Performance Analysis of Resource Sharing in Wireless Networks: Analytical and Empirical Perspectives

BY

TAN, Wee Lum

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

IN INFORMATION ENGINEERING

JANUARY 2009

Abstract

In wireless networks, the efficient sharing of scarce wireless spectral resources is important in order to provide guaranteed Quality-of-Service (QoS) to the wireless users. The effectiveness of resource sharing schemes in wireless networks are often heavily influenced by different aspects of the system behavior, such as user mobility, traffic dynamics and practical realization constraints. In this thesis, using analytical modeling and empirical measurement techniques, we investigate the impact of these system behaviors on the performance of resource sharing in wireless networks. In particular, we investigate the dynamic sharing of an access point's bandwidth resources among moving vehicles in a vehicular network, the adaptive sharing of the medium access resources among nodes with different and varying traffic loads in a wireless sensor network, and the practical implementation of network resources sharing among users and applications with different QoS requirements in 3G wireless networks.

In the first part of this thesis, we focus on Drive-thru Internet systems where access points (AP) are placed on roadsides and vehicles passing through the coverage range of the APs can download data from them. The amount of data downloaded by an individual user is affected not only by the scheduling algorithms, but also by the user dynamics, i.e. the movement of the vehicles which impacts the amount of time the vehicle spends in the AP's coverage range, as well as the number of contending vehicles for the AP's resources. We have developed practical analytical models with tractable solutions to characterize the data communication performance of a vehicle in a Drive-thru Internet system. A distinctive aspect of our models is that they combined both vehicular traffic theory and wireless network/protocol properties to investigate the effects of various system parameters on a drive-thru vehicle's data communication performance.

In the second part of this thesis, we examine resource sharing in wireless sensor networks in terms of the node access to the wireless medium. We propose an energy-efficient TDMA-based MAC protocol that significantly reduces energy consumption in the network, while efficiently handling network traffic load variations and optimizing channel utilization through a timeslots stealing mechanism and timeslots reassignment procedure. We have analytically derived the average delay performance of our MAC protocol, with and without the timeslots stealing feature. Our delay model, validated via simulations, shows that the timeslots stealing feature can substantially improve the protocol throughput in situations with varying and asymmetric traffic patterns. Simulation results show that the timeslots reassignment procedure is efficient in handling the longer timescale changes in the traffic load, while the timeslots stealing mechanism is better in handling the shorter timescale changes in the traffic patterns.

The third part of this thesis focuses on our empirical investigations into the performance of practical implementation of resource sharing schemes in 3G wireless networks. We have investigated the performance of multiple commercial 3G networks in Hong Kong, in terms of their ability to provide service guarantees to different traffic classes as well as the fairness of the radio-link scheduler in allocating the bandwidth resources to multiple data calls in a saturated network. We have also investigated the data throughput, latency, video and voice calls handling capacities of the 3G networks under saturated network conditions. Our findings point to the diverse nature of the network resources allocation mechanisms and the call admission control policies adopted by different operators. Our results also show that the 3G network operators seem to have extensively customized their network configurations in a cell-by-cell manner according to the individual site's local demographics, projected traffic demand and the target coverage area of the cell. As such, the cell capacity varies widely not only across different operators but also across different measurement sites of the same operator.

摘要

在無線網絡中，各方面的因素，如用戶移動性，交通動態和系統設計上的限制，往往會大大影響系統資源共享的模式及效率。在此論文中，我們會通過分析模型和實證測量技術，研究上述各種因素對無線網絡資源共享的影響。

我們首先研究汽車動態如何影響一個 Drive-thru Internet 系統的效能。利用分析模型，我們把汽車流量理論和無線網絡資源共享及其數據通訊性能一併融合起來分析。

我們也研究了無線傳感器網絡，MAC 通訊協議之設計和系統分析。我們提出了高能源效率的 TDMA 制式 MAC 通訊協定。透過"借用"基制和時段調動程序，有效地處理流量負荷變化的影響。在平均延誤的分析中，透過"借用"基制能在某些情況提高網絡容量達。我們的結果表明時段調動程序可有效處理較長時間上的通訊流量變化，而"借用"機制可[illegible]時間上的流量起伏。

最後，我們藉實證研究，探討商業 3G 無線網絡中的資源分配與共享。我們研究在實際網絡中，各種不同用戶和應用模式如何達到其要求之通訊服務品質保証。我們的結果指出，各無線網絡營運商都有不大相同的網絡資源分配基制和接納控制的政策。

"Now to Him who is able to do exceedingly abundantly above all that we ask or think, according to the power that works in us"
(Ephesians 3:20)

Acknowledgements

My life as a graduate student in CUHK has been an amazing and enlightening experience, made possible with the constant support and encouragement from the innumerable fantastic people around me. With the completion of my doctorate, I would like to take the opportunity to thank the many people who have been the never-ending source of ideas, inspiration, friendship, laughter and love that have kept me going through this difficult yet rewarding process.

I am very privileged to have Prof. Wing Cheong Lau and Prof. OnChing Yue as my thesis advisors. Their broad perspectives, insightful ideas and suggestions have greatly assisted me in the development of my research. Working with them has been a truly rewarding and intellectually stimulating experience. I am truly indebted to them for their help and advice, and for being not only a mentor, but also a friend.

I am also very thankful for the many friends in my church, who are always there for me with their warm friendship and for always supporting me with their prayers. I would also like to thank the many friends I have had the privilege of knowing during my years at CUHK, for my making my student life more memorable and enjoyable.

Last but definitely not least, I wish to express my heartfelt gratitude to my mother and family, for their continuous encouragement and support. I am also deeply thankful to my wife and son for all of their love, care, prayers and encouragement to me at all times. With them by my side, my life is more complete and meaningful. This thesis is dedicated to them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The past decade or so has witnessed great advancements in wireless technologies. Wireless communication devices are becoming smaller, more user friendly and more ubiquitous, fulfilling to a large extent the promise of anywhere and anytime communication. They are not only carried by people, but are also integrated into physical objects. These devices are utilized in various ways, ranging from the traditional voice communications in wireless cellular networks, to data collection and dissemination in wireless sensor networks. However in order to provide this multitude of services, the wireless networks need to have efficient allocation and management of the scarce wireless spectral resources, shared among the devices. Efficient and reliable bandwidth resources allocation is crucial in order for the networks to be able to provide guaranteed Quality-of-Service (QoS) to the wireless users. In this thesis, our focus is to understand (via analytical modeling and empirical measurements) the performance of various resource sharing schemes in wireless networks.

There are several different aspects of the system behavior that can impact the performance of resource sharing schemes in wireless networks. For example, in a mobile wireless network served by an access point (AP), the amount of bandwidth resources allocated to an individual user is affected not only by the scheduling algorithms in the AP, but also by the user dynamics, i.e. the movement of users which determines the amount of time that a user spends in the AP's coverage range, as well as the number of contending users for the AP's resources. In energy-constrained wireless networks that employ time-slotted medium access protocols to schedule data

transmissions and avoid collisions, the delay and power consumption performance is affected by how the resource sharing scheme assigns the timeslots to the users in accordance to the network traffic load and rate variations. In wireless cellular networks, resource sharing mechanisms need to consider not only the usual fairness versus overall system throughput/utility trade-offs, but also the QoS requirements of different classes of users and applications. While the performance of resource sharing schemes in wireless networks can be analyzed through analytical models and simulations, another important means is through empirical measurements of the real-life and practical resource sharing mechanisms employed in wireless networks.

To illustrate the performance of resource sharing schemes in wireless networks, in this thesis, we focus our investigations on three different types of wireless networks:

1. In Drive-thru Internet systems, users in moving vehicles can connect to a road-side AP to obtain Internet connectivity as the vehicles pass through the AP's coverage range. Under most AP resource sharing schemes or protocols, the amount of data that a passing vehicle can download from (or upload to) the AP is dependent upon the period or sojourn time of the vehicle within the AP's coverage range, and the number of other concurrent vehicles competing for the AP's bandwidth resources during its sojourn. A vehicle's sojourn time is determined by its speed and the length of the AP's coverage range. The vehicle speed is, in turn, impacted by the interactions among the vehicles on the road. This leads to interesting and important interplays between vehicular traffic parameters, wireless network settings and the vehicles' communication/data download performance within a Drive-thru Internet system. In order to provide a systematical framework to evaluate the type of communication services and the quality-of-service that these systems can provide, in this thesis, we investigate the communication performance of vehicles sharing a Drive-thru Internet system.

2. In wireless sensor networks, energy management is the most important task due to the fact that most sensor networks are designed to operate unattended for

long-term in a remote or hostile environment where battery-operated sensor nodes cannot be easily replaced or have its energy supply replenished. Typically, radio communication between sensor nodes consumes the largest amount of energy, and thus, all network communication protocols designed for sensor networks must be energy-efficient in order to optimize the network lifetime. In particular, an energy-efficient MAC protocol is important since it directly controls the operation of the transceiver and hence, is crucial for the effective operation and longevity of a wireless sensor network. In this thesis, we propose and investigate the performance of a MAC protocol that significantly reduces energy consumption, efficiently handle network traffic load variations and optimize channel utilization in a wireless sensor network.

3. In 3G wireless networks, resource sharing schemes manage and allocate the bandwidth resources among the different users and applications, with the aim of efficiently utilizing the available spectral resources while at the same time satisfying the QoS requirements of the different classes of users and applications. The performance and impact of the resource sharing schemes in 3G wireless networks is particularly critical under heavily-loaded conditions. Therefore in this thesis, we investigate the empirical performance of live commercial 3G networks under saturated conditions, using a mixture of data, video and voice traffic.

1.2 Contributions of this Thesis

The main contributions of this thesis are as follows.

1. In Chapter 3,
   - We have developed practical analytical models with tractable solutions to characterize the data download/upload performance of a vehicle in a Drive-thru Internet system, under the presence of other vehicles contending for the same wireless bandwidth resource. A distinctive aspect of our models is that it combined both vehicular traffic theory and wireless

network/protocol properties to investigate the effects of various system parameters on a drive-thru vehicle's data communication performance.

- We have derived the closed-form solution for the average number of bytes downloaded by a vehicle by the end of its sojourn through an AP's coverage range. Our model is able to handle the cases of homogeneous and heterogeneous classes of vehicles with different communication requirements and scheduling weights, as well as the case of the AP providing service at a fixed rate throughout the AP's coverage range, or at different rates depending on the vehicle's distance from the AP. Our closed-form solution is also insensitive to the distribution of the vehicle's sojourn time.
- We have also developed an analytical solution for the distribution of the number of bytes downloaded by a vehicle by the end of its sojourn through an AP's coverage range. In terms of analysis technique, we mapped the study of our vehicular data downloading process into the transient analysis of a series of Markov reward processes. Our use of Markov reward model is non-standard in the sense that we only select from the corresponding Markov chain, a subset of relevant sample paths that matches the required behavior of our vehicular flow model.
- Using our models, we have investigated the effects of the interplays and trade-offs between
  (a) Vehicular traffic characteristics (e.g. vehicle speed and road traffic densities, penetration rate of wireless-equipped vehicles) and
  (b) Wireless network configuration and protocols settings (e.g. the AP's coverage range, range-dependent transmission rate adaptation and the presence of multiple classes of drive-thru vehicles with different communication requirements and scheduling weights),

  on a drive-thru vehicle's data communication performance.

To the best of our knowledge, our work in Chapter 3 is the first to offer a unique analytical framework based on which the interplay between vehicular traffic parameters and a vehicle's data download/communication performance in a Drive-thru Internet system can be studied and optimized in a systematic,

quantitative manner.

2. In Chapter 4, we have examined resource sharing in wireless sensor networks in terms of the wireless nodes' access to the transmission medium. In particular,
   - We have proposed a TDMA-based and receiver-driven MAC protocol that significantly reduces energy consumption in the network. A drawback of a static timeslots assignment MAC protocol (with the timeslots equally distributed among the nodes) is its inefficiency in handling asymmetric traffic load across different nodes, or traffic flows with rate variations over multiple timescales. To counter this, we introduced a "timeslots stealing" mechanism into our MAC protocol to efficiently handle network traffic load variations and optimize channel utilization.
   - We have analytically derived the average delay performance of our MAC protocol, with and without the timeslots stealing feature. Our delay model, validated via simulations, shows that the timeslots stealing feature can substantially improve the protocol throughput in situations with varying and asymmetric traffic patterns.
   - We have also proposed a simple timeslots reassignment procedure to allow the receiver nodes to redistribute the timeslots among the sender nodes according to their offered traffic load. Our evaluation results show that our proposed MAC protocol should utilize the timeslots reassignment procedure to handle the longer timescale changes in the network traffic load variations, while using the timeslots stealing mechanism to adapt to the shorter timescale changes in the traffic patterns. By doing so, our MAC protocol is able to conserve the nodes' energy and also reduce the average packet latency.

3. In Chapter 5, we present the findings from an extensive measurement study on multiple commercial 3G networks.
   - We have investigated the performance of those 3G networks in terms of their ability to provide service guarantees to different traffic classes, and their data throughput, latency, video and voice calls handling capacities under saturated network conditions.

- By trading off the data saturation capacity against video and voice calls, we examined the behavior of the network resources allocation mechanisms and the call admission control policies in the 3G networks. We also investigated the fairness of the radio-link scheduler in allocating the bandwidth resources to multiple data calls in a data saturated network. Our findings point to the diverse nature of the network resources allocation mechanisms and the call admission control policies employed by different operators.
- We have studied and compared the throughput and latency performance of 3G data services in fully-loaded and lightly-loaded network conditions. We have also performed data throughput and latency measurements on a lightly-loaded HSDPA network, and compared the results with those obtained over the 3G networks. In addition, by comparing the latency performance of 3G data services in lightly-loaded network conditions with the latency performance obtained from several other access networks (e.g. ADSL, cable modem, Ethernet), we quantified the effects of the 3G network processing and queueing delay on its latency performance.

To the best of our knowledge, our work in Chapter 5 is the first public report on a large scale empirical study on the performance and capacities of commercial 3G networks carrying live data, video and voice traffic.

1.3 Outline of this Thesis

This thesis is outlined as follows. We provide a short background description on the three types of wireless networks that are the focus of our investigation in this thesis in the next chapter. We then present an analysis of the data communication performance of a vehicle in a Drive-thru Internet system in Chapter 3. In Chapter 4, we propose a TDMA-based and receiver-driven MAC protocol for wireless sensor networks, and investigate its delay and power consumption performance through analytical and simulation studies, while Chapter 5 presents the findings of our large-scale empirical study on the performance and capacities of commercial 3G networks. We conclude this thesis with a brief summary and suggestions for future work in Chapter 6.

Chapter 2

Background Study

In this chapter, we provide a brief overview of the three different wireless networks that are the focus of our study in this thesis, namely vehicular networks, wireless sensor networks and 3G wireless networks.

2.1 Vehicular Networks

Wireless vehicular networks have attracted much attention in recent years. By exploiting the significant on-board sensing, computing and communication capabilities of the vehicles, a wide range of applications can be enabled, ranging from collision avoidance and emergency message dissemination to real-time traffic condition monitoring [1]. Vehicular network applications can generally be classified into safety [2, 3] and non-safety applications [4-6]. For instance, vehicular safety applications may include automatic collision notification, heavy fog detection and notification, and other assistances for safe driving. Non-safety applications include real-time traffic congestion notification, location-based driver information services, automobile high speed Internet access, and many others.

Due to the high degree of mobility involved, vehicular network topologies are typically highly dynamic in nature. As such, there has been ample research on determining the connectivity of a vehicular network [7-9]. In a vehicular network, two types of vehicular communications are defined: vehicle-to-vehicle (V2V) communications and vehicle-to-infrastructure (V2I) communications. In V2V communications, various research efforts have dealt with the issue of fast dissemination of messages among the vehicles [2, 10-13]. In terms of efficient medium sharing and reducing the medium access delay, [3, 14-16] describe some

MAC protocols designed for V2V communications. Another important aspect of V2V communications is the routing of packets from the source vehicle to the destination vehicle, which is dealt with in [17-23].

In this thesis, our focus of investigation in vehicular networks is on Drive-thru Internet systems [5], which provide intermittent V2I communications. In such systems, users in vehicles can obtain network connectivity by temporarily connecting to a road-side networked base station or access point (AP) as the vehicle travels through the AP's coverage range. The issues and challenges in Drive-thru Internet systems and other related work will be discussed in more detail in Chapter 3 when we investigate the performance of resource sharing in such systems.

2.2 Wireless Sensor Networks

Advancements in digital electronics, embedded systems, signal processing, and wireless communications have led to the development of tiny, low-cost and low-power sensor devices or nodes, equipped with multiple parameter sensing, processing and communication capabilities [24]. A wireless sensor network is formed by deploying a large number of these sensor nodes over a region of interest. These sensor nodes often cooperate to perform various monitoring and data gathering tasks, processing the gathered information, and communicating the processed data over multiple hops to the end-user located at the data sink. Many application scenarios have been envisioned for wireless sensor networks, ranging from low data rate applications like precision agriculture monitoring systems to high data rate applications like real-time battlefield surveillance and industrial monitoring systems.

Research in wireless sensor networks has tended to focus on the issue of energy efficiency. This is due to the large number of nodes and their deployment in remote, unattended, and hostile environments, where it is usually difficult, if not impossible, to recharge or replace their batteries. Low power solutions are required in order to minimize energy consumption and extend the lifetime of the networks, without

jeopardizing reliable and efficient communications in the networks. To this end, there have been a lot of research in energy-efficient MAC protocols [25-29], data gathering protocols [30-34], network topology maintenance [35-39] and transport mechanisms [40-42] for wireless sensor networks.

In this thesis, our focus of investigation in wireless sensor networks is on the performance of energy-efficient MAC protocols. In Chapter 4, we will discuss in more detail the related work on energy-efficient MAC protocols in sensor networks.

2.3 3G Wireless Networks

WCDMA-based 3G wireless networks enable network operators to offer users a wider range of more advanced services while achieving greater network capacity through improved spectral efficiency. Due to the interest in the throughput and delay performance of applications in 3G networks, there have been quite a number of field measurement studies done, but they are mainly focused on the performance of pure data traffic under lightly-loaded or controlled environments [43-47].

As mentioned before in Chapter 1, the performance and impact of the resource sharing schemes in 3G wireless networks is particularly critical under heavily-loaded conditions. Therefore in this thesis, we investigate the empirical performance of live commercial 3G networks under saturated conditions, using a mixture of data, video and voice traffic. By performing our measurements under saturated conditions, we can also empirically obtain the actual capacity of the 3G networks, compared to previous theoretical models [48-53] which are more useful for preliminary network capacity approximation and network planning purposes.

Chapter 3

Resource Sharing in Drive-thru Internet Systems

3.1 Introduction

There have been increasing commercial and research interests in utilizing wireless technologies like WiFi and WiMax to provide Internet connectivity to users in moving vehicles. Such systems, termed Drive-thru Internet [5], operate by placing inter-connected road-side access points (APs) on city roads and trunk roads so as to enable vehicular users to obtain network connectivity by temporarily connecting to an AP as the vehicle passes through the AP's coverage range. Previous measurement studies [5, 54, 55] have shown the viability of this type of network access for vehicular users, with some [56, 57] proposing protocols to improve the inter-activity and throughput of the network access performance.

An important feature of Drive-thru Internet systems, shown in Fig. 3.1, is the multi-access sharing of the AP's bandwidth among the vehicles that are simultaneously under the coverage of the AP. Under most AP resource sharing schemes or protocols, the amount of data that a passing vehicle can download from (or upload to) the AP is dependent upon two main factors:

a) The period or sojourn time of the vehicle within the AP's coverage range
b) The dynamically changing number of other concurrent vehicles competing for the AP's bandwidth resources during its sojourn

A vehicle's sojourn time is determined by its speed and the length of the AP's coverage range. The vehicle speed is, in turn, impacted by the interactions among the vehicles on the road. This leads to interesting and important interplays between vehicular traffic

parameters, wireless network settings and the vehicles' communication/data download performance within a Drive-thru Internet system.

In this chapter, we will develop practical analytical models with tractable solutions to characterize a typical vehicle's communication/data-downloading performance in Drive-thru Internet systems. In particular, we focus on deriving the solutions for the average and the distribution of the number of bytes that a vehicle can download from an AP by the end of its sojourn. A distinctive aspect of our work is that it combines both vehicular traffic theory and wireless network/protocol properties to investigate the effects of various system parameters on the vehicle's data download performance. To be more specific, we investigate the interplays and trade-offs between:

a) Vehicular traffic characteristics (e.g. vehicle speed and road traffic densities, penetration rate of wireless-equipped vehicles)
b) Wireless network configuration and protocols settings (e.g. the AP's coverage range, range-dependent transmission rate adaptation and the presence of multiple classes of drive-thru vehicles with different communication requirements and scheduling weights)

In terms of analysis technique, we map the study of our vehicular data downloading process into the transient analysis of a series of Markov reward processes. Our use of Markov reward model is non-standard in the sense that we only select from the corresponding Markov chain, a subset of relevant sample paths that matches the required behavior of our vehicular flow model.

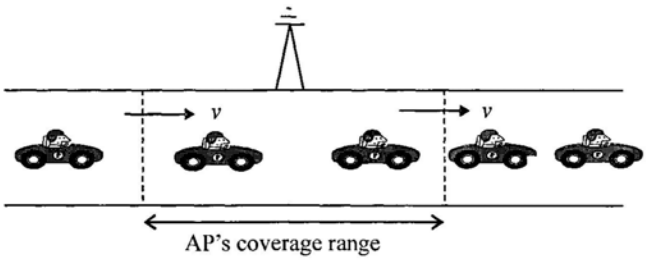


Figure 3.1 Drive-thru Internet

We believe our work offers a unique analytical framework based on which the interplay between vehicular traffic parameters and a vehicle's data download/ communication performance in a Drive-thru Internet system can be studied and optimized in a systematic, quantitative manner. Our results are useful to system architects and network planners in designing optimization algorithms that schedule and route data packets to selected APs based on vehicle routes and road traffic conditions.

The remainder of this chapter is organized as follows. In Section 3.2, we provide a discussion of related work. A description of our system model is given in Section 3.3, and we derive the expressions for the average and the distribution of the number of bytes received by a vehicle by the end of its sojourn in Section 3.4 and Section 3.5. We present our results and point out some interesting observations from them in Section 3.6, and we conclude this chapter in Section 3.7.

3.2 **Related Work**

Ott and Kutscher [5] introduced the idea of Drive-thru Internet systems and verified its feasibility through experiments. There have also been other measurement studies [54, 55, 58] and numerous proposals that try to improve the performance of Drive-thru Internet systems by modifying their MAC [57, 59], routing [60, 61], transport [56, 59] and application layer [6, 56] protocols. On the analytical front, [7-9] have concentrated on modeling the connectivity or message propagation distances in inter-vehicle communications under a highway environment. Under the context of Delay-Tolerant Networking (DTN), there have also been analytical works [62, 63] which model the mobility pattern and communication performance of vehicles traveling between multiple road-side APs or communication "throwboxes". However, these DTN-related studies all focus on the low-service-penetration regime where each road-side AP only needs to cover and interact with at most one vehicle at a time. Their emphasis was on the inter-encountering time between a vehicle and the series of APs or communication throwboxes it visited. In other words, communication resource contention among

multiple concurrent drive-thru vehicles was not considered at all. While the models in [62, 63] seem to be reasonable for analyzing the performance of low-vehicle-density DTNs like the UMass DieselNet [60, 61], they are unlikely to be applicable for commercial Drive-thru Internet systems to be deployed in an urban city like Hong Kong. In particular, with the high-density of vehicles (which are often packed with one or more bandwidth-hungry wireless users or applications), any performance model for the corresponding Drive-thru Internet systems must address the issues of (1) potential resource contention amongst multiple vehicles concurrently covered by an AP and (2) the dynamic changes in the membership of the group of vehicles being served by an AP due to vehicular movement.

An analytical framework based on fluid and stochastic vehicular traffic models was introduced in [64-66] to capture the impact of highway mobility on cellular network performance. However, these studies only focused on circuit-switched-based services, where the performance metrics of interest were mostly system-level, aggregated ones, e.g. system outage probability, call blocking rate, and base-station call-setup/hand-off loads. In contrast, our work focuses on analyzing the quality of data service, e.g. download throughput, received by individual drive-thru vehicles. In [67], researchers proposed an integrated mobility and traffic model whose parameters were learned and tracked using a Hidden-Semi-Markov model. Based on the model, call admission control and resource allocation schemes were developed. Their work is orthogonal to our work, where our objective is to establish an analytical framework with tractable solutions to quantify the interplay between vehicular traffic conditions, wireless network settings and data service performance.

Our model of a group of in-range vehicles sharing the AP's bandwidth during their sojourn is quite different to that of a processor sharing (PS) queueing system [68]. In a conventional PS system, a customer's sojourn time is dependent on its total service requirement and the number of other concurrent customers sharing the server's services. A customer will not leave the system until it has received its requested amount of service. In contrast, for our model, the sojourn (drive-thru) time of a vehicle

is deterministic but the actual amount of bandwidth resource (service) it receives (in terms of number of bytes) will be dependent on the number of other concurrent vehicles sharing the AP.

Our investigation of the amount of data that a drive-thru vehicle can download from an AP under the presence of other contending vehicles leads to a Markov reward model-based formulation in which the accumulated reward over a finite time interval represents the total amount of data downloaded by a vehicle after passing through the AP's coverage range. Markov reward models are commonly used in performability analysis (see [69-72] and references therein), where the research focus is on the evaluation of the distribution of the cumulative reward that is earned either (i) at the end of a finite time interval, or (ii) until the Markov process enters into some absorption states. Our work is different in that we are evaluating the cumulative data downloaded by the end of a vehicle's sojourn which translates into the additional requirements of: (i) the vehicle has to spend some specific amount of time in the system and (ii) at the specified departure time, the corresponding Markov process must be in some specific state(s) (determined by the number of departures of other vehicles during the tagged vehicle's drive-thru).

3.3 System Model

Vehicular traffic flow models are generally divided into two major classes [73, 74]: microscopic models that consider the behavior of each individual vehicle separately, and macroscopic models that aggregate all vehicles into a flow and describe the flow in terms of fundamental quantities including vehicle density, flow and speed. In our work, we adopt the macroscopic vehicular traffic modeling approach and combine it with a stochastic queueing model. Denote by k the vehicle density which corresponds to the number of vehicles per unit distance along the road segment. Let q be the vehicle flow which measures the number of vehicles that pass a fixed road-side observation point per unit time. Use v to represent the vehicle speed, i.e. the distance that a vehicle travels per unit time. These three variables, generally taken as average values, are

related by the fundamental relationship:

$$q = kv \tag{3.1}$$

Based on field observations [75, 76], Greenshields developed the first model that captures the speed-density relationships by assuming a linear relationship between speed v and density k,

$$v = v_f(1 - k/k_{jam}) \tag{3.2}$$

where v_f is defined as the free-flow speed corresponding to the speed used when the vehicle is all alone on the road (usually taken as the road's speed limit), and k_{jam} is the vehicle jam density at which traffic flow comes to a halt. Using Eqns. (3.1) and (3.2), the typical speed-flow-density diagram can be constructed as shown in Fig. 3.2 [77]. A speed-flow-density diagram can be used to show the effect of the increase in the vehicle density on the movements of vehicles on the road segment. Observe from the figure that the traffic flow rate is zero when there are no vehicles on the road. At this point, the first vehicle arrival can travel at the free-flow speed. At the other extreme when density becomes so high that all vehicles stop, the flow is also zero. Between these two points, we see that as density increases from zero, flow also increases due to the increasing number of vehicles on the road while speed starts to decrease due to the

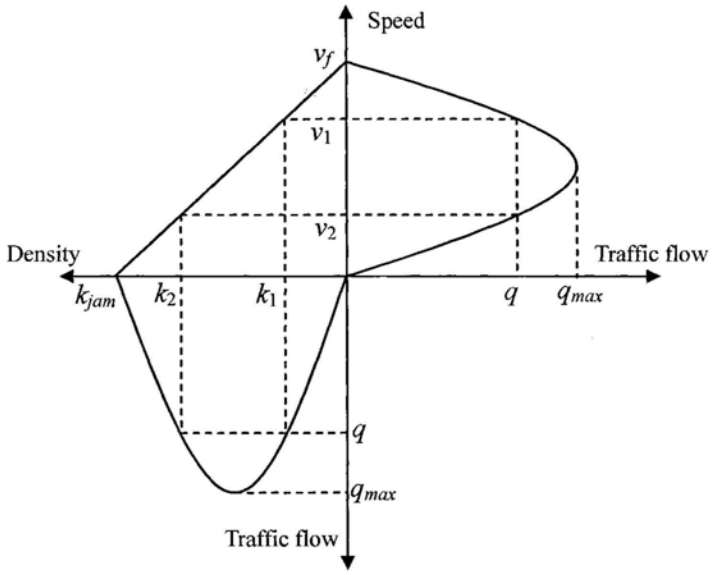


Figure 3.2 Speed-flow-density diagram

interactions among the vehicles. This phase is generally known as the free-flow phase. As density continues to increase, the traffic flow reaches capacity q_{max} when the product of density and speed results in the maximum flow. Beyond this point, if density keeps increasing, traffic becomes congested and will result in reduced traffic flow and low speeds (referred to as the congested-flow phase). Fig. 3.2 also shows that for any value of traffic flow q (other than q_{max}), there exist two different values of density k_1, k_2, and speed v_1, v_2, corresponding to the free-flow and congested-flow phases respectively.

Eqns. (1) and (2) characterize a homogeneous, equilibrium traffic flow passing through a road segment [73]. The state of the traffic flow is only dependent on the density of the vehicles. Given the vehicle density k on the road, v can be determined via Eqn. (2), and q can be obtained from Eqn. (1). Using this model of a steady-state traffic flow passing through the AP's coverage range, we can approximate the vehicle arrivals to a road-side observation point (i.e. the leading edge of an AP's coverage range) by a Poisson process[1] [77-82] with mean arrival rate λ (= q, the traffic flow rate described above). Moreover, according to the steady-state traffic flow model, each vehicle's speed v can be assumed to be the same and constant throughout its sojourn. Under these settings, we can use an M/D/C/C queueing system[2] to model the vehicular traffic flow passing through an AP's coverage range. In particular, the sojourn time of each vehicle is deterministic and is given by $T = L/v$, where L is the length of the AP's coverage range. The capacity of the system (C), i.e. the maximum number of vehicles that can be accommodated by the AP's coverage range is given by $C = k_{jam}L$. Based on this M/D/C/C queueing model, p_N, the steady-state probability that there are N vehicles simultaneously under the coverage of the AP, is given by [83]:

$$p_N = \frac{(\lambda T)^N/N!}{\sum_{i=0}^{C}(\lambda T)^i/i!} \qquad \text{for } 0 \le N \le C \tag{3.3}$$

The above model assumes uninterrupted vehicular traffic flow, i.e. there aren't any

[1] We will compare our analytical results with simulations based on other non-Poisson arrival patterns in Section 3.6.2.

[2] We will compare our analytical results with simulations based on the M/D/C queueing system in Section 3.6.4.

external means, namely traffic lights, to regulate the flow of traffic. Therefore, the traffic flow is only influenced by vehicle-vehicle interactions and the road parameters, e.g. speed limit. Furthermore, this basic model only considers unidirectional, single-lane roads, as shown in Fig. 3.1.

3.4 Average, Minimum & Maximum Number of Bytes Received by a Vehicle per Drive-Thru

For ease of illustration, assume for now that all vehicles passing through the coverage area of the AP are wireless-equipped vehicles with infinite request backlog[3]. The utilization of the AP, ρ, is defined as the probability that the AP is busy serving one or more vehicles. Since we assume that all vehicles are competing for AP's bandwidth, ρ is equivalent to the steady-state probability that there is at least one vehicle within the coverage area of the AP. The average number of bytes received by a vehicle per drive-thru (Y_{avg}) can then be computed as follows:

$$
\begin{aligned}
&Y_{avg} \\
&= B.(\text{Avg. amount of service time that a vehicle received from the AP per drive-thru}) \\
&= B.\lim_{t\to\infty}\left(\frac{\text{Total amt. of time that AP was busy serving one or more vehicles by time t}}{\text{Total number of vehicles served by time t}}\right) \\
&= B.\left(\frac{\rho}{\lambda_{eff}}\right) = \frac{B(1-p_0)}{\lambda(1-p_C)} \qquad (3.4)
\end{aligned}
$$

where B is the AP's transmission bit rate, and λ_{eff} is the effective vehicle arrival rate to the AP's coverage range, adjusted for "fictitious" vehicle loss due to blocking in the M/D/C/C queueing system. p_0 and p_C are the system idle and blocking probabilities respectively, as given in Eqn. (3.3).

Note that Eqn. (3.4) still holds even if we were to relax our assumption of deterministic vehicle sojourn times, to some other general distributions. This is due to the powerful

[3] By assuming an ideal scheduler and no MAC or signaling overhead, the amount of data received can be viewed as an upper-bound value.

insensitivity property of the M/G/C/C queueing systems [83], which states that the steady-state system size probability (Eqn. (3.3)) is only a function of the mean sojourn time, and is independent of the sojourn time distribution. Therefore, even in the case where the vehicle speeds are changing during the sojourn (for example speeding up or slowing down), as long as we know the mean sojourn time of the vehicles, we can use Eqns. (3.3) and (3.4) to compute Y_{avg}.

We can generalize Eqn. (3.4) to handle the case of multiple classes of wireless-equipped vehicles receiving different weights of service (including zero) from the AP. Let there be m different classes of vehicles, with α_i and β_i denoting respectively, the fraction of class-i vehicles and the service-weight of the class-i vehicles. If our tagged vehicle is a class-j vehicle, then the average number of bytes it receives per drive-thru is determined by the average amount of service time that it receives from the AP during its drive-thru. Using the same argument as earlier, for a class-j vehicle,

$$Y_{avg} = B.\left[\frac{\rho(j)}{\lambda_{eff}(j)}\right] = B.\left[\frac{\rho(j)}{\alpha_j \lambda(1-p_C)}\right] \tag{3.5}$$

with

$$\rho(j) = \sum_{N=1}^{C} \sum_{\substack{n_1\geq 0, n_2\geq 0, \ldots, n_j\geq 1, \ldots, n_m\geq 0; \\ n_1+n_2+\ldots+n_j+\ldots+n_m=N}} p_N \left(\prod_{i=1}^{m} \alpha_i^{n_i}\right) \cdot \left(\frac{N!}{\prod_{i=1}^{m} n_i!}\right)\left(\frac{n_j\beta_j}{\sum_{i=1}^{m} n_i\beta_i}\right) \tag{3.6}$$

where the product of the first and second term under the sign of the 2nd summation in Eqn. (3.6) represents the probability of the combination of N vehicles from among the m classes of vehicles with at least one vehicle from class-j, the third term is the multinomial coefficient representing the number of permutations of the N vehicles into the m classes, and the final term represents the fraction of the AP's resources allocated to the class-j vehicles.

To determine the range of values for the number of bytes received by our tagged vehicle by the end of its sojourn, we also compute the minimum and the maximum values of this range. The minimum number of bytes that our tagged vehicle can receive

by the end of its sojourn corresponds to the scenario where throughout its whole sojourn, our tagged vehicle shares the AP's resources with (C-1) other vehicles having service-weight $\beta_{max} = \max\{\beta_i, i \in [1,2,\ldots,m]\}$. We then have

$$Y_{min} = \left(\frac{\beta_j}{\beta_j + (C\text{-}1)\beta_{max}}\right) BT \tag{3.7}$$

On the other hand, the maximum number of bytes that our tagged vehicle can receive by the end of its sojourn corresponds to the scenario where throughout its whole sojourn, our tagged vehicle has the AP's resources all to itself, i.e. it does not have to share the AP's resources with any other vehicles. Therefore,

$$Y_{max} = BT \tag{3.8}$$

Our model in Eqn. (3.4) can also be used to handle the case where the AP's transmission bit rate received by our tagged vehicle is dependent on its distance from the AP. As an example, consider Fig. 3.3 where the AP's transmission bit rate received by the vehicle changes from B_1 to B_2, and to B_3, as the vehicle passes through the AP's coverage range. Since the vehicle's speed is assumed to be a constant v throughout its total sojourn time T, therefore the vehicle's sojourn time for each of the different ranges is given by $T_i = L_i/v$, where L_i is the distance of the range in which it receives service at a bit rate of B_i. From Eqn. (3.4), out of the total average service time that a vehicle received from the AP per drive-thru, a fraction T_i/T of it is spent receiving service at bit rate B_i. We can then compute the average number of bytes received by a drive-thru vehicle through the following relationship:

$$Y_{avg} = \frac{\rho}{\lambda_{eff}} \sum_{i=1}^{K} B_i \frac{T_i}{T} \tag{3.9}$$

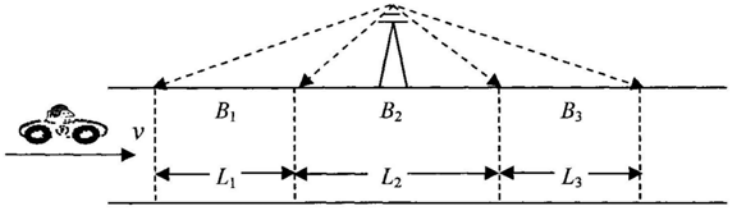


Figure 3.3 Range-dependent transmission rates

where K is the total number of ranges passed through by our tagged vehicle in which it received service at different transmission bit rates from the AP.

3.5 Distribution of the Number of Bytes Received by a Vehicle per Drive-Thru

In order to obtain a complete probabilistic characterization of the amount of data that our tagged vehicle can download by the end of a sojourn through the AP's coverage range, in this section we focus on deriving its distribution. During our tagged vehicle's sojourn, it shares the AP's bandwidth resources with other vehicles that are also concurrently within the AP's coverage range. Furthermore, from our tagged vehicle's viewpoint, its share of the AP's bandwidth resources fluctuates as a result of vehicle arrivals/departures to/from the AP's coverage range, as shown in Fig. 3.4. To aid our analysis, we assume for now that there is only a single class of wireless-equipped vehicles. Let the random variable Y_T denote our tagged vehicle's downloaded data at the end of its sojourn time T, and $N(t)$ denote the number of vehicles within the AP's coverage range during the sojourn time T. Therefore, we have

$$Y_T = \int_0^T \frac{B}{N(t)} dt \tag{3.10}$$

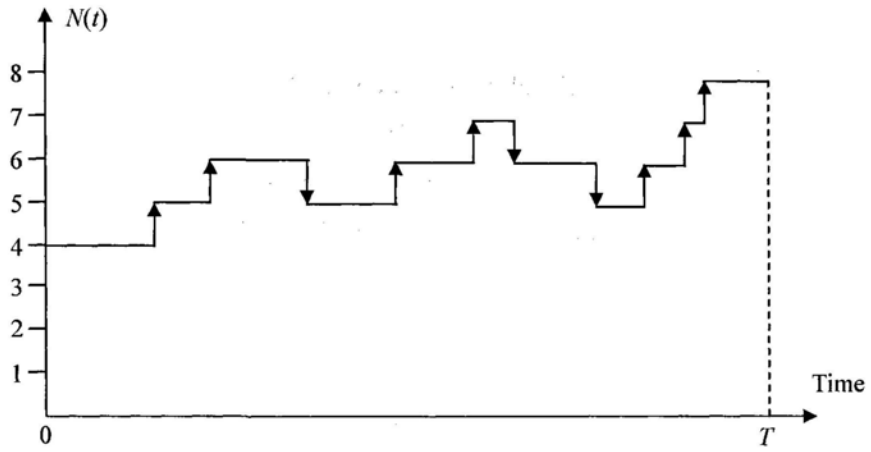


Figure 3.4 Variations in the number of vehicles in AP's coverage range during a tagged vehicle's sojourn time T

Take note that the insensitivity results with regards to the vehicle's sojourn time distributions (mentioned in the previous section) is only applicable to the computation of the average number of bytes received, Y_{avg}, and does not apply in the computation of the distribution of Y_T. Therefore, in our analytical derivation of the distribution of Y_T, we use the assumption of a steady-state vehicular traffic flow, i.e. each vehicle's speed v is the same and constant throughout its sojourn, and as such, the sojourn time of each vehicle is deterministic. In Section 3.6.3, we will compare our analytical results with simulations based on other non-deterministic vehicle sojourn time distributions.

Fig. 3.5 shows the vehicle arrivals/departures instants during our tagged vehicle's sojourn. If our tagged vehicle begins its sojourn as part of a group of i vehicles, and ends its sojourn as part of a group of j vehicles, this is equivalent to our tagged vehicle finding (i-1) vehicles already in the AP's coverage range upon its arrival, and leaving behind (j-1) vehicles upon its departure. Due to our assumption that every vehicle's speed v is the same and constant throughout its sojourn, there is no vehicle overtaking and hence, the (i-1) vehicles seen by our tagged vehicle upon its arrival will all depart before the end of its sojourn. In other words, the number of vehicle departures during our tagged vehicle's sojourn is exactly (i-1) vehicles. Similarly, the (j-1) vehicles left behind is the number of new vehicle arrivals during our tagged vehicle's sojourn. Therefore, $N(t)$ is a random process with (i-1) down-steps and (j-1) up-steps, starting with $N(0) = i$, and ending with $N(T) = j$. To derive the distribution of Y_T, we assume that the blocking probability of vehicle arrivals to the AP's coverage range is negligible and hence there is little or no blocking or losses of vehicles. Thus, with the assumption that the vehicle arrivals to the AP's coverage range is a Poisson process with rate λ ($=q$), and knowing that each vehicle's sojourn time is a constant value T, the vehicle departures from the AP's coverage range is also a Poisson process with rate λ. Hence, the arrival instants of the (j-1) vehicles and the departure instants of the (i-1) vehicles are both uniformly distributed over our tagged vehicle's sojourn period (0, T) [84]. Note that the inter-departure intervals τ_i's are dependent random variables since $(\tau_1+\tau_2+\ldots+\tau_{i-1}) \leq T$, and their joint density function is given by

$$f(\tau_1,\tau_2,\ldots,\tau_{i-1})=\frac{(i-1)!}{T^{i-1}} \tag{3.11}$$

It is clear from the above that $N(t)$ is a non-Markovian process. As such, we cannot use a standard Markov chain to model the vehicle departure process during our tagged vehicle's sojourn period (0, T). The additional complication caused by the interleaving of the vehicle arrival instants and departure instants during our tagged vehicle's sojourn leads to more difficulty in analyzing the distribution of Y_T. In the next two subsections, we discuss two approaches that we used to analyze the distribution of Y_T.

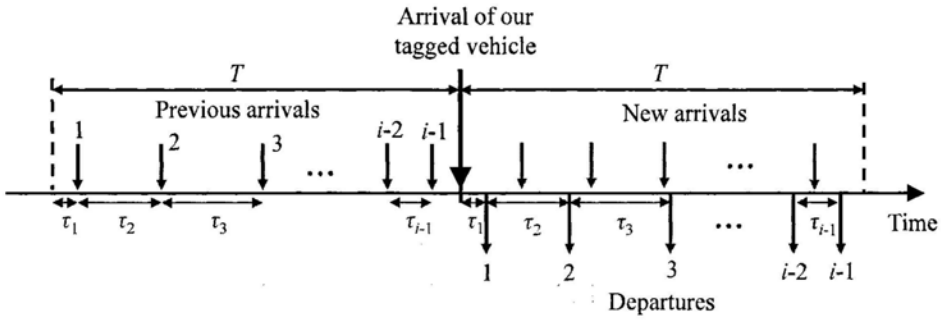


Figure 3.5 Vehicle arrivals and departures during a tagged vehicle's sojourn time T

3.5.1 The Interval-Decomposition Approach

As mentioned before, the inter-departure intervals are dependent random variables. However we note that during each inter-departure interval, there are only Poisson vehicle arrival events. Thus, a possible approach is to first decompose our tagged vehicle's sojourn period (0, T) into intervals separated by the departure instants. We can then easily analyze the distribution of data downloaded in each interval separately, conditioned on the number of vehicles at the beginning of each interval, and the number of vehicle arrivals in each interval. This is possible because with only Poisson vehicle arrivals in each interval, we know that the arrival instants are uniformly distributed over the duration of each interval, with their joint density function given by Eqn. (3.11). Finally, we can then use convolution on the distributional results from each interval to arrive at the distribution of Y_T. By defining

$$F[x,T|i,j,(i-1)] = \mathbb{P}\{Y_T \leq x \mid N(0) = i, N(T) = j, (i-1) \text{ departures}\} \quad (3.12)$$

we have that $F[x, T \mid i, j, (i\text{-}1)]$ is the distribution of Y_T conditioned on the events that our tagged vehicle begins its sojourn as part of a group of i vehicles, end its sojourn as part of a group of j vehicles, and that there are $(i\text{-}1)$ vehicle departures during the sojourn time T. This is also equivalent to $F[x, T \mid i, a, (i\text{-}1)]$, where $a = (j\text{-}1)$ is the number of new vehicle arrivals during our tagged vehicle's sojourn time T. As an approximation, we assume an M/D/∞ model for the vehicular traffic flow into the AP's coverage range, thereby allowing these vehicle arrivals to occur at any of the intervals, irrespective of the number of vehicles already present at the start of the interval.

In the following theorem, we use the notation $F^*[s, T \mid i, a, (i\text{-}1)]$ to denote the conditional Laplace-Stieltjes transform (LST) of $F[x, T \mid i, a, (i\text{-}1)]$. Using the interval-decomposition approach described above, we will compute the LST for the conditional distribution of Y_T by recursively evaluating the LST for the distribution of data downloaded in each interval separated by the vehicle departure instants.

Theorem 3.1:

The recurrence relationship that relates $F^*[s, T \mid i, a, (i\text{-}1)]$ to $F^*[s, \bullet \mid \bullet, \bullet, (i\text{-}2)]$ is given by

$$F^*[s,T \mid i,a,(i-1)]$$

$$= \int_0^T \sum_{l=0}^{a} F^*[s,\tau|i,l,0].F^*[s,T-\tau|i+l-1,a-l,(i-2)]$$

$$.\mathbb{P}\{l \text{ arrivals in } (0,\tau)\}.f(\tau)d\tau \quad (3.13)$$

where

$$F^*[s,\tau \mid i,l,0]$$

$$= \frac{1}{(sB\tau)^l}\sum_{m=0}^{l}(-1)^{m+l}\binom{l}{m}(i+m)^{l-1}[i(i+1)\ldots(i+l)]exp\left(-\frac{sB\tau}{i+m}\right) \quad (3.14)$$

$$\mathbb{P}\{l \text{ arrivals in } (0,\tau)\}$$

$$= \frac{\text{Prob}[l \text{ arrivals in } (0,\tau)]\ \text{Prob}[(a-l) \text{ arrivals in } (\tau,T)]}{\text{Prob}[a \text{ arrivals in } (0,T)]} \tag{3.15}$$

with the arrivals following a Poisson distribution, and

$$f(\tau) = \frac{i-1}{T^{i-1}}(T-\tau)^{i-2} \tag{3.16}$$

Proof:

By conditioning on the first vehicle departure instant τ, and then further conditioning on the number of vehicle arrivals in the first interval $(0, \tau)$, we obtain

$$F^*[s,T \mid i,a,(i-1)]$$

$$= \int_0^T F^*[s,T \mid i,a,(i-1);\ \tau]\, f(\tau)d\tau$$

$$= \int_0^T \sum_{l=0}^{a} F^*[s,T|i,a,(i-1);\tau,l \text{ arrivals in } (0,\tau)]\,.\,\mathbb{P}\{l \text{ arrivals in } (0,\tau)\}.f(\tau)d\tau \tag{3.17}$$

where $f(\tau)d\tau$ is the probability that the earliest of the $(i\text{-}1)$ vehicle departure instants (that are uniformly distributed in $(0, T)$) is contained in the interval $(\tau, \tau+d\tau)$, and is given by [84]:

$$f(\tau)d\tau = \frac{i-1}{T^{i-1}}(T-\tau)^{i-2}d\tau$$

Let us consider the interval (τ, T). If there had been l arrivals during the first interval $(0, \tau)$, then at the start of the next interval (τ, T), our tagged vehicle would be a part of a group of $(i+l\text{-}1)$ vehicles, and during this interval, there would be $(a\text{-}l)$ new vehicle arrivals. Let us denote

$F^*[s, \tau \mid i, l, 0]$ = LST of the distribution of data downloaded in the first interval $(0, \tau)$, conditioned on the events that our tagged vehicle starts this interval as part of a group of i vehicles and that there are l new vehicle arrivals during this interval with only vehicle arrival events, and

$F^*[s, T\text{-}\tau \mid i+l\text{-}1, a\text{-}l, (i\text{-}2)]$ = LST of the distribution of data downloaded in the interval (τ, T), conditioned on the events that our tagged vehicle starts this interval as part of a group of $(i+l\text{-}1)$ vehicles, there are $(a\text{-}l)$ new vehicle arrivals during this interval, and that there are $(i\text{-}2)$ vehicle departures during this interval.

We can then rewrite the first term under the sign of the summation in Eqn. (3.17) as $F^*[s, T \mid i, \text{a}, (i\text{-}1); \tau, l \text{ arrivals in } (0, \tau)] = F^*[s, \tau \mid i, l, 0] \cdot F^*[s, T\text{-}\tau \mid i+l\text{-}1, a\text{-}l, (i\text{-}2)]$, due to the independence of the distributions of data downloaded in the intervals $(0, \tau)$ and (τ, T) conditioned on the number of vehicles at the beginning of each interval. Hence, we can rewrite the last row of Eqn. (3.17) into Eqn. (3.13) of Theorem 3.1.

To determine the LST for the distribution of data downloaded in an interval with only vehicle arrival events, we consider Fig. 3.6. By conditioning on the inter-arrival times $(t_1, t_2, \ldots, t_l)$, we have

$$F^*[s, \tau \mid i, l, 0]$$

$$= \int_0^{\tau} \int_0^{\tau - t_1} \cdots \int_0^{\tau - t_1 - \cdots - t_{l-1}} F^*[s, \tau \mid i, l, 0;\ t_1, t_2, \ldots, t_l] \frac{l!}{\tau^l}\, dt_l \ldots dt_2 dt_1 \qquad (3.18)$$

The first term under the sign of integral in Eqn. (3.18) can be evaluated as

$$F^*[s, \tau \mid i, l, 0;\ t_1, t_2, \ldots, t_l]$$

$$= exp\left\{-s\left[\frac{B}{i} t_1 + \frac{B}{i+1} t_2 + \cdots + \frac{B}{i+l}(\tau - t_1 - t_2 - \cdots - t_l)\right]\right\} \qquad (3.19)$$

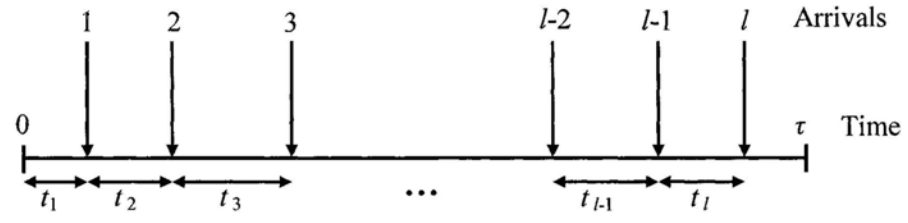


Figure 3.6 l vehicle arrivals during an interval $(0, \tau)$

By replacing Eqn. (3.19) into Eqn. (3.18), and then evaluating and simplifying the resultant expression, we will obtain Eqn. (3.14). A proof of Eqn. (3.14) by the method of induction is given in Appendix A. □

By expanding the right-hand side of Eqn. (3.13) in Theorem 3.1 recursively, and then inverting the resulting LST expression, we can obtain the conditional distribution of Y_T. We have evaluated and verified Theorem 3.1 using a small model. However, this approach does not scale with increasing number of vehicles, and as such motivates us to investigate a different method described in the next subsection. Nevertheless, we have gained valuable insights into the characteristics of our model. In particular, we have been evaluating the different sample paths characterizing the evolution of the random process $N(t)$, conditioned on the events $N(0) = i$, and $N(T) = j$. These sample paths are described by the different number of Poisson vehicle arrivals in each inter-departure interval. Each of these sample paths also satisfies the requirement that there must be (i-1) vehicle departures during the sojourn period (0, T). Using these insights, we proceed to our next approach in analyzing the distribution of Y_T.

3.5.2 The Markov Reward Process Approach

The major difficulty in the analysis of the distribution of Y_T is the requirement that there must be exactly (i-1) vehicle departures during our tagged vehicle's sojourn period (0, T). In our second approach, we take a closer look at the vehicle arrivals/departures instants during our tagged vehicle's sojourn time T, as shown in Fig. 3.5. We note that during our tagged vehicle's sojourn, there are new vehicle arrivals. From our tagged vehicle's viewpoint, these new vehicle arrivals follow a Poisson process with rate λ, and they occur independently of the current vehicle departures. A key observation from Fig. 3.5 is that the current vehicle departures are actually vehicle arrivals that arrived in the previous period T. If we assume that the blocking probability of vehicle arrivals to the AP's coverage range is negligible[4] and hence

[4] We have verified from our model simulations that even under substantial blocking conditions, this assumption has little effect on our results.

there is little or no blocking of vehicles, then the current inter-departure intervals (τ_i's) are dictated by the realization of the exponential inter-arrival intervals in the previous period T. Consequently, from our tagged vehicle's viewpoint, we can reconstruct all possible realizations of τ_i's from a Poisson arrival process of rate λ, with the additional requirement that $(\tau_1+\tau_2+\ldots+\tau_{i-1}) \leq T$.

Therefore with these insights, we can construct a 2-dimensional (2-D) continuous-time Markov chain $\{X_t, t \geq 0\}$ shown in Fig. 3.7 to model the vehicle arrivals and departures process, where X_t describes the state of the process at time t. To be more specific, the system states are described by $(d; N)$, where d represents the number of vehicle departures, and N represents the number of vehicles within the AP's coverage range. Vehicle arrivals will cause transitions from state $(d; N)$ to state $(d; N+1)$, while vehicle departures will cause transitions from state $(d; N)$ to state $(d+1; N-1)$. Observe that the 2-D Markov chain is acyclic, and the vehicle arrivals are limited by the system capacity C, which, in turn, is determined by the vehicular jam density k_{jam}. Let Q denote the infinitesimal generator matrix of the 2-D Markov chain, with its $(m, n)^{th}$ element given by

$$[Q]_{m,n} = \begin{cases} \lambda, & m=(d;N),\ n=(d;N+1), & 0 \leq d \leq i-1, & i-d \leq N \leq C-1 \\ \lambda, & m=(d;N),\ n=(d+1;N-1), & 0 \leq d \leq i-2, & i-d \leq N \leq C \\ \lambda, & m=(d;N),\ n=(>i-1;\_), & d = i-1, & i-d \leq N \leq C \\ -\lambda, & m=n=(d;N), & 0 \leq d \leq i-1, & N = C \\ -2\lambda, & m=n=(d;N), & 0 \leq d \leq i-1, & i-d \leq N \leq C-1 \\ 0, & \text{otherwise} & & \end{cases}$$

The key idea in our second approach is that the 2-D Markov chain helps us to enumerate all the relevant sample paths that would satisfy the requirement on the number of vehicle departures during our tagged vehicle's sojourn time T. Each relevant sample path requires us to specify its initial state and final state. For example in Fig. 3.7, if our tagged vehicle begins its sojourn as part of a group of i vehicles, then the initial state is $(0; i)$. Since we must have $(i-1)$ vehicle departures by time T, then the

desired final states can only be those from the set E={(i-1; 1), (i-1; 2), …, (i-1; C)} which correspond to (i-1) vehicle departures. All other sample paths that end at time T in other states where the number of vehicle departures is not equal to (i-1) will not be used during our computation of the distribution of Y_T. Finally by associating a reward rate (equal to our tagged vehicle's share of the AP's bandwidth = B/N, for the case of a single-class of vehicles with equal scheduling weight) with each state, we have a Markov reward process. Note that for computational purposes, all the states with i or more vehicle departures in Fig. 3.7 can be grouped together into one state.

For each relevant sample path, we can evaluate its contribution to the distribution of Y_T through the expression:

$$\text{Prob}\{Y_T > x, X_T = j \mid X_0 = (0; i)\} = F_{i,j}(x, T) \tag{3.20}$$

Eqn. (3.20) denotes the joint probability distribution of Y_T and the probability that the

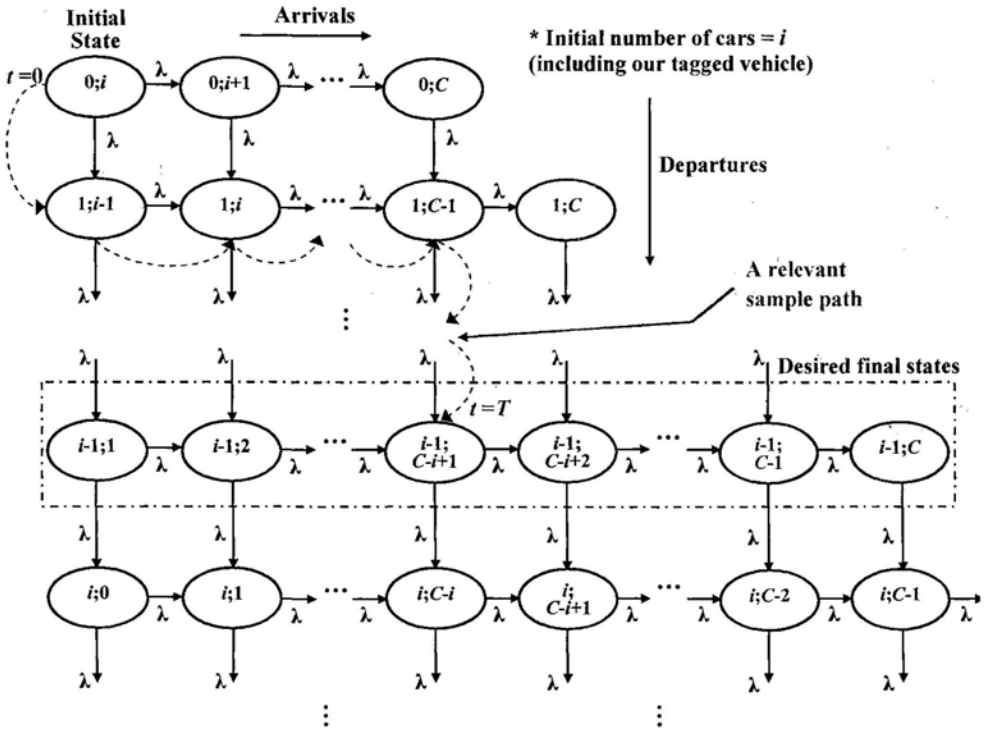


Figure 3.7 2-D Markov chain

Markov process is in state j (where $j \in E$) at time T, conditioned on the event that the Markov process was in state (0; i) at time 0. Eqn. (3.20) can be evaluated using the results from [71] where the author considered a continuous-time Markov chain with an infinitesimal generator matrix Q, and a reward rate $\rho(i)$ associated with each state i of the state space S. The number of distinct rewards is m+1, and their different values are $r_0 < r_1 < \cdots < r_{m-1} < r_m$. The state space S is partitioned into disjoint subsets A_l, $l = 0, \ldots, m$, containing the states with the same reward rates. It is shown in [71] that $F_{i,j}(x, t)$ satisfies the following hyperbolic forward partial differential equation (PDE):

$$\frac{\partial F_{i,j}(x,t)}{\partial t} = -\rho(j)\frac{\partial F_{i,j}(x,t)}{\partial x} + \sum_{k \in S} F_{i,k}(x,t)Q_{k,j} \tag{3.21}$$

By letting $F(x, t)$ denote the matrix containing the terms $F_{i,j}(x, t)$ for $i, j \in S$, a solution for the PDE in Eqn. (3.21) is given as follows.

For every $t > 0$, and $x \in [r_{h-1}t, r_h t)$, for $1 \le h \le m$,

$$F(x,t) = \sum_{n=0}^{\infty} e^{-\varphi t}\frac{(\varphi t)^n}{n!}\sum_{l=0}^{n}\binom{n}{l}x_h^l(1 - x_h)^{n-l}C^{(h)}(n,l) \tag{3.22}$$

where $x_h = \dfrac{x - r_{h-1}t}{[(r_h - r_{h-1})t]}$, $\varphi \ge \max\{-Q_{ii}, i \in S\}$, and

$C^h(n,l) = \{C^{(h)}_{A_u A_v}(n,l)\}_{0 \le u,v \le m}$ are matrices givenby the recurrence relations defined in [71]. The derivations leading to Eqns. (3.21) and (3.22) are summarized in Theorem B.1 in Appendix B.

The following theorem shows how we use the results from the contribution of each relevant sample path to compute the distribution of Y_T.

Theorem 3.2:

The distribution of Y_T is given as

$$\text{Prob}\{Y_T > x\} = \sum_{i=1}^{C}\sum_{j \in E}\frac{\text{Prob}\{Y_T > x, X_T = j \mid X_0 = (0;i)\}}{\sum_{l \in E}\text{Prob}[X_T = l \mid X_0 = (0;i)]}\mathbb{P}\{X_0 = (0;i)\} \tag{3.23}$$

where

$$\mathbb{P}\{X_0 = (0;i)\} = \frac{p_{i-1}}{\sum_{j=0}^{C-1} p_j} \tag{3.24}$$

$$\text{Prob}\{X_T = l \mid X_0 = (0;i)\} = (e^{QT})_{(0;i),l} \tag{3.25}$$

with Q, the infinitesimal generator of the 2-D Markov chain shown in Fig. 3.7, and

$$\text{Prob}\{Y_T > x, X_T = j \mid X_0 = (0;i)\} = F_{i,j}(x,T) \tag{3.26}$$

where $F_{i,j}(x, T)$ can be evaluated using Eqn. (3.22).

Proof:

By conditioning on the event that our tagged vehicle begins its sojourn as part of a group of i vehicles, we have

$$\text{Prob}\{Y_T > x\} = \sum_{i=1}^{C} \text{Prob}\{Y_T > x \mid X_0 = (0;i)\}\, \mathbb{P}\{X_0 = (0;i)\} \tag{3.27}$$

where $\mathbb{P}\{X_0=(0; i)\}$ can be interpreted as the normalized steady-state probability of our tagged vehicle finding (i-1) vehicles already in the AP's coverage range upon its arrival, and as such is given by Eqn. (3.24). The terms p_{i-1} and p_j in Eqn. (3.24) can be evaluated using Eqn. (3.3).

Further conditioning of the first term under the sign of the summation in Eqn. (3.27) on the event that the sample path in the 2-D Markov chain ends in one of the desired final states in the set E, gives us

$$\begin{aligned}&\text{Prob}\{Y_T > x \mid X_0 = (0;i)\} \\ &= \sum_{j \in E} \text{Prob}\{Y_T > x \mid X_0 = (0;i), X_T = j\}\, \mathbb{P}\{X_T = j \mid X_0 = (0;i)\}\end{aligned} \tag{3.28}$$

where

$$\mathbb{P}\{X_T = j \mid X_0 = (0;i)\} = \frac{\text{Prob}[X_T = j \mid X_0 = (0;i)]}{\sum_{l \in E} \text{Prob}[X_T = l \mid X_0 = (0;i)]} \tag{3.29}$$

is the normalized transition probability of the Markov process being in state j at time T, given that it was in state (0; i) at time 0. As for the first term under the sign of the summation in Eqn. (3.28), it is equivalent to the following

$$\text{Prob}\{Y_T > x \mid X_0 = (0;i), X_T = j\} = \frac{\text{Prob}\{Y_T > x, X_T = j \mid X_0 = (0;i)\}}{\text{Prob}[X_T = j \mid X_0 = (0;i)]} \tag{3.30}$$

Finally, using Eqns. (3.29) and (3.30) in Eqn. (3.28), and then replacing Eqn. (3.28) into Eqn. (3.27), gives us Eqn. (3.23) in Theorem 3.2. □

Referring back to Eqn. (3.8), we were able to compute the value of the maximum number of bytes that our tagged vehicle can receive by the end of its sojourn. This corresponds to the scenario where throughout its whole sojourn, our tagged vehicle has the AP's resources all to itself, i.e. it does not have to share the AP's resources with any other vehicles. Using the 2-D Markov chain, we can now compute the probability of this event occurring as

$$\begin{aligned}&\text{Prob}\{\text{Our tagged vehicle received max number of bytes at the end of its sojourn}\}\\&= \mathbb{P}\{X_0 = (0;\,1)\}.\,\mathbb{P}\{X_T = (0;\,1) \mid X_0 = (0;\,1)\}\end{aligned} \tag{3.31}$$

where $\mathbb{P}\{X_0 = (0;\,1)\}$ can be computed using Eqn. (3.24) by replacing i=1, and $\mathbb{P}\{X_T = (0;\,1) \mid X_0 = (0;\,1)\}$ can be computed using Eqn. (3.29) by replacing j=(0; 1) and i=1, with the set E={(0; 1), (0; 2), …, (0; C)}.

3.5.3 Extension to the Markov Reward Process Approach

The approach described in the previous subsection can be extended to handle the case where we have multiple classes of wireless-equipped vehicles receiving different weights of service (including zero) from the AP. For example, if there are two classes of vehicles (Class_A = a class of non-wireless-equipped vehicles, and Class_B = a class of wireless-equipped vehicles), then instead of a 2-D Markov chain, we would need to model the vehicle arrivals and departures process with a 3-D Markov chain. The system states will be described by (d_A, d_B; N_A, N_B), where d_A (d_B) represents the number of Class_A (Class_B) vehicle departures, and N_A (N_B) represents the number of Class_A (Class_B) vehicles within the AP's coverage range. Class_A vehicle arrivals will cause transitions from state (d_A, d_B; N_A, N_B) to state (d_A, d_B; N_A+1, N_B), while Class_B vehicle arrivals will cause transitions from state (d_A, d_B; N_A, N_B) to state (d_A, d_B; N_A, N_B+1). Similarly, Class_A vehicle departures will cause transitions from state (d_A, d_B; N_A, N_B) to state (d_A+1, d_B; N_A-1, N_B), while Class_B vehicle departures will cause transitions from state (d_A, d_B; N_A, N_B) to state (d_A, d_B+1; N_A, N_B-1). The

Class_A (Class_B) vehicle arrivals and departures processes are independent of each other, and are also time-invariant Poisson processes with rate $\alpha_A\lambda$ ($\alpha_B\lambda$), where $\alpha_A(\alpha_B)$ is the fraction of Class_A (Class_B) vehicles and $\alpha_A+\alpha_B = 1$. Fig. 3.8 shows the state transition diagram for the 3-D Markov chain where our tagged vehicle (which must be a Class-B vehicle) is assumed to begin its sojourn as part of a group of i Class-A vehicles and j Class-B vehicles. We see that the states are grouped into separate levels defined by the number of departures that have already occurred. Transitions in each level are due to vehicle arrivals only, while transitions between consecutive levels are due to vehicle departures. Fig. 3.8 also shows a special state denoted by "States with (>i,_) or (_,>j-1) departures", which groups together all the states that have "more than i Class_A vehicle departures" or "more than (j-1) Class_B vehicle departures". Let Q denote the infinitesimal generator matrix of the 3-D Markov chain, with its $(m, n)^{th}$ element given by:

$[Q]_{m,n} =$

$$
\begin{cases}
\alpha_A\lambda, & m=(d_A, d_B; N_A, N_B),\ n=(d_A, d_B; N_A+1, N_B),\ 0 \le d_A \le i,\ 0 \le d_B \le j\text{-}1 \\
& j\text{-}d_B \le N_B \le C\text{-}(i\text{-}d_A)\text{-}1,\ i\text{-}d_A \le N_A \le C\text{-}1\text{-}N_B \\
\alpha_B\lambda, & m=(d_A, d_B; N_A, N_B),\ n=(d_A, d_B; N_A, N_B+1),\ 0 \le d_A \le i,\ 0 \le d_B \le j\text{-}1 \\
& j\text{-}d_B \le N_B \le C\text{-}(i\text{-}d_A)\text{-}1,\ i\text{-}d_A \le N_A \le C\text{-}1\text{-}N_B \\
\alpha_A\lambda, & m=(d_A, d_B; N_A, N_B),\ n=(d_A+1, d_B; N_A-1, N_B),\ 0 \le d_A \le i\text{-}1,\ 0 \le d_B \le j\text{-}1 \\
& j\text{-}d_B \le N_B \le C\text{-}(i\text{-}d_A),\ i\text{-}d_A \le N_A \le C\text{-}N_B \\
\alpha_B\lambda, & m=(d_A, d_B; N_A, N_B),\ n=(d_A, d_B+1; N_A, N_B\text{-}1),\ 0 \le d_A \le i,\ 0 \le d_B \le j\text{-}2 \\
& j\text{-}d_B \le N_B \le C\text{-}(i\text{-}d_A),\ i\text{-}d_A \le N_A \le C\text{-}N_B \\
\alpha_A\lambda, & m=(d_A, d_B; N_A, N_B),\ n=\text{"States with (>}i\text{,\_) or (\_,>}j\text{-}1\text{) departures"}, \\
& d_A = i,\ 0 \le d_B \le j\text{-}1,\ j\text{-}d_B \le N_B \le C,\ i\text{-}d_A \le N_A \le C\text{-}N_B \\
\alpha_B\lambda, & m=(d_A, d_B; N_A, N_B),\ n=\text{"States with (>}i\text{,\_) or (\_,>}j\text{-}1\text{) departures"}, \\
& d_B = j\text{-}1,\ 0 \le d_A \le i,\ j\text{-}d_B \le N_B \le C\text{-}(i\text{-}d_A),\ i\text{-}d_A \le N_A \le C\text{-}N_B \\
-\lambda, & m=n=(d_A, d_B; N_A, N_B),\ 0 \le d_A \le i,\ 0 \le d_B \le j\text{-}1 \\
& j\text{-}d_B \le N_B \le C\text{-}(i\text{-}d_A),\ N_A = C\text{-}N_B \\
-2\lambda, & m=n=(d_A, d_B; N_A, N_B),\ 0 \le d_A \le i,\ 0 \le d_B \le j\text{-}1 \\
& j\text{-}d_B \le N_B \le C\text{-}(i\text{-}d_A)\text{-}1,\ i\text{-}d_A \le N_A \le C\text{-}1\text{-}N_B \\
0, & \text{otherwise}
\end{cases}
$$

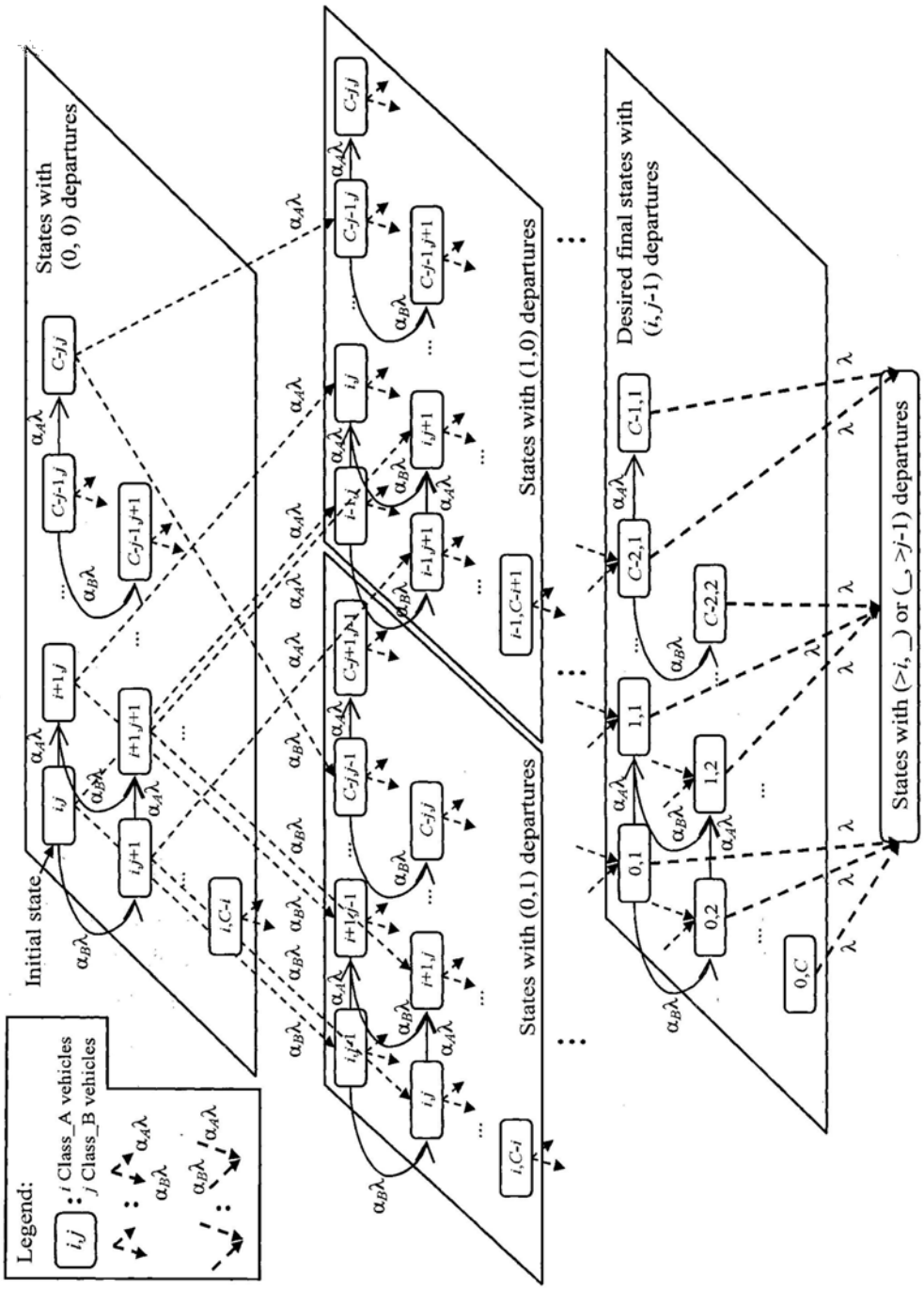


Figure 3.8 3-D Markov chain

A reward rate of (B/N_B) is also associated with each state with N_B Class_B vehicles, corresponding to our tagged vehicle's share of the AP's bandwidth resources in that state. In terms of the relevant sample paths that would satisfy the required behavior of our original vehicular flow system, if our tagged vehicle begins its sojourn as part of a group of i Class_A and j Class_B vehicles, then the initial state is $(0, 0; i, j)$. The desired final states of the sample paths would be those from the set $E = \{(i, j\text{-}1; N_A, N_B)\}$, where $N_A \geq 0$, $N_B \geq 1$, and $N_A + N_B \leq C$.

Similar to Eqn. (3.31), we can also compute the probability of the event that throughout its whole sojourn, our tagged vehicle has the AP's resources all to itself, i.e. it does not have to share the AP's resources with any other wireless-equipped Class-B vehicles. Note that there could be other non-wireless-equipped Class_A vehicles on the road during our tagged vehicle's sojourn, but these vehicles do not contend for the AP's resources. The probability of this event can be computed as

Prob{Our tagged vehicle received max number of bytes at the end of its sojourn}

$$= \sum_{i=0}^{C-1} \mathbb{P}\{X_0 = (0, 0; i, 1)\} . \sum_{k=0}^{C-1} \mathbb{P}\{X_T = (i, 0; k, 1) \mid X_0 = (0, 0; i, 1)\} \qquad (3.32)$$

where

$\mathbb{P}\{X_0 = (0, 0; i, 1)\}$ is the normalized steady-state probability of our tagged vehicle beginning its sojourn as the only Class_B vehicle, together with a group of i Class_A vehicles, and

$\mathbb{P}\{X_T = (i, 0; k, 1) \mid X_0 = (0, 0; i, 1)\}$ is the normalized transition probability of the process being in state $(i, 0; k, 1)$ at time T, given that it was in state $(0, 0; i, 1)$ at time 0, with the set $E = \{(i, 0; N_A, N_B)\}$, where $N_A \geq 0$, $N_B \geq 1$, and $N_A + N_B \leq C$.

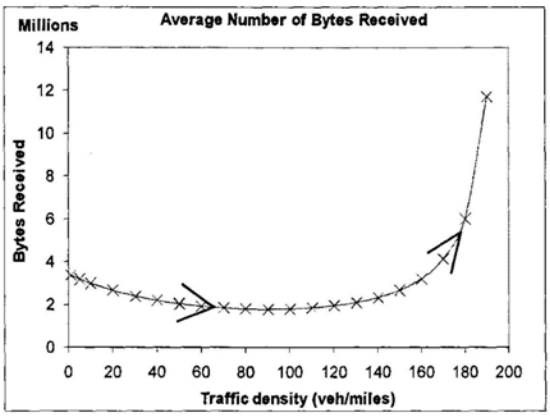
In general, the problem of computing the distribution of the number of bytes received in the case of multiple classes of wireless-equipped vehicles receiving different weights of service (including zero) from the AP, can be solved by (i) designing a different set of reward rates to be associated with the states of the Markov reward process and (ii) expanding the dimension of the corresponding Markov chain.

3.6 Results and Discussion

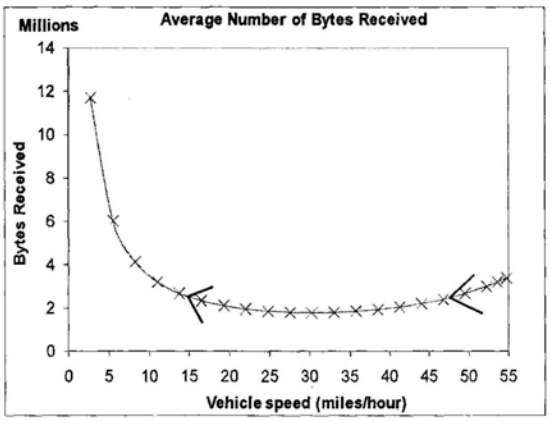
For evaluation purposes, we consider an 802.11b based AP with coverage range L = 0.038 miles (≈ 60m), and transmission rate 11Mbps. The traffic jam density k_{jam} = 200 veh/miles, the free-flow speed v_f = 55mph and the system capacity C = 8 vehicles. Unless stated otherwise, all the results are obtained for the case of a single class of vehicles in which all vehicles equally share the AP's bandwidth resources.

3.6.1 M/D/C/C Model

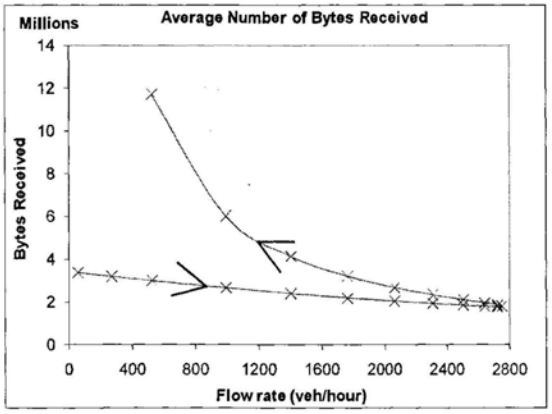
Fig. 3.9 shows the average number of bytes received by a drive-thru vehicle (Y_{avg}) as a function of the three vehicular traffic parameters, namely, vehicle density k, speed v and flow-rate q (= λ). In Fig. 3.9(a), we observe that at low traffic densities (corresponding to high vehicle speed and low traffic flow-rate), there are very few vehicles and as such there is very little contention for the AP's bandwidth resources. As the traffic density increases, bandwidth contention increases and Y_{avg} gradually decreases even though the vehicle sojourn time T has increased due to the lower vehicle speeds. This shows that at low-to-medium traffic densities, the impact of the vehicles' contention for the AP's bandwidth resources has more effect on Y_{avg} compared to that of the vehicle sojourn time. When traffic density increases to approach the jam density, the vehicles covered by the AP are virtually stand-still, i.e. v approaches zero while T approaches infinity. Since there is only a finite number of vehicles ($\leq C$) contending for the AP's bandwidth, the average number of bytes received by each "stand-still" vehicle also approaches infinity. Therefore, at high traffic densities approaching the jam density, Y_{avg} is influenced more by the vehicle sojourn time. By using the relationship between vehicle density, speed and flow-rate (refer to the description of Fig. 3.2 in Section 3.3), the effect of vehicle speed and flow rate on Y_{avg} in Fig. 3.9(b) and Fig. 3.9(c) can also be explained in the same manner.



(a)



(b)



(c)

Figure 3.9 Average number of bytes received (Y_{avg}) versus (a) traffic density; (b) vehicle speed; (c) flow rate

Fig. 3.10 shows the range of values for the number of bytes that a vehicle can receive from the AP per pass. The results here are for the case where there are two classes of vehicles, i.e. wireless-equipped and non-wireless-equipped vehicles. The fraction of the former class of vehicles can be interpreted as the service penetration rate p of wireless-equipped vehicles in the market. The values of the minimum (Y_{min}) and maximum (Y_{max}) number of bytes that can be received are evaluated using Eqns. (3.7) and (3.8). From Fig. 3.10, we make the following interesting observations:

a) For lower values of service penetration rate p, Y_{avg} is higher due to the lower number of contending vehicles for the AP's resources.
b) At low traffic densities, the values of Y_{avg} are close to that of Y_{max} for all values of p. This is due to the low number of contending vehicles and hence a vehicle is more likely to be able to receive the maximum number of bytes.
c) At higher traffic densities, we see that for high values of p, the values of Y_{avg} are closer to that of Y_{min} while for low values of p, the values of Y_{avg} are closer to that of Y_{max}. This is because at high traffic densities, there would be a lot of vehicles within the AP's coverage range. If service penetration rate is high, then there would be a lot of vehicles contending for the AP's resources, leading to lower values of Y_{avg}. On the other hand, if service penetration rate is low, then there would be fewer number of contending vehicles and as such, the values of Y_{avg} would be closer to that of Y_{max}.
d) As traffic density increases, we see that the values of Y_{min} and Y_{max} also increase due to the corresponding increase in the vehicle sojourn time.

From the results in Fig. 3.10, we are also interested in the probability that a vehicle can receive the maximum number of bytes (Y_{max}) by the end of its sojourn. Using Eqns. (3.31) and (3.32), we compute this probability for different values of the service penetration rate p. As expected, Fig. 3.11 shows that a vehicle is more likely to receive the maximum number of bytes under the low traffic density regime, as well as under the low service penetration rate regime. Under these regimes, there is little contention for the AP's resources, and as such a vehicle is more likely to receive the maximum number of bytes by the end of its sojourn.

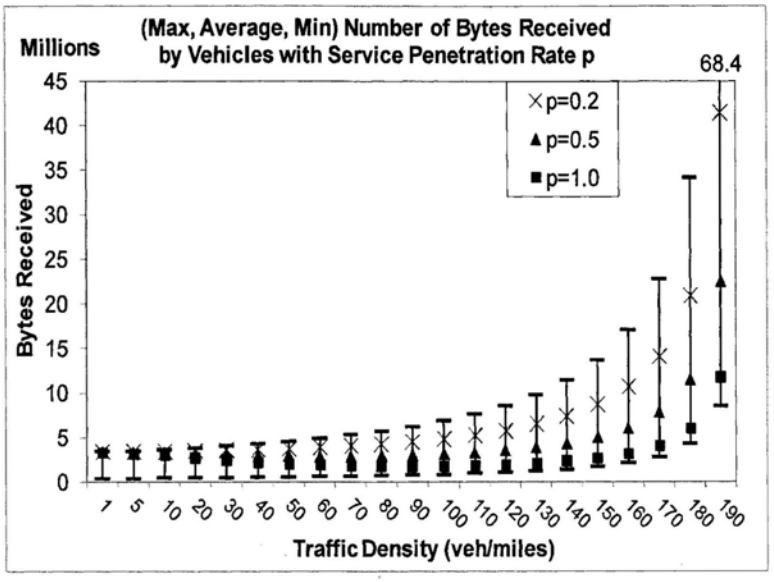


Figure 3.10 The range of values for the number of bytes received by vehicles with service penetration rate p

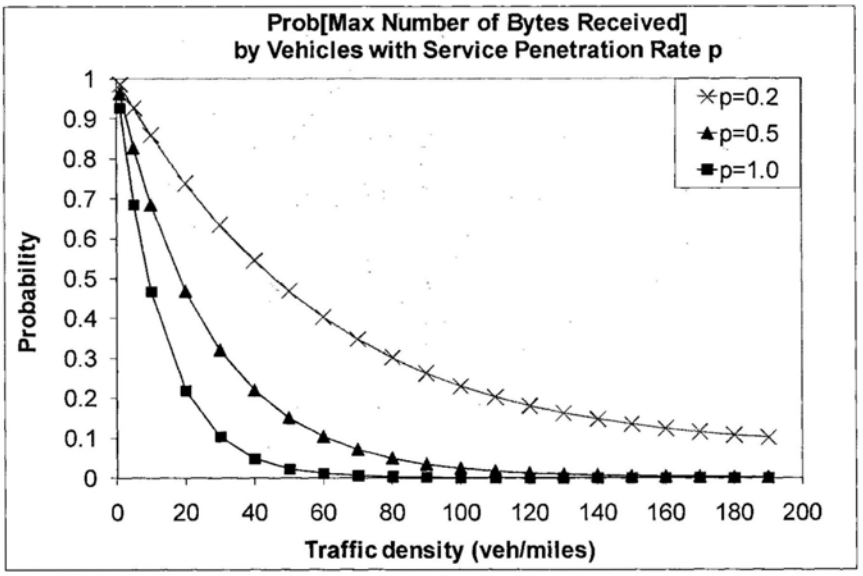
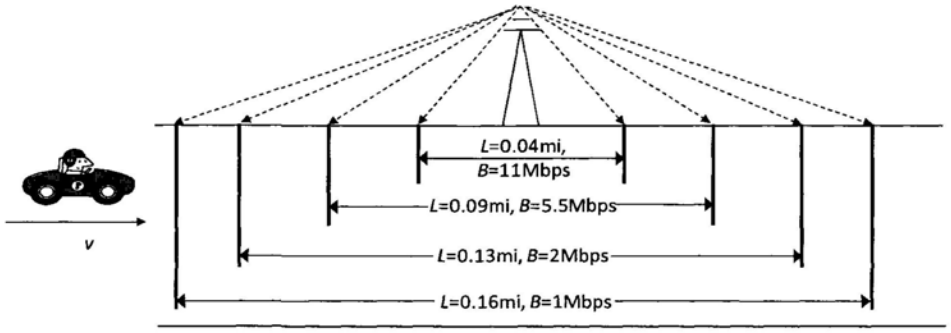


Figure 3.11 The probability of vehicles with service penetration rate p receiving the maximum number of bytes

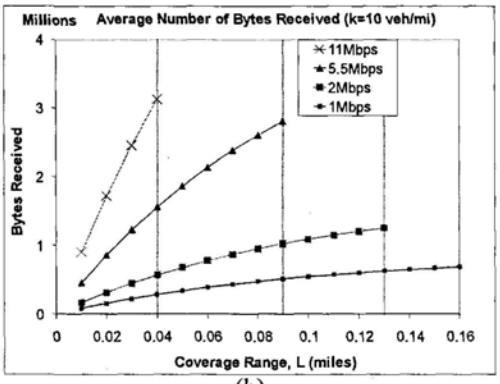
Next, we want to investigate the impact of the AP's coverage range L versus transmission rate tradeoff on Y_{avg}. The previous results have been obtained using a fixed transmission rate 11Mbps over a fixed coverage range L = 0.038 miles. In this investigation, by increasing the AP's coverage range at the expense of lower transmission bit rate, we would like to investigate the impact on Y_{avg}. We use the experimental data from [85] for the transmission ranges corresponding to the different transmission bit rates, as shown in Fig. 3.12(a). We also assume that the adaptive bit rate feature of the AP is not turned on so that our drive-thru vehicle receives service at a constant bit rate throughout its sojourn. Fig. 3.12 shows the results of this investigation, where the AP's transmission rate is 11Mbps for a coverage range up to 0.04 miles (≈ 64m), 5.5Mbps for coverage range up to 0.09 miles (≈ 145m) and so on. Fig. 3.12 shows the results for three vehicle densities, a low-density, a medium-density, and a high-density scenario. In general, we see that for all three scenarios, increasing the AP's coverage range will cause a decrease in the amount of data downloaded. This is due to two factors, namely, when we increase the AP's coverage range, the transmission rate will decrease, and more importantly, the number of contending vehicles will increase as well.

We now investigate the impact of increasing the AP's coverage range versus transmission rate tradeoff on Y_{avg}, with the AP's adaptive bit rate feature turned on. In this case, a drive-thru vehicle receives service at different bit rates depending on its distance from the AP. The average number of bytes received, Y_{avg} is computed using Eqn. (3.9). We also do a comparison with the non-adaptive bit rate case where the AP is set to transmit at a fixed bit rate of 11Mbps for L = 0.038 miles. The distances and the corresponding bit rates for the "adaptive" and "non-adaptive" cases are shown in Fig. 3.13. Fig. 3.14 shows the comparison results. We observe that except for the low traffic densities, the non-adaptive bit rate case (with shorter AP coverage range) performs better in terms of the average number of bytes received by a drive-thru vehicle. This is because the adaptive bit rate case, due to its longer AP coverage range, results in a higher number of contending vehicles for the AP's resources and as such the average number of bytes received by a drive-thru vehicle is smaller. Hence, we

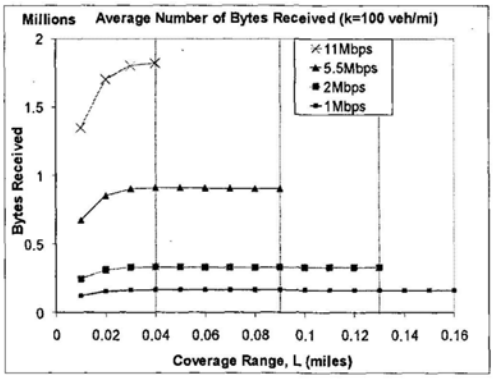
again see that increasing the AP's coverage range may not necessarily increase the amount of data downloaded by a drive-thru vehicle.



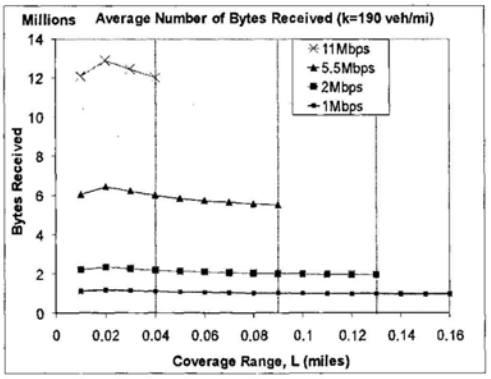
(a)



(b)



(c)



(d)

Figure 3.12 Impact of AP's coverage range versus transmission rate tradeoff on the average number of bytes received (Y_{avg}) for traffic densities
(b) 10 veh/miles; (c) 100 veh/miles; (d) 190 veh/miles

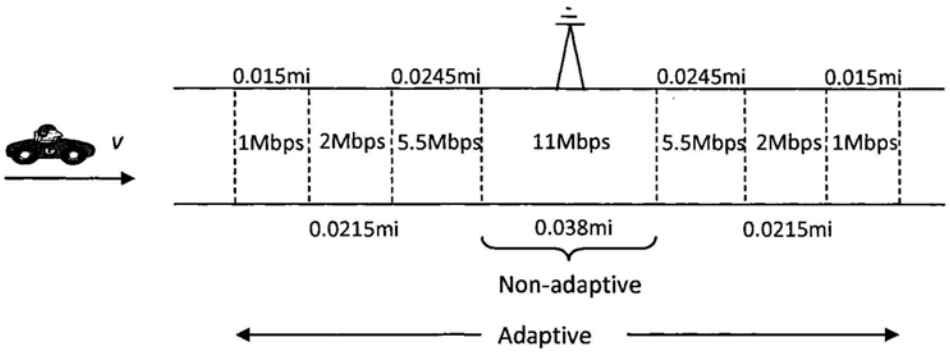


Figure 3.13 Adaptive versus non-adaptive bit rate cases

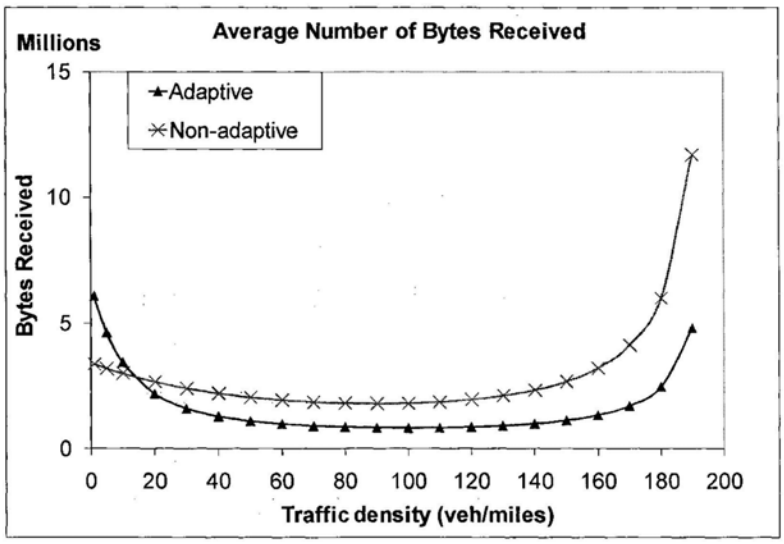
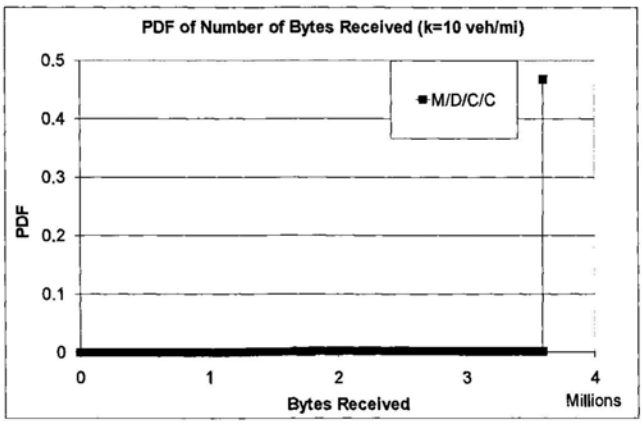


Figure 3.14 Comparison of the impact of AP's adaptive vs. non-adaptive bit rate feature on the average number of bytes received (Y_{avg})

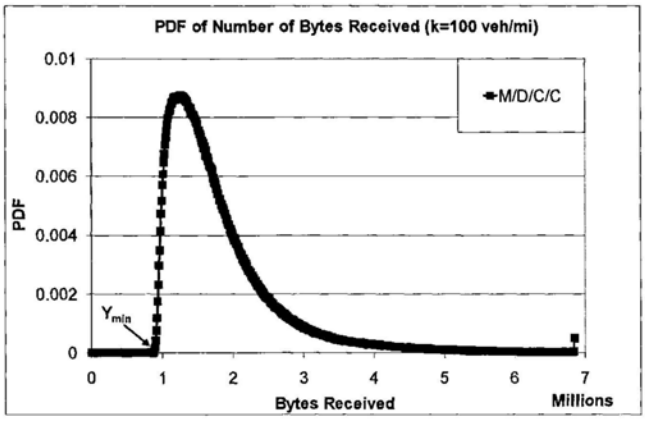
Fig. 3.15 shows the distribution of the number of bytes received (Y_T), where the results can be used to obtain a complete probabilistic characterization of the number of bytes that a vehicle can receive by the end of its sojourn. From Figs. 3.15(b) and 3.15(c), we can also determine Y_{min} as defined in Eqn. (3.7). In addition, the discrete jump observed at the maximum value of bytes received in Figs. 3.15(a) and 3.15(b) corresponds to the event that our tagged vehicle begins its sojourn by itself, and that throughout its whole sojourn, there were no other vehicle arrivals. The probabilities of this event can be computed from Eqn. (3.31).

Fig. 3.16 shows the distribution of the number of bytes received (Y_T) by vehicles with service penetration rate p, similar to the scenario explained in the discussion of Fig. 3.10. We again see the discrete jump at the maximum value of bytes received for the service penetration rates p=0.2 and p=0.5. These discrete jumps correspond to the event that our tagged vehicle begins its sojourn as the only wireless-equipped vehicle, and that throughout its whole sojourn, there were no other arrivals of wireless-equipped vehicles. The probability of this event can be computed from Eqn. (3.32). The results in Fig. 3.16 show that for a specified grade of service (e.g. 95%), the corresponding minimum number of bytes received by a drive-thru vehicle is higher for lower values of p. This is due to the fact that lower values of p translate to lower number of contending vehicles for the AP's resources.

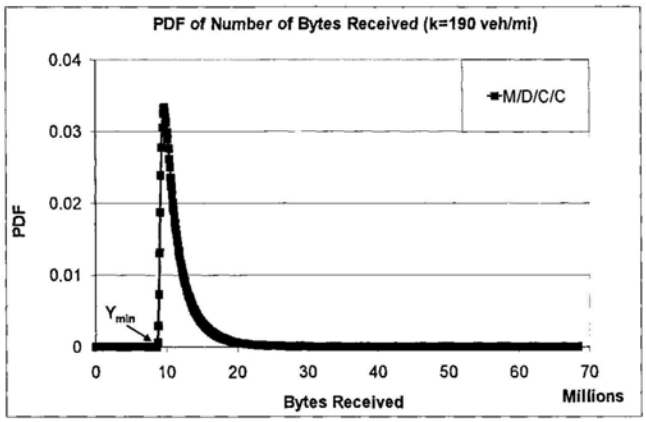
In terms of complexity, the main effort is in evaluating Eqn. (3.22) using Theorem B.1 in Appendix B. Using the algorithm proposed in [72], both the computational cost and storage complexity are $O(C^5)$ for the case of a single vehicle class using the 2-D Markov chain model. For the case of two vehicle classes using the 3-D Markov chain model, both the computational cost and storage complexity are $O(C^9)$, where C is the maximum number of vehicles that can be accommodated in the AP's coverage range. While the order of complexity may look formidable, we note that in practice, values of C are quite small due to the limited coverage range of Drive-thru Internet systems. For instance in our evaluation of the distributions, using practical WiFi coverage range of L = 0.038 miles ($\approx$ 60m), with jam density k_{jam} = 200 veh/miles, C = 8 vehicles.



(a)



(b)



(c)

Figure 3.15 Distribution of the number of bytes received (Y_T) for traffic densities (a) 10 veh/miles; (b) 100 veh/miles; (c) 190 veh/miles

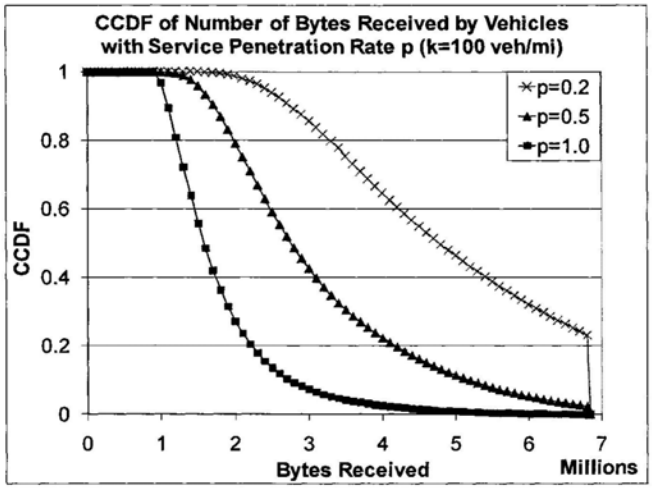


Figure 3.16 Distribution of the number of bytes received (Y_T) by vehicles with service penetration rate p, for traffic density=100 veh/miles

In our evaluations using an Intel Pentium 4 3.2GHz machine with 1GB RAM, the three distributions plotted in Fig. 3.15 (with number of data points ranging from 36 to 68) took less than 5 minutes to compute. The results in Fig. 3.15 are for eight possible initial states, with the corresponding 2-D Markov chains varying in size from nine to 37 states. On the other hand, the results in Fig. 3.16 (with p=0.2 and p=0.5) are for 36 possible initial states, with the corresponding 3-D Markov chains varying in size from 37 to 281 states. Hence, the results in Fig. 3.16 (with number of data points = 68) took longer, about 3 hours to compute. Note that in general, we are more interested in the data point corresponding to a specified grade of service or percentile, and as such the time taken to compute this data point would be much shorter. This data point can be found by first doing a linear interpolation between the values of Y_{min} and Y_{max}, and then selecting an estimated point on the linear line corresponding to the specified percentile. We can then use the binary search procedure to narrow down the range of values of the bytes received. By doing this iteratively, we can then find the value of the amount of bytes received corresponding to the specified percentile.

3.6.2 Comparison with Non-Poisson Vehicle Arrival Patterns

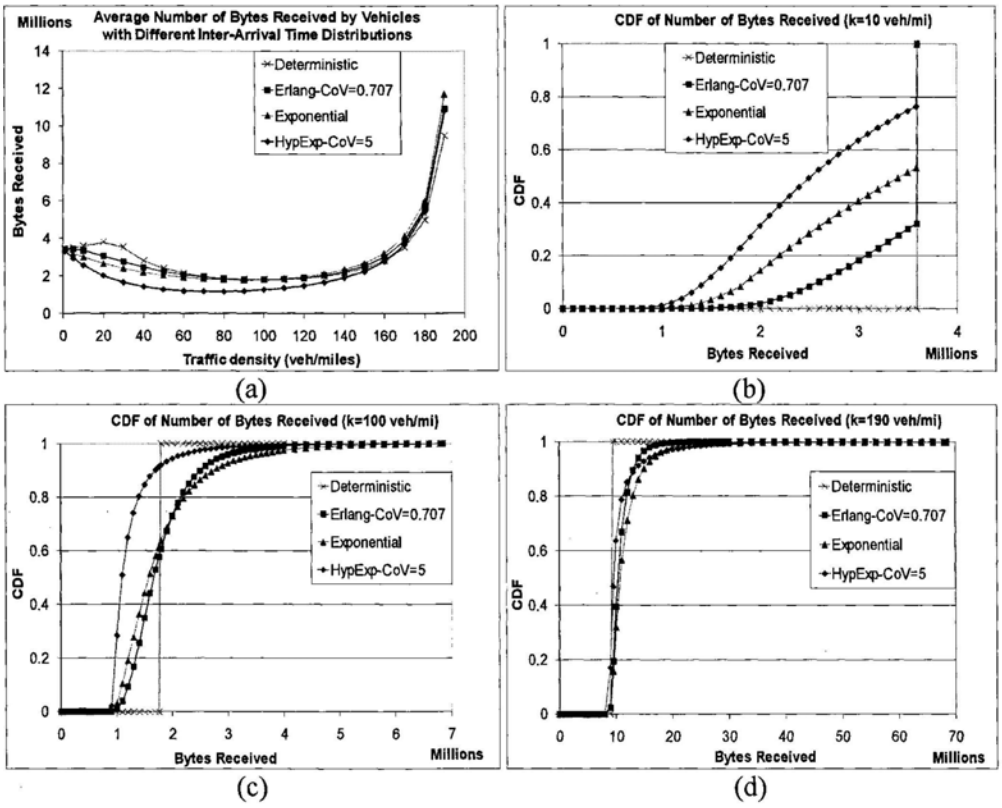


Figure 3.17 Comparison with non-Poisson vehicle arrival patterns for (a) average, and distribution of number of bytes received at traffic densities (b) 10 veh/mi; (c) 100 veh/mi; (d) 190 veh/mi

Fig. 3.17 shows the sensitivity impact of the vehicle arrival process on the amount of data downloaded by a drive-thru vehicle. We compare the results for different arrival patterns, with the average inter-arrival time for all the different arrival processes set to be the same. Qualitatively, the trend of the plots is the same for the different arrival patterns. Quantitatively, the amount of data downloaded is different compared to our model assumption of Poisson arrivals, and we see that the impact of the different arrival patterns is not uniform across the range of the traffic densities.

3.6.3 Comparison with Non-Deterministic Vehicle Sojourn Times

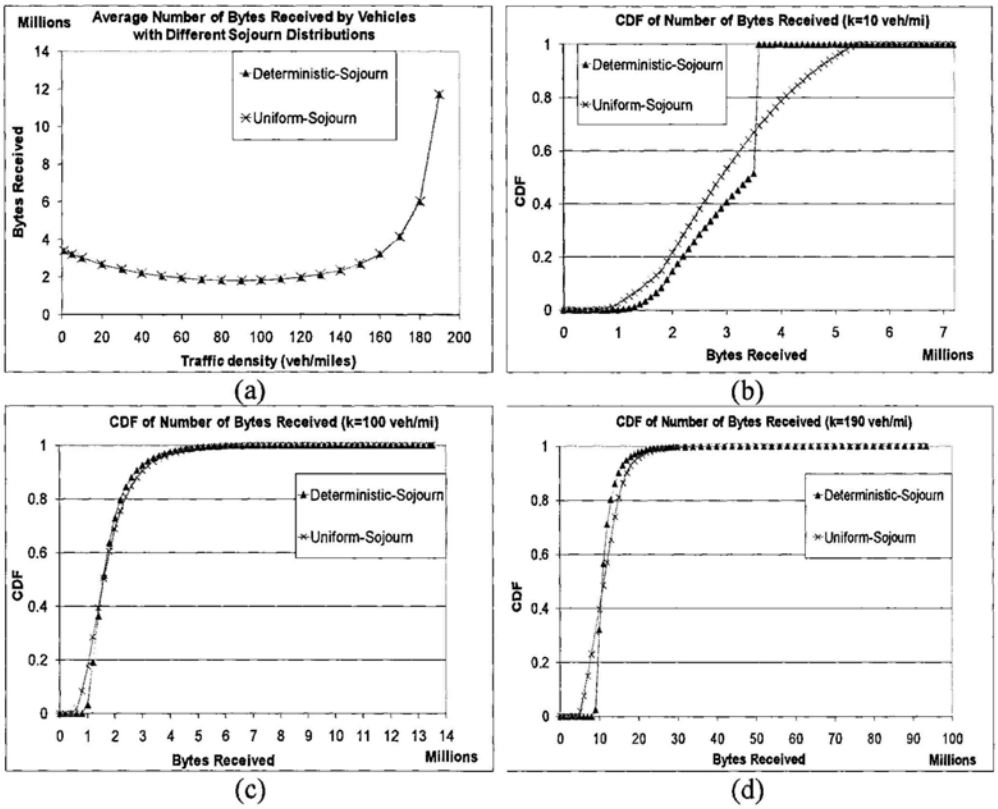


Figure 3.18 Comparison with non-deterministic vehicle sojourn times for (a) average, and distribution of number of bytes received at traffic densities (b) 10 veh/mi; (c) 100 veh/mi; (d) 190 veh/mi

Fig. 3.18 shows a comparison of the results when the vehicle sojourn times are relaxed from deterministic to be uniformly distributed. For the uniformly distributed vehicle sojourn times, the sojourn time T of each vehicle is a random variable, uniformly distributed between $[0.5L/v, 1.5L/v]$ where L is the AP's coverage range, and v is the average vehicle speed computed using Greenshields' model in Eqn. (3.2). The mean vehicle sojourn time in both cases of sojourn distributions is set to be the same. As described earlier in Section 3.4, the insensitivity property of the M/G/C/C queueing system states that the steady-state system size probability is only a function of the mean vehicle sojourn time, and is independent of the vehicle sojourn time distribution.

Therefore in our evaluation here, Fig. 3.18(a) shows that the average number of bytes received by vehicles with deterministic and uniformly distributed sojourn times is exactly the same. This is due to the fact that the mean vehicle sojourn time for the uniformly distributed case is L/v, the same as that of the deterministic case. However, the insensitivity results is only applicable to the computation of the average number of bytes received, and does not apply in the computation of the distribution of the number of bytes received. As such, as seen from Figs. 3.18(b), 3.18(c), and 3.18(d), the distribution plots for the deterministic and uniformly distributed cases do not match.

3.6.4 Comparison with M/D/C Model

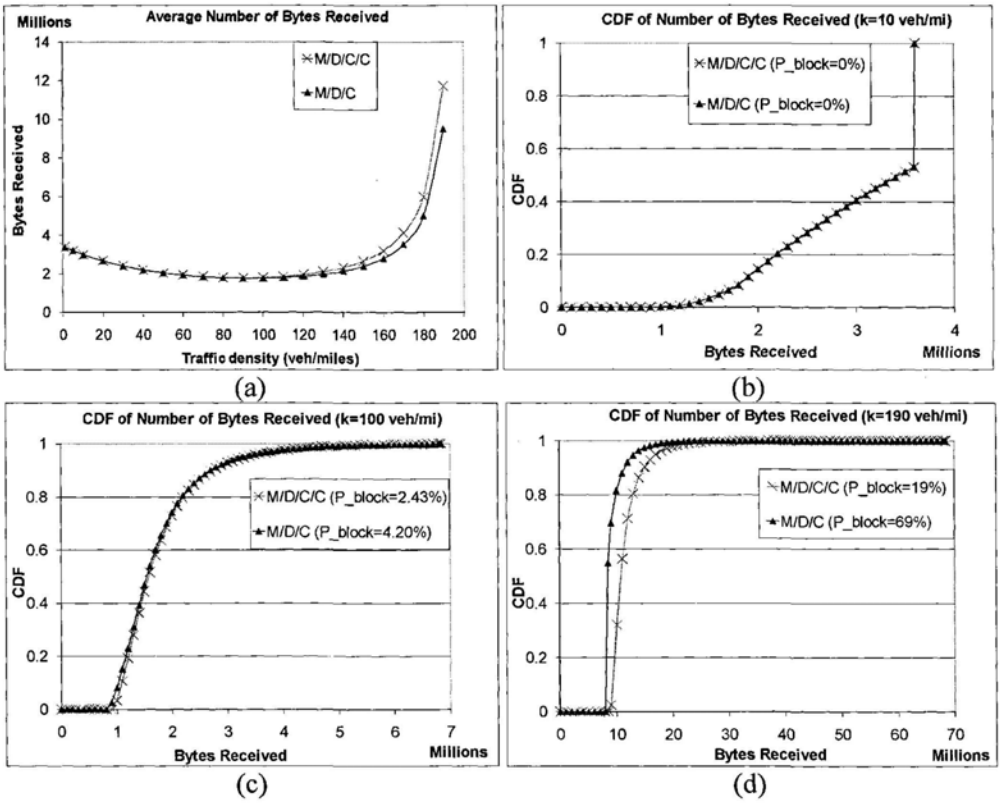


Figure 3.19 Comparison with the M/D/C model for (a) average, and distribution of number of bytes received at traffic densities (b) 10 veh/mi; (c) 100 veh/mi; (d) 190 veh/mi

Fig. 3.19 shows a comparison of the results from the M/D/C/C and M/D/C models. As seen in Figs. 3.19(a), 3.19(b) and 3.19(c), for low to medium traffic densities, the probability of blocking is negligible and as such, the average and distribution results for both models are the same. On the other hand, for high traffic densities, the probability of blocking is no longer negligible. In such a scenario, blocked vehicles in the M/D/C model are not lost but are delayed in the queue, and as such the probability of blocking for the M/D/C model would be higher compared to that of the M/D/C/C model. In addition, the system in an M/D/C model would be more likely to have the maximum number of vehicles for a higher proportion of time, compared to the system in an M/D/C/C model. Therefore, for high traffic densities, the amount of data downloaded in an M/D/C model is smaller compared to that in an M/D/C/C model, as seen in Figs. 3.19(a) and 3.19(d).

3.6.5 Some Insights from Our Results

Our results have generated some interesting insights with regards to the design and planning of Drive-thru Internet systems. For example, results from Figs. 3.13 and 3.14 suggest that in order to increase the amount of downloaded user data, it would be better to perform all the signaling and connection set-up procedures when the vehicles are located at the "low data rate" distances from the AP, and postpone all bulk data transfers until the vehicles are within the "high data rate" distances from the AP.

In addition, results from Figs. 3.10 and 3.16 shows that the amount of data downloaded by a vehicle is dependent on the traffic conditions (i.e. traffic density) on the road segment covered by the AP, and the percentage of downloading vehicles. Therefore, to enable all the vehicles in the system to jointly share the APs' resources in a fair manner and optimize the amount of data downloaded by each vehicle (based on some pre-defined objectives or service agreements), this observation suggests that it may be beneficial to perform load balancing in the Drive-thru Internet system, in the sense that data for a vehicle can only be downloaded at some specific APs along its route.

3.7 Summary

In this chapter, we have developed analytical models for a vehicle's data download performance in Drive-thru Internet systems. In particular, we have investigated the average and the distribution of the number of bytes downloaded by a vehicle by the end of its sojourn. Our model captures the impact of various parameters like road traffic density, vehicle speed, service penetration rate, AP wireless coverage range and its corresponding transmission rate, on the amount of downloaded data. To the best of our knowledge, our work provides the first analytical model that characterizes the interplay between vehicular traffic parameters and a vehicle's data download performance in a Drive-thru Internet system. Based on our results, system designers and network planners can systematically evaluate the types of communication service and the quality-of-service that Drive-thru Internet systems can support. Such capability will be crucial for the successful deployment of large-scale Drive-thru Internet systems.

Chapter 4

Resource Sharing in Wireless Sensor Networks

4.1 Introduction

In wireless sensor networks, energy management is the most important task due to the fact that most sensor networks are designed to operate unattended for long-term in a remote or hostile environment where battery-operated sensor nodes cannot be easily replaced or have its energy supply replenished. Typically, radio communication between sensor nodes consumes the largest amount of energy, and thus, all network communication protocols designed for sensor networks must be energy-efficient in order to optimize the network lifetime. In particular, an energy-efficient MAC protocol is important since it directly controls the operation of the transceiver and hence, can effectively reduce potential energy waste due to

a) *Collisions* – packet collisions and subsequent retransmissions can cause unnecessary increase in energy consumption
b) *Overhearing* – energy is wasted on reception and processing when a sensor node overhears packets that are destined for other nodes
c) *Idle listening* – sensor nodes waste energy by listening to an idle channel when there are no transmissions
d) *Overhead* – typical overheads in sensor network MAC protocols include synchronization messages, long preambles, and control messages

In this chapter, we will examine resource sharing in wireless sensor networks in terms of the wireless nodes' access to the transmission medium. We propose RMAC (Receiver-Driven Medium Access Control), a TDMA-based and receiver-driven

channel access protocol that utilizes knowledge of a node's timeslots in order to schedule collision-free transmissions. Unlike other TDMA-based MAC schemes, RMAC is designed to avoid the need for sender nodes to broadcast explicit control messages or traffic schedules before data transmissions. In RMAC, we place the ownership of the timeslots in the hands of the receiver nodes and let the receiver nodes assign their timeslots among the neighboring sender nodes. Receiver nodes will only wake-up during their own timeslots to receive any packet transmissions from the assigned neighboring sender nodes, and remain asleep for all other timeslots. Channel utilization is also improved through the provision of mechanisms that enable the "stealing" of unused timeslots. To handle the network traffic load variations, we introduce a simple timeslots reassignment procedure to redistribute the timeslots among the sender nodes according to their offered traffic load. We analytically derive the delay performance of RMAC and show that RMAC outperforms other TDMA-based and sender-driven MAC protocols in terms of the average packet delay and average power consumption metrics.

The rest of this chapter is organized as follows. In the next section, we provide an overview of related work, while Section 4.3 describes the protocol operations of RMAC together with an analysis of its delay performance. We compare the performance of RMAC with another TDMA-based and sender-driven MAC protocol, and evaluate the performance improvement of RMAC with the timeslots stealing and timeslots reassignment mechanisms in Section 4.4. We conclude this chapter in Section 4.5 by summarizing our key findings.

4.2 Related Work

Energy-efficient MAC protocols in sensor networks generally aim to eliminate the energy consumption due to idle listening by making the nodes go to sleep when they don't have any traffic to send or receive. The major challenge is to determine when and how to wake the nodes up whenever there is a packet to send or receive. S-MAC [25] and its variants [26, 86-88] solve this problem by synchronizing the sensor nodes to

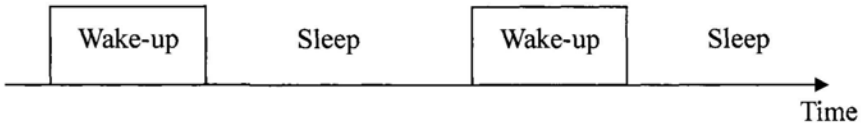


Figure 4.1 S-MAC frame format

follow a duty cycle, consisting of a wake-up period and a sleep period, as shown in Fig. 4.1. Packets are transmitted and received only during the wake-up periods (using the RTS/CTS scheme), and at other times, nodes will turn off their receivers and thus save energy. However, these schemes incur the penalty of increased channel contention during the wake-up periods, since all communication is concentrated into these wake-up periods. In the presence of medium to high traffic load, this will then lead to increased packet collisions which degrade channel utilization and further reduce battery life. Another similar scheme is DMAC [89], designed specifically for data gathering applications using unidirectional trees. Using a staggered wake-up and sleep duty cycle at each level of the tree, DMAC schedules transmissions at each hop so that the end-to-end latency is reduced. However, DMAC only works for a fixed network topology, and doesn't support other types of communication patterns.

TDMA-based MAC schemes have also been proposed due to its inherently collision-free characteristic, whereby nodes schedule collision-free transmissions based on the unique timeslots assigned to them. A TDMA-based protocol has the great advantage of avoiding all energy wastage due to collisions and idle listening since the schedule is used to determine when a node should be active and when it can go to sleep. The challenge is to devise an efficient way through which the sender nodes can inform the receiver nodes to wake-up at the appropriate timeslots to receive their transmissions.

In receiver-driven TDMA-based MAC schemes [28, 90-92], the owners of the timeslots are the receiver nodes. The schedule of timeslots in which a receiver node will wake-up is constructed and exchanged between neighboring nodes [28, 90] or computed at each node by applying some hash functions on their neighboring nodes'

ID [91, 92]. Receiver nodes will need to wake-up during their timeslots to receive any data transmissions from contending sender nodes. However, these schemes suffer from the same problems as S-MAC, i.e. the contention overhead as well as packet collisions among the sender nodes contending for the right to send data to the receiver nodes. On the other hand, in sender-driven TDMA-based MAC schemes [27, 93, 94], the owner of the timeslots are the sender nodes, which needs to send a control message or traffic schedule to inform intended receiver nodes to wake-up and receive data at specified timeslots. In TDMA-W [93], each node is assigned two slots in a frame, a "Transmit" slot for data transmission and a "Wake-up" slot for receiving control packets. Whenever a sender node has data to send to a receiver node, it sends a control packet in the "Wake-up" slot of the intended receiver node to inform it to wake-up in the "Transmit" slot of the sender node in the next frame. If collisions were to occur in the "Wake-up" slot of a receiver node, it will then have to wake up in the "Transmit" slot of all of its one-hop neighbors in order to receive the data from the contending sender nodes. In LMAC [94], each timeslot contains a control portion which is used by the slot owner to inform the intended receiver node to wake-up to receive data during the rest of the timeslot. Therefore, all nodes need to always listen to the control portion of every timeslot of their one-hop neighboring nodes. In TRAMA [27], sender nodes use traffic information to periodically construct and exchange traffic schedules which contains information on the timeslots that they will wake-up and transmit data, and the intended receiver node for each of those timeslots. TRAMA can achieve high channel utilization due to its scheduling of collision-free transmissions. However, packets in TRAMA nodes also have to endure a higher queueing delay, particularly if they arrived after the nodes' schedules have already been announced. TRAMA also has a higher level of complexity compared to other MAC protocols since it has to maintain large amounts of state on the node (e.g. neighbor lists, traffic schedules) and update that state frequently. These sender-driven TDMA-based MAC schemes [27, 93, 94] also suffer from the overhead of control messages broadcasting before any data transmissions can commence.

Z-MAC [29] is similar to previous TDMA-based MAC schemes in that sensor nodes

are assigned timeslots, with the difference that sensor nodes are also allowed to utilize timeslots they do not own through CSMA with prioritized backoff times. Z-MAC achieves high channel utilization at all times by adapting its behavior to the level of contention in the network so that in Low Contention Level (LCL) state, it behaves like CSMA, and in High Contention Level (HCL) state, like TDMA. High network contention level is detected through lost acknowledgments or congestion backoffs, and sender nodes will need to propagate Explicit Contention Notification (ECN) messages to nodes in its two-hop neighborhood to force them to conform to a TDMA schedule. However, propagating the ECN messages places an additional burden on an already busy network, which may result in more packet losses and delay while waiting for nodes to transit to the HCL state. Furthermore, Z-MAC is built on top of B-MAC [95], so that every time a node needs to send data, it must first send a long preamble to alert its receiver of an upcoming data transmission. This increases the energy consumption of senders.

4.3 Receiver-Driven Medium Access Control (RMAC)

4.3.1 Protocol Overview

RMAC is a TDMA-based MAC solution, where all the nodes are time-synchronized. Time synchronization can be achieved among the nodes in RMAC through the application of various proposed schemes [96, 97], and as such, will not be covered in our work. RMAC consists of a neighbor discovery phase, a timeslots allocation phase, and a scheduled data transmission phase. Typically, the scheduled data transmission phase is a lot longer compared to the neighbor discovery phase and the timeslots allocation phase. During neighbor discovery, every node will discover the existence of other nodes within its two-hop neighborhood [98]. In the timeslots allocation phase, whether using the distributed scheduling solution in [99] or using the IDs of itself, its one-hop and two-hop neighbor nodes as the input to multiple hash functions [91],

every node will be able to determine the timeslots allocation in a frame for the nodes within its two-hop neighborhood. The timeslots schedule ensures that every node will be allocated timeslots in a frame that are unique within its two-hop neighborhood. These timeslots are then used to schedule collision-free transmissions.

Unlike other sender-driven TDMA-based MAC schemes, in RMAC, ownership of the timeslots belongs to the receiver nodes. Therefore, receiver nodes will wake-up during their own timeslots to receive any data transmissions, and remain asleep for all other timeslots in order to save energy. Different from other receiver-driven TDMA-based MAC schemes, in RMAC, the receiver nodes will assign their timeslots to their neighboring sender nodes, thereby forming clusters of receiver nodes and sender nodes in the network. By doing so, we avoid any possible contention among sender nodes wanting to transmit data in the same timeslot. To ensure fairness in the transmission opportunities for the sender nodes, the timeslots of all the receiver nodes in a cluster can be equally divided among the sender nodes in the same cluster. The assignment of the timeslots can be achieved using the available information on the IDs of neighboring sender nodes and receiver nodes in the same cluster, and the total number of timeslots in a frame allocated to the receiver nodes in the cluster. The assignment of the timeslots is performed during the timeslots allocation phase.

During the scheduled data transmission phase, the sender nodes will utilize knowledge of their assigned timeslots to schedule their data transmissions. Receiver nodes will only wake-up during their own timeslots to listen for transmissions from their assigned neighbor nodes, and remain asleep for all other timeslots. By utilizing the timeslots in this manner, we eliminate the need for the sender nodes to inform the receiver nodes to wake-up to receive data, as well as eliminate all contention and collision overhead among the sender nodes. As a further optimization, if a sender node does not utilize its assigned timeslot, the receiver node can go back to sleep after a timeout period of no channel activity in order to reduce the idle listening overhead. In general, from a sender node's point of view, it can be assigned timeslots from multiple receiver nodes. This enables the sender node to schedule its data transmissions within the

neighborhood. From a receiver node's point of view, since it controls the assignment of its timeslots, it can control the amount of data relaying tasks that it performs. This enables the receiver nodes to conserve their energy when their power reserves are low.

4.3.2 Timeslots Stealing

In the basic RMAC protocol, when an assigned timeslot is not used, then it is wasted. In situations where there is an asymmetric traffic load or changes in the traffic load at different sender nodes, this will lead to inefficient channel utilization when timeslots assigned to lightly-loaded sender nodes are not fully utilized at the same time that heavily-loaded sender nodes do not have enough timeslots to transmit all their data packets. To increase the channel utilization, we propose a mechanism (called timeslots stealing) to allow an unused timeslot to be used by another sender node.

In RMAC, timeslots are assigned to a sender node, called the primary sender node. To enable timeslot stealing, we assign another sender node (called the secondary sender node) to every timeslot, such that if the primary sender node does not utilize the timeslot, the secondary sender node can steal the timeslot and use it for its data transmission. To do so, the secondary sender node needs to listen to the channel to determine if the primary sender node is transmitting. If after a timeout period it doesn't sense any channel activity, it can then steal the timeslot. Note that the secondary sender node cannot be a hidden node to the primary sender node in order for it to be able to detect any channel activity from the latter node.

Timeslots stealing has the advantage of increasing the channel utilization and reducing the average packet latency. The tradeoff is the increased energy consumption due to the energy expended by the secondary sender node in determining whether it can steal a timeslot. The performance improvement of timeslots stealing is also dependent on the pairing of the primary and secondary sender nodes to a timeslot.

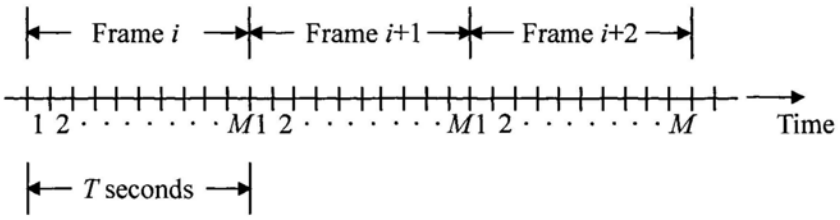


Figure 4.2 TDMA frame structure

4.3.3 Delay Analysis

4.3.3.1 Basic RMAC (without timeslots stealing)

In the basic RMAC protocol (without timeslots stealing), the timeslots of all the receiver nodes in a cluster are equally and uniformly assigned to the sender nodes in the same cluster. This means that each sender node will be assigned one timeslot per frame. This is also referred to as synchronous time division multiplexing. Fig. 4.2 shows the TDMA frame structure, where the frame duration is T sec, and the number of timeslots in a frame is M.

The delay performance of such a system has been analyzed in [100] where the packet arrivals to a sender node i is assumed to be Poisson, with rate λ_i packets/sec. The transmission duration of a packet is equal to the length of a timeslot, T/M. The service time of a packet is defined in [100] to include the time during which the packet is at the head of the queue with no transmission in progress, and its transmission time of T/M seconds. The average packet delay $E[D_i]$ for a sender node i is given by [101]:

$$E[D_i] = \frac{\overline{x_e}}{1 - \lambda_i(\overline{x_b} - \overline{x_e})} + \frac{\lambda_i(\overline{x_e^2} - \overline{x_b^2})}{2[1 - \lambda_i(\overline{x_b} - \overline{x_e})]} + \frac{\lambda_i \overline{x_b^2}}{2(1 - \lambda_i \overline{x_b})} \qquad (4.1)$$

where

$\overline{x_b}$ & $\overline{x_b^2}$ = first and second moments of the service time (X_b) of a packet that arrives at the head of the queue during a busy period, and

$\overline{x_e}$ & $\overline{x_e^2}$ = first and second moments of the service time (X_e) of a packet that initiates a busy period

From [100],

$$\overline{x_b} = T \tag{4.2}$$

$$\overline{x_b^2} = T^2 \tag{4.3}$$

$$\overline{x_e} = \bar{Y} + T/M \tag{4.4}$$

$$\overline{x_e^2} = \overline{Y^2} + 2(T/M)\,\bar{Y} + (T/M)^2 \tag{4.5}$$

$$\bar{Y} = \left(T - \frac{T}{M} - \frac{1}{\lambda_i}\right) + \frac{T exp\left[-\lambda_i\left(T - \frac{T}{M}\right)\right]}{1 - exp(-\lambda_i T)} \tag{4.6}$$

$$\overline{Y^2} = \left(T - \frac{T}{M} - \frac{1}{\lambda_i}\right)^2 + \left(\frac{1}{\lambda_i}\right)^2 + \left(T^2 - 2\frac{T}{\lambda_i}\right)\frac{exp\left[-\lambda_i\left(T - \frac{T}{M}\right)\right]}{1 - exp(-\lambda_i T)} \tag{4.7}$$

By substituting Eqns. (4.2-4.7) into Eqn. (4.1), we obtain

$$E[D_i] = \frac{T}{2} + \frac{T}{M} + \frac{\lambda_i T^2}{2(1 - \lambda_i T)} \tag{4.8}$$

If there are N sender nodes, with λ_i being the packet arrival rate to each sender node i, $i = 1, 2, \ldots, N$, then the overall average delay performance in RMAC is given by:

$$E[D] = \sum_{i=1}^{N} \frac{\lambda_i}{\sum_{j=1}^{N} \lambda_j} E[D_i] \tag{4.9}$$

We have compared the overall average delay performance in RMAC given by Eqn. (4.9) with simulation results. In our evaluation, we have a cluster of L receiver nodes and N sender nodes. Each receiver node is allocated k-slots per frame, therefore each frame has $L*k$ timeslots which are equally and uniformly assigned to the N sender nodes. We consider two Poisson-distributed traffic arrival patterns at the sender nodes, symmetric and asymmetric traffic arrival patterns. For the former, each sender node has an arrival rate of λ/N, with λ denoting the aggregate packet arrival rate to the system. For the latter, the sender nodes have been classified into low-rate, medium-rate, and high-rate nodes, and their arrival rates has a distribution ratio of 1:4:16. The parameters used are summarized in Table 4.1. Fig. 4.3 shows that the simulation results accurately match the average delay performance computed from Eqn. 4.9.

Table 4.1 Evaluation parameters

Parameters	Values
N	10
L	5
k	4 timeslots per frame
λ	1-20 packets/sec
Timeslot size	50ms
Asymmetric traffic load	4 lo-rate, 4 med-rate, 2 hi-rate nodes

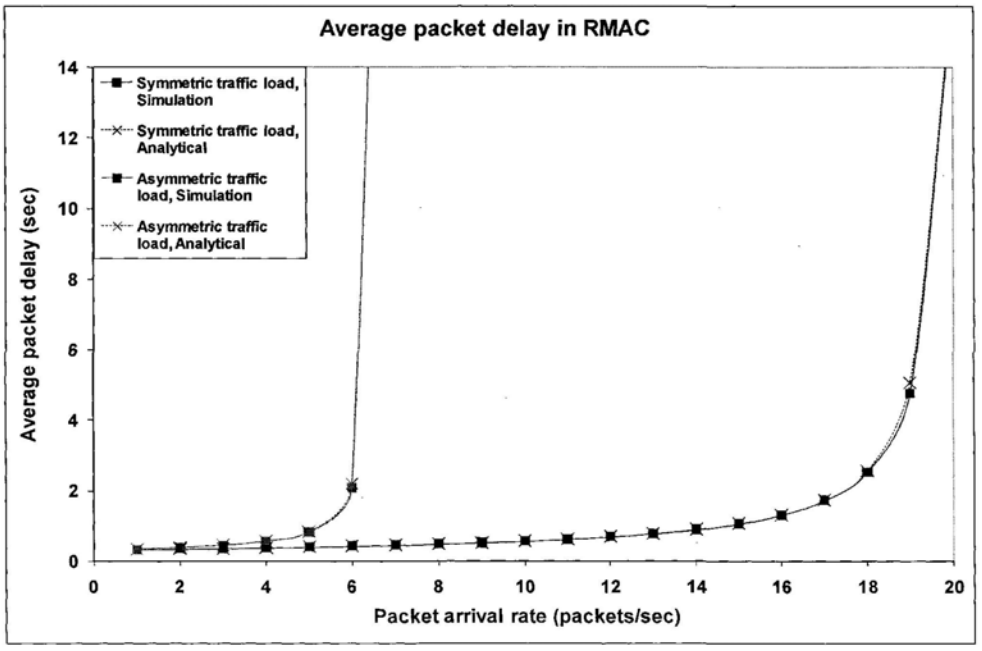


Figure 4.3 Average delay performance in RMAC without timeslots stealing

4.3.3.2 RMAC with timeslots stealing

In the basic RMAC protocol (without timeslots stealing), each sender node is assigned one timeslot per frame. In RMAC with timeslots stealing, every timeslot is assigned to a primary sender node and a secondary sender node. If the primary sender node does not utilize the timeslot, the secondary sender node can steal the timeslot. As a result, on average, the number of timeslots in a frame that a sender node can use to transmit its

packets can be more than one. To compute the average number of timeslots that a sender node has in a frame, we define the following terms:

T = frame duration

M = number of timeslots in a frame

N = number of sender nodes

λ_i = packet arrival rate at sender node i, $i = 1, 2, \ldots, N$

k_i = number of timeslots in a frame that is assigned to sender node i, in which sender node i is the primary sender node

ρ_i = utilization of sender node i

The utilization of sender node i is given by the following relationship:

$$\rho_i = \frac{\text{number of packet arrivals in a frame}}{\text{number of packets served in a frame}} = \frac{\lambda_i T}{K_i} \tag{4.10}$$

where K_i is the average number of timeslots that sender node i has in a frame. K_i can be computed as follows:

$$K_i = k_i + \sum_{j=1; j \neq i}^{N} (1 - \rho_j) k_{j;i} \tag{4.11}$$

where $k_{j;i}$ is the number of timeslots in which sender node j is the primary sender node, and sender node i is the secondary sender node. Note that Eqns. (4.10) and (4.11) can lead to a recursive relationship in the computation of K_i and ρ_i if we have a timeslots assignment in which there exist timeslots for which sender nodes i and j are the primary and secondary nodes, and timeslots for which the roles are reversed for sender nodes i and j. In this case, we can compute K_i and ρ_i using Algorithm 4.1, which uses k_i as the initial value for K_i in its computation. In our evaluation, the computation in Algorithm 4.1 converges to the correct value for ρ_i, given k_i as the initial value for K_i. As part of our future work, we plan to prove the convergence property of Algorithm 4.1 for any arbitrary initial value of K_i.

We can now proceed to analyze the delay performance of a sender node i. The average packet delay $E[D_i]$ for a sender node i is again given by Eqn. (4.1), with $\overline{x_b}$, $\overline{x_b^2}$, $\overline{x_e}$ and $\overline{x_e^2}$ having the same definition as in Section 4.3.3.1. For computational simplification,

we assume that the packet arrival process at the sender nodes is Bernoulli, rather than Poisson, and that arrivals may occur only at timeslot boundary instants. A similar approximation has been used in [102]. In our evaluation at the end of this section, we will show that the error introduced by this approximation is very limited. Fig. 4.4 shows an example of the timeslots in a frame for which a sender node i can possibly transmit its packets. The black-colored timeslots represent the timeslots for which sender node i is the primary sender node, while the grey-colored timeslots represent the timeslots for which sender node i is the secondary sender node. The total number of timeslots in a frame for which sender node i can possibly transmit its packets is (l+1), with the indexes 0, 1, ..., l, denoting the timeslots. We use $p_0, p_1, \ldots, p_l$ to represent the probabilities that sender node i can successfully transmit its packets in those respective timeslots, and $d_0, d_1, \ldots, d_l$ to represent the distances between those successive timeslots. Note that $\sum_{i=0}^{l} d_i = M$. Since sender node i is the primary sender node for timeslot 0, $p_0 = 1$. As for p_u (u=1, 2, ..., l), it is equal to (1- ρ_j), where sender node j (j=1, 2, ..., N; $j \neq i$) is the primary sender node for timeslot u, and ρ_j is the utilization of sender node j, computed using Algorithm 4.1.

Algorithm 4.1 Computation of K_i and ρ_i

Input: λ_i, T, k_i, ϵ where ϵ is the desired precision for the computation of ρ_i

1: **for** i = 1 to N **do**
2: $prev\_\rho_i = \lambda_i T / k_i$; // Initialization
3: **while** (TRUE)
4: **for** i = 1 to N **do**
5: Evaluate Eqn. (4.11) using $prev\_\rho_i$;
6: Evaluate Eqn. (4.10);
7: **if** $\bigcap_{i=1}^{N}[(\rho_i - prev\_\rho_i) < \epsilon]$
8: **return** the latest values for K_i and ρ_i;
9: **for** i = 1 to N **do**
10: $prev\_\rho_i = \rho_i$;
11: **end**

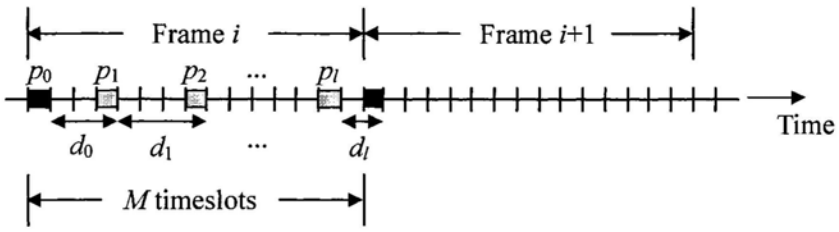


Figure 4.4 Timeslots assignment for a sender node i

To compute the average packet delay for a sender node i as given in Eqn. (4.1), we will need to evaluate the four parameters $\overline{x_b}$, $\overline{x_b^2}$, $\overline{x_e}$ and $\overline{x_e^2}$. To compute $\overline{x_b}$ and $\overline{x_b^2}$ in our model, we condition on the event that the last transmission occurred at timeslot u, i.e. a packet arrives at the head of the queue at timeslot u, $u = 0, 1, \ldots, l$. For example, in Fig. 4.4, if the last transmission occurred at timeslot 0, then with probability p_1, the packet at the head of the queue can be transmitted at timeslot 1, i.e. $X_b = d_0$. Similarly, with probability $(1\text{-}p_1)p_2$, the packet at the head of the queue can be transmitted at timeslot 2, i.e. $X_b = (d_0+d_1)$. By generalizing this argument to all possible timeslots in which sender node i can transmit its packets, we have

$$\begin{aligned}
\overline{x_b} &= E[X_b] \\
&= \sum_{u=0}^{l} E[X_b|\text{last transmission at timeslot } u]\, P[\text{last transmission at timeslot } u] \\
&= \sum_{u=0}^{l} E[X_b|\text{last transmission at timeslot } u] \frac{p_u}{\sum_{w=0}^{l} p_w} \\
&= \sum_{u=0}^{l} \sum_{v=0}^{l-u} m f_b(m) \frac{p_u}{\sum_{w=0}^{l} p_w}
\end{aligned} \tag{4.12}$$

where

$$m = \sum_{w=0}^{v} d_{(u+w)} \tag{4.13}$$

$$f_b(m) = \left[\prod_{w=1}^{v} 1 - p_{(u+w)}\right] p_{(u+v+1)\bmod(l+1)} \tag{4.14}$$

Note that when v=0 in Eqn. (4.14), we use the convention $\prod_a^b(\ldots) = 1$ if $a>b$. To evaluate $\overline{x_b^2}$, we replace m with m^2 in Eqn. (4.12).

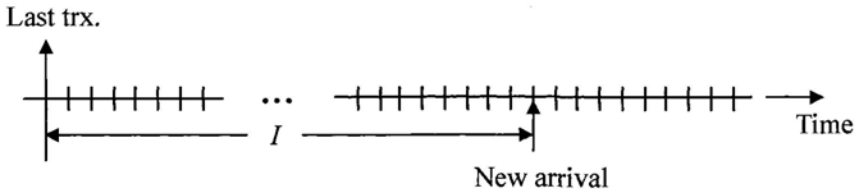


Figure 4.5 Illustration of idle period I

To compute $\overline{x_e}$ and $\overline{x_e^2}$, we again condition on the event that the last transmission occurred at timeslot u, $u = 0, 1, \ldots, l$. However, in this case, after this last transmission, there is an idle period of I timeslots, before a new packet arrival that initiates a new busy period as shown in Fig. 4.5. The slot address of the new arrival depends on the duration of the preceding idle period. The probability of a new arrival at a particular timeslot in a frame, Q_r, is then

$$\begin{aligned} Q_r &= \mathrm{P}[I = r \text{ slots}] + \mathrm{P}[I = M + r \text{ slots}] + \mathrm{P}[I = 2M + r \text{ slots}] + \ldots \\ &= (1-\lambda_i)^r\lambda_i + (1-\lambda_i)^{M+r}\lambda_i + (1-\lambda_i)^{2M+r}\lambda_i + \cdots \\ &= \frac{\lambda_i(1-\lambda_i)^r}{1-(1-\lambda_i)^M} \end{aligned} \tag{4.15}$$

where $r = 0, 1, \ldots, (M\text{-}1)$, and λ_i is the Bernoulli packet arrival rate to sender node i.

Going back to the example in Fig. 4.4, if the last transmission occurred at timeslot 0, and the new packet arrival occurred within the interval d_0, then for $0 \le r \le d_0 - 1$, the service time of the new packet, $X_e = (d_0 - r)$ with probability $Q_r p_1$. Similarly, with probability $Q_r(1\text{-}p_1)\,p_2$, the service time of the new packet $X_e = (d_0 + d_1 - r)$. By generalizing this argument to all possible timeslots at which the last transmission can occur and all possible intervals in which the new packet arrival can occur, we have

$$\begin{aligned} \overline{x_e} &= E[X_e] \\ &= \sum_{u=0}^{l} E[X_e|\text{last transmission at timeslot } u]\, P[\text{last transmission at timeslot } u] \\ &= \sum_{u=0}^{l} E[X_e|\text{last transmission at timeslot } u] \frac{p_u}{\sum_{w=0}^{l} p_w} \end{aligned}$$

$$= \sum_{u=0}^{l} \sum_{v=0}^{l} \sum_{r=idle_{lo}}^{idle_{hi}} \sum_{s=v}^{l} m f_e(m) \frac{p_u}{\sum_{w=0}^{l} p_w} \tag{4.16}$$

where

v = index for the interval in which the new packet arrival occurred

r = number of idle timeslots

$$idle_{lo} = \begin{cases} \sum_{w=u}^{v+l} d_{w \bmod (l+1)} & \text{if } u > v \\ \sum_{w=u}^{v-1} d_w & \text{if } u \leq v \end{cases} \tag{4.17}$$

$$idle_{hi} = idle_{lo} + d_v - 1 \tag{4.18}$$

$$m = idle_{lo} - r + \sum_{w=v}^{s} d_w \tag{4.19}$$

$$f_e(m) = Q_r \left[\prod_{w=v+1}^{s} (1 - p_w) \right] p_{(s+1) \bmod (l+1)} \tag{4.20}$$

$\overline{x_e^2}$ can be evaluated by replacing m with m^2 in Eqn. (4.16).

The computed values for $\overline{x_b}$, $\overline{x_b^2}$, $\overline{x_e}$ and $\overline{x_e^2}$ are then substituted into Eqn. (4.1) to obtain the average packet delay for sender node *i*. We can then use Eqn. (4.9) to evaluate the overall average delay performance in RMAC with timeslots stealing. We again verify our analytical model with simulations. We utilize the same parameters that were used previously in Section 4.3.3.1, as shown in Table 4.1. For our current evaluation, we only consider asymmetric traffic load at the sender nodes, with four low-rate sender nodes (nodes 0-3), four medium-rate nodes (nodes 4-7) and two high-rate sender nodes (nodes 8-9). The packet arrival rates at these sender nodes follow a distribution ratio of 1:4:16. For timeslots assignment, we use the assignment shown in Table 4.2. We also used both Bernoulli and Poisson packet arrivals in our simulations. Results in Fig. 4.6 show the accuracy of our analytical model, where the simulation plot with Bernoulli arrivals matches our analytical plot, while the difference between the simulation plot with Poisson arrivals and our analytical plot is

less than 15%. For comparison purposes, Fig. 4.6 also shows the average delay performance of the basic RMAC protocol without timeslots stealing. We see that RMAC with timeslots stealing improves the system capacity by more than 200%, from 6 packets/sec to 19 packets/sec.

Table 4.2 Timeslots assignment

Timeslots	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Primary sender	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
Secondary sender	8	9	8	9	8	9	8	9	4	5	8	9	8	9	8	9	8	9	6	7

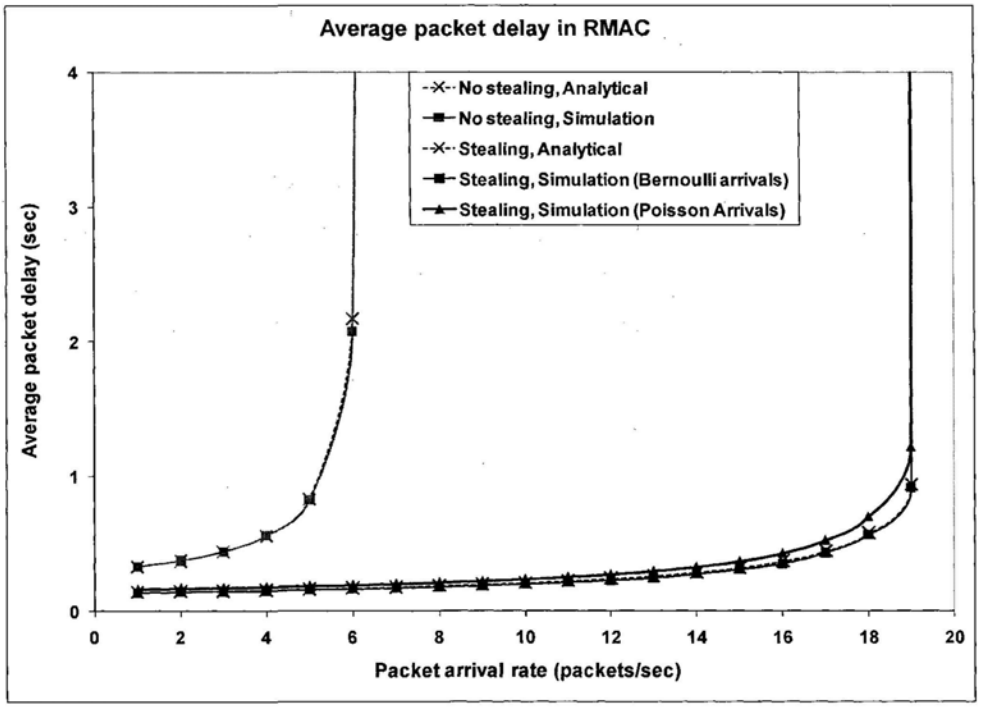


Figure 4.6 Average delay performance in RMAC with timeslots stealing

We note that the results in [100] as applied previously in the case of RMAC without timeslots stealing, cannot be applied in the case of RMAC with timeslots stealing because the analysis in [100] is for a model where every sender node has only one timeslot per frame to transmit its packets. On the other hand, the model in [102] allows a sender node to schedule its transmissions in any one of the M timeslots in a frame. The objective in [102] is to allow the system to switch between TDMA and CSMA according to the level of contention in the system. In [102], a sender node schedules transmission in its own timeslot with probability a, and in other timeslots with probability b. These probabilities are adjusted by the function $a + b(M - 1) = 1$. However, the analysis in [102] is also not suitable for RMAC with timeslots stealing due to the fact that there is still a probability of packet collision even when a sender node transmits in its own timeslot (for values of $a < 1$). Furthermore in our model, the probabilities $p_0, p_1, \ldots, p_I$ as described earlier in Fig. 4.4, are different in values and do not necessarily sum up to 1.

4.3.4 Timeslots Reassignment

An optimal timeslots assignment scheme should take into account the offered traffic load at the sender nodes. In the basic RMAC protocol, by default, the receiver nodes' timeslots in a frame are equally and uniformly assigned to the sender nodes. However, this results in sub-optimal performance when the offered traffic load is asymmetric across the different sender nodes. Timeslots stealing helps to improve the performance, but as mentioned at the end of Section 4.3.2, the performance improvement is dependent on the pairing of the primary and secondary sender nodes to the timeslots.

As a further enhancement to RMAC, we propose a simple timeslots reassignment scheme that allows the receiver nodes to redistribute the timeslots among the sender nodes according to their offered traffic load. In this scheme, the timeslots reassignment is performed after every n frames of data transmission. For every n frames of data transmission, each sender node keeps track of the number of packet arrivals. During the n^{th} frame, each sender node will piggyback this information onto their normal data

packet (if any) or generate a control packet to send this information to the nodes in the cluster. Every node will need to be awake during this n^{th} frame in order to receive this offered load information.

Based on the received offered load information of every sender node, we can calculate the packet arrival rate for each sender node as follows:

$$\text{Packet arrival rate for sender node } i,\ \lambda_i = C_i/\Delta t \tag{4.21}$$

where C_i is the number of packet arrivals in sender node i in Δt, and Δt is the time duration of the n frames of data transmission. The ratio of sender node i's packet arrival rate is given by:

$$R_i = \lambda_i/\lambda_{total} \tag{4.22}$$

where $\lambda_{total} = \lambda_1 + \lambda_2 + \ldots + \lambda_N$, N is the number of sender nodes. Letting M be the total number of timeslots in a frame, then the number of timeslots in a frame to be assigned to sender node i for the next n frames of data transmission is:

$$k_i = R_i M \tag{4.23}$$

where k_i is also subject to the following condition:

$$k_i >= 1, \text{ for all } i \tag{4.24}$$

This simple algorithm is run in every node in the cluster, and the new timeslots assignment will be effective for the next n frames.

4.4 Performance Evaluation & Discussion

4.4.1 Sender-MAC

For comparison purposes, we have implemented a simple model of a TDMA-based and sender-driven MAC scheme similar to TRAMA [27], called Sender-MAC. In our implementation of Sender-MAC, each sender node owns k-slots per frame. The first timeslot in every frame for each sender node is used to broadcast the traffic information, i.e. the sender node checks its buffer for queued packets and then announces the intended receivers for those packets for its next (k-1) timeslots. Every neighboring node of the sender nodes will need to be awake for the broadcast of this

traffic control message. To minimize the energy consumption due to this traffic broadcast, if a sender node has no packet to transmit in that frame, then it need not broadcast any traffic control message. In this case, all its neighboring nodes can also go back to sleep after a timeout period of no channel activity in order to reduce the idle listening overhead.

We have also implemented timeslots stealing in Sender-MAC. In Sender-MAC, if a sender node does not have any or enough packets queued up in its buffer, then not all of its (k-1) timeslots will be used during the frame. These unused timeslots can be used by other sender nodes. To do so, the other sender nodes can deduce from the received traffic control messages (if any), how many timeslots are unused. Then when broadcasting their own traffic control messages, the sender nodes can announce that they are stealing those unused timeslots and also announce the intended receivers for those stolen timeslots. In Sender-MAC, timeslots stealing is FCFS, i.e. whoever is slated next to broadcast its traffic control message, will have priority in stealing any unused timeslots. Once a timeslot is stolen, other sender nodes cannot steal it anymore. For Sender-MAC, timeslots stealing does not incur any additional energy expenditure because the detection of unused timeslots and the decision to steal them happens during the broadcast of the traffic control messages when all nodes are awake. However, its performance improvement is still dependent upon the order of the sender nodes broadcasting the traffic control messages.

4.4.2 Simulation Model

The simulation model that we used to evaluate the performance of RMAC and Sender-MAC is the same as the one described earlier at the end of Sections 4.3.3.1 and 4.3.3.2. We again have a cluster of L receiver nodes and N sender nodes. For RMAC, each receiver node is allocated k-slots per frame, therefore each frame has $L*k$ timeslots which are equally and uniformly assigned to the N sender nodes. For Sender-MAC, each sender node is allocated k-slots per frame (values of k range from 2 to 10), and therefore each frame has $N*k$ timeslots. We consider two

Poisson-distributed traffic arrival patterns at the sender nodes, symmetric and asymmetric traffic arrival patterns. For the former, each sender node has an arrival rate of λ/N, with λ denoting the aggregate packet arrival rate to the system. For the latter, the sender nodes have been classified into four low-rate, four medium-rate, and two high-rate nodes, and their arrival rates follow a distribution ratio of 1:4:16. As in [27], the average power consumption in transmit, receive, and sleep modes is 24.75mW, 13.5mW, and 15μW respectively. The no channel activity timeout period used by receiver nodes to reduce the idle listening overhead is set to 10% of the timeslot size. The values of the parameters used are summarized in Table 4.1. Two performance metrics are investigated, i.e. the average packet delay and the average power consumption per packet. The latter metric is the amount of power needed to transmit and receive a packet successfully, and serves as an indicator of the scheme's energy efficiency. In the following subsections, we will compare the performance of RMAC against Sender-MAC, with and without timeslots stealing. We will also show the performance of RMAC with the timeslots reassignment procedure.

4.4.3 RMAC vs. Sender-MAC (without timeslots stealing)

We compare the performance of RMAC and Sender-MAC without timeslots stealing (under symmetric and asymmetric traffic loads) in Fig. 4.7. We see that RMAC consistently performs better than Sender-MAC for all values of k in terms of the average packet delay and average power consumption per packet. This is due to RMAC's elimination of the need to periodically broadcast the traffic control message. Fig. 4.7 also shows that the performance of Sender-MAC is dependent on the value of k, with k=10 performing the best in terms of maximizing the system capacity and minimizing the power consumption.

4.4.4 RMAC vs. Sender-MAC (with timeslots stealing)

Due to the fact that in RMAC, timeslots are equally and uniformly assigned among the sender nodes, in the case of symmetric traffic load, timeslots stealing only marginally

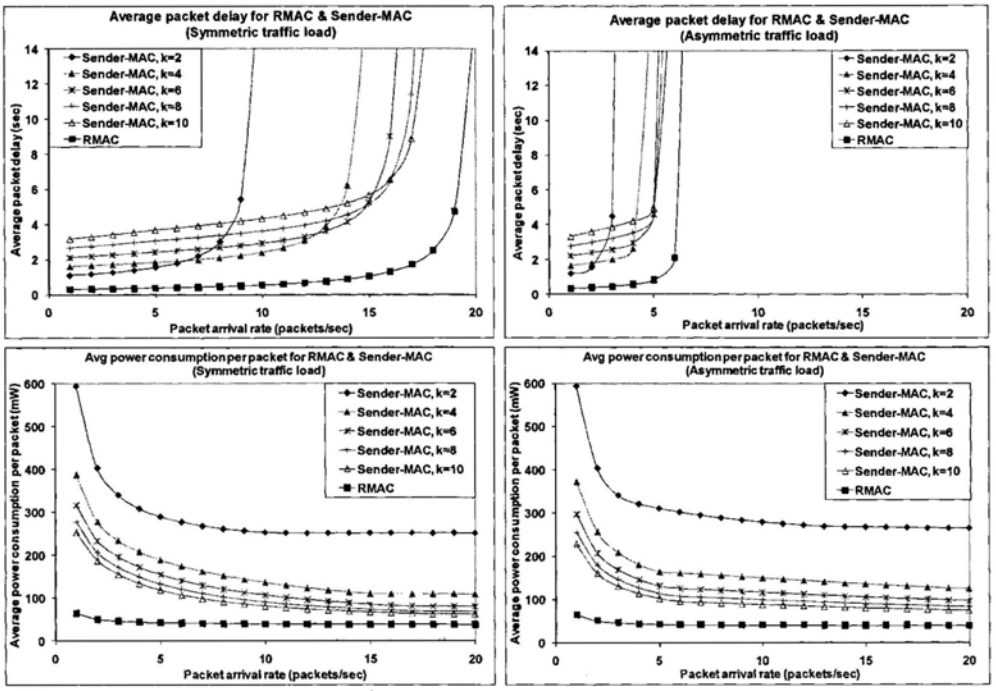


Figure 4.7 Comparing RMAC and Sender-MAC without timeslots stealing

improves the system performance. As such, we will only discuss the performance improvement of timeslots stealing under asymmetric traffic load.

Fig. 4.8 shows the effects of timeslots stealing on the performance of RMAC. We see that with timeslots stealing, the system capacity increases and average packet delay is reduced at the expense of increased power consumption. The increased power consumption is due to the overhead incurred by the secondary sender nodes in stealing the timeslots. For example at λ=19 packets/sec, the average packet delay improves substantially from 210s to 1s, while the average power consumption per packet increases from 39.29mW to 40.17mW. In addition, we see that the system capacity increases by more than 200%, from 6 packets/sec to 19 packets/sec. The results in Fig. 4.8 are obtained with the timeslots assignment shown in Table 4.2, which is the best pairing of primary and secondary sender nodes to the timeslots.

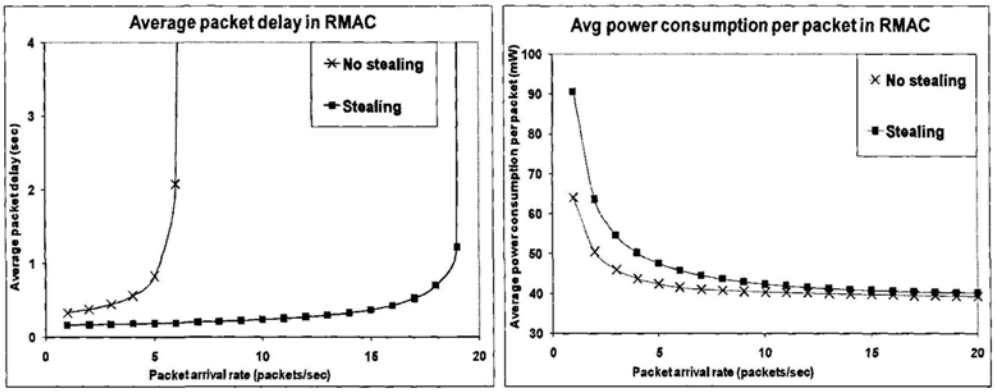


Figure 4.8 Effects of timeslots stealing on the performance of RMAC

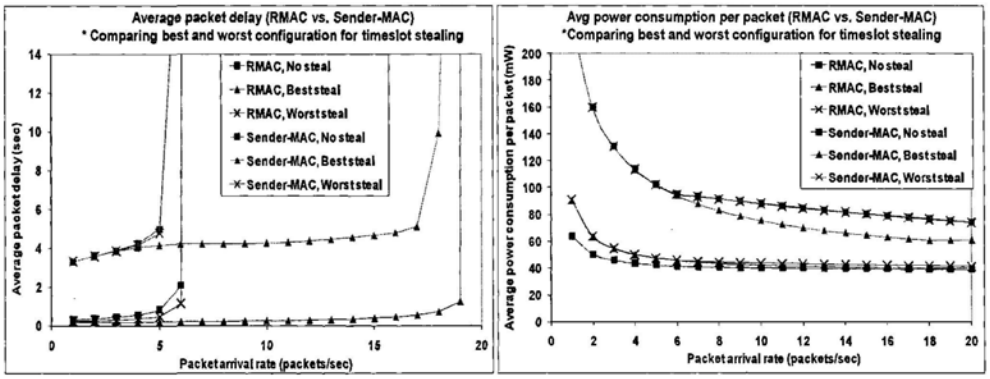


Figure 4.9 Comparing RMAC and Sender-MAC with timeslots stealing

Fig. 4.9 shows a comparison of the effects of timeslots stealing on the performance of RMAC vs. Sender-MAC. We also compare the effects of the best and worst pairings of primary and secondary sender nodes to the timeslots in RMAC, as well as the effects of the order of the sender nodes in the control messages broadcasting in Sender-MAC (with k=10). The worst pairing of primary and secondary sender nodes in RMAC is the case where the low-rate nodes are paired among themselves, the medium-rate nodes are paired among themselves, and the high-rate nodes are paired among themselves. In Sender-MAC, the best order of sender nodes in the control messages broadcasting is the case where low-rate nodes broadcast first and the high-rate nodes broadcast last, while the worst order is the case where the order of the low-rate nodes and the high-rate nodes are reversed. Overall, we see that RMAC with timeslots stealing still

performs better than Sender-MAC with timeslots stealing. We also see that in terms of the average packet delay, the worst timeslots stealing performance ("worst steal") in RMAC and Sender-MAC is similar to the case without timeslots stealing ("no steal"). However, in terms of the average power consumption per packet, "worst steal" in RMAC shows poorer performance compared to "best steal" or "no steal" cases. This is due to the unsuccessful attempts (and therefore, wasting power) by the secondary sender nodes to steal the timeslots.

4.4.5 RMAC with Timeslots Reassignment

In this section, we investigate the performance of RMAC with timeslots reassignment (with and without timeslots stealing). In our evaluation, we used an ON-OFF asymmetric traffic load at the sender nodes, with an exponentially distributed mean of 500ms for both the ON and OFF periods. The frame size is fixed at 60 timeslots, and timeslots reassignments are performed every n=1, 5, and 50 frames. At the beginning of the evaluation, the 60 timeslots are equally and uniformly assigned among the N sender nodes. After each timeslots reassignment, the timeslots are redistributed among the sender nodes according to their traffic load in the preceding n frames of data transmission. In the case of timeslots reassignment with timeslots stealing, at the beginning, each timeslot is randomly assigned to a secondary sender node that is different from the assigned primary sender node. After each timeslots reassignment, with the information on the traffic load of each sender, the estimated high-rate nodes are paired up with the estimated low-rate nodes, and vice-versa. For comparison purposes, we also investigate the performance of RMAC without timeslots reassignment (denoted as "None"). The data for the average packet delay and average power consumption per packet metrics are taken at periodic intervals of 10s.

Fig. 4.10 shows that with λ=12 packets/sec, the performance of n=50 is the best. This is because with low packet arrivals, n=50 allows the system to have a sufficiently long observation interval to estimate the offered traffic load, and therefore, is able to produce a more accurate timeslots assignment result compared to n=1 or 5. We also see

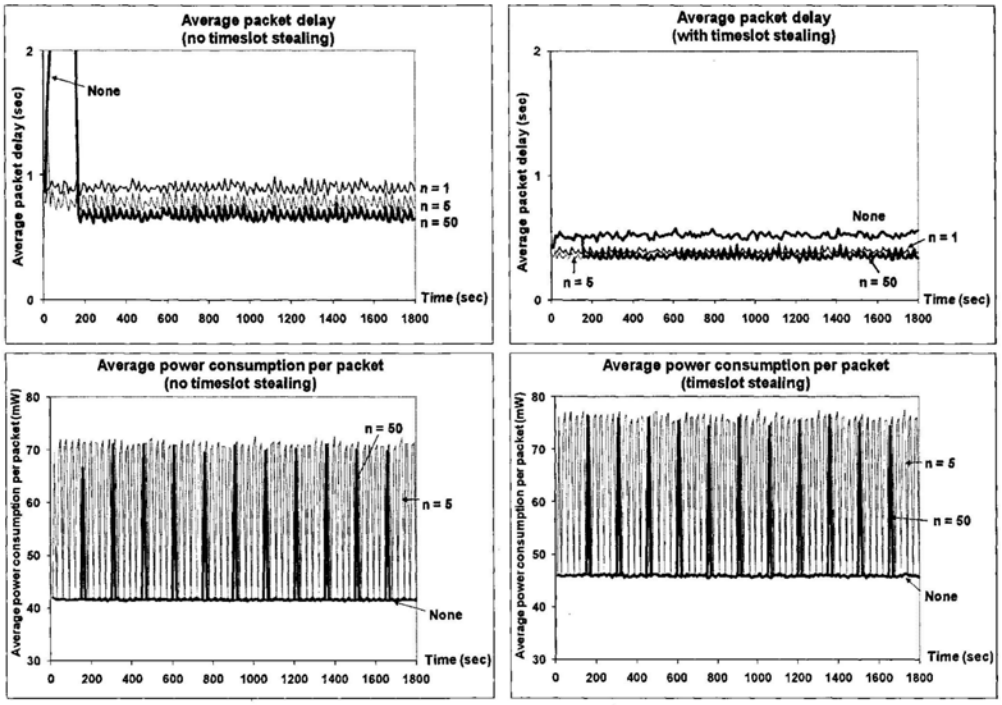


Figure 4.10 Performance of RMAC with timeslots reassignment at λ=12 packets/sec (with and without timeslots stealing)

the improvements in the average packet delay performance of all cases when timeslots stealing is enabled. In terms of the average power consumption per packet, the plot of n=1 is out of the range shown in Fig. 4.10, and we also see that for n=5 and 50, there are periodic spikes in the power consumption due to the fact that every node in the cluster needs to be awake during the n^{th} frame in order to receive the offered load information from each sender node.

Fig. 4.11 shows that at λ=35 packets/sec, with higher packet arrivals, n=1 and n=5 are able to obtain a better estimate of the offered traffic load, and thus produce a better timeslots assignment and achieve lower average packet delay compared to n=50. On the other hand, n=50 requires a long observation interval before it can produce the estimated offered traffic load, and it is not able to handle the short timescale changes in the traffic pattern at high packet arrivals. However, with timeslots stealing, n=50 is

able to reduce the average packet delay. In addition, n=50 performs better than n=1 or n=5 in terms of the average power consumption per packet. Therefore, for energy conservation purposes, it is better to spread consecutive timeslots reassignment phases farther apart and combine its operations with timeslots stealing to reduce the packet delay.

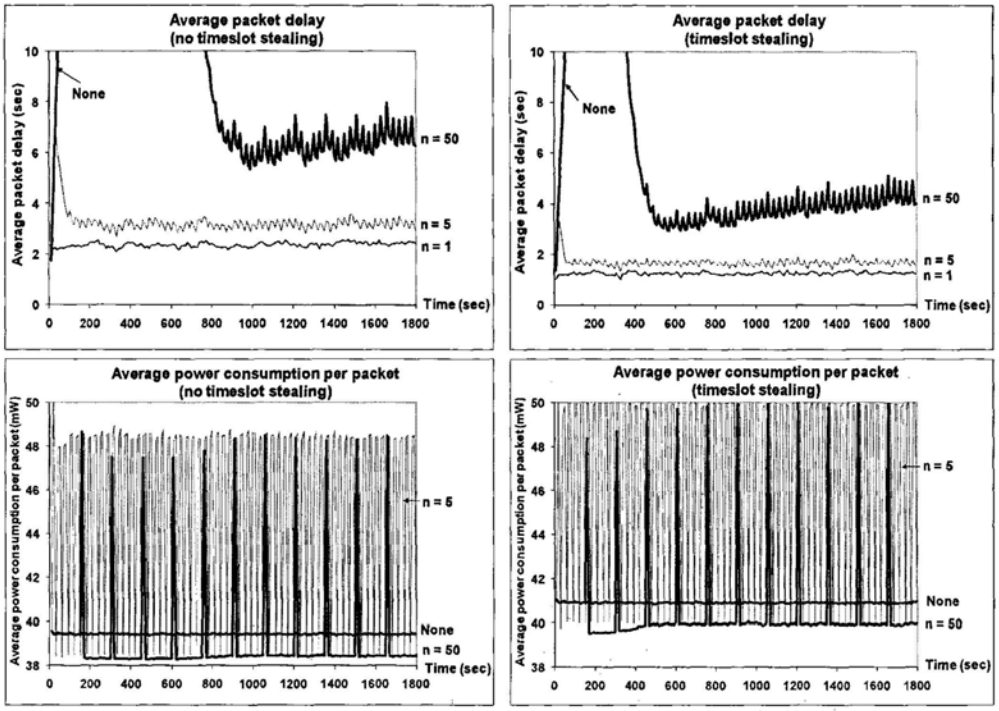


Figure 4.11 Performance of RMAC with timeslots reassignment at λ=35 packets/sec (with and without timeslots stealing)

4.5 Summary

In this chapter, we have examined resource sharing in wireless sensor networks in terms of the wireless nodes' access to the transmission medium. We have proposed an energy-efficient and receiver-driven, TDMA-based MAC protocol for wireless sensor networks. The proposed protocol is named RMAC and it uses knowledge of a node's timeslots to schedule conflict-free data transmissions and avoid explicit control

messages and traffic schedules. In RMAC, ownership of the timeslots belongs to the receiver nodes, which will then assign their timeslots to their neighboring sender nodes. Receiver nodes will only wake-up during their own timeslots to listen for transmissions from their assigned neighbor sender nodes, and remain asleep for all other timeslots. By utilizing the timeslots in this manner, RMAC eliminates the need for the sender nodes to inform the receiver nodes to wake-up to receive data, as well as eliminate all possible contention and collision overhead among multiple sender nodes wanting to transmit to the same receiver node. We have also devised a timeslots stealing mechanism in RMAC that enables unused timeslots to be "stolen" by other sender nodes in order to optimize channel utilization. We have analytically derived the delay performance of RMAC and our results show that RMAC with timeslots stealing substantially improves the system capacity in situations with varying and asymmetric traffic patterns. We have evaluated the performance of RMAC against another TDMA-based, sender-driven MAC protocol, and our results show that RMAC consistently performs better in terms of the average packet delay and average power consumption per packet metrics.

We have also proposed a simple timeslots reassignment scheme to allow the receiver nodes to redistribute the timeslots among the sender nodes according to their offered traffic load. Our results show that in the face of asymmetric, bursty (ON-OFF) traffic patterns, the timeslots reassignment scheme enables the network to efficiently reassign the timeslots among the sender nodes in accordance to the estimated traffic load. The results also show that it is better to perform the timeslots reassignment procedure infrequently. This is so that during low packet arrival rates, the network will be able to obtain a better estimate of the traffic load, and during high packet arrival rates, conserve its energy and rely on the timeslots stealing mechanism to react to the shorter timescale changes in the traffic patterns and reduce the average packet latency.

Chapter 5

Resource Sharing in 3G Networks

5.1 Introduction

UMTS-based third generation cellular networks are now in operation in many countries around the world. Built upon the WCDMA [103] technology, these 3G networks offer a substantial increase in the capacity for data and voice communications compared to the previous 2G/2.5G networks. HSDPA [104] is an upgrade to the Release 99 versions of the 3G networks, offering higher downlink speeds of up to 10Mbps. In this work, via empirical measurements, we examine the throughput and delay performance of IP-based applications over the 3G networks under lightly-loaded and saturated network conditions. We also investigate similar performance in a HSDPA network under lightly-loaded conditions. To the best of our knowledge, our work is the first large scale study of its kind that evaluates the performance of live 3G networks under saturated conditions, using a mixture of data, video and voice traffic. We have conducted field measurements on three commercial 3G networks in Hong Kong, with the measurement sites spread out over multiple different regions with a large variety of characteristics, ranging from rural to urban sites, and commercial to residential sites.

In our measurements, we examined the behavior of the network resources allocation mechanisms and the call admission control policies in the 3G networks. We also investigated the fairness of the radio-link scheduler in allocating the bandwidth resources to multiple data calls in a saturated network. By performing our measurements under saturated network conditions, we have also studied the data throughput, latency, video and voice calls handling capacities of the 3G networks. In addition, we examined the throughput and latency performance of 3G data services in

fully-loaded and lightly-loaded network conditions. We have also performed data throughput and latency measurements on a lightly-loaded HSDPA network, and compared the results with those obtained over the 3G networks. The term 3G used in this chapter refers to the Release 99 version of the UMTS standards, and unless specified otherwise, the results reported in this chapter are those of the 3G networks.

The remainder of this chapter is as follows. We will first describe the measurement methodology that we employed in our tests in the next section. Section 5.3 documents our experimental observations and experiences, while Section 5.4 presents the measurement results from a qualitative and quantitative point of view. We conclude this chapter by summarizing our key findings and results in Section 5.5.

5.2 Measurement Methodology

Our measurements are designed to investigate

a) The data, video and voice capacities of 3G networks.
b) The tradeoffs between data and video/voice traffic in 3G networks.
c) The tradeoffs between video and voice traffic in 3G networks.
d) The end-user experience in lightly-loaded 3G and HSDPA networks.

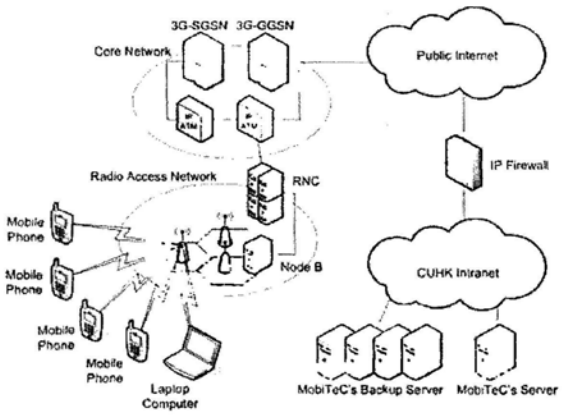


Figure 5.1 Measurement setup

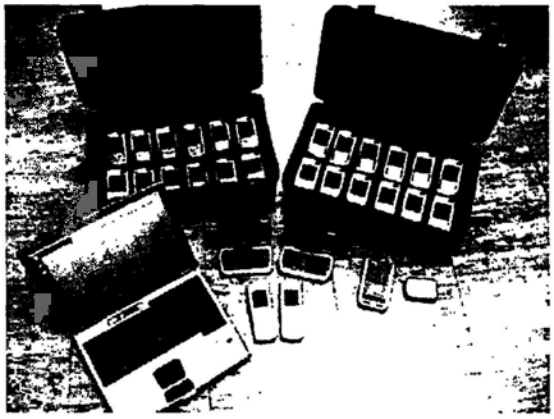


Figure 5.2 Measurement equipment

Fig. 5.1 depicts our measurement setup while Fig. 5.2 shows the equipment used in our measurements. We used a set of client-and-server programs, of which the client programs were developed on the Java Platform, Micro Edition (Java ME), while the server programs were developed on the Java Platform, Standard Edition (Java SE). For clients, we used Nokia 6680 mobile phones, which are Java-enabled, MIDP-2.0 compatible devices. For servers, we used Dell workstations, loaded with Linux kernel 2.4.21, running TCP Reno with TCP SACK enabled. We also used laptop computers equipped with 3G/HSDPA data cards to remotely control the servers from the measurement sites, as well as to conduct the end-user experience measurements. In addition, GPS devices and the NetMonitor [5] program were used during the measurement site selection process.

For the data measurements, we measured the 3G network's data capacity in both the downlink and uplink directions. Multiple TCP connections are setup between the mobile phones and the server in order to transfer bulk data in the downlink/uplink directions. During the data transfer, the current individual and aggregate data throughput values are periodically calculated and displayed by the client and server programs respectively. This is useful to help determine the data saturation point and also to keep track of the aggregate data throughput value during the different stages of

[5] http://www.symbian-freak.com/quides/netmon.htm

the measurements. In addition, we used Ethereal to capture the packet traces at the server. By analyzing the traces with the aid of tcptrace and a set of scripts that we developed, we are able to determine the throughput and delay performance of the data connections over the 3G network.

5.2.1 Measurement Tests

We defined nine measurement tests covering the whole range of services provided by the 3G networks. Table 5.1 shows the measurement tests and their definitions. For the measurement tests involving tradeoffs between data and video/voice, the network is saturated with data calls first before video/voice calls are made. This is to determine the maximum number of video/voice calls that can be made when the network is saturated with data calls. For more efficient data and air-time consumption, measurement tests involving data are performed consecutively. For example, video calls are added into the network (MT2) immediately after data saturation (MT1), and then taken away so that the network is saturated with data calls again (MT1) before voice calls are added into the network (MT3).

Table 5.1 Measurement tests

Measurement Test	Definition
MT1	Downlink – Pure data saturation
MT2	Downlink – Data vs. video
MT3	Downlink – Data vs. voice
MT4	Uplink – Pure data saturation
MT5	Uplink – Data vs. video
MT6	Uplink – Data vs. voice
MT7	Video – Pure video saturation
MT8	Video – Video vs. voice
MT9	Speed test & Ping test

For the MT7 test, video calls are made until no more video calls can be admitted into the network. Voice calls would then be made until the network is fully saturated. The MT8 test is performed immediately after that, where one video call is dropped and voice calls are then made in place of the dropped video call. Through these two tests, we can determine the network's video calls capacity as well as identify the number of voice calls that are equivalent to one video call. In all our tests involving video/voice calls, the calls are made in pairs using the handsets at the measurement site so that we can better utilize our limited handsets to saturate the network.

5.2.2 Measurement Site Selection

Working on information provided by the Office of Telecommunications Authority (OFTA) of Hong Kong, we selected potential measurement sites based on the following criteria:

a) The potential to receive a strong dominant signal from a single base station.
b) Diversity of site demographics and characteristics.
c) Accessibility to the site.
d) Proximity of base stations from different operators.

We also conducted on-site assessments, using the GPS devices and the NetMonitor program, in order to locate the physical location which has a strong dominant signal from a single base station with significantly weaker signals from other neighboring base stations. This is to avoid any complications to our measurement tests due to CDMA soft-handovers where a mobile phone could potentially be simultaneously connected to two or more base stations. During our on-site assessments, we have observed that a mobile phone simultaneously connected to two or more base stations, usually has weak received signal strengths which can cause our measurement calls to be easily disconnected. Furthermore, from the NetMonitor program, we observed that the entries on the base station list (showing the base stations that the mobile phone is connected to) are constantly vanishing and reappearing due to the weak signal strengths. Therefore, to avoid these problems while measuring the cell capacity, our

chosen measurement sites are those sites that have a strong dominant signal from a single base station. In all our measurements, the signal strength (E_c/N_0), base station ID and the GPS location are recorded down before the measurements are started.

5.2.3 Additional Considerations

On average, each measurement set (downlink measurement set, uplink measurement set, and video measurement set) took about 25 minutes to complete, and the three measurement sets were normally carried out consecutively. Due to its length, measurements were usually conducted during off-peak hours in order to minimize the interruption to normal 3G users due to the overloading of the network during our measurements, and more importantly, minimize network capacity measurement errors due to other traffic users on the network. In addition, all our measurements were conducted at fixed locations so as to eliminate the effect of mobility on our results.

We have also performed tests to verify that the cause of the low measured throughput values (< 384kbps per data call) in the 3G networks is not due to any bandwidth bottleneck at the server's network connection or a processing bottleneck at the server itself. In these verification tests, we used the mobile phone emulator program (supplied with the Java ME Wireless Toolkit) running in two laptop computers, to directly connect to the server through a 100Mbps Ethernet LAN. Twenty concurrent TCP connections were setup between the laptop clients and the server. The test results showed that the achievable throughput with the mobile phone emulator can reach up to 4.5Mbps (downlink) and 550kbps (uplink) at each of the twenty concurrent TCP connections, which is higher than the maximum downlink rate of 384kbps offered by the 3G networks. These results proved that the server is capable of serving multiple simultaneous TCP connections at a much higher rate than that seen in the measurements over the 3G networks, and therefore verifies that the server's performance is not the cause of the low throughput values observed in the 3G measurements.

Additional tests have also been conducted to determine if the 3G networks performed any data payload compression on packets passing through the networks. In these tests, we used two types of data payloads, one consisting of all "1"s and another one consisting of some random data sequences. The test results show that the achievable throughput values are approximately the same for both types of data payload. This indicates that data payload compression is not performed on packets passing through the 3G networks.

5.3 Experimental Observations & Experiences

We have conducted measurements at 170 sites over a four-month period, covering the networks of three 3G operators in Hong Kong. For ease of reference, the three operators will henceforth be referred to as Operator X, Operator Y and Operator Z in this chapter. During our measurements, numerous interesting observations were made regarding the behavior of the 3G networks during saturated conditions.

5.3.1 Determination of Saturation Point

In the MT1 and MT4 tests, data calls are added one by one into the network until the network is saturated. It is seen that the aggregate data throughput value will keep on increasing as the data calls are added, until the data saturation point where it would be at a maximum. However, it is difficult to know in advance the optimal number of data calls needed to saturate the network. One observation that helped in determining this saturation point is that when an additional data call is added into a saturated network, the aggregate data throughput value would drop from its peak value. When this extra data call is dropped, it is observed that the aggregate data throughput value will rise back again. Using this "trial-and-error" method, we are able to find the optimal number of data calls to saturate the network. Another observation that helped in determining the data saturation point is that during the MT4 test in the network of Operator Y, repeated attempts to add an additional data call after the data saturation point would fail. Therefore using this observation, if an additional data call cannot be

added into the network, then it means that the data saturation point has already been reached.

5.3.2 Tradeoffs between Data and Video/Voice Calls

During the MT2, MT3, MT5 and MT6 tests, attempts were made to add video/voice calls into the network after the data saturation point. We observed that in almost all of the MT5 tests and in some of the MT2 and MT6 tests in Operator Y's network, these attempts would fail even when more than 30 tries were attempted using more than 10 different pairs of handsets. This implication of video/voice calls not necessarily having higher priority compared to data calls could be due to the admission control policy employed in Operator Y's network, which tries to ensure some form of minimum service for the already admitted data calls (to be discussed further in Section 5.4.1.2).

Nevertheless it is generally seen that in the MT2 and MT3 tests, if a video/voice call attempt is successful, the aggregate data throughput value will drop by about 300-400kbps, and additional video/voice calls after that can be made easily without repeated attempts. On the other hand, it is also observed that during the MT5 and MT6 tests in the networks of Operator X and Operator Z, some of the existing data calls will be dropped by the network when a video/voice call attempt is successful. This is especially so when video calls are added. In view of this, the order of the MT5 and MT6 tests has been reversed; after the saturation with uplink data calls (MT4), the MT6 test is performed first before the MT5 test. This is because the addition of voice calls rarely caused the data calls to be dropped compared to the video calls, and doing the tests in this order enables us to reduce the number of attempts to reconnect the data calls.

5.3.3 Network Stability

After the data saturation point has been reached in the MT1 test, we observed that the throughput values of the data calls in Operator X's network are generally fluctuating, while the throughput values of the data calls in the networks of Operator Y and

Operator Z are relatively more stable. The throughput fluctuations observed in Operator X's network are probably due to the dynamic bandwidth resources sharing among the data calls as the radio-link scheduler tries to fairly allocate resources for each call (to be discussed further in Section 5.4.1.1). As for the MT4 test in Operator Z's network, we have observed multiple uplink data calls being dropped after the data saturation point has been reached. We have also experienced instances where all types of connections are suddenly dropped when the network is saturated. This happened in all of the three operators' networks.

5.3.4 Video and Voice Call Quality in Saturated 3G Networks

In the data-video and data-voice tradeoff tests, we observed that as long as a video/voice call is successfully added into the data-saturated network, the quality of the connected video/voice call is satisfactory. On the other hand, in the video-voice tradeoff test, the successful addition of voice calls can cause some of the video calls to sometimes experience frozen video frames and one-sided conversation, where one party can view the other party of the conversation, but not the other way round. However for our measurements, we do not consider the video/voice call quality, but rather only take into account the number of video/voice calls that can be successfully added into the network, as this parameter is more relevant to our network capacity measurement purposes.

5.4 Measurement Results

We now look into our measurement results from a qualitative and quantitative perspective. We have selected and will analyze below, several graphs that are representative of the measurement results from the 3G operators.

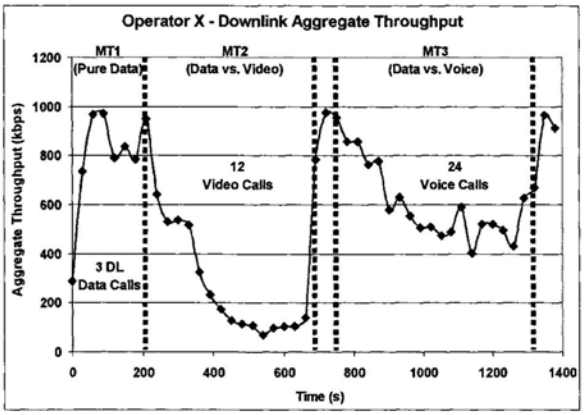
5.4.1 Performance in Saturated 3G Networks

We have generated TCP throughput and round-trip-time (RTT) graphs for all the measurement tests that we have done. The aggregate throughput graphs show the total

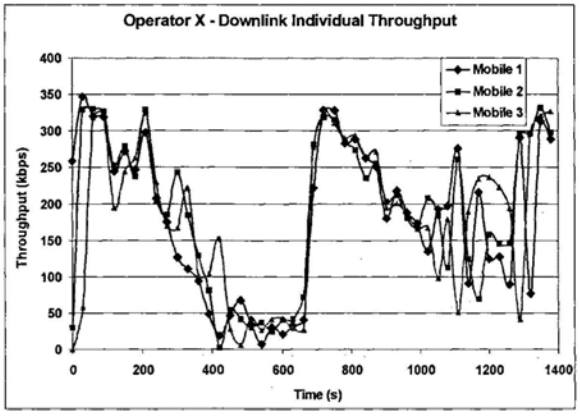
TCP throughput in the cell that is allocated to all the handsets during the measurement test. From these graphs, we can see the effects on the data calls' total throughput levels due to the addition of video and voice calls in the cell. A similar graph is the individual throughput graph which shows the TCP throughput of each handset during the measurement test, as well as provides an indication of the radio-link scheduler's fairness in terms of the bandwidth allocation to the handsets in the cell. The RTT graphs show the RTT values of each handset during the measurement test, and are typically the response times experienced by a data user in a saturated 3G network.

5.4.1.1 Downlink Measurements

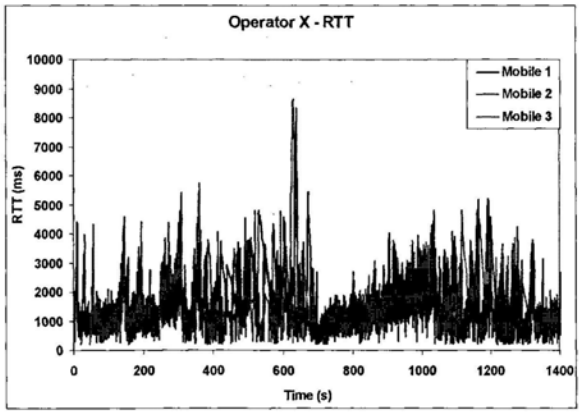
Figs. 5.3-5.5 show the downlink measurement results for the 3G operators. We can see from the TCP downlink aggregate throughput graphs in Fig. 5.3(a), Fig. 5.4(a), and Fig. 5.5(a) that when video or voice calls are added into the data-saturated cell, the TCP downlink aggregate throughput for the data calls are reduced drastically. The TCP throughput reduction is higher when video calls are added compared to voice calls because a video call requires more bandwidth (~64kbps) compared to a voice call (~12kbps). We also observe that more voice calls can be added into the data-saturated cell compared to video calls. In fact, during the MT3 tests, all the handsets have been fully utilized to make the data calls and the voice calls, and as a result, we are not able to investigate if more voice calls could be made. On the other hand, during the MT2 tests, not all the handsets were used up. For example, Operator X's results in Fig. 5.3(a) show that only 12 video calls could be successfully added into the cell, and from our measurement worksheet, we see that attempts to add more video calls failed. As a comparison, Operator X's results from the MT7 test at the same measurement site show that 13 video calls and 3 voice calls were required to fully saturate the cell. Fig. 5.3(a) shows that for the MT2 test, the TCP downlink aggregate throughput value after the addition of the 12 video calls is approximately 100kbps, which is equivalent to the sum of the bandwidth required for 1 video call and 3 voice calls. This indicates that the cell's bandwidth resources have been fully shared between the data and video calls during the MT2 test. Similar observations are seen in the results for Operators Y and Z.



(a)

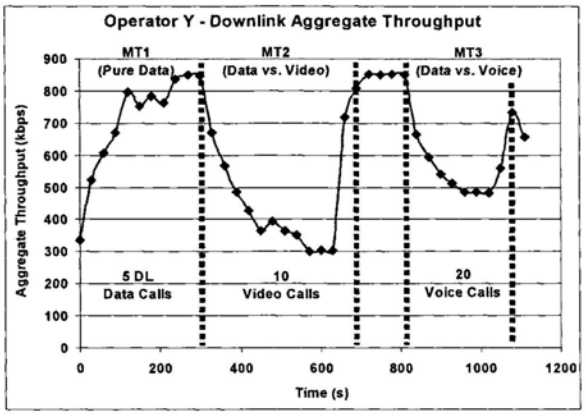


(b)

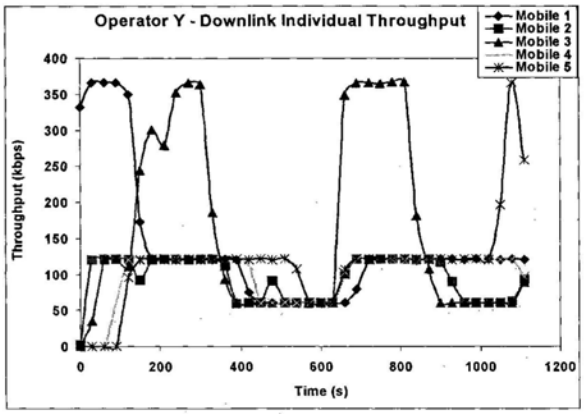


(c)

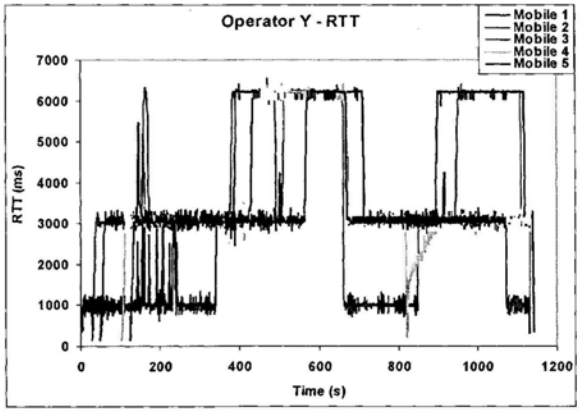
Figure 5.3 Downlink measurement results for Operator X



(a)

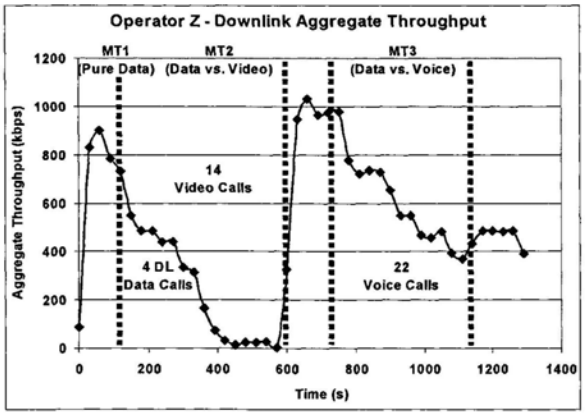


(b)

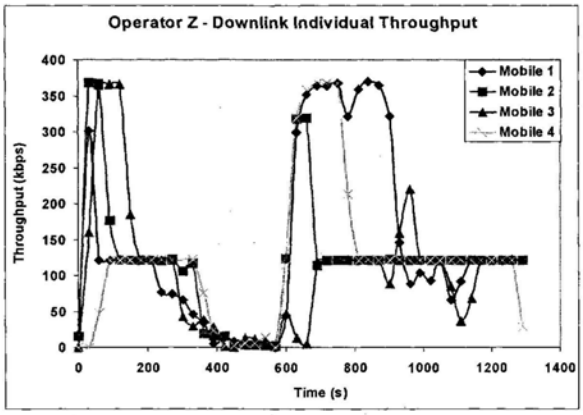


(c)

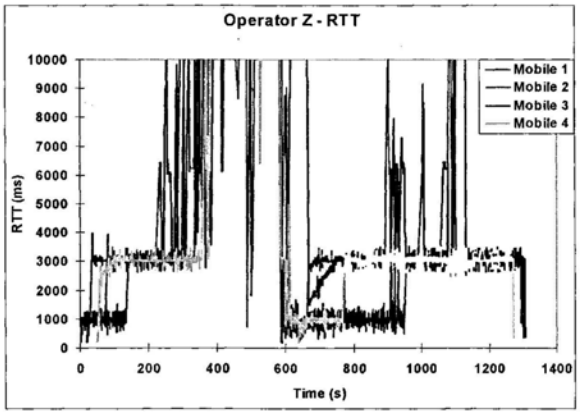
Figure 5.4 Downlink measurement results for Operator Y



(a)



(b)



(c)

Figure 5.5 Downlink measurement results for Operator Z

Observations from the MT2 tests seem to suggest that the radio-link scheduler reserves some minimum bandwidth resources for the data calls. It can be explained that the addition of the initial group of video calls is possible because the bandwidth resource consumption of the data calls at the peak aggregate throughput is much higher than the minimum data bandwidth threshold ($BW_{data\text{-}min}$). As video calls are added into the cell, bandwidth resources are taken away from the data calls (i.e. the radio bearers of the data calls are downgraded to lower bit rates), thereby causing the downlink aggregate throughput value as well as the downlink individual throughput value of the data calls to drop. Once the downlink aggregate throughput value has been reduced to $BW_{data\text{-}min}$, attempts to add in more video calls would be unsuccessful.

The TCP downlink individual throughput graphs in Fig. 5.3(b), Fig. 5.4(b), and Fig. 5.5(b) show that the addition of video/voice calls into the data-saturated cell caused the individual data calls' throughput to drop significantly from a high of 370kbps to as low as 10kbps. We also see that in Fig. 5.4(b), the TCP throughput values in Operator Y's network can be separated into three distinct levels of 370kbps, 125kbps, and 60kbps. Fig. 5.4(b) also shows that for most of the duration of the downlink measurement test, the different data calls in the cell do not have the same TCP throughput values as each other. While similar results are also obtained in Operator Z's network, Fig. 5.3(b) shows that the data calls in Operator X's network generally have the same TCP throughput values as each other. These results suggest that the radio-link schedulers employed in the networks of the three operators implement different bandwidth resources allocation mechanisms, which agrees with our knowledge that the three 3G operators are using different vendors' equipment in their networks (although we don't have access to the algorithms and their parameter values). In terms of fairness in the bandwidth resources allocation, the results suggest that the radio-link schedulers in the networks of Operator Y and Operator Z do not treat each data call in the same manner, but instead allocate different bandwidth resources to different data calls and thereby causing each data call to achieve different TCP throughput values. On the other hand, the radio-link scheduler in Operator X's network is fair in that every data call in the cell is allocated the same bandwidth resources, resulting in all of the data calls

generally achieving the same throughput values as each other. The criteria to determine the amount of bandwidth resources to allocate to the data calls could be based on the individual link quality and the amount of data backlog in the network buffer.

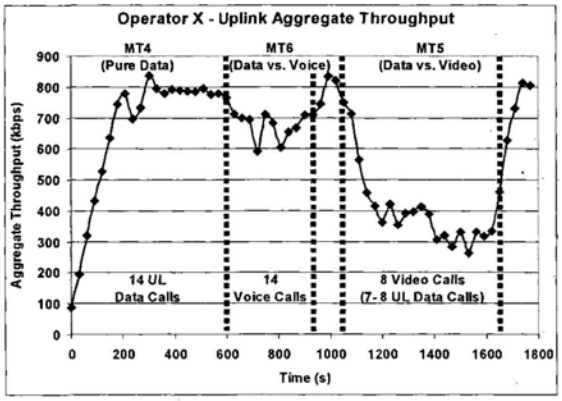
The RTT graphs in Fig. 5.3(c), Fig. 5.4(c), and Fig. 5.5(c) show that the RTT values for each data call are inversely proportional to its TCP throughput values. While Fig. 5.3(c) shows that the RTT values for the data calls in Operator X's network are generally the same as each other, Fig. 5.4(c) shows that the RTT values for the data calls in Operator Y's network can be separated into distinct levels of 1000ms, 3000ms, and 6000ms, inversely corresponding to the different TCP throughput levels observed in Fig. 5.4(b). The RTT results for Operator Z's network are similar to those in Operator Y's network in that the RTT values can be divided into distinct values of 1000ms and 3000ms. The difference is that for Operator Z, the RTT values during the MT2 test are exceptionally high, i.e. above 100s (out of the range shown in Fig. 5.5(c)), corresponding to the low TCP throughput values (~10kbps) seen in Fig. 5.5(b). It is worth noting that even with such high RTT and low TCP throughput values during the MT2 test, the data transfer across the data calls in Operator Z's network was still able to continue.

5.4.1.2 Uplink Measurements

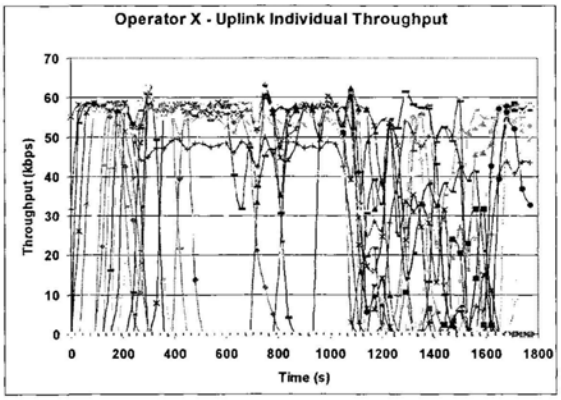
The uplink measurement results are shown in Figs. 5.6-5.8. We can see from Operator X's results in Fig. 5.6(a) that the addition of video calls during the MT5 test caused some of the data calls to be dropped. Even though there were attempts to reconnect the data calls, the number of data calls that could eventually co-exist with the video calls in the MT5 test is smaller than that in the MT4 test. From Fig. 5.6(a), we see that after 8 video calls were added into the cell, the total number of data calls dropped from 14 to about 7-8 data calls. In addition, attempts to add more video calls in the cell also failed.

Similar observations are also seen in Operator Z's results in Fig. 5.8(a). Just like in the downlink measurement results, these observations also suggest the existence of $BW_{data\text{-}min}$. Using a similar explanation as in the downlink case, the addition of the

initial group of video calls is possible because the bandwidth resource consumption of the data calls at the peak aggregate throughput is much higher than $BW_{data\text{-}min}$. As the video calls are added into the cell, bandwidth resources are taken away from the uplink data calls. Unlike downlink data calls which are allocated higher radio bearer rates, uplink data calls are generally allocated the minimum radio bearer rate of 64kbps. Hence when bandwidth is taken away, some of the uplink data calls would be dropped. The observations also suggest that more than the required number of uplink data calls were dropped, thereby causing the uplink aggregate throughput value to be lower than $BW_{data\text{-}min}$. When some of the data calls were reconnected, the uplink aggregate throughput value rose back to $BW_{data\text{-}min}$. Once this happened, attempts to add in more video calls would fail.

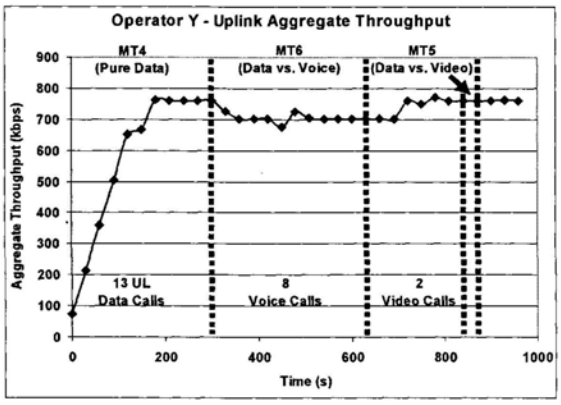


(a)

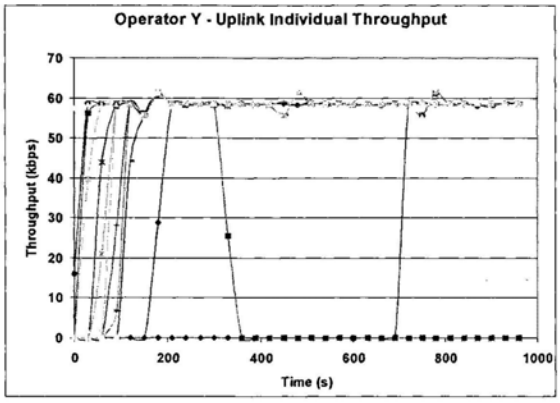


(b)

Figure 5.6 Uplink measurement results for Operator X



(a)

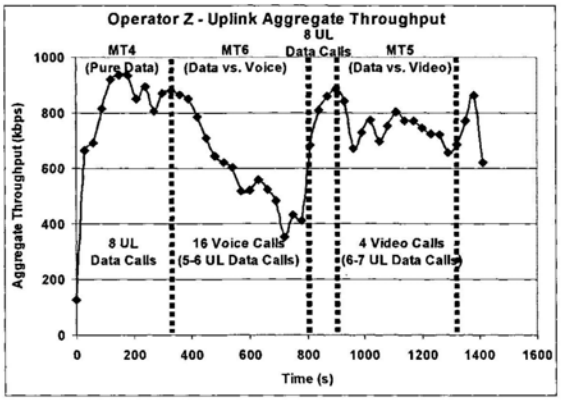


(b)

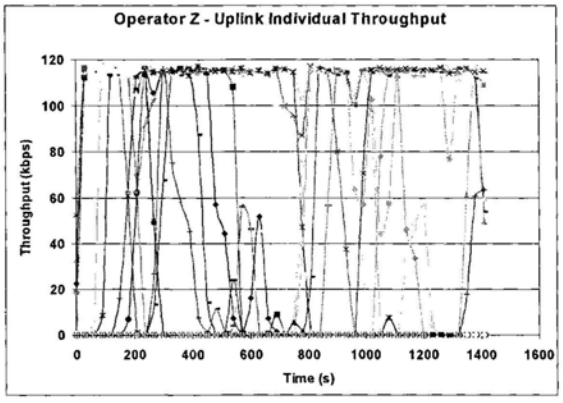
Figure 5.7 Uplink measurement results for Operator Y

Conversely, our uplink measurement results for Operator Y's network (Fig. 5.7(a)) generally show that the additions of video/voice calls would not cause the dropping of the existing data calls. In addition, from our measurement worksheet, we see that no more data calls could be added after uplink data saturation, and attempts to add more voice calls and video calls during the MT6 and MT5 tests were also unsuccessful after the initial successful addition of 8 voice calls and 2 video calls respectively. These observations suggest that in Operator Y's network, once a data call is admitted into the cell, it would not be pre-empted by new video calls. Coupled with the fact that no more data calls could be added, it implies that there is an upper limit on the number of data calls that can be admitted into the cell. It also suggests that the call admission control

policy in Operator Y's network operates in such a way as to guarantee some form of minimum service for all admitted calls. The initial group of voice/video calls could be added into the cell probably because after the upper limit of the data calls has been reached, there is still some minimal amount of bandwidth resources left to admit the voice/video calls. However, once these remaining bandwidth resources are used up, no more voice/video calls could be added.



(a)



(b)

Figure 5.8 Uplink measurement results for Operator Z

Fig. 5.6(b) shows that the maximum TCP uplink individual throughput value for the data calls in Operator X's network is about 60kbps. While similar results are obtained for the uplink measurements in Operator Y's network, we found that the data calls in Operator Z's network can achieve a throughput value up to 115kbps (Fig. 5.8(b)). Since the same handsets were used in all our tests, this throughput difference points to a different configuration setting in Operator Z's network.

5.4.2 Performance in Lightly-loaded 3G Networks

The MT9 test is designed to investigate the 3G user experience, using a laptop equipped with a 3G data card in a lightly-loaded cell scenario, as compared to our other measurement tests which were conducted with multiple handsets in a fully-loaded cell scenario. Table 5.2 shows the speed test results[6] for the three operators. As expected, the link speed to the local website is higher than that to the overseas website for all three operators. The average link speeds of the three operators across all the measurement sites are more or less the same while their maximum link speeds to the local website are still within the same order of magnitude compared to the maximum downlink throughput values seen in Fig. 5.3(b), Fig. 5.4(b), and Fig. 5.5(b). This is because the link speeds are bounded by the maximum downlink rate of 384kbps offered by the 3G networks.

Table 5.2 Comparison of link speed (in kbps)

Operator	Overseas Website			Local Website		
	Max	***Min***	***Avg.***	***Max***	***Min***	***Avg.***
X	104.3	18.0	68.7	329.4	130.6	204.6
Y	200.0	22.7	89.9	368.6	33.5	228.1
Z	142.2	42.0	82.3	233.1	164.8	199.4

[6] Connection speed to the following Internet speed test websites:
Overseas: http://www.auditmypc.com/broadband-speed-test.asp
Local(HK): http://www.netfront.net/speedtest/testspeed1.htm

Due to the blocking of ICMP messages by the 3G networks, we chose to use the TCPing[7] tool for our ping tests. TCPing is a TCP-based ping implementation for the Win32 platform. In measuring the RTT from the 3G networks to various websites located in Hong Kong and overseas, TCPing tries to setup four consecutive TCP connections with the websites. The time taken to complete each TCP connection establishment between our laptop client and the websites is the RTT. Looking through the source codes of TCPing, it is seen that between successive pings (TCP connection establishments), if the previous RTT is less than 2000ms, then TCPing will sleep for a period that is equal to the difference between 2000ms and the previous RTT value. While this sleep feature of TCPing has no effect on the achievable RTT values in 3G networks, it will be shown to have an effect on the achievable RTT values in HSDPA networks, as will be discussed later in Section 5.4.7.

Fig. 5.9 shows the ping test results, where the range of RTT values shown is from the 5th to 95th percentile, with the horizontal line representing the 50th percentile. We see that the RTT values to the overseas websites are generally higher than that to the local websites due to the long propagation delay to overseas websites. In addition, the minimum and average RTT values achieved across all local websites for (Operator X, Operator Y, Operator Z) are (187ms, 109ms, 140ms) and (385ms, 350ms, 201ms) respectively. This shows that the minimum RTT value achieved in lightly-loaded cells is more than five times to nine times lower than the lowest RTT value of 1000ms achieved in a fully-loaded cell as seen previously in Fig. 5.4(c) and Fig. 5.5(c). The RTT values from our ping tests also correspond to those reported in [43-46]. We have also performed the ping test using several broadband network providers (using ADSL, cable modem, Ethernet) in Hong Kong, and found the average RTT value to local websites to be 8.69ms, which is significantly lower than the minimum RTT values achieved with lightly-loaded 3G networks. Since the round-trip transmission time of a ping packet over the 3G wireless links is only about 9ms (assuming packet size = 50bytes, downlink BW = 200kbps, uplink BW = 60kbps, then RTT = 50bytes/200kbps

[7] http://www.elifulkerson.com/projects/tcping.php

+ 50bytes/60kbps), this clearly shows that the end-to-end data latency over the 3G wireless networks is dominated by the network processing and queueing delay which amounts to more than 100ms at a minimum.

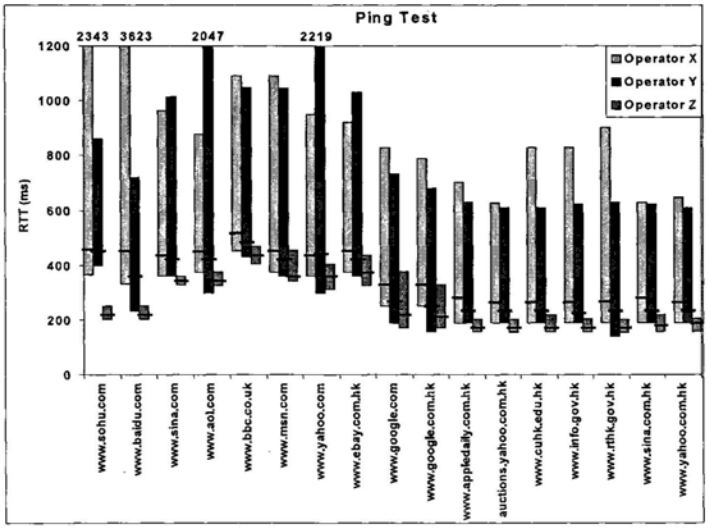


Figure 5.9 Ping test results

5.4.3 TCP Retransmissions in 3G Networks

We have calculated the percentages of TCP retransmissions in all our measurement tests involving data, where it is defined as the percentage ratio of the number of retransmitted bytes over the total number of bytes transmitted during the measurement tests. Fig. 5.10(a) shows that the majority of the measurement sites in Operator X's network have a TCP retransmission rate of 1%, with some sites even reaching up to 6%. In contrast, results for Operator Y and Operator Z show that all of the measurement sites in the networks of Operator Y and Operator Z have TCP retransmission rates below 1%, with the majority of the sites having TCP retransmission rates of 0.1% and 0.2% respectively. These results suggest that the networks of Operator Y and Operator Z may have been configured to use RLC Acknowledged Mode, thereby lowering their TCP retransmission rates. The results also suggest a low threshold value for the targeted signal-to-noise ratio in Operator X's

network, to allow for higher network capacity in terms of the data bandwidth capacity. This is in line with our capacity measurement results (to be discussed next in Section 5.4.4) which shows that Operator X generally has the higher data capacity results, compared to the other two operators.

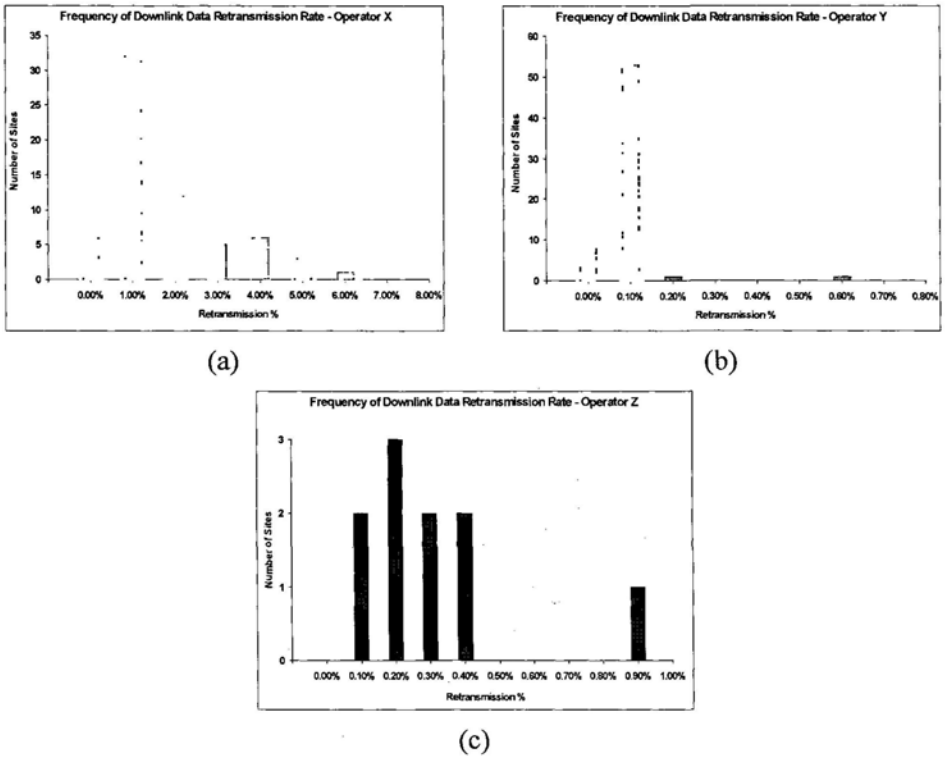


Figure 5.10 Percentages of TCP retransmissions for operators X, Y, and Z

5.4.4 Network Capacity Results

We have collated our measurement results and generated several graphs showing the 3G network capacities in terms of the downlink and uplink data bandwidth capacity, and the maximum number of video and voice calls that can be made in the cell. These results are representative of the 3G network capacity as perceived from an end-user's viewpoint.

Fig. 5.11(a) presents the measured uplink and downlink data bandwidth capacity results for the three operators. The data points represent the maximum aggregate throughput value out of all throughput values calculated over 30-seconds averaging windows throughout the whole duration (~1500s) of the downlink and uplink measurement tests, with each data point corresponding to the results for one measurement site. Generally we observe a positive correlation between the uplink and downlink data bandwidth capacities. On the whole, the uplink data bandwidth capacity is usually lower than the downlink data bandwidth capacity for a measurement site. Fig. 5.11(b) and Fig. 5.11(c) also show a positive correlation between the measured uplink data bandwidth capacity and the video and voice capacities. This is similar to the observation in Fig. 5.11(a) and we also obtained likewise results when plotting the downlink data bandwidth capacity against the video and voice capacities. In determining the voice capacity, we have utilized the MT7 and MT8 test results to workaround the limitation on the number of handsets that we have. From the MT8 test results, we can identify the number of voice calls that are equivalent to one video call. Table 5.3 shows a summary of the MT8 test results. To calculate the network's voice capacity, we use the formula "*Voice capacity* = (r x v) + o", where r = ratio of the number of voice calls to one video call, v = maximum number of video calls in the MT7 test, and o = number of voice calls in the MT7 test.

From the measurement results, the mean values and standard deviations for the four parameters of downlink and uplink data bandwidth capacities (in kbps), and the maximum number of video and voice calls that can be made for the three operators are tabulated in Table 5.4. The values are represented in the format of (Operator X, Operator Y, Operator Z). The cumulative distribution graphs for the four parameters are also plotted and shown in Fig. 5.12. Based on our trade-off test results (MT2, MT3, MT5, MT6, and MT8), we have also constructed the capacity region of a cell as shown in Fig. 5.13 for the case of heterogeneous services. A capacity region is a graphical representation of the cell capacity in terms of the mixture of circuit switched and packet switched services. The analysis in [50, 51, 53] has suggested that the boundary of the capacity region can be characterized by a linear plane, as shown in Fig. 5.13.

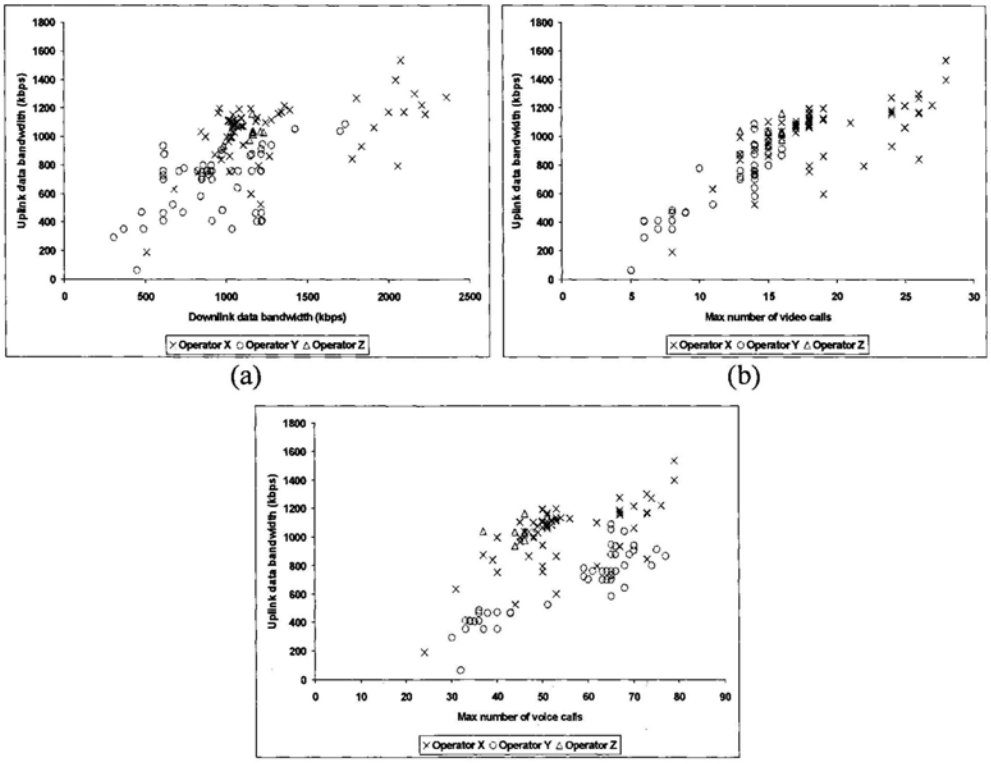


Figure 5.11 Capacity measurement results for operators X, Y, and Z

Table 5.3 Number of voice calls equivalent to one video call

Operator	Mean	Median	Standard Deviation
X	2.83	3	1.20
Y	4.52	4	1.71
Z	2.90	3	0.88

Table 5.4 Mean and standard deviations of the four-tuple parameters for operators X, Y, and Z

Parameter	Mean	Standard Deviation
Downlink data BW	(1259.9, 926.4, 1143.4)	(424.04, 296.68, 77.24)
Uplink data BW	(1041.3, 686.3, 1029.2)	(208.54, 209.56, 57.32)
Video calls	(18.88, 12.13, 15.4)	(4.38, 3.07, 0.97)
Voice calls	(54.29, 57.43, 44.9)	(11.52, 13.58, 3.0)

To verify the accuracy of this linear characterization (approximation), we have plotted in Fig. 5.14 the cross-sections of the normalized downlink/uplink bandwidth capacity versus the normalized video calls capacity for more than 40 cells of Operator X. From Fig. 5.14, we can observe that the linear capacity plane characterization turned out to be a good conservative estimation for more than 70% of the cells. The data points that fall below the linear capacity plane are relatively few and not that far from the 100% scaled capacity line. By further moving the linear capacity plane to the 90% scaled capacity line, we are able to obtain the linear capacity plane characterization of the capacity region's boundary for more than 90% of the cells. Similar results are also observed for other service tradeoffs, for the other operators.

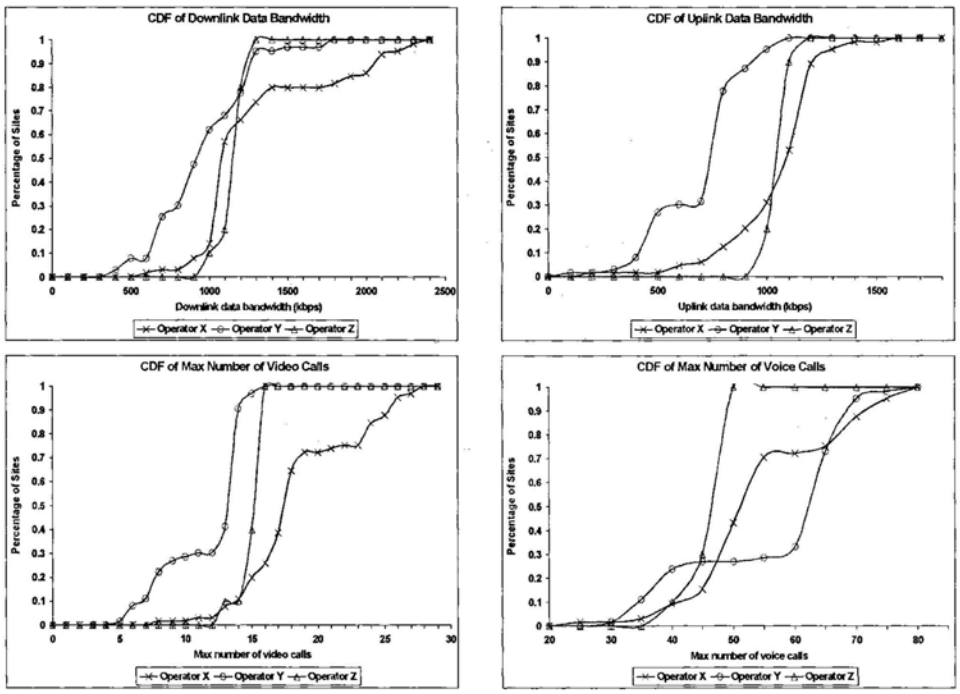


Figure 5.12 CDF of the capacity measurement results for operators X, Y, and Z

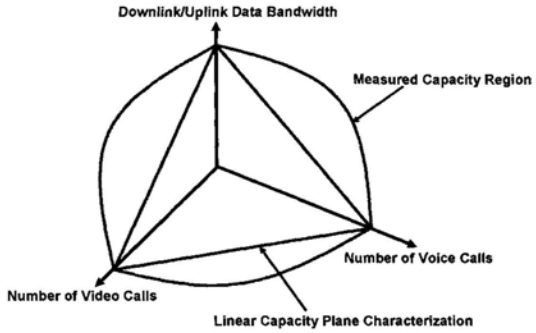


Figure 5.13 Capacity region of a cell

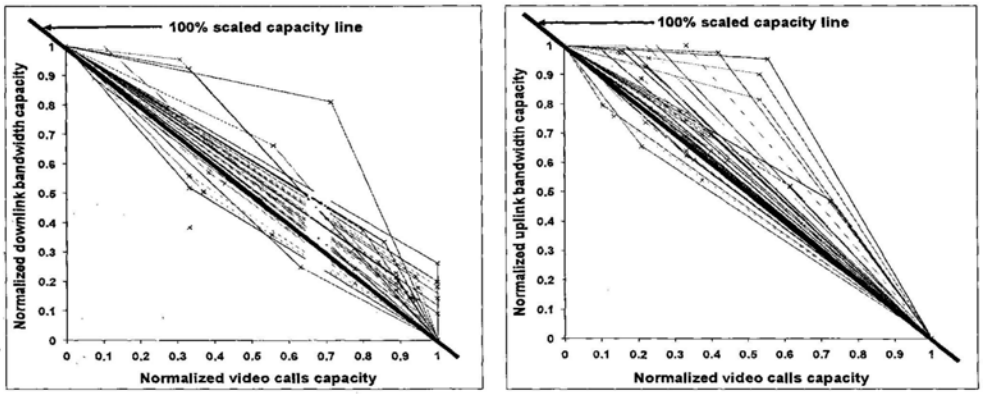


Figure 5.14 Normalized capacity trade-offs between downlink/uplink data bandwidth and the number of video calls for Operator X

5.4.5 Modeling of Measured 3G Network Capacity Results

Fig. 5.15 shows the measured cell capacities plotted with respect to the received E_c/N_0 at their respective measurement sites. The value for the E_c/N_0 parameter is the average of the received E_c/N_0 measured with six different handsets at the beginning of every measurement test. Fig. 5.15(a) and Fig. 5.15(b) show that the majority of both the downlink and uplink data bandwidth capacity results are spread across the E_c/N_0 range of -2dB to -6dB with no identifiable trend. Even for the same value of E_c/N_0, the data bandwidth capacities differ substantially across different cells and operators. Similar observations are also seen in Fig. 5.15(c) and Fig. 5.15(d) with respect to the relationship between the video or voice capacities and the received E_c/N_0. As such, it is

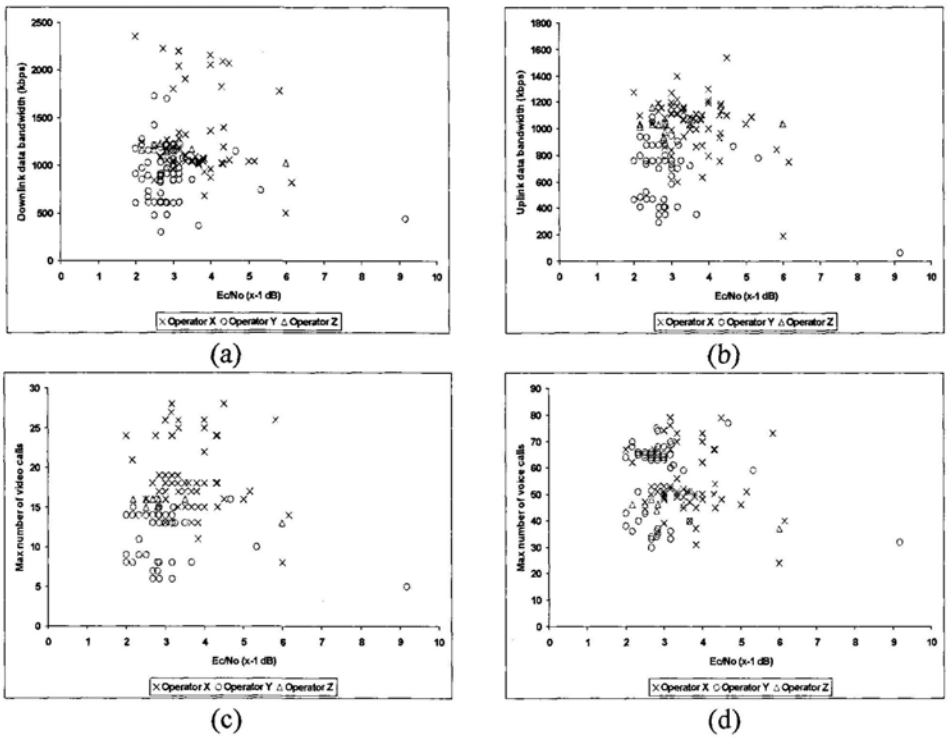


Figure 5.15 Cell capacities versus E_c/N_0 for operators X, Y, and Z

obvious that we cannot use measurements of E_c/N_0 alone to predict the network capacity.

We next try to categorize the measurement sites into different classes based on some common site characteristics like relative population densities (rural, suburban, urban) or usage nature of the sites (commercial, residential, industrial) to ascertain if we can identify any distinct clustering among the measured cell capacities results. However, Fig. 5.16 shows that even when the measurement sites are grouped into rural, suburban, and urban sites, we are still unable to identify any distinct clusters among the measured data bandwidth capacities results of Operator Y. Similar results are also obtained when measurement sites are grouped according to the usage nature of the sites, as well as when repeating these efforts on the measured video and voice capacities results, for all three operators.

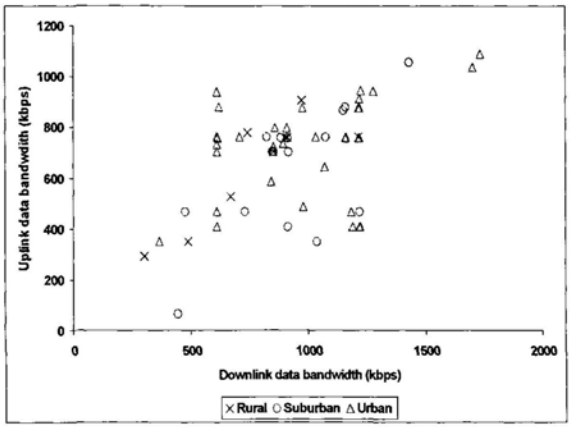
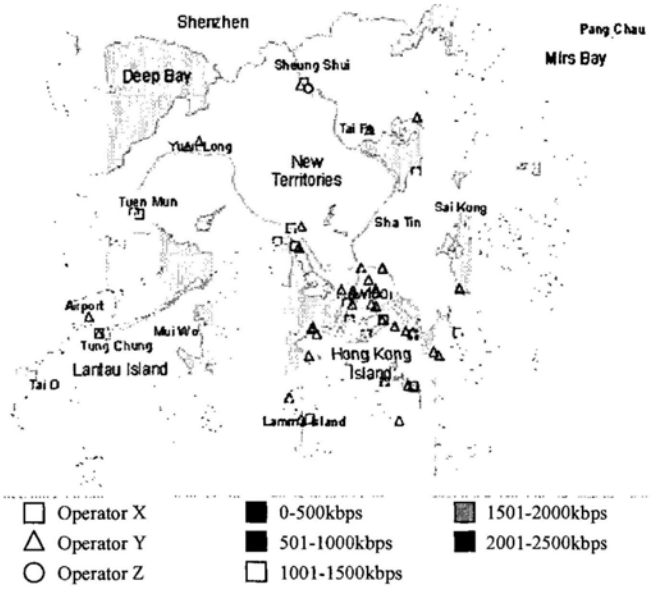


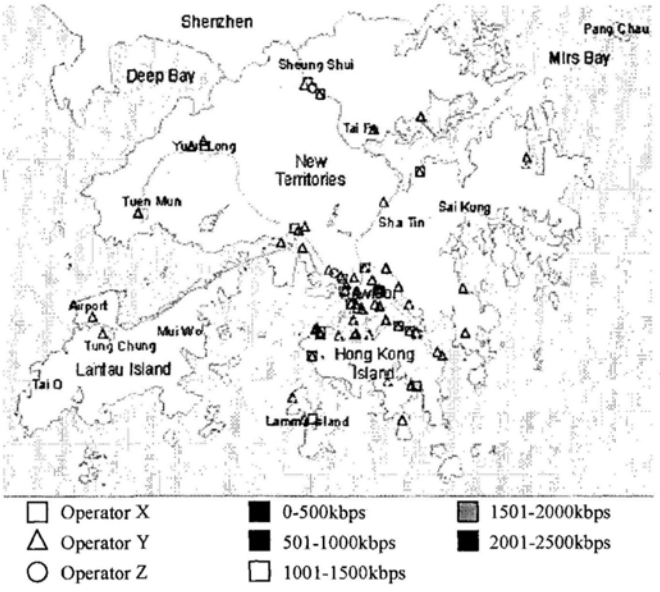
Figure 5.16 Data bandwidth capacities, with measurement sites categorized according to population densities for Operator Y

To explore further, we decided to study the capacity distribution patterns of the three operators by mapping the data, video, and voice capacity results onto their respective measurement locations in the Hong Kong map. Each measurement site's location is drawn based on their GPS location information. Since the capacity distribution maps for the data, video and voice capacities are by and large similar to one another, we will only show the capacity distribution map for the data bandwidth capacities.

Fig. 5.17 shows that the majority of the measurement sites are concentrated in the areas of Kowloon and the north of Hong Kong Island. These densely populated urban areas are the highly busy and important commercial zones in Hong Kong and hence the concentration of base stations from all three 3G operators in these areas is very high as well. This is corroborated with the cell site information used previously in the measurement site selection process described in Section 5.2.2. Covered by relatively small cells, Fig. 5.17 shows that these areas have medium to high cell capacities. On the other hand, sparsely populated rural areas (northeast part of New Territories, and Lamma Island) are often covered by relatively large cells, with low to medium cell capacities. Fig. 5.17 also shows several areas located in New Territories (e.g. Tuen Mun, Tung Chung, Sheung Shui, and the Airport) with medium to high cell capacities.



(a) Downlink data bandwidth capacities



(b) Uplink data bandwidth capacities

Figure 5.17 Distribution of data bandwidth capacities for operators X, Y, and Z

These suburban areas are often covered by relatively large cells as concluded from the low concentration of base stations there. These observations suggest that the 3G network operators seemed to have customized their network configuration in a cell-by-cell manner according to their own proprietary requirements, in order to cope with the projected traffic demand of individual sites and neighborhood.

Based on the observations in Fig. 5.17, we decided to investigate the relationship between the measured cell capacity and the cell coverage size. This relationship is supported by well-known theoretical models in CDMA systems (Chapter 8 in [103]), [105], as shown in Fig. 5.18. In our investigation, we used an estimated cell radius value (based on the cell site information that we had) to approximate the target coverage area of a cell. In our scheme, the nearest base station to the measurement site is chosen to be the targeted base station, and the cell radius is estimated to be one-half of the average distance between the targeted base station and the k-nearest base stations. We assume the use of omni-directional antenna at the base stations. However, our approximation method cannot account for other factors that can influence a cell's target coverage area, such as the use of a directional antenna, the antenna height and tilt-angle, the landscape in the area, etc. For example, we found instances where the base stations are located somewhat along a linear trajectory on highways. In these instances, k=1 would give a good estimate of the targeted cell's coverage area. On the other hand, values of $k > 1$ would give a better estimate of the cell's coverage area in instances where the k-nearest base stations are located around the targeted base station. Using these estimated cell radius values, we try to identify their correlation with the measured cell capacities, with the measurement sites grouped according to their relative population densities.

Fig. 5.19 shows the measured downlink data bandwidth capacity for Operator X, plotted with respect to the estimated cell radius, for two values of k=1 and k=3. However, we are still not able to identify any clear set of clusters among the measured data points. Similar results are also obtained for the other network capacity parameters, for all three operators. This observation reinforces the point that the actual settings or

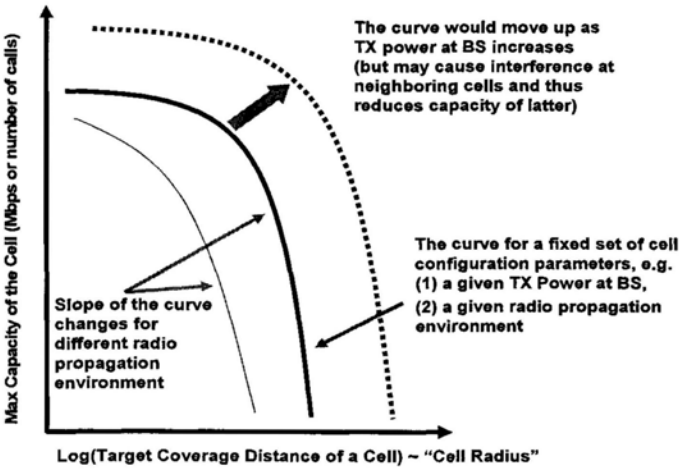


Figure 5.18 Theoretical model for cell capacity versus cell radius

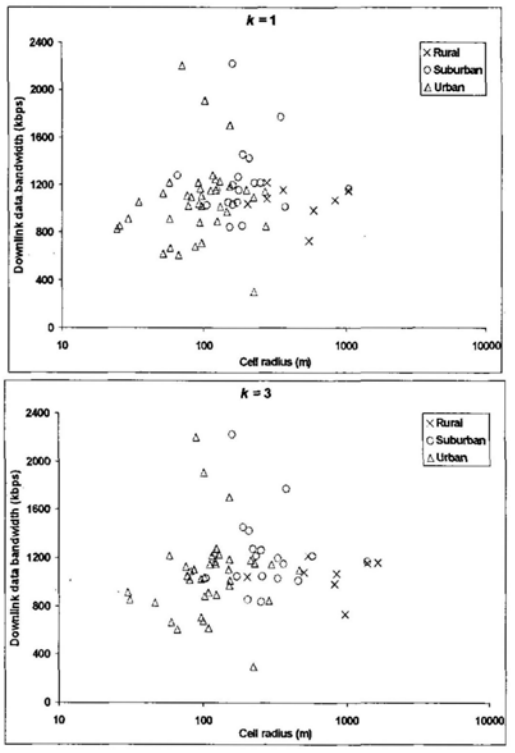


Figure 5.19 Downlink data bandwidth capacity versus cell radius for Operator X

parameters employed for each cell seemed to be very diverse across different cells, and that the cells cannot be grouped into a small number of sets for categorization purposes. It is obvious that the measured results do not fit the predictions of well-known theoretical capacity models using a reasonably small set of values for the configuration parameters, nor can the measured capacity data in one cell be used to predict the capacity of another cell.

5.4.6 Repeated Measurements

In order to investigate the relationship of the 3G network capacity with time-of-the-day, we have performed repeated measurements at several sites on different days and time-of-the-day. In general, we found that the video calls capacity in the first and second measurements do not differ much, with an average and maximum differences of 3.9% and 14.3% respectively. However, the data bandwidth capacities can vary substantially in some situations. Table 5.5 shows the results of the repeated measurements for downlink data bandwidth capacities, where each measurement result has been classified into one of the network activity factor categories of "very busy", "moderately busy", and "near idle". This classification is based on our best knowledge of the measurement site's characteristics during the time of measurement. We observe that for seven out of the ten sites, the repeated measurements were conducted during times when the network is considered to be moderately busy or nearly idle, and as such there is only a small capacity difference between the first and second measurements (average and maximum differences of 4.0% and 12.5% respectively). The only exception is the repeated measurement at Site B, a suburban measurement site located next to a warehouse and a construction workplace, where we observe a capacity difference of 35.1% even though we presumed the network loading to be moderately busy during the first and second measurements. This considerable capacity difference may have been caused by an unforeseen heavy network loading or changes in the physical and environmental surroundings between the measurement times of the first and second measurements.

Table 5.5 Downlink data bandwidth capacities (in kbps) in repeated measurements for operators X and Y (measurement times in brackets)

	Operator X				
	Site A	Site B	Site C	Site D	Site E
	Urban Residential	Suburban Warehouse/ Construction Site	Suburban Residential	Urban Market Place	Urban Ferry Pier
Moderately Busy	1034 (14:10)	1395 (14:40)	2093 (18:46)	1045 (12:54)	2193 (15:32)
	1094 (01:12)	2149 (09:25)			
Near Idle			2391 (06:17)	1052 (22:05)	2359 (03:10)
Difference	5.48%	35.09%	12.46%	0.67%	7.04%

	Operator Y				
	Site F	Site G	Site H	Site I	Site J
	Urban Commercial Offices	Urban Commercial Shopping	Rural Ferry Pier	Suburban Shopping	Urban Residential
Moderately Busy				1427 (11:10)	1156 (22:21)
				1458 (13:31)	1157 (15:23)
Near Idle	1702 (21:14)	1733 (02:09)	669 (20:02)		
			670 (15:00)		
Difference	35.55%	48.53%	0.15%	2.13%	0.09%

On the other hand, the repeated measurements at Site F and Site G were performed during periods of time when the network loading is considered to be vastly different ("very busy" vs. "nearly idle"), and thus it is not a surprise that there is a big

improvement (35.6% - 48.5%) in the downlink data bandwidth capacity between the first and second measurements. These findings indicate that different measurement times (off-peak hours vs. peak hours) can potentially affect the capacity measurement results, particularly in busy commercial zones. However, the majority of our measurements were conducted during off-peak hours and hence we were able to minimize any network capacity measurement errors due to factors beyond our control, such as the network loading. Furthermore, in the case of the data bandwidth capacity measurements, we increased the likelihood of being able to measure the network data capacity by taking the maximum aggregate throughput value out of all throughput values calculated over 30-seconds averaging windows throughout the whole duration (~1500s) of the measurement tests.

5.4.7 Performance in Lightly-loaded HSDPA Network

Operator Y's 3G network has since been upgraded (after the end of our 3G measurements) to support HSDPA with an advertised maximum downlink speed of up to 3.6Mbps. In order to determine the performance improvements over the 3G networks, we have carried out TCP bulk data transfer measurements (downlink and uplink) at several different locations in Hong Kong, using a single laptop equipped with a USB modem containing a HSDPA data card, connecting to our server over a lightly-loaded HSDPA network. Packet traces were captured at both the server and the laptop client. Analysis of the TCP downlink traces seemed to suggest that the TCP connection between the server and the laptop client is not end-to-end, but rather has been split into two at a TCP Performance Enhancing Proxy (TCP PEP) [106] in the HSDPA network. TCP PEPs are often employed at the intersection of wired and wireless networks to mitigate performance problems arising from wireless link transmission errors, or link disconnections. Nevertheless, the introduction of the TCP PEPs also leads to the violation of the end-to-end semantics of the TCP connections, additional processing delay over the TCP connections, as well as serious scalability issues in highly loaded networks [106].

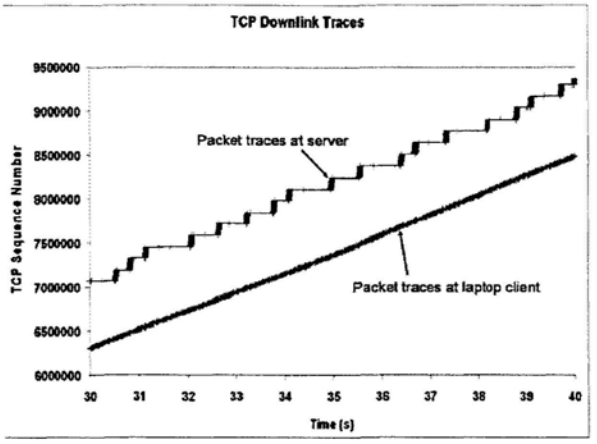


Figure 5.20 TCP downlink traces in a lightly-loaded HSDPA network

Evidences of the existence of the TCP PEP are seen in the snapshots of the packet traces in Fig. 5.20, in which we see the data packets being sent out in bursts at the server, whereas at the laptop client, the data packets are received in a continuous and regular manner. Closer inspection of the packet traces at the server revealed that the received TCP acknowledgments were periodically advertising a receiver window size of zero (TCP Zero Window), thereby preventing the server from sending more data and causing data packets to be sent in bursts when the TCP receiver window opens again. The examination of the server's packet traces also showed that the maximum advertised TCP receiver window size was 64KB, and that there were very little TCP retransmissions during the data transfer. On the other hand, inspection of the packet traces at the laptop client showed that not only the client's TCP receiver window did not close at all, but the advertised receiver window value was a steady 128KB in all the transmitted TCP acknowledgment packets (TCP Large Window option was enabled in the laptop client). All these observations seemed to suggest the existence of a TCP PEP in the HSDPA network that served to receive, acknowledge and buffer the data packets from the server, and forward the packets over the wireless link to the laptop client.

Table 5.6 TCP throughput results in a lightly-loaded HSDPA network (in kbps)

	Max	**Min**	**Average**
TCP Downlink	2970.64	972.69	1820.61
TCP Uplink	365.62	59.49	265.25

As a further validation test, we performed another test where midway through the TCP downlink data transfer, the USB modem's cable was disconnected from the laptop client. Once this happened, the client program in the laptop immediately terminated. However, the server program continued to run in the server, and analysis of the server's packet traces revealed that it was still receiving TCP acknowledgments for its transmitted packets, albeit with a decreasing TCP receiver window size that eventually closed as indicated by the reception of a TCP Zero Window segment. What was surprising is that long after the disconnection of the USB modem, the server was still sending out TCP Keep-Alive messages and receiving TCP ACK messages as replies from some entity in the HSDPA network. The interval between successive TCP Keep-Alive messages was progressively doubled until it reached a value of 120s, and thereafter the TCP Keep-Alive messages were sent out every 120s. This went on for about 1200s until the server finally received a TCP RESET message which caused the server program to terminate. Since the client program has long since ended, the receptions of the TCP ACK messages and the TCP RESET message at the server pointed to the existence of a TCP PEP in the HSDPA network.

Table 5.6 shows the TCP throughput results obtained in our measurements. We see that the TCP downlink throughput values range from a low of 972kbps to a high of 2970kbps, which is an improvement of 2.6 to 8.0 times compared to the maximum single-user TCP downlink throughput of 370kbps achieved in the 3G network. In fact, the maximum TCP downlink throughput of 2970kbps achieved by a single user in Operator Y's HSDPA network is about 1.75 times higher than the maximum aggregate TCP downlink throughput of 1.7Mbps obtained in the 3G network of Operator Y, seen previously in Fig. 5.11(a). Similarly, the TCP uplink throughput results in the HSDPA network of Operator Y, with a maximum of 365kbps, is more than 6 times higher than

the maximum single-user TCP uplink throughput of 60kbps achieved in the 3G network of Operator Y.

We have also performed the ping test (using the same TCPing tool that was used in Section 5.4.2) to several local websites and found that the achieved RTT values alternate between 60-125ms and 2000-6000ms. We also observed the data connection indicator on our USB modem automatically toggling between the "3G" and "HSDPA" modes during the ping test. This observation suggests that if there isn't enough data passing through the connection, the network will downgrade the HSDPA connection to a 3G connection and vice versa. Speculating that this "switching" observation is due to the sleep feature of TCPing when the RTT value is less than 2000ms (refer to Section 5.4.2), we decided to investigate further by removing the sleep function from TCPing and repeating the ping tests. Doing so, we found that the achieved RTT values to local websites are now consistently in the range of 60-125ms with an average of 81ms, and we also observed that the data connection indicator on the USB modem is consistently set to the "HSDPA" mode. In addition, these RTT values also correspond with the reported RTT values in [107].

The RTT observations with the HSDPA network prompted us to examine the effect of the sleep function in TCPing on the RTT values achieved over the 3G network. Using the 3G data card of Operator Y, we repeated the ping tests to local websites using the versions of TCPing with and without the sleep function. The results show that with TCPing-Sleep, the achieved RTT values are within the range of 125-969ms with an average of 175ms, while with TCPing-NoSleep, the achieved RTT values are in the range of 125-953ms with an average of 164ms. This observation indicates that the sleep function of TCPing has no effect on the achievable 3G RTT values to local websites. We also see that the minimum RTT values are similar to those reported in Section 5.4.2 for Operator Y. On the other hand, due to the smaller sample size in our new ping tests, the average RTT values are lower compared to those reported in Section 5.4.2.

For a comparison of the network processing and queueing delay in the HSDPA and 3G networks of Operator Y, we note that the round-trip transmission time of a ping packet over the wireless link in a HSDPA network is only about 2.2ms (assuming packet size=50bytes, downlink BW=2000kbps, uplink BW=200kbps, then RTT=50bytes/2000kbps + 50bytes/200kbps). Therefore, this shows that the network processing and queueing delay in the HSDPA network (~60ms at a minimum) is lower than that of the 3G network (>100ms at a minimum).

5.5 Summary

We have designed and performed a large scale empirical study on the performance of three commercial 3G networks in Hong Kong, using a mixture of data, video and voice traffic. The conclusions of this work are based on the measurement results from real operational networks, albeit with the caveat that the performance can change with different network settings, different vendor equipment as well as when operators perform software upgrades to their networks. Nevertheless, our results do provide researchers with significant, real performance results from three different operational networks. Our key findings include:

- Substantial additional processing and queueing latency for data services (>100ms at a minimum) are imposed by the 3G network when compared to wireline IP networks, even under lightly-loaded scenarios.
- Average latency for 3G data services increases to beyond 1sec in fully-loaded network conditions.
- The minimum RTT values achievable in the HSDPA network (~60ms) is about 50% of that achievable in the 3G network (~125ms) of the same operator. Similarly, the network processing and queueing latency in the HSDPA network is also about 50% lower than that of the 3G network.
- There is a maximum eight-fold increase in the TCP downlink individual throughput achieved in the HSDPA network (2970kbps), compared to that achieved in the 3G network (370kbps).

- There is a maximum six-fold increase in the TCP uplink individual throughput achieved in the HSDPA network (365kbps), compared to that achieved in the 3G network (60kbps).
- The TCP retransmission probability over the wireless link varies substantially across different operators (from <<1% to 1~6%), which points to the different operational configurations across different operators in terms of the configured RLC data transfer mode and the threshold value for the targeted signal-to-noise ratio in the 3G networks.
- Our HSDPA measurement results pointed to the existence of a TCP PEP in the HSDPA network.
- We have found that 3G networks operating under near saturated conditions exhibit highly unpredictable behavior, e.g. difficulties (requiring repeated attempts) in adding supposedly higher priority video/voice calls into a data saturated network, with a substantial number of observed cases where not even one video/voice call could be added, and multiple data calls (especially uplink data calls) being dropped unexpectedly.
- Our 3G measurement results suggest the existence of a guaranteed minimum bandwidth resource for the data services, when data calls are present in the network. The results also indicate that this guaranteed data bandwidth threshold value is different across different operators and measurement sites. Such findings further demonstrate the diverse nature of the network resources allocation mechanisms and the call admission control policies employed by different operators.
- The multi-service capacity region of a cell can be conservatively bounded by a plane with axis-intercept values equal to the maximum capacity for the corresponding single-service types, i.e. maximum data bandwidth, maximum number of video/voice calls.
- 3G network operators seem to have extensively customized their network configurations in a cell-by-cell manner in order to cope with the projected traffic demand of individual sites and neighborhood, and the target coverage area of the base station. As such, the cell capacity varies widely not only across different

operators but also across different measurement sites of the same operator.

- It is practically impossible to predict the actual capacity of a particular cell based on known theoretical models and standard parameters, even when supplemented by key field measurements such as the received signal-to-noise ratio (E_c/N_0). In other words, it is practically impossible to derive the necessary model parameters which can provide a good fit between the theoretical prediction and the measured capacity values.

To the best of our knowledge, this is the first public report on a large scale empirical study on the performance and capacities of commercial 3G networks carrying live data, video and voice traffic. Our results are useful to both network planners in planning enough capacity to support new data applications, and to application designers in designing flow control algorithms to handle the bandwidth variability as the video/voice calls arrive and terminate.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In wireless networks, the effectiveness of resource sharing schemes are often heavily influenced by different aspects of the system behavior, such as user mobility, traffic dynamics and practical realization constraints. In this thesis, using analytical models, simulations and empirical measurements, we have investigated the impact of these system behaviors on the performance of resource sharing in three types of wireless networks. In Chapter 3, we have analyzed the data communication performance of a vehicle in a Drive-thru Internet system. The amount of data downloaded by an individual user is affected not only by the AP's scheduling algorithms, but also by the user dynamics, i.e. the movement of the vehicles which impacts the amount of time the vehicle spends in the AP's coverage range, as well as the number of contending vehicles for the AP's resources. We have derived practical analytical models with tractable solutions for the average and distribution of the number of bytes downloaded by a vehicle by the end of its sojourn through an AP's coverage range. Our model quantifies the impact of road traffic density, vehicle speed, service penetration rate, AP wireless coverage range and the corresponding AP transmission rate, on the amount of downloaded data. The accuracy of our analytical models has been validated via simulations. To the best of our knowledge, our work provides the first analytical model that characterizes the interplay between vehicular traffic parameters and a vehicle's data communication performance in a Drive-thru Internet system.

In Chapter 4, we have proposed and analyzed the performance of a TDMA-based and receiver-driven MAC protocol (RMAC) in wireless sensor networks. RMAC utilizes knowledge of a node's timeslots in order to schedule collision-free transmissions in the

network. Channel utilization is also improved through the provision of mechanisms that enable the "stealing" of unused timeslots. We analytically derived the delay performance of RMAC (with and without timeslots stealing) and show that the timeslots stealing feature can substantially improve the network capacity in situations with varying and asymmetric traffic patterns. In addition, we proposed a simple timeslots reassignment procedure to redistribute the timeslots among the sender nodes according to their estimated offered traffic load. Our results show that it is better to perform the timeslots reassignment procedure infrequently. This is so that during low packet arrival rates, the network will be able to obtain a better estimate of the traffic load and hence produce a better timeslot assignment while at the same time allowing the network to conserve its energy. During high packet arrival rates however, it can rely on the timeslots stealing feature in order to react to shorter timescale changes in the traffic patterns and reduce the average packet latency.

In Chapter 5, we have presented the findings of a large-scale measurement study on the performance and capacities of live commercial 3G networks, carrying data, video and voice traffic. We have studied the performance of the networks in terms of their ability to provide service guarantees to different traffic classes as well as the fairness of the radio-link scheduler in allocating the bandwidth resources to multiple data calls in a saturated network. We have also investigated the data throughput, latency, video and voice calls handling capacities of the 3G networks under saturated network conditions. Our findings demonstrate the diverse nature of the network resources allocation mechanisms and the call admission control policies adopted by different operators. Our results also point out that the 3G network operators seem to have extensively customized their network configurations in a cell-by-cell manner in order to cope with the projected traffic demand of individual sites and neighborhood, and the target coverage area of the base station. As such, the cell capacity varies widely not only across different operators but also across different measurement sites of the same operator. To the best of our knowledge, our work is the first large-scale study of its kind that evaluates the performance of live 3G networks under saturated conditions, using a mixture of data, video and voice traffic.

6.2 Future Work

There are several interesting directions for future research based on the work described in this thesis. We list our suggestions below.

In Chapter 3, our approach for computing the distribution of the number of bytes downloaded by a drive-thru vehicle can be adapted for more general scenarios including: (a) the presence of multiple classes of vehicles with different transmission rates, (b) the support of dynamic rate adaptation during a drive-thru, and (c) multiple-lane, bi-directional roads. This can be accomplished by (i) designing a different set of reward rates to be associated with the states of the Markov reward process and/or (ii) expanding the dimension of the corresponding Markov chain. It will also be interesting to try to relax the Poisson arrival assumption to model more bursty vehicular arrivals, e.g. using Markov-modulated Poisson processes. Our analytical models and solutions can also be used as the basic (per-node) building blocks in formulating various network-wide optimal resource allocation, scheduling and data routing problems, such as those in [108], in which packets are scheduled and delivered to selected APs within a network of inter-connected Drive-thru Internet systems based on vehicle routes, road traffic conditions as well as the communication requirements of individual vehicles.

In Chapter 4, we had used Algorithm 4.1 to recursively compute the utilization ρ_i of a sender node i, given an initial value for the average number of timeslots K_i that the sender node i has in a frame. It would be useful to prove that this algorithm always converges to the correct value of ρ_i, given any arbitrary initial value of K_i. Other than the analysis of the delay performance of our proposed MAC protocol, it would also be useful to analyze the power consumption performance of the protocol, with and without the timeslots stealing feature. The average amount of power consumed in a frame is dependent on the utilization of the individual nodes, their traffic load, as well as the average number of timeslots that a node has in a frame.

In Chapter 5, we mentioned that for our 3G network capacity measurement purposes, we do not consider the video/voice call quality, but rather only take into account the number of video/voice calls that can be successfully added into the network. While this is acceptable for circuit-switched type of services which take up a fixed amount of bandwidth, it may not be true for an all IP-based services envisioned in the future. Therefore for our future work, we intend to focus on the investigation of the QoS of voice and video calls (as perceived by the end-users) under varying load in the network. Recent advances in the research of the human perception system has resulted in a number of algorithms that can use various objective link/network/content-level measurements to derive and estimate the corresponding "subjective" service quality perceived by human users. Such schemes include those on perceptual voice quality estimation [109, 110], image quality assessments [111], and video quality assessments [112, 113]. We intend to adapt and extend these state-of-the-art Multimedia Quality Assessment (MQA) schemes to support our investigation of the Quality-of-Experience (QoE) as perceived by the end-users.

Appendix A -

A Proof for Eqn. (3.14)

Recall Eqn. 3.14, reproduced below,

$F^*[s,\tau \mid i,l,0]$

$$= \frac{1}{(sB\tau)^l}\sum_{m=0}^{l}(-1)^{m+l}\binom{l}{m}(i+m)^{l-1}[i(i+1)\ldots(i+l)]exp\left(-\frac{sB\tau}{i+m}\right) \qquad \text{(A.1)}$$

$F^*[s, \tau \mid i, l, 0]$ is defined as the LST of the distribution of data downloaded in the interval $(0, \tau)$, conditioned on the events that our tagged vehicle starts this interval as part of a group of i vehicles and that there are l new vehicle arrivals during this interval with only vehicle arrival events. In this Appendix, we will prove Eqn. (A.1) by using the method of induction.

For l=0, Eqn. (A.1) gives us

$$F^*[s,\tau \mid i,0,0] = exp\left(-\frac{sB\tau}{i}\right) \qquad \text{(A.2)}$$

Let Y_τ denote the amount of data downloaded by our tagged vehicle as part of a group of i vehicles, in an interval $(0, \tau)$ with no vehicle arrival nor departure events. Given the preceding conditions, Y_τ is exactly $(B\tau/i)$. Therefore, the LST of Y_τ is $\exp(-sB\tau/i)$, which is what Eqn. (A.1) gives us with l=0, as shown in Eqn. (A.2).

Next, if we assume Eqn. (A.1) is true for l arrivals, then for $(l+1)$ arrivals, Eqn. (A.1) gives us

$F^*[s,\tau \mid i,l+1,0]$

$$= \frac{1}{(sB\tau)^{l+1}}\sum_{m=0}^{l+1}(-1)^{m+l+1}\binom{l+1}{m}(i+m)^{l}[i(i+1)\ldots(i+l+1)]exp\left(-\frac{sB\tau}{i+m}\right) \qquad \text{(A.3)}$$

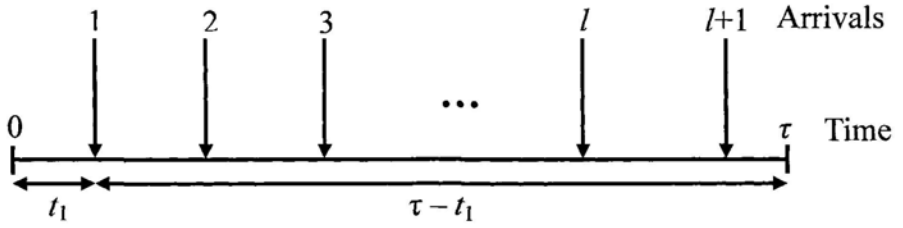


Figure A.0.1 (l+1) vehicle arrivals during an interval (0, τ)

To evaluate F*[s, τ | i, l+1, 0], we condition on the first vehicle arrival instant t_1, as shown in Fig. A.1.

$$F^*[s,\tau \mid i,l+1,0]$$

$$= \int_0^\tau F^*[s,\tau \mid i,l+1,0;t_1]\, f(t_1)dt_1$$

$$= \int_0^\tau F^*[s,\tau \mid i,l+1,0;t_1]\, \frac{l+1}{\tau^{l+1}}(\tau - t_1)^l dt_1$$

$$= \int_0^\tau exp\left(-\frac{sBt_1}{i}\right) F^*[s,\tau - t_1 \mid i+1,l,0]\, \frac{l+1}{\tau^{l+1}}(\tau - t_1)^l dt_1 \qquad \text{(A.4)}$$

where the last line above is obtained due to the independence of the distributions of data downloaded in the intervals (0, t_1) and (t_1, τ) conditioned on the number of vehicles at the beginning of each interval.

By replacing Eqn. (A.1) into the second term of Eqn. (A.4), doing the integration, and simplifying the resultant expression, we obtain

$$F^*[s,\tau \mid i,l+1,0]$$

$$= \frac{1}{(sB\tau)^{l+1}}\left\{\sum_{m=0}^{l}(-1)^{m+l+1}\binom{l+1}{m+1}(i+m+1)^l[i(i+1)\ldots(i+l+1)]exp\left(-\frac{sB\tau}{i}\right)\right.$$

$$\left. - \sum_{m=0}^{l}(-1)^{m+l+1}\binom{l+1}{m+1}(i+m+1)^l[i(i+1)\ldots(i+l+1)]exp\left(-\frac{sB\tau}{i+m+1}\right)\right\} \qquad \text{(A.5)}$$

We will consider the two terms in Eqn. (A.5) separately.

For the second term,

$$\sum_{m=0}^{l}(-1)^{m+l+1}\binom{l+1}{m+1}(i{+}m{+}1)^{l}[i(i{+}1)\ldots(i{+}l{+}1)]exp\left(-\frac{sB\tau}{i{+}m{+}1}\right)$$

$$=(-1)^{l+1}\binom{l+1}{1}(i{+}1)^{l}[i(i{+}1)\ldots(i{+}l{+}1)]exp\left(-\frac{sB\tau}{i{+}1}\right)$$

$$+(-1)^{l+2}\binom{l+1}{2}(i{+}2)^{l}[i(i{+}1)\ldots(i{+}l{+}1)]exp\left(-\frac{sB\tau}{i{+}2}\right)$$

$$+\cdots$$

$$+(-1)^{2l+1}\binom{l+1}{l+1}(i{+}l{+}1)^{l}[i(i{+}1)\ldots(i{+}l{+}1)]exp\left(-\frac{sB\tau}{i{+}l{+}1}\right)$$

$$=\sum_{m=1}^{l+1}(-1)^{m+l}\binom{l+1}{m}(i{+}m)^{l}[i(i{+}1)\ldots(i{+}l{+}1)]exp\left(-\frac{sB\tau}{i{+}m}\right) \qquad \text{(A.6)}$$

For the first term,

$$\sum_{m=0}^{l}(-1)^{m+l+1}\binom{l+1}{m+1}(i{+}m{+}1)^{l}[i(i{+}1)\ldots(i{+}l{+}1)]exp\left(-\frac{sB\tau}{i}\right)$$

$$=(-1)^{l+1}[i(i{+}1)\ldots(i{+}l{+}1)]exp\left(-\frac{sB\tau}{i}\right)\sum_{m=0}^{l}(-1)^{m}\binom{l+1}{m+1}(i{+}m{+}1)^{l} \qquad \text{(A.7)}$$

We consider the summation term in Eqn. (A.7) as follows:

$$\sum_{m=0}^{l}(-1)^{m}\binom{l+1}{m+1}(i{+}m{+}1)^{l}$$

$$=\sum_{m=0}^{l}(-1)^{m}\binom{l+1}{m+1}\sum_{j=0}^{l}\binom{l}{j}(m{+}1)^{j}i^{l-j} \qquad \text{(using Binomial Theorem)}$$

$$=(l+1)\sum_{m=0}^{l}(-1)^{m}\binom{l}{m}\frac{1}{m{+}1}\sum_{j=0}^{l}\binom{l}{j}(m{+}1)^{j}i^{l-j}$$

$$=(l+1)\sum_{m=0}^{l}(-1)^{m}\binom{l}{m}\sum_{j=0}^{l}\binom{l}{j}(m{+}1)^{j-1}i^{l-j}$$

$$= (l+1)\left\{\left[\sum_{m=0}^{l}(-1)^m\binom{l}{m}\frac{i^l}{m+1}\right]+\left[\sum_{m=0}^{l}(-1)^m\binom{l}{m}\sum_{j=1}^{l}\binom{l}{j}(m+1)^{j-1}i^{l-j}\right]\right\} \tag{A.8}$$

For the expression in the first term of Eqn. (A.8), it can be evaluated by using Eqn. (5.41) in [114] which states that

$$\sum_m \binom{l}{m}\frac{(-1)^m}{x+m} = x^{-1}\binom{x+l}{l}^{-1}$$

Therefore,

$$\sum_{m=0}^{l}(-1)^m\binom{l}{m}\frac{i^l}{m+1} = i^l\binom{1+l}{l}^{-1} \tag{A.9}$$

For the expression in the second term of Eqn. (A.8), it can be evaluated by using Eqn. (5.42) in [114] which states that

$$\sum_m \binom{l}{m}(-1)^m(a_0 + a_1 m + \cdots + a_l m^l) = (-1)^l l!\, a_l \quad , \quad \text{integer } l \geq 0$$

where $(a_0 + a_1 m + \cdots + a_l m^l)$ is a polynomial in m of degree l.

Since the expression $\sum_{j=1}^{l}\binom{l}{j}(m+1)^{j-1}i^{l-j}$ in the second term of Eqn. (A.8) is only a polynomial in m of degree (l-1), therefore the coefficient a_l=0. Hence, the whole expression in the second term of Eqn. (A.8) evaluates to zero.

Therefore, by using this information and Eqn. (A.9), Eqn. (A.8) can be evaluated to become

$$\sum_{m=0}^{l}(-1)^m\binom{l+1}{m+1}(i+m+1)^l$$

$$= (l+1)\left\{i^l\binom{1+l}{l}^{-1}+0\right\}$$

$$= i^l$$

Replacing the above into Eqn. (A.7), we obtain

$$\sum_{m=0}^{l}(-1)^{m+l+1}\binom{l+1}{m+1}(i+m+1)^{l}[i(i+1)\dots(i+l+1)]exp\left(-\frac{sB\tau}{i}\right)$$

$$=(-1)^{l+1}i^{l}[i(i+1)\dots(i+l+1)]exp\left(-\frac{sB\tau}{i}\right)$$

Then by replacing the above and Eqn. (A.6) into Eqn. (A.5), we obtain

$$F^{*}[s,\tau\mid i,l+1,0]$$

$$=\frac{1}{(sB\tau)^{l+1}}\left\{(-1)^{l+1}i^{l}[i(i+1)\dots(i+l+1)]exp\left(-\frac{sB\tau}{i}\right)\right.$$

$$\left.-\sum_{m=1}^{l+1}(-1)^{m+l}\binom{l+1}{m}(i+m)^{l}[i(i+1)\dots(i+l+1)]exp\left(-\frac{sB\tau}{i+m}\right)\right\}$$

$$=\frac{1}{(sB\tau)^{l+1}}\left\{(-1)^{l+1}i^{l}[i(i+1)\dots(i+l+1)]exp\left(-\frac{sB\tau}{i}\right)\right.$$

$$\left.+\sum_{m=1}^{l+1}(-1)^{m+l+1}\binom{l+1}{m}(i+m)^{l}[i(i+1)\dots(i+l+1)]exp\left(-\frac{sB\tau}{i+m}\right)\right\}$$

$$=\frac{1}{(sB\tau)^{l+1}}\sum_{m=0}^{l+1}(-1)^{m+l+1}\binom{l+1}{m}(i+m)^{l}[i(i+1)\dots(i+l+1)]exp\left(-\frac{sB\tau}{i+m}\right)$$

which is equivalent to what Eqn. (A.1) gives us with (l+1) arrivals, as shown in Eqn. (A.3).

□

Appendix B -

A Method to Evaluate Eqn. (3.26)

Recall Eqn. 3.26, reproduced below,

$$\text{Prob}\{Y_T > x, X_T = j \mid X_0 = (0; i)\} = F_{i,j}(x, T) \tag{B.1}$$

which describes the joint probability distribution of the accumulated data Y_T over a sojourn period $(0, T)$, and the probability that the Markov process is in state j at time T, conditioned on the event that the process was in state $(0; i)$ at time 0. A method to evaluate this expression can be found in [71]. In this Appendix, we give a summary of the important results in [71] that are useful for our work.

In [71], Sericola considered a continuous-time Markov chain $\{X_t, t \geq 0\}$ with infinitesimal generator matrix Q, and a reward rate $\rho(i)$ associated with each state i of the state space S. The random variable Y_t is defined by

$$Y_t = \int_0^t \rho(X_u)du$$

The number of distinct rewards is $m+1$, and their different values are $r_0 < r_1 < \cdots < r_{m-1} < r_m$. The state space S is partitioned into disjoint subsets containing the states with the same reward rates. States in the subsets A_l, $l = 0, \ldots, m$, have the same reward rate r_l. Without loss of generality, r_0 is set to 0. This can be done by considering the random variable $Y_t - r_0 t$ instead of Y_t, and the reward rates $r_l - r_0$ instead of r_l.

Let P denote the transition probability matrix of the uniformized discrete-time Markov chain $\{Z_n, n \geq 0\}$ associated to the Markov chain $\{X_t\}$. With the same initial distribution, P is related to Q by the relationship $P = I + Q/\varphi$, where I is the identity matrix and φ satisfies $\varphi \geq \max\{-Q_{ii}, i \in S\}$. The rate φ is the rate of the Poisson process $\{N_t, t \geq 0\}$, independent of Z, that counts the number of transitions of the process $\{Z_{N_t}, t \geq 0\}$ over $[0, t)$. It is well known that the processes $\{Z_{N_t}\}$ and X are stochastically equivalent.

For every $i, j \in S$, and $t > 0$, the function $F_{i,j}(x, t)$ is defined by

$$F_{i,j}(x, t) = \text{Prob}\{Y_t > x, X_t = j \mid X_0 = i\} \tag{B.2}$$

Let $F(x, t)$ denote the matrix containing the terms $F_{i,j}(x, t)$ for $i, j \in S$. Using the partition $A_0, \ldots, A_m$, the matrices Q, P, and $F(x, t)$ can be written, for $u, v = 0, \ldots, m$, as

$$Q = \{Q_{A_u A_v}\}; \quad P = \{P_{A_u A_v}\}; \quad F(x, t) = \{F_{A_u A_v}(x, t)\}$$

Denote $N(t, t+s)$ to be the number of transitions of the Poisson process N_t during the interval $[t, t+s)$. By conditioning on the number of transitions in the interval $[t, t+s)$, we have

$$\begin{aligned}\text{Prob}\{Y_{t+s} > x, X_{t+s} = j \mid X_0 = i\} = \quad & \text{Prob}\{Y_{t+s} > x, X_{t+s} = j, N(t, t+s) = 0 \mid X_0 = i\} \\ & + \text{Prob}\{Y_{t+s} > x, X_{t+s} = j, N(t, t+s) = 1 \mid X_0 = i\} \\ & + \text{Prob}\{Y_{t+s} > x, X_{t+s} = j, N(t, t+s) \geq 2 \mid X_0 = i\} \end{aligned} \tag{B.3}$$

Without going into the details, in [71], the three terms in the right hand side of Eqn. (B.3) are evaluated separately to give the following relationships:

$$\text{Prob}\{Y_{t+s} > x, X_{t+s} = j, N(t, t+s) = 0 \mid X_0 = i\} = (1 - \varphi s)\, F_{i,j}(x - \rho(j)s, t) + o(s) \tag{B.4}$$

$$\lim_{s \to 0} \frac{\text{Prob}\{Y_{t+s} > x, X_{t+s} = j, N(t, t+s) = 1 \mid X_0 = i\}}{s} = \varphi \sum_{k \in S} F_{i,k}(x, t) P_{k,j} \tag{B.5}$$

$$\text{Prob}\{Y_{t+s} > x, X_{t+s} = j, N(t, t+s) \geq 2 \mid X_0 = i\} = o(s) \tag{B.6}$$

where the notation $o(s)$ denotes any function that goes to zero with s faster than s itself.

Putting together these three terms back into Eqn. (B.3), we obtain

$$\begin{aligned}\frac{F_{i,j}(x, t+s) - F_{i,j}(x, t)}{s} = & \frac{(1 - \varphi s) F_{i,j}(x - \rho(j)s, t) - F_{i,j}(x, t)}{s} \\ & + \frac{\text{Prob}\{Y_{t+s} > x, X_{t+s} = j, N(t, t+s) = 1 \mid X_0 = i\}}{s} + \frac{o(s)}{s}\end{aligned}$$

Be letting s tend to 0, we get

$$\frac{\partial F_{i,j}(x, t)}{\partial t} = -\rho(j) \frac{\partial F_{i,j}(x, t)}{\partial x} - \varphi F_{i,j}(x, t) + \varphi \sum_{k \in S} F_{i,k}(x, t) P_{k,j}$$

Since $P = I + Q/\varphi$, we obtain

$$\frac{\partial F_{i,j}(x,t)}{\partial t} = -\rho(j)\frac{\partial F_{i,j}(x,t)}{\partial x} + \sum_{k\in S} F_{i,k}(x,t)Q_{k,j} \tag{B.7}$$

Eqn. (B.7) is a hyperbolic forward partial differential equation describing the behavior of the pair (Y_t, X_t), having a unique solution [115] on the domain $\mathbb{E}$

$$\mathbb{E} = \left\{(x,t); t > 0 \text{ and } x \in \bigcup_{l=1}^{m}(r_{l-1}t, r_l t)\right\}$$

with the initial condition given by

$$F_{i,j}(0,t) = \text{Prob}\{X_t = j \mid X_0 = i\} - \text{Prob}\{Y_t = 0, X_t = j \mid X_0 = i\}$$

A solution for Eqn. (B.7) is given by the following theorem.

Theorem B.1:

For every $t > 0$, and $x \in [r_{h-1}t, r_h t)$, for $1 \le h \le m$,

$$F(x,t) = \sum_{n=0}^{\infty} e^{-\varphi t}\frac{(\varphi t)^n}{n!}\sum_{l=0}^{n}\binom{n}{l}x_h^l(1 - x_h)^{n-l}C^{(h)}(n,l) \tag{B.8}$$

where

$$x_h = \frac{x - r_{h-1}t}{(r_h - r_{h-1})t}$$

and $C^h(n,l) = \{C^{(h)}_{A_u A_v}(n,l)\}_{0\le u,v\le m}$ are matrices given by the following recurrence relations:

For $0 \le u \le m$, and $h \le v \le m$:

for $n \ge 0$:

$$C^{(1)}_{A_u A_v}(n,0) = (P^n)_{A_u A_v} \quad \text{and} \quad C^{(h)}_{A_u A_v}(n,0) = C^{(h-1)}_{A_u A_v}(n,n) \text{ for } h > 1$$

for $1 \le l \le n$:

$$C^{(h)}_{A_u A_v}(n,l) = \frac{r_v - r_h}{r_v - r_{h-1}}C^{(h)}_{A_u A_v}(n,l-1) + \frac{r_h - r_{h-1}}{r_v - r_{h-1}}\sum_{w=0}^{m} C^{(h)}_{A_u A_w}(n-1,l-1)\,P_{A_w A_v}$$

For $0 \leq u \leq m$, and $0 \leq v \leq h-1$:

for $n \geq 0$:

$$C^{(m)}_{A_u A_v}(n,n) = 0_{A_u A_v} \quad \text{and} \quad C^{(h)}_{A_u A_v}(n,n) = C^{(h+1)}_{A_u A_v}(n,0) \text{ for } h < m$$

for $0 \leq l \leq n-1$:

$$C^{(h)}_{A_u A_v}(n,l) = \frac{r_{h-1} - r_v}{r_h - r_v} C^{(h)}_{A_u A_v}(n,l+1) + \frac{r_h - r_{h-1}}{r_h - r_v} \sum_{w=0}^{m} C^{(h)}_{A_u A_w}(n-1,l) P_{A_w A_v}$$

Proof: See [71]. □

In Theorem B.1, all computed quantities are between 0 and 1, and require only additions and multiplications of non-negative quantities. Therefore, Theorem B.1 yields a computational method that avoids any overflow problem, and leads to a stable algorithm whose precision can be specified in advance [72]. By letting ε be the desired precision for the computation of $F(x, T)$, we can define the integer N_{min} by

$$N_{min} = min\left\{ n \geq 0 \left| \sum_{i=0}^{n} e^{-\varphi t} \frac{(\varphi t)^i}{i!} \geq 1 - \varepsilon \right. \right\}$$

We can then rewrite Eqn. (B.8) as

$$F(x,t) = \sum_{n=0}^{N_{min}} e^{-\varphi t} \frac{(\varphi t)^n}{n!} \sum_{l=0}^{n} \binom{n}{l} x_h^l (1 - x_h)^{n-l} C^{(h)}(n,l) + e(N_{min})$$

where $e_{i,j}(N_{min}) \leq \varepsilon$, for every $i, j \in S$.

Bibliography

[1] M. L. Sichitiu and M. Kihl, "Inter-vehicle communication systems: a survey," *IEEE Communications Surveys & Tutorials,* vol. 10, pp. 88-105, 2008.

[2] X. Yang, L. Liu, N. H. Vaidya and F. Zhao, "A vehicle-to-vehicle communication protocol for cooperative collision warning," *Proceedings of the First Annual International Conference on Mobile and Ubiquitous Systems (MOBIQUITOUS),* 2004.

[3] Q. Xu, T. Mak, J. Ko and R. Sengupta, "Medium access control protocol design for vehicle–vehicle safety messages," *IEEE Transactions on Vehicular Technology,* vol. 56, pp. 499-518, 2007.

[4] A. Nandan, S. Das, B. Zhou, G. Pau and M. Gerla, "AdTorrent: digital billboards for vehicular networks," *Proceedings of IEEE/ACM International Workshop on Vehicle-to-Vehicle Communications (V2VCOM),* 2005.

[5] J. Ott and D. Kutscher, "Drive-thru Internet: IEEE 802.11b for "automobile" users," *Proceedings of the 23th IEEE International Conference on Computer Communications (INFOCOM),* 2004.

[6] A. Balasubramanian, B. N. Levine and A. Venkataramani, "Enhancing interactive web applications in hybrid networks," *Proceedings of the 14th ACM International Conference on Mobile Computing and Networking (Mobicom),* 2008.

[7] W. L. Jin and W. W. Recker, "Instantaneous information propagation in a traffic stream through inter-vehicle communication," *Transportation Research Part B,* vol. 40, pp. 230-250, 2006.

[8] H. Wu, R. Fujimoto and G. Riley, "Analytical models for information propagation in vehicle-to-vehicle networks," *Proceedings of the IEEE 60th Vehicular Technology Conference (VTC Fall),* 2004.

[9] S. Yousefi, E. Altman, R. El-Azouzi and M. Fathy, "Analytical model for connectivity in vehicular ad hoc networks," *Research Report, the LIA, Université d'Avignon,* vol. Available at: http://www.lia-avignon.fr/fich_art/1004-VanetConnectivity.pdf, 2007.

[10] L. Briesemeister, L. Schafers and G. Hommel, "Disseminating messages among highly mobile hosts based on inter-vehicle communication," *Proceedings of the IEEE Intelligent Vehicles Symposium,* 2000.

[11] L. Wischhof, A. Ebner and H. Rohling, "Information dissemination in self-organizing intervehicle networks," *IEEE Transactions on Intelligent Transportation Systems,* vol. 6, pp. 90-101, 2005.

[12] H. Wu and R. Fujimoto, "MDDV: a mobility-centric data dissemination algorithm for vehicular networks," *Proceedings of the First ACM International Workshop on Vehicular Ad Hoc Networks (VANET),* 2004.

[13] J. Zhao and G. Cao, "VADD: vehicle-assisted data delivery in vehicular ad hoc networks," *IEEE Transactions on Vehicular Technology,* vol. 57, pp. 1910-1922, 2008.

[14] H. Menouar, F. Filali and M. Lenardi, "A survey and qualitative analysis of mac protocols for vehicular ad hoc networks," *IEEE Wireless Communications,* vol. 13, pp. 30-35, 2006.

[15] Y. Zang, L. Stibor, G. R. Hiertz and H. J. Reumerman, "Vehicular wireless media network (VWMN): a distributed broadband MAC for inter-vehicle communications," *Proceedings of the 2nd ACM International Workshop on Vehicular Ad Hoc Networks (VANET),* 2005.

[16] J. Zhu and S. Roy, "MAC for dedicated short range communications in intelligent transport system," *IEEE Communications Magazine,* vol. 41, pp. 60-67, 2003.

[17] V. Dumitrescu and J. Guo, "Context assisted routing protocols for inter-vehicle wireless communication," *Proceedings of the IEEE Intelligent Vehicles Symposium,* 2005.

[18] C. Lochert, H. Hartenstein, J. Tian, H. Fussler, D. Hermann and M. Mauve, "A routing strategy for vehicular ad hoc networks in city environments," *Proceedings of the IEEE Intelligent Vehicles Symposium,* 2003.

[19] C. Lochert, M. Mauve, H. Füßler and H. Hartenstein, "Geographic routing in city scenarios," *ACM SIGMOBILE Mobile Computing and Communications Review,* vol. 9, pp. 69-72, 2005.

[20] Z. Mo, H. Zhu, K. Makki and N. Pissinou, "MURU: a multi-hop routing protocol for urban vehicular ad hoc networks," *Proceedings of the Third Annual International Conference on Mobile and Ubiquitous Systems (MOBIQUITOUS),* 2006.

[21] V. Naumov and T. R. Gross, "Connectivity-aware routing (CAR) in vehicular ad-hoc networks," *Proceedings of the 26th IEEE International Conference on Computer Communications (Infocom),* 2007.

[22] J. Tian, L. Han, K. Rothermel and C. Cseh, "Spatially aware packet routing for mobile ad hoc inter-vehicle radio networks," *Proceedings of the IEEE Intelligent Transportation System Conference,* 2003.

[23] N. Wisitpongphan, F. Bai, P. Mudalige, V. Sadekar and O. Tonguz, "Routing in sparse vehicular ad hoc wireless networks," *IEEE Journal on Selected Areas in Communications,* vol. 25, pp. 1538-1556, 2007.

[24] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine,* vol. 40, pp. 102-114, 2002.

[25] W. Ye, J. Heidemann and D. Estrin, "Medium access control with coordinated adaptive sleeping for wireless sensor networks," *IEEE/ACM Transactions on Networking,* vol. 12, pp. 493-506, 2004.

[26] T. van Dam and K. Langendoen, "An adaptive energy-efficient MAC protocol for wireless sensor networks," *Proceedings of the 1st International Conference on Embedded Networked Sensor Systems (SenSys),* 2003.

[27] V. Rajendran, K. Obraczka and J. J. Garcia-Luna-Aceves, "Energy-efficient, collision-free medium access control for wireless sensor networks," *ACM/Kluwer Wireless Networks,* vol. 12, pp. 63-78, 2006.

[28] L. Campelli, A. Capone, M. Cesana and E. Ekici, "A receiver oriented MAC protocol for wireless sensor networks," *Proceedings of the IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS),* 2007.

[29] I. Rhee, A. Warrier, M. Aia, J. Min and M. L. Sichitiu, "Z-MAC: A hybrid MAC for wireless sensor networks," *IEEE/ACM Transactions on Networking,* vol. 16, pp. 511-524, 2008.

[30] W. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy-efficient communication protocol for wireless sensor networks," *Proceedings of 33rd Hawaii International Conference on System Science (HICSS),* 2000.

[31] W. R. Heinzelman, J. Kulik and H. Balakrishnan, "Adaptive protocols for information dissemination in wireless sensor networks," *Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom),* 1999.

[32] C. Intanagonwiwat, R. Govindan and D. Estrin, "Directed diffusion: a scalable and robust communication paradigm for sensor networks," *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking (MobiCom),* 2000.

[33] R. C. Shah and J. M. Rabaey, "Energy aware routing for low energy ad hoc sensor networks," *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC),* 2002.

[34] Y. Xu, J. Heidemann and D. Estrin, "Geography-informed energy conservation for ad hoc routing," *Proceedings of the 7th Annual International Conference on Mobile Computing and Networking (MobiCom),* 2001.

[35] C. Alberto and E. Deborah, "ASCENT: adaptive self-configuring sensor networks topologies," *IEEE Transactions on Mobile Computing,* vol. 3, pp. 272-285, 2004.

[36] S. Bandyopadhyay and E. J. Coyle, "An energy efficient hierarchical clustering algorithm for wireless sensor networks," *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM),* 2003.

[37] C. Hsin and M. Liu, "Network coverage using low duty-cycled sensors: random & coordinated sleep algorithms," *Proceedings of the Third International Symposium on Information Processing in Sensor Networks (IPSN),* 2004.

[38] A. Karnik and A. Kumar, "Distributed optimal self-organisation in a class of wireless sensor networks," *Proceedings of the 23rd AnnualJoint Conference of the IEEE Computer and Communications Societies (INFOCOM),* 2004.

[39] C. Schurgers, V. Tsiatsis, S. Ganeriwal and M. Srivastava, "Optimizing sensor networks in the energy-latency-density design space," *IEEE Transactions on Mobile Computing,* vol. 1, pp. 70-80, 2002.

[40] S. J. Park, R. Vedantham, R. Sivakumar and I. F. Akyildiz, "A scalable approach for reliable downstream data delivery in wireless sensor networks," *Proceedings of the 5th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc),* 2004.

[41] Y. Sankarasubramaniam, Ö B. Akan and I. F. Akyildiz, "ESRT: event-to-sink reliable transport in wireless sensor networks," *Proceedings of the 4th ACM International Symposium on Mobile Ad Hoc Networking & Computing (MobiHoc),* 2003.

[42] C. Y. Wan, A. T. Campbell and L. Krishnamurthy, "Pump-slowly, fetch-quickly (PSFQ): a reliable transport protocol for sensor networks," *IEEE Journal on Selected Areas in Communications,* vol. 23, pp. 862-872, 2005.

[43] K. Pentikousis, M. Palola, M. Jurvansuu and P. Perala, "Active goodput measurements from a public 3G/UMTS network," *IEEE Communications Letters,* vol. 9, pp. 802-804, 2005.

[44] P. Reichl, M. Umlauft, J. Fabini, R. Lauster and G. Pospischil, "Project WISQY: A measurement-based end-to-end application-level performance comparison of 2.5G and 3G networks," *Proceedings of the 4th Annual Wireless Telecommunications Symposium,* 2005.

[45] M. Kohlwes, J. Riihijarvi and P. Mahonen, "Measurements of TCP performance over UMTS networks in near-ideal conditions," *Proceedings of the 61st IEEE Vehicular Technology Conference (VTC Spring),* 2005.

[46] J. M. Cano-Garcia, E. Gonzalez-Parada and E. Casilari, "Experimental analysis and characterization of packet delay in UMTS networks," *Proceedings of the 6th International Conference on Next-Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN),* 2006.

[47] A. Mäder, B. Wagner, T. Hoßfeld, D. Staehle and H. Barth, "Measurements in a laboratory UMTS network with time-varying loads and different admission control strategies," *Proceedings of the 4th International Workshop on Internet Performance, Simulation, Monitoring and Measurement,* 2006.

[48] J. S. Evans and D. Everitt, "On the teletraffic capacity of CDMA cellular networks," *IEEE Transactions on Vehicular Technology,* vol. 48, pp. 153-165, 1999.

[49] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver Jr. and C. E. Wheatley III, "On the capacity of a cellular CDMA system," *IEEE Transactions on Vehicular Technology,* vol. 40, pp. 303-312, 1991.

[50] I. Koo, J. H. Ahn, J. A. Lee and K. Kim, "Analysis of Erlang capacity for the multimedia DS-CDMA systems," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences,* vol. 82, pp. 849-855, 1999.

[51] V. K. Paulrajan, J. A. Roberts and D. L. Machamer, "Capacity of a CDMA cellular system with variable user data rates," *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM),* 1996.

[52] A. M. Viterbi and A. J. Viterbi, "Erlang capacity of a power controlled CDMA system," *IEEE Journal on Selected Areas in Communications,* vol. 11, pp. 892-900, 1993.

[53] J. R. Yang, Y. Y. Choi, J. H. Ahn and K. Kim, "Capacity plane of CDMA system for multimedia traffic," *Electronics Letters,* vol. 33, pp. 1432-1433, 1997.

[54] V. Bychkovsky, B. Hull, A. Miu, H. Balakrishnan and S. Madden, "A measurement study of vehicular internet access using in situ Wi-Fi networks," *Proceedings of the 12th Annual International Conference on Mobile Computing and Networking (Mobicom),* 2006.

[55] D. Hadaller, S. Keshav, T. Brecht and S. Agarwal, "Vehicular opportunistic communication under the microscope," *Proceedings of the 5th International Conference on Mobile Systems, Applications and Services (Mobisys)*, 2007.

[56] J. Ott and D. Kutscher, "A disconnection-tolerant transport for drive-thru Internet environments," *Proceedings of the 24th IEEE International Conference on Computer Communications (INFOCOM)*, 2005.

[57] A. Balasubramanian, R. Mahajan, A. Venkataramani, B. Levine and J. Zahorjan, "ViFi: Interactive WiFi connectivity for moving vehicles," *Proceedings of ACM SIGCOMM*, 2008.

[58] R. Mahajan, J. Zahorjan and B. Zill, "Understanding WiFi-based connectivity from moving vehicles," *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, 2007.

[59] J. Eriksson, H. Balakrishnan and S. Madden, "Cabernet: vehicular content delivery using WiFi," *Proceedings of the 14th ACM International Conference on Mobile Computing and Networking (Mobicom)*, 2008.

[60] N. Banerjee, M. D. Corner and B. N. Levine, "An energy-efficient architecture for DTN throwboxes," *Proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM)*, 2007.

[61] A. Balasubramanian, B. Levine and A. Venkataramani, "DTN routing as a resource allocation problem," *Proceedings of ACM SIGCOMM*, 2007.

[62] X. Zhang, J. K. Kurose, B. N. Levine, D. Towsley and H. Zhang, "Study of a bus-based disruption-tolerant network: mobility modeling and impact on routing," *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking*, 2007.

[63] N. Banerjee, M. D. Corner, D. Towsley and B. N. Levine, "Relays, base stations, and meshes: enhancing mobile networks with infrastructure," *Proceedings of the 14th ACM International Conference on Mobile Computing and Networking*, 2008.

[64] K. K. Leung, W. A. Massey and W. Whitt, "Traffic models for wireless communication networks," *IEEE Journal on Selected Areas in Communications*, vol. 12, pp. 1353-1364, 1994.

[65] W. A. Massey and W. Whitt, "A stochastic model to capture space and time dynamics in wireless communication systems," *Probability in the Engineering and Information Sciences*, vol. 8, pp. 541–569, 1994.

[66] N. Antunes, A. Pacheco and R. Rocha, "Traffic modelling for broadband wireless networks: the highway scenario," *Advances in Performance Analysis,* vol. 3, pp. 109-136, 2000.

[67] H. Kobayashi, S. Z. Yu and B. L. Mark, "An integrated mobility and traffic model for resource allocation in wireless networks," *Proceedings of the 3rd ACM International Workshop on Wireless Mobile Multimedia (WoWMoM),* 2000.

[68] T. J. Ott, "Sojourn-time distribution in the M/G/1 queue with processor sharing," *Journal of Applied Probability,* vol. 21, pp. 360-378, 1984.

[69] R. M. Smith, K. S. Trivedi and A. V. Ramesh, "Performability analysis: measures, an algorithm, and a case study," *IEEE Transactions on Computers,* vol. 37, pp. 406-417, 1988.

[70] E. S. e Silva, H. R. Gail and R. V. Campos, "Calculating transient distributions of cumulative reward," *ACM SIGMETRICS Performance Evaluation Review,* vol. 23, pp. 231-240, 1995.

[71] B. Sericola, "Occupation times in Markov processes," *Stochastic Models,* vol. 16, pp. 479-510, 2000.

[72] M. Bladt, B. Meini, M. F. Neuts and B. Sericola, "Distributions of reward functions on continuous-time Markov chains," *Proceedings of the Fourth International Conference on Matrix-Analytic Methods: Theory and Applications,* 2002.

[73] A. Klar, R. D. Kuehne and R. Wegener, "Mathematical models for vehicular traffic," *Surveys on Mathematics for Industry,* vol. 6, pp. 215-239, 1996.

[74] M. A. P. Taylor, W. Young and P. W. Bonsall, *Understanding Traffic Systems: Data, Analysis and Presentation.* Avebury Technical, 1996.

[75] J. D. Fricker and R. K. Whitford, *Fundamentals of Transportation Engineering: A Multimodal Systems Approach.* Prentice Hall, 2004.

[76] B. D. Greenshields, "A study of traffic capacity," *Highway Research Board Proceedings,* vol. 14, pp. 448-477, 1935.

[77] T. van Woensel and N. Vandaele, "Modeling traffic flows with queueing models: a review," *Asia Pacific Journal of Operational Research,* vol. 24, pp. 435, 2007.

[78] D. J. Buckley, "Road traffic headway distributions." *Australian Road Research Board Proceedings,* vol. 1, pp. 153-187, 1962.

[79] L. C. Edie, "Traffic delays at toll booths," *Journal of the Operations Research Society of America,* vol. 2, pp. 107-138, 1954.

[80] R. Jain and J. M. Smith, "Modeling vehicular traffic flow using M/G/C/C state dependent queueing models," *Transportation Science,* vol. 31, pp. 324-336, 1997.

[81] A. J. Miller, "A queueing model for road traffic flow," *Journal of Royal Statistical Society, Series B (Methodological),* vol. 23, pp. 64-75, 1961.

[82] T. V. Woensel and N. Vandaele, "Empirical validation of a queueing approach to uninterrupted traffic flows," *4OR: A Quarterly Journal of Operations Research,* vol. 4, pp. 59-72, 2006.

[83] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory.* ,3rd ed.John Wiley & Sons, 1998.

[84] S. M. Ross, *Stochastic Processes.* ,2nd ed.John Wiley & Sons, 1996.

[85] G. Anastasi, E. Borgia, M. Conti and E. Gregori, "Wi-fi in ad hoc mode: a measurement study," *Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications (PerCom),* 2004.

[86] J. Ai, J. Kong and D. Turgut, "An adaptive coordinated medium access control for wireless sensor networks," *Proceedings of the International Symposium on Computers and Communications,* 2004.

[87] P. Lin, C. Qiao and X. Wang, "Medium access control with a dynamic duty cycle for sensor networks," *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC),* 2004.

[88] C. K. Nguyen and A. Kumar, "A wireless-sensor-network energy-efficient medium-access-control protocol with overhearing avoidance," *Proceedings of 3rd IEEE Consumer Communications and Networking Conference (CCNC),* 2006.

[89] L. Gang, B. Krishnamachâri and C. Raghavendra, "An adaptive energy-efficient and low-latency MAC for data gathering in sensor networks," *Proceedings of the 4th International Workshop on Algorithms for Wireless, Mobile, Ad Hoc and Sensor Networks (WMAN 04),* 2004.

[90] T. Zheng, S. Radhakrishnan and V. Sarangan, "PMAC: an adaptive energy-efficient MAC protocol for wireless sensor networks," *Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium,* 2005.

[91] K. Balachandran, J. H. Kang and W. C. Lau, "Adaptive sleeping and awakening protocol (ASAP) for energy efficient adhoc sensor networks," *Proceedings of IEEE International Conference on Communications (ICC),* 2005.

[92] G. Halkes and K. Langendoen, "Crankshaft: an energy-efficient MAC protocol for dense wireless sensor networks," *Proceedings of European Conference on Wireless Sensor Networks,* 2007.

[93] Z. Chen and A. Khokhar, "Self organization and energy efficient TDMA MAC protocol by wake up for wireless sensor networks," *Proceedings of the 1st IEEE Sensor and Ad Hoc Communications and Networks (SECON) Conference,* 2004.

[94] L. F. W. van Hoesel and P. J. M. Havinga, "A lightweight, medium access protocol (LMAC) for wireless sensor networks," *Proceedings of the 2004 International Conference on Networked Sensing Systems (INSS),* 2004.

[95] J. Polastre, J. Hill and D. Culler, "Versatile low power media access for wireless sensor networks," *Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems (SenSys),* 2004.

[96] J. Elson and D. Estrin, "Time Synchronization for Wireless Sensor Networks," *Proceedings of the 15th IEEE International Parallel & Distributed Processing Symposium,* 2001.

[97] S. Ganeriwal, R. Kumar and M. B. Srivastava, "Timing-sync protocol for sensor networks," *Proceedings of the 1st International Conference on Embedded Networked Sensor Systems (SenSys),* 2003.

[98] M. J. McGlynn and S. A. Borbash, "Birthday protocols for low energy deployment and flexible neighbor discovery in ad hoc wireless networks," *Proceedings of the 2nd ACM International Symposium on Mobile Ad Hoc Networking & Computing (MobiHoc),* 2001.

[99] I. Rhee, A. Warrier, J. Min and L. Xu, "DRAND: distributed randomized TDMA scheduling for wireless ad-hoc networks," *Proceedings of the 7th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc),* 2006.

[100] S. Lam, "Delay analysis of a time division multiple access (TDMA) channel," *IEEE Transactions on Communications,* vol. 25, pp. 1489-1494, 1977.

[101] P. D. Welch, "On a generalized M/G/1 queueing process in which the first customer of each busy period receives exceptional service," *Operations Research,* vol. 12, pp. 736–752, 1964.

[102] A. Ephremides and O. A. Mowafi, "Analysis of a hybrid access scheme for buffered users-probabilistic time division." *IEEE Transactions on Software Engineering,* vol. 8, pp. 52-60, 1982.

[103] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications.*3rd ed.John Wiley & Sons, 2004.

[104] "High-Speed Downlink Packet Access (HSDPA) Overall Description Stage 2," *TS 25.308 V 5.6.0,* Sept. 2004.

[105] V. V. Veeravalli and A. Sendonaris, "The coverage-capacity tradeoff in cellular CDMA systems," *IEEE Transactions on Vehicular Technology,* vol. 48, pp. 1443-1450, 1999.

[106] J. Border, M. Kojo, J. Griner, G. Montenegro and Z. Shelby, "Performance enhancing proxies intended to mitigate link-related degradations," *IETF RFC 3135,* June 2001.

[107] H. Holma and J. Reunanen, "3GPP Release 5 HSDPA measurements," *Proceedings of the IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC),* 2006.

[108] S. Yoon, H. Q. Ngo and C. Qiao, "On shooting a moving vehicle with data flows," *Proceedings of the IEEE Workshop on Mobile Networking for Vehicular Environments,* 2007.

[109] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* 2001.

[110] L. Carvalho, E. Mota, R. Aguiar, A. F. Lima, J. N. de Souza and A. Barreto, "An E-Model implementation for speech quality evaluation in VoIP systems," *Proceedings of the 10th IEEE Symposium on Computers and Communications (ISCC),* 2005.

[111] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing,* vol. 13, pp. 600-612, 2004.

[112] Z. Wang, L. Lu and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication,* vol. 19, pp. 121-132, 2004.

[113] S. Channappayya, K. Seshadrinathan and A. C. Bovik, "Video quality assessment with motion and temporal artifacts considered," *EE Times,* vol. Available at http://users.ece.utexas.edu/~kalpana/pub/ks_eetimes07.pdf, December 2007.

[114] R. L. Graham, D. E. Knuth and O. Patashnik, *Concrete Mathematics.* Addison-Wesley, 1989.

[115] I. G. Petrovsky, *Lectures on Partial Differential Equations. Translated by A. Shenitzer.* Interscience Publishers Inc, 1954.