searching step efficiently, while discarding those undesirable ones.

Since the scheduling problem has a decomposed structure in the time domain, instead of optimizing all the N flows in N time slots simultaneously, we solve the problem in an iterative manner by reducing the searching space of scheduling states from $2^{N\cdot N}$ to 2^N. In this way, each step of the scheduling problem becomes,

Problem 6. *(P6)*

$$\max_{u_{k,i}\in\{0,1\}} \sum_{i=1}^{N} \rho_{k-1,i} R_{U_{k,i}}. \tag{5.7}$$

Another advantage of this approach is that it maintains the long-term fairness among competing flows, because the dynamic weight of each flow $\rho_{k-1,i}$ depends on the scheduling decisions $U_1, U_2, \cdots, U_{k-1}$ in the past $k-1$ time slots, thus the iterative allocation structure allows us to update the weights slot by slot.

Note that the decomposed version of the scheduling problem (P6) is still non-convex and NP-hard. To improve the efficiency of the GSA, in each searching step the state should move towards a better one. For our problem, the metric to determine whether or not the scheduling state $\tilde{U_k}$ is better than U_k is the corresponding system throughputs, i.e., to compare $\sum_{i=1}^{N} R_{U_{k,i}^{\sim}}$ with $\sum_{i=1}^{N} R_{U_{k,i}}$. For fast convergence to a better state, we propose a single-flip scheme as follows. If the profit of adding flow i is greater than the degradation of throughput it causes to other pre-selected flows, this flow will be temporarily added (flip $u_{k,i}$ from zero to one); otherwise, this flow will be temporarily removed (flip $u_{k,i}$ from one to zero).

Based on this idea, below we propose the single-flip global search algorithm (S-GSA). For the k-th time slot, we first initialize the scheduling vector $U_k = \vec{0}$ and calculate the weighted fair queuing coefficients $\rho(k-1, i)$. Next, the local optimal value $u_{k,1}$ is evaluated. To this end, we let $u_{k,1}$ be 0 or 1 while keeping all the other variables $u_{k,2}, u_{k,3}, \cdots, u_{k,N}$ fixed, and choose the value of $u_{k,1}$ according to the superior weighted sum of all flows' data rates in U_k. The same criterion is used to set $u_{k,2}$ while keeping all other $u_{k,i}$ $(i \neq 2)$ fixed. All other variables $u_{k,3}, u_{k,4}, \cdots u_{k,N}$ are locally optimized in a similar manner. The above process is repeated until U_k converges (i.e., no flip of single $u_{k,i}$ will result in a higher weighted sum of flow data rates). A step by step description of the S-GSA algorithm is shown in Algorithm 1.

Algorithm 1 Single-flip GSA (S-GSA)

1: initialize the dynamic weight of each flow, $\rho_{0,i} = \frac{1}{\epsilon}\ \forall\ i$;
2: **for** $k = 1$ to N (time slot) **do**
3: update $\rho_{k-1,i} = w_i/(\sum_{l=1}^{k-1} R_{U_{l,i}} + \epsilon)^{\alpha}$ for each flow based on $U_1, U_2, \cdots, U_{k-1}$
4: initialize the vector $U_k = \vec{0}$
5: **repeat**
6: **for** $i = 1$ to N (i is the flow index) **do**
7: set $u_{k,i}$ according to $\arg\max_{u_{k,i} \in \{0,1\}} \{\rho_{k-1,i} R_{U_{k,i}} + \sum_{l=1, l \neq i}^{N} \rho_{k-1,l} R_{U_{k,l}}\}$
8: **end for**
9: **until** $U_k = [u_{k,1}\ u_{k,1}\ \cdots\ u_{k,N}]$ converges
10: **end for**

The proposed S-GSA algorithm reduces the exponential searching complexity for

each slot from 2^N to $O(T_1 N)$, where T_1 is the number of iterations for U_k to converge. The value of T_1 is typically smaller than the polynomial function of N. From the simulation results of a network with up to 40 flows, T_1 was found to be smaller than 5. Thus, the computation load of S-GSA is tractable and feasible for making realtime scheduling decisions.

Remarks: The S-GSA algorithm ensures that we cannot get a better scheduling solution by adding or removing one flow to share the time slot. However, it is possible to get a better solution by adding or removing more than one flows. Therefore, the S-GSA can only find locally optimal solution and cannot ensure the global optimality of the solution. On the other hand, for each time slot, an individual flow that is not scheduled to transmit is considered unfavorable (namely, adding it will reduce the total throughput), and the profit of adding a group of such unfavorable flows is statistically low. Similarly, each flow being scheduled to transmit in the slot is considered favorable (namely, removing it will reduce the total throughput), the profit of removing a group of such favorable flows is also low.

5.3.4 Scheduling Algorithm for Constrained Problem Using a Dual Method

In this section, we present an algorithm for the constrained optimal scheduling problem (P5). The constraints give rise to the difficulties for the scheduling problem. With the previous S-GSA algorithm, some unlucky flows cannot get the minimum throughput required, even the long-term fairness is ensured. This is because, in the rate-adaptive wireless networks the total throughput can be achieved is unknown (unlike many other scheduling or job-assignment problems) and proportional fairness among flows cannot guarantee a minimum throughput of a flow.

Using duality variables ensures that the solution locates in the feasible region (thus satisfying the constraints). However, using this approach, it is difficult if not impossible to obtain the global optimal solution of the primal problem, which can be obtained only if the number of time slots goes to infinite ($N \to \infty$) under the time-sharing condition [111] or the jointly concave condition [65]. Hence, our algorithm is not aimed at the optimal solution of the duality variables and it will terminate when the constraints are satisfied, that can significantly reduce the computational complexity. Consider the constrained optimization problem (P5), we follow the definition of Lagrangian dual [10] and take the Lagrangian function as our objective function.

The corresponding Lagrangian function is given by,

$$L(U;\lambda) = \sum_{k=1}^{N}\sum_{i=1}^{N} \rho_{k-1,i} \cdot R_{U_{k,i}} + \sum_{i=1}^{N} \lambda_i [\sum_{k=1}^{N} R_{U_{k,i}} - R^i_{\min}]. \tag{5.8}$$

where $\lambda = [\lambda_1, \lambda_2, \cdots \lambda_N]$ is the Lagrangian dual variable.

Define the dual function $g(\lambda)$ as the maximization of the Lagrangian function

$$g(\lambda) = \max_{U} L(U;\lambda), \tag{5.9}$$

and the Lagrangian dual optimization problem can be formulated as

Problem 7. *(P7)*

$$\min \quad g(\lambda) \tag{5.10}$$

$$\text{s.t.} \quad \lambda \geq 0. \tag{5.11}$$

If the primal problem (P5) is convex, then the duality gap is zero and the maximum value of the primal problem (P5) and the minimum value of the dual problem (P7) converge at the same optimal solution. If the primal problem is non-convex, the dual problem provides an upper bound, which is not always tight, for the primal solution. However, the dual problem is always convex w.r.t. λ regardless of the primal problem. This allows a direct optimization of $g(\lambda)$ by some dual update methods. The main idea here is to minimize $g(\lambda)$ by updating all components of λ along a specific direction, which can be found using the existing optimization tools.

To this end, the gradient descent method is employed (sub-gradient for the non-differentiable case), and the updating direction corresponding to λ_i is given by

$$\mathbf{d} = -(\sum_{k=1}^{N} R_{U_{k,i}} - R^i_{\min}). \tag{5.12}$$

Next, we update all λ's along their gradient descent directions with a step sequence s^l. Considering $\lambda \geq 0$, the update is performed as,

$$\lambda^{l+1} = \max(0, \lambda^l - s^l \cdot (\sum_{k=1}^{N} R_{U_{k,i}} - R^i_{\min})), \tag{5.13}$$

where s^l is a scalar sequence and is square summable [85]. Thus, we choose $s^l = 1/l$

as the update step size. Ideally, we can employ this algorithm to get the optimal solution of the dual problem and then solve the primal problem with the leverage of the dual problem. However, evaluating the dual objective function, which is to maximize the Lagrangian function to all λ, is difficult if not impossible. Nevertheless, the dual problem is still very important to ensure that the solution is feasible when we use the S-GSA algorithm for the constrained scheduling problem, since λ may further increase the weights of those starving flows. In what follows, we employ a local search algorithm to approximately evaluate the dual objective function

$$\begin{aligned} g(\lambda) &= \max_{U} L(U;\lambda) \\ &= \max_{U}\{\sum_{k=1}^{N}\sum_{i=1}^{N}\rho_{k-1,i}\cdot R_{U_{k,i}} \\ &\quad + \sum_{i=1}^{N}\lambda_i[\sum_{k=1}^{N}R_{U_{k,i}} - R^i_{\min}]\} \\ &= \sum_{k=1}^{N}\max_{U_k}\{\sum_{i=1}^{N}(\rho_{k-1,i}+\lambda_i)R_{U_{k,i}}\} - \sum_{i=1}^{N}\lambda_i R^i_{\min}. \end{aligned} \tag{5.14}$$

Since the scheduling problem has a decomposed structure, we set

$$\tilde{g}(\lambda) = \max_{U_k}\{\sum_{i=1}^{N}(\rho_{k-1,i}+\lambda_i)R_{U_{k,i}}\} \tag{5.15}$$

and the S-GSA algorithm can be modified to locally optimize $g(\lambda)$. Different from the original S-GSA algorithm, we need to use the dual update to ensure our solution is feasible, i.e. the minimum throughput requirements are satisfied. It is easy to show that λ_i increases if the constraint of flow i is violated. In (5.15), the throughput of each flow is multiplied by $(\rho_{k-1,i}+\lambda_i)$. Thus, the flows with a larger value of λ_i will get a larger weights. In the following iteration, these flows will have a better chance to get resources. A sub-optimal solution of the constrained optimization problem is then obtained by combining the S-GSA and the dual update technology, which is named Single-flip Dual-update GSA (SD-GSA)z.

The computational complexity of the SD-GSA is $O(T_2T_1N^2)$. It is higher than S-GSA due to the update of λ. The inner-loop of SD-GSA is used for evaluating $\tilde{g}(\lambda)$, similar to S-GSA, and it has the same computational complexity as S-GSA, $O(T_1N^2)$. The outer-loop is used to update the dual variables λ_i.

5.4 Performance Evaluation

5.4.1 Simulation Setting

In the simulation, we set a UWB network with random network topology. In the network, all nodes were randomly deployed in a $10 \times 10m^2$ region and the communication pairs were selected randomly, i.e., randomly selected one node from the sender set and one from the receiver set to form a communication pair. The shadowing gain from each sender to any receiver was generated randomly according to the log-normal distribution.

Let all senders use the same transmission power level for transmission. The location of each node is known, and the average signal and interference power were calculated based on the physical layer model discussed before. We tested our algorithm with different node density, by varying the active node number from 4 (2 flows) to 80 (40 flows). For each density (with the same number of flows), we repeated 1000 times using Monte Carlo simulations, and calculated the average results with different network topologies and wireless channels.

We implemented the heuristic algorithm based on the exclusive region (Ex-region) proposed in [62], since, to the best of our knowledge, it is the best algorithm so far solving the same concurrent scheduling problem. The exclusive region size was set to 2 meters, which was the optimal value calculated according to the analysis in [62]. We used the same network configurations to compare the performance of all scheduling algorithms.

Two performance metrics were considered, the total network throughput and the fairness index. To show the performance gain, network throughputs were normalized to the throughput using the TDMA scheme with two flows. Fairness was measured by the widely used Jain's fairness index which is defined as $\frac{(\sum_{i=1}^{N} R_i)^2}{N \cdot \sum_{i=1}^{N} R_i^2}$ [38].

5.4.2 Performance of Scheduling without Constraints

We first evaluate the performance of the S-GSA algorithm for the unconstrained scheduling problem (P4) with different node density and different value of α.

As shown in Figs. 5.2 and 5.3, the throughput degrades as α increases while the fairness has a reverse trend. This is because that a larger value of α will give the

starving flows larger weights, thus more time slots will be allocated to the flows with bad channel conditions. When the value of α is moderate, around 0.4 to 0.6, the network throughput using the S-GSA algorithm is much higher than that with the exclusive region based algorithm, and both algorithms are found to achieve similar fairness level. The throughput with S-GSA increases with network density, and it can outperform TDMA and the exclusive region based algorithm by 1300% and 100% with 40 flows, respectively (with $\alpha = 0.4$). The major weakness of the exclusive region based algorithm is that the procedure to select the concurrent transmission subset is not guaranteed to reach a better scheduling state. In addition, the exclusive region size is difficult to be optimized in a dynamic wireless environment.

As shown in Fig. 5.3, the fairness index with all of the scheduling algorithms including TDMA are low. This is because the received signal powers of different flows vary significantly, so their throughputs have large variation even when there is no mutual interference. On the other hand, there is always a tradeoff between throughput and fairness, and the advantage of S-GSA is that we can adjust the parameter α to flexibly make the tradeoff.

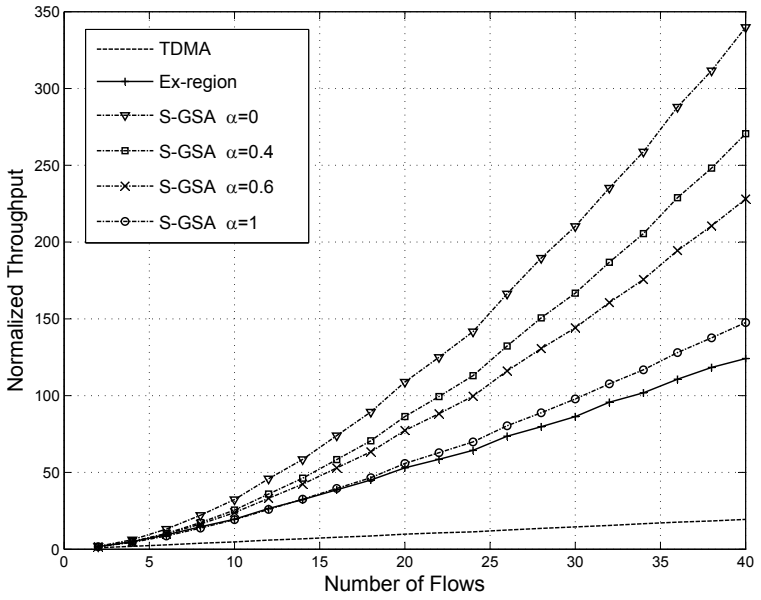


Figure 5.2: Normalized network throughput.

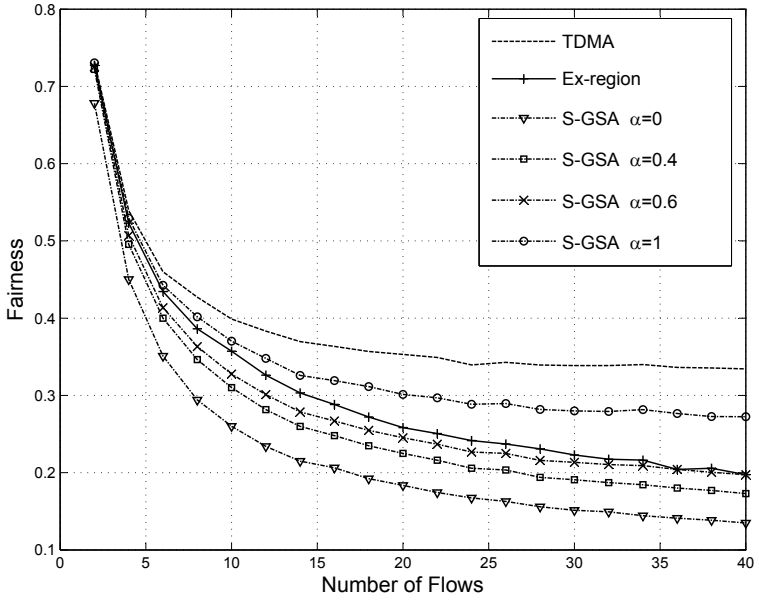


Figure 5.3: Fairness index.

5.4.3 Performance of Scheduling with Constraints

Next, the performance of SD-GSA for the constrained scheduling problem is evaluated. In SD-GSA, the dual method is employed to solve the constrained non-convex optimization problem. The Lagrangian dual variables are used to ensure that all the constraints are satisfied.

Using SD-GSA, simulation results show that the minimum throughput requirements of all flows are satisfied; however, the overall network throughput degrades slightly compared with that with S-GSA, as shown in Fig. 5.4. This is because, with SD-GSA, more time slots are allocated to the flows with small received signal power to satisfy their minimum throughput requirements, so the total network throughput is degraded.

The minimum flow throughput requirements cannot be arbitrarily set since the data rates of some flows with bad channel conditions are very low even without interference. Thus, the minimum throughput requirements should be carefully set to be feasible. In the simulation, we use 1.15 times the flow throughput achieved using the TDMA scheme as the minimal throughput requirement, so the requirement could be feasible *w.h.p*, and the resulting minimum flow throughput can be 80% higher than that with the TDMA scheme (because of the quantized time resources, thus much higher than 15%). In general, it is difficult to evaluate whether or not

the throughput constraints are feasible, since the data rate of each flow depends on the network topology and the scheduling decision. What is the capacity of the rate-adaptive wireless network is still an open issue.

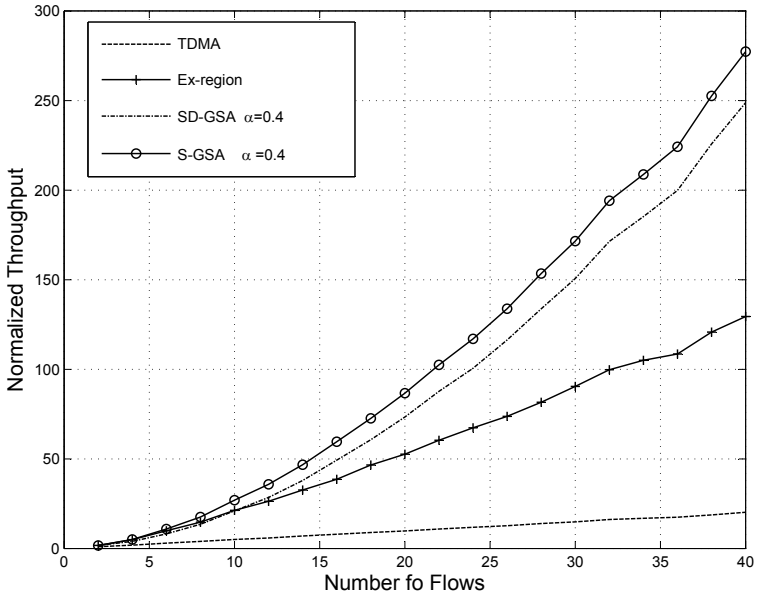


Figure 5.4: Normalized network throughput, with minimum per-flow throughput constraint.

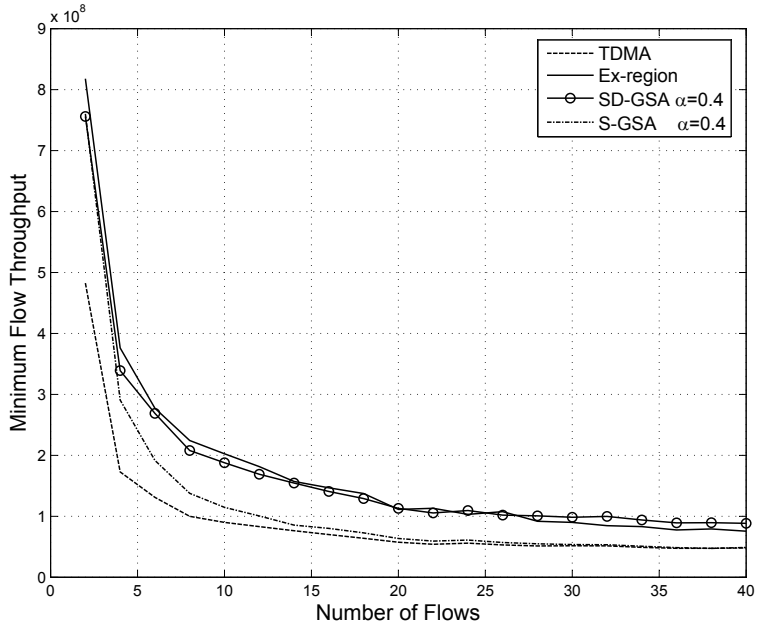


Figure 5.5: Minimal flow throughput.

Fig. 5.5 shows the minimal flow throughput among all competing flows. The results show that the minimal flow throughput with SD-GSA can be higher than that with TDMA and S-GSA, so it can achieve better max-min fairness. The improvement is not as significant as the total network throughput, since the algorithm will not allocate excessive resources for those flows with bad channel conditions. Otherwise, the total network throughput will be sacrificed.

5.5 Conclusion

In this chapter, we have investigated the scheduling of concurrent transmission problem for rate-adaptive wireless networks. For the unconstrained and constrained optimal scheduling problems, we have proposed S-GSA and SD-GSA algorithms. Extensive simulation results have shown that the proposed algorithms can significantly out-perform the previous heuristic algorithm, while the computation cost of the proposed algorithms is low and tolerable.

5.6 Symbol List

Symbol	Explanation
$R_(U_{k,i})$	throughput of flow i in slot k
$(u_{k,i})$	schedule result of user i in slot k
d	transceiver distance
$\rho_{k-1,i}$	fairness control parameter
w_i	weight of flow i
α	parameter to make the tradeoff between fairness and throughput i
ϵ	parameter used to prevent zero denominator in fairness calculation
$L(U;0)$	Lagrangian function
λ	Lagrangian dual variable
$g(\lambda)$	maximization of the Lagrangian function
$\mathbf{d}$	sub-gradient

Chapter 6

Conclusions and Future work

In this dissertation, we have investigated three important and fundamental problems in wireless communication/network systems: (1) digital modulation schemes including network modulation, which takes the network topology into account, and hexagonal modulation that utilizes the hexagon based constellation mapping, (2) environment-aware clock skew estimation and dynamic compensation, which takes the environment factor into account and improve synchronization accuracy while reduce communication overhead, and (3) practical concurrent communication scheduling in broad band wireless communication systems. The following outlines the contributions achieved and the possible future research directions.

6.1 Network Modulation

We have proposed the NM and EAR schemes that take the network topology into account and can explore the spatial diversity and the broadcast nature of wireless networks. The NM scheme can superimpose bits with different SNR requirements on a single transmitted symbol and thus it can naturally support two users simultaneously without introducing new resources. Besides, this property can be used in relay communication, which is also motivated by the spatial diversity of wireless systems. The EAR scheme can further enhance the performance of relay communication, which leverages the useful information contained in the corrupted packets and reduces the amount of information being relayed. In addition, as the NM scheme is based on the bit mapping/re-mapping of the conventional QAM schemes, it does not require specialized communication hardware and is ready to be implemented. The effectiveness

is verified by extensive simulation results.

It is concluded that topology-aware PHY layer design will be a promising direction with many open issues for further study. First, in this dissertation, we group users mainly considering the estimated channel conditions. For heterogeneous applications, the scheduling of users should consider the different utility functions for different applications. Second, the admission region of single-hop and multi-hop wireless networks with NM and EAR needs to be redefined in the new setting. Third, the interactions of NM, EAR and the protocols of other layers need further investigation. Fourth, in this work, the receiver demodulates and decodes bits from one transmission. It is possible to combine various user cooperative diversity techniques with NM/EAR to further improve the system performance by fully utilizing the spatial diversity gain. Finally, many new cross-layer optimization problems can be formulated, e.g., jointly optimize NM/EAR with relay selection/scheduling and routing, as the topology-aware PHY layer design provides one more dimension of freedom to improve the network performance.

6.2 Hexagonal Modulation and Non-binary Error Control Coding

We have proposed to adopt the hexagonal modulation and multiplex bits and trits for wireless systems. We have also introduced the ternay error control coding to provide the error protection for trits. The H-QAM based AMC schemes can achieve much higher overall system throughput. Extensive simulation results have verified the efficacy of the proposed schemes. Moreover, we have also implemented a prototype system that demonstrates the feasibility of the proposed schemes.

In this dissertation, we examine four hexagonal modulation schemes as examples, but others are possible and the design should depend on the particular application. For instance, H-QAM with denser constellation points may be desirable for higher SNR scenarios. Second, it is possible to improve the constellation mapping to further reduce the BER, and the non-binary trellis coded modulation techniques can be a promising extension. In addition, we employed a ternary convolutional code with code rates of 1/2 and 3/4 (the latter obtained by puncturing). The design of ternary error correcting codes has received very little attention in the literature, so this is a promising area for further investigation.

6.3 Environment-aware Clock Skew Estimation and Synchronization

We have investigated the estimation and compensation of clock skew to improve the clock synchronization performance and proposed to utilize the correlation between clock skew and temperature to accurately estimate the clock skew. Based on the clock skew v.s. temperature estimation results, we have proposed a two-phase clock synchronization scheme, the environment-aware external clock synchronization phase and clock self-calibration phase. The proposed scheme utilizes the temperature information to estimate clock skew and then removes the effect of clock skew directly. It can improve the clock synchronization accuracy and reduce the computational complexity and communication overhead, which is verified by simulation and experimentation with a Mica2 sensor nodes testbed.

It is a promising research direction to take the environment factors into account to improve the performance of communication/network system. How to further improve the accuracy of the clock skew especially with an online estimation remains an open issue. Meanwhile, more research work about the uncertainty analysis of TACSC scheme under imperfect parabolic parameter estimation is beckoned since it can give the error bound of TACSC and thus the re-synchronization can be triggered appropriately according to the synchronization requirement. Besides the temperature, further research is needed to investigate the impact of other environment factors (such as the power supply voltage, humidity, electromagnetic interference, and vibration) on clock skew, which are treated as noise in this work as their influence is typically less than that of the temperature.

6.4 Scheduling for Concurrent Transmission

In this dissertation work, we have investigated the concurrent transmission scheduling problem in broad band network. It allows the transmission of multiple flows within collision regions. We have proposed simple yet effective algorithms to conduct the scheduling efficiently and significantly improve the network throughput compared with the existing solutions.

Although the proposed algorithms can achieve a much higher network throughput; however, how to calculate the network throughput with concurrent transmission

and rate adaptation still remains an open issue. Besides, the concurrent scheduling problem in fading environment will also be an interesting research problem. Given the popularity of rate-adaptive PHY layer technologies in wireless systems, more research work is needed for this difficult and important problem.

Bibliography

[1] A. Agatep. Reed-Solomon solutions with spartan-II FPGAs. *White Paper: Spartan-II Family*, 2000.

[2] R. Ahlswede, N. Cai, S.-Y.R. Li, and R.W. Yeung. Network information flow. *IEEE Transactions on Information Theory*, 46(4):1204 –1216, Jul 2000.

[3] I.F. Akyildiz, X. Wang, and W. Wang. Wireless mesh networks: a survey. *Computer Networks*, 47(4):445–487, 2005.

[4] M. Alicherry, R. Bhatia, and L. Li. Joint channel assignment and routing for throughput optimization in multi-radio wireless mesh networks. In *Proc. ACM Mobicom'05*, pages 58–72, 2005.

[5] L. Auler and R. d'Amore. Adaptive Kalman Filter for Time Synchronization over Packet-Switched Networks: An Heuristic Approach. In *IEEE COMSWARE'07*, pages 1–7, 2007.

[6] F. Aune. Cross-layer design tutorial. *Norwegian University of Science and Technology, Dept. of Electronics and Telecommunications*, 2004.

[7] A. Bletsas. Evaluation of Kalman filtering for network time keeping. *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, 52(9):1452–1460, 2005.

[8] H. Blom and Y. Bar-Shalom. The interacting multiple model algorithm for systems with Markovianswitching coefficients. *IEEE Transactions on Automatic Control*, 33(8):780–783, 1988.

[9] E. Blossom. Gnu radio: tools for exploring the radio frequency spectrum. *Linux journal*, 2004(122):4, 2004.

[10] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[11] M.F.U. Butt, R. Zhang, S.X. Ng, and L. Hanzo. Superposition coding aided bi-directional relay transmission employing iteratively decoded self-concatenated convolutional codes. In *IEEE VTC 2010-Spring*, pages 1–5, May 2010.

[12] L. Cai, Y. Luo, S. Xiang, and J. Pan. Scalable modulation for scalable wireless videocast. In *IEEE INFOCOM'10, mini-symposium*, San Diego, CA, USA, Mar. 2010.

[13] R. Cendrillon, W. Yu, M. Moonen, J. Verlinden, and T. Bostoen. Optimal multi-user spectrum management for digital subscriber lines. *IEEE Transactions on Communications*, 54(5):922–933, 2006.

[14] K. Cheun and W.E. Stark. Optimal selection of Reed-Solomon code rate and the number of frequency slots in asynchronous FHSS-MA networks. *IEEE Transactions on Communications*, 41(2):307–311, 2002.

[15] Cisco. Cisco visual networking index: Global mobile data traffic forecast update, 2011-2016, white paper, 2012.

[16] T. Cover. Broadcast channels. *IEEE Transactions on Information Theory*, 18(1):2–14, 1972.

[17] R.L. Cruz and A.V. Santhanam. Optimal routing, link scheduling and power control in multihop wireless networks. In *IEEE INFOCOM 2003.*, volume 1, pages 702–711, March/April 2003.

[18] A. Demers, S. Keshav, and S. Shenker. Analysis and simulation of a fair queueing algorithm. *Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 1–12, 1989.

[19] I.B. Djordjevic and B. Vasic. Nonbinary LDPC codes for optical communication systems. *IEEE Photonics Tech. Lett.*, 17(10):2224–26, Oct. 2005.

[20] M. Elaoud and P. Ramanathan. Adaptive use of error-correcting codes for real-time communication in wireless networks. In *IEEE INFOCOM'98*, volume 2, pages 548–555, 1998.

[21] J. Elson, L. Girod, and D. Estrin. Fine-grained network time synchronization using reference broadcasts. *ACM SIGOPS Operating Systems Review*, 36(SI):147–163, 2002.

[22] K. Engdahl and K.S. Zigangirov. Comparison analysis of the hexagonal multilevel QAM and rectangular multilevel QAM. *Problems of Information Transmission C/C of Problemy Peredachi Informatsii*, 34:322–331, 1998.

[23] M. Ettus. Ettus research, LLC. *Online information on USRP board, available at* `http://www.ettus.com`, 2008.

[24] Jr. Forney, G., R. Gallager, G. Lang, F. Longstaff, and S. Qureshi. Efficient modulation for band-limited channels. *IEEE Journal on Selected Areas in Communications*, 2(5):632 – 647, Sep. 1984.

[25] H. N. Gabow. An efficient implementation of edmonds' algorithm for maximum matching on graphs. *J. ACM.*, 23(2):221–234, April 1976.

[26] S. Ganeriwal, R. Kumar, and M.B. Srivastava. Timing-sync protocol for sensor networks. In *ACM Sensys'03*, 2003.

[27] M. Gast. *802.11 wireless networks: the definitive guide*. O'Reilly Media, 2005.

[28] B. Grünbaum and G.C. Shephard. *Tilings and patterns*. WH Freeman & Co., 1986.

[29] P. Gupta and P.R. Kumar. The capacity of wireless networks. *IEEE Transactions on Information Theory*, 46(2):388–404, 2000.

[30] B.R. Hamilton, X. Ma, Q. Zhao, and J. Xu. ACES: adaptive clock estimation and synchronization using Kalman filtering. In *ACM MobiCom'08*, pages 152–162. ACM New York, NY, USA, 2008.

[31] S. Han, J.M. Cioffi, and J. Lee. Tone injection with hexagonal constellation for peak-to-average power ratio reduction in OFDM. *IEEE Commun. Lett.*, 10(9):646–648, Sept. 2006.

[32] S.H. Han, J.M. Cioffi, and J.H. Lee. On the use of hexagonal constellation for peak-to-average power ratio reduction of an ODFM signal. *IEEE Transactions on Wireless Communications*, 7(3):781–786, 2008.

[33] B. Hayes. Third base. *American Scientist*, 89(6):490, Nov./Dec. 2001.

[34] Epson Electroincs Inc. SE2412CT-ND data sheet. Available at `http://www.eea.epson.com/portal/pls/portal/docs/1/1407484.PDF`.

[35] Seiko Instruments Inc. Quartz Crystal Product Catalogue. Available at `http://www.sii.co.jp/compo/catalog/crystal_en.pdf`.

[36] Seiko Instruments Inc. Quartz Crystal Product Catalogue. Available at `http://www.spezial.pl/Crystal_Products200810_E.pdf`.

[37] N.K. Jablon. Joint blind equalization, carrier recovery and timing recovery for high-order QAM signal constellations. *IEEE Trans. on Signal Processing*, 40(6):1383–98, 1992.

[38] R. Jain, A. Durresi, and G. Babic. Throughput fairness index: An explanation. In *ATM Forum Contribution 99*, volume 45. Citeseer, 1999.

[39] K. Jamieson and H. Balakrishnan. PPR: Partial packet recovery for wireless networks. In *ACM SIGCOMM'07*, Kyoto, Japan, August 2007.

[40] Y. Jin, D.G. Le, and G.W. Bai. Kalman filter-based adaptive link-layer FEC mechanism for wireless sensor networks. *Application Research of Computers*, 27(4):1412–1415, 2010.

[41] D. Katabi, H. Rahul, and S. Jakubczak. SoftCast: One video to serve all wireless receivers. Technical Report MIT-CSAIL-TR-2009-005, CSAIL, MIT, Feb. 2009.

[42] S. Katti, D. Katabi, H. Balakrishnan, and M. Medard. Symbol-Level Network Coding for Wireless Mesh Networks. In *ACM SIGCOMM'08*, Seattle, WA, August 2008.

[43] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Médard, and J. Crowcroft. XORs in the air: practical wireless network coding. *IEEE/ACM Transactions on Networking (TON)*, 16(3):497–510, 2008.

[44] A.E. Khandani, J. Abounadi, E. Modiano, and L. Zheng. Cooperative routing in static wireless networks. *IEEE Transactions on Communications*, 55(11):2185–2192, 2007.

[45] H. Khlifi and J.C. Gregoire. Estimation and removal of clock skew from delay measures. In *IEEE LCN'04*, 2004.

[46] S. Kim, R. Fonseca, and D. Culler. Reliable transfer on wireless sensor networks. In *IEEE SECON'04*, pages 449–459, 2004.

[47] D.E. Knuth. *The Art of Computer Programming - Volume 2: Seminumerical Algorithms*. Addison-Wesley, 1980.

[48] T. Koike and S. Yoshida. Space-time Trellis-coded ternary PSK for mobile communications. *Electronics Letters*, 40(16):1011–1012, 2004.

[49] S.C. Krishnan, R. Panigrahy, and S. Parthasarathy. Error-correcting codes for ternary content addressable memories. *IEEE Trans. on Computers*, 58(2):275–279, 2009.

[50] D. Krueger, JR Cruz, and R. He. A ternary coded magnetic recording system. In *IEEE ICC'95*, volume 1, pages 638–641, 1995.

[51] JN Laneman, DNC Tse, and GW Wornell. Cooperative diversity in wireless networks: Efficient protocols and outage behavior. *IEEE Transactions on Information Theory*, 50(12):3062–3080, 2004.

[52] J.W. Lee, R.R. Mazumdar, and N.B. Shroff. Non-convex optimization and rate control for multi-class services in the Internet. *IEEE/ACM Transactions on Networking*, 13(4):827–840, 2005.

[53] M. Leng and Y.C. Wu. On clock synchronization algorithms for wireless sensor networks under unknown delay. *IEEE Transactions on Vehicular Technology*, 59(1):182–190, 2010.

[54] C. Li, G. Yue, X. Wang, and M. Khojastepour. Ldpc code design for half-duplex cooperative relay. *Wireless Communications, IEEE Transactions on*, 7(11):4558–4567, 2008.

[55] X.R. Li and V.P. Jilkov. Survey of maneuvering target tracking. Part V: Multiple-model methods. In *Conference on Signal and Data Processing of Small Targets*, volume 41, pages 559–581, 2005.

[56] X.Y. Li, A. Nusairat, Y. Wu, Y. Qi, J.Z. Zhao, X. Chu, and Y. Liu. Joint Throughput Optimization for Wireless Mesh Networks. *IEEE Transactions on Mobile Computing*, 8(7):895–909, 2009.

[57] Z. Li, W. Chen, C. Li, M. Li, X. Li, and Y. Liu. Flight: Clock calibration using fluorescent lighting. *ACM Mobicom'12*, 2012.

[58] C. Liang, W. Lenan, and R. Piche. Posterior Cramer-Rao lower bound for mobile tracking in mixed LOS/NLOS conditions. In *EUSIPCO'09*, 2009.

[59] K.C.J. Lin, N. Kushman, and D. Katabi. ZipTx: Harnessing partial packets in 802.11 networks. In *ACM MobiCom'08*, pages 351–362, 2008.

[60] S. Lin and D.J. Costello. *Error control coding*. Prentice-Hall Englewood Cliffs, NJ, 1983.

[61] X. Lin and N. Shroff. Joint rate control and scheduling in multihop wireless networks. In *Decision and Control, 2004. CDC. 43rd IEEE Conference on*, volume 2, pages 1484–1489, Dec. 2004.

[62] K.H. Liu, L. Cai, and X. Shen. Exclusive-region based scheduling algorithms for UWB WPAN. *IEEE Transactions on Wireless Communications*, 7(3):933–942, 2008.

[63] K.H. Liu, L. Cai, and X.S. Shen. Multi-class utility-based scheduling for UWB networks. *IEEE Transactions on Vehicular Technology*, 57(2):1176–1187, 2008.

[64] Y. Luo and L. Cai. Throughput maximization for user cooperative wireless systems with adaptive modulation. In *IEEE ICC'10*, 2010.

[65] Z. Luo and S. Zhang. Dynamic spectrum management: Complexity and duality. *IEEE Journal of Selected Topics in Signal Processing*, 2(1):57, 2008.

[66] V. Mahinthan, L. Cai, J.W. Mark, and X. Shen. Maximizing cooperative diversity energy gain for wireless networks. *IEEE Transactions on Wireless Communications*, 6(7):2530–2539, June 2007.

[67] V. Mahinthan, L. Cai, J.W. Mark, and X. Shen. Partner selection based on optimal power allocation in cooperative-diversity systems. *IEEE Transactions on Vehicular Technology*, 57(1):511–520, 2008.

[68] M. Maroti, B. Kusy, G. Simon, and Á. Lédeczi. The flooding time synchronization protocol. In *ACM Sensys'04*, pages 39–49. ACM New York, NY, USA, 2004.

[69] MAXIM. 32.768kHz Temperature-Compensated Crystal Oscillator. Datasheet available at `http://datasheets.maxim-ic.com/en/ds/DS32kHz.pdf`.

[70] D.L. Mills. Internet time synchronization: The network time protocol. *IEEE Transactions on Communications*, 39(10):1482–1493, 1991.

[71] SB Moon, P. Skelly, and D. Towsley. Estimation and removal of clock skew from network delaymeasurements. In *IEEE INFOCOM'99*, volume 1, 1999.

[72] S.J. Murdoch. Hot or not: Revealing hidden services by their clock skew. In *ACM CCS'06*, pages 27–36. ACM New York, NY, USA, 2006.

[73] C.D. Murphy. High-order optimum hexagonal constellations. In *IEEE PIMRC'00.*, volume 1, pages 143–146. IEEE, 2000.

[74] K.L. Noh, Q.M. Chaudhari, E. Serpedin, and B.W. Suter. Novel clock phase offset and skew estimation using two-way timing message exchanges for wireless sensor networks. *IEEE Transactions on Communications*, 55(4):766–777, 2007.

[75] B. O'hara and A. Petrick. *IEEE 802.11 handbook: a designer's companion*. Standards Information Network, 2005.

[76] V. Pejovic. OFDM Example. *Available at* ***http://www.cs.ucsb.edu/~veljko/downloads/ofdm_example.tar.gz***, 2011.

[77] D. Pisinger. Where are the hard knapsack problems? *Computers and Operations Research*, 32(9):2271–2284, 2005.

[78] P. Popovski and E. De Carvalho. Improving the rates in wireless relay systems through superposition coding. *IEEE Transactions on Wireless Communications*, 7(12):4831–4836, 2008.

[79] B. Radunovic and J.Y. Le Boudec. Optimal power control, scheduling, and routing in UWB networks. *IEEE Journal on Selected Areas in Communications*, 22(7):1252–1270, 2004.

[80] A. Rajeswaran, G. Kim, and R. Negi. A scheduling framework for UWB & cellular networks. *Mobile Networks and Applications*, 11(1):9–20, 2006.

[81] T. Richardson. Error floors of LDPC codes. In *Proceedings of the annual Allerton conference on communication control and computing*, pages 1426–1435, 2003.

[82] T. Schmid, Z. Charbiwala, R. Shea, and M.B. Srivastava. Temperature compensated time synchronization. *Embedded Systems Letters, IEEE*, 1(2):37–41, 2009.

[83] S. Sen, S. Gilani, S. Srinath, S. Schmitt, and S. Banerjee. Design and implementation of an approximate communication system for wireless media applications. *ACM SIGCOMM Computer Communication Review*, 40(4):15–26, 2010.

[84] S. Sharma, Y. Shi, Y.T. Hou, H.D. Sherali, and S. Kompella. Cooperative communications in multi-hop wireless networks: Joint flow routing and relay node assignment. In *IEEE INFOCOM'10*, 2010.

[85] NZ Shor, K.C. Kiwiel, and A. Ruszcayski. *Minimization methods for non-differentiable functions*. Springer-Verlag New York, Inc. New York, NY, USA, 1985.

[86] B. Sklar. *Digital communications*, volume 2. Prentice Hall, 2001.

[87] R. Sugihara and R. Gupta. Clock synchronization with deterministic accuracy guarantee. *Wireless Sensor Networks*, pages 130–146, 2011.

[88] M. Tanahashi and H. Ochiai. A multilevel coded modulation approach for hexagonal signal constellation. *IEEE Trans. Wireless Commun.*, 8(10):4993–97, Oct. 2009.

[89] Crossbow Technology. MICA2 data sheet available at *http://www.xbow. .com/Products/Product_pdf_files/Wireless_pdf /MICA2_Datasheet.pdf*.

[90] Crossbow Technology. MPR-MIB Series Users Manual. Available at `http://www.xbow.com/Support/Support_pdf_files/MPR-MIB_Series_Users_Manual.pdf`.

[91] Crossbow Technology. MTS-MDA Sensor Board Users Manual. Available at `http://www.xbow.com/Support/Support_pdf_files/MTS-MDA_Series_Users_Manual.pdf`.

[92] P. Tichavsky, C. Muravchik, and A. Nehorai. Posterior cramér-rao bounds for discrete-time nonlinear filtering. *IEEE Transactions on Signal Processing*, 46(5):1386–1396, 1998.

[93] T. Tran, T. Nguyen, and B. Bose. A joint network-channel coding technique for single-hop wireless networks. In *IEEE NetCod'08*, pages 1–6, 2008.

[94] M.B. Uddin and C. Castelluccia. Toward clock skew based wireless sensor node services. In *ICST WICON'10*, 2010.

[95] R. Urbanke. On multiple-access communication. *PhD thesis, Washington University, St. Louis*, August 1995.

[96] EN 300 421 v1.1.2. Digital Video Broadcasting (DVB), Framing structure, channel coding and modulation for 11/12 GHz satellite services. ETSI Standard, Aug. 1997.

[97] D. Veitch, S. Babu, and A. Pàsztor. Robust synchronization of software clocks across the internet. In *ACM SIGCOMM'04*, pages 219–232. ACM New York, NY, USA, 2004.

[98] M.C. Vuran and I. F. Akyildiz. Cross-layer packet size optimization for wireless terrestrial, underwater, and underground sensor networks. In *IEEE INFOCOM'08*, pages 226–230, 2008.

[99] S. Wang and B. K. Yi. Optimizing enhanced hierarchical modulations. In *IEEE Globecom'08*, Nov./Dec. 2008.

[100] Greg Welch and Gary Bishop. An introduction to the kalman filter, 1995.

[101] Y.C. Wu, Q. Chaudhari, and E. Serpedin. Clock synchronization of wireless sensor networks. *IEEE Signal Processing Magazine*, 28(1):124–138, 2011.

[102] C.Q. Yang, E. Hossain, and V.K. Bhargava. On adaptive hybrid error control in wireless networks using Reed-Solomon codes. *IEEE Transactions on Wireless Communications*, 4(3):835–840, 2005.

[103] L. Yang and G.B. Giannakis. Ultra-wideband communications: an idea whose time has come. *IEEE Signal Processing Magazine*, 21(6):26–54, 2004.

[104] Z. Yang, L. Cai, A. Gulliver, and J. Pan. Beyond powers of two: Non-binary coding and hexagonal modulation for wireless communications. *Submitted to IEEE Trans. on Wireless Communication, 2013.*

[105] Z. Yang, L. Cai, Y. Liu, and J. Pan. Environment-aware clock skew estimation and synchronization for wireless sensor networks. In *IEEE INFOCOM'12*, 2012.

[106] Z. Yang, L. Cai, and W. Lu. Practical scheduling algorithms for concurrent transmissions in rate-adaptive wireless networks. In *IEEE INFOCOM'10*, 2010.

[107] Z. Yang, Lin Cai, Yuanqian Luo, and Jianping Pan. Topology-aware modulation and error-correction coding for cooperative networks. *IEEE Journal on Selected Areas in Communications*, 30(2):379–387, 2012.

[108] Z. Yang, L. He, L. Cai, and J. Pan. Temperature-assisted clock synchronization and self-calibration for sensor networks. *Submitted to IEEE Trans. on Wireless Communication, 2013.*

[109] Z. Yang, Y. Luo, and L. Cai. Network modulation: A new dimension to enhance wireless network performance. In *IEEE INFOCOM'11*, April 2011.

[110] Z. Yang, J. Pan, and L. Cai. Adaptive clock skew estimation with interactive multi-model Kalman filters for sensor networks. In *IEEE ICC'10*, 2010.

[111] W. Yu and R. Lui. Dual methods for nonconvex spectrum optimization of multicarrier systems. *IEEE Transactions on Communications*, 54(7):1310, 2006.

[112] J. Yuan, Z. Li, W. Yu, and B. Li. A cross-layer optimization framework for multihop multicast in wireless mesh networks. *IEEE Journal on Selected Areas in Communications*, 24(11):2092, 2006.

[113] L. Zhang, Z. Liu, and H. Xia. Clock synchronization algorithms for network measurements. In *IEEE INFOCOM'02*, volume 1, 2002.

[114] S. Zhang, S. Liew, and P. Lam. Hot topic: physical-layer network coding. In *ACM MobiCom'06*, pages 358–365, New York, NY, USA, 2006.

[115] Z. Zhong, P. Chen, and T. He. On-demand time synchronization with predictable accuracy. In *IEEE INFOCOM'11*, 2011.

Appendix A

PCRLB for AMKF

A.1 PCRLB for AMKF

Here we provide a brief outline for the derivation of PCRLB of AMKF as the estimation performance bound. The clock skew estimation error bound can not only offer a performance benchmark but also determine the synchronization period for a given synchronization accuracy and provide important insights for clock synchronization protocol design [105].

With an appropriate filter parameter setting, the estimation covariance matrix from the Kalman filter can be used to indicate the estimation accuracy and convergence, which is bounded by the Posterior Cramer-Rao Lower Bound (PCRLB) [92]. In this section, we derive the PCRLB for the clock skew estimation and clock skew fusion, which can be used as a benchmark for performance evaluation.

Denote the state to be estimated and the measurement at time instance k as x_k ($\in R^n$) and z_k ($\in R^m$), respectively. The posterior Fisher information matrix (FIM) for X_k is defined as

$$J\left(X_k\right) \triangleq -E\left[\frac{\partial^2 \log p\left(X_k, Z_k\right)}{\partial X_k \left(\partial X_k\right)'}\right] = -E\left[\frac{\partial^2 \log p\left(X_k | Z_k\right)}{\partial X_k \left(\partial X_k\right)'}\right]. \tag{A.1}$$

where $X_k = [x_1^T, x_2^T \cdots x_k^T]^T$, $Z^k = [z_1^T, z_2^T \cdots z_k^T]^T$, and $p(X_k, Z_k)$ are the joint probability density function (PDF) of X_k and Z_k, respectively. Denote by $\hat{X}_k$ the estimation of state vector X_k. The estimate covariance is bounded by the PCRLB as

$$P_{X_k} = E[(\hat{X}_k - X_k)(\hat{X}_k - X_k)'] \geq \Phi_k \triangleq J_k^{-1}\left(X_k\right), \tag{A.2}$$

$$\Phi_k = \begin{bmatrix} J_k^{11} - (J_k^{12} J_k^{22^{-1}} J_k^{21})^{-1} & -(J_k^{11} - (J_k^{12} J_k^{22^{-1}} J_k^{21})^{-1}) J_k^{12} J_k^{22^{-1}} \\ -J_k^{22^{-1}} J_k^{21} (J_k^{11} - (J_k^{12} J_k^{22^{-1}} J_k^{21})^{-1}) & (J_k^{22} - (J_k^{21})^T (J_k^{11})^{-1} J_k^{12})^{-1} \end{bmatrix}. \tag{A.4}$$

with the equality holds when $\hat{X}_k = E(X_k|Z_k)$. The difference of the two matrices $P_{X_k} - \Phi_k$ is a positive semidefinite matrix. Since $X_k = [(X_{k-1})^T, x_k^T]^T$, Tichavsky et al. showed in[92] that the posterior FIM $J(X_k)$ can be decomposed into block matrices as

$$J(X_k) = \begin{bmatrix} J_{X_{k-1},X_{k-1}} & J_{X_{k-1},x_k} \\ J_{X_{k-1},x_k}^T & J_{x_k,x_k} \end{bmatrix} = \begin{bmatrix} J_k^{11} & J_k^{12} \\ J_k^{21} & J_k^{22} \end{bmatrix}. \tag{A.3}$$

Applying the blockwise inversion to $J(X_k)$ above, Φ_k can be expressed as (A.4). Similarly, P_{X_k} can also be decomposed into block matrices

$$P_{X_k} = E \begin{bmatrix} P_{X_{k-1}} & (\hat{X}_{k-1} - X_{k-1})(\hat{x}_k - x_k)^T \\ (\hat{X}_{k-1} - X_{k-1})(\hat{x}_k - x_k^T)^T & P_{x_k} \end{bmatrix}, \tag{A.5}$$

where $E[P_{x_k}]$ is the estimated covariance of x_k. As $P_{X_k} - \Phi_k$ is positive semidefinite, the corresponding PCRLB is

$$P_{x_k} = E[(\hat{x}_k - x_k)(\hat{x}_k - x_k)^T] \geq \phi_k, \tag{A.6}$$

where ϕ_k is the lower-right corner block of Φ_k and $\phi_k = (J_k^{22} - (J_k^{12})^T (J_k^{11})^{-1} J_k^{12})^{-1}$. Thus, the corresponding posterior FIM for x_k is

$$J(x_k) = J_k^{22} - (J_k^{12})^T (J_k^{11})^{-1} J_k^{12}. \tag{A.7}$$

It is not easy to evaluate the equation above. However, given the fact that

$$p(X_k, Z_k) = p(z_k|x_k) p(x_k|x_{k-1}) p(X_{k-1}, Z_{k-1}), \tag{A.8}$$

let $X^k = [(X_{k-2})^T, x_{k-1}^T, x_k^T]^T$, we have,

$$\begin{aligned} J(X_k) &= J([(X_{k-2})^T, x_{k-1}^T, x_k^T]^T) \\ &= \begin{bmatrix} J_{k-1}^{11} & J_{k-1}^{12} & 0 \\ (J_{k-1}^{12})^T & J_{k-1}^{22} + D_{k-1}^{11} & D_{k-1}^{12} \\ 0 & (D_{k-1}^{12})^T & D_{k-1}^{22} \end{bmatrix}, \end{aligned} \tag{A.9}$$

where

$$D_{k-1}^{11} = -\frac{\partial^2 \log p(x_k|x_{k-1})}{\partial x_{k-1}\partial(x_{k-1})'},$$
$$D_{k-1}^{12} = -\frac{\partial^2 \log p(x_k|x_{k-1})}{\partial x_{k-1}\partial(x_k)'},$$
$$D_{k-1}^{22} = -\frac{\partial^2 \log p(x_k|x_{k-1})}{\partial x_k\partial(x_k)'}.$$

Similar to (A.4)-(A.6), we can express $J(x_k)$ based on (A.9) as

$$J(x_k) = D_{k-1}^{22} - (D_{k-1}^{12})^T(J(x_{k-1}) + D_{k-1}^{11})^{-1}D_{k-1}^{12}, \tag{A.10}$$

where $J(x_{k-1}) = J_{k-1}^{22} - (J_{k-1}^{12})^T(J_{k-1}^{11})^{-1}J_{k-1}^{12}$.

Then $J(x_k)$ can be calculated recursively, since it is not difficult to evaluate D_{k-1}^{11}, D_{k-1}^{12} and D_{k-1}^{22} in linear Gaussian systems. Given a general linear Gaussian system,

$$\begin{aligned} x_{k+1} &= Ax_x + v_k, \\ z_k &= Hx_k + \omega_k, \end{aligned} \tag{A.11}$$

where A is the state transition matrix, H is the observation matrix, v_k and ω_k are the process noise and observation noise respectively, and the corresponding covariance matrices are Q and R. Then, we can rewrite D_{k-1}^{11}, D_{k-1}^{12} and D_{k-1}^{22} as

$$\begin{aligned} D_{k-1}^{11} &= A^TQ^{-1}A, \\ D_{k-1}^{12} &= -A^TQ^{-1}, \\ D_{k-1}^{22} &= Q^{-1} + H^TR^{-1}H. \end{aligned} \tag{A.12}$$

Substituting (A.12) into (A.10), we have

$$\begin{aligned} J(x_k) = Q^{-1} + HR^{-1}H^T - \\ Q^{-T}A(J(x_{k-1}) + A^TQ^{-1}A)^{-1}A^TQ^{-1}. \end{aligned} \tag{A.13}$$

We can further simplify it by applying the matrix inversion lemma [58], and we obtain

$$J(x_k) = (Q + AJ(x_{k-1})^{-1}A^T)^{-1} + HR^{-1}H^T. \tag{A.14}$$

Then we can evaluate $J(x_k)$ recursively.