Resource Management in E-health Systems

by

Qinghua Shen

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2015

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

E-health systems are the information and communication systems deployed to improve quality and efficiency of public health services. Within E-health systems, wearable sensors are deployed to monitor physiology information not only in hospitals, but also in our daily lives under all types of activities; wireless body area networks (WBANs) are adopted to transmit physiology information to smartphones; and cloud servers are utilized for timely diagnose and disease treatment. The integrated services provided by E-health systems could be more convenient, reliable, patient centric and bring more economic healthcare services.

Despite of many benefits, e-health systems face challenges among which resource management is the most important one as wearable sensors are energy and computing capability limited, and medical information has stringent quality of service (QoS) requirements in terms of delay and reliability. This thesis presents resource management mechanisms, including transmission power allocation schemes for wearable sensors, Medium Access Control (MAC) for WBANs, and resource sharing schemes among cloud networks, that can efficiently exploit the limited resources to achieve satisfactory QoS.

First, we address how wearable sensors could energy efficiently transmit medical information with stringent QoS requirements to a smart phone. We first investigate how to provide worst-case delay provisioning for vital physiology information. Sleep scheduling and opportunistic channel access are exploited to reduce energy consumption in idle listening and increase energy efficiency. Considering dynamic programming suffers from curse of dimensionality, Lyapunov optimization formulation is established to derive a low complexity two-step transmission power allocation algorithm. We analyze the conditions under which the proposed algorithm could guarantee worst-case delay. We then investigate the impacts of peak power constraint and statistical QoS provisioning. An optimal transmission power allocation scheme under a peak power constraint is derived, and followed by an efficient calculation method. Applying duality gap analysis, we characterize the upper bound of the extra average transmission power incurred due a peak power constraint. We demonstrate that when the peak power constraint is stringent, the proposed constant power scheme is suitable for wearable sensors for its performance is close to optimal. Further, we show that the peak power constraint is the bottleneck for wearable sensors to provide stringent statistical QoS provisioning.

Second, WBANs can provide low-cost and timely healthcare services and are expected to be widely adopted in hospitals. We develop a centralized MAC layer resource management scheme for WBANs, with a focus on inter-WBAN interference mitigation and sensor

power consumption reduction. Based on the channel state and buffer state information reported by smart phones deployed in each WBAN, channel access allocation is performed by a central controller to maximize the network throughput. Note that sensors have insufficient energy and computing capability to timely provide all the necessary information for channel resource management, which deteriorates the network performance. We exploit the temporal correlation of body area channel such that channel state reports from sensors are minimized. We then formulate the MAC design problem as a partially observable optimization problem and develop a myopic policy accordingly.

Third, cloud computing is expected to meet the rising computing demands. Both private clouds, which aim at patients in their regions, and public clouds, which serve general public, are adopted. Reliability control and QoS provisioning are the core issues of private clouds and public clouds, respectively. A framework, which exploits the abundant resource of private clouds in time domain, to enable cooperation among private clouds and public clouds, is proposed. Considering the cost of service failure in e-health system, the first time failure probability is adopted as reliability measures for private clouds. An algorithm is proposed to minimize the failure probability, and is proven to be optimal. Then, we propose an e-health monitoring system with minimum service delay and privacy preservation by exploiting geo-distributed clouds. In the system, the resource management scheme enables the distributed cloud servers to cooperatively assign the servers to the requested users under a load balance condition. Thus, the service delay for users is minimized. In addition, a traffic shaping algorithm is proposed, which converts the user health data traffic to the non-health data traffic such that the capability of traffic analysis attacks is largely reduced.

In summary, we believe the research results developed in this dissertation can provide insights for efficient transmission power allocation for wearable sensor, can offer practical MAC layer solutions for WBANs in hospital environment, and can improve the QoS provisioning provided by cloud networks in e-health systems.

Acknowledgements

First and foremost, I would like to express my deepest and sincerest gratitude to professor Xuemin (Sherman) Shen, for his continuous guidance, encouragement and kind support throughout my Ph.D study. Professor Shen has been an invaluable mentor to me from research to life. He has such deep understanding of research and enthusiasm for high quality research. From him, I learnt not just focus on finding the right solution, but to raise the right question. I also learnt that the quality of research not only lies in the idea, but also in the way to presenting the whole piece. Professor Shen has also shown great care to our students. He is always available to offer us help and advice from research to career development. From the bottom of my heart, I thank prof. Shen for everything that I have achieved during my Ph.D study at university of waterloo, and am about to achieve in my future career.

I would also like to thank Professor Weihua Zhuang, who had a particularly significant impact on my Ph.D study. She provided me with my first research opportunities, and introduced me stochastic process and radio resource management. I am immensely grateful that Professor Zhuang shared her valuable insights on my research topic. The meetings with Professor Zhuang help to train my critical thinking ability and improve my communication skills. Together with professor Shen, they set the standard for who I want to be as a researcher and engineer.

Furthermore, I am fortunate to have a thesis committee that has not only been available for examining this thesis, but also provides constructive reviews and suggestions. Along with professor Shen, these are professor Liangliang Xie, professor Liping Fu, professor Oleg Michailovich, professor from university of waterloo, and professor Cheng Li from memorial university of newfoundland.

I am grateful to the Broadband Communications Research (BBCR) Lab members. My deep appreciation goes to Professor Xiaodong Lin and Dr. Xiaohui Liang for the fruitful collaborations on e-health related research. I am also very grateful to other BBCR Lab members: Dr. Hao Liang, Dr. Tom (Hao) Luan, Dr. Muhammad Ismail, Dr. Khadige Abboud, Dr. Ning Lu, Dr. Zhongming Zheng, Dr. Miao Wang, Dr. Jian Qiao, Neda Mohammadizadeh, Kuan Zhang, Xiaoxia Zhang, Ran Zhang, and Qianqiao Shao.

Special thanks to Dr. Henry Luo at Care in Motion Inc., Dr. Jeff Ungar, Dr. Jeff Koller, Dr. Bryan Brains, Heather Sullens and Dr. William Athas at Apple Inc..

Dedication

This is dedicated to my wife Weiwei Li.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

E-health is defined as healthcare practice supported by information and communication technologies (ICTs) [1]. The goal of e-health is to help physicians, hospitals and governments to provide healthcare services to patients in a more convenient, efficient, reliable and economical way compared to current public heath services. The advancements of sensor technologies, communication networks and computing technologies speed up this evolution process. Specifically, the advancement of wearable sensors enables vital physiology monitoring not only in hospitals, but also anywhere anytime in our daily lives. Wireless body area networks (WBANs) are proposed and developed to support real time medical information transmission from wearable sensors to smart phones. Moreover, with the wide adoption of cloud computing technologies, the data gathered from continuous monitoring could be stored and analyzed for illness prediction and treatments.

Yet many benefits, E-health systems are required to meet the stringent Quality of Service (QoS) requirements, such as reliability, delay and power consumption requirements, of medical applications. However, limited available resources, such as the energy and computing capability of wearable sensors and network bandwidth, pose challenges in designing systems to meet QoS of medical applications.

1.1 E-health Systems

According to American Heritage Dictionary, Healthcare is defined as " the prevention, treatment, and management of illness and the preservation of mental and physical well-being through the services offered by the medical and health professions" [2]. However,

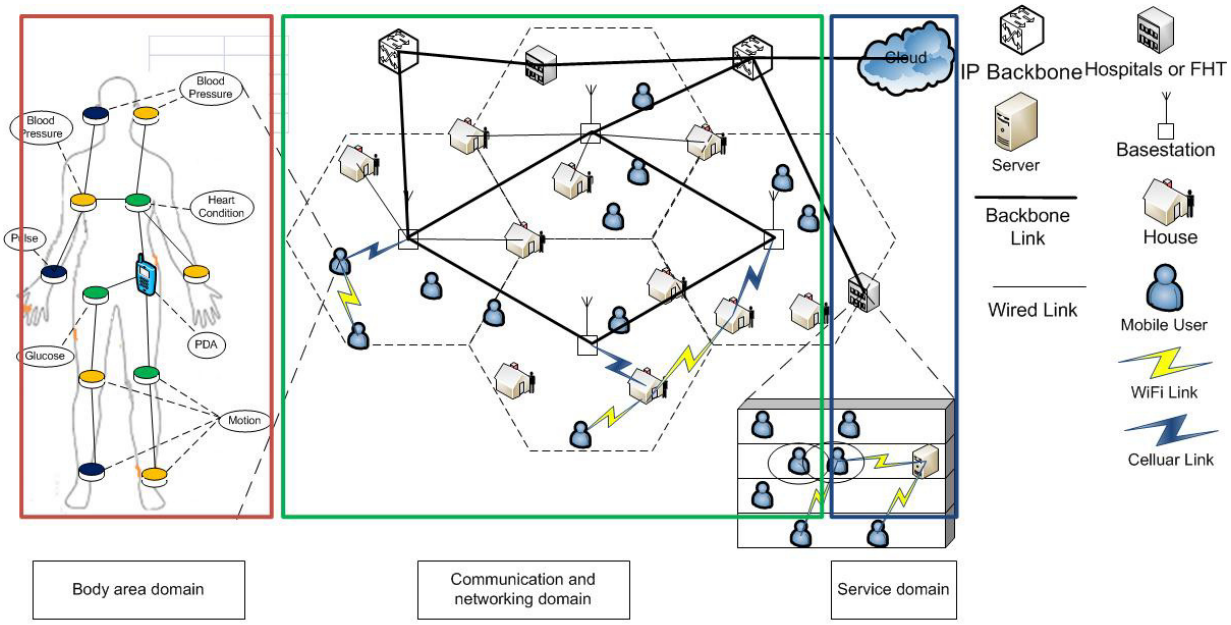


Figure 1.1: Illustration of e-health systems

current healthcare system does not provide satisfied QoS in all aspects. For example, according to Ontario Ministry of Health and Long-term Care, the provincial average waiting time and total time spent in emergency room for complex conditions are 4 hours and 9.6 hours, respectively [3]. What is more, as pointed out by the former president of the Canadian Medical Association, Dr. Jeffrey Turnbull, current hospital centric healthcare is inefficient in dealing with chronic conditions [4], such as heart diseases and diabetes. Firstly, it lacks the ability to identify chronic diseases in the early stage. Secondly, chronic conditions need to be treated in a long-term ward instead of shortly in hospitals.

E-health provides a promising solution to existing healthcare issues. The adoption of ICTs helps to improve access to healthcare, reduce waiting and processing time, and lower the cost. To realize these goals, communication is indispensable for real time health information delivery, whereas computing is required for diagnoses and predictions through signal processing and pattern recognition.

The e-health systems have been categorized into three domains: body area domain, communication and networking domain and service domain [5]. Body area domain has two tasks: one is to gather physiology information and the other is to store it in a smart phone for transmission. As illustrated in Fig. 1.1, body area domain is made of WBANs.

Each WBAN is composed of several sensors and a smart phone. The sensors can be either wearable sensors or implant sensors. Sensors monitor different physiology signals. Thus, the traffic generated by sensors are various in terms of rate and priority. Vital signals, such as heart rate, blood pressure, respiration rate and peripheral capillary oxygen saturation (SpO2), contain crucial information, thus should be given higher priority. Electromyography (EMG), on the other hand, has less importance compared to vital signals, thus can be given lower priority. The smart phone gathers medical information from sensors through WBAN. IEEE802.15.6 [6] is proposed to serve as the communication technology for WBANs.

In communication and networking domain, smart phones organize all personal medical information and report to remote servers in a timely, reliable and security manner. Major smart phone platforms have provided this function. Apple Inc. announced HealthKit^{TM} to collect physiology information and store it in remote servers. Google published Android Fit^{TM} and Microsoft launched Health^{TM}. The communication and networking domain adopts technologies, including wireless network, such as cellular networks and WiFi, and wired Internet Protocol (IP) backbones. For medical traffic, QoS, in terms of delay and reliability, is the most important design consideration. These requirements indicate the needs for tailored designs for medical traffic transmission.

The tasks of service domain are to provide storage for massive health related data, real time diagnoses and early stage diseases detection. Massive health related data services require storage for huge amount of data and timely data access whenever needed. Real time diagnoses and diseases detections could adopt learning algorithms with high computing complexity [7, 8]. To serve the computing intensive and storage intensive applications, cloud computing is proposed to be utilized in e-health systems. Cloud computing is made of shared network of configurable computing resources, such as servers, storage and applications, to enable ubiquitous and on-demand access [9]. Based on deployment methods, cloud computing can be categorized into private clouds and public clouds. Private clouds are clouds operated by a single organization, whereas public clouds are clouds that are open for public usage. Both private clouds and public clouds have been adopted in e-halth systems. Provincial governments in Canada opt to private clouds to suit local needs, such as project E-health Ontario [10]. Federal government of U.S. intends to build national wide electronic health record access systems with public clouds [11].

Several scenarios show the benefits of e-health systems. First is long term remote monitoring [12, 13]. Consider a person with chronic heart diseases at home with a WBAN. For such a patient, continuous monitoring is crucial for early stage heart failure detection. A sensor on body collects heart related physiology data, such as Electrocardiography (ECG) and blood pressure. Information is delivered to remote servers in real time. Through data

analysis, heart arrhythmia, even heart failure, could be predicted by clouds using ECG data[14, 15]. Medical interventions could be provided once heart performance deteriorates. This application helps to reduce the harm caused by sudden heart failure, the cost for frequent clinic visits and enable health services delivery in a continuous and long term fashion.

Second scenario is the information collection in hospital environments. Consider a hospital, where many patients move around, each carries a WBAN. In such scenario, it is challenging to gather all necessary information from patients and provide timely instructions, such as determining the priorities of patients. With the help of e-health systems, smart phones collect and forward information to local servers, which can determine patient priority and direct patients to appropriate departments without waiting for instructions from nurses. This application has the potential to reduce the workload of medical practitioners, cut the waiting time and processing time in hospitals.

1.2 Resource Management in E-health Systems

The e-health systems can only be launched with provisioning of medical level QoS. The metrics of requirements include lifetime of sensors [16], the transmission power, the delay suffered by medical traffic. Compared to exiting data services, medical services have more stringent requirements. However, the e-health systems suffer from dramatic changing network topology and intermittent connectivity, making QoS provisioning a very challenging task. As a result, proper management of the limited resources, including energy, bandwidth and computing resources, is necessary in e-health systems. The following subsections present the key ingredients of e-health systems and the challenges faced by them.

1.2.1 Transmission Power Allocation for Sensors

The limited size of wearable sensors leads to limited battery that can be carried. In the mean time, for the purpose of continuous monitoring, a sensor is expected to work for a long time. For example, wearable sensors are expected to sustain a weekend usage without charging [17]. The expectation on long lifetime and the limited battery call for proper power allocation [18].

Three features of wearable sensors make the energy efficiency transmission challenging. Firstly, for health data, there is a need on stringent QoS provisioning in terms of delay. Data includes vital signal information needs to be guaranteed with a worst-case delay [19].

Secondly, different from mobile terminals, such as smart phones, tablets, which are capable of cellular communication or WiFi communication, wearable sensors are extremely computing capability constrained. Thus, complex signal processing algorithms and transmission scheduling algorithms are not suitable to be implemented on sensors. Thirdly, for the safety of human skins, the peak transmission power is also limited, to prevent overheating phenomenon, which causes skin irritations and burning [20].

Several methods have been devised to increase energy efficiency through transmission power allocation [21, 22, 23, 24]. First, for wireless transmission over dynamic channels, opportunistic transmission [25], which exploits the dynamics of channel states through transmitting more bits when channel state is good and less when channel state is poor, is widely adopted. For wearable sensors, two factors need to be considered to utilize opportunistic transmission. One is the trade-off between delay and energy efficiency, since waiting a long duration for potential good channel could cause violations of delay requirements. The other is the impact of peak transmission power. Based on water-filling transmission policy, more energy should be allocated when channel is in good condition. However, due to the peak transmission power constraint, a sensor may not be able to fully utilize the good channel. Secondly, exploiting the fact that active radio system consumes energy [26] and vital signals typically have low data rate, completely shut down the sensor radio, also know as the sleep mode, is proposed [27]. This method raises the problems when to put sensor to sleep and when to wake it up. We use the term sleep scheduling to refer to the algorithms that addresses this problem. For medical traffic transmission over wireless body area channel, we should consider the possibility that after a period of sleep, a sensor wakes up, but have to face a long period of bad channel conditions that may cause transmission failure and delay violations. In summary, energy efficient and low complexity transmission power allocation schemes for wearable sensors are desired.

1.2.2 Medium Access Control for WBANs

In hospital, multiple WBANs coexist as each patient carries one WBAN. Each patient gathers medical information through WBAN on the body with stringent QoS provisioning [28, 29, 30]. Since nearby WBANs share the same wireless medium, the transmissions of different WBANs could interfere with each other, resulting in packet loss and energy waste for sensors, and more importantly, violation of delay requirements. Thus, medium access control is needed.

Several factors make the MAC design for WBANs in hospital challenging [31]. Firstly, considering the limited transmission power and the channel fading, not all sensors are

within the communication range of each other. As a result, the network is distributed in nature. In such a network, the notorious hidden terminal problem will likely exist. Secondly, the body area channel is prone to errors due to the body movement. As a result, even if a WBAN is granted access to wireless medium, the channel between the sensor and the smartphone could be in bad state, resulting in waste of channel resources. MAC design aiming at maximizing network throughput calls for the consideration of dynamic channel. Thirdly, all patients share the same rights to access the wireless resources. Thus, fairness is demanded in MAC layer resource management. In summary, MAC layer resource management requires tailored design in hospital environments.

Existing MAC layer protocols have limitations when applied to WBANs [32, 33, 34, 35, 36]. Firstly, existing works on hidden terminal cancellation rely on control signal exchange between transmitters and receivers, such as the black burst based scheme, which adopts a busy tone to jam the channel in order to gain access. However, these schemes consume a lot of energy for transmitters, thus not suitable for wearable sensors. Secondly, body area channel dynamics call for consideration when allocating channel access. MAC protocol for wireless terminals, such as laptops and tablets, does not take channel into account. With channel dynamics, MAC needs to intelligently avoid allocating channel access to link with poor channel condition. This increases the complexity of MAC layer design. Moreover, the transmitter, namely the wearable sensors, and the receivers, namely the smartphones, have asymmetrical capabilities in terms of energy supply and computational power. This feature has not been explored by previous works where terminals on both sides are considered to have similar capabilities. In summary, how to utilize the unique characteristics of WBANs to design MAC for hospital environment applications needs to be investigated.

1.2.3 Resource Management in Cloud Networks

Cloud computing is expected to meet increasing computing demand, such as storage and computing tasks, for e-health applications [37]. Based on targeted users, current computing service providers can be classified into public clouds and private clouds, such as SIEMENS Healthcare Private Cloud [38], which provides personal health information storage, monitoring and access management. When it comes to medical applications, availability and reliability of cloud computing services are of the utmost importance [38]. However, current cloud service provider face delay and reliability issues in order to meet medical grade requirements.

The deployment of public cloud is progressing dramatically due to several potential benefits for users: easy to deploy, fast to scale and flexible to use [39, 40]. As the unavoidable communication delay from the backbone network hurts the revenue of the public

service providers in an unacceptable manner, geographically deploy storage and computing servers have become key technologies to improve QoS in cloud computing [41]. However, to geographically deploy data centers and servers faces obstacles. Medical computing facilities are required to be built with a stringent standard [42], thus it is expensive for public service providers to maintain suitable environments to accommodate these servers at different locations.

At the same time, private cloud, referred as internal data centres and computing facilities of a business or other organizations provide specific services for targeted users, suffers from under-utilization problem during off-peak periods and poor services during burst periods. Service provisioning during burst period is particularly difficult since when the burst will come and how large the burst will be are hard to predict. The IT administrators face the dilemma between extremely expensive computing facilities capable of dealing with burst requests and consequences from unsatisfactory services. Replacing all private cloud with public cloud can help improve the reliability of service for enterprises. But this approach wastes the resource of current private cloud facilities.

The reliability and availability of the computing systems are of critical importance in medical application. The result of failure could be fatal. For example, when a user with near heart congestion condition requests for diagnoses, late decision could result in failure in heart congestion prevention. However, as discussed above, private clouds have limited computing capability, which makes it unreliable when a burst of demand arrives [43]. In fact, most private clouds in market are designed to have three times of computing capacity compared to average demand. On the other hand, services provided by public clouds suffer from communication delay due to distance. For public cloud, geo-distributed cloud network has been considered as an effective way to reduce delay. However, geo-distributed cloud requires proper load balancing among computing facilities [44]. Moreover, for medical applications, privacy is considered as the top priority when using public cloud [38]. Many attackers in the internet could violate the privacy of a patient and make profit by exploiting the private information, such as targeted advertisement. One of the attack that can harm the privacy of a patient is traffic analysis (TA) attack [45]. TA attack is a type of inference attack that aims at determining traffic type through learning patterns of the communication, even if all data is encrypted.

Previous works on improving the reliability of private cloud suggest to utilize public cloud for burst demand arrival [46]. The idea is whenever local demand exceeds the computing capacity of private clouds, the newly arrived requests are directed to public clouds. This, however, does not fully utilize the abundant computing capacity of private clouds in average sense. Thus, how to design proper resource management schemes to fully exploit the resource of private clouds for reliability improvement needs investigation. For geo-

distributed clouds, previous works [47, 48, 44] on load balancing are designed for less delay sensitive applications compared to medical applications. Thus, a load balancing scheme with delay taken into account is needed.

1.3 Research Contributions

The objective of this research is to devise resource management schemes to achieve reliable and efficient e-health systems. As discussed above, existing work has limitations in applying to e-health applications. We focus on transmission power allocation for wearable sensors, MAC protocols for WBANs in hospitals and resource management schemes for cloud networks. Specifically, we devote to investigate transmission power allocation, which is to exploit the channel dynamics to reduce the average power consumption while providing a worst-case delay guarantee, and to build understanding of the impacts of the peak power constraint and statistical QoS provisioning; to incorporate channel dynamics and utilize the capability of smart phone in MAC design to support medical information collection in a hospital environment; to improve the reliability of private clouds through cooperation with public clouds; and to preserve privacy against traffic analysis, and balance load in geo-distributed clouds. The research contributions are discussed in the following.

1.3.1 Energy Expenditure Analysis of QoS Provisioning

To design proper transmission power allocation schemes for wearable sensors, we aim to characterize the tradeoff between energy expenditure and QoS requirements, such as delay requirements and peak power constraints. Specifically, our objectives are as follows:

- Analyze the tradeoff between worst-case delay and average transmission power for vital medical information transmission over dynamic body area wireless channel, and propose a low complexity algorithm suitable for wearable sensors [49].
- Analyze the impacts of the peak power constraint on energy efficiency and the impacts of statistical QoS provisioning.

For vital signal, including heart rate, respiration rate, oxygen saturation and blood pressure, whose data rate is much smaller compared to channel capacity, we investigate how to utilize the sleep and opportunistic transmission to improve energy efficiency, while guaranteeing a worst-case delay. For sensors, idle listening consumes comparable energy

as transmission, thus sensors are to put radio off for energy saving. This requires a control over when the sensor should be radio on and when the sensor should be radio off. Generally, the longer a sensor is put to radio off, the larger the delay suffered by data in buffer. In the meantime, if there is limited data in the buffer, a sensor is set to radio on and the channel is detected in good condition, the good channel state could not be fully utilized due to insufficient data. To achieve energy saving by exploiting propagation channel quality with a worst-case delay requirement poses challenges in developing a scheduling policy. We address this problem with a Lyapunov optimization formulation and propose a two-step scheduling algorithm. We prove that the algorithm can provide worst-case delay guarantee under certain conditions. Theoretical analysis and simulation results are presented to demonstrate the tradeoff between the transmission delay and energy consumption.

For certain physiology monitoring, such as implant camera and EMG monitor, the data rate is considerably larger than the data rate required for vital signal monitoring. For theses applications, we investigate the impacts of the peak transmission power and statistical QoS provisioning. With peak transmission power constraint, the good channel state may not be fully utilized for insufficient power. We formulate the minimization of average power with a peak power constraint and an average transmission rate constraint problem as a convex optimization problem, derive an optimal solution and propose an efficient calculation method. Applying duality gap analysis, we characterize the upper bound of the extra average transmission power incurred due to a peak power constraint. Further, we show that a constant power scheme is suitable for wearable sensors to support application with large data rate. To support delay requirements, we incorporate a QoS provisioning constraint, in the form of statistical delay guarantee, into our formulation, and derive the solution accordingly. Through simulations, we shown that peak power is the bottleneck for wearable sensors to support stringent statistical QoS provisioning.

1.3.2 MAC Solution with Partial Network States

The objective of studying MAC for WBANs is to reduce interference in an energy efficient way to satisfy the QoS requirement of medical applications in hospital. Note that in a hospital environment, due to the limited space, multiple WBANs would coexist in a region and share the channel to support physiology information collection from wearable sensors by smartphones. This inevitably incurs severe inter-WBAN interference which, if without an appropriate design, would significantly reduce the network throughput and, more importantly, incur high power consumption of wearable sensors. Therefore, an efficient channel resource management scheme in MAC layer is crucial. On addressing this issue, in this work, we develop a centralized MAC layer resource management scheme for WBANs,

with a focus on inter-WBAN interference mitigation and sensor power consumption reduction. Based on the channel state and buffer state information reported by smart phones deployed in each WBAN, channel access allocation is performed by a central controller to maximize the network throughput. Note that sensors have strict limitations on energy and computing capability and cannot timely provide all the necessary information for channel resource management, which deteriorates the network performance. We exploit the temporal correlation of body area channel such that channel state reports from sensors are minimized. We formulate the network design as a partially observable optimization problem and develop a myopic policy accordingly. Simulation results are presented to demonstrate the effectiveness of our proposed policy.

1.3.3 Cooperation Schemes for Cloud Networks

To improve QoS provisioning capability of cloud networks, we intend to enable cooperation through proper incentive scheme design. Specifically, our objectives are as follows:

- Propose a scheme to enable the cooperation between private clouds and public clouds such that the reliability of private clouds is improved and the delay suffered by public clouds is reduced [50].
- Investigate the load balancing problem for geo-distributed clouds with delay minimization and privacy preservation [51].

First, a cooperation framework is proposed through exploiting unique features of existing clouds. First feature is that private clouds are geographically distributed, whereas the second is public clouds can be regarded to possess infinite computing resources available [41]. The cooperation contains two key ideas: 1) The geographically distributed private clouds offer parts of their capacities to help a public cloud provide services to its nearby users. The public cloud rewards this help. In other word, private clouds turn their resources that can not be accumulated into rewards that can be accumulated through the cooperation with a public cloud. 2) With sufficient reward, the public cloud offers to serve excess requests to the private clouds to improve their reliability and scalability. To adopt reward is to eliminate selfish private clouds. This cooperation scheme improves reliability of private clouds and reduces the delay of public clouds at the same time.

Second, we propose a resource management scheme to achieve the minimized service delay and the reduced communication costs. We first derive a sufficient condition in resource management to ensure the stability of cloud servers. Considering this condition, we design

the resource management scheme: each server only redirects the requests to others who have shorter queue lengths; and the number of redirected requests must be proportioned to the difference of their queue lengths and reciprocal to the service delay between them. We also prove the proposed resource management scheme satisfies the derived sufficient condition in balanced state. In addition, we compare the scheme with two other alternatives using joint the short queue (JSQ) and distributed control law (DCL), both of which are proven to be stable. Through extensive simulations, we show that our scheme achieves a much smaller average service delay than the JSQ-based and DCL-based schemes.

Third, we propose a traffic shaping algorithm to prevent the health data of users from being detected by the TA attackers. We focus on the health data traffic generated by e-health monitoring systems, such as heart rate and blood pressure, which are typically modelled as deterministic processes [52]. We analyze the statistical differences between health data traffic and non-health data traffic. Our proposed shaping algorithm is designed such that the distribution of the shaped health data traffic is the same as the distribution of the non-health data traffic; and the autocorrelation of the shaped health data traffic is close to the autocorrelation of the non-health data traffic. Note that, the proposed algorithm introduces a delay, referred as shaping delay, on the user side which is related to the privacy requirement. We provide the numeric results on this relation. Then, we model the shaping delay by the D/M/1 queue, and consider the shaping delay into the resource management scheme. The simulation results show that our resource management scheme is still efficient with the shaping delay.

1.4 Outline of the Thesis

Chapter 2 and Chapter 3 are devoted to the transmission power allocation problem. Chapter 2 formulates the energy efficiency transmission with worst-case delay provisioning problem. An algorithm is proposed and proven to be able to guarantee worst-case delay. Chapter 3 investigates the impacts of peak power and statistical QoS provisioning on transmission power allocation. Chapter 4 studies how to utilize the dynamics of body area channel to maximize the network throughput in MAC layer. In this chapter, we show that the proposed MAC could achieve satisfactory performance in terms of throughput and fairness. Chapter 5 proposes a cooperation framework between private clouds and public clouds. Through survival analysis, we show that the cooperation scheme enhances the reliability of private clouds. Chapter 6 presents a traffic shaping algorithm to protect privacy from traffic analysis and a load balancing algorithm, which is proven to be able to stabilize cloud networks. The efficiency of the proposed scheme is evaluated by analysis

and simulations based on the geographic and population of Canada. Finally, Chapter 7 summarizes the results of the thesis and outlines a number of potential avenues of future research. Possible extensions include developing fully distributed MAC for WBANs, and resource management in cloud networks.

Chapter 2

Transmission Power Allocation with Worst-case Delay Provisioning

Data collection from a single sensor is an important and common scenario of e-healthcare applications in WBANs. Sensors with multiple functions are becoming popular nowadays. For example, Apple Watch and Fitbit can monitor heart rate, count steps at the same time. Most chronic patients need only one sensor to monitor his or her specific physical condition. In developing communication protocols for this scenario, energy consumption of the wearable sensor is an essential consideration. Meeting a worst-case delay requirement for medical data transmissions makes the protocol development a challenging issue.

In the literature, two approaches for wireless communication and sensor networks have been proposed for energy efficiency:

- Opportunistic communication [21]: The idea is to transmit as long as the channel is in a good condition, through exploiting dynamic fluctuations of channel gain;
- Sleep scheduling [53]: With a low traffic load, radios of sensor nodes are turned off to reduce idle listening.

Two facts motivate us to consider both approaches in WBANs: 1) The channel gain of body area wireless channel varies over time; 2) A sensor on body consumes energy in a listening state. The side effect of these approaches is extra the delay in data transmission. For example, a transmitter using opportunistic transmission policy, holds data when the channel is in a bad condition [54]. Thus, the transmission delay of data in buffer increases.

For medical applications in WBAN, the tradeoff between delay and energy consumption needs to be studied.

In this work, we investigate the problem of minimizing average power consumption over a time varying wireless body area channel with a worst-case delay requirement and random traffic arrivals. The minimizing average power consumption problem can be solved using dynamic programming (DP). However, DP suffers from the curse of dimensionality. Its computation complexity grows exponentially with the number of system states, including channel states, buffer states, which causes difficulty in implementation on sensors. In this work, we formulate the problem based on the Lyapunov optimization theory [55]. A two-step power allocation algorithm is proposed, which utilizes both sleeping mode and opportunistic transmission for energy efficiency. The algorithm is suitable for sensors because: 1) it does not require *a-priori* information of the channel gain and data arrival rate; and 2) it follows a fixed threshold policy for the first step, which has lower computation complexity when compared with DP. In performance analysis, we study the conditions under which the proposed algorithm can guarantee a worst-case delay and the tradeoff between the worst-case delay and power consumption. Computer simulations results are presented to demonstrate the performance of the proposed algorithm.

2.1 Literature Review

Research on sleep scheduling originates from sensor networks [27]. In a sensor network, the power consumption in a listening state is comparable to that in a transmission state. Sensor radios are thus turned off for energy saving. For sensor networks, the delay under consideration is the time difference between the time instance a packet is generated to the time instance the packet is received by a sink. Since routes from a sensor to a sink are usually multi-hop, most work on controlling delay for sensors with sleep scheduling focuses on ensuring the connectivity of the multi-hop networks [56]. Moreover, work for sensor networks generally considers the channels to be time invariant. In summary, this line of work approaches the sleep scheduling problem in MAC and routing layers, thus can not be directly applied to WBANs.

The fundamental tradeoff between average power consumption and average delay over a fading channel is investigated in [21]. It is concluded that any scheduling policy, requiring transmission power smaller than $O(1/V)$ plus the minimum power, must have an average queueing delay greater than or equal to $\Omega(\sqrt{V})$, where V is a weighting factor. The power delay tradeoff problem based on a static-channel assumption under a worst-case delay requirement is investigated in [57]. In [22, 23], time varying channel models and a

worst-case delay requirement are considered. A suboptimal policy for Shannon capacity based power consumption model is proposed in [22], while an optimal policy is developed for a monomial power consumption model in [23]. Both work assumes data arrives at the beginning of each time frame, and is required to be transmitted at the end of the time frame. Both schedulers are controlled based on the observations of the instantaneous channel gain, buffer length, and remaining time to deadline.

Existing policies with a worst case delay assume either the channel is static or the traffic is static. In WBAN, the channel dynamics have been observed by extensive experiments [58]. So does the traffic dynamics. Hence, dynamics of both channel and traffic are necessary for practical formulation. However, randomness in channel and traffic makes the problem complex. When DP is adopted, the computation complexity grows exponentially with both the number of states for channel and traffic. Specifically, with channel dynamics, an optimal scheduler shall consider both current channel information and future channel states. Moreover, without the traffic dynamics, all the data in the buffer share the same deadline. With dynamics of the traffic, data in the buffer have different remaining time to go. As a result, a scheduler considers how much data from different urgent levels need to be transmitted. This increases the complexity of the problem.

Delay constrained optimization problems are investigated in [59, 60] using a framework to combine stability and optimization techniques, called Lyapunov optimization theory [55]. It takes account of both the stability of the system and other system utility maximization objectives, such as fairness, and energy consumption. Based on the theory of Lyapunov, which studies the stability issues, Lyapunov optimization theory further considers other system utility maximization objectives, such as fairness, and energy consumption. This framework can yield an algorithm is developed to stabilize the system and drive the utility to an optimal value. The performance in terms of delay and other utilities is evaluated in the form of upper bounds. Given a weighting factor V and a quadratic-form Lyapunov function, it is shown that the algorithm designed can push the utility performance to an optimal value within $O(\frac{1}{V})$ at the cost of an average delay increases within $O(V)$.

2.2 System Model

The WBAN system model under consideration is presented in this section. Elements in the WBAN are described from the perspectives of their limitations and functionalities. A medical traffic model with a worst-case delay requirement is introduced, followed by channel models and an energy consumption model for reliable transmission.

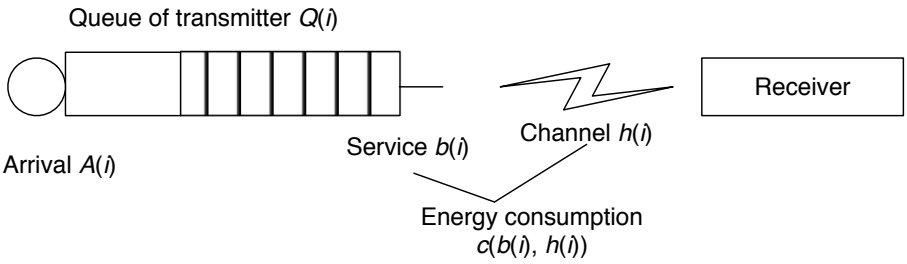


Figure 2.1: Model for continuous monitoring

Network Model

Fig. 2.1 illustrates the model for the remote monitoring application under consideration. Consider a body area network, consisting of a smart phone and a wearable sensor, where the sensor is the transmitter and the smart phone is the receiver. The sensor has a stringent requirement on power consumption, and a maximum transmission power P_{max}. In comparison with the sensor, the smart phone has more power supply and more signal processing capability. As a result, we assume that the smart phone is always turned on for simplicity, and is in charge of the data collection.

The system time is partitioned into slots of constant duration, denoted by T. A narrow band frequency channel is adopted. Assume the channel gain is reciprocal, constant over each time slot, but changes independently from slot to slot [61]. At the beginning of each time slot, smart phone sends out a pilot signal with constant transmit power for the sensor to evaluate current channel state based on the received signal power level. The duration of the pilot is αT, where $\alpha(\alpha < 1)$ is the ratio of pilot duration to time slot duration. The sensor can either listen to this pilot and then transmit some data or stay in sleeping mode. During the sleeping mode, the sensor only inactivates its radio. Hence, a transmission policy, denoted by μ, consists of two decision variables for each time slot. One is sleep decision during time slot i, defined as

$$s(i) = \begin{cases} 0, & \text{if the sensor sleeps;} \\ 1, & \text{if the sensor wakes up.} \end{cases} \tag{2.1}$$

If the sensor decides to wake up, it makes the second decision, which is how much data the sensor should transmit during time slot i, denoted by $b(i)$.

Data Traffic Model and Delay Requirements

The requirements for medical traffic are application based. To facilitate analysis, all different medical data are assumed to have the same requirement in this work. Let $A(i)$ be the number of information bits arrived on time slot i, with a maximum A_{max}. The numbers, $A(i)$, are independently and identically distributed (i.i.d) and ergodic. The reason for the random arrival rate is that to get accurate vital signal from raw data, such as obtaining heart rate from Photoplethysmogram (PPG) signal, the complexity of the algorithm depends on the quality of the PPG signal [62]. In general, the poorer the signal quality, the higher the algorithm complexity. The difference in the algorithm complexity will result in the difference in processing time, thus the difference in the arrival time of vital signal data. In reality, the quality of the PPG signal is highly correlated with the motion status of the person [63]. Since a person could move randomly, we assume a random arrival for vital signal.

The delay experienced by data is the sum of the delay in the buffer of sensor and the transmission delay. Compared to the delay in the buffer, the transmission time can be omitted [21]. Thus we focus on the delay in the buffer. Let $Q(i)$ be the number of bits in the buffer of the sensor at the beginning of time slot i. The data in the buffer is processed in the First Come First Served (FCFS) principle. The buffer backlog is empty at the time slot 0, $Q(0) = 0$. We have

$$Q(i+1) = \max[Q(i) - b(i)s(i), 0] + A(i)\,. \tag{2.2}$$

The max operator ensures that the number of bits in the buffer is non-negative. The buffer decreases $b(i)s(i)$ bits is the buffer is not empty and increases $A(i)$ bits in the end of each slot.

A worst-case delay requirement is considered for time critical medical applications. Worst-case delay is the maximum time that a data experiences in the sensor buffer. Let D^{μ}_{max} be the worst-case delay experienced by any data in the buffer under policy μ. Let D_{max} be the worst-case delay allowed, typically from 100ms to 250ms for medical data.

Channel Models

Let $h(i)$ be the channel state in time slot i, defined as the channel path loss between transmitted signal power $P_T(i)$ and received signal power $P_R(i)$. Generally, the channel path loss is a constant for a 10ms duration [61][64] in WBANs. Compared to the delay requirements of medical traffic, the channel state variations can be utilized for opportunistic transmission.

The channel path loss consists of a distance between the transmitter and receiver dependent component and a random component. The random component is ergodic, i.i.d over time slots, and follows Gaussian distribution [61]. Both minimum h_{min} and maximum value h_{max} of the path loss value are given. The statistical information of the channel state including mean and variance is available to the sensor for decision making. Under a Gaussian distribution model, the sensor has the probability density function (pdf) of the random component based on its mean and variance.

Energy Consumption Model

Energy consumption of the sensor node contains two parts. One comes from a listening state, and the other is the required transmission power for reliable communication. The power consumption in the listening state is constant, denoted by P_L. As a result, $\alpha P_L T$ energy is consumed for accessing the channel state during a waking-up time slot. Let m be the ratio of P_L to P_{max}. To ensure reliable communication at rate $r(i)$ in time slot i, the required received signal to noise ratio is $k'[r(i)]^n, k' > 0, n > 1, (n, k' \in R)$, where k' is a scaling factor and n an index factor. This monomial type function gives good approximation for most practical transmission [65]. Let $P_T(h(i), b(i))$ denote the required transmission power for the sensor to transmit $b(i)$ bits of data for channel state $h(i)$. At a result, the required transmission power is

$$P_T(h(i), r(i)) = k \cdot [h(i)][r(i)]^n, \tag{2.3}$$

where $r(i) = \frac{b(i)}{(1-\alpha)T}$, $k = k' \times P_n$ and P_n is the noise power at the receiver end. Let $c(i)$ be the total energy consumption during a time slot. It is the sum of the energy spent in a listening state and transmission state during a waking-up time slot i

$$c(i) = s(i)\{\alpha P_L + (1-\alpha)k \cdot h(i)[r(i)]^n\}T\ . \tag{2.4}$$

2.3 Problem Formulation

In this section, we first formulate the problem of minimizing the average energy consumption of a sensor with a worst-case delay requirement. The formulation does not provide a relationship between the decision variables, including sleeping decision $s(i)$ and transmitted bit $b(i)$, and the delay constraints. Thus, we transform the worst-case delay constraint into a buffer occupancy constraint.

The problem is characterized by making decisions at each time slot to minimize a time average cost. We transform the problem using Lyapunov optimization theory. The benefits are two fold: 1) it decomposes a time average objective into objectives for each time slot; 2) it is capable of capturing the trade-off between different metrics.

Define $\bar{c}$ as the time average of the expectation of the energy cost of a wearable sensor

$$\bar{c} \triangleq \lim_{I\to\infty} \frac{1}{I} \sum_{i=0}^{I-1} \mathbb{E}\{c(i)\}. \tag{2.5}$$

The problem of minimizing energy cost for the sensor with a worst-case delay requirement is formulated as:

$$\mathbf{P1} \qquad \min_{s(i),b(i)} \quad \bar{c} \tag{2.6}$$

$$\text{s.t.} \qquad s(i) \in \{0,1\}, b(i) \leq b_{max}. \tag{2.7}$$

$$D^{\mu}_{max} < D_{max}, \tag{2.8}$$

where D^{μ}_{max} is the worst-case delay under a policy denoted by μ. In **P1**, the constraint (2.7) describes the feasible region of each decision variable. The constraint (2.8) guarantees the worst-case delay is finite and bounded. To describe the relationship between a worst-case delay constraint and the transmission decision variables, we first transform the worst-case delay constraint.

Worst-Case Delay Constraint Transform

The delay experienced in the buffer can be linked to the transmission decision variables through the buffer occupancy. The little's law describes the relationship between average delay and average buffer occupancy. However, there is no direct link between maximum buffer occupancy and maximum delay in the buffer. We follow the steps in [66] to construct the relationship between queue occupancy and a worst-case delay. A virtual queue, denoted by $Z(i)$, is introduced. The virtual queue increases at a virtual arrival rate $\varepsilon_{1(Q(i)>0)}$, which is a constant arrival data ε if actual queue $Q(i)$ is not empty:

$$\varepsilon_{1(Q(i)>0)} = \begin{cases} 0, & \text{if } Q(i) = 0; \\ \varepsilon, & \text{otherwise}. \end{cases} \tag{2.9}$$

The virtual queue updates according to

$$Z(t+1) = \max[Z(i) - b(i)s(i) + \varepsilon_{1(Q(i)>0)}, 0] \,. \tag{2.10}$$

Ensure both the occupancies of virtual queue and actual queue having an upper bound yields a worst-case delay.

lemma 1 [66] *Suppose the system is controlled by a policy μ, so that $Z(i) \leq Z_{max}$, $Q(i) \leq Q_{max}$ for all i, for some positive constants Z_{max} and Q_{max}. Then all data in the queue is transmitted with a maximum delay of D^{μ}_{max}, given by*

$$D^{\mu}_{max} \triangleq \lceil (Q_{max} + Z_{max})/\varepsilon \rceil \tag{2.11}$$

where function $\lceil x \rceil$ is the minimum integer larger than x.

Based on Lemma 1, the worst-case delay constraint (2.8) can be transformed to a buffer occupancy constraint as:

$$Q(i) < Q_{max}, \quad Z(i) < Z_{max}, \quad \text{for any } i. \tag{2.12}$$

Lyapunov Optimization Formulation

The objective of **P1** is time average minimization of an energy cost. To decompose this objective into optimization goals for each time slot, we adopt Lyapunov optimization theory [55]. The Lyapunov theory studies the stability of a dynamical system through each difference in the time domain of a non-negative function of the buffer occupancy, called drift. Through making decisions based on a drift-plus-penalty function, Lyapunov optimization theory introduces a method to control the delay of such a system and other utilities at the same time.

Define $\Theta(i) \triangleq [Q(i); Z(i)]$ as the vector of system states. Define a quadratic form Lyapunov function of the system states

$$L(\Theta(i)) \triangleq \frac{1}{2}[Q(i)^2 + Z(i)^2] \,. \tag{2.13}$$

Define the one-step Lyapunov drift, denoted by $\triangle(\Theta(i))$, as the difference in time of Lyapunov functions in the form of expectation given current system states $\Theta(i)$:

$$\triangle\left(\Theta(i)\right) \triangleq \mathbb{E}[L(\Theta(i+1)) - L(\Theta(i))|\Theta(i)] \,. \tag{2.14}$$

In the following, the upper bound of a drift-plus-penalty function for each time slot is derived. According to Lyapunov optimization theory, delay constrained average cost minimization problem can be transformed to a problem of minimizing the upper bound of the drift-plus-penalty function for each time slot. The logic of transform is two fold:

- To minimize a Lyapunov drift in every time slot is to control the buffer occupancy, thus the delay in the buffer;
- To minimize a penalty function in every time slot is to minimize the energy consumption.

Let $\triangle(\Theta(i))^{sup}$ denote the upper bound for Lyapunov drift $\triangle(\Theta(i))$. It is obtained through the derivation of the upper bound of Lyapunov function. To get the upper bound of Lyapunov function, we first remove the max[ ] operator in (2.2) and (2.10) through enlarging their right-hand items:

$$\begin{aligned} L(\Theta(i+1)) \leq & \frac{1}{2}[Q(i)^2 + A(i)^2 + 2A(i)Q(i) - 2Q(i)b(i)s(i) \\ & + Z(i)^2 + 2b(i)^2 s(i)^2 + \varepsilon^2 + 2Z(i)\varepsilon \\ & - 2b(i)s(i)(z(i)+\varepsilon)] \,. \end{aligned} \tag{2.15}$$

Using equation (2.15), we can obtain the upper bound for Lyapunov drift

$$\begin{aligned} \triangle(\Theta(i)) \leq & \mathbb{E}\{B_{max} - Q(i)[b(i)s(i) - A(i)] \\ & - Z(i)[b(i)s(i) - \varepsilon] | \Theta(i)\} = \triangle(\Theta(i))^{sup} \end{aligned} \tag{2.16}$$

where $b_{max} = \lceil (1-\alpha)T[\frac{P_{max}}{k(1-\alpha)h_{min}}]^{\frac{1}{n}} \rceil$ is the maximum number of bits that can be transmitted during a time slot, and $B_{max} = \frac{1}{2}[A_{max}^2 + 2b_{max}^2 + \varepsilon^2]$ is a finite constant.

The upper bound of a drift-plus-penalty function is the summation of the upper bound of a Lyapunov drift and a weighted cost function. Thus, the objective (2.6) under delay constraint (2.12) is transformed to the following problem in every time slot:

$$\min \qquad \triangle(\Theta(i))^{sup} + V\mathbb{E}\{c(i)|\Theta(i)\} \tag{2.17}$$

where V is a non-negative weighting factor. The weighting factor is used to balance the tradeoff between the delay and energy consumption. In practice, if a sensor gives higher priority to delay over energy consumption, V should be set to be small. Otherwise, V should be given a large value.

Substitute $\triangle(\Theta(i))^{sup}$ and $c(i)$ in (2.17) with (2.16) and (2.4), respectively. The transformation of **P1** using Lyapunov optimization theory is:

$$\begin{aligned}\mathbf{P2}\ \max_{s(i),b(i)}\ &\mathbb{E}\{Q(i)[b(i)s(i)-A(i)]+Z(i)[b(i)s(i)-\varepsilon]\\&-Vs(i)(\alpha P_L+(1-\alpha)kh(i)(\frac{b(i)}{(1-\alpha)T})^n)T|\Theta(i)\}\end{aligned} \tag{2.18}$$

$$\text{s.t.}\quad s(i)\in\{0,1\}, b(i)\leq \min[b_{max}, Q(i)]. \tag{2.19}$$

Problem **P2** is a nonlinear integer programming problem with two decision variables. Compared with **P1**, the objective of **P2** is a function of the decision variables $s(i)$ and $b(i)$. The summation of the first two items is the upper bound for Lyapunov drift $\triangle(\Theta(i))^{sup}$, and the second item is the weighted energy consumption.

2.4 Algorithm Design

Based on the objective in **P2** for each time slot and the fact that a sensor makes decision on $s(i)$ first, we propose an online algorithm in the following. For an online algorithm, when making the decision on whether or not to wake up, a sensor does not know current channel state. Hence, we propose a two-step algorithm to solve **P2**.

Step 1 (*Sleep Scheduling*): Consider problem **P3**.

$$\begin{aligned}\mathbf{P3}\ \max_{s(i)}\ &\mathbb{E}\{[-V(\alpha P_L+(1-\alpha)kh(i)(\frac{b(i)}{(1-\alpha)T})^n)T)\\&+b(i)Q(i)+b(i)Z(i)]s(i)|\Theta(i)\}\,.\end{aligned} \tag{2.20}$$

P3 is constructed by picking up items related to waking-up decision $s(i)$ in **P2**. Solve **P3** according to a threshold policy:

$$s(i)=\begin{cases}1, & \text{if } Q(i)+Z(i)\geq V\Upsilon_{min};\\ 0, & \text{otherwise}\end{cases} \tag{2.21}$$

where $\Upsilon=\frac{T}{b(i)}\{\alpha P_L+(1-\alpha)k\cdot h(i)[\frac{b(i)}{(1-\alpha)T}]^n\}$, and Υ_{min} is the expectation on the minimum of Υ. To calculate Υ_{min}, we first minimize Υ for every possible channel state by choosing $b(i)$. Then, we take expectation of the minimum of Υ on the sample space of the channel states. The symbol Υ_{min} is a fixed value for a given link. Thus, the sensor calculates Υ_{min}

once. The threshold $V\Upsilon_{min}$ influences the upper bounds of $Q(i)$ and $Z(i)$, thus the worst-case delay in buffer. Their relationship is shown in Theorem 1 in performance analysis section.

Step 2 (*Opportunistic Transmission*): If $s(i) = 1$, the sensor determines $b(i)$ after acquiring the channel state $h(i)$ based on the pilot signal. Consider problem **P4**:

$$\mathbf{P4} \max_{b(i)} \; b(i)[Q(i) + Z(i)] - Vkh(i)[(1-\alpha)T]^{1-n} b(i)^n \tag{2.22}$$

$$\text{s.t. } b(i) \leq \min[b_{max}, Q(i)]. \tag{2.23}$$

Solve **P4** to obtain $b(i)$ by:

$$b(i) = \begin{cases} b(i)^*, & \text{if } b(i)^* \leq \min[b_{max}(h(i)), Q(i)]; \\ \min[b_{max}(h(i)), Q(i)], & \text{otherwise} \end{cases} \tag{2.24}$$

where $b(i)^* = \lceil (1-\alpha)T[\frac{Q(i)+Z(i)}{kVn \cdot h(i)}]^{\frac{1}{n-1}} \rceil$, and $b_{max}(h(i))$ is the maximum number of bits that can be transmitted in a time slot given a channel state $h(i)$.

2.5 Performance Analysis

Algorithm derived from **P2**, which minimizes the upper bound of drift-plus-penalty, does not ensure a non-positive drift. Thus, whether or not the lengths of actual queue and virtual queue increase to infinity can not be determined using the Lyapunov theory. Instead, we investigate the queue occupancy problem by studying the evolution of the actual queue and virtual queue under the proposed algorithm. We examine the conditions under which the proposed algorithm can guarantee a worst-case delay. Also, the tradeoff between the worst-case delay and energy consumption is presented in this section.

Lemma 1 states that a worst-case delay exists if both the actual queue and virtual queue are bounded. Hence, an algorithm can provide a worst-case delay if it can ensure the existence of upper bounds for both queues. Define two conditions as follows:

$$(1-\alpha)T[\frac{\Upsilon_{min}}{kn \cdot h(i)}]^{\frac{1}{(n-1)}} \geq \max[A_{max}, \varepsilon] \tag{2.25}$$

$$b_{max}(h(i)) \geq \max[A_{max}, \varepsilon]. \tag{2.26}$$

theorem 1 *If conditions (2.25)-(2.26) hold, then worst-case upper bounds exist for actual queue and virtual queue as follows:*

$$Q(t) \leq V\Upsilon_{min} + A_{max} \, . \tag{2.27}$$

$$Z(t) \leq V\Upsilon_{min} + \varepsilon \, . \tag{2.28}$$

Theorem 1 indicates that the proposed algorithm can provide a worst-case delay guarantee under the conditions (2.25)-(2.26). We provide the proof by induction for the existence of the upper bound for actual queue as stated in theorem 1. The proof for the virtual queue bound is similar and is omitted.

Proof 1 *The proof is achieved by considering all possible queue evolution situations:*

(1) At $i = 0$, $Q(i) = 0$. This is due to our assumption that the sensor buffer is empty at the beginning. Then $Q(i+1) \leq A_{max}$. The upper bound holds.

(2) Suppose at time slot i, $Q(i) \leq V\Upsilon_{min}$, then at time slot $i+1$, $Q(i+1) \leq V\Upsilon_{min} + A_{max}$. The upper bound still holds.

(3) Suppose at time i, $V\Upsilon_{min} \leq Q(i) \leq V\Upsilon_{min} + A_{max}$. In this case, $s(i)$ is set to 1 by our algorithm, and $b(i)$ is determined by (2.24). Suppose $b(i) = b(i)^*$, and consider the assumption $V\Upsilon_{min} \leq Q(i)$, we have

$$b(i) \geq (1-\alpha)T[\frac{\Upsilon_{min}}{kn \cdot h(i)}]^{\frac{1}{(n-1)}} \, . \tag{2.29}$$

Condition (2.25) ensures the upper bound for this case.

Otherwise, suppose $b(i)$ equals $Q(i)$, then $Q(i+1) = A(i) \leq A_{max}$. The bound holds. Or suppose $b(i)$ equals $b_{max}(h(i))$, then condition (2.26) ensures the upper bound holds.

(4) Queue $Q(t)$ evolves from zero. Thus $Q(i)$ will not exceed $V\Upsilon_{min} + A_{max}$.

Theorem 1 indicates that, to combine opportunistic transmission and sleep scheduling for energy saving with a worst-case delay requirement, there exists some prerequisites. Condition (2.26) indicates that the maximum transmission amount, given any channel state, should be no less than the maximum data arrival amount. It is a necessary condition to guarantee a worst-case delay without data dropping for any algorithm. It is also a

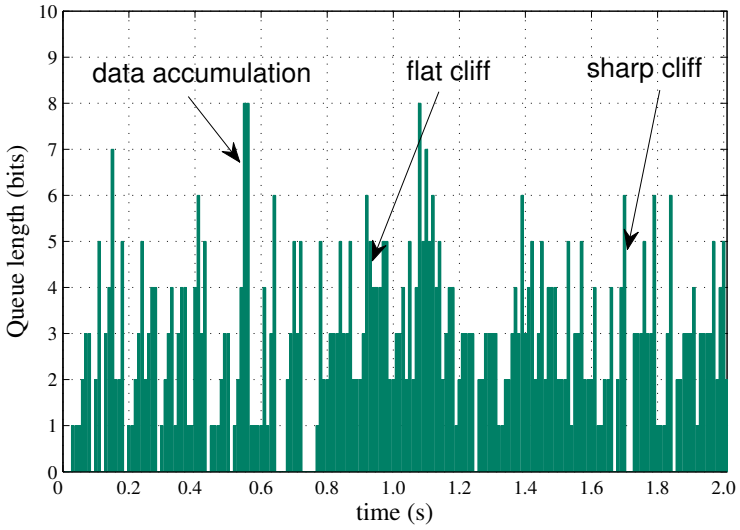


Figure 2.2: Queue evolutions

sufficient condition to prove such an algorithm exists. Given condition (2.26) is true, condition (2.25) is a sufficient condition under which our proposed algorithm can benefit from both sleeping mode and opportunistic transmission for energy saving, with a worst-case delay constraint.

theorem 2 *Given the minimum power consumption P^* that the system can achieve, the average power consumption of our proposed algorithm P_{ave} satisfies: $P_{ave} \leq P^* + C/V$, where C is a constant, at the cost of a worst-case delay increases within $O(V)$.*

The proof is similar to that in [66] using Lyapunov optimization formulation, and is omitted for brevity.

2.6 Numerical Results

In this section, numerical results are presented to demonstrate the performance of the proposed algorithm via simulations. Simulation setup is first presented, followed by discussion on the numerical results. In addition to the delay and average power consumption performance, we present a waking-up ratio, which is the fraction of time slots in which the sensor wakes up among the number of total time slots.

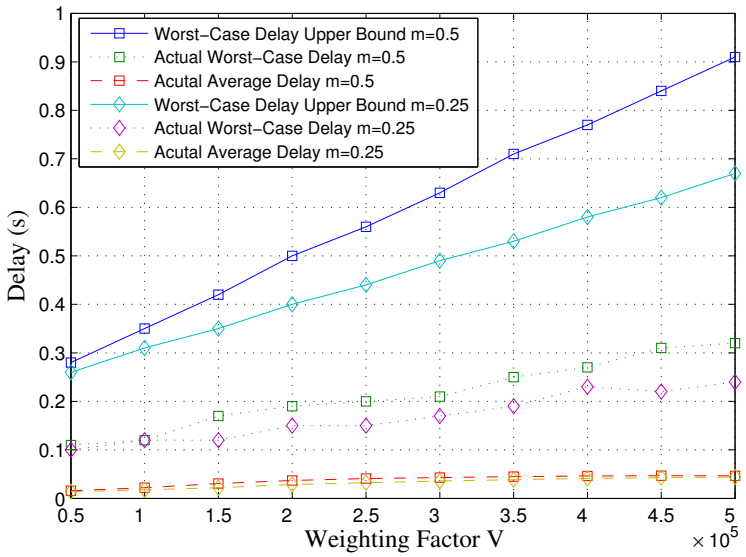


Figure 2.3: Impact of weighting factor on delay

For the channel model, we choose the model suggested by IEEE 802.15 task group 6 under the frequency band 2.4GHz [61]:

$$h(i)[dB] = a \times log_{10}(d) + b + N(i) \tag{2.30}$$

where a and b are scaling factors, N is a Gaussian random variable with zero mean and standard deviation σ_N, and d is the average distance between the sensor and the smart phone. In each time slot, a value is chosen for $N(i)$. The wearable sensor generates a data flow according to a Poisson process with average rate λ. Simulation parameters are chosen according to [20, 65, 61, 67, 68], given in Table 2.1. We choose a, b and σ_N based on the

Table 2.1: System Parameters

Parameter	Value	Parameter	Value
a	29.3	b	-16.8
σ_N	6.89	d	30 cm
α	1/8	n	2.67
T	10 ms	P_{max}	13 dBm
P_n	-120 dBm	λ	100 bps
k'	0.043		

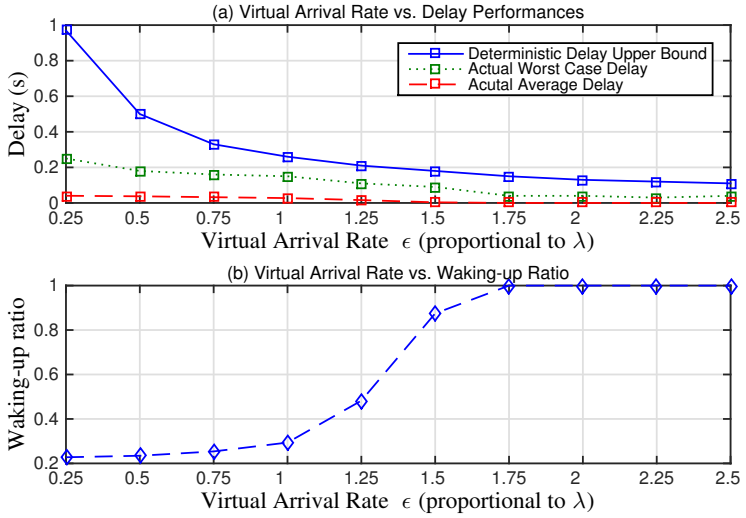


Figure 2.4: Impact of virtual arrival rate

body surface to body surface channel for 2.4GHz from the experiments in anechoic chamber [20]. The distance is chosen to be 30cm based on the assumption that the wearable sensor is on the left wrist and the smart phone is in the right hand side pocket. The slot duration is chosen to be 10ms for experiments in [61] show that channels are stable within a 5-10ms duration with a probability larger than 90% and less stable over 10ms. The peak power constraint is chosen to be 13dbm, whereas the noise power is assumed to be -120dbm. The power consumption parameters k' and n are chosen based on [65].

Fig. 2.2 shows the queue evolution at the sensor buffer under the proposed algorithm. It is observed that, when the queue is nearly empty, the sensor accumulates data, which saves the energy in a listening state, corresponding to the data accumulation in Fig. 2.2. When the summation of actual queue length and virtual queue length exceeds a threshold, the sensor wakes up and attempts to transmit. Our algorithm controls the sensor to transmit more data when the channel condition is good and less data when channel condition is poor, corresponding to the sharp cliff and flat cliff in Fig. 2.2, respectively.

The impacts of the system parameters, weighting factor V and the ratio of the power consumption in a listening state to the maximum transmission power m, on three delay metrics (worst-case delay upper bound, actual worst-case delay, and average delay) are shown in Fig. 2.3. As weighting factor V or ratio m increases, all the three delay metrics increase at different rates. The reason is that, when ratio m increases, the threshold

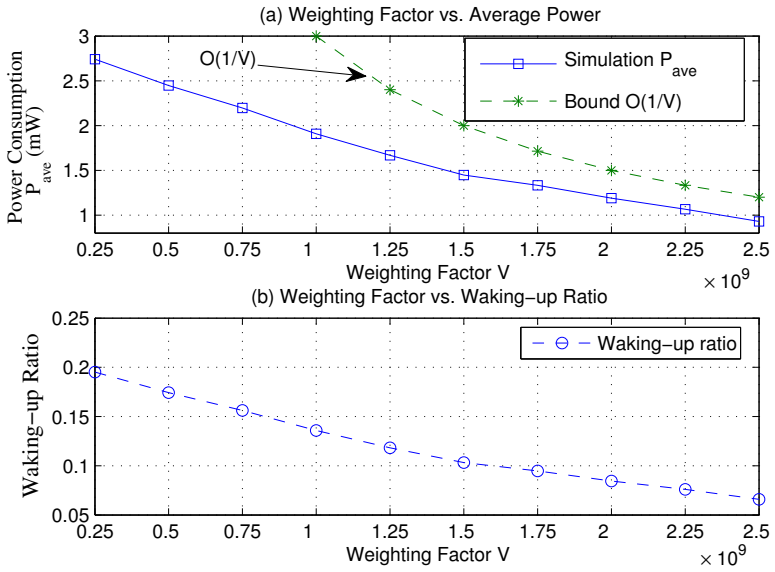


Figure 2.5: Impact of weighting factor on average power consumption

$V\Upsilon_{min}$ increases while the queue depleting rate remains the same. When weighting factor V increases, the threshold $V\Upsilon_{min}$ increases linearly with V. At the same time, the queue depleting rate decreases, at a speed slower than the order of $V^{(-\frac{1}{n-1})}$ as in (2.24). In both cases, the sensor accumulates more data before waking up, leading to a larger delay. The results in Fig. 2.3, which show that the increase rate of the actual worst-case delay can be bounded by a linear function of V, are consistent with Theorem 2.

The impacts of virtual arrival rate ε on the delay performance and waking-up ratio are shown in Fig. 2.4. It is clear that the delay metrics decrease as ε increases. When ε is larger, the virtual queue length increases faster while the threshold $V\Upsilon_{min}$ remains the same. Thus, the sensor wakes up more frequently, as shown in the waking-up ratio curve in Fig. 2.4 (b), leading to a smaller delay.

The average power consumption performance of our proposed algorithm is shown in Fig. 2.5. As the weighting factor V increases, the average power consumption P_{ave} and the waking-up ratio decrease. When weighting factor V increases, the worst-case delay that the sensor can tolerate increases. Thus, the buffer holds more data before the sensor wakes up. The reduction in waking-up ratio reduces the energy cost in a listening state. After the sensor wakes up, it has more data to transmit when the channel is in a good condition and can choose to transmit less otherwise, which utilizes the dynamics of channel

gain. The minimal power P^* the system can achieve is not given in this work. Instead, Fig. 2.5 show that P_{ave} is bounded by $O(\frac{1}{V})$. It is a sufficient condition to ensure that the inequality $P_{ave} \leq P^* + C/V$ holds. The results are consistent with Theorem 2.

2.7 Summary

In this chapter, we investigate the energy efficient power allocation problem with a worst-case delay requirement in a WBAN. A two-step power allocation algorithm is proposed based on the Lyapunov optimization formulation. At the first step, the sensor decides whether or not to wake up based on the current queue state and information of system statistics. If the sensor is awake, it first estimates the channel condition based on a pilot signal from the smart phone, then decides how much data it will transmit based on the information of queue state and current channel state. We show the conditions for our algorithm to have a worst-case delay limit. The tradeoff between energy consumption and delay is demonstrated in performance analysis. Numerical results indicate the effectiveness of our algorithm and the correctness of our analysis in two aspects: 1) The queue evolution shows the proposed algorithm can utilize sleep scheduling and opportunistic communication for energy saving; 2) A tradeoff $[O(V); O(\frac{1}{V})]$ between energy consumption and worst-case delay constraint is achieved.

Chapter 3

The Impacts of Peak Power Constraints and Statistical QoS Provisioning

In the previous chapter, we show that to provide the worst-case delay provisioning, the peak power is required to support the maximum data arrival rate given any channel state as shown in equation (2.26). Applications, such as EMG and EEG monitoring, consume larger data rate compared to vital signal monitoring [69]. When the data rate increases, the demand for peak power increases. For wearable sensors, a peak transmission power constraint is imposed to protect human skins from burns and irritations caused by overheated sensors. In this case, peak transmission power could be the bottleneck to support QoS requirements specified by EMG and EEG monitoring applications.

The peak transmission power constraint could reduce the energy efficiency, and lower the supported transmission rate. Take the water-filling scheme [24, 70] for example. The water-filling scheme, which is a threshold based policy, is proven to achieve maximal transmission rate under average transmission power constraint. When the channel gain is larger than a threshold, power is allocated based on the channel gain. The power allocated is a non-decreasing function of the channel gain. When the channel gain is smaller than the threshold, no power is allocated. The threshold, referred to as water-level, is calculated based on the average power constraint and the channel statistics. With a peak transmission power constraint, water-filling scheme may not feasible. Firstly, when channel gain is high, a transmitter may not have sufficient power to utilize the good channel. Secondly, since the transmission rate obtained from channels with high channel gain is reduced, to

support a targeted transmission rate, a transmitter is required to transmit over channels with low channel gain, leading to further energy efficiency reduction. For a wearable sensor with limited battery, it is desired to characterize the impacts of peak power constraints on the average transmission power consumption.

QoS provisioning in terms of statistical delay provisioning is desired for applications, such as EMG and EEG monitoring. These physiology information does not have the same priority as vital signals, yet a large delay could bring unsatisfactory user experiences. For example, EMG monitoring is utilized for gesture recognition, which can be used for interaction with computer [71]. The statistical delay requirement can be formulated using effective capacity [72]. Effective capacity of a service process is the maximum constant rate that the service process supports given a statistical delay requirement. Statistical delay provisioning has been shown to have significant impacts on energy efficiency, especially when the requirement becomes stringent [72]. Yet, the impacts of peak power constraints on the statistical delay provisioning has not been well understood.

Above facts motivate us to build qualitative and quantitive understanding of the impacts of peak transmission power constraint and statistical delay provisioning. Our contributions are three folds. Firstly, the optimal scheme for the power minimization problem under a peak power constraint is derived, and an efficient calculation method is proposed. Secondly, applying duality gap analysis, the impacts of the peak power constraint in terms of the upper bound of the extra average power consumption incurred compared to the water-filling scheme is characterized. We also propose a low complexity constant power allocation scheme that is suitable for wearable sensors. Through simulations, we validate the accuracy of the upper bound and show that the average power consumption of the constant power scheme is close to the optimal one. Thirdly, the impacts of statistical QoS provisioning is investigated. We show that peak power constraint is the bottleneck for wearable sensors to support stringent statistical QoS provisioning.

3.1 Literature Review

The transmission power allocation under average power constraints has been studied extensively. The optimal solution to the maximal transmission rate problem is the water-filling scheme [24]. [73] summarizes and generalizes water-filling strategies to a class of resource allocation problems, where multiple carriers are considered. Note that the water-filling scheme is not a closed form solution, which contains a summation over all possible channel states. The search for optimal water level could be too complex for wearable sensors to be solved in a real time manner. Exploring the fact that in large SNR regime, the power-rate

function is not sensitive to SNR, an interactive constant power based scheme is proposed and analyzed in [74].

Peak power constraint has been considered for transmission power allocation.[75] presents a geometric based approach to solve water-filling related problems. When the peak power constraint is imposed, the method in [75] can avoid the computing complexity in finding the optimal water level. This method depends on the formulation that average transmission power is known and the objective is to maximize transmission rate. However, when the transmission rate is a constraint and the objective is to minimize the average transmission power consumption, the geometric based approach can not be applied directly. Moreover, previous work has not investigated this question: given a peak power constraint, how much more average transmission power is required. In this work, we intend to design low complexity schemes for wearable sensors under the peak power constraint, and characterize the extra average power consumption incurred due to the peak power constraint.

Statistical delay provisioning has been studied for transmission power allocation utilizing effective capacity concept. The optimal scheme for maximizing the transmission rate subject to a given statistical QoS constraint is presented in [76]. [76] uses the formulation that average transmission power is given and the objective is to maximize effective capacity, thus the impact of statistical delay provisioning on peak power is not fully investigated. In this work, we focus on how statistical delay provisioning impacts the peak power and average power consumption.

3.2 System Model

Consider a link between a smart phone and a sensor. The sensor collects physiology signals, such as EMG and ECG, and transmits these information to the smart phone. The sensor has peak transmission power limit due to regulation. Let P_{max} denote the peak transmission power. The time domain of the system is slotted into duration with the same size T.

Consider the channel gain is identically and independently distributed over time slots. The channel power gain during the nth time slot is denoted by $v(n)$. We assume a finite state channel model with K channel states. Let $v_k, k \in \{1, 2, ..., K\}$ denote the channel gain in the kth state. Without loss of generality, we place the v_k in ascending order, namely $v_1 \leq v_2 \leq ... \leq v_K$. The probability of channel gain to be v_k is denoted by p_k with $\sum_{k=1}^{K} p_k = 1$. The transmission channel is modeled as follows. Let $Y(n), X(n)$ and $N(n)$ denote the received signal, transmitted signal and noise signal at the nth time slot,

we have

$$Y(n) = \sqrt{v(n)}X(n) + N(n). \tag{3.1}$$

The noise power at the receiver is denoted by σ^2.

Let $\bar{A}$ denote the targeted average traffic rate. Consider traffic with large arrival rate. Under this scenario, a wearable sensor does not have the luxury to be radio OFF for energy saving. Thus, a sensor is always radio ON. At the beginning of each time slot, the sensor obtains the channel condition through calculating RSSI of the pilot signal sent by the smart phone. The energy cost of acquiring channel gain is omitted in this formulation for the sensor is always radio ON. With channel gain information, the sensor decides how much power to spend during current time slot. Since arrival rate is large, we assume the buffer of sensor is always not empty. This approximation helps us focusing on understanding the impacts of peak transmission power. Without the consideration of buffer, channel state determines the power allocation decision.

Let the power allocated for channel gain v_k denoted by S_k. When transmission power is chosen to be S_k, $r(S_k)$ bits information can be reliably transmitted. The function $r(\cdot)$ is assumed to be a non-negative, increasing, strictly concave function, which is referred to as power-rate function. In this work, we adopt $r(S_k) = B\log_2(1 + S_k v_k/\sigma^2)$ [77]. Note that $\log(x)$ is a concave function with respect to x. Thus, $r(S_k)$ is a concave function with respect to S_k.

3.3 An Optimal Scheme under a Peak Power Constraint

In this section, we formulate the transmission power allocation under a peak power constraint problem as an optimization problem. Since the sensor has limited energy, the objective of the optimization problem is set to minimize average power consumption. We show that the formulated problem is a convex problem, and use Karush Kuhn Tucker (KKT) [78] method to obtain an optimal solution.

3.3.1 Problem Formulation

The objective of the problem is to minimizely the average transmission power. The average transmission power can be written as $\sum_{k=1}^{K} p_k S_k$. Thus, the objective is

$$\min_{S_k} \sum_{k=1}^{K} p_k S_k. \tag{3.2}$$

There are two constraints: one is on the average transmission rate, and the other is on the transmission power. First, the average transmission rate is required to be no less than the average data arrival rate $\bar{A}$. The average transmission rate is calculated over all possible channel states, namely $\sum_{k=1}^{K} p_k r(S_k)$. Thus, the constraint on transmission rate is

$$\sum_{k=1}^{K} p_k r(S_k) \geq \bar{A}. \tag{3.3}$$

The second constraint is on the transmission power S_k, which needs to be nonnegative and smaller than peak power P_{max}. We have

$$S_k \leq P_{max}, \text{ for } k \in \{1, .., K\}. \tag{3.4}$$

$$S_k \geq 0, \text{ for } k \in \{1, .., K\}. \tag{3.5}$$

Note that equation (3.4) and (3.5) each represents K constraints.

In summary, objective (3.2) and constraints (3.3), (3.4) and (3.5) constitute our optimization problem. Note that, we assume $\sum_{k=1}^{K} p_k r(P_{max}) > \bar{A}$. The assumption ensures that there exists feasible solutions. Specifically, if the sensor uses the peak transmission power over all channels, the sensor can provide the traffic rate as required. Without this assumption, the formulated problem may be infeasible.

3.3.2 Problem Transformation Via KKT

In this subsection, we show that the formulated problem is a convex optimization problem and adopt KKT method to transform the formulation. To start with, we show that the formulated problem is a convex optimization problem. Firstly, the objective function, $\sum_{k=1}^{K} p_k S_k$, is a piece-wise function with respect to control variable S_k, thus a convex function.

Transform the constraints (3.3), (3.4) and (3.5) in the form of $f(x) \leq 0$ as follows:

$$-\sum_{k=1}^{K} p_k r(S_k) + \bar{A} \leq 0. \tag{3.6}$$

$$S_k - P_{max} \leq 0, \text{ for } k \in 1, .., K. \tag{3.7}$$

$$-S_k \leq 0, \text{ for } k \in 1, .., K. \tag{3.8}$$

We show the left hand sides of constraints (3.6,3.7,3.8) are convex functions as follows. Since $r(S_k)$ is a concave function with respect to S_k, $-\sum_{k=1}^{K} p_k r(S_k) + \bar{A}$ in (3.6) is convex with respect to S_k. Note that a linear function is a convex function, thus $S_k - P_{max}$ in (3.7) and $-S_k$ in (3.8) are convex functions with respect to S_k.

The objective function and constraints are differentiable with respect to S_k. Based on above convex and differentiable characteristics, we adopt KKT conditions to derive an optimal solution.

Introducing Lagrange multiplier λ for constraint (3.6), multiplier μ_k, $k = 1, 2, ..., K$ for (3.7) and multiple η_k, $k = 1, 2, ..., K$ for (3.8). The Lagrangian is

$$L(S_k, \lambda, \mu_k, \eta_k) = \sum_{k=1}^{K} p_k S_k + \sum_{k=1}^{K} \eta_k (S_k - P_{max}) + \sum_{k=1}^{K} \mu_k (-S_k) + \lambda(-\sum_{k=1}^{K} p_k r(S_k) + \bar{A}). \tag{3.9}$$

The KKT conditions are constituted of stationarity, primal feasibility, dual feasibility and complementary slackness [78]. The stationarity condition requires $\frac{\partial L}{\partial S_k} = 0$. Substitute L with equation (3.9), we obtain

$$p_k - \mu_k + \eta_k - \lambda p_k \frac{\partial r(S_k)}{\partial S_k} = 0. \tag{3.10}$$

Since $\frac{\partial r(S_k)}{\partial S_k} = \frac{\lambda p_k}{\frac{\sigma^2}{v_k} + S_k} \frac{1}{\ln 2}$ when $v_k \neq 0$, we have

$$p_k - \mu_k + \eta_k = \frac{\lambda p_k}{\frac{\sigma^2}{v_k} + S_k} \frac{1}{\ln 2} \tag{3.11}$$

Transform equation (3.11), the transmission power that satisfies the stationarity condition can be represented by

$$S_k = \frac{\lambda p_k}{p_k - \mu_k + \eta_k} \frac{1}{\ln 2} - \frac{\sigma^2}{v_k}, \tag{3.12}$$

when $p_k - \mu_k + \eta_k \neq 0$.

Based on dual feasibility, all multipliers should be nonnegative. We have

$$\mu_k \geq 0, \text{ for } k \in \{1, ..., K\} \tag{3.13}$$

$$\eta_k \geq 0, \text{ for } k \in \{1, ..., K\} \tag{3.14}$$

$$\lambda \geq 0 \tag{3.15}$$

Based on complementary slackness, we have

$$\mu_k(-S_k) = 0, \text{ for } k \in \{1, ..., K\} \tag{3.16}$$

$$\eta_k(S_k - P_{max}) = 0, \text{ for } k \in \{1, ..., K\} \tag{3.17}$$

$$\lambda[-\sum_{k=1}^{K} p_k \log(1 + \frac{S_k v_k}{\sigma^2}) + \bar{A}] = 0 \tag{3.18}$$

Solve equations (3.12) to (3.18) for S_k, μ_k and η_k for $k \in \{1, ..., K\}$ and λ. Note that there are $3K + 1$ unknown variables, and $3K + 1$ equations by (3.12), and (3.16) to (3.18). Inequalities (3.13), (3.14) and (3.15) constitute $2K + 1$ constraints which regulate the feasible region.

3.3.3 The Optimal Solution and Calculation Method

In this section, we derive the optimal solution to the transformed problem. We first solve equations (3.12) to (3.18). The solution, containing summation operator, is not a closed form solution. We then present an efficient calculation method.

The Optimal Structure

To start with, we argue that the optimal solution is achieved when $\sum_{k=1}^{K} p_k r(S_k) = \bar{A}$.

$$\sum_{k=1}^{K} p_k r(S_k) = \bar{A}. \tag{3.19}$$

The reason is that, based on the power-rate function, the required power is an increasing function of transmission rate. Thus, the minimal transmission power is achieved when the supported transmission rate equals to target rate.

To obtain S_k, λ needs to be calculated first. Replace S_k in equation (3.19) with (3.12), we have

$$\sum_{k=1}^{K} p_k \log_2[\frac{v_k}{\sigma^2 \ln 2} \frac{\lambda p_k}{(p_k - \mu_k + \eta_k)}] = \bar{A}. \tag{3.20}$$

Equation (3.20) contains $2K$ unknown parameters, namely μ_k and η_k, $k = 1, ..., K$. Next, we use dual feasibility equations (3.13)-(3.14), and complementary slackness equations (3.16)-(3.17) to remove μ_k and η_k.

The strategy to remove μ_k is presented as follows. According to (3.16), μ_k and S_k cannot be nonzero for the same k. When $S_k = 0$, a sensor refrains from transmission, and does not contribute to the transmission rate. When $S_k \neq 0$, we have $\mu_k = 0$. In this case, the item μ_k can be removed from (3.20). Since a sensor may refrain from transmission when channel gain is small, and v_k is placed in ascending order, we assume a sensor starts transmission if channel gain is no less than v_{k_s}. We refer to v_{k_s} as cutoff threshold. As a result, we can simplify equation (3.20) as

$$\sum_{k=k_s}^{K} p_k \log_2[\frac{v_k}{\sigma^2 \ln 2} \frac{\lambda p_k}{(p_k + \eta_k)}] = \bar{A}. \tag{3.21}$$

With above simplification, K unknown parameters are replaced by a single parameter $k_s \in [1, K]$.

The strategy to remove η_k is presented as follows. According to (3.17), η_k and $S_k - P_{max}$ cannot be nonzero for the same k. When $S_k = P_{max}$, a sensor uses constant power for transmission. In this case, the transmission rate equals to $\log_2(1 + \frac{P_{max} v_k}{\sigma^2})$. When $S_k < P_{max}$, we have $\eta_k = 0$. Assume a sensor uses peak transmission power if channel gain is no less than v_{k_m}. v_{k_m} is referred to as Max-on threshold. Thus, we can simplify equation (3.21) as

$$\sum_{k=k_s}^{k_m-1} p_k \log_2[\frac{v_k \lambda}{\sigma^2 \ln 2}] + \sum_{k=k_m}^{K} p_k \log_2(1 + \frac{P_{max} v_k}{\sigma^2}) = \bar{A}. \tag{3.22}$$

With above simplification, K unknown parameters are replaced by a single parameter $k_m \in [2, K]$.

In summary, the optimal power allocation structure is given by:

$$S_k = \begin{cases} 0, & \text{if } v_k < v_{k_s} \\ \frac{\lambda}{\ln 2} - \frac{\sigma^2}{v_k}, & \text{if } v_{k_s} \leq v_k < v_{k_m} \\ P_{max}, & \text{if } v_k \geq v_{k_m}, \end{cases} \tag{3.23}$$

where the multiplier λ, the cutoff threshold v_{k_s} and the Max-on threshold v_{k_m} need to satisfy the average transmission rate condition (3.22). We will present an algorithm to obtain the value for them in next section.

The solution presented in equation (3.23) is segmented into three sections. The first segment points out under what condition the sensor does not transmit. The third segment illustrates under what condition the sensor transmits with peak power. The second segment illustrates how much power a sensor uses otherwise. The second segment has a water-filling structure, where the transmission power is determined by the gap between a water level $\frac{\lambda}{\ln 2}$ and an item reflects the channel gain $\frac{\sigma^2}{v_k}$.

When P_{max} is large enough that S_k is always smaller than P_{max}. In this case, $S_k - P_{max} \neq 0$. Based on equation (3.17), $\eta_k = 0$, for $k \in \{1, ..., K\}$. As a result, the transmission power in equation (3.12) can be written as

$$S_k = \frac{\lambda}{1 - \frac{\mu_k}{p_k}} \frac{1}{\ln 2} - \frac{\sigma^2}{v_k}. \tag{3.24}$$

Consider $S_k \mu_k = 0$, we have

$$S_k = \begin{cases} 0, & \text{if } v_k < v_{k_s} \\ \frac{\lambda}{\ln 2} - \frac{\sigma^2}{v_k}, & \text{if } v_{k_s} \leq v_k. \end{cases} \tag{3.25}$$

Equation (3.25) can be considered as the solution to the problem, which has no constraint on the peak transmission power. As expected, it is the same as the water filling scheme.

The structure (3.23) and condition (3.22) are insufficient to produce a unique solution. The relationship between the water level λ and the max-on threshold v_{k_m} can be exploited to produce a unique solution. First, for the transmission power in the second segment, its value should be smaller than P_{max}. Thus we have

$$\frac{\lambda}{\ln 2} - \frac{\sigma^2}{v_k} < P_{max}, \text{for } k = k_s, ..., k_{m-1}. \tag{3.26}$$

Since left hand side of inequality (3.26) is an increasing function of v_k, we have

$$\frac{\lambda}{\ln 2} - \frac{\sigma^2}{v_{k_{m-1}}} < P_{max}. \tag{3.27}$$

Moreover, if the gap structure of the second segment is used to determine the transmission power for channel with channel gain no less than v_{k_m}, the transmission power will

be larger than P_{max}. Based on this fact, we have

$$\frac{\lambda}{\ln 2} - \frac{\sigma^2}{v_{k_m}} \geq P_{max}. \tag{3.28}$$

The combination of inequalities (3.27), (3.28), and condition (3.22) provides a unique solution with structure (3.23).

Calculation Method

In the following, we present an efficient algorithm to obtain the multiplier λ, the cutoff threshold v_{k_s} and the Max-on threshold v_{k_m}. We exploit the following characteristics: 1) the average transmission rate constraint, namely equation (3.22), is to be satisfied; 2) the water level should satisfy (3.27) and (3.28). The algorithm first determines the Max-on threshold v_{k_m}, and then determines the multiplier λ. Note that, when the multiplier is set, the cutoff threshold v_{k_s} is determined due to the constraint that transmission power is non-negative.

The search for Max-on threshold v_{k_m} is as follows. Set search range as $[v_s, v_e]$, where $v_s = v_1$ and $v_e = v_K$. Set $v_{k_m} = v_t$ with $v_t = \lfloor (v_s + v_e)/2 \rfloor_v$, where $\lfloor x \rfloor_v$ is the largest number in sequence v_k smaller than x. Then the transmission power S_k for all $k \geq t$ is set to be P_{max}. We set the multiplier as $\lambda_t^{max} = \ln 2(P_{max} + \frac{\sigma^2}{v_{t-1}})$ and $\lambda_t^{min} = \ln 2(P_{max} + \frac{\sigma^2}{v_t})$, respectively. If inequality (3.27) is satisfied, when the multiplier is λ_t^{max}, the provided average transmission rate should be no less than $\bar{A}$, namely,

$$\sum_{k=1}^{K} p_k r(S_{v_t}^{\lambda_t^{max}}) \geq \bar{A}, \tag{3.29}$$

where $S_{v_t}^{\lambda_t^{max}}$ is the transmission power scheme with multiplier λ_t^{max}. If (3.29) does not hold, it suggests more channels need to be utilized to support the target transmission rate $\bar{A}$. In this case, we can set $v_s = v_t$ and test again.

When (3.29) is satisfied, we need to check inequality (3.28). If (3.28) is satisfied, when the multiplier is λ_t^{min}, the provided average transmission rate should be less than $\bar{A}$, namely,

$$\sum_{k=1}^{K} p_k r(S_{v_t}^{\lambda_t^{min}}) < \bar{A}, \tag{3.30}$$

where $S_{v_t}^{\lambda_t^{min}}$ is the transmission power scheme with multiplier λ_t^{min}. If (3.30) does not hold, it suggests current allocation scheme produces more transmission rate than needed. As a

result, peak power should be allocated to less channels. In this case, we can set $v_e = v_t$ and test again. Above procedures are summarized in Algorithm 1.

Algorithm 1: Determine the Max-on threshold

(1) Set search range as $[v_s, v_e]$, where $v_s = v_K$ and $v_e = v_1$.
(2) Set $v_{k_m} = v_t$ with $v_t = \lfloor (v_s + v_e)/2 \rfloor_v$, where $\lfloor x \rfloor_v$ is the largest number in sequence v_k smaller than x.
(3) Set the multiplier as $\lambda_t^{max} = \ln 2(P_{max} + \frac{\sigma^2}{v_{t-1}})$ and check whether (3.29) holds.
(4) If (3.29) does not hold, update $v_s = v_t$ and return to step (2). Otherwise, set the multiplier as $\lambda_t^{min} = \ln 2(P_{max} + \frac{\sigma^2}{v_t})$, and check whether (3.30) holds.
(5) If (3.30) does not hold, update $v_e = v_t$ and return to step (2). Otherwise, end search and output v_t.

The complexity of algorithm 1 consists of two factors: one is searching for next possible Max-on threshold, and the other is judging whether criteria are met. Searching part shares the structure of binary search, thus is logarithmic efficient with respect to the number of channel states K, whereas judging part calculates the summation of the transmission rate contributed by each channel, thus is linearly efficient with respect to the number of channel states K. In summary, the algorithm executes in $\mathrm{O}(K \log K)$ time.

So far, we have already obtained the optimal Max-on threshold v_{k_m}. The remaining problem of determining multiplier λ is the same as the water filling scheme. In this case, only those channels with channel gain smaller than v_{k_m} are considered. The required transmission rate is $\bar{A} - \sum_{k=k_m}^{K} p_k \log_2(1 + \frac{P_{max} v_k}{\sigma^2})$.

3.4 Performance Analysis via a Constant Power Scheme

In the section, we present the analysis on the impacts of the peak power constraint under the convex optimization framework. Specifically, we first characterize the upper bound of the gap between the average power of any power allocation strategy and of the optimal water filling scheme via duality gap analysis. The gap is referred to as average power loss compared to water filling scheme. We then propose a constant power scheme given a peak power constraint. Since it is suboptimal compared to the optimal scheme (3.23), the upper bound of average power loss of the constant power scheme is also the upper bound of the average power loss of the optimal scheme (3.23).

3.4.1 Duality Gap Analysis

In this subsection, we characterize the gap between the average power of any power allocation strategy and of the optimal water filling strategy via the duality gap analysis.

The duality gap, denoted by Γ, is the difference between the primal objective and the dual objective. Here, the problem under investigation is the problem that gives water filling scheme. It is the problem with objective (3.2) and constraints (3.6) and (3.8). So the primal objective is $\sum_{k=1}^{K} p_k S_k$. The dual objective is given by replacing the stationary condition into the Lagrangian equation. The Lagrangian equation and the stationary condition can be obtained by setting $\eta_k = 0$ in (3.9) and (3.10), respectively. Then the dual objective can be written as

$$g(\lambda, \mu_k) = \sum_{k=1}^{K} \left(\frac{\lambda}{1 - \frac{\mu_k}{p_k}} \frac{1}{\ln 2} - \frac{\sigma^2}{v_k}\right)(p_k - \mu_k) + \lambda\left(-\sum_{k=1}^{K} p_k \log_2\left(\frac{\lambda}{1 - \frac{\mu_k}{p_k}} \frac{1}{\ln 2} \frac{v_k}{\sigma^2}\right) + \bar{A}\right), \quad (3.31)$$

where $g(\lambda, \mu_k)$ is the dual objective function. Note that solving μ_k and λ to obtain the lowest dual objective is equivalent to solving the original optimization problem. To find a simple bound, our goal is to represent the duality gap Γ with power allocation decision S_k. Γ is calculated as $\sum_{k=1}^{K} p_k S_k - g(\lambda, \mu_k)$. Assuming that the allocation strategy satisfies the transmission rate constraint with equality, namely $-\sum_{k=1}^{K} p_k \log_2(1 + S_k v_k/\sigma^2) + \bar{A} = 0$. Thus, we have

$$\begin{aligned} \Gamma &= \sum_{k=1}^{K} S_k \mu_k \\ &= \sum_{k=1}^{K} S_k p_k \frac{(S_k + \frac{\sigma^2}{v_k}) \ln 2 - \lambda}{(S_k + \frac{\sigma^2}{v_k}) \ln 2}. \end{aligned} \quad (3.32)$$

The dual gap Γ by equation (3.32) is a function of λ, where λ needs to be determined. Since larger λ produces smaller Γ, a large λ is desired. At the same time, to satisfy dual feasibility, we have $\mu_k \geq 0$, namely $p_k \frac{(S_k + \frac{\sigma^2}{v_k}) \ln 2 - \lambda}{(S_k + \frac{\sigma^2}{v_k}) \ln 2} \geq 0$. As a result, λ should be no larger than $(S_k + \frac{\sigma^2}{v_k}) \ln 2$. Thus, the largest λ is

$$\lambda = \min_k (S_k + \frac{\sigma^2}{v_k}) \ln 2. \quad (3.33)$$

Substituting the λ in equation (3.32) with equation (3.33) gives the following:

$$\Gamma = \sum_{k=1}^{K} S_k p_k [1 - \frac{\min_k (S_k + \frac{\sigma^2}{v_k})}{S_k + \frac{\sigma^2}{v_k}}]. \quad (3.34)$$

theorem 3 *For the minimization average transmission power problem, if the transmission rate constraint is satisfied with equality, then the average transmission power using S_k is at most Γ w away from the optimal water filling solution, where Γ is expressed in (3.34).*

Note that when the water filling scheme is used, $S_k + \frac{\sigma^2}{v_k}$ is a constant for channels that are chosen for transmission. In this case, $\Gamma = 0$. The optimal scheme (3.23) is not a closed formed solution. Thus, replacing S_k in (3.34) with (3.23) does not give intuitive explanations on how P_{max} influences the gap Γ. We address this issue via a constant power scheme in the following.

3.4.2 The Upper Bound via a Constant Power Scheme

In this subsection, we propose a constant power allocation scheme under a peak power constraint. The upper bound of the gap between the constant power scheme and the water filling scheme is analyzed based on theorem 3. Since the constant power is suboptimal compare to the optimal scheme (3.23), the upper bound also holds for the optimal scheme.

We turn to constant power allocation schemes not only for the convenience on analysis, but also for the fact that they are appealing to wearable sensors. As pointed out in [74], constant power allocation schemes can significantly lower the computation complexity and simplify transmitter design.

A constant power allocation scheme is designed as such: the sensor uses constant power S_c for transmission if the channel state is better than a cutoff threshold v_{k_c}; otherwise, the sensor does not transmit. It can be written as

$$S_k = \begin{cases} 0, & \text{if } v_k < v_{k_c} \\ S_c, & \text{if } v_{k_c} \leq v_k. \end{cases} \tag{3.35}$$

In [74], the cutoff threshold is chosen the same as that of water filling schemes. In this work, to provide a target transmission rate, v_{k_c} is chosen such that its index k_c satisfies $\sum_{k=k_c}^{K} p_k \log_2(1 + \frac{S_c v_k}{\sigma^2}) = \bar{A}$.

Note that, the power rate function with form $\log_2(1 + SNR)$ is sensitive to low SNR. Thus, for power saving purpose, it is desired for a sensor refrained from transmiting over the channels with low channel gain, especially those channels are not utilized by the water filling schemes. When the peak power constraint P_{max} is smaller than the water level, the best strategy is to avoid channels with low channel gain is to use the peak power for transmission. Thus, we set $S_c = P_{max}$.

Substituting the constant power scheme (3.35) into the dual gap equation (3.33) gives

$$\begin{aligned}\Gamma &= \sum_{k=k_c}^{K} P_{max} p_k [1 - \frac{\min_k (P_{max} + \frac{\sigma^2}{v_k})}{P_{max} + \frac{\sigma^2}{v_k}}] \\ &= \sum_{k=k_c}^{K} P_{max} p_k [\frac{\frac{\sigma^2}{v_k} - \min_k (\frac{\sigma^2}{v_k})}{P_{max} + \frac{\sigma^2}{v_k}}] \\ &= \sum_{k=k_c}^{K} p_k (\frac{P_{max}}{P_{max} + \frac{\sigma^2}{v_k}})(\frac{\sigma^2}{v_k} - \frac{\sigma^2}{\max_k (v_k)}).\end{aligned} \tag{3.36}$$

Equation (3.36) characterizes the upper bound of the gap between the proposed average transmission power of a constant power allocation scheme with a peak power constraint and the water filling scheme without peak power constraint. Since $\frac{P_{max}}{P_{max} + \frac{\sigma^2}{v_k}} < 1$, we have $\Gamma < \sum_{k=k_c}^{K} p_k (\frac{\sigma^2}{v_k} - \frac{\sigma^2}{\max_k (v_k)})$. We can see that the smaller the k_c, the larger the upper bound for more items are summed. Moreover, since $\frac{\sigma^2}{v_k} - \frac{\sigma^2}{\max_k (v_k)}$ increases with decreases of v_k, the right hand side of the inequality increases faster when the channel gain v_k is smaller.

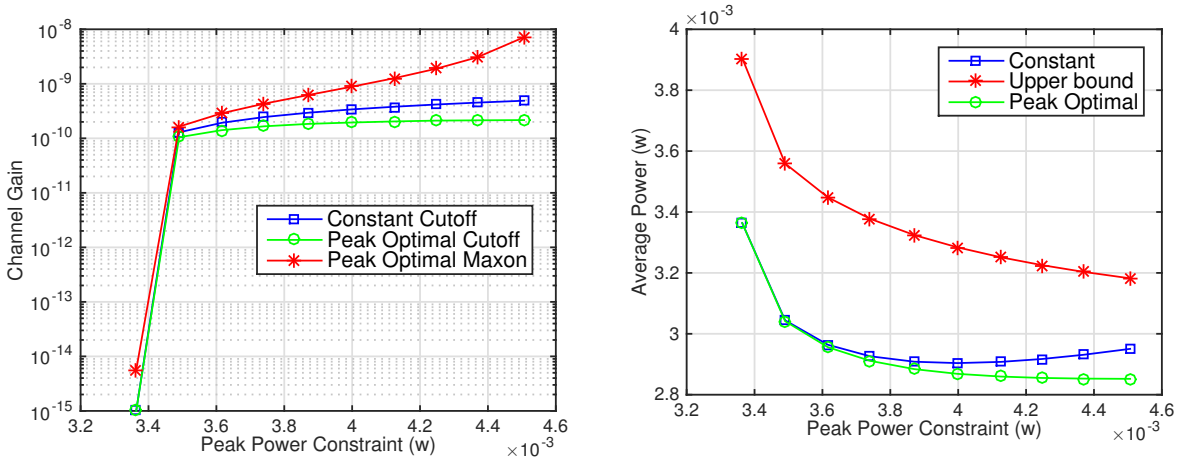
3.4.3 Numerical Results

In this subsection, we present the numerical results using body area channel to show: 1) the proposed constant power scheme is close to the optimal scheme (3.23) in terms of average transmission power consumption, especially when the peak power constraint is stringent; 2) the impacts of P_{max} on average transmission power consumption.

For the channel model, we choose the same model used in previous chapter from the IEEE 802.15 task group 6 [61]:

$$v(d) = -(a \times \log_{10}(d) + b + N) \tag{3.37}$$

where a and b are scaling factors, N is a Gaussian random variable with zero mean and standard deviation σ_N, and d is the direct distance between the sensor and smart phone. The target transmission rate A is set to 2.5Mbps. The range of P_{max} is set as follows. The upper limit is chosen as the maximal power used in the water filling scheme. When the peak power constraint is larger than the upper limit, the water filling scheme can be used. The lower limit is set as the minimal constant power needed when all channels are used.



(a) Cutoff Threshold

(b) Average Power

Figure 3.1: Constant Power vs Optimal Scheme with a Peak Power Constraint

When the peak power constraint is smaller than the lower limit, no constant power schemes can provide required transmission rate. Remaining simulation parameters are chosen the same as previous chapter given in Table 2.1.

Constant Power vs Optimal Scheme (3.23)

We compare the average transmission power consumption and the thresholds of the constant power and the optimal scheme (3.23) in Fig. 3.1.

Fig. 3.1(a) shows the cutoff thresholds for the constant power scheme and the optimal scheme, and the Max-on threshold for the optimal scheme. From the Fig. 3.1(a), firstly, we can see that the cutoff thresholds of constant power scheme are larger than the optimal scheme, namely the optimal scheme utilizes channels with smaller channel gain for transmission. The reason is that, the constant scheme always uses the peak transmission power, and thus has less channel usage. Secondly, the Max-on thresholds of the optimal scheme are larger than the cutoff thresholds of the constant power scheme. With the decreases of the peak power, the differences between the cutoff thresholds of the constant scheme and the Max-on thresholds of the optimal scheme decrease. The reason is that, when peak power is reduced, to support a target transmission rate, peak power is used for more channels with small channel gain, reducing the differences between constant power scheme and the optimal scheme.

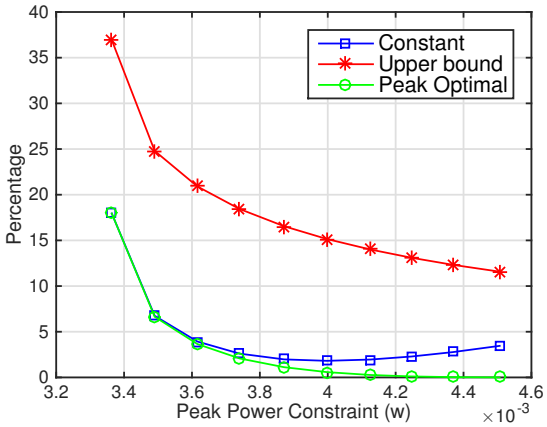


Figure 3.2: Impacts of Peak Power Constraint on Average Power

Fig. 3.1(b) shows the average transmission power of the constant power scheme, the optimal scheme and the upper bound. From the Fig. 3.1(b), we can see that with the decrease of the peak power, the average transmission power of the constant scheme and the optimal scheme increases, and the differences between the constant scheme and the optimal scheme decrease. The reason is that, with the decrease of peak power, more channels with poor channel condition are utilized, leading to the increase of average transmission power. And with the decrease of peak power, the optimal scheme uses peak power for transmission over more channels. As a result, the average transmission power of the optimal scheme is closer to that of the constant power scheme. This result suggests that when the peak power constraint is stringent, constant power scheme is a good choice for wearable sensors for its performance is close to the optimal one.

Impacts of Peak Power Constraint

We show the impacts of peak power constraint via the extra average power required compared to the water filling scheme in Fig. 3.2. As we can see from Fig. 3.2, as the peak power decreases, the extra average power required increases. Specifically, for the optimal scheme, with 1 mw decrease in peak power, the percentage increases from zero to about 18%, whereas for the constant power scheme, the percentage increases from about 4% to about 18%. Moreover, when the peak power is close to the upper limit, the extra average power required increases at a slow pace, whereas when the peak power is close to the lower limit, the average power increases at a fast pace. The reason is that, when approaching the

lower limit, channels with low channel gain are to be utilized. Compare to channels with high channel gain, channels with low channel gain require more power to generate the same transmission rate. This result suggests that peak power constraint could cause significant average power consumption increase, especially for applications which have utilized most of channels for transmission.

Accuracy of the Upper Bound

The accuracy of the upper bound can be determined from Fig. 3.1(b) and Fig. 3.2. Firstly, the upper bound can reflect the increase of average transmission power as a result of the decrease of peak power. Specially, the upper bound can reflect the trend that when the peak power is close to lower limit, the average transmission power increases at a fast pace. Secondly, the upper bound applies to both the constant scheme and the optimal scheme. The reason is that, the upper bound is derived based on constant scheme, which has larger average power consumption than the optimal scheme. Thus, the upper bound also applies to the optimal scheme. Thirdly, when compared to the average power of the water filling scheme, the estimation error ranges from 10% to 20%. Above results suggest that the upper bound we derived could be served as a guideline when designing transmission power allocation schemes for wearable sensors.

3.5 The Impact of Statistical QoS Provisioning

In this section, we study the impacts of statistical QoS provisioning on peak power and average power consumption. The QoS provisioning in this section refers to delay performance provisioning. We first review the concept of statistical QoS provisioning, effective capacity, and the optimal transmission power allocation scheme. Through simulations, we show that peak power could be the bottleneck for wearable sensors to support stringent statistical QoS provisioning. In the end, we discuss the possibility of reducing peak power given statistical QoS provisioning requirements.

3.5.1 Statistical QoS and Effective Capacity

Statistical QoS provisioning means the tail distribution of the delay random process can be bounded by an exponential distribution. For a dynamic queueing system with stationary

and ergodic arrival and service processes, the queue length $Q(t)$ converges such that

$$-\lim_{x\to\infty}\frac{\log(\Pr\{Q(\infty)>x\})}{x}=\theta, \tag{3.38}$$

where the parameter θ, a positive value, represents the exponential decay rate of the tail distribution [79]. The larger the value of θ, the more stringent is the delay requirement. In particular, when $\theta\to 0$, the system can be considered to have no delay constraint, as the problem investigated in previous section. On the other hand, when $\theta\to\infty$, any delay is intolerable, meaning traffic needs to be transmitted once it arrives. In literature, the exponential decay rate θ is referred to as QoS exponent [80].

Effective capacity is proposed to evaluate the maximum constant arrival rate that a service process can support given a statistical QoS requirement specified by θ [81, 82]. We present the definition of effective capacity according to [83, 80]. Let $R[i], i=1,2,...$ be a discrete time stationary and ergodic stochastic service process and $R_s[t]\triangleq\sum_{i=1}^{t}R[i]$ be the partial sum of the service process $R[i]$. The effective capacity of the service process, denoted by $E_c(\theta)$ is

$$E_c(\theta)=-\lim_{t\to\infty}\frac{\log(\mathbb{E}\{e^{-\theta R_s[t]}\})}{\theta t}. \tag{3.39}$$

When the service process is an uncorrelated process, the effective capacity $E_c(\theta)$ can be simplified to

$$E_c(\theta)=-\frac{\log(\mathbb{E}\{e^{-\theta R[i]}\})}{\theta}. \tag{3.40}$$

To utilize the concept of effective capacity to characterize the statistical QoS requirement, we replace the average transmission rate constraint (3.3) by an effective capacity constraint as

$$-\frac{\log(\mathbb{E}\{e^{-\theta r[i]}\})}{\theta}\geq\bar{A}. \tag{3.41}$$

The expectation can be written as $\mathbb{E}\{e^{-\theta r(i)}\}=\sum_{k=1}^{K}p_k e^{-\theta r(S_k)}$ under our finite state channel model. Transform the constraints (3.41) in the form of $f(x)\leq 0$ as

$$\sum_{k=1}^{K}p_k e^{-\theta r(S_k)}-e^{-\theta\bar{A}}\leq 0. \tag{3.42}$$

Given a positive QoS exponent θ, the exponent of the first item $-\theta r(S_k)$ is a convex function with respect to transmission power S_k, thus the left hand side of (3.42) is a log-convex function, also a convex function. As a result, we can apply KKT to obtain the optimal solution.

3.5.2 Power Allocation Schemes with Statistical QoS Provisioning

We derive the optimal transmission power allocation scheme with statistical QoS provisioning based on the above problem formulation. The derivation based on a KKT approach.

The Lagrangian changes to

$$\begin{aligned} L(S_k, \lambda, \mu_k, \eta_k, \theta) = &\sum_{k=1}^{K} p_k S_k + \sum_{k=1}^{K} \eta_k (S_k - P_{max}) \\ &+ \sum_{k=1}^{K} \mu_k (-S_k) + \lambda (\sum_{k=1}^{K} p_k e^{-\theta r(S_k)} - e^{-\theta \bar{A}}), \end{aligned} \tag{3.43}$$

According to the stationarity condition $\frac{\partial L}{\partial S_k} = 0$, we have

$$p_k - \mu_k + \eta_k + \lambda p_k (\frac{\partial e^{-\theta r_e(S_k)}}{\partial S_k}) = 0, \tag{3.44}$$

where

$$\frac{\partial e^{-\theta r_e(S_k)}}{\partial S_k} = (-\theta \ln 2)(\frac{v_k}{\sigma^2})(1 + \frac{S_k v_k}{\sigma^2})^{(-\theta \ln 2 - 1)}. \tag{3.45}$$

We use β to denote $\theta \ln 2$, and γ_k to denote $\frac{v_k}{\sigma^2}$ for simplicity. Thus, equation (3.44) can be written as

$$p_k - \mu_k + \eta_k - \lambda p_k \beta \gamma_k (1 + S_k \gamma_k)^{(-\beta - 1)} = 0. \tag{3.46}$$

Equation (3.46), together with complementary slackness and dual feasibility, can be used to obtain the optimal transmission power allocation scheme with statistical QoS provisioning.

The Impacts of Statistical QoS Provisioning

We first show the impacts of statistical QoS provisioning on average transmission power consumption. To do so, we derive the optimal transmission strategy for the QoS provisioning problem without peak power constraints. In other word, we consider $P_{max} \to \infty$, namely the multiplier $\eta_k = 0$ for $k = 1, ..., K$.

$$S_k = \begin{cases} 0, & \text{if } v_k < v_{k_s} \\ \frac{\gamma_0^{\frac{1}{\beta+1}}}{\gamma_k^{\frac{\beta}{\beta+1}}} - \frac{1}{\gamma_k}, & \text{if } v_{k_s} \le v_k. \end{cases} \tag{3.47}$$

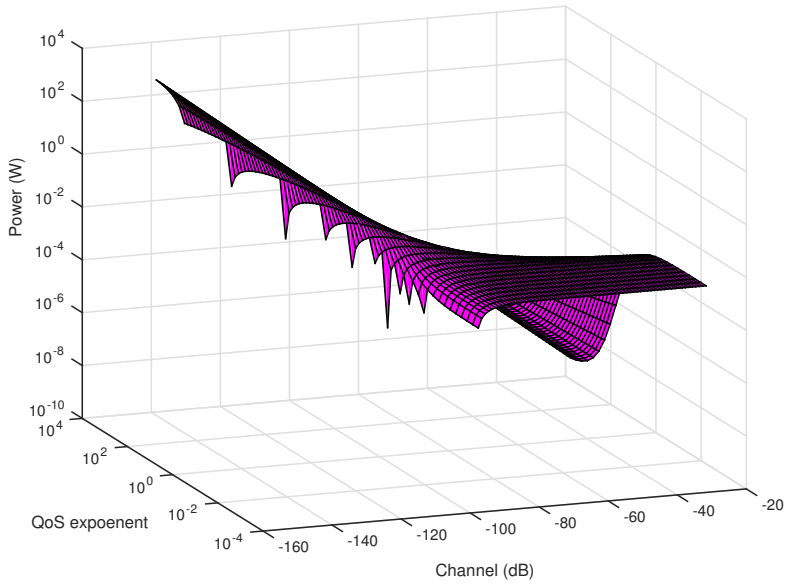
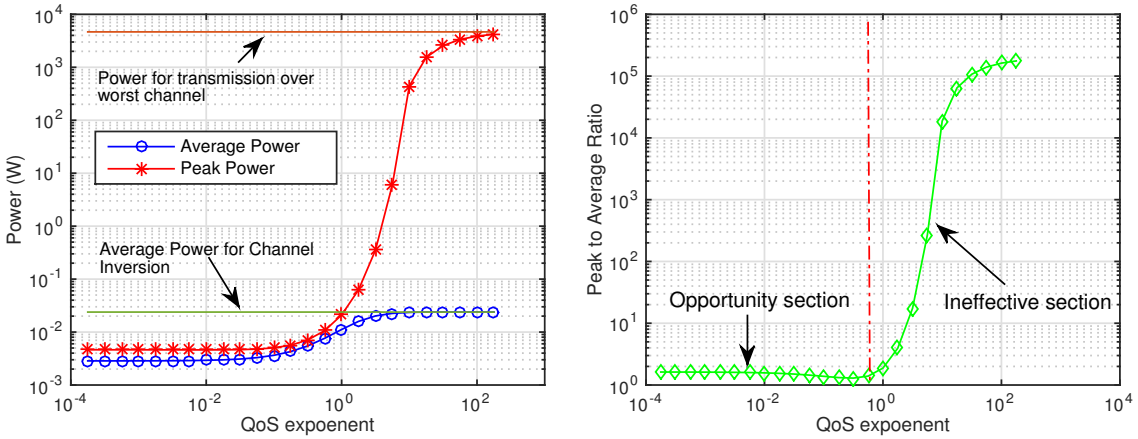


Figure 3.3: Power Allocation with Statistical Delay Provisioning

where $\gamma_0 = \lambda\beta$. λ and v_{k_s} can be obtain through solving $\sum_{k=k_s}^{K} p_k e^{-\theta r(S_k)} - e^{-\theta \bar{A}} = 0$. The structure of the optimal solution has been revealed in [76]. Our goal here is to gain insights on how to design power allocation scheme for wearable sensors.

With QoS provisioning constraint, the solution to the power minimization problem does not have a fixed water level for different channel gain v_k. Instead, the level is channel state dependent. Fig. 3.3 shows an example of optimal transmission power with a fixed target transmission rate, as a function of channel gain and QoS exponent. We can see the trend that with increase of QoS exponent, channels with lower channel gain are utilized, and the peak power increases. Specifically, when QoS exponent θ increases from 10^{-4} to 10^{2}, the peak power increases from about $2 \times 10^{-3} w$ to about $10^{3} w$. Moreover, when QoS exponent is smaller than 10^{-4}, peak power is allocated to the channels with highest channel gain, whereas when QoS exponent is larger than 10^{2}, peak power is allocated to the channels with lowest channel gain. It is reasonable, for when QoS exponent is large, data arrived needs to be transmitted immediately and successfully. Thus, when channel is in a bad state, high transmission power is needed.

We show the impacts of QoS exponent on average power and peak power in Fig. 3.4, where Fig. 3.4(a) shows the average power and peak power v.s. QoS exponent, and Fig.



(a) Average Power vs Peak Power

(b) Peak to Average Ratio

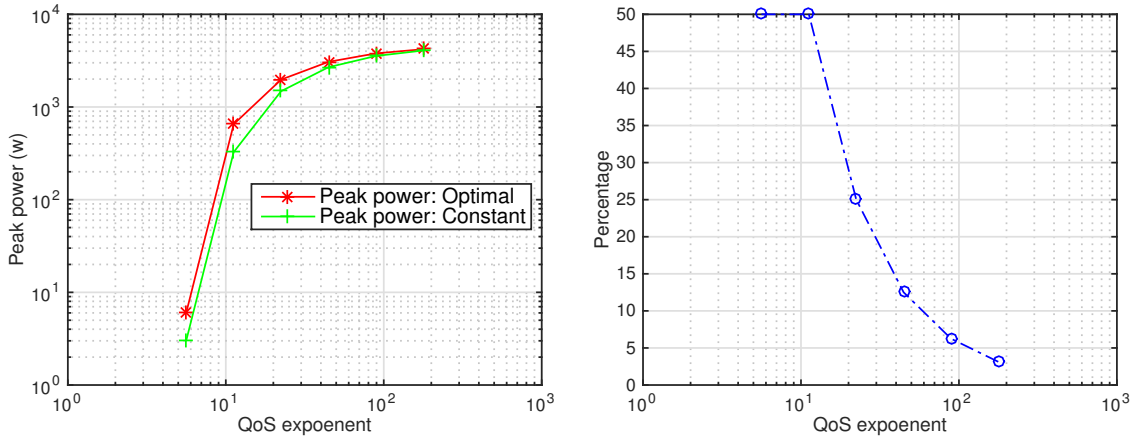
Figure 3.4: Impacts of QoS Exponent on Average power and Peak Power

3.4(b) shows the peak to average ratio v.s. QoS exponent. As we can see from Fig. 3.4(a), the difference between peak power and average power changes slowly when QoS exponent θ is smaller than 10^{-1}, and increases fast when QoS exponent θ is larger than 10^{0}. This trend can also be observed from 3.4(*b*). As shown in Fig. 3.4(b), the peak to average ratio increases from less than 2 to more than 10^{3} when QoS exponent increases from 0.1 to 10.

We refer the section where average power and peak power increase slowly to as opportunity section. The reason is that with small increase of power, the QoS exponent can be increased significantly. We refer the section where peak power increases considerably as ineffective section. The reason is that to improve delay performance in terms of QoS exponent, a system is required to increase peak power significantly. Based on above findings, we suggest that a wearable sensor should be designed to operate near the junction of opportunity section and ineffective section. Thus, the wearable sensor can provide QoS provisioning in an energy efficient way.

Peak Power Reduction

In the ineffective section, peak power becomes the bottleneck for a wearable sensor to provide statistical QoS provisioning. The potential strategy to reduce the peak power is to use a constant power for transmission over all channels. Note that the constant power is chosen as the minimal constant power required such that the statistical delay is



(a) Peak Power: Minimal Average Power vs Constant (b) Percentage of Peak Power Reduction

Figure 3.5: Peak Power Reduction

guaranteed. Using this strategy, the peak power equals to average power consumption. In fact, this strategy is the best a sensor can do to reduce delay when a peak power constraint is imposed.

We show the possibility of reducing required peak power with high QoS exponents in Fig. 3.5, where Fig. 3.5(a) shows the peak power of the minimal average power scheme and of the constant power scheme, and Fig. 3.5(b) shows the percentage of peak power reduction when the constant power scheme is adopted. We can see that, when QoS exponent is close to 10, the constant power scheme can cut 50% peak power, and when QoS exponent increases to 2×10^2, the peak power reduction decreases to less than 5%. The reason is that, with the increase of QoS exponent, the statistical delay requirement converges to a deterministic delay requirement. To guarantee a deterministic delay, the peak power should be able to support transmission of the maximal data arrival rate. Thus, the difference in peak power between the constant scheme and the minimal average power scheme reduces under a high QoS exponent.

3.6 Summary

In this chapter, we have investigated the impacts of peak power constraint and statistical QoS provisioning on transmission power allocation for wearable sensors. We characterize the tradeoff between peak power and average transmission power consumption via duality

gap analysis. An upper bound of the extra average power incurred due to a peak power constraint is derived. Through the analysis of the upper bound, we conclude that when the peak power constraint is stringent, constant power scheme is suitable for wearable sensors for its performance is close to optimal. Further, we show that the peak power constraint is the bottleneck for wearable sensors to provide stringent statistical QoS provisioning.

Chapter 4

MAC for WBANs

WBANs in hospitals could provide continuous monitoring of the physical conditions of patients, thus reducing the workload of medical practitioners and cost of healthcare[49, 16]. However, the wide adoption of WBANs in hospitals faces fundamental challenges to provide the guaranteed communication services for critical medical traffic. As pointed out in [84], to perform rapid medical response, the vital signals, including heart rate, blood pressure, respiratory rate, temperature, pulse oximetry, and level of consciousness, are essential and should be monitored in real time. This dictates that the medical traffic must be delivered with short delays yet high accuracy [85, 86]. In addition, the wearable sensors are typically power limited. This requires the communication protocols to be energy efficient with minimal transmission errors and retransmissions. The sensors are also computing capability limited. With limited data buffer size in the sensors, the real-time data that cannot be transmitted in a given period would be dropped, leading to a high report dropping ratio. In a nutshell, unlike traditional home networks, the e-health systems require more efficient communication services due to the distinguished challenges imposed by the critical medical traffic and resource limited sensors. Since the hospitals are typically space limited, it is common that multiple WBANs deployed for different patients coexist in the same region and inter-WBAN interference is severe. Therefore, an efficient MAC layer resource management is crucial to provision the desired service quality as imposed by e-health systems.

In this work, we aim to develop a centralized MAC protocol for WBANs in hospital environment. In specific, we consider multiple WBANs coexist in a region and contend the channel for transmissions. Each WBAN is composed of wearable sensors which continuously transmit the collected data via a smartphone. Note that due to the variations of the body area channel, the transmission link between a sensor and the smart phone may

not be always available. In order to maximize the network throughput and reduce packet drop ratio, an intuition is to let the WBAN with good channel quality yet long cached data in sensors transmit. However, this requires the real-time channel and buffer state information of all WBANs. Since the on-body sensors are constrained in computing and energy resource, such information cannot be always accurately measured and provided by sensors, which inevitably leads to the inefficient use of channels.

To address this issue, we exploit the temporal channel correlations to guide the MAC resource management. Specifically, we represent the channel state using a belief state, and use this metric to manage the channel access. The belief state which is not chosen for transmissions is updated according to the statistical information. Only the sensor chosen from the WBAN needs to report its current state such that the total channel state reports from sensors are minimized. Given the incomplete information, we formulate the throughput maximization problem as a partially observable optimization problem. To solve this problem, we first analyze the dynamics of the belief states and buffer states. We then construct a myopic policy and investigate its drawbacks in incurring packet dropping. We further propose a modified myopic policy through approximating the future impact of current decision. Finally, we compare our proposed policy with Round Robin (RR) to demonstrate its effectiveness.

4.1 Literature Review

In this section, we review works on resource management in MAC layer for WBANs, and works on partial observable optimal control problem.

The resource management in MAC layer has long been a hot and important research issue in WBAN. In a hospital environment, [87] develops a fuzzy logic learning algorithm to adjust the MAC layer control parameters based on real time network information. To support medical traffic transmission for coexisting WBANs, IEEE 802.15.6 is proposed to eliminate inter-WBANs interference through either collaborative way or non-collaborative ways, such as beacon shifting and channel hopping [88]. To support QoS provisioning for emergency traffic, the IEEE 802.15.6 adopts two schemes [89]. First, it specifics an Exclusive Access Period (EAP) for the traffic with the highest priority. Second, the standard adopts a scheme to freeze the procedure to double contention window for odd times of failure, which aims to reduce the average delay. However, both schemes have limitations in crowded hospitals. To adopt EAP in supporting emergency traffic, all the sensors in the interference range should be synchronized and agree to the same frame structure. The distributed and mobile natures of WBANs make it hard to achieve synchronization

and consensus. Without this agreement, an emergency traffic during EAP from a WBAN could collide with lower priority traffic from another WBAN. Moreover, the idea to freeze contention window double procedure for odd times of failure favors low priority traffic when network is near saturated. Considering low priority traffic has larger contention window, doubling contention window downgrades the performance of low priority traffic more. Thus, emergency traffics do not benefit from the freezing scheme. In summary, to support emergency traffic in hospitals, modifications and improvements of current standard are needed.

Interference mitigation schemes designed for WiFi based network, such as busy tone scheme [90, 91, 92], are not suitable for WBANs as the energy consuming control signals would quickly drain up the battery power of sensors. To reduce energy consumption of sensors, [93] formulates the scheduling problem using the game theory and proposes a heuristic cooperative scheduling policy. Considering a network, which can tolerate concurrent transmission of multiple WBANs, [94] proposes a low-complexity scheduling scheme inspired by the random incomplete coloring scheme. The variations of body area channel are not considered in these works.

In this chapter, we formulate the throughput maximization problem as a partial observable optimization problem due to the incomplete network state information available. The partially observable optimization problem has been studied extensively. The researches started from scheduling over a single random process. For a random process governed by a Markov chain [95, 96] show that a sufficient statistic information for optimal scheduling policy is the conditional probability distribution of current state given previous control decisions and observations. Moreover, the convex property of corresponding value function is proved in [95], which is the key to obtain an optimal policy. [97] studied the property of optimal policies for scheduling over multiple random processes.

4.2 System Model

In this section, we first describe the system model from the aspects of network, channel and traffic.

4.2.1 Network model

Consider a network consisting of N_w WBANs and one central controller, as shown in Fig. 4.1. The central controller is in charge of allocating the channel resources to WBANs,

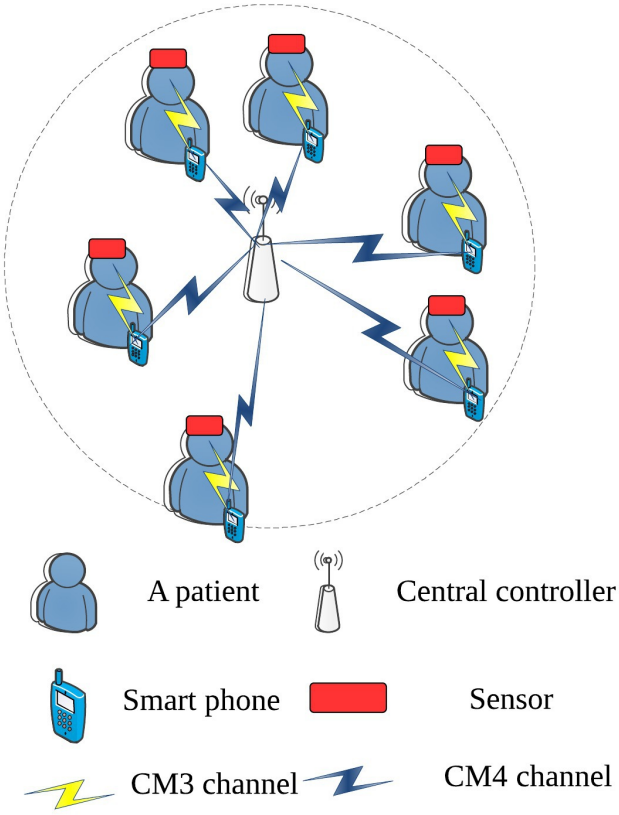


Figure 4.1: System Model

and forwards the medical information through wired network. Each WBAN is equipped on a single patient. Within each WBAN, there is a smart phone and one sensor. Only one sensor is considered in this work since most vital signals can be monitored by a single sensor nowadays [98]. The smart phone is in charge of collecting information from the sensor in the same WBAN, and transmits those information to the central controller. The sensors on body are power and computing capacity limited, whereas smart phones have sufficient power supply and computing capacity. As a result, sensors are set to turn radio on at predefined time instance, and put radio off for the most of time. In the contrast, smart phone is always turned on. The time of the system is partitioned into slots. The duration of each time slot is denoted as T.

4.2.2 Channel Model

The medical traffic is transmitted in two hops from the sensor to the central controller, as shown in Fig. 4.1. Following the standard [99, 100], we denote the first hop body surface to body surface channel as CM3 channel, and the second hop body surface to external channel as CM4 channel. The CM4 channel is modelled as free space wireless channel, and the transmissions of the second hop are error-free due to the ample transmission power and clear channel conditions.

The CM3 channel is represented by following two features [99, 101],

- severe path loss and variations of loss due to the absorption of human body;
- temporal correlations of channel between neighbouring time slots.

To capture above features and without the loss of generality, in the work, we adopt the Gilbert Elliot (GE) model [101, 102], as shown in Fig. 4.2, with two channel states: ON channel state with error-free transmissions, and OFF channel state with unsuccessful transmissions.

Let $C_i(n)$ denote CM3 channel state for the ith WBAN over the nth time slot. $C_i(n) = 1$ if the channel is in the ON state, and otherwise $C_i(n) = 0$. Let R_i^c and Π_i^c denote the probability transition matrix and stationary distribution of the CM3 channel for ith WBAN, respectively. According to the GE model, the probability transition matrix R_i^c can be represented as

$$R_i^c = \begin{bmatrix} 1 - g_i & g_i \\ b_i & 1 - b_i \end{bmatrix}, \tag{4.1}$$

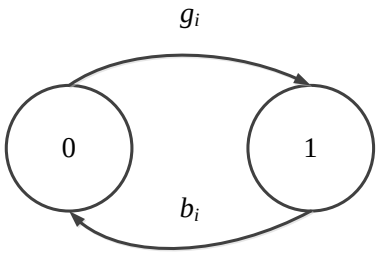


Figure 4.2: ON-OFF wireless channel model

where g_i is the conditional probability that the channel changes from the OFF state to the ON state, and $g_i \triangleq \Pr\{C_i(n) = 1 | C_i(n-1) = 0\}$ for $n \in \{1, 2...\}$. b_i is the conditional probability that the channel changes from the ON state to the OFF state, and $b_i \triangleq \Pr\{C_i(n) = 0 | C_i(n-1) = 1\}$ for $n \in \{1, 2...\}$. The corresponding stationary distribution is $\Pi_i^c = [\frac{b_i}{b_i+g_i}, \frac{g_i}{b_i+g_i}]$. In the GE model, the property that a channel tends to stay in its current state means the channel is positively correlated, i.e., $1 > b_i + g_i$.

4.2.3 Traffic Model

The medical data are collected and summarized as a report by the sensor and transmitted to the smartphone at the end of each time slot. A report contains multiple health condition information, as required by rapid medical response [12]. Let N_p denote the number of packets in a report. The transmission of reports from the sensor to the smartphone follows the sum of Bernoulli process with the arrival rate λ_i for the ith WBAN. As health conditions of patients change much slower than the channel variations, we only consider the scenario with arrival rate λ_i smaller than $\frac{1}{T}$. Due to the limited computing capability of sensors, we assume that the sensor buffer can only store limited packets. With the loss of generality, we consider that a sensor can only cache one report at each time slot. If a new report arrives

whereas the previous report is still cached in the buffer, the previous report is evicted from the buffer and replaced by the new report. Let $q_i(n) \in \{0, 1\}$ denote the buffer state of the ith WBAN at the beginning of time slot n, where $q_i(n) = 0$ represents that the sensor buffer of the ith WBAN is empty at the beginning of time slot n, and otherwise $q_i(n) = 1$.

4.2.4 Channel Access Scheme

The goal of MAC is to maximize the network throughput with modest energy consumption of sensors. The channel access scheme is described as follows. At the beginning of each time slot, the central controller sends out a beacon to choose a WBAN for transmission during this time slot. Since only one WBAN is scheduled, inter-WBANs interference is avoided. Let the $s(n)$ denote the index of the WBAN be chosen during time slot n. Suppose $s(n) = i$. Then the smart phone of the ith WBAN sends a beacon to the sensor in the same WBAN. If the CM3 channel is in ON state and there is one report available for transmission, namely $C_i(n) = 1$ and $q_i(n) = 1$, a successful transmission is made. If a successful transmission from the sensor to the smart phone is made, the buffer of the sensor is emptied. Then the smart phone forwards the report to the central controller. Otherwise, only the channel state is reported to the central controller for future scheduling.

Two types of information about network state are available for the central controller. One type is the statistics of the random processes of the CM3 channels and the medical report event arrival. In practice, the central controller can obtain these statistics through learning over a period of time. Specifically, the smart phone in each WBAN can learn the channel statistics of that WBAN first, and then forward the statistics information to the central controller. Adaptive learning algorithms could be applied to improve learning accuracy in real time. In this work, we assume the central controller can obtain accurate statistics information of the body area channels. We will consider the impact of imperfect and delayed information on MAC design in the future. The other type is the partial real time information of network state. As described in the channel access scheme, the central controller has the information of the WBAN it chooses at the end of each time slot. Thus, the real time information about the network state is partially available.

In order to make proper decision at each time slot, the central controller maintains belief states of the channel and buffer states of all WBAN based on both statistical information and partial real time information. Let $\Omega(n) \triangleq [\omega_1(n), ..., \omega_{N_w}(n)]$ denote the belief states of the channel states of all WBANs at the beginning of time slot n, where $\omega_i(n)$ is the belief state of the channel state of ith WBAN over time slot n. The belief states evolve as follows. If real time information of the ith WBAN is available, the central controller updates its

belief state on the ith WBAN based on real information. Otherwise, the central controller updates its belief based on statistical information. In summary, the belief state evolution can be written as

$$\omega_i^c(n+1) = \begin{cases} 1-b_i, & S(n)=i, C_i(n)=1; \\ g_i, & S(n)=i, C_i(n)=0; \\ T(\omega_i^c(n)), & S(n)\neq i, \end{cases} \tag{4.2}$$

where $T_i^c(\gamma)$ is an evolution operator of the belief state of channel of the ith WBAN. For ON-OFF channel model, the operator is

$$T_i^c(\gamma) = \gamma(1-b_i) + (1-\gamma)g_i. \tag{4.3}$$

As shown in [95], above belief state is a sufficient statistic that depicts current channel state given the channel state is a Markov process.

4.3 Problem Formulation

In this section, we formulate the problem as a partially observable optimization problem. Then, we investigate the value function of the proposed problem for policy design.

4.3.1 Reward and Objectives

We first design a reward to facilitate decision making for the central controller. The reward should favor higher throughput and disfavor packet drop. We design the reward as follows. If the ith WBAN is chosen and a successful transmission is made, $B_i(n)$ units of rewards are collected by the network. Let $R_i(n)$ denote the reward obtained in slot n given the ith WBAN is chosen, we set $R_i(n)$ as

$$R_i(n) = C_i(n)q_i(n)B_i(n). \tag{4.4}$$

The reward shown in equation (4.4) is designed as such: if the channel of the chosen WBAN is in the OFF state or the buffer of the chosen WBAN is empty, the reward is set to zero. Otherwise, $B_i(n)$ amount of reward will be accumulate. We set $B_i(n)$ equal to the probability of exact one medical report arrival since last successful transmission. Thus, if a WBAN is not given the channel access for a long time, the reward the WBAN can give is small, for many packets may have lost.

The control problem for the central controller is to choose which WBAN for channel access at each time slot. A control policy for this problem can be denoted by $\pi : \Omega(n) \rightarrow s(n)$, a function that maps the belief state $\Omega(n)$ to the action $s(n)$. The goal of the central controller is to maximize the average reward of the network over infinite horizon, which is a common measure in communication system [97]. Thus, the control problem can be written as

$$\mathbf{P5} \quad \max_{\pi} \ \mathbb{E}[\lim_{K\rightarrow\infty} \frac{1}{K} \sum_{j=1}^{K} R_{\pi(\Omega(j))}(j)|\Omega(1)]. \tag{4.5}$$

Let π^* denote the optimal solution to P1, it can be written as

$$\pi^* = \arg \ \max \ \mathbb{E}[\lim_{K\rightarrow\infty} \frac{1}{K} \sum_{j=1}^{K} R_{\pi(\Omega(j))}(j)|\Omega(1)]. \tag{4.6}$$

Since the real time full network state is not observable, the proposed problem **P5** is a partially observable optimization problem. If only the channel state is considered, **P5** turns to a partially observable Markov decision process problem (POMDP) [95]. POMDP has larger state space compared to its observable counterparts, making it generally difficult to solve. Our problem is more difficult than POMDP since we consider a random traffic arrival. Problem **P5** falls into the dynamic programming problem category, thus we study the value function of **P5** in the next subsection for policy design.

4.3.2 Value Function

Value function analysis breaks an optimization problem over multiple periods into sub-problems at different points in time. Let $V_n(\Omega(n))$ denote the value function at time slot n. It is the maximum expected reward that the network can gain at time slot n. Consider the central controller choose the ith WBAN at the beginning of time slot n and update the information state at the end of time slot n, the reward can be obtained from time slot n consists of two parts: the expected immediate reward $\mathbb{E}[R_i(n)]$ and the maximum expected reward from time slot $n+1$, namely $V_{n+1}(\Omega(n+1)|s(n) = i, C_i(n))$. Thus, the value function of P1 at time slot n can be written as

$$\begin{aligned} V_n(\Omega(n)) \quad &= \max\nolimits_{s(n)1,\ldots,N_w}\{\mathbb{E}[R_i(n)] + \\ &\omega_i(n)V_{n+1}(\Omega(n+1)|i,1) + \\ &(1-\omega_i(n))V_{n+1}(\Omega(n+1)|i,0\}. \end{aligned} \tag{4.7}$$

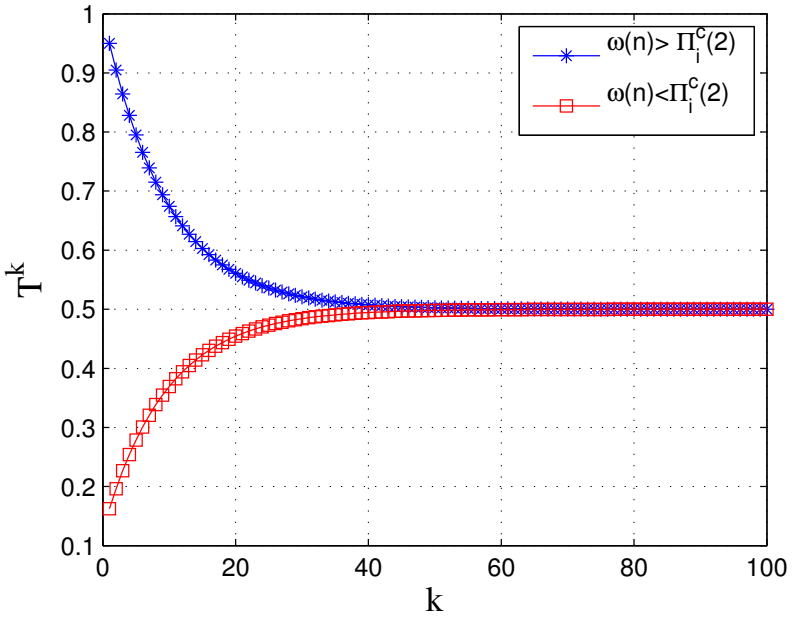


Figure 4.3: Evolution of belief state of channel

For a POMDP, the value function is piecewise linear and convex [96]. However, for a general partial observable problem. This property of the value function may not hold. The equation (4.7) can be solved backwards to obtain the value of $V_1(\Omega(1))$ and the optimal policy π^*. However, due to the exponentially increased computation complexity, the value function and the optimal policy can not be obtained in real time by the central controller.

4.4 Policy Design

Since obtaining the optimal solution to P1 is difficult, we first investigate the properties of channel and buffer dynamics. Based on the analysis, we propose a policy.

4.4.1 Properties of the System Dynamics

To provide insights to this problem, we investigate the properties of the channel and buffer dynamics.

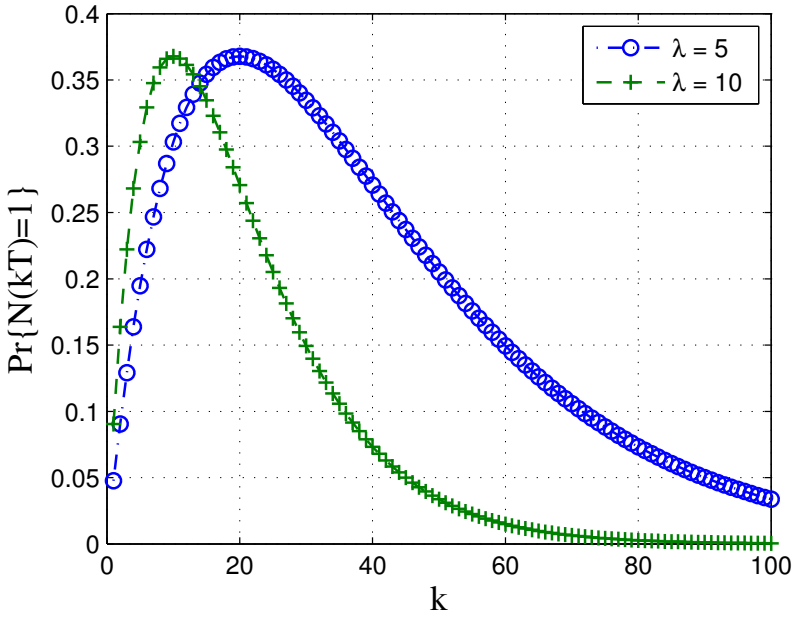


Figure 4.4: Evolution of one event arrival probability

Firstly, we study, given the belief state of the ith WBAN is ω_i at any time slot, what the belief state will be after k consecutive time slots during which the ith WBAN is not chosen by the central controller. Let $(T_i^c(\omega(n)))^k \triangleq \Pr\{C(n+k)=1|w(n)\}(k=0,1,2,...)$ denote the belief state evolution for k consecutive unobserved time slots, we have [97]

$$T_i^c(\omega(n)))^k = \frac{g_i}{b_i+g_i} - \frac{(1-b_i-g_i)^k(g_i-(b_i+g_i))\omega(n)}{b_i+g_i}. \tag{4.8}$$

Given positive correlation of CM3 channel, namely $1-b_i-g_i>0$, we show an example of how $T_i^c(\omega(n)))^k$ changes over time in Fig. 4.3. As we can observe from Fig. 4.3, the belief state will eventually converge to $\Pi_i^c(2)$, the stationary probability that the channel is in the ON state. This result suggests that a proper policy should work in the following way. If the ith WBAN is chosen and the channel of the ith WBAN is in the ON state, in order to utilize the ON channel state, the central controller should prefer to choose this WBAN again in near future. In contrast, if the ith WBAN is chosen and the channel of the ith WBAN is in the OFF state, in order to avoid wasting channel resource, the central controller should prefer not to choose this WBAN in near future.

Secondly, we study, given initial state $q_i(n)=0$, the probability of exact one report

arrives during a duration of k consecutive time slots without transmission. This is the reward we set for a successful transmission. Let $N(kT)$ denote the number of arrived reports during a period of kT. For a sum of Bernoulli process, the probability of m events arrive during a period of kT is $\frac{e^{-\lambda kT}(\lambda kT)^m}{m!}$. Thus, given initial state $q_i(n) = 0$, the probability that exact one report arrives during a duration of k time slots is

$$\Pr\{N(kT) = 1\} = e^{-\lambda kT}(\lambda kT). \tag{4.9}$$

We can derive from equation (4.9) that the probability has a maximum value, which is achieved at the point $\frac{1}{\lambda}$. Since the system is slotted, the corresponding number of time slot k can be either the minimum integer larger than $\frac{1}{\lambda T}$ or the maximum integer smaller than $\frac{1}{\lambda T}$.

An example of how $\Pr\{N(kT) = 1\}$ changes over time is shown in Fig. 4.4. As we can observe from Fig. 4.4, the probability of exact one report arrival increases to its peak value as time goes and then decreases. This result suggests that a control policy should have the following property. If the central controller identifies the buffer of the ith WBAN is empty, it needs to wait a period of time to revisit the ith WBAN for a new report arrival. However, if the duration is larger than $\frac{1}{\lambda}$, the probability of report loss increases, leading to smaller reward. Thus, the central controller should not wait too long for a revisit.

4.4.2 A Modified Myopic Policy

In this subsection, we first construct a myopic policy. Through analysis, we point out the myopic policy will incur high report dropping. Based on the previous analysis on system dynamics, we propose a modified myopic policy to address the report dropping issue through approximating the expected future reward.

If only the dynamics of channel is considered, as shown in [97], the optimal policy is to stick to the WBAN with the ON channel state. Specifically, if a WBAN is found to be in the ON channel state, the central controller should keep choosing this WBAN until the channel turns to the OFF state. With the consideration of random report arrival in WBANs, above policy is no longer optimal. As shown in Fig. 4, after a successful transmission, the probability that there is a report arrival is low. Thus, even with the ON channel state in previous slot, the central control should prefer not to choose this WBAN.

Since obtaining an optimal policy requires high complexity, in this work, we first develop a myopic policy. The objective of a central controller is simplified to maximize the expected reward for current time slot based on the belief states, and ignores the impact of current

decision on the future reward. Let π^m denote the myopic policy. It can be written as

$$\pi^m = \arg\ \max\ \mathbb{E}[R_{\pi(\Omega(j))}(j)|\Omega(1)]. \tag{4.10}$$

The myopic policy π^m has two issues. Firstly, since the belief states of the channel converge, given a homogeneous network setting, where all WBANs share the same statistics, the central controller needs a scheme to choose from multiple WBANs with the same expected rewards, which is not addressed in the myopic policy. Secondly, for a myopic policy, once a WBAN has not be chosen for more than $\frac{1}{\lambda T}$ consecutive time slots, the chance that the central controller will choose this WBAN decreases since the expected reward reduces. This, however, will cause report dropping for that WBAN. This problem is rooted in the myopic philosophy. In contrast, an optimal control policy, which considers future reward, does not have such issue. The reason is that, when a WBAN is not chosen for more than $\frac{1}{\lambda T}$ consecutive time slots under an optimal policy, the central controller will have larger tendency to choose this WBAN. Otherwise, the expected future reward will be smaller, leading to a smaller total reward. In other word, after $\frac{1}{\lambda T}$, the myopic policy significantly deviates from the optimal policy.

We propose a modified myopic policy to address above issues. Firstly, when multiple WBANs have the same maximum expected reward, the central controller chooses one based on a random picker. A random picker is a picker chooses based on a random number it generates from a probability density function. In this work, we use uniform distribution. Secondly, the impact of future reward is considered. Since the complexity to obtain the accurate future reward is high, the future reward is approximated heuristically. Specifically, we increase the expected reward of current time slot for those WBANs have been waiting more than $\frac{1}{\lambda T}$ time slots as

$$R_i(\tau) = \frac{w\tau\lambda T + 1}{\lambda T} C_i(\tau) q_i(\tau) B_i(\tau), \tag{4.11}$$

where τ is the number of slot that the ith WBAN has been waiting more than $\frac{1}{\lambda T}$, and w is scaling factor. The part $\frac{w\tau\lambda T+1}{\lambda T}$ increases as the time goes. Thus, the WBANs that have been waiting more than $\frac{1}{\lambda T}$ time slots will have increasing chances to be chosen. This helps solve the second issue introduced by the myopic policy.

4.5 Simulation results

This section evaluates the performance of the proposed MAC layer resource management through simulations using Matlab.

4.5.1 Simulation Setup

We simulate a scenario similar to that shown in Fig. 4.1. A central controller is placed at the center of the network. There are total N_w patients, with each installed a WBAN for health monitoring. The CM3 channels are simulated using the GE model, where as the CM4 channels are simulated to be error-free for packet transmissions. The initial channel states of the patients are generated randomly based on the stationary distribution of the channel states. In practice, the arrival rates for vital signals, such as blood pressure and heart rate, are usually less than $100kbps$, whereas the arrival rates for ECG and EMG signals could be near $1Mbps$ [85]. Thus, the network can be either in unsaturated condition or congested condition. This motivates us to evaluate the effectiveness of the proposed resource management under both unsaturated network condition and congested network condition. Given the normalized service capacity, the congested network condition can be represented as

$$\sum_i \lambda_i T > 1, \tag{4.12}$$

whereas the unsaturated network condition as

$$\sum_i \lambda_i T < 1. \tag{4.13}$$

In the simulation, we varies the report arrival rate to change the network condition. Let λ_c and λ_u denote the report arrival rate for congested network condition and unsaturated network condition, respectively. Let N_w^c and N_w^u denote the number of patients under congested network condition and unsaturated network condition, respectively.

For simplicity, we consider a homogeneous network, where the report arrival rates and the channel statistics of all WBANs are the same. As such, we omit the index of parameters in the following. The detailed settings can be found in Table 6.2.

In each experiment, we compare our proposal with the round robin (RR) scheme [103] and a myopic (Myo) policy [104]. The RR scheme is chosen since it is simple and starvation free. In the RR scheme, the central controller assigns the channel access opportunity to WBANs in a circular order. Let $s_{rr}(n)$ denote the WBAN chosen in time slot n. It can be written as

$$s_{rr}(n) = n \mod N, \tag{4.14}$$

where $\mod$ is the modulo operator. We set $kN \mod N = N, k \in \{0, 1, 2, ..\}$ since the network index starting from 1 in our work. The myopic algorithm is chosen since if only channel dynamics is considered, it has been proven to be optimal [104]. The myopic algorithm is to choose the WBAN with the best belief state of channel.

Table 4.1: System Parameters for Simulation

Parameter	Definition	Value
b	probability that channel turns good	0.15
g	probability that channel turns bad	0.05
N_w^c	number of patients in cong. network	10
N_w^u	number of patients in unsat. network	15
λ_u	arrival rate for unsaturated scenario	5 per sec
T	slot duration	10 ms
λ_c	arrival rate for congested scenario	30 per sec
w	scaling factor for modified reward	10

We denote our proposal as MyoMo throughout this section. In each simulation, we report the number of successful transmissions, number of reports dropped and number of wasted transmission opportunities using our proposal and existing proposals. An wasted transmission opportunity event occurs if a WBAN is chosen but without any successful transmission, either due to channel in the OFF state or empty buffer. The performance improvements of MyoMo and RR over Myo are also reported. They are calculated as ratio of comparing the performance differences between an algorithm and Myo to the performance of Myo in percentage. Each simulation is conducted over a 1000s duration.

4.5.2 Performance Evaluation

We report the simulation results from the following three aspects: 1) network throughput in terms of number of successful transmissions; 2) number of reports dropped; and 3) channel utilization in terms of number of wasted transmission opportunities.

Unsaturated Scenario

The performance of three algorithms, namely RR, Myo and MyoMo, under unsaturated network condition, is shown in Fig. 4.5. It can be seen from Fig. 4.5(a) that in terms of the number of successful transmissions, Myo has the worst performance with about 450 less successful transmissions than RR and MyoMo. MyoMo performs slightly better than RR with about 50 more successful transmissions in average. We can observe from Fig. 4.5(b) that Myo causes about 400 more dropped reports than RR and MyoMo, and most WBANs drop less reports under MyoMo than under RR. The trends that Myo performs worse than RR and MyoMo can also be observed from Fig. 4.5(c). The improvements

of RR and MyoMo over Myo under unsaturated scenario are shown in Fig. 5(f). It can be seen from Fig. 5(f), compared to Myo, MyoMo and RR have 82% and 77% more successful transmissions, around 15% less dropped reports, and around 10% less wasted transmission opportunities. The reason why Myo has the worst performance is that only the channel state is considered by Myo. Thus, Myo always chooses the WBAN with the best channel. However, the chosen WBAN could have an empty buffer, leading to a high number of wasted transmission opportunities. With a low channel utilization, the number of reports dropped is high and the number of successful transmissions is low. The reason that RR performs very close to MyoMo is as follows. Under unsaturated network condition, a wasted transmission opportunity is more likely caused by an empty buffer than a channel in the OFF state. In RR algorithm, each WBAN needs to wait for N_w^u time slots for a transmission opportunity. When N_w^u, the number of WBANs in the network, is sufficiently large, the probability of an empty buffer is small. In other word, RR performs as an algorithm that aims to avoid an empty buffer. As a result, RR can effectively reduce the number of wasted transmission opportunities and improve the number of successful transmissions. Meanwhile, under unsaturated network condition, the reward adopted by MyoMo shown in equation (4.11) is dominant by the consideration of buffer. In this case, MyoMo can be regarded as an algorithm that aims to avoid an empty buffer, which is similar to RR. As a result, RR performs similarly to MyoMo under unsaturated scenario.

The differences among WBANs in terms of the number of channel in the ON state and the number of report arrivals are shown Fig. 5(d) and Fig. 5(e), respectively. In Fig. 5(d), 0 in the y-axis represents the average number of channel in the ON state. It can be seen from Fig. 5(d) that the differences in the number of channel in the ON state among different WBANs can be more than 1000. The reason is twofold: 1) the initial channel states are generated randomly based on the stationary distribution of the channel states; and 2) the channel states evolve according to a probability transition matrix as shown equation (4.2). It can be seen from Fig. 5(e) that the number of report arrivals is different for different WBANs. These differences are the result of the random Poisson number generation method we adopted using Matlab. The differences in the number of channel in the ON state and the number of report arrivals cause the variations in performance among WBANs under the same algorithm.

Saturated Scenario

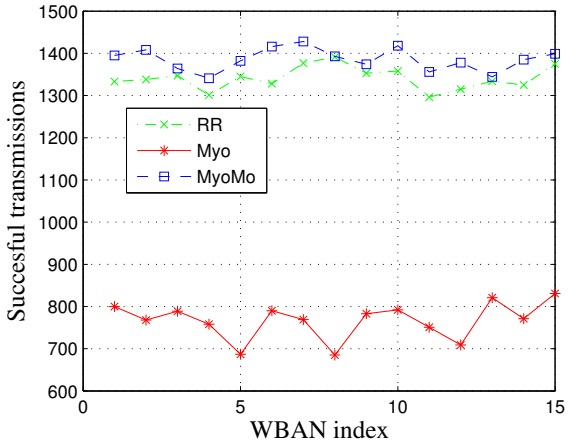
The performance of three algorithms, namely RR, Myo and MyoMo, under congested network condition, are shown in Fig. 4.6. It can be seen from Fig. 4.6(a), in terms of the number of successful transmissions, MyoMo outperforms RR and Myo with about 300

more successful transmissions, and Myo has the worst performance. But the performance difference between Myo and RR is smalle. Above trends can also be seen in report dropping performance from Fig. 4.6(b) and channel utilization performance from Fig. 4.6(c). The improvements of RR and MyoMo over Myo under congested scenario is also shown in Fig. 6(f). It can be seen from Fig. 6(f), compared to Myo, MyoMo and RR have 20% and 8% more successful transmission, 2% and 1% less report dropping event, and 6% and 3% less wasted transmission opportunities. It can be concluded that MyoMo outperforms RR under congested scenario. The reason is that, under congested scenario, the proportion of wasted transmission opportunities caused by empty buffer reduces, whereas the proportion caused by channel in the OFF state increases. Compare to RR, MyoMo exploits the temporal channel correlation through utilizing the belief state of channel in making decision. Thus, MyoMo is less likely to choose a WBAN with channel in the OFF state, leading to a better performance.

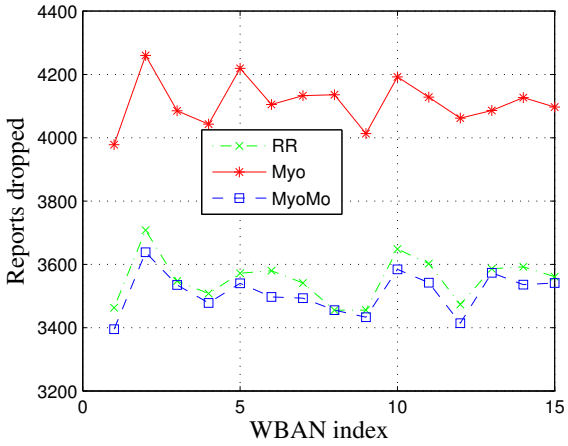
From Fig. 4.6 and Fig. 4.5, we can see that the number of successful transmissions under congested scenario is larger than that under unsaturated scenario, whereas the number of wasted transmission opportunities under congested scenario is smaller than that under unsaturated scenario. The reason is that, under congested scenario, the wasted transmission opportunities due to empty buffer are greatly reduced, leading to a higher number of successful transmissions.

4.6 Summary

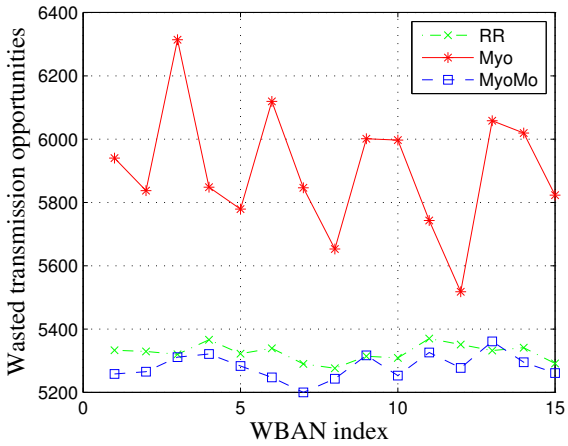
In this chapter, we have proposed a MAC layer resource management scheme for WBANs [105]. Using both analysis and extensive simulation results, we have demonstrated the effectiveness of the proposal in increasing network throughput and enhancing the channel utilization under both the unsaturated and congested network conditions. In the future, we intend to investigate distributed MAC protocol for WBANs. Due to the mobility nature of WBANs, a centralized controller is not always available. This calls for development of a MAC scheme that is able to handle the inter-WBAN interference and is operated in a distributed manner.



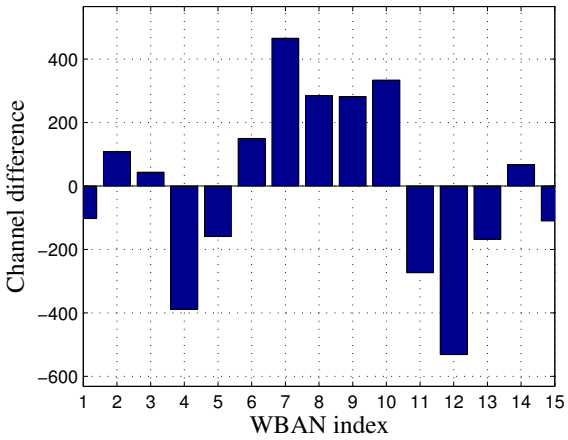
(a) Successful transmissions



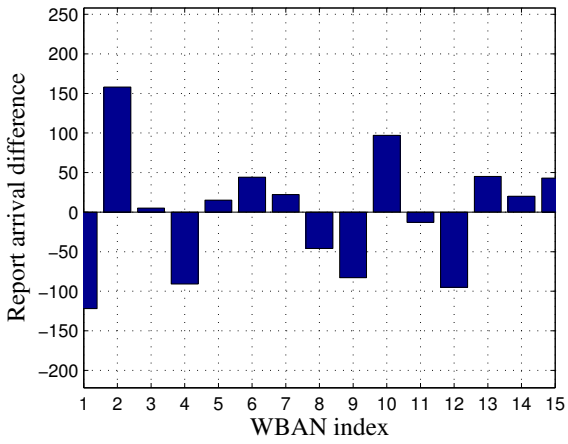
(b) Report dropping



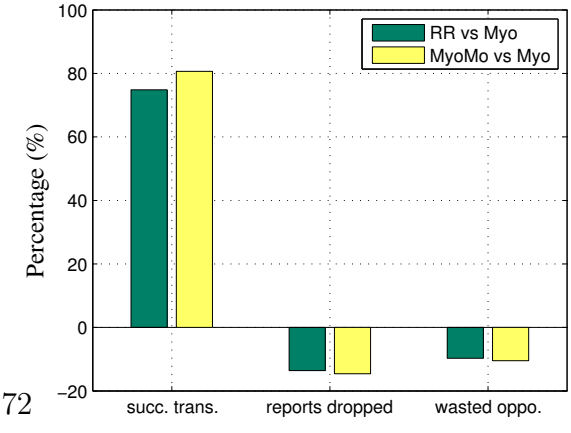
(c) Waste transmission opportunities



(d) Channel difference

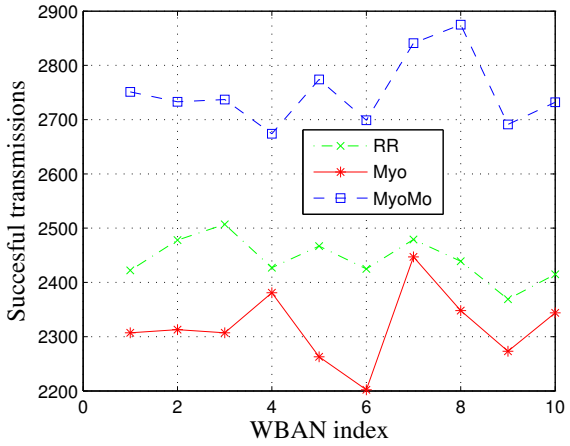


(e) Report arrival difference

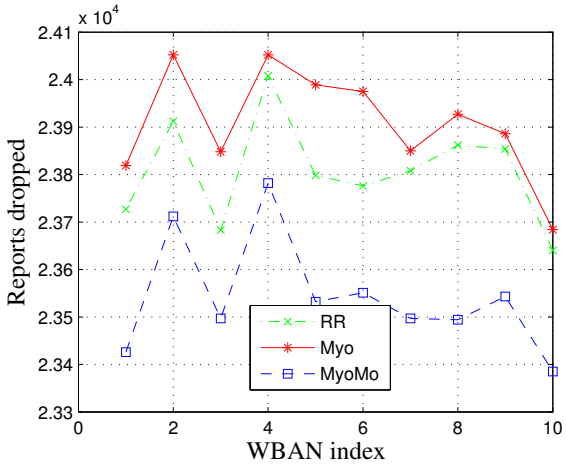


(f) Improvement in percentage

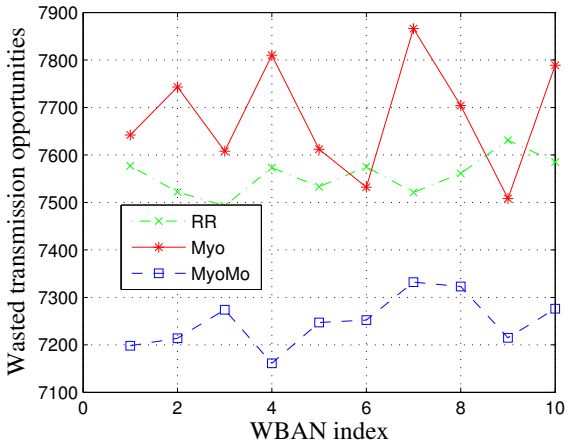
Figure 4.5: Performance comparison for unsaturated scenario



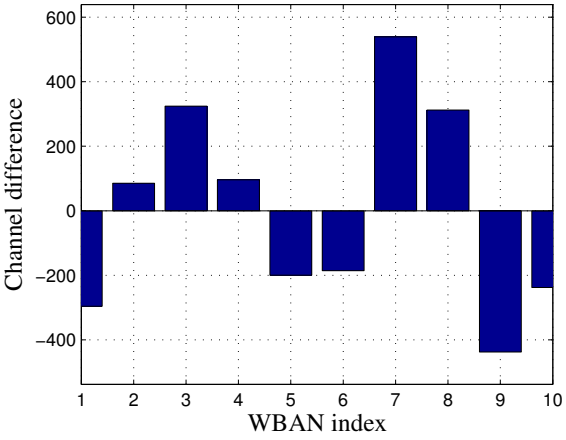
(a) Successful transmissions



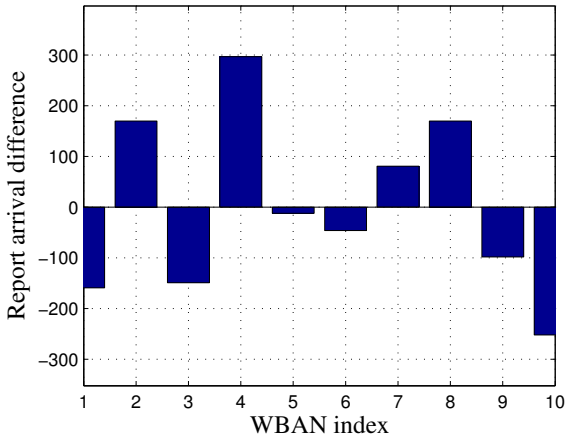
(b) Report dropping



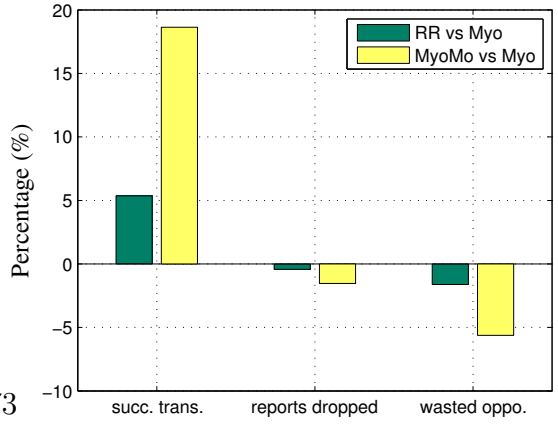
(c) Waste transmission opportunities



(d) Channel difference



(e) Report arrival difference



(f) Improvement in percentage

Figure 4.6: Performance comparison for congested scenario

Chapter 5

Reliability Enhancement for Private Clouds

In this chapter, we propose a novel cooperation framework to solve the reliability issue of private clouds and reduce the delay suffered by public clouds. Our contributions are twofold.

First, a cooperation framework is proposed through exploiting unique features of existing clouds. First feature is that private clouds are geographically distributed, whereas the second is public clouds can be regarded to possess infinite computing resources available [41]. The cooperation contains two key ideas: 1) the geographically distributed private clouds offer parts of their capacities to help a public cloud provide services to its nearby users. The public cloud gives reward for this help. In summary, the abundant resources of private clouds are utilized through investing them in a public cloud. 2) With sufficient reward, the public cloud offers to serve excess requests to the private clouds to improve their reliability and scalability. To adopt reward is to eliminate selfish private clouds. This cooperation scheme improves reliability of private clouds and reduces the delay of public clouds at the same time.

Second, considering the potential fatal results of failure in e-health systems, we adopt a stringent reliability measurement, which is the probability that a failure never happens. Based on the proposed framework, we develop an algorithm for private clouds to decide how many computing resources to be shared at each time to avoid failure. Using stochastic control theory, we prove that our designed strategy is optimal for this stringent reliability requirements. Numerical results demonstrate that our proposed scheme improves the reliability of private clouds over non-cooperation scheme significantly.

5.1 Literature Review

In this section, we review related studies on QoS improvement of cloud computing in e-health. Then, we introduce the basics of risk control using stochastic control.

5.1.1 Private Cloud for e-health

The state of cloud computing in e-health systems is summarized in [106]. QoS provisioning is identified as one of the challenges faced by exiting cloud models [107]. The works on improving QoS of public clouds consists focus on design geographically distributed clouds and allocate contents among these clouds to reduce delay and provide consistent QoS. An efficient resource management algorithm for distributed clouds is designed in [47] to minimize the maximum latency between selected distributed clouds. To reduce the cost of utility without degrading the QoS, an task allocation among distributed clouds problem is studied in [44]. However, these approach cannot be applied to e-health. Medical computing facilities are required to be built with a stringent standard [42], thus it is expensive for public service providers to maintain servers at different locations under medical standards.

The challenges faced by private clouds are summarized in[43]. For medical services, the losses associated with service agreement violation far outweigh any cost savings [43]. Thus, for the private clouds, the primary goal is to reduce the probability of failure. In order to improve the reliability of a private cloud, hybrid clouds model, which adopts public clouds as the backup servers when the private cloud cannot provide sufficient resource by itself, is studied in [46]. However, this scheme does not provide solution to delay problem of public clouds. Within one clouds, the performance of services is analyzed in [108] using a M/G/m/m+r queuing Systems. The security and privacy concerns on the cooperation among clouds have been addressed in [109] through a scheme to partition works dependent on their security levels.

5.1.2 Failure Probability Control

To minimize the failure probability under random demand process is a risk control problem. For an insurance company, it needs to set a proper price of premium and invest its assets to reduce the probability that the claims from the users surpass the cash the company earned. This problem is critical to small insurance company with limited reserve. Two methods have been applied to control the risk. One is to invest cash the company has. The other is to reinsurance the contract to a bigger insurance company. Stochastic control can

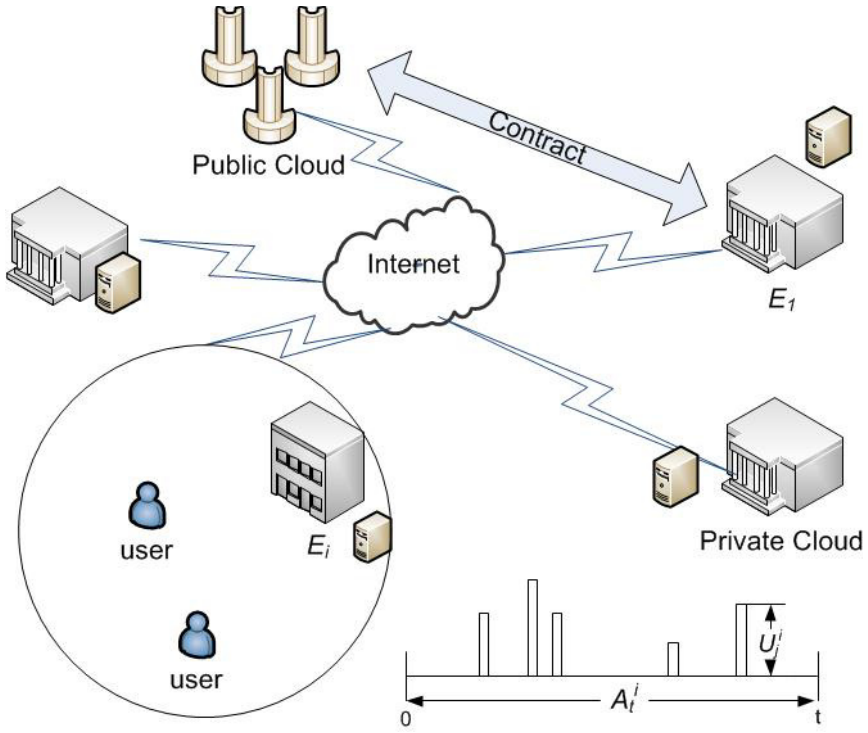


Figure 5.1: System model for clouds

be applied to solve the risk control problem. Under a proportional risk sharing scheme, an optimal policy has been proposed in [110]. Further, the optimal policy for an excess loss risk sharing agreement is derived in [111]. General results for risk control related stochastic control are summarized in [112, 113].

5.2 System Model

We consider the resource management of clouds for medical applications in continuous time. As shown in Fig. 6.1, a public cloud and N_p private clouds locate in different geographic regions. Each region has only one private cloud. Let E^i denote the ith private cloud. To provide satisfactory services to end users, each cloud limits the maximum number of virtual machine (VM) it can support simultaneously. The maximum number of VM is used to depict the computing capacity of each cloud. The computing capacity of public cloud is considered as infinity [41]. The computing capacity of each private cloud is limited.

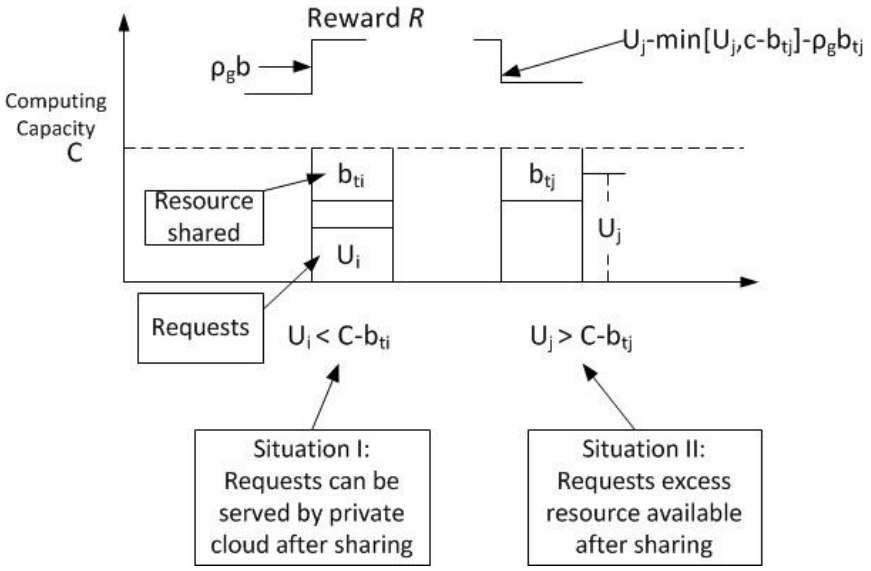


Figure 5.2: Reward Dynamics

Denote by C_i, $i \in 1, ..., N_p$, the computing capacity of private cloud E^i.

The requests for each private cloud come from the region where the private cloud locates. Let A_t^i be the number of requests for the jth private cloud in a time interval $(0, t]$. A_t^i is a Poisson process with arrival rate λ^i. Each request consists of a series of tasks from different users. Each user could have several tasks. Each task requires one VM. In this work, tasks are regarded as flash jobs since we consider resource management in continuous time. Let U_j^i denote the number of tasks of the jth request for private cloud E^i. U_j^i, $j = 1, 2, ..$, is a positive random i.i.d variable independent of A_t^i. In summary, demand process is determined by two random variables A_t^i and U_j^i, as shown in Fig. 6.1. The requests for the public cloud originate from all regions. If the size of requests exceeds the computing capacity of the private cloud, the extra part can not be served by current private cloud.

5.3 Resource Sharing among Clouds

5.3.1 A Cooperation Framework

We propose a cooperation scheme aiming at improving the reliability of private clouds, as well as reducing the delay suffered by public clouds. The cooperation scheme is motivated by the facts that: 1) private clouds are geographically deployed and have abundant resources in average sense; and 2) public clouds can be regarded to possess infinite computing resources. Thus, we propose to store those abundant computing resources of private clouds through public cloud and fetch them when the private cloud needs. The fact public cloud have resources, which can be regarded as infinity, makes this storage concept reasonable. Plus, public clouds are allocated with geographically distributed computing resources, which helps to reduce delay.

To enable cooperation, the ith geo-distributed private cloud decides to share $b^i(t)$ amount of resource to serve the requests to public cloud at its region at time t. The decision for time t is made at time $t-$, and is an agreement between the private cloud and the public cloud, which cannot be violated. In return, the public cloud rewards $\rho_g^i b^i(t)$ amount of resource, where ρ_g^i is a scaling factor. ρ_g^i reflects the cooperation inclination between the public cloud and ith private cloud. Practically, the private cloud with larger capacity has less incentive to cooperate, namely if $C_i > C_j$, $\rho_g^i < \rho_g^j$. The accumulated reward of ith private cloud at time t is denoted by $R^i(t)$.

Given the randomness of the arrival process, the accumulated reward is random. The failure probability is associated with accumulated reward. To minimize the failure probability, the shared resource $b^i(t)$ shall be chosen according to the dynamics of the accumulated reward. Thus, the decision can be represented as a feedback equation:

$$b^i(t) = b^i(R_{t-}^{b^i}), \tag{5.1}$$

where $R_{t-}^{b^i}$ is the reward process under strategy $b^i(t)$. The optimal strategy depends on the cumulative reward the private cloud has at each time only and not on the history of the cumulative reward.

Two situations could happen under the proposed framework. Suppose the jth request arrives at time t, 1): when the size of the request $U_j^i(t)$ is smaller than the remaining capacity after cooperation, the reward increases by $\rho_g^i b^i(t)$, as shown by Situation I in Fig. 5.2; 2): When the size of the requests $U_j^i(t)$ is larger than the available computing capacity $C_i - b^i(t)$ at time t, the private cloud uses $U_j^i(t) - \min[U_j^i(t), C_i - b^i(t)s]$ amount of reward

to serve the extra part, as shown by Situation II in Fig. 5.2. Conisder above situations, the reward process evolves as

$$R_b^i(t) = R_b^i(0) + \rho_g^i \int_0^t b^i(s)ds - \sum_{j=1}^{A_t^i} [U_j^i - \min(U_j^i, C^i - b^i(t))], \tag{5.2}$$

where $R^i(0)$ is the initial reward that the public cloud give to the ith private cloud, $\min(x, y)$ is the minimum of x and y. Assume ρ_g^i is smaller than 1 for every $i, (i \in 1, ..., N_p)$, $\rho_g^i < 1$. Otherwise, the optimal policy is to share all computing capacity with the public cloud, which is not reasonable due to privacy and delay concerns. $U_j^i - \min(U_j^i, C^i - b^i(t))$ can be written in the form of a sign function $(U_j^i - C^i + b^i(t))^+$, where

$$x^+ = \begin{cases} 0, & \text{if } x < 0; \\ x, & \text{otherwise.} \end{cases} \tag{5.3}$$

5.3.2 Definition of Failure Probability

Consider the potential fatal results of failure in medical services, the probability of first failure of a private cloud is adopted as reliability measurement. A private cloud provides services to its users based on service level agreements. An event that the service agreement is violated is called a failure event. Specifically, the event a task within a request to private cloud cannot be served is considered as a violation in this work. Let τ be the first time a failure event occurs. The failure probability is defined as the probability that τ is finite. Without cooperation, the failure occurs when the capacity of a private cloud is not able to serve requests. Let τ_{nc}^i denote the first time a failure occurs without cooperation for ith private cloud, we have

$$\tau_{nc}^i = \inf\{t \geq 0 : U_i(t) > C_i\}. \tag{5.4}$$

Let P_{fi}^{nc} denote the failure probability without cooperation for the ith private cloud, we have:

$$P_{fi}^{nc} = P\{\tau_{nc}^i < \infty\}. \tag{5.5}$$

Without cooperation, the only way to reduce failure probability is to increase the service capacity C_i to infinity.

With cooperation, the ith private would experience a failure event at time t if the cumulated reward is below 0 at time t, $R_b^i(t) < 0$. If $R_b^i(t) > 0$, even if the size of requests is larger than the remaining resource, the private cloud can turn to public cloud.

Let τ_b^i denote the first time of failure occurs for the ith private cloud under the proposed framework. It is defined as:

$$\tau_b^i = \inf\{t \geq 0 : R_b^i(t) < 0\} \tag{5.6}$$

Then the corresponding failure probability is $P\{\tau_b^i < \infty\}$.

In the following, we formulate the minimize failure probability problem for the private clouds. We adopt dynamic programming and derive the Hamilton-Jacobi-Bellman (HJB) equation. Since each private cloud has the same reward updating equation, we omit the index of private clouds for simplicity.

5.4 Objective of Private Clouds

The objective of private clouds is to minimize the failure probability under the proposed framework as shown in P6.

$$\textbf{P6} \quad \min_b \quad P\{\tau_b < \infty\} \tag{5.7}$$

As stated in equation (5.1), the control variable in (5.7) is determined by how much reward is left. The HJB equation derived from **P6** is hard to solve. Instead, we consider survival probability. Let $\delta_b(s)$ denote the survival probability given there is s amount of rewards left and strategy b is adopted, we have:

$$\delta_b(s) = P\{\tau_b = \infty | R(t) = s\} \tag{5.8}$$

The minimization of failure probability can be transformed to the maximization of the survival probability. Consider P7:

$$\textbf{P7} \quad \max_b \quad \delta_b(s). \tag{5.9}$$

Let $\delta(s)$ denote the optimal value with $\delta(s) = \sup_b\{\delta_b(s)\}$.

5.5 Derivation of HJB Equation

The objective is in the form of probability. We derive the HJB equation heuristically. Consider reward update equation (5.2) within a short time interval $(0, \Delta]$, during which a fixed strategy b is adopted. Consider the arrival process, we have:

1. With probability $1 - \lambda\Delta + o(\Delta)$, there is no request in time interval $(0, \Delta]$. Then the cumulated reward at time Δ can be written as:

$$R_\Delta = s + \rho_g b\Delta. \tag{5.10}$$

2. With probability $\lambda\Delta$, there is exactly one request with size U in time interval $(0, \Delta]$. The cumulated reward is:

$$R_\Delta = s + \rho_g b\Delta - \mathrm{E}[(U - C + b)^+]. \tag{5.11}$$

The probability of more than one request come is within an order than the first infinitesimal of time duration Δ, thus can be omitted. The survival probability under strategy b can be calculate by taking expectations and averaging over all possible request sizes. We have

$$\begin{aligned}\delta_b(s) =&(1 - \lambda\Delta + o(\Delta))\delta_b(s + \rho_g b\Delta) + o(\Delta)\\ &+ \lambda\Delta\{\delta_b(s + \rho_g b\Delta - \mathrm{E}[(U - C + b)^+])\}.\end{aligned} \tag{5.12}$$

For $\Delta \to 0$ we have

$$0 = \lambda\mathrm{E}[\delta_b(s - (U - C + b)^+) - \delta_b(s)] + \rho_g b\delta_b(s)'. \tag{5.13}$$

Maximizing over all possible values for decision b, we obtain the HJB equation of our problems:

$$0 = \sup_b\{\lambda\mathrm{E}[\delta(s - (U - C + b)^+) - \delta(s)] + \rho_g b\delta(s)'\}. \tag{5.14}$$

Before solving the HJB equation, we investigate the property of the solution. Intuitively, when the reward goes to infinity, the failure probability becomes zeros. In fact, as stated in Lemma 2, the condition that reward goes to infinity is the necessary condition for the failure probability to be nonzero.

lemma 2 *For any strategy b, with probability 1, either failure occurs or $R_b(t) \to \infty$ as $t \to \infty$.*

Lemma 2 can be stated as a joint probability of two events is zero. First event is the reward process under a strategy b, R_b, is bounded. Second event is the first time of failure τ_b under strategy b goes to infinity.

Proof 2 *Choose a constant B satisfies $P\{U > B\} > 0$. Assume a reward process $R_b(t)$ under an arbitrary strategy b is bounded by a constant $M > 0$, namely $R_b(t) \leq M$. We show the reward process $R_b(t)$ goes to zero before time goes to infinity. In other word, the first time of failure $\tau_b < \infty$. That is*

$$P\{R_b(t) \leq M \text{ for all } t > 0 \text{ and } \tau_b = \infty\} = 0. \tag{5.15}$$

Let n be the number of requests, whose size are larger than B, during an interval of length 1. The probability of n satisfies $n = \lceil \frac{M+\rho_g C}{B-C} \rceil$ is positive, where $\lceil x \rceil$ is the minimum integer larger than x. Given the request process is stationary and independent, with probability 1 there are more than n requests during an interval $[t, t+1]$. Consider the reward process $R_b(t) \leq M$,

$$R_b(t) \leq M + \rho_g C - n(B - C) < 0. \tag{5.16}$$

Thus, given the reward process is bounded, the first time of failure $\tau_b < \infty$. This proves the lemma.

Since the reward increases to infinity is the necessary condition for the failure probability to be zero, we only consider strictly increasing functions. To solve the equation (5.14) with a strictly increasing solution, we rewrite the equation in a standard derivation equation form as

$$\delta'(s) = \inf_b \{\lambda \frac{\delta(s) - \delta(s - \mathrm{E}(U - C + b)^+)}{\rho_g b}\}. \tag{5.17}$$

5.6 Strategy Design

In this section, we design a strategy for the problem and prove that the strategy we design is the optimal solution for our problem defined. It consists of two steps: first to design a strategy and to show the strategy is a solution to equation (5.17); second to prove the strategy proposed maximizes the survival probability.

5.6.1 Existence of a Solution

In this subsection, we prove the existence of a solution of equation (5.17). To do so, we construct a sequence and show the sequence converges to a function which is a solution. To claim the sequence converges, we show the sequence is monotonous and bounded. We refer previous work [114] [112],[111] for preliminaries.

theorem 4 *Assume the request size distribution Q is continuous. There exists a nondecreasing solution $V(s)$ of the HJB equation (5.17), which is continuous on $[0,\infty)$, continuously differentiable on $(0,\infty)$.*

Proof 3 *Define a sequence $V_n(s)$ with $V_0(s) = \delta_0(s)$, which is the failure probability with full cooperation. For full cooperation, the private cloud chooses $b = C$. And the sequence is obtained through recursion as*

$$V'_{n+1}(s) = \inf_b \{\frac{V_n(s) - V_n(s - E(U - C + b)^+)}{\frac{1}{\lambda}\rho_g b}\}. \tag{5.18}$$

We prove the the sequence $V'_n(s)$ is a decreasing sequence through induction. We first show the deceasing property holds for $V'_1(s) \leq V'_0(s)$. For $n = 0$, we have

$$V'_0(s) = \lambda \frac{V_0(s) - V_0(s - E(U)^+)}{\rho_g C}, \tag{5.19}$$

which is derived from equation (5.13).

Consider $n = 1$ for equation (5.18), we get

$$V'_1(s) = \inf_b \{\lambda \frac{V_0(s) - V_0(s - E(U - C + b)^+)}{\rho_g b}\} \tag{5.20}$$

Obviously, $V'_1(s) \leq V'_0(s)$. We then show given $V'_n(s) \leq V_{n-1}(s)'$ for all $s \geq 0$, $V'_{n+1}(s) \leq V_n(s)'$ holds. For all b, we have

$$\begin{aligned} V'_{n+1}(s)\rho_g b &\leq \lambda E[V_n(s) - V_n(s - (U - C + b)^+)] \\ &= \lambda E[\int_{s-(U-C+b)^+}^{s} V'_n(u)du] \\ &\leq \lambda E[\int_{s-(U-C+b)^+}^{s} V'_{n-1}(u)du] \\ &= \lambda E[V_{n-1}(s) - V_{n-1}(s - (U - C + b)^+)]. \end{aligned} \tag{5.21}$$

Divide $\rho_g b$ on both sides, we have

$$V'_{n+1}(s) \leq \frac{\lambda E[V_{n-1}(s) - V_{n-1}(s - (U - C + b)^+)]}{\rho_g b} \tag{5.22}$$

Since b is arbitrary, we can conclude from equation (5.22) and equation (5.18) that $V'_{n+1}(s) \leq V_n(s)'$. *In summary,* $V'_n(s)$ *is a decreasing sequence.*

Now we show $V'_n(s) > 0$, *for all* n. *Suppose*

$$s_0 = \inf\{s : V'_n(s) = 0 < \infty\}. \tag{5.23}$$

Then, we have

$$0 = \inf_b\{V_{n-1}(s_0) - V_{n-1}(s_0 - E(U - C + b)^+)\}. \tag{5.24}$$

Thus, $V_{n-1}(s_0) = V_{n-1}(s_0 - E(U - C + b)^+)$. V_{n-1} *is continuous for* $[s_0 - E(U - C + b)^+, s_0]$ *and differential for* $(s_0 - E(U - C + b)^+, s_0)$. *Based on Rolle's theorem, there exists* $s_1, s'_1 \in [s_0 - E(U - C + b)^+, s_0]$, *with* $V_{n-1}(s_1) = V_{n-1}(s'_1)$. *Choose* b *such that* $s'_1 = s_1 - E(U - C + b)^+$. *Then* $V'_n(s_1) = 0$, *which is contradicted to the assumption* s_0 *is the minimum value for* $V'_n(s) = 0$. *Thus, we conclude* $V'_n(s) > 0$.

Since $V'_n(s) > 0$, *we conclude from bounded convergence* $g(s) = \lim_{n\to\infty} V'_n(s)$ *exists.*

Define $V(s)$ *as:*

$$V(s) = 1 + \int_0^s g(s)ds. \tag{5.25}$$

Obviously, $V(s)$ *is a nondecreasing continuous function, which fulfills equation (5.17). So far, we prove the existence of a solution to the HJB equation.*

We now proof the derivation of $g(s)$ *is continuous. Consider* $x, y > 0$, *we have*

$$\begin{aligned} |g(x) - g(y)| \leq \sup_b |\lambda\{&\frac{E[V(x) - V(x - (U - C + b)^+)]}{\rho_g b} \\ &- \frac{E[V(y) - V(y - (U - C + b)^+)]}{\rho_g b}\}|. \end{aligned} \tag{5.26}$$

We conclude $g(s)$ *is continuous by the continuity of* $V(s)$.

5.6.2 Verification of the Optimal Strategy

Previously, we construct a function through recursion and prove the approach we construct the function is a solution to the HJB equation. However, whether the solution is the maximum survival probability one can obtain is not clear. In this subsection, we show that the approach we construct can achieve the optimal. Specifically, we show that the strategy derived from the minimizer $b(s)$ in equation (5.17) maximizes the survival probability.

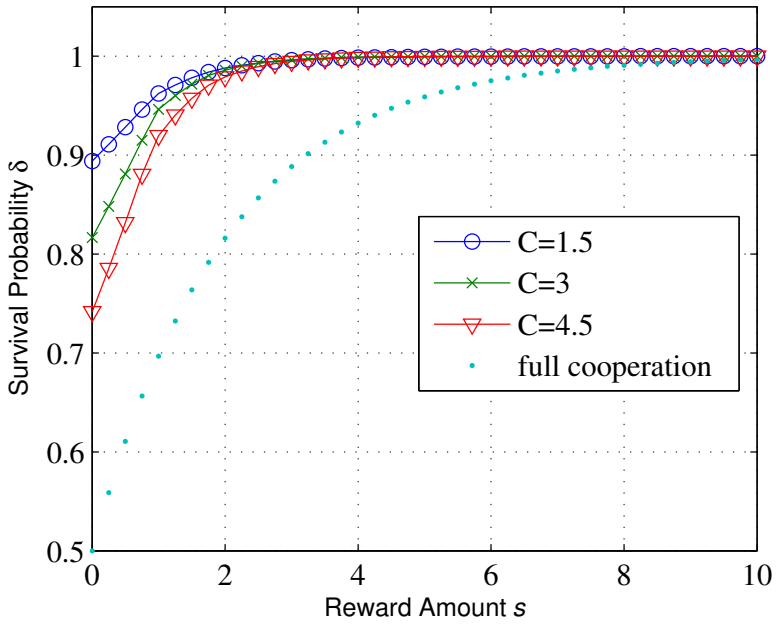


Figure 5.3: Survival Probability

theorem 5 *An optimal strategy, which maximizes the survival probability, is given by* $b^* = b(R^*)$*, where* $b(R)$ *is the strategy that minimize equation (5.17), and* R^* *is the reward process under the optimal strategy.*

The detailed proof to this theorem is similar to that in [112] chapter 2, and is omitted for brevity. The gists of the proof is to to show the solution to the HJB equation forms a martingale and those process derived by arbitrary strategies are supermartingales.

5.7 Numerical Results

In this section, we present the numerical results to demonstrate the effectiveness of our proposed framework and strategy in enhancing the reliability of private clouds for e-health applications. Simulation setup is first presented, followed by discussion on the numerical results.

Let request size U obeys exponential distribution, with mean $\frac{1}{m}$. Set $m = 1$. We omit the magnitude in this work. This simplification does not influence the study on

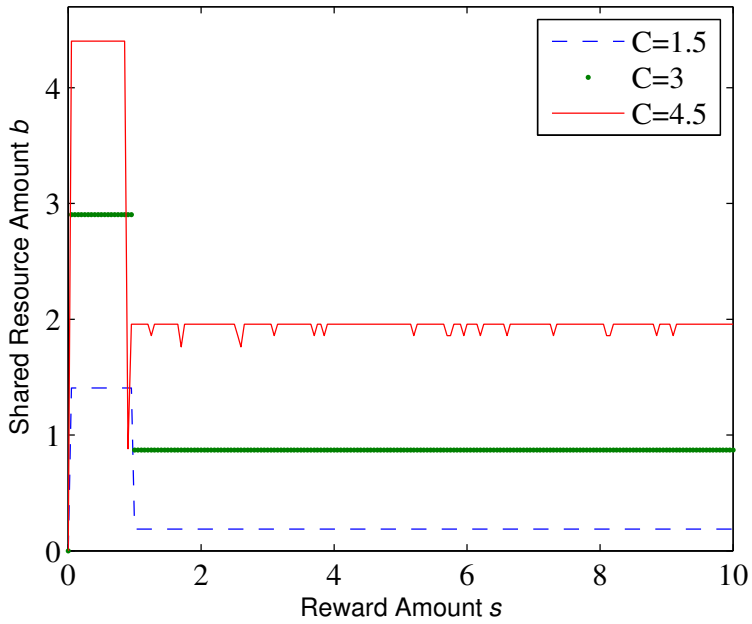
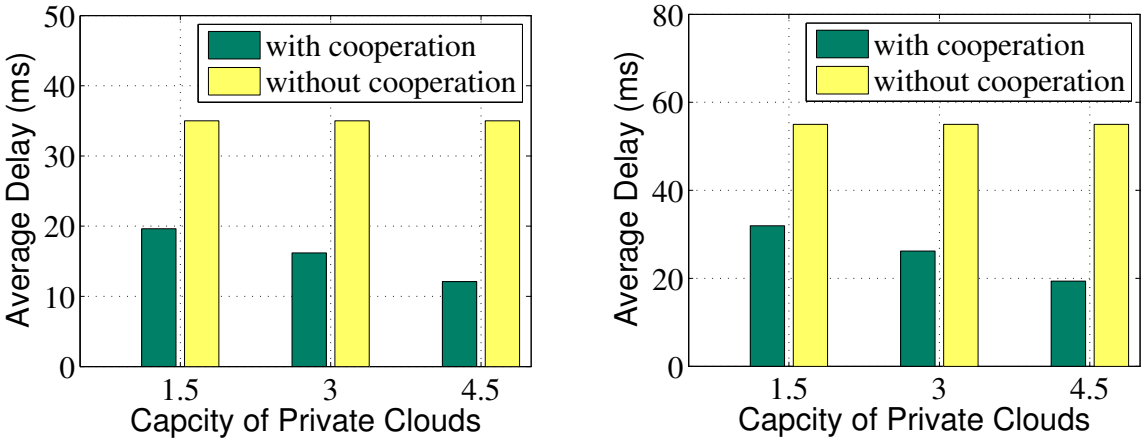


Figure 5.4: Optimal Cooperation Strategy

survival probability. Normalize the computing capacity and shared resource to the mean of the request size. The cooperation inclination scaling factor ρ_g is a decreasing function of computing capacity of each private cloud. In this work, we choose inverse function to describe the relationship, $\rho_g = \frac{1}{C}$.

The impact of the computing capacity of private clouds, C, on the survival probability δ is shown in Fig. 5.3. First of all, without cooperation, the survival probability of private clouds with bounded computing capacity is zero. It is clear that, our proposed framework demonstrates significant improvement with survival probabilities increase to at least 50%. The survival probability δ under optimal cooperation strategy increases as the reward amount s increases. When the computing capacity C is larger than average request size $\frac{1}{m}$, survival probability under optimal strategy is always better than that of full cooperation.

The cooperation inclination scaling factor ρ_g is always smaller than 1, which discounts the shared resource. It is reasonable for services provided by public clouds has longer delay than those provided by local servers. The survival probability for private cloud with computing capacity $C = 1.5$ is better than that with $C = 4.5$, namely, our strategy performs better when computing capacity is scarce. The reason is that given the same request



(a) Distance $L = 1500$km

(b) Distance $L = 2500$km

Figure 5.5: Delay Performance of Public Cloud

statistics, the private cloud with more computing resource values the service provided by public cloud less, thus has less incentives to cooperate. Practically, the reason for less cooperation inclination can be delay or privacy concerns.

The optimal strategies under different computing capacity amounts C, are shown in Fig. 5.4. The optimal strategy demonstrates itself as a threshold policy. Given the request process, a private cloud cooperates more when the reward amount s is below a threshold and less when s is beyond the threshold. The reason for the threshold coincides for different computing capacity C is that we choose the inverse function to describe the relationship between computing capacity and cooperation inclination scaling factor.

The average delay performances of public clouds are shown in Fig. 6.8, with distance between local region and the public cloud equal to 1500 km and 2500km, respectively. It can be observed from Fig. 5.5(a) and Fig. 5.5(b), the average delay under the proposed cooperation scheme is smaller than that without cooperation. It is the result that parts of requests are served by local private clouds. As shown in Fig. 5.5(a), with cooperation, the average delay decreases with the increase of the computing capacity of local private clouds. This is reasonable for private cloud with larger capacity shares more resources as shown in Fig. 5.4. Comparing Fig. 5.5(a) and Fig. 5.5(b), we can conclude that our proposed scheme achieve more delay reduction when the distance between local region and the public cloud is larger.

5.8 Summary

In this chapter, we have proposed a framework to enhance the reliability of private clouds in e-health applications. The framework exploits the time domain abundant resource of private clouds to motivate the public clouds to cooperate in improving the reliability of private clouds. Both private clouds and public clouds have proper incentives in the proposed scheme. The problem of how to allocate resource for private clouds to minimize failure probability has been investigated under random demands. A policy constructed through recursion is proved to provide optimal solutions. Numerical results have been provided to show the effectiveness of our proposed framework.

Chapter 6

Resource Allocation in Geo-distributed Clouds

In this chapter, we propose an e-health monitoring system supported by geo-distributed clouds. The geo-distributed clouds consist of many cloud servers which are geographically deployed over a large region [115]. The proposed e-health monitoring system consists of two parts, a resource management scheme for servers and a traffic shaping algorithm for users. The servers are initialized with the same resource management scheme. When users require to connect to the system, the local server (geographically-close to the users) handles the request and checks the workloads of other servers. It then runs the resource management scheme and responds to the users with the assigned servers. After receiving the responses, users apply a traffic shaping algorithm on their health data before transmitting the data to the assigned servers. Traffic shaping algorithm hides the original health data and preserves user privacy. Specifically, our contributions are twofold.

First, we propose a resource management scheme to achieve the minimized service delay and the reduced communication costs. We first derive a sufficient condition in resource management to ensure the stability of cloud servers. Considering this condition, we design the resource management scheme: each server only redirects the requests to others who have shorter queue lengths; and the number of redirected requests must be proportioned to the difference of their queue lengths and reciprocal to the service delay between them. We also prove the proposed resource management scheme satisfies the derived sufficient condition in balanced state. In addition, we compare the scheme with two other alternatives using joint the short queue (JSQ) and distributed control law (DCL), both of which are proven to be stable. Through extensive simulations, we show that our scheme achieves a much smaller average service delay than the JSQ-based and DCL-based schemes.

Second, we propose a traffic shaping algorithm to prevent the health data of users from being detected by the TA attackers [116]. We focus on the health data traffic generated by e-health monitoring systems, such as heart rate and blood pressure, which are typically modelled as deterministic processes [52]. We analyze the statistical differences between health data traffic and non-health data traffic. Our proposed shaping algorithm is designed such that: the distribution of the shaped health data traffic is the same as the distribution of the non-health data traffic; and the autocorrelation of the shaped health data traffic is close to the autocorrelation of the non-health data traffic. We propose to preserve the autocorrelations of the target process. Note that, the proposed algorithm introduces a delay, referred as shaping delay, on the user side which is related to the privacy requirement. We provide the numeric results on this relation. Then, we model the shaping delay by the D/M/1 queue, and consider the shaping delay into the resource management scheme. The simulation results show that our resource management scheme is still efficient with the shaping delay.

6.1 Literature Review

In this section, we review the related works in resource management and privacy preservation for e-health monitoring systems.

6.1.1 Resource management for Cloud Network

E-health monitoring systems have attracted great attention recently, and their applications have been developed widely [117, 118]. Due to the surging computing and storage demands from these applications, geo-distributed clouds have been regarded as promising solutions[119, 120]. In geo-distributed clouds for e-health monitoring systems, resource management acts as a critical component to provide timely and reliable services [106]. Previous works on resource management for geo-distributed clouds have two objectives: one is to reduce the service delay for users and the other is to reduce the cost for service provider. From a user's perspective, paper [47] proposed a centralized resource management scheme for geo-distributed clouds to minimize the service delay among selected servers, and a heuristic algorithm to partition a requested resource among the chosen servers. By exploiting the characteristics of social influences, paper [48] proposed an online resource management scheme to efficiently migrate contents, and redirect user requests to appropriate servers for timely responses. To reduce the operating cost for service providers, a

scheme that distributes requests among geo-distributed clouds to utilize the spatial differences in electricity price is proposed in [44]. For service providers, load balance is also an important requirement for its crucial role played to maintain the stability of all servers. As pointed out by paper [121], without proper resource management, requests may be redirected to a single server, leading to congestions. Paper [122] designed a distributed scheme for geo-distributed clouds, which stabilizes all the servers. In this paper, we study the resource management in geo-distributed clouds for e-health monitoring systems, where both average service delay and stability of clouds are considered as design objectives.

6.1.2 Privacy Preservation

The flourish of e-health monitoring systems faces the challenges in privacy preservation [123, 124]. TA attacks have been recognized as effective methods to reveal the type of users' health data [45]. Two countermeasures have been proposed, one algorithm is padding and the other algorithm is traffic shaping [125]. Padding algorithms obfuscate the packet length and rate by padding random amount of plaintext. The drawback of padding algorithms is that a large amount of bandwidth is required. In e-health monitoring systems, sensors on human body have limited energy and communication capabilities [16]. Thus padding algorithms are not suitable for e-health monitoring systems. Traffic shaping algorithm shapes the distribution of a traffic [125]. The key of this algorithm is to randomly sample a predefined matrix for the distribution transformation. As indicated by [126], this algorithm is not effective in preserving a user's privacy for it does not consider the time dependency of a random process. In this paper, we plan to design an efficient traffic shaping algorithm for e-health monitoring systems by addressing the above problems.

6.2 System Model

We consider geo-distributed clouds in e-health monitoring systems. As shown in Fig. 6.1, N cloud servers locate in different geographic regions. Each region $\{1, \cdots, N\}$ has one server. The server in i-th region is denoted by S_i. The service capacity is evaluated by the number of virtual machines (VMs) a server has. Thus, the service capacity is limited. We consider time is slotted. At different time slot, servers have different available service capability due to the dynamic allocation. Let $\mu_i(t)$, $i \in \{1, ..., N\}$ denote the available service capacity of server S_i during time slot t.

The service requests for a server come from the users in the region that the server locates. Let $Q_i(t)$ denote the number of waiting requests (queue length) of server S_i at

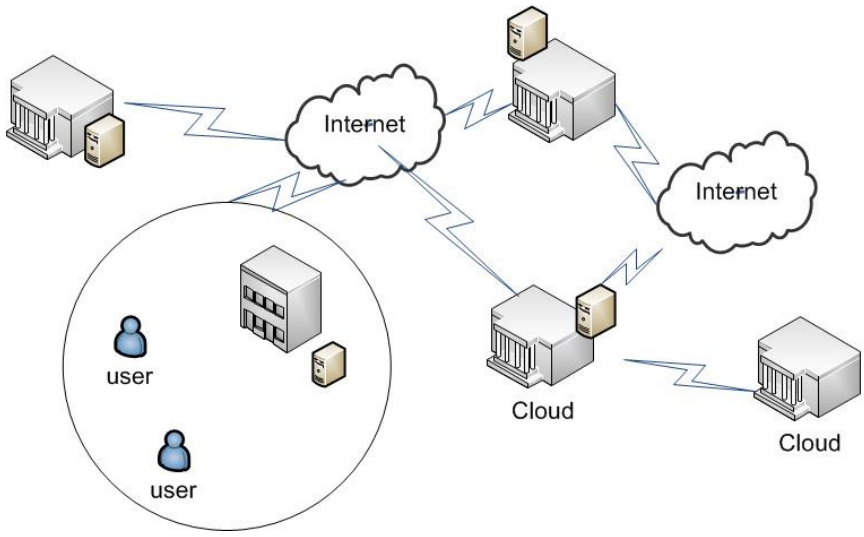


Figure 6.1: Geo-distributed Clouds Environment

the beginning time slot t. Let $A_i(t)$ denote the number of arrival requests for the server S_i during time slot t. $A_i(t)$ is considered as a Poisson process with arrival rate λ_i. For each arrival request, it contains certain amount of traffic. We consider the traffic from one request is a deterministic process with constant traffic arrival rate λ_m [52].

The service delay for any request includes two parts. One is the shaping delay due to the traffic shaping algorithm. The other one is the communication delay. Let D^p denote the shaping delay by the traffic shaping algorithm, and $D^c_{i,j}$ denote the communication delay between region i and region j. As indicated by [47], the communication delay in geo-distributed clouds cannot be negligible. We consider the communication delay over the Internet by measuring the geographic distance, i.e., communication delay increases linearly with the geographic distance [44]. Let $L_{i,j}$ denote the distance between region i and region j. From [44], we consider the slop of the linear function as

$$\frac{\delta(time)}{\delta(distance)} \approx 0.02ms/km \tag{6.1}$$

The communication delay $D^c_{i,j}$ can be calculated as:

$$D^c_{i,j}(ms) = 0.02(ms/km) \times L_{i,j}(km) + 5(ms). \tag{6.2}$$

We consider the TA attacks in the geo-distributed clouds environment. Such attacks aim to analyze the traffic statistics to determine the specific type of the health data. We

measure the capability of TA attacks by using the Kullback-Leibler (K-L) divergence [127]. K-L divergence is also referred as relative entropy to measure the difference between two probability distributions. Let P denote the distribution of health data traffic and Q denote the distribution of non-health data traffic. Let $D_{KL}(P||Q)$ denote the K-L divergence. We have

$$D_{KL}(P||Q) = \int \ln\{\frac{f_p(x)}{f_q(x)}\} f_p(x)dx, \tag{6.3}$$

where $f_p(x)$ and $f_q(x)$ are the probability density functions of distributions P and Q, respectively. The K-L divergence reaches its minimum when $f_p(x) = f_q(x)$. When two distributions P and Q are the same, the capability of TA attacks is reduced to the minimum.

6.3 An E-Health Monitoring System

In this section, we propose an e-health monitoring system with minimum service delay and privacy preservation. The system consists of two parts, the traffic shaping algorithm and the resource allocation scheme. The traffic shaping algorithm converts the health data traffic to non-health data traffic such that the capability of the TA attacks is largely reduced. The resource allocation scheme considering load balance as a necessary condition aims to minimize the service delay.

6.3.1 Traffic Shaping

In this subsection, we propose an effective traffic shaping algorithm to preserve users' privacy against TA attacks. We choose voice traffic as target traffic for two reasons: different from other common internet traffic, voice traffic is not heavy tailed and thus consumes less bandwidth; and voice traffic is given higher priority than data traffic in communication protocols [128], which helps health data to reduce the medium access time when competing with other traffic.

We demonstrate why existing traffic shaping algorithm is not suitable for time dependent random process. Consider a voice source. Due to its characteristics that voice source could be divided into talk spurt and silent period, voice traffic is modeled using ON-OFF model. Let α and β denote the average ON and OFF period of voice, respectively. During the talk spurt, the voice source generates packets with length L_{voice} with packet inter-arrival time t_a, whereas during silent period, no packet is generated. In existing traffic shaping algorithm in use, the distribution of voice traffic is preserved in the following way.

Whenever a packet is available for transmission, with probability $\frac{\alpha}{\alpha+\beta}$ the packet is transmitted. Thus, the probability of the output traffic in ON state is $\frac{\alpha}{\alpha+\beta}$, which is the same as the voice traffic. However, the average length of ON period might not be α. Namely, the time dependency feature of voice traffic is not shown in the shaped traffic. As a result, a TA attacker could use autocorrelation to distinguish the shaped traffic from real voice traffic.

Traffic Shaping Algorithm

Since the arrival rate of health data λ_m could be larger than that of a single voice source, the target traffic could be a traffic containing multiple voice sources. Given a constant health data arrival rate λ_m, a user first decides the number of voice traffic in the target traffic, denoted by N_v. The choice of N_v shall satisfy the condition that the average traffic rate of the target traffic should be no less than the average arriving rate of health data. Otherwise, the shaping delay caused by this algorithm could not be limited, since the departure rate is less than the arrival rate. Given the utilization factor ρ_v of a voice traffic equals to $\frac{\alpha}{\alpha+\beta}$ and the traffic rate λ_v of a voice traffic during talk spurt, the number of voice traffic should satisfy:

$$N_v \geqslant \lceil \frac{\lambda_m}{\rho_v \lambda_v} \rceil, \tag{6.4}$$

where $\lceil x \rceil$ is the minimum integer greater than x.

After choosing the number of voice sources N_v, the user accumulates traffic in its buffer and then transmits them according to the traffic generation rate of N_v voice sources. The traffic rate of N_v voice sources is a binomial process with each voice source in ON state with probability $\frac{\alpha}{\alpha+\beta}$. Thus, the probability of the traffic generating rate equals $i\lambda_v$ is

$$\Pr\{r = i\lambda_v\} = C_{N_v}^{i} (\frac{\alpha}{\alpha+\beta})^{i} (\frac{\beta}{\alpha+\beta})^{(N_v - i)}, \tag{6.5}$$

where $C_{N_v}^{i}$ is equal to $\frac{N_v!}{i!(N_v-i)!}$. For each time slot, a user chooses a traffic generating rate based on the probability described by equation (6.5), and uses the rate to transmit.

6.3.2 Resource Allocation

In this subsection, we design a resource allocation scheme for the e-health monitoring systems with stabilized server queues and reduced service delay.

A server receives requests from the users in the local region and performs the resource allocations. Specifically, the server first collects the queue length $Q_j(t)$ from other servers $j \in \{1, ..., N\}$. Considering the service delay and the queue length, the server then determines the allocation strategy where some requests will be redirected to other servers. Let $A_i^U(t)$ denote the number of request arriving at server S_i during time slot t after the redirections are made. The queue length of server $i \in \{1, ..., N\}$ at time slot $t+1$ can be represented as

$$Q_i(t+1) = \max[Q_i(t) + A_i^U(t) - \mu_i(t), 0]. \tag{6.6}$$

Then, the local cloud feedbacks its decision to end users. Each user directs its traffic directly to the assigned server.

Resource Allocation Constraints

We present the stabilization concept and explain the importance of stabilizing all servers. Based on the stability condition, to ensure the stability of server S_i, we need resource allocation scheme such that

$$E_t[A_i^U(t)] \leqslant E_t[\mu_i(t)], \tag{6.7}$$

where $E_t[x]$ is the expectation of random process x over t.

In e-health monitoring systems, failure or overload of any server could cause fatal results. Thus, the resource allocation scheme for health data must achieve the stability condition for all servers, i.e., the equation (6.7) needs to be satisfied for any $i \in \{1, ..., N\}$. To design resource allocation scheme satisfying above conditions is difficult. For each server could redirect parts of its requests to other servers, which requires a scheme to consider the interactions among different servers.

To solve this problem, we first investigate a sufficient condition to achieve stability. Considering this condition, we then propose a delay aware algorithm.

A Sufficient Condition

We derive a sufficient condition, which ensures the stability for all servers in the geo-distributed clouds environment. We start from the definition of stability.

Definition 1 *Suppose a process $q(t)$ has an equilibrium q_e, if for every $\epsilon > 0$, there exists a $\delta = \delta(\epsilon) > 0$ such that, if $||q(0) - q_e|| < \delta$, then $||q(t) - q_e|| < \epsilon$, for every $t \geqslant 0$.*

From definition 1, we can see that, if a process $q(t)$ reaches q_e at time slot n, namely $q(n) = q_e$, any scheduling policy that guarantees $q(n+1) - q(n) = 0$ can stabilize $q(t)$. Thus, $\Delta q = q(n+1) - q(n) = 0$ is a sufficient condition for a process to achieve stability.

Let a vector $[Q_1(t), ..., Q_N(t)]'$ be the queue length of the interactive servers, denoted by $\overrightarrow{Q(t)}_{1\times N}$. The sufficient condition could be represented as:

$$\Delta\overrightarrow{Q(t)} = \overrightarrow{Q(t+1)} - \overrightarrow{Q(t)} = \vec{0}, \text{given } \overrightarrow{Q(t)} = \overrightarrow{Q_e}. \tag{6.8}$$

Since our purpose is to design a resource allocation scheme that makes decisions based on current queue length, the change in queue length of all servers can be presented by

$$\Delta\overrightarrow{Q(t)} = \mathbf{U}\overrightarrow{Q(t)}, \tag{6.9}$$

where $\mathbf{U}$ is a $N \times N$ matrix, and $U_{i,j}$ represents the number of requests that server S_i redirected to server S_j. Based on this interpretation, the sufficient condition for multiple interactive queues (6.8) can be written as

$$\mathbf{U}\overrightarrow{Q_e} = \overrightarrow{0}. \tag{6.10}$$

Equation (6.10) could also be written as a system of equations

$$\begin{cases} U_{11}Q_{e1} + U_{12}Q_{e2} + ... + U_{1N}Q_{eN} = 0 \\ \vdots \\ U_{11}Q_{e1} + U_{12}Q_{e2} + ... + U_{1N}Q_{eN} = 0 \end{cases} \tag{6.11}$$

To design a resource allocation scheme, which ensures the stability of all servers, is to determine the value of each U_{ij} such that the system of equations (6.11) is satisfied. The system of equations (6.11) has N equations and $N \times N$ unknown. Thus, there is more than one solution to equations (6.11). In other word, there are multiple schemes, which can satisfy the constraints.

Resource Allocation Scheme

In this subsection, we design a resource allocation scheme satisfies equations (6.11). The idea is based on the fact that 0 is an eigenvalue of any Laplacian matrix corresponding to vector $[1, 1, ..., 1]_{1\times N}$ [122, 129]. In addition, the resource allocation scheme is designed to minimize service delay.

In a resource allocation scheme, we need to design a matrix $\mathbf{U}$ to satisfy condition (6.10). We present how eigenvalue could facilitate scheme design. An eigenvector of a square matrix $\mathbf{U}$ is a vector $\overrightarrow{e_u}$ satisfies:

$$\mathbf{U}\ \overrightarrow{e_u} = m_u \overrightarrow{e_u}, \tag{6.12}$$

where m_u is the corresponding eigenvalue. For Laplacian matrix, 0 is always a eigenvalue corresponding to $\overrightarrow{1}$. Consider an equilibrium state of all servers is balanced, namely $\overrightarrow{Q_e} = q_e \overrightarrow{1}$, then any $\mathbf{U}$, which is a Laplacian matrix, can stabilize all servers in the geo-distributed clouds. Thus, designing a resource management scheme such that $\mathbf{U}$ is a Laplacian matrix can achieve load balance. Based on this observation, we design a resource management scheme in the following.

The scheme design utilizes two facts: to stabilize all servers, the server with shorter queue length shall serve more requests; to reduce delay, a request prefers servers with less service delay. Thus, a good design should have two characteristics: requests should only be directed to servers with shorter queues; and the amount of redirected requests shall be an increasing function of queue length difference, and be a decreasing function of service delay.

6.4 Performance Analysis

This section evaluates the performances of our proposed traffic shaping algorithm and resource management scheme.

6.4.1 Performance of the Traffic Shaping Algorithm

In this subsection, we present the analysis of the shaping delay and the privacy preservation of the proposed traffic shaping algorithm.

Shaping Delay Performance

The arrival rate of health data is a constant, whereas the departure rate of the shaping algorithm is a random process obeys binomial distribution. Since the binomial distribution converges to a Poisson distribution as the number of tests goes to infinity, we approximate the service process as a poisson process in this work. Based on this approximation, the

Algorithm 2: Resource Management Scheme

1) For each server S_i: the server measures the communication delay $D^c_{i,j}$ between itself and server S_j for all servers $j \in [1, N], j \neq i$;
2) Based on link information, each server S_i sets the the privacy requirements $D^{i,j}_{KL}$, in terms of K-L divergence, for different link $L_{i,j}$, then calculates the shaping delay $D^p_{i,j}$ incurred for privacy preservation requirements. Specifically, in our proposed privacy preservation scheme, a server S_i chooses the number of voice traffic N_v such that

$$D_{KL}(N_v) \leqslant D^{i,j}_{KL}; \tag{6.13}$$

a server S_i uses the result from equation (6.13) to calculate the shaping delay $D^p_{i,j}$ for each link.
3) A server S_i calculates the service delay for accessing each server S_j, $D_{i,j} = D^c_{i,j} + D^p_{i,j}$.
4) A server S_i updates the buffer length information of other servers $Q_j(t)$, and redirects request according to

$$Q_i(t+1) = Q_i(t) + \sum_j M_{i,j}(t), \tag{6.14}$$

where

$$M_{i,j}(t) = \begin{cases} \frac{(Q_i(t)-Q_j(t))A_i}{D_{i,j}M_i^{max}} & \text{Otherwise} \\ -M_{j,i} \text{ for } j \in (Q_i(t) < Q_j(t)), \end{cases} \tag{6.15}$$

where $M_i^{max} = \sum_{j \in (Q_i(t) > Q_j(t))} \frac{(Q_i(t)-Q_j(t))}{D_{i,j}}$.

delay introduced by the traffic shaping algorithm could be analyzed through a D/M/1 queue. Based on the analysis in [130], the queue stationary distribution given utilization factor $\rho = \frac{\lambda_m}{N_v \lambda_v \rho_v} < 1$ is given by

$$\pi_i = \begin{cases} 0, & \text{when } i = 0; \\ (1-\delta)\delta^{(i-1)}, & \text{when } i > 0, \end{cases} \tag{6.16}$$

where δ is the smallest absolute value of all solutions to equation

$$\rho = -\frac{1-\delta}{\ln \delta}. \tag{6.17}$$

Further, the average shaping delay introduced by the shaping algorithm could be calculated based on equation (6.16). Let $D_m(N_v)$ denote the average shaping delay with N_v voice sources; it can be calculated as[130]:

$$D_m(N_v) = \frac{1}{N_v \rho_v \lambda_v} \frac{\delta}{1-\delta}. \tag{6.18}$$

Privacy Preservation

In the following, we present the analysis of the privacy preservation capability, which is measured by K-L divergence, of our proposed traffic shaping algorithm.

To analyze the privacy preservation performance of our proposed traffic shaping algorithm based on K-L divergence, we need to know the distribution of the target traffic and the distribution of our algorithm output. The target traffic obeys Poisson as discussed before, whereas the distribution of the output is unknown. We present the analysis on the distribution of the output as follows.

The output of our proposed algorithm has the same distribution as that of the output of D/M/1 queue. Given utilization factor $\rho < 0.2$, the output of D/M/1 queue is not Poisson, namely the distribution of the time between two consecutive departure does not obey exponential distribution [131]. Low utilization factor ρ represents a high service rate compared to arrival rate, thus smaller average waiting time. However, in this case, the leakage risk is high, since the output deviates from the target traffic significantly. As utilization factor ρ goes from 0.2 to 1, the difference between the output and Poisson process diminishes, and is 0 when ρ is 1 [131]. We do not consider the situation where utilization factor $\rho > 1$, for the queue is stable under this condition. Motivated by above

facts, we adopt Poisson process to approximate the output of our proposed traffic shaping algorithm for utilization factor $\rho \in (0.2, 1)$.

Based on above analysis, the privacy preservation performance of our proposed traffic shaping algorithm can be described by the K-L divergence of two Poisson processes. Since the inter-arrival time is the identity of a Poisson process, we consider the K-L divergence of two exponential distributions, which represents the inter-arrival time of the algorithm output and the target traffic, respectively. The average inter-arrival time of the algorithm output is equal to that of the input, namely the average inter-arrival time of the health data $\frac{1}{\lambda_m}$. The average inter-arrival time of the target process is $\frac{1}{N_v \rho_v \lambda_v}$. The K-L divergence between them is

$$\begin{aligned} D_{KL}(P||Q) &= \int_0^\infty \lambda_p e^{-\lambda_p x} \ln(\frac{\lambda_p e^{-\lambda_p x}}{\lambda_q e^{-\lambda_q x}}) dx \\ &= \int_0^\infty \lambda_p e^{-\lambda_p x} (\ln(\frac{\lambda_p}{\lambda_q}) + (-\lambda_p + \lambda_q)x) dx \\ &= \ln(\frac{\lambda_p}{\lambda_q}) \int_0^\infty \lambda_p e^{-\lambda_p x} dx + E_p[(-\lambda_p + \lambda_q)x] \\ &= \ln(\frac{\lambda_p}{\lambda_q}) + \frac{-\lambda_p + \lambda_q}{\lambda_p}, \end{aligned} \tag{6.19}$$

where $\lambda_p = \lambda_m$ and $\lambda_q = N_v \rho_v \lambda_v$.

6.4.2 Performance of the Resource Management Scheme

In the following, we show that the resource scheme proposed could stabilize all servers.

theorem 6 *If the network operates under our proposed algorithm 2, the network will stay in balanced state.*

Proof 4 *We prove theorem 2 through showing the equation (6.14) satisfies the sufficient condition (6.9).*

Equation (6.14) could be written as

$$Q_i(t+1) - Q_i(t) = \sum_j M'_{i,j}(t)(Q_i(t) - Q_j(t)), \tag{6.20}$$

where $M'_{i,j}(t) = \frac{M_{i,j}(t)}{(Q_i(t) - Q_j(t))}$.

The equation (6.20) could also be written in the form of equation (6.9). Thus, the relationship between $M'_{j,i}$ and $U_{i,j}$ could be described by

$$\sum_j U_{i,j}(t)Q_j = \sum_j M'_{i,j}(Q_i(t) - Q_j(t)). \tag{6.21}$$

Solve equation (6.21), we obtain

$$U_{i,j}(t) = \begin{cases} -M'_{i,j}, & forj \neq i \\ \sum_{j \neq i} M'_{i,j}(t) & forj = i. \end{cases} \tag{6.22}$$

Based on equation (6.15) and equation (6.22), it is easy to verify the matrix $\boldsymbol{U}$ generated under our propose scheme is a Laplacian matrix. As a result, condition (6.10) is satisfied when the distributed clouds are in balanced state.

6.5 Performance Evaluation

In this section, we evaluate our proposed traffic shaping algorithm and resource management scheme through simulations. We first choose to compare our proposed traffic shaping algorithm with an existing traffic shaping algorithm. We are interested in autocorrelation feature preservation, and the tradeoff between delay performance and privacy preservation ability. Then we compare the proposed resource management scheme with the JSQ-based approach [132] and DCL-based approach [122]. We are interested in delay performance and the queue length performance.

6.5.1 Simulation Setup

We consider a geo-distributed clouds environment where cloud servers are deployed in Canada. As shown in Fig. 6.2, the servers are placed on $N = 17$ cities (regions) in Canada. The distance between any two regions $L_{i,j}$ for $i, j \in \{1, ...N\}$ are measured using Google Maps. The request arrival rate of i-th server λ_i is chosen to be proportional to the population of the region, whereas the service rate is chosen to be proportional to the number of hospital of that region. In order to evaluate our resource management schemes, we initialize servers with different queue lengths. Note that, a saturated network is defined [122] when

$$\sum \lambda_i = \sum \mu_i. \tag{6.23}$$

Figure 6.2: Map of Major Cities in Canada

We slightly increase the arrival rates such that the above equation (6.23) could be satisfied. The detailed settings can be found in Table 6.1.

For each service request, we set the arrival rate as 30kbps [52]. For traffic shaping target, we choose a coded version voice traffic according to GSM 6.10 codec. The average duration of ON state, OFF state and the average arrival rate are listed in Table 6.2.

In existing traffic shaping algorithm, a matrix, which is designed based on the distributions of the source traffic and target traffic, is sampled randomly to shape the traffic distribution. To shape a medical traffic with a constant arrival rate into ON-OFF traffic, a vector $[1, 1]$ is sample based on probability $[\frac{\alpha}{\alpha+\beta}, \frac{\beta}{\alpha+\beta}]$.

In the JSQ scheme [132], the server with the shortest queue is chosen to serve the users. Specially, server i chooses the j^*th server to redirect its requests, based on

$$j^* = \text{argmin}_{j\in\{1,\cdots,N\}} Q_j(t). \tag{6.24}$$

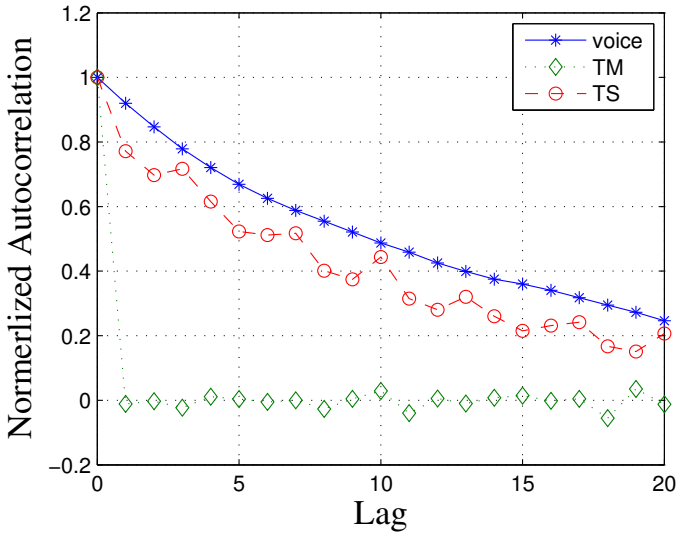


Figure 6.3: Autocorrelations of Voice, TM and TS

In the DCL algorithm, the amount of traffic from server j to server i is calculated as

$$U_{i,j}(t) = \frac{(Q_i(t) - Q_j(t))A_i}{\sum_{j \in N_{i-}} (Q_i(t) - Q_j(t))}. \tag{6.25}$$

6.5.2 Traffic Shaping Algorithm Evaluation

In this subsection, we provide simulation results to show: 1) our proposed traffic shaping algorithm can preserve the autocorrelation features of the target process; and 2) there is a tradeoff between shaping delay and privacy leakage risk. We use TM and TS to denote the traffic shaping algorithm in [125] and in this work, respectively.

Fig. 6.3 shows simulation results of the normalized autocorrelation of voice traffic, a medical traffic shaped by TM, and shaped by TS with lags no larger than 20. The results show that TS outperforms TM significantly in preserving the autocorrelation features of voice traffic. As it can be observed that, the autocorrelation of the output of algorithm TM is almost 0 when lag is larger than 0. That is to say, the time dependency feature of a target process, in terms of autocorrelation, is not preserved in TM, as explained in Sec. IV part A. In comparison, the autocorrelation of TS is similar to that of a target traffic. The reason is that, TM shapes the traffic based on the time dependent features of a target

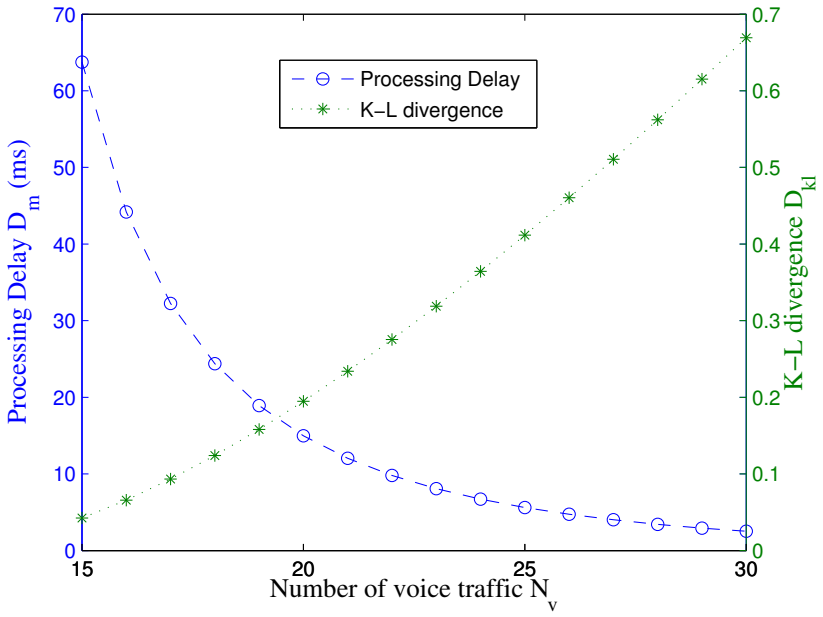


Figure 6.4: Tradeoff Between Privacy Preservation and Shaping Delay

traffic. Specifically, TM shapes the medical traffic to mimic the ON-OFF behaviour of voice traffic.

Consider a TA attacker runs a classifier, which chooses the changing rate of the autocorrelation as the classification characteristic [133]. As we can observe from Fig. 3., the decreasing rate of the autocorrelation of voice increases slowly and smoothly. The decreasing rate of autocorrelation of TS is similar as that of voice only with small turbulences. In comparison, the autocorrelation of TM decreases sharply and remains almost constant. In this case, the TM could be identified by the TA attacker, whereas TS is hard to detect. When a classifier is adopted by a TA attacker, the autocorrelation between the shaped traffic and the voice traffic needs to have significant similarity to avoid being identified[134]. Thus, we can conclude the improvement of TS over TM in terms of autocorrelation is important for privacy preservation.

Fig. 6.4 shows the tradeoff between shaping delay and privacy leakage risk using our proposed traffic shaping scheme. The results are obtained through numerical simulation based on the analysis on the average shaping delay and privacy preservation capability in Section V. It can be observed that as the increase of the number of the chosen voice source, the shaping delay of our proposed algorithm decreases, where as the K-L divergence

increases. The reason is that, when the number of voice source adopted increases, the number of voice traffic that are in ON state increases, leading to shorter time for the health data waiting to be packeted and transmitted. However, in this case, the probability of insufficient health data packets increases at the same time, leading to a larger K-L divergence, i.e., higher privacy leakage risk.

6.5.3 Resource Management Scheme Evaluation

In this subsection, we provide simulation results to demonstrate two benefits of our proposed scheme: 1) reduce the delay suffered by traffic; and 2) achieve load balance among different servers. We use RAS denote the resource management scheme designed in this work.

Average Service Delay

The service delay performances of three algorithms, namely JSQ, DCL and RAS, are shown in Fig. 6.5. It can be seen from Fig. 6.6, the average service delay of all requests under three algorithms are 120ms, 87ms and 56ms, respectively. The reason why our algorithm has smaller average service delay is that, in our algorithm, the amount of requests redirected to other clouds is reciprocal to the service delay among two clouds. This method limits the number of requests to be redirected to a remote cloud server, thus introducing less delay for the requests. The average service delay suffered by the requests to each cloud under algorithm JSQ, DCL and RAS, are shown in Fig. 6.5(a), Fig. 6.5(b) and Fig. 6.5(c), respectively. As we can observe that, the average service delay for requests to each cloud under the JSQ is higher than that under DCL and RAS. The reason is that, JSQ always pours all the requests to the cloud with smallest queue length. Thus, when the cloud with smallest queue length is far away, the delay is significant large. In comparison, both DCL and RAS redirect requests to all other servers with smaller queue length, thus avoid the situation to direct all requests to the remote cloud.

Queue Length

The average queue length for all clouds under algorithm JSQ, DCL and RAS, are shown in Fig. 6.7. It can be seen that, the average queue length under JSQ is higher than that under DCL and RAS. The reason is that, JSQ is designed to maximize the throughput of the distributed clouds. So its algorithm is designed to avoid the situation where any buffer

is empty. Thus, the average queue length is the highest. We can also observe that, the average queue length under RAS is comparable to the average queue length under DCL. This proves the ability of our proposed algorithm in stabilizing the cloud networks.

The queue dynamics of all clouds under the algorithm JSQ, DCL and RAS, are shown in Fig. 6.8. The queue dynamics over each iteration for cloud at St. John, Quebec, Toronto and Regina are shown in Fig. 6.8(a), Fig. 6.8(b), Fig. 6.8(c) and Fig. 6.8(d), respectively. In can be seen that, compared to JSQ, both DCL and RAS perform better in terms of eliminating backlogs and ensuring the stability of all clouds. And our proposed algorithm is comparable to DCL, in terms of maintaining the stability of all clouds.

6.6 Summary

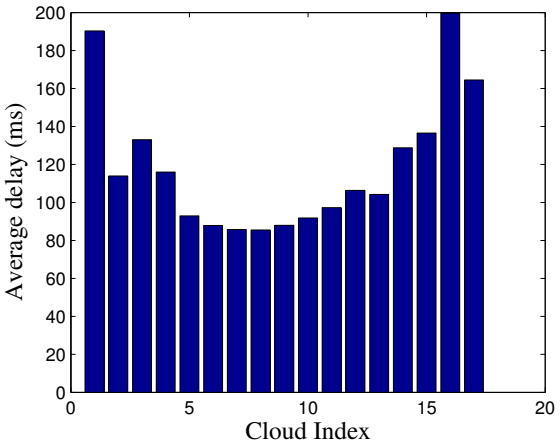
In this chapter, we have explored geo-distributed clouds to propose an e-health monitoring system with minimum service delay and privacy preservation. We have provided the numerical analysis and simulation results to demonstrate the effectiveness of the system. For our future work, we will extend this work by studying a more general and complicated case where users have random medical requests and diverse privacy preservation requirements.

Table 6.1: Network Parameters

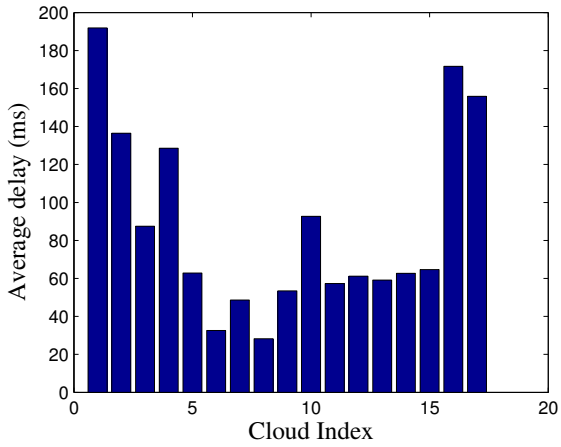
City Name	arrival rate	service rate	initial buffer
S.t. Johns	2	2	200
Charlottetown	1	1	100
Halifax	4	9	50
Fredericton	1	1	250
Quebec	6	7	200
Montreal	17	33	100
Ottawa	9	9	200
Toronto	55	24	100
Winnipeg	7	9	50
Regina	2	2	250
Saskatoon	3	3	100
Edmonton	11	12	100
Calgary	11	11	250
Vancouver	23	23	200
Victoria	3.5	8	100
Whitehorse	0.3	1	50
Yellowknife	0.2	1	150

Table 6.2: Traffic Parameters

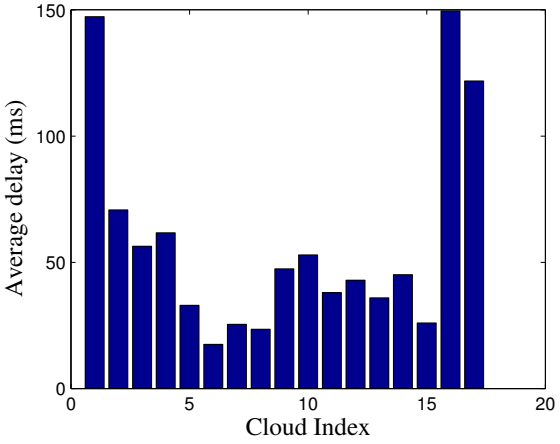
Parameter	Value	Parameter	Value
α	352 ms	β	650 ms
A_v	4kbps	A_m	30kbps



(a) Delay performance of JSQ



(b) Delay performance of DCL



(c) Delay performance of RAS

Figure 6.5: Average Service Delay Performance

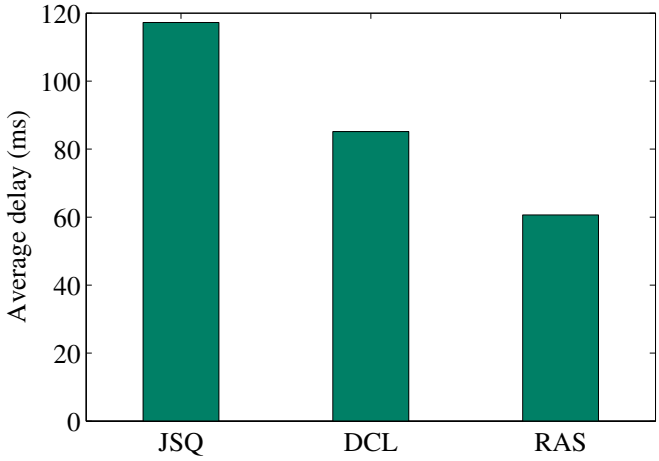


Figure 6.6: Average Service Delay Comparison

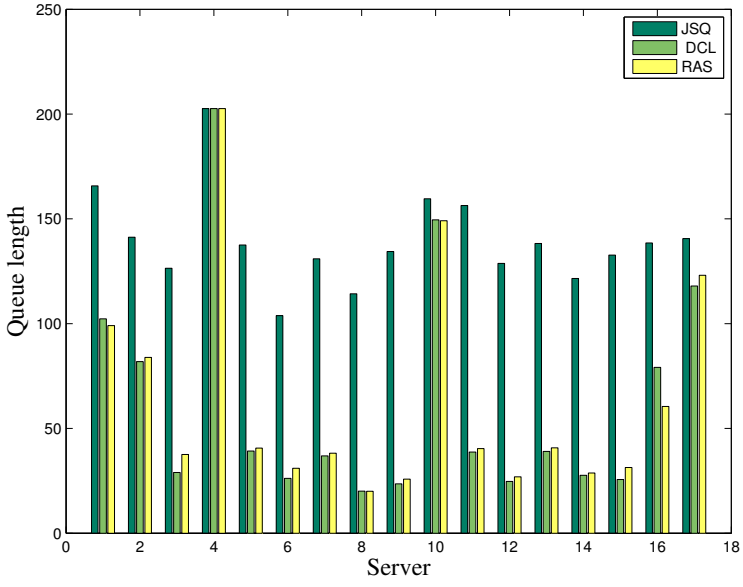
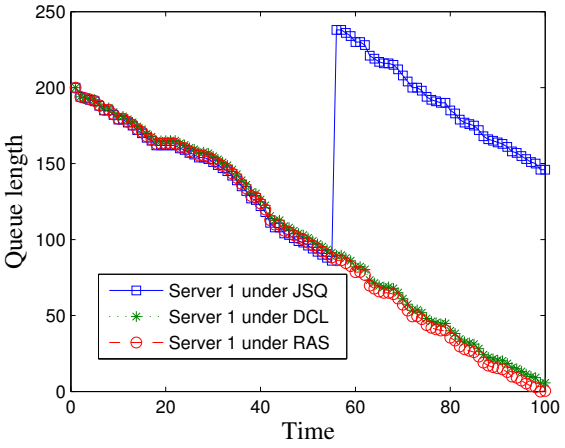
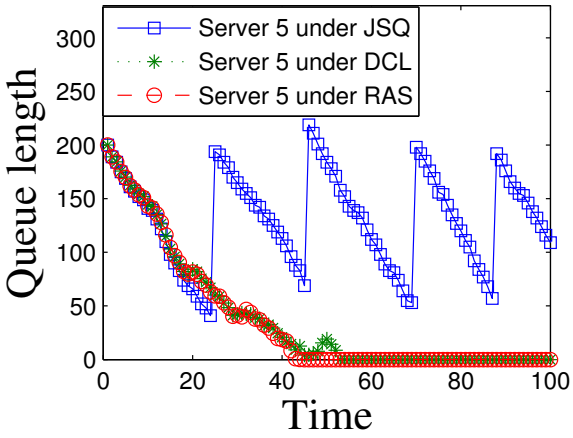


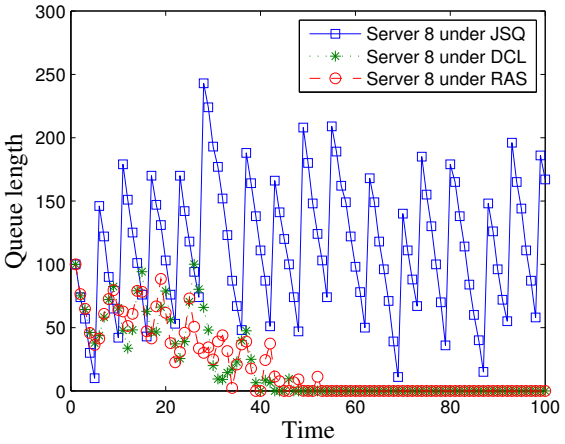
Figure 6.7: Comparison of Average Queue Length of JSR, DCL and RAS



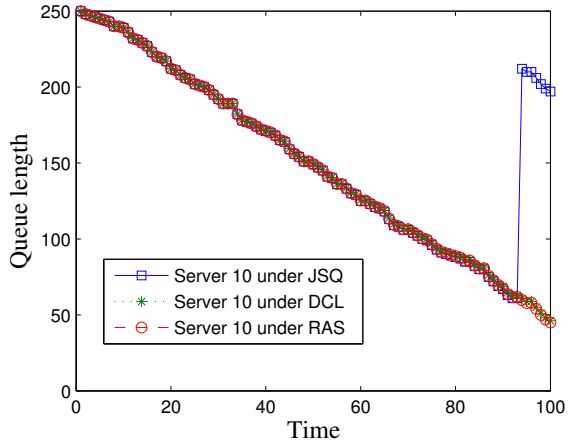
(a) Server at St. John



(b) Server at Quebec



(c) Server at Toronto



(d) Server at Regina

Figure 6.8: Queue Dynamics

Chapter 7

Conclusions and Further Work

In this final chapter, I will revisit and summarize the topics discussed so far, then discuss a few ideas for moving the research on resource allocation in e-health systems forward. Section 7.1 summarizes the main results of this thesis. Section 7.2 describes a few possible avenues of future research that could grow out of the work presented here.

7.1 Major Research Results

Motivated by the importance and challenges of e-healthcare systems, this research seeks to develop novel, practical and effective resource management schemes to provide efficient, reliable and low-cost healthcare services. Three research topics have been studied, namely energy efficient transmission power allocation for wearable sensors with QoS provisioning, medium access control for WBANs with interference management and throughput maximization, and resource management in clouds networks for reliability improvements and delay minimization. The proposed schemes are able to address challenges due to: 1) the limited capability of wearable devices; 2) dynamic body area channel in WBANs; 3) the randomness of computing requests and communication delay over in cloud networks. Specifically, the main contributions of this research are summarized as follows.

- Transmission power allocation with worst-case delay provisioning: We have proposed a transmission scheduling scheme for vital physiology signals with worst-case delay provisioning. The proposed scheme achieves the delay provisioning by considering a virtual queue, which increases when the actual queue is not empty. Meanwhile,

the energy efficiency for sensor is improved through reducing energy consumption during idle listening and utilizing opportunistic channel access through transmitting over channel in good state. The proposed scheme is derived based on Lyapunov optimization framework. The conditions for our algorithm to have a worst-case delay limit are studied. The trade-off between energy consumption and the worst-case delay is investigated in performance analysis and showed in numerical results. We expect the algorithm developed to inspire the transmission scheduling for wearable devices with vital physiology monitoring function.

- The impacts of the peak transmission power and statistical QoS provisioning: We have derived the optimal transmission power allocation scheme under a peak power constraint, and proposed an efficient calculation method. Applying duality gap analysis, we characterize the upper bound of the extra average transmission power incurred due a peak power constraint. Through the analysis of the upper bound, we conclude that when the peak power constraint is stringent, a proposed constant power scheme is suitable for wearable sensors for its performance is close to optimal. Further, we show that the peak power constraint is the bottleneck for wearable sensors to provide stringent statistical QoS provisioning.

- MAC for WBANs over Dynamic Body Area Channel: We have proposed a centralized MAC scheme for WBANs in hospital. Due to partial information of the channel state of individual WBANs, we formulate a partial observable optimization problem for network throughput optimization. We investigate two properties of the network, namely time dependency of the channel states and buffer occupancy which depends on traffic arrival and departure process. Based on above network characteristics, a modified myopic policy is proposed to address the fairness issues of a myopic policy. The performance of the algorithm is evaluated under both unsaturated network and congested network conditions. Compared with existing approaches such as Round Robin scheme, our proposed algorithm can significantly improve the network throughput and enhance channel utilization.

- Reliability improvement of private clouds through cooperation: We have proposed a cooperation framework to address the distinct challenger facing different clouds. It is inspired by the fact that private clouds are geographically deployed and public clouds can be considered to possess infinite computing resources. In our framework, private clouds are designed to serve parts of local requests for public clouds, and rewarded by receiving help with excess requests. We adopt stochastic control theory to address the failure minimization issues for private clouds under random demand process. We

prove the optimality of a policy constructed through recursion. Numerical and simulation results are presented to demonstrate that our proposed scheme can improve the reliability of private clouds, as well as reduce average delay of public clouds.

- Delay minimization for geo-distributed clouds: We have proposed an e-health monitoring system with minimum service delay and privacy preservation by exploiting geo-distributed clouds. In the system, the resource management scheme enables the distributed cloud servers to cooperatively assign the servers to the requested users under the load balance condition. Thus, the service delay for users is minimized. In addition, a traffic shaping algorithm is proposed. The traffic shaping algorithm converts the user health data traffic to the non-health data traffic such that the capability of traffic analysis attacks is largely reduced. Through the numerical analysis, we show the efficiency of the proposed traffic shaping algorithm in terms of service delay and privacy preservation. Furthermore, through the simulations, we demonstrate that the proposed resource management scheme significantly reduces the service delay compared to schemes using joint the short queue stratagies.

7.2 Future Work

Resource management in e-health system is a broad and expanding research area. With the advancement of wearable sensors and understanding of how to utilize continuous monitoring results, more applications with new requirements will emerge. Thus, there are still open issues to be investigated:

- Transmission power allocation: The statistics of body area channel have been shown to be posture dependent. Channel gains under postures, such as running and rowing, have strong correlations and high variations, whereas channel gains under driving have small variations. To support QoS provisioning for medical traffic and achieve energy efficiency for sensors, the impacts of posture state dependent body area channel on transmission power allocation call for investigation. For example, since state-of-art wearable sensors are equipped with activity monitoring function, the posture dependent features of body area channel could be utilized to develop low complexity and yet efficient transmission power allocation schemes for wearable sensors.

- Distributed MAC for WBANs: In this research, a centralized MAC is proposed for WBANs with medical information transmission. Due to the mobility nature of WBANs, a centralized controller is not always available. This calls for development

of a MAC scheme that is able to handle the inter-WBAN interference and is operated in a distributed manner. It is well known that the CSMA/CA MAC scheme suffers from severe unfairness and starvation problems in multi-hop wireless networks. The starvation is not only caused by spatial bias referred to as Flow-in-the-Middle, but also by a generic coordination problem of CSMA-based scheme referred to as Information Asymmetry. For medical application, starvation could cause failure in healthcare services, thus a distributed MAC protocol that could reduce starvation and improve network throughput is desired.

- Geo-distributed clouds:In this research, the resource management scheme for geo-distributed clouds is designed for homogeneous traffic. With the advancement of sensors, various applications and monitoring data are expected to emerge in near future. These applications could generate traffic in various rates and require heterogeneous QoS provisioning in terms of delay, jitter and privacy. How to provide differentiated services to various applications in cloud networks is an importance topic, which calls for further study.

References

[1] "E-health," 2015 (Last accessed July-2015). [Online]. Available: http://www.who.int/trade/glossary/story021/en/

[2] "Healthcare," 2015 (Last accessed June-2015). [Online]. Available: https://www.ahdictionary.com/word/search.html?q=health+care&submit.x=0&submit.y=0

[3] "Ontario wait times," 2015 (Last accessed August-2015). [Online]. Available: http://www.ontariowaittimes.com/er/en/Data.aspx?LHIN=0&city=toronto&pc=&dist=0&hosptID=0&str=&view=0&period=0&expand=0

[4] J. Simpson, "The real problem with canadian health care," 2012 (Last accessed July-2013). [Online]. Available: http://news.nationalpost.com/full-comment/jeffrey-simpson-the-real-problem-with-canadian-health-care

[5] X. Liang, M. Barua, L. Chen, R. Lu, X. Shen, X. Li, and H. Luo, "Enabling pervasive healthcare through continuous remote health monitoring," *IEEE Wireless Communications*, vol. 19, no. 6, pp. 10–18, 2012.

[6] "Ieee standard for local and metropolitan area networks - part 15.6: Wireless body area networks," *IEEE Std 802.15.6*, pp. 1–271, 2012.

[7] S. Fahn, "Psychogenic movement disorders." Wiley-Blackwell, 2012.

[8] D. A. Nowak and G. R. Fink, "Psychogenic movement disorders: aetiology, phenomenology, neuroanatomical correlates and therapeutic approaches," *NeuroImage*, vol. 47, no. 3, pp. 1015–1025, 2009.

[9] P. Mell and T. Grance, *The NIST definition of cloud computing.* National Institute of Standards and Technology Computer Security Division, 2011.

[10] P. C. Webster, "Canada's ehealth software "tower of babel"," *Canadian Medical Association Journal*, vol. 182, no. 18, pp. 1945–1946, 2010.

[11] R. Garrie and P. Paustian, *Mhealth regulation, legislation, and cybersecurity.* Springer, 2014.

[12] J. Welch, J. Moon, and S. McCombie, "Early detection of the deteriorating patient: the case for a multi-parameter patient-worn monitor," *Biomedical Instrumentation & Technology*, vol. 46, no. s2, pp. 57–64, 2012.

[13] H. T. Cheng and W. Zhuang, "Bluetooth-enabled in-home patient monitoring system: early detection of alzheimer's disease," *IEEE Wireless Communications*, vol. 17, no. 1, pp. 74–79, 2010.

[14] Z. Syed, B. Scirica, C. Stultz, and J. Guttag, "Electrocardiographic prediction of arrhythmias," in *Computers in Cardiology*, 2009, pp. 565–567.

[15] J. Sriram, M. Shin, T. Choudhury, and D. Kotz, "Activity-aware ecg-based patient authentication for remote health monitoring," in *ICMI*, 2009, pp. 297–304.

[16] X. Liang, X. Li, Q. Shen, R. Lu, X. Lin, X. Shen, and W. Zhuang, "Exploiting prediction to enable secure and reliable routing in wireless body area networks," in *IEEE Proc. INFOCOM*, 2012, pp. 388–396.

[17] X. Huang, H. Shan, and X. Shen, "On energy efficiency of cooperative communications in wireless body area network," in *IEEE Proc. WCNC*, 2011, pp. 1097 –1101.

[18] S. Xiao, A. Dhamdhere, V. Sivaraman, and A. Burdett, "Transmission power control in body area sensor networks for healthcare monitoring," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 1, pp. 37–48, 2009.

[19] H. Lee, K.-J. Park, Y.-B. Ko, and C.-H. Choi, "Wireless lan with medical-grade qos for e-healthcare," *IEEE Journal of Communications and Networks*, vol. 13, no. 2, pp. 149–159, 2011.

[20] T. Aoyagi, J.-i. Takada, K. Takizawa, N. Katayama, T. Kobayashi, K. Y. Yazdandoost, H.-b. Li, and R. Kohno, "Channel model for wearable and implantable wbans," *IEEE 802.15-08-0416-04-0006*, 2008.

[21] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *IEEE Transactions on Information Theory*, vol. 48, no. 5, pp. 1135–1149, 2002.

[22] J. Lee and N. Jindal, "Energy-efficient scheduling of delay constrained traffic over fading channels," *IEEE Transactions on Wireless Communications*, vol. 8, no. 4, pp. 1866–1875, 2009.

[23] ——, "Delay constrained scheduling over fading channels: Optimal policies for monomial energy-cost functions," in *IEEE Proc. ICC*, 2009, pp. 1–5.

[24] A. J. Goldsmith and P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1986–1992, 1997.

[25] S.-h. Kuo and J. K. Cavers, "Energy optimal scheduler for diversity fading channels with maximum delay constraints," *IEEE Transactions on Wireless Communications*, vol. 8, no. 11, pp. 5520–5529, 2009.

[26] M. Halgamuge, M. Zukerman, K. Ramamohanarao, and H. L. Vu, "An estimation of sensor energy consumption," *Progress In Electromagnetics Research B*, vol. 12, pp. 259–295, 2009.

[27] W. Ye, J. Heidemann, and D. Estrin, "An energy-efficient mac protocol for wireless sensor networks," in *IEEE Proc. INFOCOM*, 2002, pp. 1567–1576.

[28] A. Soomro and D. Cavalcanti, "Opportunities and challenges in using wpan and wlan technologies in medical environments," *IEEE Communications Magazine*, vol. 45, no. 2, pp. 114–122, 2007.

[29] S. Ullah, B. Shen, S. Riazul Islam, P. Khan, S. Saleem, and K. Sup Kwak, "A study of mac protocols for wbans," *Sensors*, vol. 10, no. 1, pp. 128–145, 2009.

[30] S. Ullah and K. Sup Kwak, "Performance study of low-power mac protocols for wireless body area networks," in *IEEE Proc. PIMRC*. IEEE, 2010, pp. 112–116.

[31] S. Ullah, H. Higgins, B. Braem, B. Latre, C. Blondia, I. Moerman, S. Saleem, Z. Rahman, and K. S. Kwak, "A comprehensive survey of wireless body area networks on phy, mac, and network layers solutions," *Journal of Medical Systems*, pp. 1–30, 2010.

[32] "Wireless lan medium access control (mac) and physical layer (phy) specifications: Amendment 8: Medium access control (mac) quality of service enhancements," *IEEE Std 802.11e*, 2005.

[33] I. Demirkol, C. Ersoy, and F. Alagoz, "Mac protocols for wireless sensor networks: a survey," *IEEE Communications Magazine*, vol. 44, no. 4, pp. 115–121, 2006.

[34] P. Huang, L. Xiao, S. Soltani, M. W. Mutka, and N. Xi, "The evolution of mac protocols in wireless sensor networks: a survey," *IEEE Communications Surveys Tutorials*, no. 99, pp. 1–20, 2012.

[35] H. Liang and W. Zhuang, "Double-loop receiver-initiated mac for cooperative data dissemination via roadside wlans," *IEEE Transactions on Communications*, vol. 60, no. 9, pp. 2644–2656, 2012.

[36] H. T. Cheng and W. Zhuang, "Novel packet-level resource allocation with effective qos provisioning for wireless mesh networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 2, pp. 694–700, 2009.

[37] A. Abbas and S. Khan, "A review on the state-of-the-art privacy-preserving approaches in the e-health clouds," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1431–1441, 2014.

[38] SIEMENS, "The siemens healthcare private cloud," 2013 (Last accessed June-2015). [Online]. Available: http://www.downloads.siemens.com/download-center/Download.aspx?pos=download&fct=getasset&id1=A6V10595685

[39] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 4, 2008, pp. 63–74.

[40] I. Menache, A. Ozdaglar, and N. Shimkin, "Socially optimal pricing of cloud computing resources," in *EAI Proc. VALUETOOLS*, 2011, pp. 322–331.

[41] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[42] R. Kramme and H.-P. Uhlig, "Technical safety of electrical medical technology equipment and systems," in *Springer Handbook of Medical Technology*. Springer, 2011, pp. 35–47.

[43] A. Khajeh-Hosseini, I. Sommerville, and I. Sriram, "Research challenges for enterprise cloud computing," *arXiv preprint arXiv:1001.3257*, 2010.

[44] A. Qureshi, "Power-demand routing in massive geo-distributed systems," Ph.D. dissertation, Massachusetts Institute of Technology, 2010.

[45] L. Buttyán and T. Holczer, "Traffic analysis attacks and countermeasures in wireless body area sensor networks," in *IEEE Proc. WOWMOM*, 2012, pp. 1–6.

[46] R. V. den Bossche, K. Vanmechelen, and J. Broeckhove, "Cost-optimal scheduling in hybrid iaas clouds for deadline constrained workloads," in *IEEE Proc. CLOUD*, 2010, pp. 228–235.

[47] M. Alicherry and T. V. Lakshman, "Network aware resource allocation in distributed clouds," in *IEEE Proc. INFOCOM*, 2012, pp. 963–971.

[48] Y. Wu, C. Wu, B. Li, L. Zhang, Z. Li, and F. C. M. Lau, "Scaling social media applications into geo-distributed clouds," in *IEEE Proc. INFOCOM*, 2012, pp. 684–692.

[49] Q. Shen and W. Zhuang, "Energy efficient scheduling for delay constrained communication in wireless body area networks," in *IEEE Proc. GLOBECOM*, 2012, pp. 262–267.

[50] Q. Shen, X. Liang, X. Shen, and X. Lin, "Recce: A reliable and efficient cloud cooperation scheme in e-healthcare," in *IEEE Proc. GLOBECOM*, 2013, pp. 2736–2741.

[51] Q. Shen, X. Liang, X. Shen, X. Lin, and H. Y. Luo, "Exploiting geo-distributed clouds for a e-health monitoring system with minimum service delay and privacy preservation," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 430–439, 2014.

[52] A. Ahmad, A. Riedl, W. J. Naramore, N.-Y. Chou, and M. S. Alley, "Scenario-based traffic modeling for data emanating from medical instruments in clinical environment," in *WRI Proc. CSIE*, 2009, pp. 529–533.

[53] G. Lu, N. Sadagopan, B. Krishnamachari, and A. Goel, "Delay efficient sleep scheduling in wireless sensor networks," in *IEEE Proc. INFOCOM*, vol. 4, 2005, pp. 2470–2481.

[54] P. Nuggehalli, V. Srinivasan, and R. Rao, "Energy efficient transmission scheduling for delay constrained wireless networks," *IEEE Transactions on Wireless Communications*, vol. 5, no. 3, pp. 531–539, 2006.

[55] M. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[56] B. J. Choi and X. Shen, "Adaptive asynchronous sleep scheduling protocols for delay tolerant networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 9, pp. 1283–1296, 2011.

[57] B. Prabhakar, E. Uysal Biyikoglu, and A. El Gamal, "Energy-efficient transmission over a wireless link via lazy packet scheduling," in *IEEE Proc. INFOCOM*, 2001, pp. 386–394.

[58] D. Smith, L. Hanlen, J. Zhang, D. Miniutti, D. Rodda, and B. Gilbert, "Characterization of the dynamic narrowband on-body to off-body area channel," in *IEEE Proc. ICC*, 2009, pp. 1–6.

[59] M. Neely, "Opportunistic scheduling with worst case delay guarantees in single and multi-hop networks," in *IEEE Proc. INFOCOM*, 2011, pp. 1728–1736.

[60] C. ping Li and M. Neely, "Energy-optimal scheduling with dynamic channel acquisition in wireless downlinks," *IEEE Transactions on Mobile Computing*, vol. 9, no. 4, pp. 527–539, 2010.

[61] D. Miniutti, L. Hanlen, D. Smith, A. Zhang, D. Lewis, D. Rodda, and B. Gilbert, "Narrowband channel characterization for body area network," *IEEE Protocol*, July 2008.

[62] Z. Zhang, Z. Pi, and B. Liu, "Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 522–531, 2015.

[63] R. Yousefi, M. Nourani, S. Ostadabbas, and I. Panahi, "A motion-tolerant adaptive algorithm for wearable photoplethysmographic biosensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 670–681, 2014.

[64] V. Chaganti, L. Hanlen, and T. Lamahewa, "Semi-markov modeling for body area networks," in *IEEE Proc. ICC*, 2011, pp. 1–5.

[65] M. Zafer and E. Modiano, "Delay-constrained energy efficient data transmission over a wireless fading channel," in *IEEE Proc. ITA*, 2007, pp. 289–298.

[66] M. Neely, S. Tehrani, and A. Dimakis, "Efficient algorithms for renewable energy allocation to delay tolerant consumers," in *IEEE Proc. SmartGridComm*, 2010, pp. 549–554.

[67] G. G. Messier and I. G. Finvers, "Traffic models for medical wireless sensor networks," *IEEE Communications Letters*, vol. 11, no. 1, pp. 13–15, 2007.

[68] E. Reusens, W. Joseph, G. Vermeeren, D. Kurup, and L. Martens, "Real human body measurements, model, and simulations of 2.45 ghz wireless body area network communication channel," in *IEEE Proc. BSN*, 2008, pp. 149–152.

[69] P. Salvador, A. Nogueira, and R. Valadas, "Markovian models for medical signals on wireless sensor networks," in *IEEE Proc. ICC*, 2009, pp. 1–5.

[70] L. Li and A. J. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels. i. ergodic capacity," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 1083–1102, 2001.

[71] F. Kerber, P. Lessel, and A. Krüger, "Same-side hand interactions with arm-placed devices using EMG," in *ACM Proc. CHI*, 2015, pp. 1367–1372.

[72] M. C. Gursoy, D. Qiao, and S. Velipasalar, "Analysis of energy efficiency in fading channels under qos constraints," *IEEE Transactions on Wireless Communications*, vol. 8, no. 8, pp. 4252–4263, 2009.

[73] D. P. Palomar and J. R. Fonollosa, "Practical algorithms for a family of waterfilling solutions," *IEEE Transactions on Signal Processing*, vol. 53, no. 2-1, pp. 686–695, 2005.

[74] W. Yu and J. Cioffi, "Constant-power waterfilling: performance bound and low-complexity implementation," *IEEE Transactions on Communications*, vol. 54, no. 1, pp. 23–28, 2006.

[75] P. He, L. Zhao, S. Zhou, and Z. Niu, "Water-filling: a geometric approach and its application to solve generalized radio resource allocation problems," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3637–3647, 2013.

[76] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Transactions on Wireless Communications*, vol. 6, no. 8, pp. 3058–3068, 2007.

[77] K. Tutuncuoglu and A. Yener, "Optimum transmission policies for battery limited energy harvesting nodes," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 1180–1189, 2012.

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

[79] A. Abdrabou and W. Zhuang, "Stochastic delay guarantees and statistical call admission control for ieee 802.11 single-hop ad hoc networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 10, pp. 3972–3981, 2008.

[80] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630–643, 2003.

[81] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass markov fluids and other atm sources," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 424–428, 1993.

[82] L. Liu and J.-F. Chamberland, "On the effective capacities of multiple-antenna gaussian channels," in *IEEE Proc. ISIT*, 2008, pp. 2583–2587.

[83] D. Wu and R. Negi, "Effective capacity-based quality of service measures for wireless networks," *Mobile Networks and Applications*, vol. 11, no. 1, pp. 91–99, 2006.

[84] J. Crozier, J. Reid, G. Welch, K. Muir, and W. Stuart, "Early carotid endarterectomy following thrombolysis in the hyperacute treatment of stroke," *British Journal of Surgery*, vol. 98, no. 2, pp. 235–238, 2011.

[85] M. Chen, S. Gonzalez, A. Vasilakos, H. Cao, and V. Leung, "Body area networks: a survey," *Mobile Networks and Applications*, vol. 16, no. 2, pp. 171–193, 2011.

[86] H. Su and X. Zhang, "Battery-dynamics driven tdma mac protocols for wireless body-area monitoring networks in healthcare applications," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 4, pp. 424–434, 2009.

[87] B. Otal, L. Alonso, and C. Verikoukis, "Highly reliable energy-saving mac for wireless body sensor networks in healthcare systems," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 4, pp. 553–565, 2009.

[88] K. S. Kwak, S. Ullah, and N. Ullah, "An overview of ieee 802.15.6 standard," in *IEEE Proc. ISABEL*, 2010, pp. 1–6.

[89] S. Rashwand, J. Misic, and H. Khazaei, "Performance analysis of ieee 802.15.6 under saturation condition and error-prone channel," in *IEEE Proc. WCNC*, 2011, pp. 1167 –1172.

[90] P. Wang, H. Jiang, and W. Zhuang, "A new MAC scheme supporting voice/data traffic in wireless ad hoc networks," *IEEE Transaction on Mobile Computing*, vol. 7, no. 12, pp. 1491–1503, 2008.

[91] H. Jiang, P. Wang, and W. Zhuang, "A distributed channel access scheme with guaranteed priority and enhanced fairness," *IEEE Transactions on Wireless Communications*, vol. 6, no. 6, pp. 2114–2125, 2007.

[92] P. Wang and W. Zhuang, "A token-based scheduling scheme for wlans supporting voice/data traffic and its performance analysis," *IEEE Transactions on Wireless Communications*, vol. 7, no. 5-1, pp. 1708–1718, 2008.

[93] L. Wang, C. Goursaud, N. Nikaein, L. Cottatellucci, and J. Gorce, "Cooperative scheduling for coexisting body area networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 123–133, 2013.

[94] S. H. Cheng and C. Y. Huang, "Coloring-based inter-wban scheduling for mobile wireless body area networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 2, pp. 250–259, 2013.

[95] R. Smallwood and E. Sondik, "The optimal control of partially observable markov processes over a finite horizon," *Operations Research*, vol. 21, no. 5, pp. 1071–1088, 1973.

[96] E. Sondik, "The optimal control of partially observable markov processes over the infinite horizon: Discounted costs," *Operations Research*, vol. 26, no. 2, pp. 282–304, 1978.

[97] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5547–5567, 2010.

[98] H. Luo, "Wearable mini-size intelligent healthcare system," Patent WO2 008 098 346 A1, 2008.

[99] K. Takizawa, T. Aoyagi, J.-i. Takada, N. Katayama, K. Yekeh, K. Yazdandoost, and T. Kobayashi, "Channel models for wireless body area networks," in *IEEE Proc. EMBS*, 2008, pp. 1549–1552.

[100] E. Reusens, W. Joseph, B. Latré, B. Braem, G. Vermeeren, E. Tanghe, L. Martens, I. Moerman, and C. Blondia, "Characterization of on-body communication channel

and energy efficient topology design for wireless body area networks," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 6, pp. 933–945, 2009.

[101] S. Rashwand and J. Misic, "Channel and error modeling for wireless body area networks," *MONET*, vol. 19, no. 3, pp. 276–286, 2014.

[102] P. Sadeghi, R. Kennedy, P. Rapajic, and R. Shams, "Finite-state markov modeling of fading channels - a survey of principles and applications," *IEEE Signal Processing Magazine*, vol. 25, no. 5, pp. 57–80, 2008.

[103] X. Shen, H. Yu, J. Buford, and M. Akon, *Handbook of peer-to-peer networking*. Springer, 2010, vol. 1.

[104] Z. Qing, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5431–5440, 2008.

[105] Q. Shen, X. Shen, T. Luan, and J. Liu, "Mac layer resource allocation for wireless body area networks," *ZTE Communications*, vol. 3, pp. 13–21, 2014.

[106] S. Ahuja, S. Mani, and J. Zambrano, "A survey of the state of cloud computing in healthcare," *Network and Communication Technologies*, vol. 1, no. 2, pp. 12–19, 2012.

[107] M. Lin, A. Wierman, L. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," in *IEEE Proc. INFOCOM*, 2011, pp. 1098–1106.

[108] H. Khazaei, J. Misic, and V. B. Misic, "Performance analysis of cloud computing centers using m/g/m/m+ r queuing systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 5, pp. 936–943, 2012.

[109] K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan, "Sedic: privacy-aware data intensive computing on hybrid clouds," in *ACM Proc. CCS*, 2011, pp. 515–526.

[110] H. Schmidli, "Optimal proportional reinsurance policies in a dynamic setting," *Scandinavian Actuarial Journal*, no. 1, pp. 55–68, 2001.

[111] C. Hipp and V. Michael, "Optimal dynamic xl reinsurance," *Astin Bulletin*, vol. 33, no. 2, pp. 193–208, 2003.

[112] H. Schmidli, *Stochastic control in insurance*. Springer, 2008.

[113] Ø. B. R. G. Odd Aalen, Per Kragh Andersen and N. Keiding, "History of applications of martingales in survival analysis," *Electronic Journal for History of Probability and Statistics*, vol. 5, no. 1, pp. 1–28, 2009.

[114] D. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific Belmont, 1995, vol. 1.

[115] A. Assaad and D. Fayek, "General hospitals network models for the support of e-health applications," in *IEEE Proc. NOMS*, 2006, pp. 1–4.

[116] P. Branch and J. But, "Rapid and generalized identification of packetized voice traffic flows," in *IEEE Proc. LCN*, 2012, pp. 85–92.

[117] H. Tawfik, O. Anya, and A. Nagar, "Understanding clinical work practices for cross-boundary decision support in e-health," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 4, pp. 530–541, 2012.

[118] M. Masud, S. Hossain, and A. Alamri, "Data interoperability and multimedia content management in e-health systems," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1015–1023, 2012.

[119] L. Constantinescu, J. Kim, and D. D. Feng, "Sparkmed: a framework for dynamic integration of multimedia medical data into distributed m-health systems," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 1, pp. 40–52, 2012.

[120] M. Barua, X. Liang, R. Lu, and X. Shen, "Espac: Enabling security and patient-centric access control for ehealth in cloud computing," *International Journal of Security and Networks*, vol. 6, no. 2/3, pp. 67–76, 2011.

[121] A. Leivadeas, C. Papagianni, and S. Papavassiliou, "Efficient resource mapping framework over networked clouds via iterated local search-based request partitioning," *IEEE Transaction on parallel distributed System*, vol. 24, no. 6, pp. 1077–1086, 2013.

[122] S. Manfredi, F. Oliviero, and S. Romano, "A distributed control law for load balancing in content delivery networks," *IEEE/ACM Transactions on Networking*, vol. 21, no. 1, pp. 55–68, 2013.

[123] R. Lu, X. Lin, and X. Shen, "Spoc: a secure and privacy-preserving opportunistic computing framework for mobile-healthcare emergency," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 3, pp. 614–624, 2013.

[124] X. Lin, R. Lu, X. Shen, Y. Nemoto, and N. Kato, "Sage: a strong privacy-preserving scheme against global eavesdropping for ehealth systems," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 4, pp. 365–378, 2009.

[125] C. Wright, S. Coull, and F. Monrose, "Traffic morphing: An efficient defense against statistical traffic analysis," in *ISOC Proc. NDSS*, 2009.

[126] K. Dyer, S. Coull, T. Ristenpart, and T. Shrimpton, "Peek-a-boo, i still see you: Why efficient traffic analysis countermeasures fail," in *IEEE Symposium on SP*, 2012, pp. 332–346.

[127] X.-B. Li and S. Sarkar, "Against classification attacks: A decision tree pruning approach to privacy protection in data mining," *Operations Research*, vol. 57, no. 6, pp. 1496–1509, 2009.

[128] W. Song, W. Zhuang, and Y. Cheng, "Load balancing for cellular/wlan integrated networks," *IEEE Network*, vol. 21, no. 1, pp. 27–33, 2007.

[129] M. Liu and B. Liu, "A note on sum of powers of the laplacian eigenvalues of graphs," *Elsevier Applied Mathematics Letters*, vol. 24, no. 3, pp. 249–252, 2011.

[130] B. Jansson, "Choosing a good appointment system? a study of queues of the type (d, m, 1)," *INFORMS Operations Research*, vol. 14, no. 2, pp. 292–312, 1966.

[131] C. Pack, "The output of a d/m/1 queue," *SIAM Journal on Applied Mathematics*, vol. 32, no. 3, pp. 571–587, 1977.

[132] S. T. Maguluri, R. Srikant, and L. Ying, "Stochastic models of load balancing and scheduling in cloud computing clusters," in *IEEE Proc. INFOCOM*, 2012, pp. 702–710.

[133] D. Michie, D. Spiegelhalter, and C. Taylor, *Machine learning, neural and statistical classification*. Ellis Horwood, 1994.

[134] D. Barber, *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[135] B. Zhen, M. Kim, T. J.-i., and K. Ryuji, "Characterization and modeling of dynamic on-body propagation at 4.5 ghz," *IEEE Antennas and Wireless Propagation Letters*, vol. 8, pp. 1263–1267, 2009.

[136] M. Zafer and E. Modiano, "Optimal rate control for delay-constrained data transmission over a wireless channel," *IEEE Transactions on Information Theory*, vol. 54, no. 9, pp. 4020–4039, 2008.

[137] Z. Li, B. Wang, C. Yang, Q. Xie, and C.-Y. Su, "Boosting-based emg patterns classification scheme for robustness enhancement," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 3, pp. 545–552, 2013.