UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

# VOICE OVER IP NETWORKS: QUALITY OF SERVICE, PRICING AND SECURITY

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

DOCTOR OF PHILOSOPHY

By

LING WANG
Norman, Oklahoma
2008

VOICE OVER IP NETWORKS: QUALITY OF SERVICE, PRICING AND SECURITY

A DISSERTATION APPROVED FOR THE
SCHOOL OF ELECTICAL AND COMPUTER ENGINEERING

BY

Dr. Pramode K. Verma, Chair

Dr. James J. Sluss, Jr.

Dr. Hazem H. Refai

Dr. Monte P. Tull

Dr. William O. Ray

## Acknowledgments

I would like to express my sincere gratitude to my advisor, Dr. Pramode Verma, for his continuous support and inestimable guidance, and also friendship during my Master and PhD studies. I am also very grateful to my respected committee members, Dr. James Sluss, Dr. Refai Hazem, Dr. Monte Tull and Dr. William Ray, for their comments and suggestions concerning this dissertation.

## Table of Contents

## List of Illustrations

## List of Tables

## Abstract

The growth of the Internet over the past decade together with the promise of lower costs to the customer has led to the rapid emergence of Voice over Internet Protocol (VoIP). As a real-time application in large scale packet switched networks, VoIP networks face many challenges such as availability, voice quality and network security. This dissertation addresses three important issues in VoIP networks: Quality of Service, pricing and security.

In addressing Quality of Service (QoS), this dissertation introduces the notion of delay not exceeding an upper limit, termed the bounded delay (rather than the average delay), to measure the Quality of Service in VoIP networks. Queuing models are introduced to measure performance in terms of bounded delays. Closed form solutions relating the impact of bounding delays on throughput of VoIP traffic are provided. Traffic that exceeds the delay threshold is treated as lost throughput. The results addressed can be used in scaling resources in a VoIP network for different thresholds of acceptable delays. Both single and multiple switching points are addressed. The same notion and analysis are also applied on jitter, another important indicator of the VoIP QoS

This dissertation also develops a pricing model based on the Quality of Service provided in VoIP networks. It presents the impact of quality of VoIP service demanded by the customer on the transmission resources required by the network using an analytical approach. The price to be paid by the customer is based on the throughput meeting this criterion and the network transmission resources required. In particular, the impact of Quality of Service presented can be used in the design of

VoIP networks in a way that would provide fairness to the user in terms of quality of service and price while optimizing the resources of the network at the same time.

This dissertation also extends and applies the delay throughput analysis developed for VoIP networks in assessing the impact of risks constituted by a number of transportation channels, where the risk associated with each channel can be quantified by a known distribution. For VoIP security, this dissertation mainly focuses on the signaling authentication. It presents a networking solution that incorporates network-based authentication as an inherent feature. The authentication feature that we propose introduces a range of flexibilities not available in the PSTN. Since most calls will likely terminate on the network of another service provider, we also present a mechanism using which networks can mutually authenticate each other to afford the possibility of authentication across networks. Finally, this dissertation explores areas for future research that can be built on the foundation of research presented.

# Chapter 1. Introduction

**Abstract:** This chapter introduces Voice over IP networking and the three major driving forces leading to its rapid development. In comparison to the traditional calls through PSTN, VoIP has several advantages. However, VoIP networks face many challenges as well, including quality of service and security. This chapter concludes by identifying the scope and contributions of this dissertation in three aspects of VoIP networks: Quality of Service, pricing and security.

## 1.1 Motivation for VoIP Networks

Telephone systems have evolved over the last century. This evolution has comprised moving from analog to digital systems and from the circuit switching technology to packet switching systems. More recently, the potential benefits of converged networks in addition to lower costs for the customer has led to increasing use of Voice over IP [1].

VoIP emerges as one of the most disruptive services in the telecommunications industry driven by a large available market, high availability of broadband access, and consumer's willingness to accept the new technology to cut telephony costs while potentially reaping the benefits of a converged network.

As one example, Verizon, a major telecommunications service provider, stood out in 3Q:05 with 7.5 percent year-over-year decline in the number of its PSTN customers, which we suspect highlights the competitive losses to VoIP services from

multiple systems operators ( or MSOs, e.g., Comcast, Time Warner, and Cablevision) in Verizon's Northeast markets [2]. Cisco also shipped over 980,000 residential VoIP gateways in 1Q 2005 equivalent to a 35% growth year- over-year).

Figure 1.1 shows the surge in the U.S VoIP residential subscribers in the first three quarters of 2005, indicating that the VoIP market will grow aggressively. In other words, it can be seen that carriers are losing more and more primary lines to VoIP substitutions.

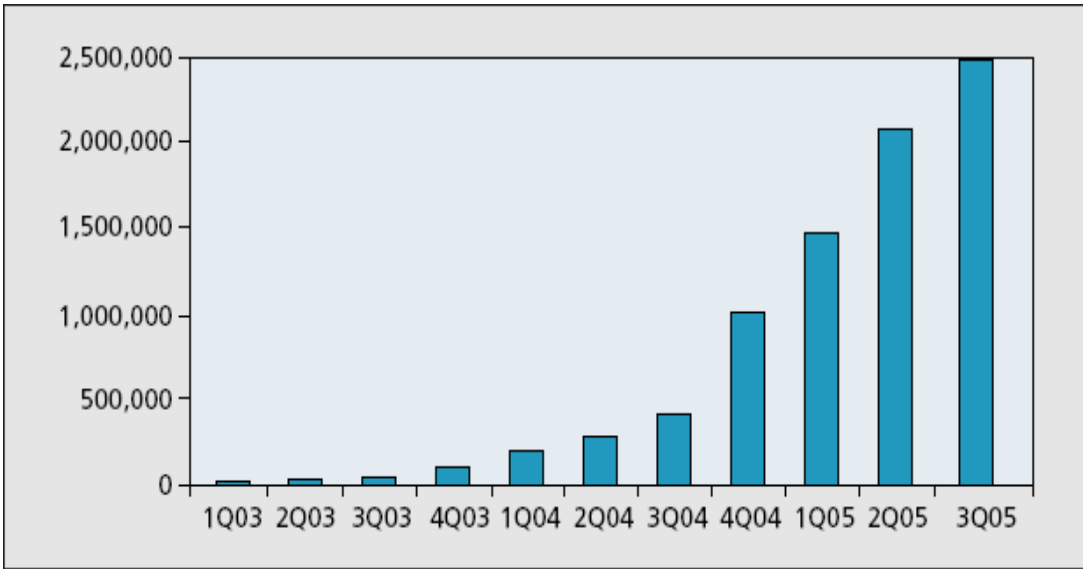


**Figure 1. 1: U.S. VoIP residential subscribers [3]**

According to the Telecommunications Industry Association, in Arlington, Va., in 2006, more than half of all the new private branch exchanges installed were IP based. And the number of residential VoIP subscribers is expected to rise 12-fold, to about 12 million, by 2009, industry analysts project. By that time, the total U.S. revenue for business and residential VoIP products and services will be nearly $21 billion, up from $2.5 billion today, says Aaron Nutt, an analyst at Atlantic-ACM, a

unit of Boston-based ACM Group Inc., which specializes in telecommunications consulting and market research [4].

The keen interest of the telecommunication industry in VoIP telephony is overwhelming, in spite of its relatively poor voice quality and lower overall availability of the Internet to support voice traffic compared to the traditional circuit-switched telephony. Three major factors could explain such interest and the rapid development and deployment of the VoIP services. These are addressed as follows:

- Business Drive
- Technology innovation
- Customer Comfort

### 1.1.1 Business Drive

The primary driving factor of VoIP networks is the business benefits, which are summarized as follows,

- **Consolidated voice and data network expenses**

A single integrated telecommunications network with a common switching and transmission system is created to carry both the data traffic and the voice traffic. The integration of data and voice use the bandwidth and the equipment more efficiently. Voice packets share bandwidth among multiple logical connections alongside the data packets.

In traditional circuit-switched telephony system like the PSTN, in order to combine different 64-kbps channels into high-speed upper links for transmission, a

substantial amount of equipment is needed. In the packet-switched IP telephony system such as VoIP, in order to integrate the voice traffic and data traffic, the IP network utilizes statistical multiplexing. This consolidation represents substantial savings on capital equipment and operations costs [5].

- **Increased revenues from new services**

VoIP networks not only support real-time voice communication, but also enable new services, such as instant messaging, unified messaging (voicemail to email), video conferencing and distributed games. These new services differentiate companies and service providers in their respective markets. Simultaneously, these new services encourage employees of the businesses using such services to improve productivity and, therefore, profit margins.

- **Flexible pricing structures**

The establishments of VoIP networks push the service providers to build new pricing models. Since the IP network is shared by voice and data, and the bandwidth is dynamically allocated, the resource consumption of the network is not, generally, measured in duration or distance as circuit-switched telephony. Dynamic allocation gives service providers the flexibility to meet the needs of their customers in ways that bring them the greatest benefits [5].

### 1.1.2 Technology Innovation

The second major reason for the rapid growth of VoIP networks is the development of technologies and new innovations, which make the VoIP networks feasible.

As we know, the main challenge of VoIP networks is that they could not provide the same Quality of Service as PSTN does. However, the maturation of various codecs (voice coders and decoders) and high speed Digital Signal Processors that perform voice packetization and compression, greatly improve the voice quality over the IP infrastructure.

Another aspect of technology innovations is the rapid emergence of new user applications and access to new communications devices, such as wireless devices, videophones, multimedia terminals, and personal digital assistants (PDAs). Upon demand by customers, service providers can much more easily enable new devices, offer larger volume of communications, and serve more subscribers in a packet switched environment. Further, VoIP services have been aggressively marketed as a component of a compelling voice/data/video services bundle.

### 1.1.3 Customer Comfort

As we have mentioned above, access to new devices and advanced applications will satisfy the customer and make their communication much more convenient and effective. VoIP has the potential to provide almost equal voice quality but multi-function service when compared to the traditional PSTN service, without increasing the price the customer pays. In VoIP networks, long-distance charges—

including charges for international calls—can be transformed into a flat monthly fee or, in the case of the advertising-supported service providers, eliminated entirely [6]. Indeed, the flat-fee charging model already exists for the Internet where most users pay a fixed monthly fee independent of the number of bits they send or receive. The amount of information transmitted or received is limited only by the access speed [7]. We might say that the potential savings on long-distance costs was one of the driving forces behind the migration to converged voice and data networks [5].

## 1.2 Comparison of PSTN and VoIP Networks

Over the last 100 years, voice services have been provided through PSTN impressively. The PSTN core network uses 64 kb/s digital channels to provide dedicated end-to-end circuit connections in each direction. Voice terminals have both analog and digital access to the PSTN. Analog and digital telephone sets are generally connected to the central office with copper wire, called subscriber loops. Digital terminals use the ISDN protocols to access the network services.

In contrast to PSTN, VoIP technology digitizes voice and transmits data in frames over IP networks, where the reception of packets is not guaranteed because it relies on best-effort transport architecture. VoIP is rapidly establishing itself as an attractive technology for telephony with the maturation of related technologies and the recovery of the telecommunication industry.

PSTN has made very impressive achievement in terms of coverage, reliability, and ease of use. The number of current telephones lines is estimated to be 1 billion. People can hear the dial tone whenever they pick up the phone, and they expect to be connected to any destination in less than a minute. The availability of telephone

service in such plain old telephone system (POTS) is 99.999%, also referred to as a five-nine's reliability.

At the present time, the Internet does not offer the same degree of reliability as the PSTN due to a variety of reasons: The complexity of multiple protocols, lack of standardization, multiplicity of equipment vendors and service providers, varying operating systems and network management systems, can cause lack of end-to-end interoperability. Also, packet switched networks experience variable delays in the transmission process. In contrast, the PSTN doesn't suffer from variable delays, although it can experience blocking when all the available circuits are being used by other calls. The public Internet is collectively available only 61% of the time [8]. The best private data networks are available about 94% of the time, on average, meaning that a user can be without the digital equivalent of dial tone about 22 days per year [9].

The PSTN-based Emergency-911 (E-911) services report the exact location of the telephone. Increasingly, this service is also required in VoIP. The location of a VoIP user is obtained by updating in the E911 database. VOIP has not yet been regulated because IP-based telephony services are not regarded as traditional telephony services. [10].

## 1.3 Major Challenges in VoIP Networks

### 1.3.1 Quality of Service

Quality of Service is one of the most important concerns in voice communications [11], determined by many factors, such as packet loss, speech coding options, delay, echo and jitter. The connection-oriented, circuit-switched network

provide each user with dedicated bandwidth for the duration of each call, which results in extremely low delay and jitter, minimum disruption due to “noise” on the connections. High quality provided by the PSTN and private PBX-based networks drives telephone users to expect high QoS of the VOIP [12-13].

As we know, VOIP uses different codecs to interwork, and codecs affect the quality of voice in a significant way, so it is especially important to measure the quality of a voice call in a standardized manner. One such measure is through the Mean Opinion Score (MOS) [Ref.]. According to ITU-T, opinion rating is generally used to assess subjective quality, which is the measurement based on a large number of users’ perception of service quality under various conditions. One of the most frequently used opinion scale is shown in the following table [14].

| MOS | Opinion | Description |
|---|---|---|
| 5 | Excellent | Greater than Toll Quality |
| 4 | Good | Toll Quality |
| 3 | Fair | Mobile Phone Quality |
| 2 | Poor | - |
| 1 | Bad | - |

**Table 1. 1: ACR or MOS Scale**

Generally speaking, VOIP systems should achieve toll quality or near toll quality, low delay, and have good resilience to packet loss but a lower cost [15].

### 1.3.2 Pricing

An important issue in designing pricing policies for today’s networks is to balance the trade-off between traffic engineering and economic efficiency [16, 17]. A recent work [18] has addressed the impact of multiple hops (or switches between the ingress and egress switching nodes) on the grade of service offered by a circuit switched telephone network. A similar approach is adopted in the context of packet

switching. Accordingly, the grade of service is replaced by the threshold delay which is an appropriate measure for perceived Quality of Service in a packet switched network. In a circuit switched network, an incomplete call is lost and does not generate any revenue. In the packet switched networks, there are no calls that are lost as such; however, some of the packets may suffer delays above the acceptable delay bound and are, similarly, not considered to provide effective throughput. Just as a caller in a circuit switched network does not pay for an incomplete call, the VoIP caller over a packet switched network in our construct does not pay for packets that suffer an unacceptable level of delay. The new pricing scheme proposed is based on the cost of lost opportunity vs. the cost of consumption of resources, which is the contemporary practice.

### 1.3.3 Security

Voice over IP applications are generally designed to function over the global Internet, although such solutions can be offered over private IP networks, such as enterprise networks. Instances of violations of security over the Internet are common occurrences that affect individuals, businesses as well as government operations. Voice over IP has not yet suffered many security violations, but the potential for attacks on security is truly large [19]. Possibly the mass of end points connected to VoIP today is below the threshold that would attract miscreants. However, faking the identity of the caller can be easily accomplished using the Internet. This can result in the unsuspecting callee passing sensitive information to the caller. Additionally, the negative impact of SPAM over Internet Telephony can be easily comprehended.

## 1.4 Scope and Contribution of the Dissertation

Delay is the most important parameter in voice communications due its effect on interactivity for real-time applications. The average delay has been traditionally used as a key measure of the network performance. However, from a user's perspective, the upper bound of delay is far more important. The notion of delay not exceeding an upper limit, termed the bounded delay (rather than the average delay), is introduced as a measure of the Quality of Service in VoIP networks. Closed form solutions relating the impact of bounding delays to specified levels on throughput of VoIP traffic is developed, which can be used in scaling resources in a VoIP network for different thresholds of bounded delays. Similar results are presented for jitter in a multi-hop network.

Analytical results are developed relating the quality of VoIP service to transmission resources required by the network. The price to be paid by the customer is based on the throughput meeting this criterion and the network transmission resources required. The results provide fairness to the user in terms of quality of service and price, while optimizing the resources of the network at the same time.

Risk Analysis models are developed which can assess the cumulative impact of risk constituted by a number of channels, where the risk associated with each channel can be quantified by a known distribution. Further, a mechanism using which VoIP networks can mutually authenticate each other to afford authentication across networks is presented in this dissertation. This proposed authentication scheme introduces a range of flexibilities not available in the PSTN.

## 1.5 Organization of the Dissertation

This dissertation consists of three main parts. The first part (Chapters 2 to 5) deals with the Voice over IP networks. Chapter 2 gives a thorough introduction to VoIP networks, including voice quality, transport, Network QoS, call signaling and security. Chapter 3 provides an overview of the applicable queuing theory and presents an analytical model for delay-throughput analysis. Chapters 4 and 5 provide closed form solutions relating the impact of bounding delays on throughput for M/M/1 and M/D/1 model respectively. Detailed analytical solutions and simulation results are presented as well.

The second part of the dissertation (Chapter 6) addresses the impact of Quality of Service on resource consumption and proposes a pricing scheme whereby the customers are fairly charged for various levels of QoS. In particular, the impact of Quality of Service presented can be used in the design of VoIP networks in a way that would provide fairness to the user in terms of quality of service and price while optimizing the resources of the network at the same time.

The third part of the dissertation focuses on the VoIP network security. Chapter 7 reviews the potential attacks and current countermeasures on VoIP networks, including related algorithms and techniques. Chapter 8 presents a new authentication scheme that incorporates network-based authentication as an inherent feature, which supports a range of flexibilities not available in the PSTN. A mechanism using which networks can mutually authenticate each other to afford the possibility of authentication across networks is also presented.

# Chapter 2. Voice over Internet Protocol

**Abstract**: This chapter deals with the technical aspects of Voice over Internet Protocol (VoIP). First, the chapter presents an overview of the architecture and protocols involved in implementing VoIP networks. After the overview, the chapter discusses the various factors that affect a high quality VoIP call. Further, the chapter introduces various codecs and the engineering tradeoffs between delay and bandwidth. Finally, the chapter gives a detailed explanation of the currently widely-used VoIP call signaling protocol, SIP.

## 2.1 VoIP Architecture

### 2.1.1 VoIP System

VoIP calls can take place between phone-to-phone, PC-to-PC, and phone-to-PC. The VoIP system configuration is shown in Figure 2.1 is a representative scenario. In the PC-to-PC call, as an example, once the media path is established, the analog signal is sampled at 8 kHz. These samples are then encoded in an 8-bit binary format. The encoded samples are put into UDP packets of different sizes and sent at a constant rate. The reverse process takes place at the receiver PC: the speech samples are extracted from the packet and then put into the play-out buffer as the analog signal.

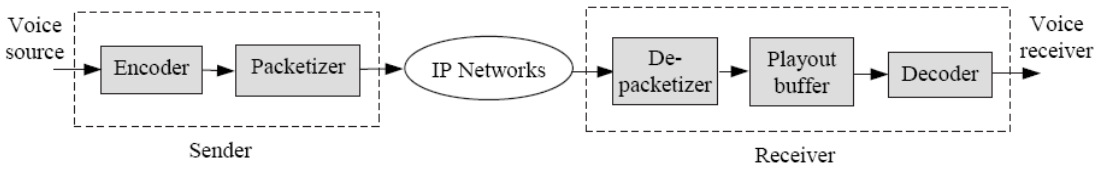


**Figure 2. 1: Conceptual diagram of a VoIP network [20]**

## 2.1.2 VoIP Protocol Structure

Since 1990's, the dominant commercial architecture uses the Internet protocol suite TCP/IP, whereas VoIP uses RTP/UDP/IP. Figure 2.2 gives complete communication network architecture.

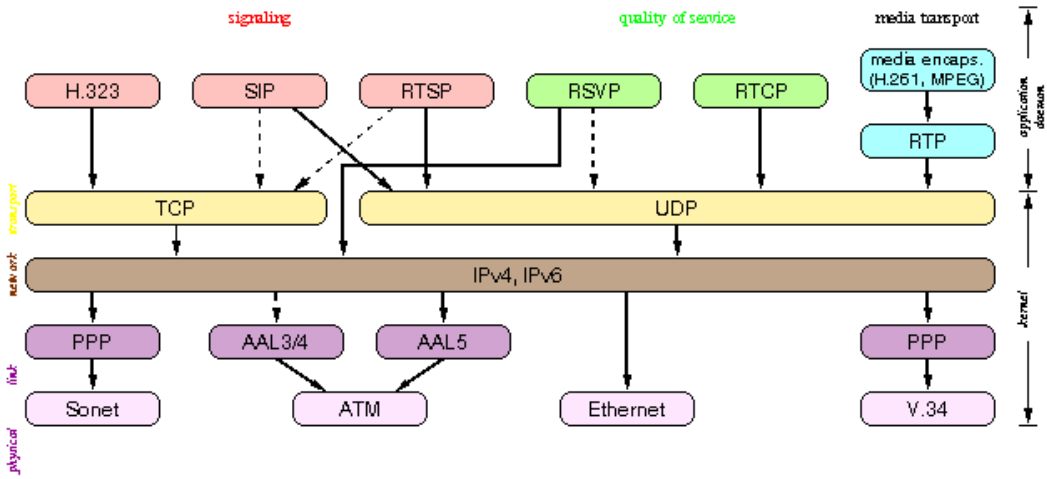


**Figure 2. 2: Internet protocol stack [21]**

As we know, the Internet Protocol (IP) deals only with the connectionless delivery of the packets, which is based on a best-effort service. Transmission Control Protocol (TCP) is a reliable connection oriented control protocol above IP. The TCP has the following characteristics.

- *Reliable*

Each Transmission of data is acknowledged by the receiver, and retransmission is needed to ensure packet receipt.

- *Connection oriented*

A virtual connection is established before the any user data is transferred.

- *Full Duplex*

The transmission is provided in both directions.

- *Rate Adjustment*

The transmission rate increases when no congestion is detected; the transmission rate reduces quickly when the sender does not receive positive acknowledgments from the receiver.

Despite these features, the TCP/IP is not suitable for real-time communications, such as speech, because the acknowledgment/retransmission feature would lead to excessive delays [1].

In contrast to TCP, User Datagram Protocol (UDP) is classified as unreliable connectionless protocol, which does not provide sequencing and acknowledgement. Without flow control and error recovery, UDP simply sends and receives IP traffic between users in an Internet.

The Real-Time Protocol (RTP), used in conjunction with UDP, provides end-to-end network transport functions for applications transmitting real-time data, such as audio and video, over unicast and multicast network services [22]. RTP standardizes the packet format by including the sequence numbers and timestamps, which is convenient to multimedia applications. It should be emphasized that RTP in itself does not provide any mechanism to ensure timely delivery of data or provide other quality of service guarantees [23]. Indeed, RTP encapsulation can be only seen at the end user, and is not distinguishable from IP packets without RTP at the intermediary routers.

A companion protocol RTCP does support the features as follows,

- Monitor the link
- Separate packets sent on a different port number

- Exchange information about losses and delays between the end systems
- Packets sent in intervals determined based on number of end systems and available bandwidth

However, a continuous stream of RTP/UDP/IP packet is offered most VoIP applications as shown in the Figure 2.3.

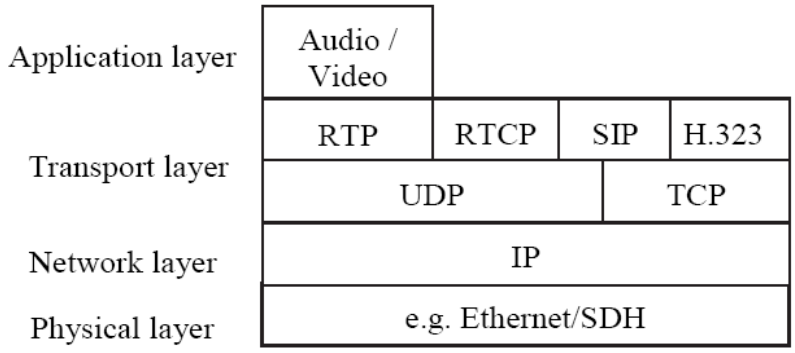


**Figure 2. 3: VoIP protocol structure**

As far as VoIP call signaling protocols are concerned, there are peer-to-peer control-signaling protocols such as H.323 protocol suite [24] and SIP [25], master–slave control-signaling protocols such as Media Gateway Control Protocol (MGCP) [26-27], and Megaco/H.248 [28]. In this dissertation, we implemented SIP signaling in our VoIP test-bed and addressed its security features.

## 2.2 Quality of Service

Quality of Service (QoS) is a measure of the voice quality experienced by the user. The network service provider uses it for bandwidth management over the IP network in order to ensure that transmission resources consistent with the expected QoS are available. Management of network resources is becoming increasingly important as more services are added on to the Internet.

VoIP becomes an attractive solution for the future, since the packet-switching technology has several advantages in both cost and architectural aspects over the circuit- switching technology. However, questions remain as to whether the voice quality provided in VoIP networks can meet the high standards provided by the PSTN that users have become accustomed to, and would expect from any competing service. The quality of speech perceived by the VoIP user is ultimately determined by parameters such as delay, jitter and packet loss [29].

### *A. Delay*

Due to the interactive nature of the voice communications, delay becomes the primary concern of the QoS in VoIP networks. It is composed of transmission delay, queuing delay, processing delay and propagation delay [30]. The transmission delay is dependent on the channel capacity in bits per second (bps). Due to the network congestion, Queuing delay is the time that the packets queued in the buffer before being processed. Processing delay is occurred at the end points, i.e. process packets headers, code/decode voice signals. Propagation delay depends on the tansmission medium, such as coax, fiber, or wireless channel. The propagation delay is generally negligible when compared to the other components of delay in an end-to-end VoIP scenario.

International Telecommunications Union-Telephony (ITU-T) Recommendation G.114 [31] provides one-way transmission delay specifications for voice. The specification is presented in Table 2.1. It has been shown that an end-to-end delay of over 150 milliseconds (ms) is intolerable to VoIP users, and the delay between successive packets must be lower than 20 ms for uninterrupted and smooth

hearing [33]. Studies have shown that several techniques such as Weighted Fair Queuing, Weighted Round Robin, Priority Round Robin, Priority Queuing, or Classed based Queuing [34], can reduce the network delay.

| Delay | Impact | Pre-Condition |
|---|---|---|
| Below150 *ms* | Acceptable for most user applications | Adequate echo control for connections of one-way delay more than 25 *ms*, as described in G.131 [32] |
| 150-400 *ms* | Acceptable for international calls | |
| Above 400 *ms* | Unacceptable for general network planning purposes, especially in the case of transporting voice in packet switched networks. | |

**Table 2. 1: Delay Specifications for Voice [31]**

In this dissertation, we focus on the impact of the queuing delay on VoIP networks. Since voice traffic has higher priority over data traffic, the queuing behavior of the voice packets is analyzed independently from the data packets in this dissertation. It is well known that an aggregate of voice (and CBR video) sources is reasonably accurately modeled by a Poisson arrival process and that queuing delays in consecutive nodes are more or less statistically independent [35]. This dissertation models two scenarios represented by the M/M/1 and M/D/1 queuing disciplines, and develops one method of calculating the throughput under a specified threshold of the total queuing delay through a VoIP network of $N$ nodes. In addition, the analytical results addressed are used in scaling resources in a VoIP network for different thresholds of acceptable delays.

***B. Jitter***

Jitter is delay variation. It can lead to the gaps in the playout of the voice stream. The jitter can be compensated by maintaining a playout buffer at the receiver side [36], which processes the incoming packets in such a way that early packets have

more delay and late packets have less delay. This means that the received voice stream can be recovered at a steady rate. In addition, arriving voice packets that exceed the maximum length of the jitter buffer will be discarded.

***C. Packet loss***

From an end-to-end point of view, the overall packet loss includes the network packet loss and the late arrival packet loss dropped at the jitter buffer. Packet loss can introduce audio distortion because of voice skips and clipping. Moreover, it can also introduce considerable impairment to voice signals. Typically, a packet loss rate of more than 5% is unacceptable for the VoIP users [37]. In order to reach the equivalent level of voice quality in a PSTN, the threshold rate of packet loss should be set below 1% in VoIP networks.

There are two methods to correct packet loss in packet switched networks, one by using Forward Error Correction (FEC), the other by using the packet loss concealment (PLC) algorithm [38]. The FEC method requires data redundancy and allows the reconstruction of lost data [39-40]. The disadvantage of this approach is that it causes overhead bits and, therefore, additional delay. PLC method, as implied in its name, conceals the packet loss. It uses a variety of techniques to recover the missing packets, such as silence substitution, packet repetition, waveform substitution, and pitch waveform replication [41].

## 2.3 VoIP Implementation

Aiming to satisfy the voice quality expectations of the customers provided in VoIP networks, laboratory experiments on how various network impairments affect

the voice quality [42], were carried out. This section describes the project that measures the voice quality under different network impairments in the VoIP networks.

### 2.3.1 VoIP Test Bed

A SIP-based VoIP test-bed in our lab has been implemented, by interconnecting the OU-Tulsa network to sip.edu by a peering arrangement. We configured the CISCO 2600 routers as media gateways, and installed MySQL 4.0.21 open source database as the location database. SIP Express Router (SER) from www.iptel.org is installed and configured as the SIP proxy server. Figure 2.4 shows the implemented VoIP infrastructure for the OU-Tulsa TCOM Lab.

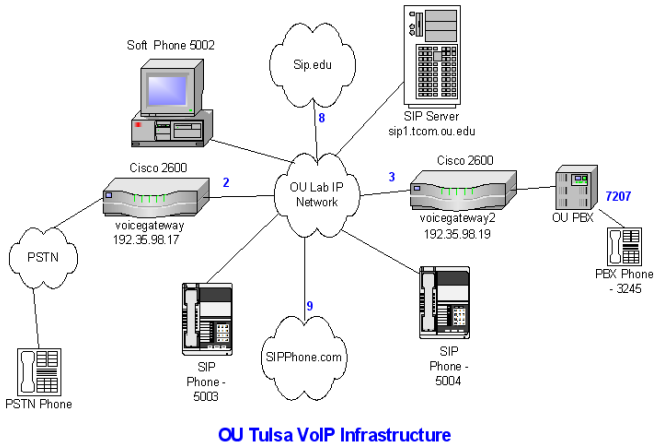


**Figure 2. 4: OU TCOM-Lab VoIP Infrastructure**

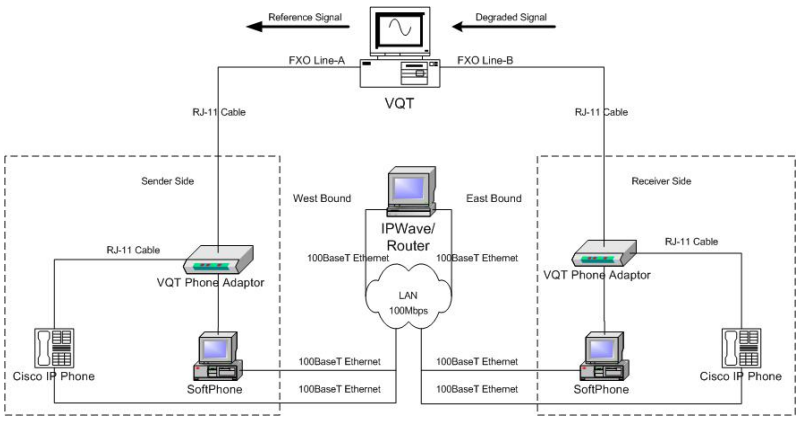


**Figure 2. 5: Physical setup of the test-bed**

In order to investigate the effects of various network impairments on the voice channel in the VoIP networks, we set up the following test-bed to measure the perceived speech quality. The test-bed consists of IPWave [43], Voice Quality Tester (VQT) [44] and the original VoIP network, as shown in Figure 2.5.

IPWave and Agilent VQT are running on Windows-NT operating system. IPWave is a network impairment generator to emulate the real world network conditions. It divides the network into Westbound and Eastbound and functions as a gateway. It introduces various network impairment conditions to the IP traffic from Westbound to Eastbound and vice versa. These impairments include the packet loss, delay-jitter, out-of-order packets, and error in packets. The Agilent VQT is, an objective speech quality measurement system, used to predict the MOS of the perceived speech quality, by means of Perceptual Speech Quality Measurement (PSQM) algorithm [45]. In order to connect the FXO line of the Agilent VQT to either hard phone headset or soft phone PC's sound card, the Agilent VQT phone adapter [46] is used.

### 2.3.2 Measurement of Voice Quality

Voice quality is inherently subjective because it is determined by the listener's perception. The subjective voice quality is measured by objective measurement techniques, by testing the Mean Opinion Score (MOS).

The perceived speech quality is measured in the way as shown in Figure 2.6. The Agilent VQT captures the perceptual domain representation of two signals, namely, a reference signal which is input to the test-bed, and a degraded signal which is the output of the test-bed. It uses Perceptual Speech Quality Measurement (PSQM)

to analyze the voice quality in terms of MOS, which is widely accepted as a norm for voice quality rating.

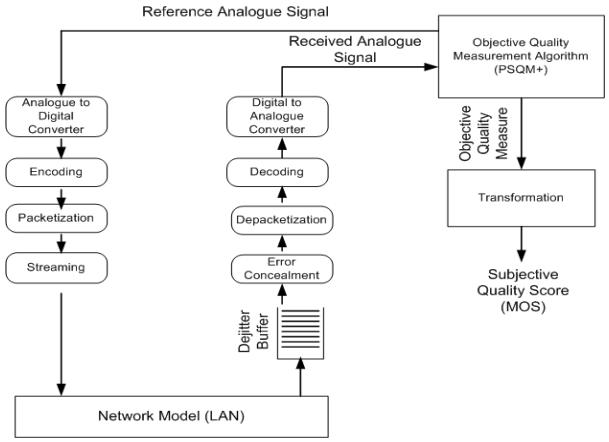


**Figure 2. 6: Block Diagram of the measurement [47-48]**

Different experiments were conducted to determine the influence of packet loss on the voice quality. In this measure, we only apply one codec G.711-$\mu$ Law selected from the Cisco hard phone. We take into account three models, which are periodic packet loss, random packet loss and burst packet loss model. The results of these three loss models are shown in Figure 2.7, Figure 2.8 and Figure 2.9 respectively. By comparing three figures, we can see that the voice quality decreases as the amount of packet loss increases. It also shows that burst packet loss has more influence on the perceived voice quality.

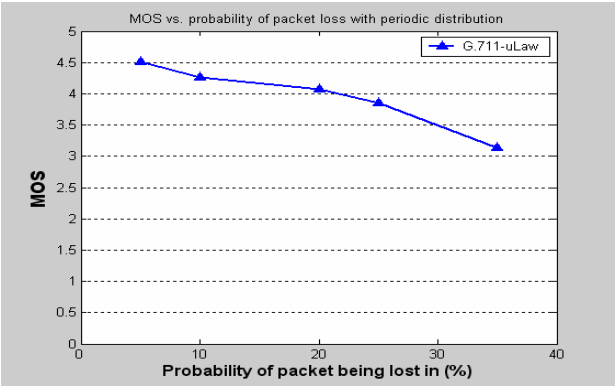


**Figure 2. 7: Periodic Loss Model**

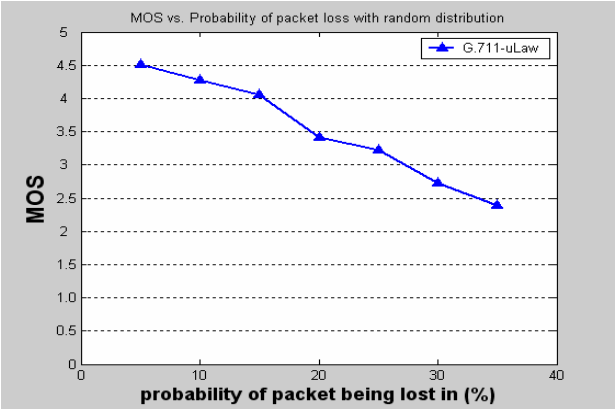


**Figure 2. 8: Random Loss Model**

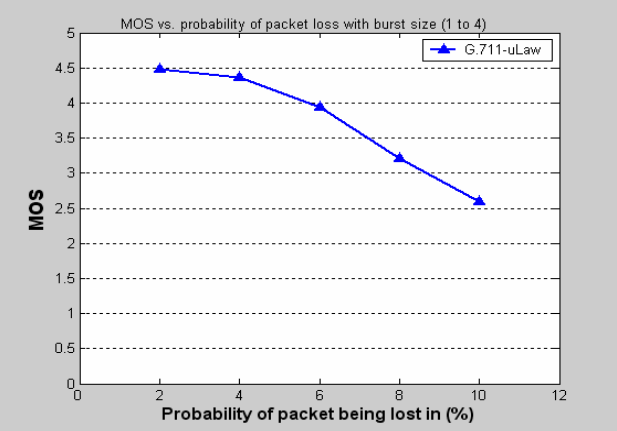


**Figure 2. 9: Burst Loss Model**

## 2.4 Session Initiation Protocol

This part introduces the Session Initiation Protocol (SIP) specification and provides the important aspects of SIP application in Voice over IP networks.

### 2.4.1 Background

SIP originated to distribute the multimedia content by the Internet Engineering Task Force (IETF) in 1996. Since SIP was standardized to be adopted in Voice over Internet Protocol (VoIP) in 1999 as [RFC2543], it has evolved significantly and now

it covers a wide range of real-time collaboration functionalities [49]. In this chapter, we will only focus on the latest standard as [RFC3261].

SIP is an end-to-end, client-server session signaling protocol. It is designed to establish presence, locate users, set up, modify and tear down voice and video sessions across the packet-switched networks. Borrowing from ubiquitous Internet protocol, such as the hypertext transfer protocol (HTTP) and simple mail transfer protocol (SMTP), SIP is text-encoded, programmable and highly extensible [50]. Due to its simplicity and extensibility, as well as the new created features, SIP is not limited to the IP telephony, SIP messages can convey arbitrary signaling payload, session description, instant messages, JPEGs, any MIME type. SIP uses Session Description Protocol (SDP) [51] for media description.

## 2.4.2 SIP Network Elements

### *A. User Agent*

User agents are end entities in SIP-based Networks, to connect each other and negotiate a session characteristics. User agents can be both hardware and software. Take example of the SIP-based VoIP testbed in our lab, we use Cisco 7960 SIP phone as shown in figure 2.10. It usually, but not necessarily, resides on a user's computer in form of a user application [52]. It can be also PSTN gateways, cellular phones, PDAs and so on.

**Figure 2. 10: A Cisco 7960 SIP Phone**

In terms of functionalities, UA can be categorized into User Agent Client (UAC) and User Agent Server (UAS). UAC and UAS are logically separated but physically combined in the same end point. UAC behaves on the behalf of the client to originate the call and receive the response, whereas UAS behaves on the behalf of the server to listen to the incoming calls and respond the request. For example, to initiate a call session, an INVITE message is sent by the caller's UAC, and received by the callee's UAS. In the opposite, to terminate the session, a BYE message is sent by callee's UAC and received by caller's UAS.

***B. SIP Server***

Based on the functionalities, SIP servers are logically classified into three components as, Registrar, Proxy Server, and Redirect Server.

Registrar is one of the SIP servers used to initialize and keep the record of the user agent. It accepts the REGISTER requests and maintains the information of the users' AoR (Address of Record) including various kinds of SIP URL address binding to the same user. Registrar also indicates the current address with the first priority where the user wants to send the request and receive the response [53].

Proxy server plays a very important role in processing the SIP signaling messages. It receives the request from the users and looks up in the location server where all the records of the users are kept, to find the destination address. And then the SIP server forwards the request by interpreting, and modifying certain parts of the INVITE message, such as Via. Proxy server can be classified as statefull proxy server and stateless proxy server.

### 2.4.3 SIP Messages

SIP messages are divided into two types depending on the direction of the messages. The SIP message sent from the client to the server is Request message, while from the server to the client is Response message. Table 2.2 and Table 2.3 give examples of the Request and Response SIP messages respectively.

| Method | Description |
|---|---|
| **INVITE** | Initiates a call, changes call parameters (re-INVITE) |
| **ACK** | Confirms a final response for INVITE |
| **BYE** | Terminates a call |
| **CANCEL** | Cancels searches and "ringing" |
| **OPTIONS** | Queries the capabilities of the other side |
| **REGISTER** | Registers with the Location Service |
| **INFO** | Sends mid-session information that does not modify the session state |

**Table 2. 2: Request Methods Example**

| Type | Class | Description | Examples | |
|---|---|---|---|---|
| | | | **Code** | **Meaning** |
| **Provisional** | 1xx | In Progress | 100 | Trying |
| | | | 180 | Ringing |
| **Final** | 2xx | Success | 200 | OK |
| | 3xx | Redirection | 300 | Multiple choices |
| | | | 301 | Moved permanently |
| | | | 302 | Moved temporarily |
| | 4xx | Client Error | 400 | Bad request |
| | | | 401 | Unauthorized |
| | | | 403 | Forbidden |
| | | | 408 | Request time-out |
| | | | 480 | Temporarily unavailable |
| | | | 481 | Call/Transaction does not exist |
| | | | 482 | Loop detected |
| | 5xx | Server Error | 500 | Server error |
| | 6xx | Global Failure | 600 | Busy everywhere |
| | | | 603 | Decline |
| | | | 604 | Does not exist anywhere |
| | | | 606 | Not acceptable |

**Table 2. 3: Response Example**

The SIP message consists of three main parts: start line, header and message body. Each SIP message begins with a Start Line to convey the message type and the protocol version. SIP headers are borrowed from the syntax and semantics of HTTP header fields, to convey more message attributes. The message body can be either Session Description Protocol (SDP) or Multipurpose Internet Mail Extensions (MIME). Here is an example of the INVITE message:

```
INVITE sip:bob@nice.com SIP/3.0
Via: SIP/3.0/UDP 192.2.4.4:5060
To: Bob < sip:555-6666@nice.com>
From: Aline <sip:555-1234@nice.com > ;
tag=203 941 885
Call-ID: b95c5d87f7721@192.2.4.4
Cseq: 26 563 897 INVITE
Contact: < sip:555-1234@192.2.4.4>
Content-Type: application/sdp
Contact-Length: 142

v=0
o=Alice 53655765 2353687637 IN IP4
128.3.4.5
s=Call from Alice
c=IN IP4 192.2.4.4
M=audio 3456 RTP/AVP 0 3 4 5
```

### 2.4.4 SIP Transactions

SIP transaction is a sequence of SIP messages ranging from the request to all responses to that request. SIP is *transactional,* because the SIP messages are arranged into transactions although they are sent independently. Figure 2.11 gives two examples of the meaning of the SIP transactions. In the first example, ACK message is within the transaction initiated by an INVITE message, while in the second example, the ACK is not considered as the part of the transaction because the final response is a 2xx successful response.

As addressed in RFC3261 [25], the transaction identifier is expressed as the branch parameter inside the Via header fields. However, since the previous SIP RFC2543 calculates the transaction identifier as the hash of all important message header fields (that included To, From, Request-URI and CSeq) [54], the compatible feature should be provided for the backwards support.

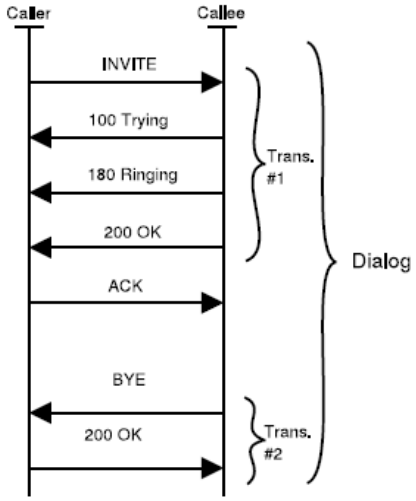


**Figure 2. 11: SIP Transactions and Dialogs**

## 2.4.5 SIP Dialogues

SIP dialog represents the peer-to-peer relationship between two end user agents. Also shown in Figure 2.11, the two transactions are not treated independently, but related in the way belonging to the same *dialog*. Being identified by From and To tag, and Call-ID, SIP Dialogs facilitate proper sequencing and routing of messages between the user agents [25]. Also, we have known that the command sequence (*Cseq*) contains an integer and a method name. This *Cseq* number increments for each new request, which actually means the *CSeq* number identifies a transaction. To some degree, *a dialog is a sequence of transactions* [52].

### 2.4.6 Typical SIP Scenarios

To understand the SIP signaling much better, two scenarios to illustrate the SIP message flow are presented.

One is a redirection scenario as shown in Figure 2.12. Upon receiving the INVITE message from the user agent A, the redirect server responds with 302 (Moved Temporarily), indicating the user agent B is temporarily available at an alternate address expressed in the Contact header. Sometimes, the duration of validity of these addresses is also included. After returning the acknowledgement to the redirect server, the user agent A sends a second INVITE message directly to the user agent B, by using the routing information pushed back from the redirect server. With the aid in locating the target of the request from the redirect server, the procedure becomes simple and quick. In other words, the redirect server yields high-performance. It is worth to note that the second INVITE message has different is CSeq value compared to the first INVITE message, however, the To and From headers, Call-ID and dialog identifiers remain the same. The following sequence of the signaling is common in each scenario: once the user agent B picks up the phone, 200 OK message is sent back to the user agent B; and the media flow is established after the user agent B receives the acknowledgement.

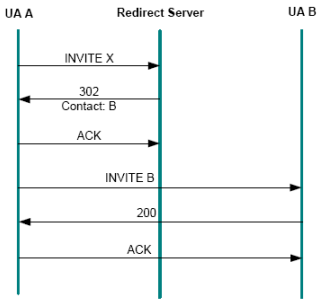


**Figure 2. 12: Signaling flow with Redirect Server [55]**

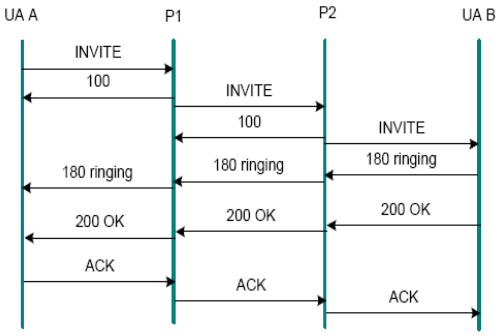


**Figure 2. 13: Signaling flow with Proxy server [55]**

The other scenario as shown in Figure 2.13 is that the request traverses multiple proxy servers before reaching the destination. The main difference from the first scenario is in that after making routing decision, each intermediary proxy server modifies the INVITE message and then forwards it to the next proxy server; Response routes through the same set of proxies in the reverse order.

## 2.5 Summary

This chapter has provided a brief overview of the VoIP networks from different perspectives. The laboratory implementation and studies on the measurement of the voice quality have been discussed. The popular VoIP signaling procedure SIP has been described in details. In the next section, we will present the analytical model for delay-throughput analysis used in this dissertation.

# Chapter 3. Traffic Characterization

**Abstract:** This chapter presents the traffic characterization of the packet-switched networks. It provides an overview of queuing theory, including the terminology and notations, which will be used in the analysis presented in the succeeding chapters. In order to study the relationship between delay and throughput, mathematical models characterized by the queuing theory are also presented. Analysis of these models is discussed in Chapter 4 and Chapter 5.

In order to analyze a telecommunication network, that carries both data and voice traffic, an analytical model is needed to capture the main characteristics of the system, i.e., the statistical properties of the traffic. Traffic modeling is necessary and important because it can help find out optimal network configuration and sizing without having to build prototypes [56]. Mathematical tools, such as stochastic processes, queuing theory, and numerical simulation are used to model the traffic. Since this dissertation is focused on the queuing delay in VoIP networks, queuing theory is mainly used to analyze the voice traffic.

Despite the fact that data networks such as the Internet are drastically different from legacy public switched telephone networks, the long held paradigm in the communication and networking research community has been that data traffic — analogous to voice traffic — is adequately described by certain Markovian models which are amenable to analysis and efficient control [57]. Hence, in this dissertation, Voice over IP networks is modeled in a way to provide a predefined quality of service, and optimize the capacity of system.

## 3.1 Packet-Switched Networks Model

The traditional telephone network has been designed as a hierarchical system layered by the local exchange (LEX) and transit exchanges (TEX). LEX is the subscriber switch which connects the individual to the network. TEX is a transit exchange connecting to several LEX's. The transit exchanges themselves form a hierarchy. The highest level switches of this hierarchy are fully interconnected with each other.

A packet-switched data network is a packet distribution network following the store-and-forward principle at the packet level, as shown in the Figure 3.1. It can be modeled as a stochastic flow system, where the flow is the packet and the channel is the network link. The data is segmented into packets and can be transmitted through different paths through the network. The packet switched network introduces queuing delay at each node. Voice packets are much more sensitive to delay and jitter than data packets. In this dissertation, the aim is to reduce the queuing delay by analyzing performance through the application of queuing theory and appropriate provisioning of network resources.

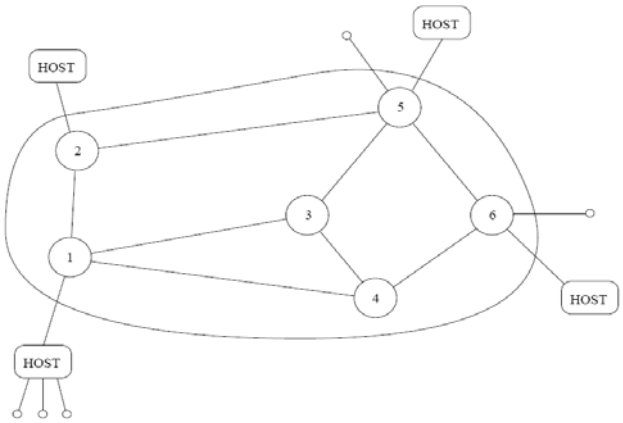


**Figure 3. 1: Packet-Switch Data Network [58]**

## 3.2 Queuing Model

In this section, a mathematical model is described to represent the packet-switched data networks which support voice traffic [59]. An upper bound for end-to-end delay is chosen as an objective measure of the quality of service. Thus, an idealized model is built, which retains the essential performance characteristics of real networks. As mentioned in Chapter 2, there are many sources of the packet delay. In this dissertation, the other delays, mostly uncontrollable delays such as the propagation and processing delays. Queuing delay is the variable component of the end-to-end delay causing considerable degradation in real-time voice communication. This dissertation only focuses on the queuing delay.

### 3.2.1 Queuing Specification

The queuing model used in this dissertation was developed by D.G. Kendall ([60]) and represented as A/B/n, which is widely used in the literature. A represents the distribution of the arrival process, B implies the distribution of the service time, and n is the number of servers. It needs to be mentioned that B is directly related to the length of the packets. To provide more information of the queuing system, a more complete specification is formatted as A/B/n/K/S/X, where K is the buffer size or the capacity of the link, S is the population of the arrival packets, and X denotes the queuing discipline.

The standard notations of frequently used traffic processes are shown in the Table 3.1.

| Notation | Description |
|---|---|
| **M** (Markovian) | Exponential time intervals (Poisson arrival process, exponentially distributed service times) |
| **D** (Deterministic) | Constant time intervals |
| **Ek** | Erlang-k distributed time intervals (E1 = M) |
| **G** (General) | Arbitrary distribution of time intervals (may include correlation) |

**Table 3. 1: Kendall's Notations**

### 3.2.2 Assumptions of the Queuing Model

Generally speaking, analytical modeling of real systems for queuing analysis is a complicated task. To simplify the analysis and yet get a realistic evaluation of the end-to-end delay of packetized voice traffic [61], the following assumptions have been made:

- All channels are assumed to be noiseless [62]. The data rate of transmission over the channel does not exceed the channel capacity.
- The buffer size is unlimited. It implies that the storage capacity at each node is infinite to provide waiting positions in the queues and sustain transient congestions.
- The transmission is considered only over one or more point-to-point links. Each packet is delivered to the single destination.
- The processes of the arrival and the service are mutually independent. The assumption of preservation of the Poisson arrival process at an intermediate node is a strong assumption made in this paper. We justify it based on the following observation. At an intermediate node, while the message arrival process from a single node might not be Poisson, the collective interarrival times and the lengths of messages generated by the entire population of subscribers exhibit an independence, since the length of a message generated by any particular subscriber

is completely independent of the arrival times of messages generated by the other subscribers [62]. It's actually the multiplicity of paths that lead to a single node and multiplicity of exit off that node that considerably reduces the dependency between interarrival times and lengths of messages as they enter and exit the various nodes within the network. Extensive simulations of networks in [62] validate this simplifying assumption which makes the mathematical analysis tractable.

- The arrival process of the packet is poisson, and the service diciplipline is negative exponentially distribution.
- For multiple hop traffic, the waiting times at each node are independent. With this assumption (approximation), the end-to-end delay can be computed by the convolution of the waiting times at each node [63].

### 3.2.3 Statistical properties of traffic

As stated in the last section, the arrival traffic is assumed to follow the Poisson discipline. The Poisson process is generally considered to be a good model for the aggregate traffic from a large number of similar and independent users [64]. Although there is no quantitative data to determine the actual distribution of the real traffic, certain data obtained by Molina [65] for telephone traffic correspond very well to the assumption [62]. Table 3.2 shows the distributions of the Poisson process.

Regarding the service process, two cases are considered. The first case is M/M/1 model, where the service time follows the exponential distribution with parameter $\mu$. It also indicates that the packet length is exponentially distributed. The

M/M/1 model has long been the typical model for queuing analysis due to its simplicity, for such widely varying applications as the terminal to computer communications, shared Local Area Networks, and airline reservations [66]. However, special emphasis is put on the second case which is M/D/1 model, with the constant service time. M/D/1 model is of the main interest since constant service time will often be the case for many applications. It is suitable for the networks with fixed packet length, such as ATM and VoIP networks.

| **Distribution** | **Exponential** | **Erlang-k** | **Poisson** |
|---|---|---|---|
| **Description** | Interval between two successes or from a random point until next success | Time interval until k'th success | Number of successes in a time interval t |
| **Formula** | $f(t) = \lambda e^{-\lambda t}, t \geq 0$<br>$m_1 = \frac{1}{\lambda}$<br>$\sigma^2 = \frac{1}{\lambda^2}$ | $f(t \mid k) = \frac{(\lambda t)^{k-1}}{(k-1)!}\lambda e^{-\lambda t}$<br>$t \geq 0$<br>$m_1 = \frac{k}{\lambda}$<br>$\sigma^2 = \frac{k}{\lambda^2}$ | $f(x \mid t) = \frac{(\lambda t)^x}{x!}e^{-\lambda t}$<br>$t \geq 0$<br>$m_1 = \lambda t$<br>$\sigma^2 = \lambda t$ |

**Table 3. 2: Distributions of the Poisson Process**

## 3.3 Analysis of the Delay Bound

Most of the work of this dissertation is devoted to improve the throughput within a predefined upper delay bound, in particular, the queuing delay. The waiting time distribution is used to analyze the upper delay bound. For simplicity, the queue is following the discipline FIFO (First–In First–Out), also called FCFS (First–Come First–Served). Packets arriving first to the node will be served first. Also it is assumed that there is single server at each node. In the following, the waiting time distributions

for both M/M/1 and M/D/1 models are derived respectively. Table 3.3 summarizes the notations that are used in queuing analysis and delay calculations throughout the dissertation.

| Symbol | Meaning |
|---|---|
| $n$ | Number of the nodes in the network |
| $h$ | Number of the hops in the network |
| $\lambda$ | Mean packet arrival rate |
| $\frac{1}{\mu}$ | Mean length of each packet (bits) |
| $C$ | Transmission capacity of each link (bps) |
| $\mu C$ | Service rate (packets per second) |
| $\rho$ | Utilization factor $\frac{\lambda}{\mu C}$ |
| $W_n$ | End-to-end queuing delay (n hops) |
| $t$ | End-to-end threshold delay |
| $F_W(t)$ | Distribution of waiting time $P\{W \le t\}$ |
| $f_W(t)$ | Probability density function |
| $\gamma_n$ | Throughput of $n$- hop network |
| $D$ | End-to-end delay |
| $D_i$ | Delay introduced by the ith hop |
| $D^{(i)}$ | The i-th moment of the delay |
| $\sigma_D^2$ | Jitter (delay variance) |

**Table 3. 3: Notations Used in This Dissertation**

### 3.3.1 M/M/1 Model

Let us consider a typical customer. Assume the number of customers waiting in the queue at the arrival time is denoted by $X^*$. In other words, it is the queue length seen by an arriving customer. The Poisson arrival process the PASTA-property (Poisson Arrivals See Time Averages),

$$P\{X^* = i\} = P\{X = i\} = \pi_i \tag{3.1}$$

where $X$ is the number of packets in the node at an arbitrary time.

Due to the memoryless property of the exponential distribution, the remaining service time $S_1^*$ of the packet in service also follows exponential distribution with the parameter $\mu$, and is independent of other packets in the queue. Service times of these packets waiting in the queue are IID (Identically and Independently Distributed). For the assumed the FIFO queuing discipline, the total waiting time for the typical packet is given as,

$$W = S_1^* + S_2 + ... + S_i \tag{3.2}$$

where $\tau_1 = S_1^*$ and $\tau_n = S_1^* + S_2 + ... + S_i$ , $n \geq 2$.

Let us construct a Poisson (Point) process [67] $\tau_n$ as shown in Figure 3.2,

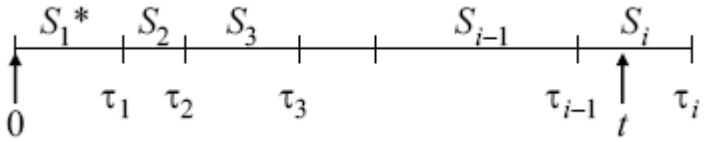


**Figure 3. 2: Poisson Process**

Therefore, we have

$$P\{W = 0\} = P\{X^* = 0\} = \pi_0 = 1 - \rho \tag{3.3}$$

Since $X^* = i$, we have $W > t \Leftrightarrow \tau_i > t$. Hence,

$$P\{W > t\} = \sum_{i=1}^{\infty} P\{W > t \mid X^* = i\} P\{X^* = i\} = \sum_{i=1}^{\infty} P\{\tau_i > t\}\pi_i = \sum_{i=1}^{\infty} P\{\tau_i > t\}(1-\rho)\rho^i \tag{3.4}$$

Also the Poisson process can be seen as a counter process $A(t)$ corresponding to $\tau_n$, which follows that

$$\tau_i > t \Leftrightarrow A(t) \leq i - 1 \tag{3.5}$$

On the other hand, it is already known that $A(t)$ follows the Poisson distribution with parameter $\mu t$ . From equations (3.4) and (3.5), we have

$$P\{W > t\} = \sum_{i=1}^{\infty} P\{\tau_i > t\}(1-\rho)\rho^i = \sum_{i=1}^{\infty}\sum_{j=0}^{i-1} \frac{(\mu t)^j}{j!} e^{-\mu t}(1-\rho)\rho^i$$

$$= \rho \sum_{j=0}^{\infty} \frac{(\mu t\rho)^j}{j!} e^{-\mu t}(1-\rho) \sum_{i=j+1}^{\infty} \rho^{i-(j+1)}$$

$$= \rho \sum_{j=0}^{\infty} \frac{(\mu t\rho)^j}{j!} e^{-\mu t} = \rho e^{\mu t\rho} e^{-\mu t} = \rho e^{-\mu(1-\rho)t} \quad (3.6)$$

It is interesting to find out that the distribution of the waiting time $W$ can be presented as the product of two independent random variables; one follows Bernoulli distribution with parameter $\rho$ and another follows Exponential distribution with parameter $\mu(1-\rho)$.

$W = JD$, where $J \sim Bernoulli(\rho)$ and $D \sim Exp(\mu(1-\rho))$ (3.7)

Thus we have the average waiting time as

$$E(W) = E(J)E(D) = \rho \frac{1}{\mu(1-\rho)} = \frac{\rho}{\mu(1-\rho)} \quad (3.8)$$

After derivation the waiting time distribution, the throughput of the VoIP network where all packets that undergo a queuing delay not exceeding the threshold delay $t$, can be derived as:

$$\gamma = \lambda p \quad (3.9)$$

where $\lambda$ is the arrival rate of packets, $p$ is the probability of the waiting time less than the threshold delay $t$, which equals to $1 - P\{W > t\}$.

Also, $p$ is referred to the normalized throughput, i.e., throughput expressed as a function of the incident traffic, which can be given as:

$$p = \frac{\gamma}{\lambda} \tag{3.10}$$

This relationship between the delay threshold and the throughput remains the same for the M/D/1 model.

### 3.3.2 M/D/1 Model

Several formulas have been proposed to numerically evaluate the waiting time distribution for M/D/1 model. For small waiting times, equation (3.11) is used for numerical evaluation. For larger waiting times, there are two different cases. One with the integral value of the waiting time *t,* equations (3.12) and (3.13) are used for calculation. The other with non-integer waiting time, the waiting time distribution is expressed in terms of integral waiting times, as shown in equation (3.14).

As given by Erlang in 1909, the distribution function of waiting time can be written in a closed form [68],

$$P(W \le t) = (1-\lambda)\sum_{j=0}^{T} \frac{[\lambda(j-t)]^j}{j!} e^{-\lambda(j-t)} \tag{3.11}$$

where $t = T + \tau$, i.e. $T = \lfloor t \rfloor$, where $\lfloor t \rfloor$ is the largest integer less than or equal to *t*.

Iversen (1982 [69]) has shown that for an integral value of t, we have

$$P\{W \le t\} = p(0) + p(1) + ... + p(t) \tag{3.12}$$

where $p(i)$ is the state probability. To be accurate, the state probabilities are calculated by using a recursive formula based on Fry's equations of state as [70],

$$p(i+i) = \frac{1}{p(0,h)}\{p(i) - [p(0) + p(1)]p(i,h) - \sum_{j=2}^{i} p(j) \bullet p(i-j+1,h)\} \tag{3.13}$$

For non-integral value of *t*, it is expressed as $t = T + \tau$. Therefore we have

$$P(W \le T + \tau) = e^{\lambda\tau} \sum_{j=0}^{T} \frac{(-\lambda\tau)^j}{j!} P\{W \le T - j\} \qquad (3.14)$$

And $P\{W \le T - j\}$ can be calculated by equation (3.11).

In this dissertation, equation (3.11) is mostly applied for analysis of the M/D/1 system.

### 3.3.3 Comparison of M/M/1 and M/D/1 Models

Figure 3.3 shows the distribution of the queuing delay for both single M/M/1 and M/D/1 models. The parameter is different values of $\lambda$. Without losing any generality, the mean service time is scaled to the unit 1. As shown in the Figure 3.3, the similarities of both models are:

- The normalized throughput increases with an increase in the threshold delay, reaching 100% asymptotically.
- Both M/M/1 and M/D/1 systems consistently show a higher normalized throughput for a smaller $\lambda$ or, equivalently, smaller $\rho$.

The difference between M/M/1 and M/D/1 is that for given value of $\lambda$, M/D/1 system has higher normalized throughput than the M/M/1 system.

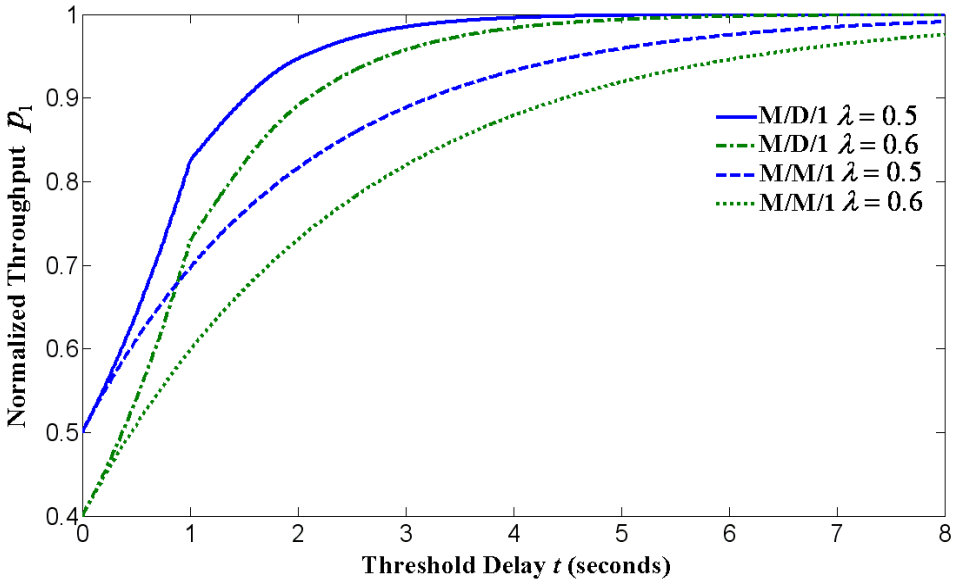


**Figure 3. 3: The Normalized Throughput of Both M/M/1 and M/D/1**

## 3.4 Summary

This chapter has provided a brief overview of the traffic modeling and queuing theory, which are applied in the work of this dissertation. The waiting time distributions for both M/M/1 and M/D/1 have been discussed. In the following chapter, the M/M/1 modeled VoIP networks are presented, including a closed form solution relating the impact of the bounded delay on the throughput.

# Chapter 4. Impact of Bounded Delays on Resource Consumption in Packet Switched Networks with M/M/1 Traffic

**Abstract:** This chapter presents a closed form solution relating the impact of bounded delays on throughput in VoIP networks modeled by M/M/1. Traffic that exceeds the delay threshold is treated as lost throughput. The results addressed can be used in scaling resources in a VoIP network for different thresholds of acceptable delays. Both single and multiple switching points are addressed, as well as the related simulation results. The contents of this chapter have been published, in part, in [71].

## 4.1 Introduction

Unlike the PSTN, the Internet was not specifically designed for voice transmission. As such, VoIP calls can suffer from delay, jitter or packet loss [72]. As described in Chapter 2, the legacy circuit-switched network allocates dedicated bandwidth to each call resulting in virtually no delay due to queuing and no jitter. PSTN provides a consistently high level of voice quality, called toll quality [73]. VoIP packets in packet-switched networks undergo varying amounts of delay at each transit hop. Therefore, delay becomes a significant parameter in VoIP networks. The delay addressed in this chapter is the variable queuing delay [13], the time each voice packet has to wait at each router in the path of a VoIP connection.

### 4.1.1 Average Delay versus Bounded Delay

The average delay of voice packets is an indicator of the QoS [74] of a VoIP network. However, the average delay is not meaningful in real-time human conversations. Suppose two services with the same average delay of *300 ms* are available. One ranges from 100 *ms* to 500 *ms*, the other from 200 *ms* to 400 *ms*. The 200-400 *ms* bounded delay system provides a much higher voice quality than the 100-500 *ms* bounded delay system. The upper delay bound is a key factor in the choice of the systems. Accordingly, from the Service Providers' point of view, a system that limits the upper bound of the queuing delay to a low value is very important.

In our model, we assume an infinite buffer at each node. Packets are thus not lost, but delayed. As mentioned earlier, we use delay to characterize the Quality of Service. While contemporary literature largely uses mean delay to characterize the QoS, we propose to use an upper bound of the delay as a measure of the same. Traffic that suffers a delay higher than the bound is considered lost and does not constitute effective throughput [75]. We treat the 'lost throughput' as being equivalent to a call that was blocked in a circuit-switched network [76].

### 4.1.2 Organization of this chapter

This chapter addresses the impact of upper delay bounds on throughput in VoIP Networks. The rest of the chapter is organized as follows: Section 4.2 describes a simple IP telephony system based on the SIP protocol; Section 4.3, 4.4 and 4.5 address the theorems for optimization and analytical proof for VoIP networks

consisting of one-hop, two-hop or multiple-hop respectively; Section 4.6 presents the simulation results and performances; Section 4.7 presents the conclusion of our work.

## 4.2 The SIP-based VoIP Network Model

Figure 4.1 shows an example VoIP network where the end points are connected to a LAN and VoIP calls a routed through the LAN, WAN or gateway to the remote end point. Session Initiation Protocol (SIP) has been used as the signaling protocol in Figure 4.1. Not being limited to IP telephony, SIP messages can convey arbitrary signaling payload, session description, instant messages, JPEGs, any MIME types. SIP uses the Session Description Protocol (SDP) for media description.

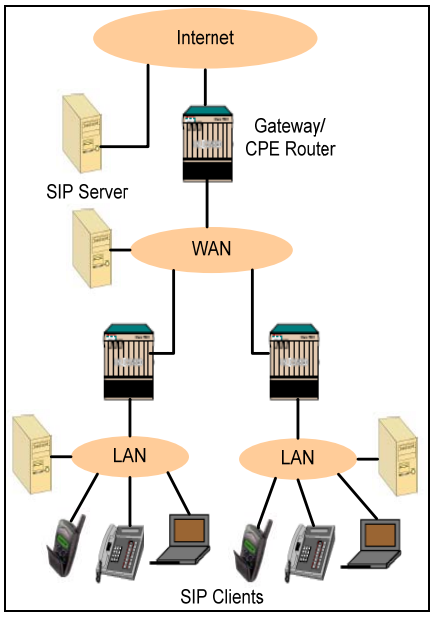


**Figure 4. 1: SIP-based VoIP Scenario**

## 4.3 A Single-hop VoIP Network

As discussed on Section 4.2, a typical voice packet would pass through several hops before arriving at the destination. We consider the voice traffic served by a single as well as multiple hops.

### 4.3.1 A single switching hop VoIP network

Consider a LAN shown in Figure 4.2 with a SIP server that functions as a VoIP network. We model the VoIP packets arriving at the SIP server as M/M/1 traffic [77-78].

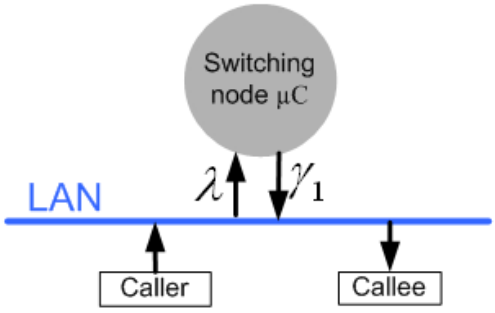


**Figure 4. 2: Single Switching Hop Network**

The distribution of the waiting time *W* can be written as [58]:

$$F_W(t) = P\{W \le t\} = 1 - \frac{\lambda}{\mu C} e^{-(\mu C - \lambda)t} = 1 - \rho e^{-\mu C(1-\rho)t} \tag{4.1}$$

The throughput of the single-hop switching system where all packets that undergo a queuing delay less than the threshold delay can be expressed as:

$$\gamma_1 = \lambda \bullet P\{W \le t\} = \lambda[1 - \frac{\lambda}{\mu C} e^{-(\mu C - \lambda)t}] \tag{4.2}$$

### 4.3.2 Analysis of a single-hop VoIP network

We first prove the following theorem that characterizes the behavior of a single-hop VoIP network.

*Theorem 1*: The throughput $\gamma_1$ of a VoIP server, where the arriving traffic follows the M/M/1 discipline, and all packets incurring a queuing delay higher than *t*

are discarded, is maximized for a mean packet arrival rate $\lambda_0$ such that the following transcendental equation:

$$\lambda_0(2+\lambda_0 t) = \mu C e^{(\mu C-\lambda_0)t}$$

is satisfied. The maximized throughput under this condition is given by

$$\gamma_{\max} = \frac{\lambda_0(1+\lambda_0 t)}{2+\lambda_0 t}$$

*Proof:* We note from equation (4.2) that $\gamma_1$ is continuous on the closed interval $[0, \mu C]$. Thus, if $\gamma_{\max}$ is an extreme value of $\gamma_1$ corresponding to $\lambda_0$ on that interval, then one of the following two statements is true: a) $\gamma'(\lambda_0) = 0$, or b) $\gamma'(\lambda_0)$ does not exist.

The first-order derivative of (4.2) is:

$$\frac{d\gamma_1}{d\lambda} = 1 - \frac{\lambda}{\mu C}(2+\lambda t)e^{-(\mu C-\lambda)t} \tag{4.3}$$

which exists. The second-order derivative of (4.2) is:

$$\frac{d^2\gamma_1}{d\lambda^2} = -[\frac{2(1+\lambda t)}{\mu C} + \frac{\lambda t(2+\lambda t)}{\mu C}]e^{-(\mu C-\lambda)t} \tag{4.4}$$

which is negative. We can now obtain the maximum throughput $\gamma_{\max}$ by putting the first derivative of $\gamma_1$ equal to zero. In other words, $\gamma_1$ will be maximized for the specific $\lambda_0$ such that

$$1 - \frac{\lambda_0}{\mu C}(2+\lambda_0 t)e^{-(\mu C-\lambda_0)t} = 0 \tag{4.5}$$

Equation (4.5) can be rewritten as,

$$\lambda_0(2+\lambda_0 t) = \mu C e^{(\mu C-\lambda_0)t} \tag{4.6}$$

The corresponding maximum throughput $\gamma_{\max}$ can be computed from equation (4.2) and (4.6) as

$$\gamma_{\max} = \frac{\lambda_0(1+\lambda_0 t)}{2+\lambda_0 t} \tag{4.7}$$

This proves the *Theorem 1*.

### 4.3.3 Discussion

Figure 4.3 shows plots of the throughput $\gamma$ for varying levels of the incident traffic $\lambda$. The capacity [79] of the server $\mu C$ and $t$ are used as parameters. It can be seen that given a fixed $\mu C$, the served traffic increases as the threshold of delay time $t$ increases. Also, given a fixed $t$, the served traffic increases as the service rate $\mu C$ increases. Further, we note that the served traffic increases as $\lambda$ increases reaching a peak at $\gamma_{\max}$ and eventually declining to zero. It follows that if the capacity of the network were fixed, the maximum throughput [80] for a given value of $t$ can be computed and the maximum allowable incident traffic known. The relationship developed will allow sizing the resources needed against known requirements of threshold delay and incident traffic.

It is instructive to compare the throughput performance illustrated in Figure 4.3 with that of a corresponding M/M/1 system where no served traffic is discarded. In that case, the served traffic asymptotically reaches the server capacity. The decline of the served traffic to zero in our case is due to the fact that as the incident traffic

approaches the server capacity, the queuing time increases indefinitely resulting in the residual traffic declining to zero.

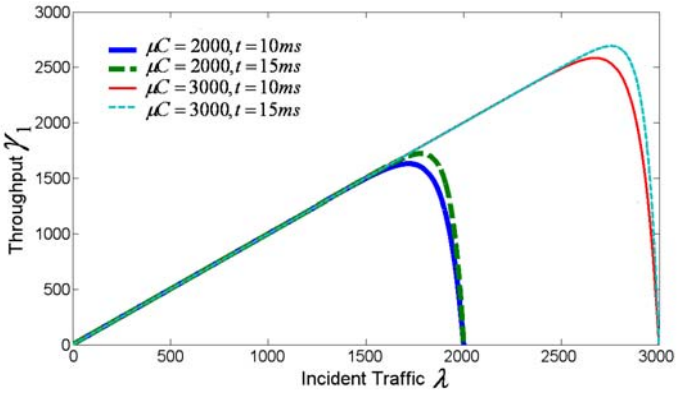


**Figure 4. 3: Single-hop Network Throughput**

## 4.4 Two-hop Tandem network

In general, a VoIP call will go through a number of hops instead of a single-hop as shown in Figure 4.2. In this section, we analyze a two-hop VoIP network.

### 4.4.1 Two-hop Tandem network

A two-hop tandem network is shown in Figure 4.4. Assume that the two links have the same capacity *C*. The analysis of throughput as a function of the threshold *t* and the incident traffic $\lambda$ can be carried out as shown below.

The arrival process of the traffic incident on the second link can be assumed to be Poisson as well [58, 81]. In our analysis of the two-hop system, we assume that all traffic served by the first hop forms the incident traffic for the second hop, *even if it was delayed beyond the threshold t.* In other words, the policy of discarding traffic with a delay higher than *t* is executed by the exit node. The probability density function (PDF) of waiting time at the first node can de derived from (4.1) as:

$$f_{W_1}(t) = \mu C\rho(1-\rho)e^{-\mu C(1-\rho)t} \tag{4.8}$$

A LST (Laplace-Stieltjes Transform) for the waiting time PDF $f_{W_1}(t)$ is given by

$$F(s) = \int_0^\infty f_{W_1}(t)e^{-st}dt = \rho\frac{\mu C(1-\rho)}{s+\mu C(1-\rho)} \tag{4.9}$$

The Laplace transform of the waiting time distribution for the two nodes in tandem $F_C(s)$, can be now calculated as [82]:

$$F_c(s) = F(s)\times F(s) = \rho^2[\frac{\mu C(1-\rho)}{s+\mu C(1-\rho)}]^2 \tag{4.10}$$

The waiting time pdf of the two-hop network can now be computed as:

$$f_c(t) = \rho^2[\mu C(1-\rho)]^2 te^{-\mu C(1-\rho)t} \tag{4.11}$$

The relation between pdf and tail-end distribution is given by

$$F_c(t) = \int_0^t f_c(x)dx \tag{4.12}$$

From (4.11) and (4.12) we have

$$P(W_2 \le t) = \rho^2\{1-e^{-\mu C(1-\rho)t}[1+\mu C(1-\rho)t]\} \tag{4.13}$$

where $P(W_2 \le t)$ represents the probability distribution function of the total delay.

### 4.4.2 Analysis of two-hop VoIP network

The maximum throughput of a two-hop M/M/1 system where each node is characterized by the same service rate $\mu C$ and *the total threshold delay is t*, can now be developed as follows. We have

$$\gamma_2 = \lambda \cdot P(W_2 \le t) = \lambda^3 (\frac{1}{\mu C})^2 \{1 - e^{-(\mu C - \lambda)t}[1 + (\mu C - \lambda)t]\} \quad (4.14)$$

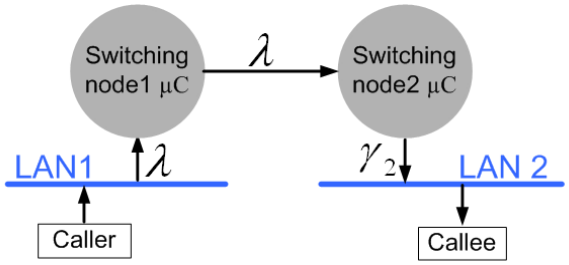


**Figure 4. 4: A Tandem VoIP Network**

From (4.14), we can get:

$$\frac{d\gamma_2}{d\lambda} = 3\lambda^2 (\frac{1}{\mu C})^2 \{1 - e^{-(\mu C - \lambda)t}[1 + (\mu C - \lambda)t]\} - \lambda^3 (\frac{t}{\mu C})^2 (\mu C - \lambda) e^{-(\mu C - \lambda)t} \quad (4.15)$$

The maximum value of $\gamma_2$ occurs when $\frac{d\gamma_2}{d\lambda} = 0$, or when $\lambda_0$ on the interval [0, $\mu C$] satisfies

$$3e^{(\mu C - \lambda_0)t} = 3 + (3 + \lambda_0 t)(\mu C - \lambda_0)t \quad (4.16)$$

The maximum value of $\gamma_2$, or $\gamma_{2\max}$ can be derived from (4.14) and (4.16) as

$$\gamma_{2\max} = \frac{\lambda_0^4}{3} \cdot (\frac{t}{\mu C})^2 \cdot (\mu C - \lambda_0) \cdot e^{-(\mu C - \lambda_0)t} \quad (4.17)$$

### 4.4.3 Discussion

Figure 4.5 plots the throughput $\gamma_2$ for two different service rates and two different thresholds *t*. Relative to the single -hop network, one can readily observe the sharp decline in the network throughput if two links in tandem, each with an identical capacity, were to serve the same incident traffic while maintaining the same end-to-end threshold delay. By comparing the throughput performance, we note that the maximum achievable throughput of the two-hop VoIP network is consistently lower

than that of the single-hop network for any specified level of threshold delay. Using the analytical results presented in this chapter, for specified values of the design parameters, namely the threshold delay and the incident traffic, the needed capacity of both the single-hop and the two-hop networks can be analytically evaluated. In the following analysis, we extend the results to an *n*-hop network.

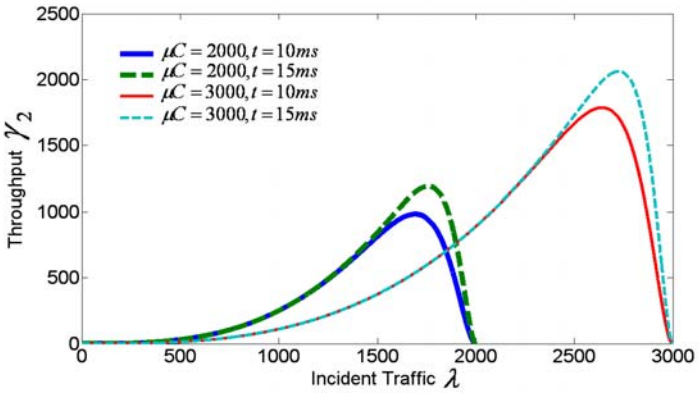


**Figure 4. 5: Two-hop Network Throughput**

## 4.5 Multiple-hop network

VoIP networks could have multiple hops. Each voice packet transmitted over the packet-switched IP network would, generally speaking, transit over multiple hops connected in series.

### 4.5.1 Analysis of multiple-hop VoIP network

As in the literature [62], the traffic is analyzed under the assumption that voice packets continue to follow the Poisson discipline at each intermediate node. Since we are interested in quantifying the impact of a multiplicity of hops, we can, without loss of generality, assume the server capacity and the arriving traffic at each node is identical. Only the last or the exit node drops the packets that have suffered delay

higher than the threshold delay *t*. The impact of multiple hops on throughput is captured in *Thereom 2*.

*Thereom 2*: The maximum throughput of a n-hop VoIP network, where each node is characterized as an individual M/M/1 system characterized with the same service rate $\mu C$ and the threshold delay *t*, is given as:

$$\gamma_{n\max} = \frac{\lambda_0^{n+2}}{(n-1)!} \cdot (\frac{t}{\mu C})^n \cdot (\mu C - \lambda_0) \cdot e^{-(\mu C - \lambda_0)t}$$

This condition holds when the incident traffic $\lambda_0$ satisfies the following condition,

$$e^{(\mu C - \lambda_0)t} = \frac{\lambda_0 t[(\mu C - \lambda_0)t]^{n-1}}{(n+1)\cdot(n-1)!} + \sum_{k=1}^{n} \frac{[(\mu C - \lambda_0)t]^{n-k}}{(n-k)!}$$

*Proof*: From (4.11), the density function $f_3(t)$ of the waiting time for the three-hop network, can be obtained by convolution of the pdf associated with the first two-hop $f_2(t)$ and the third hop $f_1(t)$.

$$f_3(t) = \int_{-\infty}^{\infty} f_2(x) f_1(t-x)dx = \frac{1}{2} t^2 [\mu C\rho(1-\rho)]^3 e^{-\mu C(1-\rho)t} \tag{4.18}$$

Continuing this convolution for each following hop, we can deduce the pdf for total waiting time in n-hop network.

$$f_n(t) = \frac{t^{n-1}}{(n-1)!} [\mu C\rho(1-\rho)]^n e^{-\mu C(1-\rho)t} \tag{4.19}$$

From (4.12) and (4.19) we have

$$P(W_n \le t) = \rho^n (1 - e^{-\mu C(1-\rho)t} \sum_{k=1}^{n} \frac{[\mu C(1-\rho)t]^{n-k}}{(n-k)!}) \tag{4.20}$$

Therefore, the throughput for the *n*-hop network is given as,

$$\gamma_n = \lambda \cdot P(W_n \le t) = \frac{\lambda^{n+1}}{(\mu C)^n}(1 - e^{-(\mu C-\lambda)t}\sum_{k=1}^{n}\frac{[(\mu C-\lambda)t]^{n-k}}{(n-k)!}) \quad (4.21)$$

Let $\frac{d\gamma_n}{d\lambda} = 0$, we have

$$e^{(\mu C-\lambda_0)t} = \frac{\lambda_0 t[(\mu C-\lambda_0)t]^{n-1}}{(n+1)\cdot(n-1)!} + \sum_{k=1}^{n}\frac{[(\mu C-\lambda_0)t]^{n-k}}{(n-k)!} \quad (4.22)$$

And the maximum throughput of n-hop network can be obtained as:

$$\gamma_{n\max} = \frac{\lambda_0^{n+2}}{(n-1)!}\cdot(\frac{t}{\mu C})^n \cdot(\mu C-\lambda_0)\cdot e^{-(\mu C-\lambda_0)t} \quad (4.23)$$

This proves Theorem 2.

### 4.5.2 Discussion

Compared to the single-hop network which can achieve a hundred percent throughput if the upper delay bound goes to infinite, for the *n*-hop network, the maximum value of throughput can only reach $\rho^n$%.

## 4.6 Simulation Results

This section is intended to corroborate the analytical results presented above through the simulation of actual single-hop and multi-hop VoIP networks. The MATLAB simulator and ANSI C are used.

### 4.6.1 Simulation Scenario

In the platform, the system is considered at the network layer, i.e., the layer above the physical and the MAC layers. Each hop in VoIP networks is assumed to have a buffer with infinite memory. Packets waiting in the buffer are transmitted

following the First-In-First-Out (FIFO) discipline. The incident traffic at each node is characterized by the Poisson distribution with the arrival rate $\lambda$. For M/M/1, the service rate follows negative exponential distribution with average value $\mu$. The incoming packets are generated randomly and independently, and the probability of generating *n* packets at time interval *t* is given by

$$P_n(t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \tag{4.24}$$

which is the Poisson distribution.

The time interval for the next generated packet is obtained using [83],

$$t = -\frac{1}{\lambda}\log(1-x) \tag{4.25}$$

where *x* is the uniform random number between 0 and 1.

At each simulation iteration a packet is generated. The separation between two successive generated packets is determined by (4.25). On the other hand, the number of served packets within these time intervals is logged as well. Hence, the buffer status at each simulation iteration will be updated by adding 1 packet and subtracting the number of packet served within the time interval between the current generated packet and the previous one. The simulation assumed unlimited buffer size (There is no overflow situation) and took into consideration the underflow condition. The simulation logs the buffer status (number of packets in the buffer) along with the simulation time, and is used to calculate the delay statistics.

### 4.6.2 Simulation results

Two scenarios have been considered, the first one is for one-hop network, and

the other is for two-hop networks with identical mean arrival rate $\lambda$ for each buffer and identical mean service time $\frac{1}{\mu}$. Figure 4.6 shows the cumulative distribution function (CDF) of the waiting time in and 4.7 describe the single-hop system, and Figures 4.8 and 4.9 the two-hop system. As shown in the figures, the simulation results closely match the analytical results derived in the previous sections.

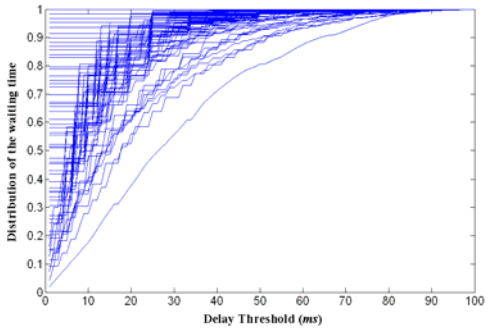


**Figure 4. 6: Simulation of the delay distribution function of the single-hop**

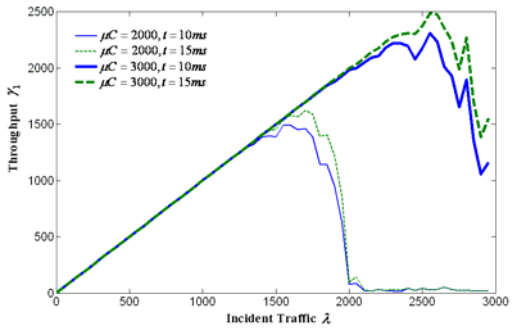


**Figure 4. 7: Simulation of the throughput in the single-hop network**

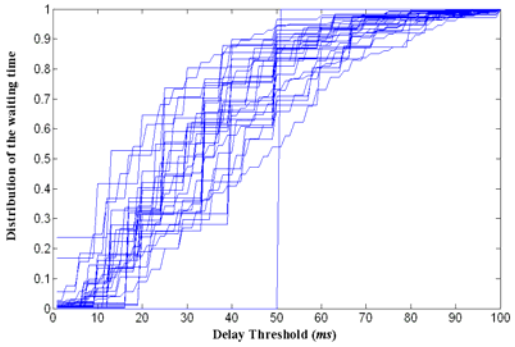


**Figure 4. 8: Simulation of the delay distribution function of the two-hop**

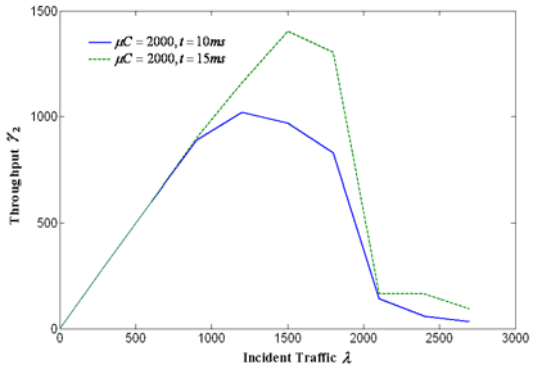


**Figure 4. 9: Simulation of the throughput in the two-hop network**

## 4.7 Conclusion

This chapter has presented a closed form solution relating the impact of bounding delays in a VoIP network to a specified limit. Our focus has been on computing throughput of a network if the traffic that suffered higher than a specified threshold delay did not constitute throughput. The results obtained will be useful in sizing up resources of the network so as to meet a specified maximum end-to-end delay criterion. The results obtained can be used to compute the maximum traffic bearing capacity of the network. The negative impact of transiting through a large number of hops between the source and the destination can be readily observed. In the next chapter, we will provide the analytical solution and simulation performance of another model M/D/1.

# Chapter 5. Impact of Bounded Delay on Resource Consumption-M/D/1 Model

**Abstract:** This chapter investigates the impact of bounded delays on throughput in VoIP networks modeled by M/D/1. Chapter 4 modeled the VoIP traffic as an M/M/1 system. Although the M/M/1 system is easier to analyze, the M/D/1 assumption for analyzing VoIP system might be more realistic in some cases. In contrast to contemporary literature that largely focuses on average delay to estimate the Quality of Service, the proposed model focuses on an upper bound of delay, referred to as delay threshold in this chapter. Traffic that exceeds the delay threshold is treated as lost throughput. The results obtained can be used in scaling resources in a VoIP Network for attaining specified levels of throughput under different thresholds of acceptable delays. Both single-hop and multi-hop transfers are addressed. The theoretical analysis presented in this chapter is further corroborated by simulation. The contents of this chapter have partially been published in [84].

## 5.1 Introduction

As mentioned in Chapter 4, there are two important parameters that capture the performance of a VoIP network: The end-to-end delay and the throughput. In contemporary literature on VoIP networks [85], the average delay has been almost always used as a measure of the delay. However, as we know, the delay will increase

as the packets traverse multiple hops, and a large delay is unacceptable for the real-time applications [86]. Therefore, most real-time applications would benefit more from the lower upper bound delay than the lower average delay. Accordingly, this chapter addresses the impact of acceptable upper delay bounds on throughput in VoIP Networks. We define effective throughput of the network as throughput that has suffered delay below the specified maximum.

The rest of the chapter is organized as follows: Section 5.2 addresses the queuing network model based on M/D/1 system; Section 5.3 presents the analytical results for VoIP networks consisting of single hop, two hops and multiple hops respectively; Section 5.4 confirms the results of the analytical solution by simulation; Section 5.5 presents the conclusion of our work.

## 5.2 Analytical Network Model

Most analyses involving packet switched systems assume that packets are negative exponentially distributed [87]. Since VoIP networks are largely deployed to capture and transport formatted information, it's more appropriate to assume a constant length for the size of each message. The M/D/1model can be more appropriate for an IP network transporting voice where all voice codecs produce flows of packets of the same size, implying that all voice packets have the same deterministic service time [88]**.** The M/D/l model can be extended to multiple channels as M/( $D_1 + D_2 + ... + D_n$ )/1 model shown in [89-90] for a VoIP network where the voice flows are produced by *n* types of codecs. This chapter will focus on the VoIP networks based on M/D/1 queues. Each of these queues is served using the

FCFS (First Come First Serve) discipline.

An M/M/1 system is characterized by two parameters, the average packet arrival rate $\lambda$ and the average service time $\frac{1}{\mu}$. An equivalent M/D/1 system would have a constant service time of $\frac{1}{\mu}$ and an arrival rate of $\lambda$. We use the parameter *h* to indicate the number of hops between the source and the destination. We further assume that, for multi-hop traffic, the traffic at each transit node continues to follow the Poisson model for the arrival process [91-92].

## 5.3 Delay-Throughput Analysis

### 5.3.1 A single-hop Network

We first consider a single-hop system, where the packets are transmitted from the source node to one of the adjacent nodes.

The distribution function of the waiting time $W_1$ in a single-hop M/D/1 system is given by [93],

$$p_1 = P(W_1 \le t) = (1-\lambda)\sum_{j=0}^{T} \frac{[\lambda(j-t)]^j}{j!} e^{-\lambda(j-t)} \tag{5.1}$$

where $t = T + \tau$, i.e., $T = \lfloor t \rfloor$, where $\lfloor t \rfloor$ is the largest integer less than or equal to *t*.

Therefore, the (residual) throughput of the single-hop network, i.e., the throughput that only includes packets with a queuing delay not exceeding the threshold delay, can be expressed as:

$$\gamma_1 = \lambda p_1 \tag{5.2}$$

where $\lambda$ is the arrival rate of packets.

The normalized throughput, i.e., throughput expressed as a function of the incident traffic can be given as:

$$p_1 = \frac{\gamma_1}{\lambda} \tag{5.3}$$

Figure 5.1 shows the waiting time distribution for M/D/1 system with different values of $\lambda$. The mean service time (which represents the length of packets) is 1 in each case. As shown in the Figure 5.1, the normalized throughput increases with an increase in the threshold delay, reaching 100% asymptotically. We also note from Figure 5.1 that the M/D/1 system consistently shows a higher normalized throughput for a smaller $\lambda$ or, equivalently, smaller $\rho$.

Using equations (5.2) and (5.3), Figure 5.2 depicts the throughput $\gamma_1$ as a function of the incident traffic $\lambda$ with several values of the threshold delay *t* as a parameter. It can be seen that, given the same incident traffic $\lambda$, the throughput increases as the threshold delay *t* increases. It is also interesting to find out that, for each threshold delay *t*, the throughput initially increases with $\lambda$, reaching a peak $\gamma_{\max}$, and eventually declines to zero. In other words, for a given *t*, the throughput is maximized for a specific value of $\lambda$. Unfortunately, due to the nature of equation (1), the specific value of $\lambda$ corresponding to the maximum value of throughput cannot be evaluated in an explicit form. However, it can be numerically evaluated indicating that, for a pre-specified threshold delay, throughput is maximized at a specific level of incident traffic that can be numerically derived. The interplay among the threshold delay, throughput and the incident traffic can be used

in order to design a cost efficient VoIP network that ideally meets the specified needs of the user. We also note from Figure 5.2 that the throughput falls much more sharply as $\lambda$ increases, for higher value of *t*. The design parameters of a VoIP network thus have a very high sensitivity to $\lambda$ for higher value of *t*.

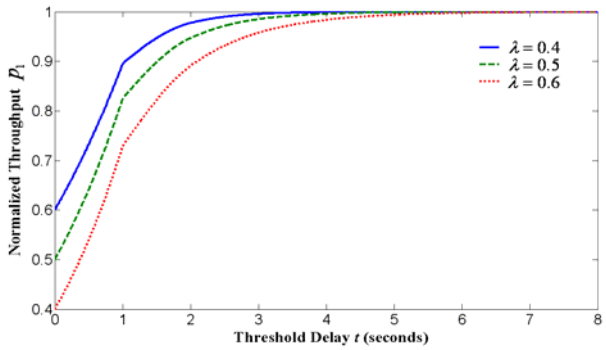


**Figure 5. 1: The waiting time distribution of a single-hop network**

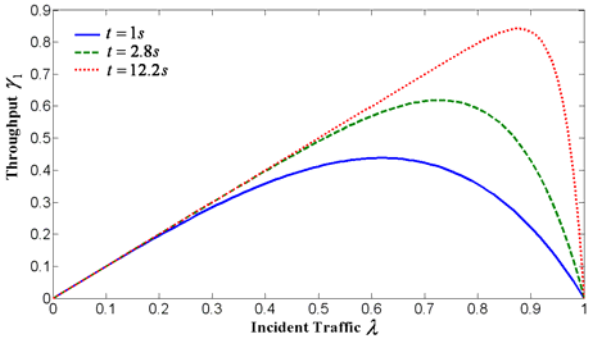


**Figure 5. 2: The throughput of a single-hop network (M/D/1)**

The results depicted in Figure 5.2 are somewhat anti-intuitive, although not difficult to comprehend under the stipulations of this investigation. If all served packets, irrespective of delay, were acceptable, the throughput would not show a decline and would reach 100% asymptotically. However, if the packets that suffered a delay higher than *t* were to be discarded and did not constitute throughput, the residual throughput would decline because a larger fraction of the packets would suffer delays higher than the threshold delay as the utilization factor (or the incident

traffic) increased. The existence of a peak at which the VoIP network would deliver the maximum throughput is an important parameter that can be used to size the resources of the network.

### 5.3.2 Two-hop Tandem network

In general, packets will go through a number of hops instead of a single hop. We first consider a two-hop tandem network. We assume identical service time distribution and arrival process for each node [58]. For purposes of analysis and insight into design, but without any loss of generality, we also assume that the first node does not drop any packets, even though it might have exceeded the delay threshold *t.* The exit node will drop any packets that have cumulatively suffered a total delay higher than *t.*

The PDF (Probability Density Function) of waiting time at the first node can be derived from equation (5.1),

$$f_1(t) = (1-\lambda)[\sum_{j=0}^{T} \frac{\lambda^{j+1}}{j!}(j-t)^j e^{-\lambda(j-t)} - \sum_{j=0}^{T} \frac{\lambda^j}{(j-1)!}(j-t)^{j-1} e^{-\lambda(j-t)}] \quad (5.4)$$

A LST (Laplace-Stieltjes Transform) for the waiting time PDF is given by

$$F_1(s) = \int_0^\infty f_{W_1}(t) e^{-st} dt = (1-\lambda)\sum_{j=0}^{T} (-\lambda)^j e^{-js} [\frac{\lambda}{(s-\lambda)^{j+1}} + \frac{1}{(s-\lambda)^j}] \quad (5.5)$$

The Laplace transform of the total (combined) waiting time PDF is equal to the product of the Laplace transforms of the waiting time PDF associated with each node. Thus,

$$F_2(s) = F_1^2(s) = (1-\lambda)^2 \{\sum_{j=0}^{T} (-\lambda)^j e^{-js} (j+1) [\frac{1}{(s-\lambda)^j} + \frac{2\lambda}{(s-\lambda)^{j+1}} + \frac{\lambda^2}{(s-\lambda)^{j+2}}]\} \quad (5.6)$$

The time-domain PDF of the two-node network can be found as,

$$f_2(t) = (1-\lambda)^2 \sum_{j=0}^{T} \frac{(-\lambda)^j}{j!} \{ j(j+1)(t-j)^{j-1} + 2\lambda(j+1)(t-j)^j + \lambda^2 (t-j)^{j+1} \} e^{\lambda(t-j)} \quad (5.7)$$

The relation between the PDF and tail-end distribution is given by,

$$F_c(t) = \int_0^t f_c(x)dx \quad (5.8)$$

From (5.7) and (5.8) we have,

$$p_2 = P(W_2 \le t) = (1-\lambda)^2 \sum_{j=0}^{T} \frac{1}{j!} \{ (j+1)[\lambda(j-t)]^j - [\lambda(j-t)]^{j+1} \} e^{-\lambda(j-t)} \quad (5.9)$$

where $p_2$ is the normalized throughput of a tandem (two-hop) system, and we have

$$\gamma_2 = \lambda p_2 \quad (5.10)$$

Figure 5.3 depicts the normalized throughput as a function of the threshold delay *t* with $\lambda$ as a parameter. It can be seen that given the same threshold delay, the normalized throughput of the two-hop system is higher for a smaller $\lambda$. In other words, a higher fraction of traffic suffers delay not exceeding the threshold delay under light low conditions. (As the threshold delay increases, the residual throughput can, however, increase even though a smaller fraction of packets constitute the residual throughput.) It can also be seen from the distribution of the waiting time in a two-hop M/D/1 system, that the normalized throughput $p_2$ can reach a maximum value of 100% as the threshod delay increases to infinity.

Figure 5.4 shows plots of the throughput $\gamma_2$ for varying levels of the incident traffic $\lambda$. The threshold delay *t* is used as the parameter. It can be seen that, as in the single-hop network case, for a given $\lambda$, the throughput is always higher (or

asymptotically equal to) for a higher value of the threshold delay. Also, and again similar to the single-hop network, we can readily observe that for a given end-to-end threshold delay, the throughput increases as $\lambda$ increases, until it reaches a maximum, and then declines to zero when the incident traffic approaches the service rate, i.e., when $\rho = 1$. Both $\gamma_{2\max}$ and the corresponding $\lambda$ can be numerically calculated from equations (5.9) and (5.10).

A comparative visualization of the single and multi-hop VoIP network performance is appropriate at this point. By comparing the throughput performance, we note that the maximum achievable throughput of the two-hop VoIP network is consistently lower than that of the single-hop network for any specified level of threshold delay. Using the analytical results presented in this chapter, for specified values of the design parameters, namely the threshold delay and the incident traffic, the needed capacity of both the single-hop and the two-hop networks can be network can be analytically evaluated. The additional transmission capacity required in the two-hop case can then be compared against, for example, the higher power requirement of the single-hop case since not all the nodes can access each other with relatively low power needed for a two-hop network.

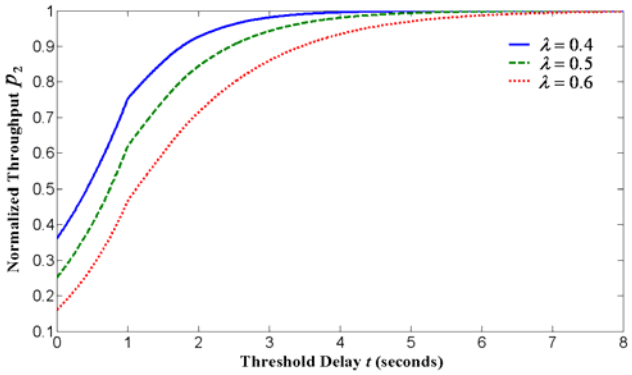


**Figure 5. 3: The waiting time distribution of a two-hop network**

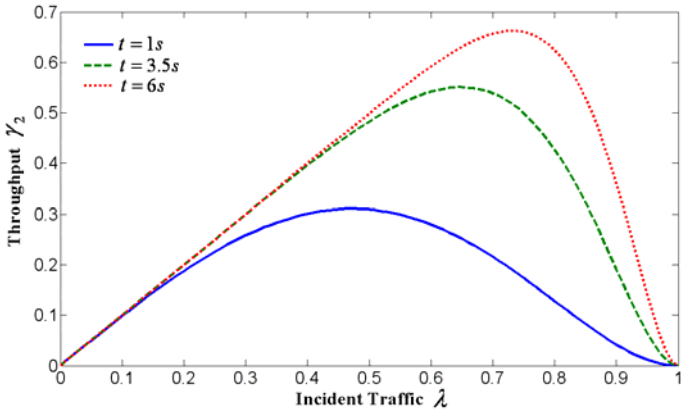


**Figure 5. 4: The throughput of a tandem network as a function of the incident traffic**

### 5.3.4 Multi-hop Network

VoIP networks could have multiple nodes for reasons cited earlier. In such networks, each packet transmitted over the VoIP can, generally speaking, transit over multiple hops connected in series. As in the two-hop case, we analyze the throughput under the assumption that packets continue to follow the M/D/1 discipline at each intermediate node [62]. Further, only the last or the exit node drops the packets that have suffered a cumulative queuing delay higher than the threshold delay *t*.

The end-to-end queuing delay in a multi-hop network is the sum of the waiting times in each hop along the series of *h*-hop [94]. Since we assume constant service time for each packet, we have,

$$W_h = \underbrace{W_1 + W_1 + \ldots + W_1}_{h} = h \times W_1 \tag{5.11}$$

The PDF of the end-to-end delay $f_h(t)$ can be derived from the convolution of the PDF of the waiting time in each hop, i.e.,

$$f_h(t) = \underbrace{f_1(t) \otimes f_1(t) \otimes \ldots \otimes f_1(t)}_{h} \tag{5.12}$$

The LST of the convolution will be the multiplication of the LST at each hop, i.e.,

$$F_h(s) = \underbrace{F_1(s) \times F_1(s) \times ... \times F_1(s)}_{h} = F_1^h(s) \tag{5.13}$$

By inverting the LST in (5.13), the end-to-end distribution will be obtained from (5.8)

$$p_h = P(W_h \le t) = (1-\lambda)^h \sum_{j=0}^{T} \sum_{i=0}^{h-1} \frac{(-1)^i}{i!\, j!} \binom{h+j-1}{i+j} [\lambda(j-t)]^{i+j} e^{-\lambda(j-t)} \tag{5.14}$$

Using equation (5.14), one can evaluate the throughput of a multi-hop VoIP network where packets suffering an end-to-end delay higher than *t* are discarded as:

$$\gamma_h = \lambda p_h \tag{5.15}$$

As in the single-hop or multi-hop cases, given a threshold delay, the optimum capacity of each node can be determined. Alternatively, given a network of specified capacity, its effective throughput for any specified threshold delay can be numerically evaluated. The throughput performance of the multi-hop network continues to deteriorate as the number of hops increases. As in the case of the two-hop network, a trade off among the performance parameters and the resource requirements can be made leading to an informed choice among the topologies under consideration.

## 5.4 Simulation Results

This section is intended to corroborate the analytical results presented above through the simulation of actual single-hop and multi-hop VoIP networks [95].

In the platform, the system is considered at the network layer, i.e., the layer above the physical and the MAC layers. Each node in VoIP networks is assumed to

have a buffer with infinite memory. Packets waiting in the buffer are transmitted following the First-In-First-Out (FIFO) discipline. The incident traffic at each node is characterized by the Poisson distribution with the arrival rate $\lambda$. For M/D/1, the service rate is fixed and is equal to $\mu$. The incoming packets are generated randomly and independently, and the probability of generating *n* packets at time interval *t* is given by

$$P_n(t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \tag{5.16}$$

which is the Poisson distribution.

The time interval for the next generated packet is obtained using [83],

$$t = -\frac{1}{\lambda}\log(1-x) \tag{5.17}$$

where *x* is the uniform random number between 0 and 1.

The simulation logs the buffer status (number of packets in the buffer) along with the simulation time and is shown in Figure 5.5, and is used to calculate the delay statistics. The delay is proportional to the number of packets queued in the buffer. Two scenarios have been considered, the first one is for one-hop network, and the other is for two-hop networks with identical arrival rate $\lambda$ for each buffer and identical service time $1/\mu$. Figures 5.6 and 5.7 describe the single-hop system, and Figures 5.8 and 5.9 the two-hop system. As shown in the figures, the simulation results closely match the analytical results derived in the previous sections.

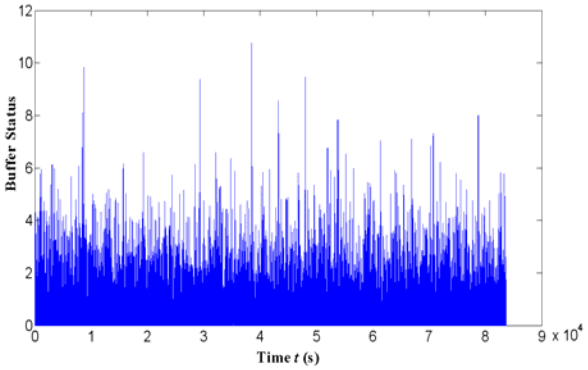


**Figure 5. 5: Buffer Status**

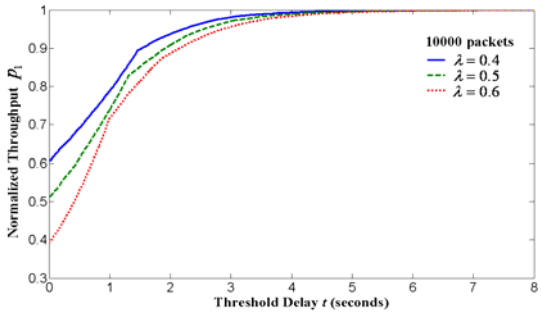


**Figure 5. 6: Simulation result of the waiting time distribution**

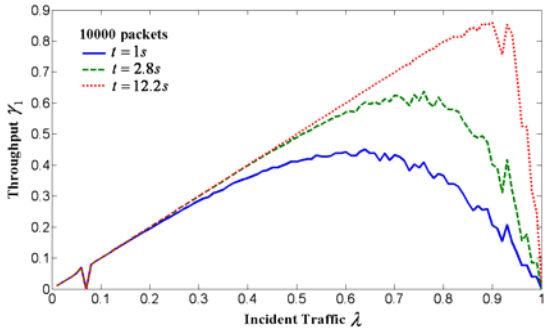


**Figure 5. 7: Simulation result of the throughput for the single-hop network**

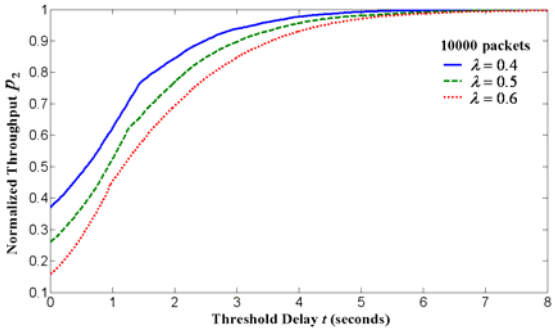


**Figure 5. 8: Simulation result of the waiting time distribution**

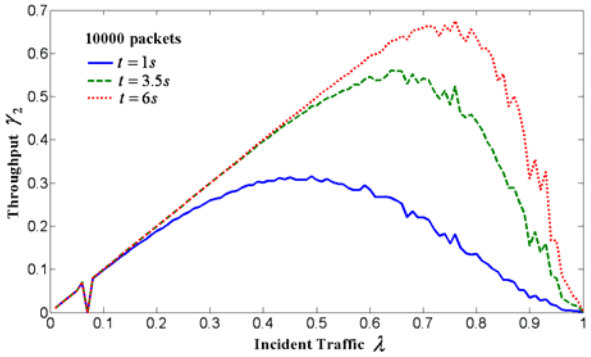


**Figure 5. 9: Simulation result of the throughput for the two-hop network**

## 5.5 Conclusion

This chapter has presented a closed form solution relating the impact of bounded delays on throughput in VoIP networks modeled by M/D/1. Our focus has been on computing throughput of a network if the traffic that suffered higher than a specified threshold delay did not constitute throughput. The results obtained will be useful in sizing up resources of the network so as to meet a specific throughput requirement with a specified maximum end-to-end delay criterion. The results obtained can be used to compute the maximum traffic bearing capacity of a network or in the design of a VoIP network with optimum throughput if the performance parameters and resource constraints were specified. The negative impact of transiting through a large number of hops between the source and the destination in terms of reduced throughput can be readily observed. The analytical results derived in this chapter have been further corroborated by simulating example networks and comparing the simulation results with the analytical results.

# Chapter 6. Impact of Bounded Jitter on Resource Consumption in Multi-hop Networks

**Abstract:** Jitter is, potentially, the largest source of degradation in the quality of voice in VoIP systems. This Chapter presents an analytical solution to the following question: How is jitter impacted by the number of hops that VoIP packets travel over and, if the end-to-end jitter were to be bounded to a predefined value, how would the resources in the network need to be scaled up as the number of hops increases? The chapter also provides a way to compute the traffic handling capability of a multi-hop resource constrained network under a defined limit of end-to-end jitter.

## 6.1 Introduction

Jitter is a potential source of quality degradation that can considerately reduce the QoS of VoIP communication. Among the parameters that define end-to-end performance, jitter is potentially the most significant. Accordingly, its containment is a major factor the design of VoIP networks [96-97]. Jitter is characterized in terms of the variance of the interval between successive packets at the receiving end *relative to that at the transmit end*. Hence, a reasonable measure of jitter in VoIP systems is in terms of the variance of the packet delay [98-99]. In this chapter, we evaluate the variance of the delay for both single-hop and multi-hop networks. The primary objective of this chapter is to present an analytical solution that answers the following question: How is jitter impacted by the number of hops that VoIP packets travel over and, if the end-to-end jitter were to be bounded to a predefined value, how would the

resources in the network need to be scaled up as the number of hops increases? The chapter also provides a way to compute the traffic handling capability of a multi-hop resource constrained network under a defined limit of end-to-end jitter. Since our focus is on relating the impact of the number of hops to QoS degradation, we make several simplifying assumptions without sacrificing the generality of our findings. We apply a Markov structure in deriving the jitter steady-state statistics for the multiplexed un-correlated traffic [100-101]. We assume that the buffer capacity at each node is infinite.

The chapter is organized as follows. Section 6.2 presents the assumptions, and generalizes the results to a two-hop and then to a multi-hop network. It also derives jitter as a function of traffic statistics in a single-hop network, i.e., when the source and the destination are separated by a single router. Section 6.3 analyzes the impact of the number of hops from a variety of perspectives. Section 6.4 presents our conclusions.

## 6.2 Jitter Analysis

### 6.2.1 Single-hop Model

We first consider voice traffic served by a single server which is our model for a single-hop network. We model the arriving VoIP packets as M/M/1 traffic [76-78].

In this model, the jitter can be estimated by the variance of the delay $\sigma_D^2$ [102], given as,

$$\sigma_D^2 = \frac{1}{\mu^2 C^2 (1-\rho)^2} \tag{6.1}$$

The following section addresses how jitter accumulates as the number of hops increases.

### 6.2.2 Two-hop and Multi-hop Model

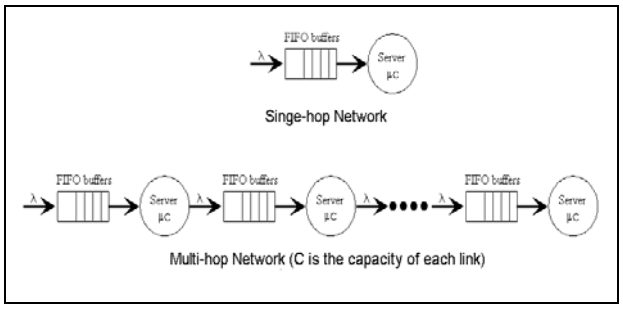


**Figure 6. 1: Single-hop and Multi-hop Networks**

Figure 6.1 shows single and multi-hop networks graphically. As in the literature [62], we compute the variance of the delay under the assumption that voice packets continue to follow the Poisson discipline for the arrival process at each intermediate hop. Since we are primarily interested in quantifying the impact of a multiplicity of hops on the accumulated jitter, we can, without loss of generality, assume that the server capacity and the traffic at each hop are identical. For the two-hop model, we have,

$$D = D_1 + D_2 \tag{6.2}$$

where $D_1$ and $D_2$ are the packet delays introduced by the first and the second server, respectively. Further, since,

$$\sigma_D^2 = D^{(2)} - (D^{(1)})^2 \tag{6.3}$$

and since the delay at each hop $D_1$ and $D_2$ are independent of each other, we can write, for the two-hop system,

$$\sigma_D^2 = [D_1^{(2)} - (D_1^{(1)})^2] + [D_2^{(2)} - (D_2^{(1)})^2] = \sigma_{D_1}^2 + \sigma_{D_2}^2 \quad (6.4)$$

which is the resultant jitter of a two-hop system.

Due to the assumption of identical traffic at each hop, the end-to-end jitter of packets that have traversed *n* hops can be obtained as,

$$\sigma_D^2 = \sum_{i=1}^{n} \sigma_{D_i}^2 = n\sigma_{D_1}^2 \quad (6.5)$$

This shows that the delay variance, and hence the jitter of the single-hop network, is lower by a factor of *n* compared to the *n*-hop network with the same bandwidth at each hop, and identical incident traffic at each node or server. We now investigate how the accumulated jitter affects the capacity needed for a multi-hop network, if the jitter were to be equal to a corresponding single-hop system.

From (6.1) and (6.5), we can express the jitter of a multi-hop network as,

$$\sigma_D^2 = \frac{n}{\mu^2 C^2 (1-\rho)^2} \quad (6.6)$$

Section 6.3 presents the cumulative impact of the number of hops on jitter in graphical form.

## 6.3 Impact of the number of hops on jitter

This section presents the impact of the findings in Section 6.2 on the end-to-end jitter from a number of perspectives.

### 6.3.1 Capacity Requirement for a n-hop Network as a function of n with a pre-defined jitter upper bound

From equation (6.6), since $\rho = \lambda / \mu C$ , we can write, after some algebraic simplification,

$$C = \frac{1}{\mu}(\lambda + \sqrt{\frac{n}{\sigma_D^2}}) \tag{6.7}$$

Equation (6.7) shows that if the end-to-end jitter were to be fixed to $\sigma_D^2$, and the traffic parameters given as $\lambda$ and $\mu$, the capacity needed at each hop can be evaluated as a function of the number of hops, n. Figure 6.2 depicts the relationship graphically. The traffic parameters used are $\lambda = 2000$ and $\mu = 1$ . The two jitter bounds used as parameters are, $\sigma_D^2 = 0.0025(s^2)$ and $\sigma_D^2 = 0.01(s^2)$ . It can be observed from Figure 6.2 that that the capacity needed *for each hop* in an n-hop network increases as the pre-defined jitter upper bound decreases. For example, for n =1, the capacity needed is 2010 *bps* for a jitter bound $\sigma_D^2 = 0.01(s^2)$ while we need a capacity of 2020 bps if the jitter bound were to be reduced to $\sigma_D^2 = 0.0025(s^2)$ , a difference of 10 bps. The corresponding figures for *n* = 4, require double the additional capacity (20 bps) to maintain the jitter bound to the lower value of $\sigma_D^2 = 0.0025(s^2)$ .

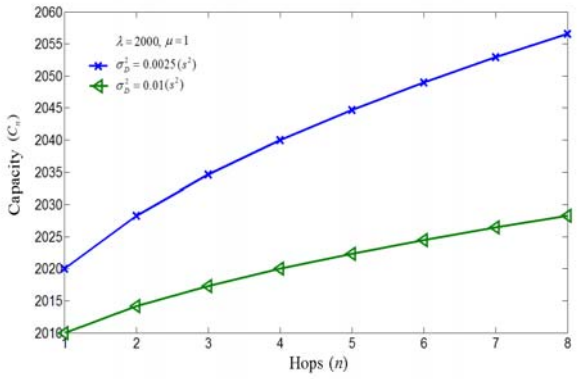


**Figure 6. 2: Capacity required as a function of n with jitter as a parameter**

### 6.3.2 The impact of the Utilization Factor on the Capacity per hop needed for a pre-defined upper bound on end-to-end jitter

In this section, we will focus on the impact of the number of hops on the capacity needed using $\rho$ as the parameter while $\mu$ and $\lambda$ are kept constant. Equation (6.6) can be rearranged to yield,

$$C = \frac{1}{\mu(1-\rho)}\sqrt{\frac{n}{\sigma_D^2}} \tag{6.8}$$

Figure 6.3 shows the results graphically. In particular, it illustrates that the capacity needed increases dramatically as the utilization factor increases. For example, the capacity needed at each hop in an eight-hop network is more than twice if the same traffic were to traverse a single hop and if the end-to-end jitter were to remain identical in each case.

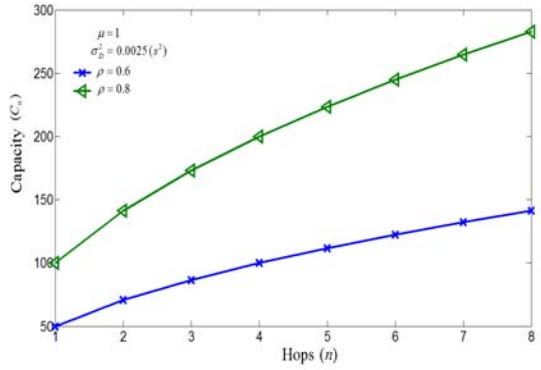


**Figure 6. 3: Capacity required as a function of n with different utilizations**

Figure 6.4 focuses on how the capacity needed varies as a function of the utilization factor. As shown previously, the smaller the value of the jitter upper bound, the more the capacity needed at each hop. Also, the capacity needed at each hop increases slowly initially, but the increase is much more rapid as the number of hops

or as the utilization factor increases. The following section addresses the impact on throughput in a resource constrained network if the jitter were to be bounded to an upper value.

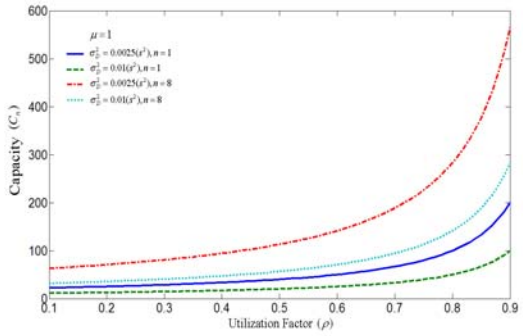


**Figure 6. 4: Capacity requirements as a function of $\rho$, with n and $\sigma_D^2$ used as parameters**

## 6.3.3 The impact on Throughput of a Resource-constrained multi-hop network with a pre-defined upper bound on end-to-end jitter

In the case of a resource constrained network, where the transmission capacity on each hop is limited to a pre-defined value *C*, the end-to-end jitter can be controlled to remain within a pre-defined upper bound, if the traffic is reduced as the number of hops n increases. Again, from equation (6.6), we can write,

$$\lambda = \mu C - \sqrt{\frac{n}{\sigma_D^2}} \tag{6.9}$$

If the end-to-end jitter were limited to $\sigma_D^2$, then for $\mu C$ = 2000, the capacity needed can be evaluated as a function of *n*. The two jitter bounds used are the same as in Figure 6.2, $\sigma_D^2 = 0.0025(s^2)$ and $\sigma_D^2 = 0.01(s^2)$. The results are graphically depicted in Figure 6.5.

It can be readily observed that, in order to keep the end-to-end jitter bounded to $0.01s^2$, the incident traffic must be reduced from 1990 packets per second in a

single-hop network to 1980 in the four-hop network. Similarly, for a lower end-to-end jitter bound of $0.0025s^2$, the incident traffic must be reduced from 1980 packets per second in the single-hop network to 1960 in the four-hop network, a reduction of twice the capacity compared to the previous case. The relationship developed in this section can be used to size the throughput (or the traffic carrying capacity) of a network serving multi-hop traffic with a specified maximum jitter bound and transmission resources limited to C bps.

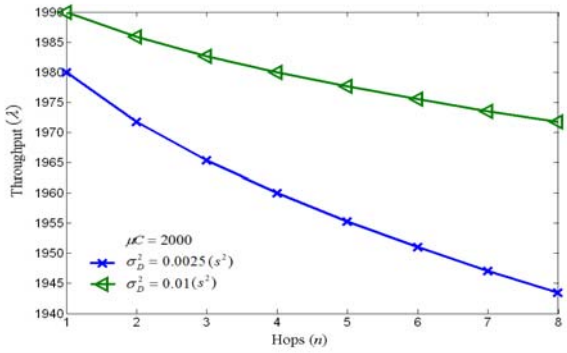


**Figure 6. 5: Throughput of the n-hop network under an upper bound jitter and specified transmission capacity C at each hop**

## 6.4 Conclusions

This chapter has derived quantitative relationship among the number of hops and end-to-end jitter in a packet switched network. Since jitter is an important factor in determining the Quality of Service of VoIP, the relationships derived can be used in sizing the resources of the network as the number of hops is increased. Alternatively, if the network resources were fixed, the relationships developed can be used to size the traffic bearing capacity of the network to serve VoIP traffic while keeping the end-to-end jitter limited to a pre-defined bound.

# Chapter 7 Cost and Quality in Packet Switched Networks: An Abstract Approach

**Abstract:** The notion of quality is an important factor in the design of any complex communication system. Since communication systems are increasingly dependent on packet switching as the underlying transport mechanism, this chapter explores the impact of levels of quality that a packet switched system can provide on cost. Several interesting results are presented that result in the notion of an optimum level of load that a packet switched system can handle given a defined upper bound on latency that a specified percentage of packets would suffer. The results presented offer a fresh insight into the relationship between cost and quality and can be used to optimize the throughput of a packet switched system under specified levels of quality. The contents of this chapter have been published in [103].

## 7.1 Introduction

The Internet is a prime example of a common user network that uses packet switching as its transport fabric. The growing reach of the Internet over the past decade together with the promise of lower costs to the customer has led to, for example,, the rapid emergence of Voice over IP [104]. VoIP, being a real-time service, requires that qualities of service (QoS) parameters, such as delay, jitter or packet loss are tightly controlled. While the relationship among these parameters can be developed using contemporary analytical techniques, the fact that a packet switched network with a specified level of transmission resources has an optimum

capacity to handle traffic with a specified level of QoS is not known. This chapter develops that relationship.

VoIP is an emerging service over the Internet. For most of the analysis and examples in this chapter, we use VoIP as the underlying application over a packet switching transport fabric.

The legacy circuit-switched network allocates dedicated 64bits/s bandwidth to each voice call resulting in virtually no delay due to queuing and no jitter. Generally speaking, PSTN provides a high level of voice quality, called toll quality [105]. VoIP packets in packet-switched networks undergo varying amounts of delay at each transit node. Delay in VoIP networks comes from three sources: processing, queuing and propagation. Therefore, to adequately meet the performance requirements of real-time application, delay becomes a significant parameter in VoIP networks. International Telecommunications Union-Telephony (ITU-T) Recommendation G.114 [31] provides one-way transmission delay specifications for voice. The delay addressed in this chapter is the *variable queuing delay*, the cumulative value of the delay each voice packet has to suffer at each router in the path of a VoIP connection.

The average delay of voice packets has been traditionally used as an indicator of the QoS of a VoIP network [1]. However, the average delay is not meaningful in real-time human conversations. Suppose two services with the same average delay of *300 ms* are available. One ranges from 100 *ms* to 500 *ms*, the other from 200 *ms* to 400 *ms*. The system with the 200-400 *ms* bounded delay provides a higher voice quality than the 100-500 *ms* bounded delay system. The upper delay bound is thus a key factor in determining the Quality of Service of VoIP systems. Accordingly, from

the user's point of view, a system that limits the upper bound of the queuing delay to a low value is very important. We note that a system with bounded upper delay necessarily controls both the delay as well as the jitter. Taken from the service provider's perspective, a larger allowable range of delay variation provides more flexibility in terms of resource provisioning. The interplay between the needs of the user in terms of the Quality of Service and the requirement it places on the service provider in terms of its impact on transmission resources to be provisioned is the primary issue addressed in this chapter.

A reasonable way to characterize the Quality of Service (QoS) in VoIP networks [106, 72] is by limiting the end-to-end queuing delay to an acceptable upper bound. We adopt this approach in this chapter. Packets that exceed the upper bound of delay (termed the threshold delay) are not counted as constituting effective throughput. The customer only pays for the effective throughput constituted by those packets that are within the delay threshold that characterizes the QoS received [107, 108]. As expected, VoIP traffic suffers higher queuing delay with increasing utilization of the link [79]. However, a higher utilization will also result in higher effective throughput under our construct. Increasing the quality of service is thus tantamount to lowering the throughput that the packet switched service and, accordingly, a lower chargeable throughput. We explore the possibility of an optimum utilization of the link that results in the maximum throughput while at the same time keeping the QoS (as measured through the threshold delay), within the defined bound. Further, this chapter also explores the impact of the number of hops on the end-to-end queuing delay of a voice packet. The resulting relationship leads to

several interesting results that can be used to price Voice over IP services offered over a multi-hop network in a way that maximizes the utilization of the network ( thus maximizing its throughput) while, at the same time, keeping the end-to-end queuing delay within a defined upper bound. We use the results to appropriately price VoIP services at equitable levels that are consistent with the resources consumed in order to achieve the contracted QoS.

The rest of the chapter is organized as follows: Section 7.2 presents a QoS based pricing model; Section 7.3 presents a mathematical model of VoIP networks; Section 7.4 presents the analytical results of pricing for the single-hop VoIP networks; Section 7.4 and 7.5 analyze the two-hop and multi-hop networks, respectively, as well as the pricing schemes; Section 7.6 consisting of one-node, two-node or multiple-node respectively; Section 7.7 presents the conclusion of our work.

## 7.2 A QoS Based Pricing Model

An important issue in designing pricing policies for today's networks is to balance the trade-off between traffic engineering and economic efficiency [16, 17]. A recent work [109] has addressed the impact of multiple hops (or switches between the ingress and egress switching nodes) on the grade of service offered by a circuit switched telephone network. The analytical results presented in that chapter have led to proposing a new pricing scheme based on the cost of lost opportunity vs. the cost of consumption of resources, which is the contemporary practice. We adopt a similar approach in this chapter in the context of packet switching. Accordingly, the grade of service is replaced by the threshold delay which is an appropriate measure for perceived Quality of Service in a packet switched network. In a circuit switched

network, an incomplete call is lost and does not generate any revenue. In the present analysis of packet switched networks, there are no calls that are lost as such; however, some of the packets may suffer delays above the acceptable delay bound and are, similarly, not considered to provide effective throughput. Just as a caller in a circuit switched network does not pay for an incomplete call, the VoIP caller over a packet switched network in our construct does not pay for packets that suffer an unacceptable level of delay.

## 7.3 Mathematical Model of a VoIP Network

A typical voice packet would pass through several nodes before arriving at the destination. We consider the voice traffic served by a single server as well as multiple servers. A server effectively constitutes a node of the network. Two or more servers are connected by transmission lines.

Consider a LAN shown in Figure 4.2 with a VoIP (e.g., a SIP) server that functions as a VoIP network. We model the VoIP packets arriving at the SIP server as M/M/1 traffic [76]. A two-node tandem network is shown in Figure 4.4. As in the literature [62], we analyze the quality of service under the assumption that voice packets continue to follow the Poisson discipline at each intermediate node. Since our major interest in this chapter is to understand the impact of multiple nodes on quality of services, without loss of generality, we can assume the server capacity and the arriving traffic at each node to be identical.

We assume that all traffic served by the first node forms the incident traffic for the second node, *even if it was delayed beyond the threshold t.* In other words, the policy of discarding traffic with a delay higher than *t* is executed by the exit node.

## 7.4 Analysis of the Single-hop VoIP Network

This section considers the impact of bounding delays on the capacity needed and the resulting throughput of the single-hop VoIP network.

### 7.4.1 Threshold Delay, Resource Consumption and Throughput

For an M/M/1 system, the probability that a voice packet suffers a delay less than $t$ is given by [58, 81],

$$p_1 = P\{W_1 \le t\} = 1 - \rho e^{-\mu C_1 (1-\rho)t} \tag{7.1}$$

Therefore, the throughput of the single-hop VoIP network where all packets that undergo a queuing delay not exceeding the threshold delay can be expressed as:

$$\gamma_1 = \lambda p_1 \tag{7.2}$$

where $\lambda$ is the arrival rate of packets.

The normalized throughput, i.e., throughput expressed as a function of the incident traffic can be given as:

$$p_1 = \frac{\gamma_1}{\lambda} \tag{7.3}$$

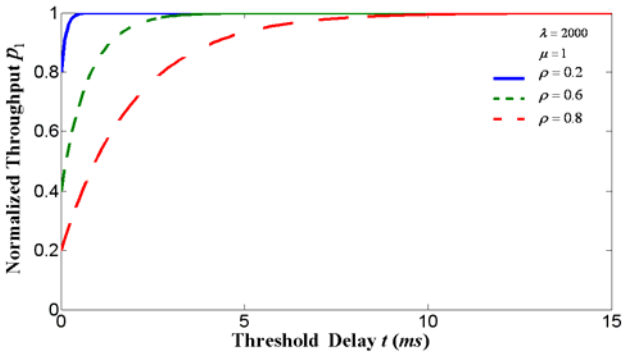


**Figure 7. 1: Normalized Throughput of Single-hop Network**

Using equation (7.1), Figure 7.1 depicts the normalized throughput as a function of the threshold delay $t$ with several values of the utilization factor

$\rho$ ( $= \frac{\lambda}{\mu C_1}$ ) as a parameter. The other parameters used are shown in the figure, $\lambda = 2000$ and $\mu = 1$, which are applied through the whole chapter. It can be seen that, given the same traffic intensity $\rho$ , the normalized throughput increases as the threshold delay increases, while given the same threshold delay *t*, the normalized throughput increases as the traffic intensity decreases. In other words, the normalized throughput of light traffic load is higher than that of the heavy traffic load. In general, the customer would like a lower value of the threshold delay *t*, while the service provider's profit will increase as *t* increases, because it would result in a higher throughput, which is paid for by the customer.

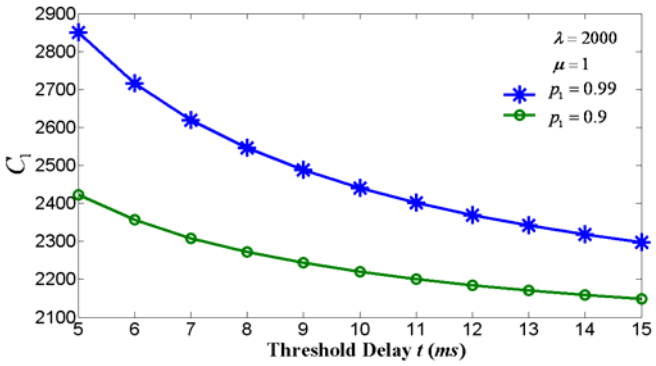


**Figure 7. 2: Capacity as a function of threshold delay**

Equation (7.1) can also be used to numerically solve for the capacity $C_1$ needed for varying threshold delay t, with the normalized throughput $p_1$ used as a parameter, given values of $\lambda$ and $\mu$ as shown in Figure 7.2. It can be seen, as expected, that the capacity needed reduces as the threshold delay increases. In other words, a higher value of threshold will require a lower capacity. From Figure 7.2, we can see that for a threshold delay of 8*ms*, a 99% throughput is obtained for a specific

value of $C_1 = 2550$ *bps*. On the other hand, using a link capacity of 2280 *bps*, only 90% packets are received with the same threshold delay.

From Figure 7.1, it can be seen that the *relative gain* in normalized throughput reduces as the threshold delay is increased, even as the normalized throughput monotonically increases. It would be interesting to determine if the resulting throughput per unit of bandwidth used shows a point of inflexion.

From equations (7.1) and (7.2), we have,

$$\frac{\gamma_1}{C_1} = \frac{\lambda}{C_1}[1 - \frac{\lambda}{\mu C_1} e^{-(\mu C_1 - \lambda)t}] \tag{7.4}$$

A point of inflexion can be determined by putting

$$\frac{d(\gamma_1 / C_1)}{dC_1} = 0 \tag{7.5}$$

which results in:

$$\mu C_{1opt} e^{(\mu C_{1opt} - \lambda)t} = \lambda(2 + \mu C_{1opt} t), \tag{7.6}$$

where $C_{1opt}$ is the optimimum capacity needed for the system to reach the maximum throughput per unit of bandwidth. Equation (7.4) shows a maximum value because

$$\frac{d^2(\gamma_1 / C_1)}{dC_1^2} = \frac{2\lambda}{C_1^3}\{1 - \frac{\lambda}{2\mu C_1}[(\mu C_1 t + 2)^2 + 2] e^{-(\mu C_1 - \lambda)t}\} \tag{7.7}$$

is negative.

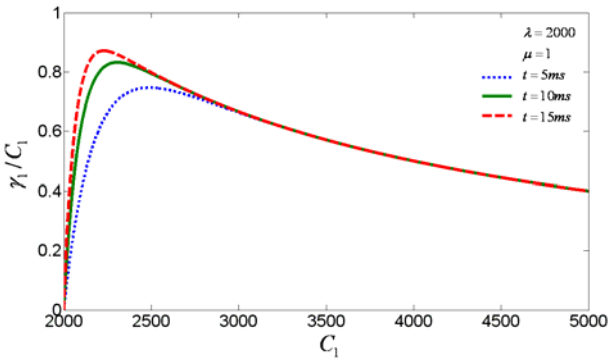


**Figure 7. 3: Throughput/capacity as a function of capacity**

Using equation (7.4), $\frac{\gamma_1}{C_1}$ is plotted as a function of the bandwidth $C_1$ with several threshold delays used as a parameter, as shown in Figure 7.3. From the service provider's perspective, the capacity $C_1$ is best utilized when the maximum $\frac{\gamma_1}{C_1}$ is at its peak. Thus, given a threshold delay t, the capacity $C_{1opt}$ is the optimum capacity that the network requires. At this capacity, the user's requirement of delay are met while the throughput delivered per unit of bandwidth (or resource consumption) is maximized. From Figure 7.3, we also notice that the maximized throughput per unit of bandwidth used increases as the threshold delay increases. In other words, a higher threshold delay results in a better utilization of the bandwidth.

We can evaluate the normalized throughput corresponding to the optimized capacity as follows. This will allow us to compute the throughput for which the customer will be actually charged.

From equations (7.6) and (7.1) we get,

$$p_{1opt} = 1 - \frac{1}{2 + \mu C_{1opt} t} \tag{7.7}$$

where $p_{1opt}$ is the normalized throughput corresponding to the optimum utilization of the bandwidth. . Figure 7.4 plots the optimum capacity along with the normalized throughput as a function of the threshold delay. It can be easily observed that, as the threhold delay increases, the optized capacity reduces while the normalized throughput increases.

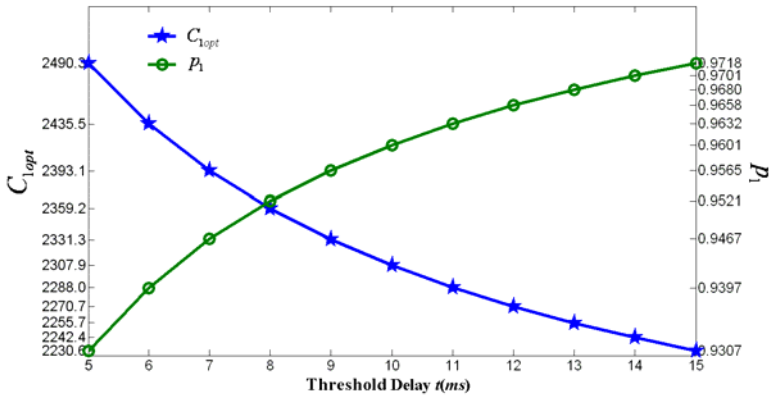


**Figure 7. 4: Optimized capacity and normalized throughput as functions of threshold delay**

It would be interesting to observe the impact of unit capacity increase on the corresponding reduction in the threshold delay $t$ . Suppose the capacity necessary for supporting the threshold delay $t_0$ is $C_0$. If a specific customer were to subscribe for a higher QoS typified by a lower threshold delay *t*, where *t* is less than $t_0$ ($t < t_0$), then in order to achieve the same throughput, the new capacity $C_1$ needed can be obtained from equation (7.2) as follows;

$$\gamma_1 = \lambda[1 - \frac{\lambda}{\mu C_1} e^{-(\mu C_1 - \lambda)t}] = const \qquad (7.8)$$

Assuming $\lambda$ and $\mu$ are fixed, the relation between the single-hop capacity $C_1$ and the threshold delay *t* can be computed by differentiation over *t* on both sides of equation (7.8),

$$\{\frac{\lambda^2}{\mu C_1^2}\frac{dC_1}{dt}+\frac{\lambda^2}{\mu C_1}[\mu t\frac{dC_1}{dt}+(\mu C_1-\lambda)\}e^{-(\mu C_1-\lambda)t}=0 \tag{7.9}$$

With simplification, we have,

$$\frac{dC_1}{dt}=-\frac{(\mu C_1-\lambda)C_1}{1+\mu C_1 t} \tag{7.10}$$

The negative value illustrates the inverse relationship between an additional unit of capacity and the threshold delay. Using equation (10), Figure 7.5 shows the impact of threshold delay $t$ on $\Delta C$, with $\rho$ and $\Delta t$ as parameters. It can be seen that for the single-hop network, a smaller $\Delta C$ is required in order to reduce the threshold delay by $\Delta t$ as t increases to maintain the same throughput. We also observe that given $t$ and $\Delta t$, a lighter traffic load (smaller $\rho$) results in larger capacity increase $\Delta C$ compared to when the network carries a heavy traffic load.

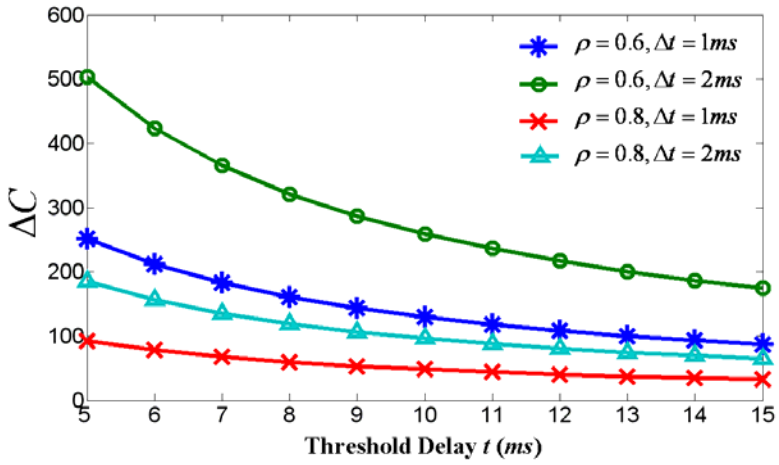


**Figure 7. 5: $\Delta C$ as a function of threshold delay**

## 7.4.2 Pricing for Single-Hop Network

As we have seen already, from equation (7.6), given the characteristics of the incident traffic in terms of $\lambda$ and $\mu$, and a given requirement of the Quality of

Service in terms of the threshold delay *t*, the capacity needed to fulfill that requirement can be numerically evaluated.

In our VoIP network model, the discarded traffic will not generate any revenue from the customer. The service provider can only charge the customer of the served traffic [110]. The amount of served traffic $\lambda p_1$ will determine the revenue for the service provider.

In the previous analysis, the single-hop VoIP network can function most efficiently when the capacity is best utilized [111]. In our pricing model for providing services with different QoS in the single-hop VoIP network, we treat the used capacity as the underlying cost and the throughput as the revenue. For the service provider, $C_{1opt}$ represents the cost of providing service. The throughput achieved under this condition is $\lambda p_{1opt}$. The cost of unit service (assuming bandwidth equals cost) is thus $\frac{C_{1opt}}{\lambda p_{1opt}}$. A customer achieving the throughput $\gamma$ will thus be charged an amount equal to $\frac{C_{1opt}}{\lambda p_{1opt}} \times \gamma$.

If the service provider were to choose another threshold delay $t'$, the necessary bandwidth $C'_{1opt}$ as well as the corresponding $p'_{1opt}$ can be similarly calculated. Then the unit price for throughput changes to $\frac{C'_{1opt}}{\lambda p'_{1opt}}$, and an individual customer realizing the throughput $\gamma$ will be charged an amount equal to $\frac{C'_{1opt}}{\lambda p'_{1opt}} \times \gamma$ .

We use the parameters presented in Figure 7.4 as a specific example. We have $\lambda = 2000$, $\mu = 1$ and $t = 5ms$. We compute $C_{1opt} = 2490.3$ and $p_{1opt} = 0.9307$. Therefore, the unit price can be obtained as $\frac{C_{1opt}}{\lambda p_{1opt}} = \frac{2490.3}{2000 \times 0.9307} = \$1.34/bps$. If we increased the threshold delay to $t' = 6ms$, $C_{1opt}$ decreases to be $C'_{1opt} = 2435.5$, and the normalized throughput increases to be $p'_{1opt} = 0.9397$. The new unit price can be calculated the same way as before $\frac{C'_{1opt}}{\lambda p'_{1opt}} = \$1.3/bps$, which is less than the unit price charged for the first service with 1*ms* less delay threshold.

In view of the results presented in equations (7.6) through (7.7) and the above example, we can state the pricing strategy for a single-hop VoIP network as follows.

For the traffic parameters $\lambda$, $\mu$ and and the quality of service defined by the threshold delay *t* associated with user's needs, compute the optimum channel capacity $C_{1opt}$ from equation (7.6). The norminal unit price for the network is computed as,

$$\text{Unit price} = \frac{C_{1opt}}{\lambda p_{1opt}} \tag{7.11}$$

For a particular customer generating traffic at the rate of $\lambda_1$ ($\lambda_1 < \lambda$), the price can be determined as:

$$\frac{C_{1opt}}{\gamma} \times \gamma_1 = \frac{C_{1opt}}{\lambda p_{1opt}} \times \lambda_1 p_{1opt} = C_{1opt} \frac{\lambda_1}{\lambda} \tag{7.12}$$

which is proportional to the price of the total incident traffic.

## 7. 5 Analysis of Two-hop VoIP Network

### 7.5.1 Comparisons of two-hop and single-hop traffic performance

We now consider the throughput of a VoIP network with two hops between the sending and receiving nodes. The arrival process of the traffic incident on the second node is assumed to be Poisson as well [78]. As described in Section 7.2, the last or the exit node drops the packets that have suffered delay higher than the threshold delay *t*. The probability density function (pdf) of the waiting time $f_{W_2}(t)$ in the two-hop network can be computed by convolving the corresponding the pdf's of the waiting time at each node $f_{W_1}(t)$ [82]. Therefore, the waiting time distribution can be given as shown in (4.13),

$$p_2 = P(W_2 \le t) = \rho^2\{1 - e^{-\mu C_2(1-\rho)t}[1 + \mu C_2(1-\rho)t]\} \tag{7.13}$$

The throughput can now be given as

$$\gamma_2 = \lambda p_2 \tag{7.14}$$

Therefore, the normalized throughput for two-hop system can be expressed as

$$p_2 = \frac{\gamma_2}{\lambda} \tag{7.15}$$

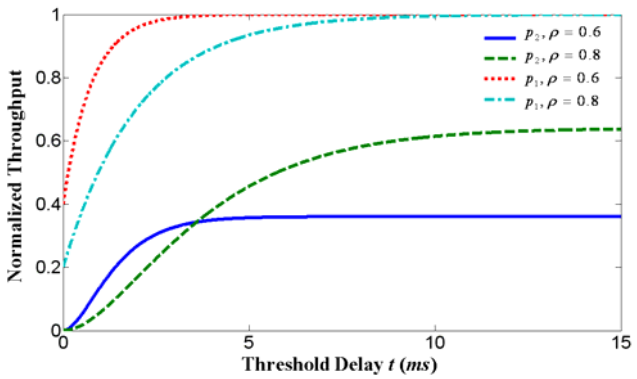


**Figure 7. 6: Normalized throughput for single-hop and two-hop systems**

We have plotted equations (7.1) and (7.13) in Figure 7.6 for specified values of $p_2$ and $\rho$. The same figure also reproduces corresponding curves for the single node network. We observe the following two characteristics: the normalized throughput for both the single-hop and two-hop networks increases as a function of *t;* however, in contrast to the single-hop network, the two-hop network normalized throughput never reaches 100%. This can also be observed from equation (7.13), where $p_2$ will always remain lower than $\rho^2$. A two-hop network can thus never achieve 100% throughput if we measured throughput as served traffic that suffers delay below a specified upper bound, *t.*

Another interesting observation from Figure 7.6 is that, for the two-hop network, a lower utilization factor results, initially, in a higher normalized throughput; however, the higher utilization factor leads to a higher normalized throughput beyond a certain value of the threshold delay. The point of intersection where the higher utilization starts delivering a higher normalized throughput can be numerically evaluated from equation (7.13).

We further note that if the two-hop network were to meet the latency requirement, since it has two links, its capacity cost would be $2C_1$ because of the two hops involved. (We assume that the cost of capacity is independent of distance and is directly proportional to the number of hops). However, since the carried two-hop traffic is always lower than the carried single-hop traffic with identical capacity in each hop, as from equation (7.13) and as illustrated in Figure 7.6, it follows that the two-hop network should be priced more than twice the single-hop price. Mathematically, from (7.1), (7.2), (7.13) and (7.14), we have,

$$\frac{\gamma_2}{\gamma_1} = \frac{\lambda(\frac{\lambda}{\mu C_1})^2\{1-e^{-(\mu C_1-\lambda)t}[1+(\mu C_1-\lambda)t]\}}{\lambda[1-\frac{\lambda}{\mu C_1}e^{-(\mu C_1-\lambda)t}]} \tag{7.16}$$

which is always less than 1.

The actual price for VoIP services that involve two links instead of one will be determined by the enhanced capacity needed in each of the two links that would result in identical threshold delay applicable to the single-hop network. Suppose this capacity were $C_2$ for each of the two links in the two-hop network. Then, for the throughput to be identical, we must have

$$\gamma_2 = \gamma_1 \tag{7.17}$$

Or

$$(\frac{\lambda}{\mu C_2})^2\{1-e^{-(\mu C_2-\lambda)t}[1+(\mu C_2-\lambda)t]\} = 1-\frac{\lambda}{\mu C_1}e^{-(\mu C_1-\lambda)t} \tag{7.18}$$

From (7.18), the value of $C_2$ can be numerically evaluated.

Figure 7.7 plots $\frac{C_2}{C_1}$ as a function of the threshold delay *t*. We observe that the relative capacity per link for the two-hop network, decreases as the threshold delay increases. Further, given *t*, the required $\frac{C_2}{C_1}$ increases as the normalized throughput increases.

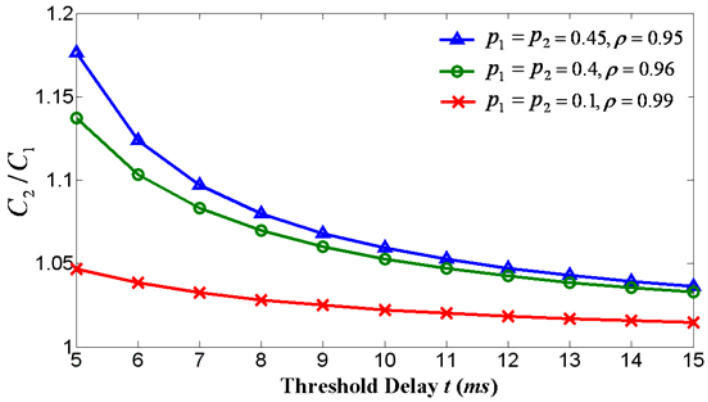


**Figure 7. 7: Relative capacity of tandem network as a function of threshold delay t**

The maximum normalized throughput $p_{2opt}$ can be obtained by putting the first derivative of $p_2$ equal to zero. In other words, $p_2$ will be maximized for the specific $C_2$ such that

$$\frac{dp_2}{dC_2} = 0, \quad \text{where} \quad \frac{d^2 p_2}{dC_2^2} < 0 \tag{7.19}$$

(The inequality can be easily shown to be true.)

(7.19) results in the following transcendental equation:

$$2e^{(\mu C_{2opt} - \lambda)t} = 2 + (2 + \mu C_{2opt} t)(\mu C_{2opt} - \lambda)t \tag{7.20}$$

$C_{2opt}$ is the optimum capacity, at which the normalized throughput $p_2$ of a two-hop VoIP Network, characterized by a fixed arriving traffic $\lambda$ and all packets incurring a queuing delay higher than $t$ are discarded, is maximized.

## 7.5.2 Pricing for Two-Hop Network

Similar to the pricing model for the single-hop network, for the two-hop VoIP network with design parameters $\lambda$, $\mu$ and $t$, the unit price for the two-hop traffic can

be observed to be $\$ \frac{2C_{2opt}}{p_{2opt}}$/bps, where $p_{2opt}$ is the normalized throughput using $C_{2opt}$ as the capacity of each of the two links. With another chosen threshold delay $t^{'}$, the necessary bandwidth $C_{2opt}^{'}$ per link and maximium normalized throughput $p_{2opt}^{'}$ can also be calculated from equations (7.20) and (7.13). In that case a different unit price $\$ \frac{2C_{2opt}^{'}}{\lambda p_{2opt}^{'}}$ will be charged to the customer subscribing to this service. The total price will be $\frac{2C_{2opt}^{'}}{\lambda p_{2opt}^{'}} \times \gamma$ based on the served traffic $\gamma$ .

As an example, given $\lambda = 2000$, $\mu = 1$ and threshold delay $t = 5ms$, the numerical results are calculated as $C_{2opt} = 2649.9$ and $p_{2opt} = 0.4757$. Therefore, the unit price will be $\frac{2C_{2opt}}{\lambda p_{2opt}} = \$5.57/bps$. Recall from the example in Section 7.4 applicable to the single-hop traffic under the identical conditions that $C_{1opt} = 2490.3$ and the unit price $\frac{C_{1opt}}{\lambda p_{1opt}} = \$1.34/bps$ . After comparision of $\frac{C_{2opt}}{C_{1opt}} = 1.064$ and $\frac{2C_{2opt}/\lambda p_{2opt}}{C_{1opt}/\lambda p_{1opt}} = 4.1567$, we observe that the unit price for two-hop traffic is more than four times the unit price for single-hop traffic. The increased price arises because of two reasons: (1) Two links of identical capacity ar needed, and (2) The bandwidth of each of the two links is higher because the required throughput for links of identical capacity is much lower in the case of two links in tandem. We consider another example when the threshold delay changes to $t^{'} = 6ms$. We have

$C'_{2opt} = 2571.8$ and $p'_{2opt} = 0.5180$. The unit price = $\frac{2C'_{2opt}}{\lambda p'_{2opt}} = \$4.9649/bps$, which is less than the capacity needed for the lower threshold delay of 5 *ms*. Compared to the single-hop traffic with the same threshold delay $t' = 6ms$, $C'_{1opt} = 2435.5$ and $\frac{C'_{1opt}}{\lambda p'_{1opt}} = \$1.3/bps$, we have $\frac{C'_{2opt}}{C'_{1opt}} = 1.056$ and $\frac{2C_{2opt}/\lambda p_{2opt}}{C_{1opt}/\lambda p_{1opt}} = 3.8192$. The price for a higher threshold delay thus shows an improvement compared to the price for a lower threshold delay case. The example also shows that both the ratios of the capcity per each link and the unit price of two-hop network relative to single-hop network decrease as the threshold delay increases. With the analysis and examples presented above, we can now state the pricing strategy for a two-hop VoIP network as follows.

For the design parameters, $\lambda$, $\mu$ and $t$ associated with a two-hop network, compute the optimum channel capacity $C_{2opt}$ from equation (7.20). The unit price for the network is now computed as,

$$\text{Unit price} = \frac{2C_{2opt}}{\lambda p_{2opt}} \tag{7.21}$$

For a specific customer with incident traffic of $\lambda_2$ ($\lambda_2 < \lambda$), the price will be proportional to the price of the total incident traffic as

$$\$\frac{2C_{2opt}}{\lambda p_{2opt}} \times \lambda_2 p_{2opt} = \$2C_{2opt}\frac{\lambda_2}{\lambda} \tag{7.22}$$

## 7.6 Analysis of Multi-hop VoIP Network

For the multi-hop network, we assume that voice packets continue to follow the Poisson discipline at each intermediate node [80]. Under this assumption, $p_h$ is computed by the convolution of probability density function (pdf) of waiting time associated with the first ($h$-1)-hop and the $h$ th hop, as derived in (4.20). We get,

$$p_h = P(W_h \le t) = \rho^h (1 - e^{-\mu C(1-\rho)t} \sum_{k=1}^{h} \frac{[\mu C(1-\rho)t]^{h-k}}{(h-k)!}) \tag{7.23}$$

As before, we can obtain the throughput of the multi-hop network as:

$$\gamma_h = \lambda p_h \tag{7.24}$$

As a function of $p_1$ with an identical threshold delay t, $p_h$ can be expressed as:

$$p_h = \rho^h \{1 - \frac{1-p_1}{\rho} \sum_{k=1}^{h} \frac{[\ln(\frac{\rho}{1-p_1})]^{h-k}}{(h-k)!}\} \tag{7.25}$$

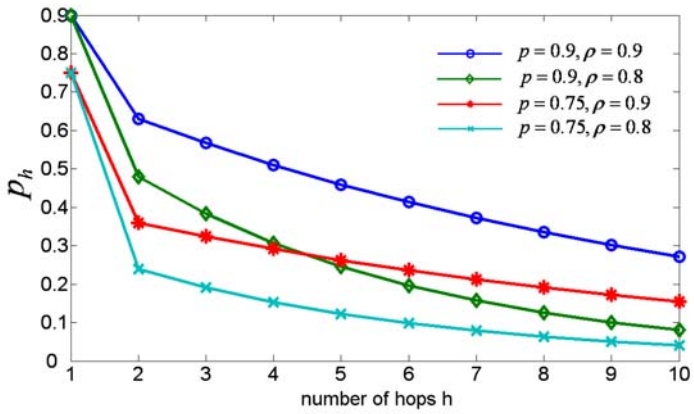


**Figure 7. 8: $p_h$ as a function of h with $p$ and $\rho$ as parameters**

Using equation (7.25), Figure 7.8 plots $p_h$ as a function of the number of hops for given $p$ and $\rho$. We can readily observe the sharp decline in the served traffic

from one hop to two hops, and relatively smoother decrease from two hops to multiple hops.

Also, $\frac{C_h}{C_1}$ as the function of number of hops $h$ is numerically evaluated from the equations (7.1) and (7.23), with design parameters $\lambda = 2000$ and $\mu = 1$. Figure 7.9 shows that, for example, for a given delay bound of 10*ms* with four-hop traffic, the transmission resources needed at each hop is 1.08 times as much as that of the single-hop traffic. With the same $1.08C_1$ capacity of each link in five-hop VoIP system, voice packets that that have the same relative throughput will experience a delay threshold of 15 *ms*.

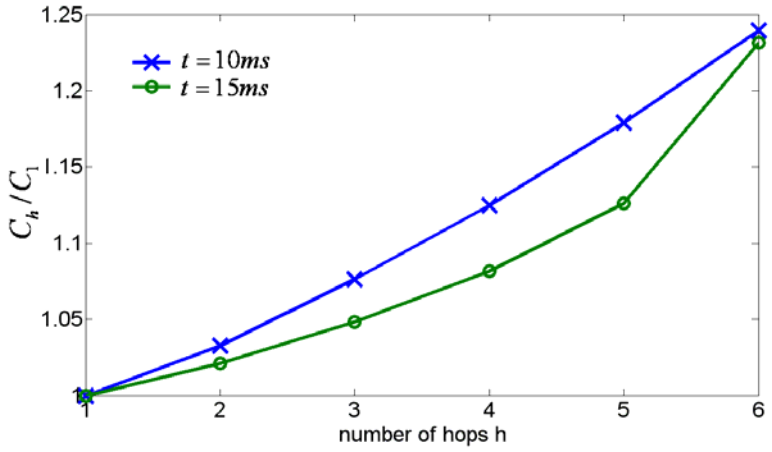


**Figure 7. 9: Price of h-hop traffic**

As discussed in Section 7.2, the customers are only charged for the served traffic. For multi-hop voice traffic, the pricing will depend on the number of hops as well as on the relative change in capacity of each hop compared to the single hop case with identical QoS. The procedure for pricing for multi-hop VoIP transport can now be stated as follows.

1) Given the values of $\lambda$ , $\mu$ and threshold delay *t*, compute the optimized transmission resources $C_{hopt}$ needed at each hop in *h*-hop system;

2) Compute the normalized throughput $p_{hopt}$ in case of $C_{hopt}$ ;

3) The unit price will be obtained as $\frac{hC_{hopt}}{\lambda p_{hopt}}$;

4) The price for the served traffic $\gamma$ will be $\frac{hC_{hopt}}{\lambda p_{hopt}} \times \gamma$ .

We note that the price levels derived above are applicable for a network that has optimized its transport resources so as to match the level of demand as specified through the incident traffic and the quality of service as specified through the threshold delay.

## 7.7 Conclusion

This chapter has presented the impact of quality of service demanded by a customer on the transport capacity of a VoIP network. We have developed analytical results that can lead to a determination of the optimum network capacity needed for prescribed levels of quality of service. The customer's pricing is based on the served traffic that meets the quality of service as specified through the threshold delay. Separate results are derived single-, two- as well multi-hop networks. The deleterious impact of multiple hops on the resources needed to maintain the same quality of service as the single-hop network can be quantitatively assessed as a result of the investigation reported in this chapter.

# Chapter 8 Cumulative Impact of Inhomogeneous Channels on Risk

**Abstract:** Chapter 4 developed a mechanism to evaluate the impact of more than one telecommunication channel in tandem on the distribution of delays. This chapter applies the mechanism developed in chapter 4 in assessing the impact of risks constituted by a number of channels, where the risk associated with each channel can be quantified by a known distribution. More specifically, this chapter considers flows of containerized traffic in a cascade of channels with diverse risk characteristics. Each channel is characterized by a probability distribution function relating the probability of loss being less than a specified value to the magnitude of the loss. The cumulative impact of cascading channels is then evaluated as a closed form solution in terms of the characteristics of the constituent channels with dissimilar risk characteristics. The results presented in this analysis can be used to shape the risk characteristics of individual channels through, for example, additional investment in order to maximize the impact of such investments. The contents of this chapter have partially been published in [112].

## 8.1 Introduction

This chapter applies the mechanism developed in chapter 4 in assessing the impact of risks constituted by a number of channels, where the risk associated with each channel can be quantified by a known distribution. The risk associated with a channel is in a mathematical sense equivalent to the delay experienced by a communication channel. The previous analysis has focused on computing the probability of communication channels in tandem having the cumulative delay

bounded to a specified value. This chapter extends that analysis and applies it to a situation where one needs to evaluate the probability that the cumulative risk is bounded to a specified value [113, 114]. The specific example considered in this chapter is the flow of goods across international boundaries. This is actually typified by the flow of containerized cargo through a number of transportation channels, e.g., roads, ocean, or air. The flow of containerized traffic will, generally speaking, also encounter a variety of gates that include national boundaries, customs and other government-mandated check points. Each of the modalities of transportation and the gates encountered by the container during its transit from the source to the destination presents a varying risk profile [115]. Each would impact the end-to-end risk characteristics of containerized traffic flow in complex ways.

This chapter develops a closed form solution relating the end-to-end risk behavior of containerized traffic flow in terms of the characteristics of each of the channels or gates. Understanding the end-to-end risk characteristics is important from the perspective of business because the business needs to develop a predictable model for risk in order to sufficiently insure its cargo and factor the costs of such insurance in the pricing model [116]. From a national security perspective, each government or national security agency needs to carefully weigh in the costs of improving safety and security against the predicted enhancement in attaining such security on an end-to-end basis. While the impact of an investment toward the improvement in the characteristics of a single channel or gate might be easily understood, this understanding is not sufficient in terms of evaluating the impact of that investment from an overall risk mitigation perspective when several channels in sequence are

involved. Accordingly, it might not yield the best ‘bang for the buck’ invested. This chapter characterizes the end-to-end risk profile in terms of the characteristics of each of the constituent elements in order to maximize the impact of investment toward improving the end-to-end risk characteristics.

The conventional way of understanding risk is in terms of the expectation of loss which is a product of the probability of loss and the average amount of loss. This information is insufficient if one were to assess the probability of loss remaining bounded within a predefined threshold. This idea is very similar to the concept developed in chapter 4, where we showed that for telecommunication applications, the average delay is not a meaningful parameter. Accordingly, this chapter addresses the probability associated with specified bounds of risk rather than the value of the average risk associated with a number of transportation channels/gates in tandem. More specifically, it addresses the scenario that a cargo goes through when it traverses a number of channels that are inhomogeneous in their risk characteristics [117]. The risk characteristics of a single channel are first addressed followed by two channels in tandem. Then, the results are generalized to include a number of cascaded channels with inhomogeneous risk characteristics. The findings of the investigation are illustrated through a number of examples.

## 8.2 The Single Channel Model

The model defines the risk characteristics of each intermediate traffic channel and each gate encountered by the in-transit cargo individually. It is assumed that the risk characteristic of each element (whether a channel or a gate) is independent. The

risk characteristics of each element are defined by a single variable $\lambda$ which is exponentially distributed.

The probability density function of this variable is given by:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0 \\ 0, x < 0 \end{cases} \tag{8.1}$$

The cumulative distribution function is given by:

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, x \geq 0 \\ 0, x < 0 \end{cases} \tag{8.2}$$

An interpretation of equation (8.2) would be that the probability of a loss of magnitude *x* or less is *F(x)*. The boundary conditions of equation (8.2) are verified by the fact that all losses are bounded by infinity and zero. Using the properties of the exponential distribution, the mean loss of a channel $=1/\lambda$. Figure 8.1 shows the distribution function for three different values of $\lambda$ =0.5, 1, and 1.5. The curve with a higher value of the mean loss (lambda=0.5) rises more slowly as expected. Figure 8.1 shows the distribution function for three different values of $\lambda$ =0.5, 1, and 1.5.

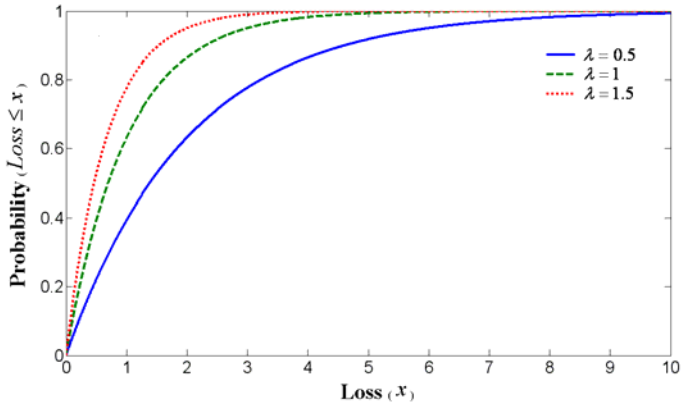


**Figure 8. 1: Loss Characteristics of a single channel**

## 8.3 The Two-channel Model

A two-channel model is considered, say, consisting of road and air transport channels. The probability density function of the two channels is defined as:

$$f_1(x) = \lambda_1 e^{-\lambda_1 x} \tag{8.3}$$

$$f_2(x) = \lambda_2 e^{-\lambda_2 x} \tag{8.4}$$

The combined probability density function f(x) of both the channels can be evaluated by convolving the two constituent probability density functions $f_1(x)$ and $f_2(x)$ [118, 119]. We have,

$$f(x) = f_1(x) \otimes f_2(x) = \int_{-\infty}^{\infty} f_1(\xi) f_2(x-\xi) d\xi = \int_{-\infty}^{\infty} \lambda_1 e^{-\lambda_1 \xi} \lambda_2 e^{-\lambda_2 (x-\xi)} d\xi$$

$$= \lambda_1 \lambda_2 e^{-\lambda_2 x} \int_0^x e^{(\lambda_2 - \lambda_1)\xi} d\xi = \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} e^{-\lambda_2 x} (e^{(\lambda_2 - \lambda_1)x} - 1)$$

$$= \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} (e^{-\lambda_1 x} - e^{-\lambda_2 x}) \tag{8.5}$$

For $\lambda_1 = \lambda_2 = \lambda$,

$$f(x) = \int_{-\infty}^{\infty} \lambda e^{-\lambda \xi} \lambda e^{-\lambda (x-\xi)} d\xi = \int_0^x \lambda^2 e^{-\lambda x} d\xi = \lambda^2 x e^{-\lambda x} \tag{8.6}$$

After some simplification, equation (8.5) can be expressed as,

$$f(x) = \begin{cases} \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} (e^{-\lambda_1 x} - e^{-\lambda_2 x}), \ (\lambda_1 \neq \lambda_2) \\ \lambda^2 x e^{-\lambda x}, (\lambda_1 = \lambda_2 = \lambda) \end{cases} \tag{8.7}$$

$$F(x) = \begin{cases} 1 - \frac{1}{\lambda_2 - \lambda_1} (\lambda_2 e^{-\lambda_1 x} - \lambda_1 e^{-\lambda_2 x}), (\lambda_1 \neq \lambda_2) \\ 1 - (1 + \lambda x) e^{-\lambda x}, (\lambda_1 = \lambda_2 = \lambda) \end{cases} \tag{8.8}$$

From equation (8.8), it is noticed that $\lambda_1$ and $\lambda_2$ are interchangeable. Therefore, the sequence of the channels does not affect the end-to-end loss. In other words, the end-to-end loss characteristic is determined entirely by the loss characteristics of each channel and is independent of their sequence.

**Examples:**

Figure 8.2 presents the loss characteristics of four different situations, each with two channels and the following loss characteristics.

Case 1: $\lambda_1 = 0.75, \lambda_2 = 1.25$

Case 2: $\lambda_1 = 0.5, \lambda_2 = 1.5$

Case 3: $\lambda_1 = 0.25, \lambda_2 = 1.75$

Case 4: $\lambda_1 = 1, \lambda_2 = 1$

Note that in each case the sum of $\lambda_1$ and $\lambda_2$ has been kept at a constant value, namely, $\lambda_1 + \lambda_2 = 2$.

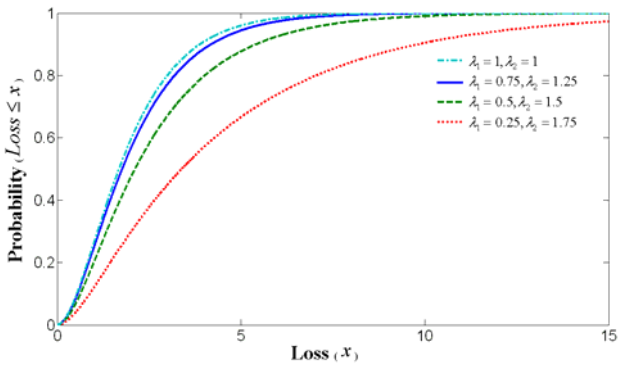


**Figure 8. 2: Loss Characteristics of two channels in tandem**

It is evident from Figure 8.2 that closer values of $\lambda_1$ and $\lambda_2$ result in curves that represent better end-to-end loss characteristics, i.e., the curves rise faster than those where $\lambda_1$ and $\lambda_2$ widely vary.

It is noted in passing that the mean value of loss of two channels in tandem is equal to $\frac{1}{\lambda_1}+\frac{1}{\lambda_2}$. This sum will obviously increase as $\lambda_1$ and $\lambda_2$ diverge while their sum remains constant. In other words, the cumulative loss characteristics of two channels in tandem are consistent with the loss experienced by each of the two channels. Let

$$\lambda_1 + \lambda_2 = C \text{ (a constant)} \tag{8.9}$$

It is intended to show that $\frac{1}{\lambda_1}+\frac{1}{\lambda_2}$ is minimized when $\lambda_1 = \lambda_2$. Let

$$\lambda_2 - \lambda_1 = \Delta \tag{8.10}$$

It is assumed, without loss of generality that $\lambda_2 > \lambda_1$ , i.e. $\Delta > 0$ . From equations (8.9) and (8.23), we have $\lambda_2 = \frac{C+\Delta}{2}$ and $\lambda_1 = \frac{C-\Delta}{2}$.

After some algebraic simplification, it is

$$\frac{1}{\lambda_1}+\frac{1}{\lambda_2} = \frac{4C}{C^2-\Delta^2} \tag{8.11}$$

For $C$ and $\Delta > 0$, it can be easily shown that (8.11) is minimized when $\lambda_1 = \lambda_2$.

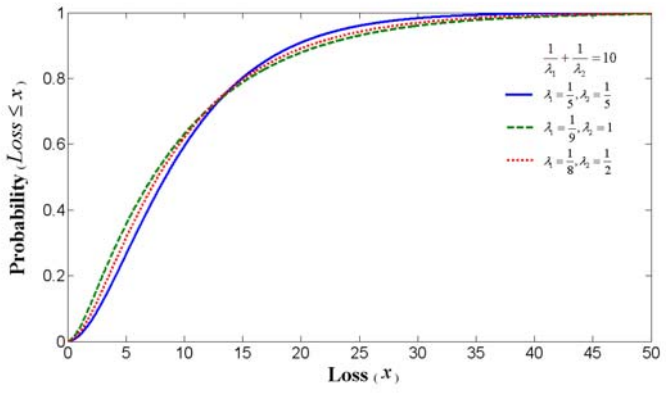


**Figure 8. 3: Loss Characteristics of two channels with the same end-to-end mean loss**

Figure 8.3 plots a number of curves for two channels in random where the sum $\frac{1}{\lambda_1}+\frac{1}{\lambda_2}$ is kept a constant while varying the individual $\lambda_1$ and $\lambda_2$. It can be observed from Figure 8.3 that, as long as the mean end-to-end loss experienced is kept to be a constant, the variance in their cumulative loss characteristics is moderate.

We can now ask ourselves the following question: Given identical end-to-end mean loss, is it preferable to have a single channel or two channels in tandem? A surprisingly elegant result is presented in Theorem 1.

***Theorem 1*:** Compared to two identical channels in tandem, each having a loss parameter equal to $\lambda$, a single-channel with the parameter $\frac{\lambda}{2}$ has superior cumulative loss characteristics up to a mean loss value that can be numerically evaluated. Beyond this point, the two-channel model with identical mean loss is superior.

*Proof:*

It is noted that the mean loss value for each model is identical since $\frac{1}{\lambda}+\frac{1}{\lambda}=\frac{1}{\frac{\lambda}{2}}$. We also have, for its single-channel model,

$$F_1(x)=1-e^{-\frac{\lambda}{2}x} \tag{8.12}$$

And for the two-channel model,

$$F_2(x)=1-(1+\lambda x)e^{-\lambda x} \tag{8.13}$$

For very small values of *x*, we have

$$F_1(x)\approx\frac{\lambda}{2}x-\frac{\lambda^2x^2}{8} \tag{8.14}$$

$$F_2(x) \approx \frac{\lambda^2 x^2}{2} \tag{8.15}$$

And therefore, for such values of *x*, $F_1(x) > F_2(x)$

The two loss curves intersect when $F_1(x) = F_2(x)$, i.e., when from equations (8.12) and (8.13)

$1 - e^{-\frac{\lambda}{2}x} = 1 - (1+\lambda x)e^{-\lambda x}$, or when

$$(1+\lambda x)e^{-\frac{\lambda}{2}x} - 1 = 0 \tag{8.16}$$

Equation (8.16) can be numerically evaluated.

For large values of *x*, we have,

$$F_2(x) - F_1(x) = e^{-\frac{\lambda}{2}x} - (1+\lambda x)e^{-\lambda x} \tag{8.17}$$

The ratio of the two terms on the R.H.S. of equation (8.17) can be written as,

$$\frac{e^{-\frac{\lambda}{2}x}}{(1+\lambda x)e^{-\lambda x}} = \frac{e^{\frac{\lambda}{2}x}}{1+\lambda x} \tag{8.18}$$

The problem then becomes to compare the relative values of $e^{\frac{\lambda}{2}x}$ and $1+\lambda x$. For a given value of $\lambda$, it is further noted that since the *slope* of an exponential function, $\frac{\lambda}{2}e^{\frac{\lambda}{2}x}$ increases with *x*, while the slope of a straight line is constant $\lambda$, there cannot be more than two points of intersection between a straight line and an exponential curve. The two channel model thus has better loss characteristics than the single channel model at higher values of loss.

## 8.4 The n-Channel Model

For *n*-channel model, there are two cases. One supposes that each channel has the same $\lambda$, then

$$f_n(x) = \frac{x^{n-1}}{(n-1)!}\lambda^n e^{-\lambda x} \tag{8.19}$$

$$F_n(x) = 1 - \sum_{i=0}^{n-1} \frac{(\lambda x)^i}{i!} e^{-\lambda x} \tag{8.20}$$

The other supposes $\lambda$, as

$$f_3(x) = \frac{\lambda_1\lambda_2}{(\lambda_2-\lambda_1)(\lambda_3-\lambda_1)}e^{-\lambda_1 x} + \frac{\lambda_1\lambda_2}{(\lambda_1-\lambda_2)(\lambda_3-\lambda_2)}e^{-\lambda_2 x} + \frac{\lambda_1\lambda_2}{(\lambda_1-\lambda_3)(\lambda_2-\lambda_3)}e^{-\lambda_3 x} \tag{8.21}$$

$$f_n(x) = \sum_{i=1}^{n} \frac{\prod_{i=1}^{n-1}\lambda_i}{\prod_{j=1,\, j\neq i}^{n}(\lambda_j-\lambda_i)} e^{-\lambda_i x} \tag{8.22}$$

$$F_n(x) = 1 - \sum_{i=1}^{n} \frac{\prod_{i=1}^{n-1}\lambda_i}{\prod_{j=1,\, j\neq i}^{n}(\lambda_j-\lambda_i)} e^{-\lambda_i x} \tag{8.23}$$

## 8.5 Conclusion

This chapter has provided a closed form solution to the general problem of assessing the probability of bounding loss in a cascade of channels when the risk characteristics of each channel, modeled as an exponential loss model, are known. The results obtained can be utilized to shape the loss characteristics of individual channel.

# Chapter 9. A Network Based Authentication Scheme for VoIP

**Abstract:** This chapter presents a VoIP networking solution that incorporates network-based authentication as an inherent feature. The proposed authentication feature introduces a range of flexibilities not available in the PSTN. Since most calls will likely terminate on the network of another service provider, a mechanism using which networks can mutually authenticate each other to afford the possibility of authentication across networks is also presented. [120]

## 9.1 Introduction

Telephone systems have evolved over the last century. This evolution has comprised moving from analog to digital systems using the circuit switching technology to packet switching systems. More recently, the potential benefits of converged networks in addition to lower costs for the customer has led to increasing use of Voice over IP [104]. This chapter focuses on just one aspect of Voice over IP, namely, mutual authentication of the calling and called parties. Mutual authentication between the caller-callee pair has been a distinguishing characteristic of the PSTN for over a century, but more so over the last twenty years or so when several applications use authentication provided by the telephone network for identification of the parties for sensitive transactions.

Voice over IP applications are generally designed to function over the global Internet, although such solutions can be offered over private IP networks, such as enterprise networks. Instances of violations of security over the Internet are common occurrences that affect individuals, businesses as well as government operations.

Voice over IP has not yet suffered many security violations, but the potential for attacks on security is truly large. Possibly the mass of end points connected to VoIP today is below the threshold that would attract miscreants. However, faking the identity of the caller can be easily accomplished using the Internet. This can result in the unsuspecting callee passing sensitive information to the calling party. Additionally, the negative impact of SPAM over Internet Telephony can be easily comprehended.

What level of security would two persons conversing over the emerging converged network need? There are three primary properties of secure communication: Secrecy, Authentication, and Integrity. For real-time voice communication among known individuals, authentication is guaranteed by their mutual recognition of each other through what is broadly known as speaker recognition. However, there are two caveats to relying on this as an exclusive authentication mechanism. First and foremost, authentication needs to take place prior to the commencement of communication, because it indeed might form the basis of call acceptance. Second, the caller might be led to a machine such as a human-machine interactive system where speaker recognition is crucial to the task being performed.

Authentication, as a generally available feature, is more important than secrecy in so far as voice communication is concerned. While authentication can be relegated to the end points (just as secrecy or integrity might be), it is considered extremely important from the security of the network perspective to positively identify and log the party that initiates the call. Most of the security issues in the

Internet exist because of its inability to positively identify the party that initiates the call giving rise to an intruder escaping identification by the network.

With the increasing demand for the VoIP application, Session Initiation Protocol (SIP) [25] has been developed by IETF for establishing real-time calls and conferences over IP networks. SIP is an end-to-end client-server signaling protocol to establish presence, locate users, set up, modify and tear down sessions. As a traditional text-based Internet protocol, it resembles the hypertext transfer protocol (HTTP) and simple mail transfer protocol (SMTP). Not being limited to IP telephony, SIP messages can convey arbitrary signaling payload, session description, instant messages, JPEGs, any MIME type. SIP uses Session Description Protocol (SDP) [51] for media description.

Despite many advantages of SIP, it is subject to various kinds of Denial of Service attacks. User authentication is becoming more and more important to prevent unauthorized users from using someone else's identity to fool other users or accounting and charging systems. Further, mutual authentication of the calling-callee pair is an essential requirement for the successful execution of several applications executed over a voice platform.

This chapter presents an authentication scheme that can be implemented within the SIP security framework. It is organized as follows. In Section 9.2, the importance of authentication is described. Current methods for authentication are discussed in Section 9.3. Section 9.4 outlines new requirements of both users and networks. Our proposed scheme meeting these requirements is addressed in Section 9.5. The chapter is concluded in Section 9.6.

## 9.2 Authentication

### 9.2.1 Need for Authentication

This section presents an expanded notion of authentication for both the calling and the called parties. It is anticipated that positive identification of the calling party will become an essential requirement for the callee to willingly engage in a conversation or engage in an application to be executed over a voice platform. At the present time, the calling line identification (CLID) serves this purpose. In VoIP, the CLID can be easily manipulated and therefore its use for the purpose will be increasingly suspect. By the same token, the calling party needs to be sure of the identity of the callee as well, e.g., when a bank employee or a credit card company calls an account holder to inquire about a specific transaction. This would imply that the instrument terminating the call and/or the party receiving the call is positively identified by the network. For maximum protection, both the calling and the called parties might need to be additionally identified using a biometric contrivance.

### 9.2.2 Mutual Authentication vs. Network Authentication

Authentication can, in general, be performed either on an end-to-end or identification by the network at points of ingress/egress basis. A case for the latter is made in this chapter. If the network were to authenticate an end user, the total number of authentication functions needed would be a maximum of n vs. n(n-1)/2 for the other case. End-points engaged in multi-party communication would need to identify themselves just once as opposed to each end-point identifying every other end-point on a pair-wise basis. Higher speed of operation as well as a large reduction in

computational overhead and key management issues can be anticipated as a result of network based authentication.

A prerequisite for the network based authentication is, of course, the existence of the network as a trusted agent of each end-point. In general, such a notion is opposite the concept of the Internet where the network can be easily fooled as far as the identity of an endpoint is concerned. In fact, this lack of ability has resulted in the growing threat of hackers of the Internet. It is also noted that there is a class of users for whom anonymity must be guaranteed, assuming the party they wish to communicate with honors such anonymity. In order to assure anonymity, in our proposed scheme the network does not pass the identity on to any other user even though it has verified the credentials of the specific user wishing to communicate. Exceptions must also be granted for a caller in emergency, possibly through the intervention of a live agent. The scheme proposed in this chapter accommodates the anonymity needs of the caller and callee independently while recognizing that communication is not guaranteed among incompatible end points.

## 9.3 Comparative Analysis of Existing Authentication Schemes

There are four commonly used solutions for authentication in the SIP framework: Digest-authentication, IPSec, TLS and S/MIME, which are analyzed and compared in Table 9.1.

Digest authentication [121] is a challenge-response based scheme to offer one-way authentication to the first requested server, while the intermediary servers have no idea whether the user is verified or not. It gives opportunity to a malicious user to masquerade in the “next-hop”, and make it vulnerable to spam, man-in-the-middle

and DoS attacks. In addition, adopting the digest authentication scheme in an IP telephony system only authenticates the communication device lacking relationship between its user and the SIP URL.

IPSec provides security services at the IP layer by enabling a system to select required security protocols, determine the algorithms to use for services, and put in place any cryptographic keys required to provide the requested services. IP Encapsulating Security Payload (ESP) [122] in the tunnel mode is preferred for IP tunneling across the Internet, although it involves substantial overhead. With IPSec used in tunnel mode, payload efficiency (ratio of payload to total packet size) of a 40-bypte VoIP packet drops from 50% to less than 30% [123], since the RTP/UDP/IP header is 40 bytes for IPv4.

Transport Layer Security (TLS) [124] provides a reliable end-to-end secure channel over connection oriented protocol. Both ends of the channel are identified by X.509 certificate [125] exchange. Making use of TLS to secure SIP signaling is transparent, which allows a signaling message at the application layer to be encrypted by TLS and then transferred through the TCP connection. If a TLS connection is requested, a SIP Secure URI (SIPS) is used. TLS is impractical to deploy in a wide area network since the TLS is built upon connection-oriented TCP protocol, restricting itself to limited applications, while most VoIP applications offer a continuous stream of RTP/UDP/IP packets. Further, if one hop along the path does not support TLS, the transit trust loses its meaning [126].

| Scheme | Key | Type | Application | Flaw |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| ***Digest*** | Pre-shared Key (PSK) | One-way Challenge -response | END-to-END | Integrity & Confidentiality not support |
| ***IPSec*** | PKI & PSK | Network Layer | HOP-by-HOP | Considerable Overhead |
| ***TLS*** | PKI | Transport Layer | HOP-by-HOP | Over TCP not UDP |
| ***S/MIME*** | PKI | S/MIME contents | END-to-END | Large size message |

**Table 9. 1: Comparison of Four Authentication Schemes**

Secure Multipurpose Internet Mail Extension (S/MIME) [127] has been developed for electronic messaging applications to provide origin privacy and message integrity. Conventionally, S/MIME used in SIP is for user-to-user or user-to-proxy authentication. In our proposed scheme, S/MIME will be adopted to encrypt the entire SIP message for hop-by-hop authentication instead of using an external protocol such as IPSec or TLS. Our proposed scheme could not be replaced by Digest authentication, which needs a pre-shared secret key between all users and between all users and proxy servers.

Some other solutions are proposed in the literature based on combinations of these four schemes. Johan Bilien has recommended solutions to secure VoIP using S/MIME, TLS and MIKEY for SIP signaling to provide end-to-end authentication and session key distribution, using SRTP on payload to protect voice media [128, 129]. NTT Network Service Systems Laboratories has proposed two approaches to

provide security on both SIP signaling and media streams for end-to-end communication in different scenarios [130]. In the following section, the enhanced view of authentication in VoIP is proposed.

## 9.4 Proposed Requirements of Authentication

Our proposed scheme offers the caller and callee options for authentication independently. In other words,

- The caller might choose to remain anonymous to the callee (not network).
- The callee might choose to reject anonymous calls.
- Both caller and callee can choose protection against eavesdropping, i.e., invoke encryption of the message.
- Callee can ask for more details related to caller's identity leading to a higher level authentication, such as website, birthday, or last transaction.

It is also important that the proposed network based authentication scheme accommodate authentication across one or more networks.

## 9.5 Proposed Schemes for Authentication

### 9.5.1 Proposed Scheme

Our proposed scheme assumes intra-networks with trust relationships between hops. S/MIME is applied to the SIP message including headers and bodies, which requires each proxy server to decrypt and encrypt the SIP message using respective public/private key pairs of the two communicating parties. It means that both user-to-

proxy and proxy-to-proxy authentication are based on S/MIME to continue passing the transit trust one by one. Figure 9.1 shows the overall operation and flow of messages within and between the two parties, which are the User Agent and ingress proxy server. Details are as follows,

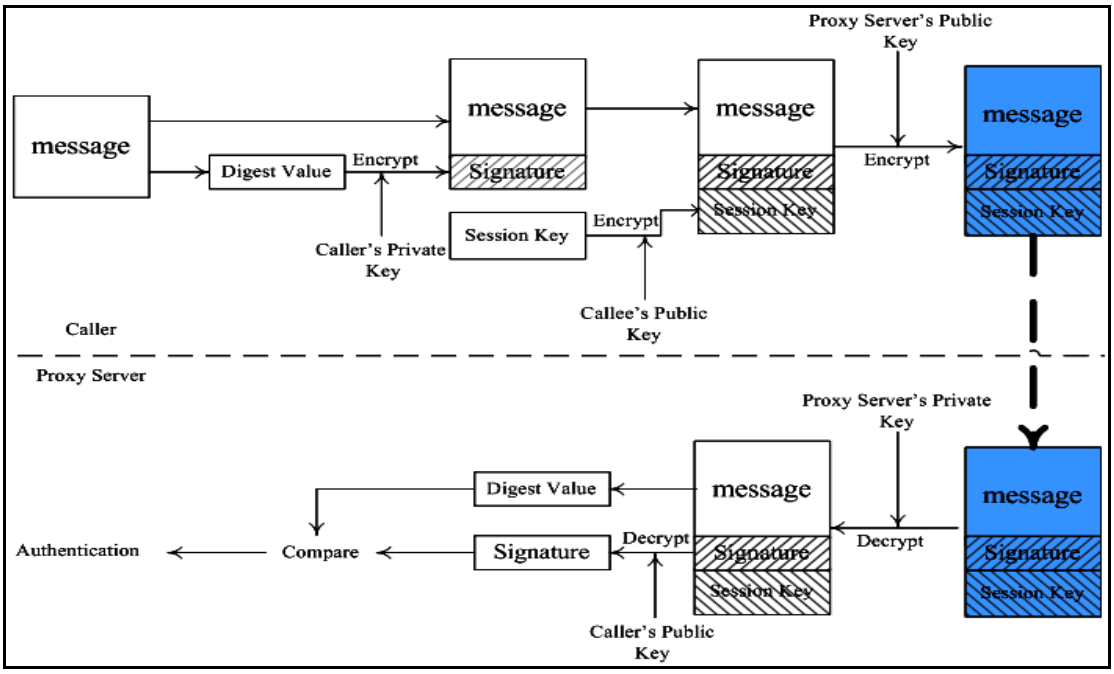


**Figure 9. 1: S/MIME INVITE Message**

1. The signature of the SIP message is created by encrypting the digest of the message with the caller's private key. The proxy server authenticates the caller using the caller's public key.

2. The caller randomly generates the session key protecting voice data against eavesdropping if so desired. Options are available to the caller/callee as described in Section 9.5.2. The session key is generated by applying 3DES on pseudorandom number, and securely sent to the callee by encrypting it with

callee's public key, which can be retrieved from X.509 Certificate hierarchy [131].

3. The SIP message, caller's signature and the encrypted session key are concatenated and encrypted using public key of the proxy server.

4. The proxy server decrypts the received message with its private key and the signature with caller's public key. It compares the recomputed digest value of the message with the decrypted signature. If they are matched, it is a valid signature and the caller's identity is verified. At the same time, authentication of the proxy server is also proved since only it has the private key to get the message. Note that the session key will not be disclosed to the proxy server because only the callee knows his/her private key.

Each proxy server along the signaling path executes the same authentication procedure as depicted above, but with its own private key for signature and the public key of the next-hop server for encryption, resulting in end-to-end authentication. The callee finally decrypts the session-key with his/her private key. The session key will be applied for protection of media flow. However, it is not a requirement for end-to-end communication. The two users' preferences for anonymity or desires for confidentiality are indicated in the matrix discussed below.

### 9.5.2 Enhanced Performance

In order to meet users' need as indicated in Section 9.4, one new header is introduced in the SIP signaling message Call-Type to offer more options. The Call-Type header is defined as an 8-bit field. The first/last four bits are set respectively by

caller/callee to indicate their preference for making/receiving an anonymous or confidential call. In addition, the callee has one more option to ask for further verification of caller's identity with more information. Each four-digit group is assigned as below in Table 9.2.

Although the proposed system requires only 3 bits, a 4-bit field is recommended for the caller/callee with the last bit reserved for future use. As Table 9.2 shows, the caller has 6 options while the callee has 8 options. There are 48 valid combinations for the network to decide the properties of the call to be established.

For anonymity, the proxy server that receives the first INVITE message directly from the caller, checks the first bit. If it is 1, the server will remove From and Contact headers from the decrypted message and add appropriate information in the Via header field; next, the server signs and encrypts the modified message, and then passes it on to the next hop.

For more information required, the proxy server that receives response directly from the callee, checks all the 8 bits in the way depicted in Table 9.3. If the 8-bit matches XXXXX11X, this proxy server returns 401 Unauthorized including the Call-Type and indicates that the request requires additional authentication. Once 401 Unauthorized is received by the caller, he will generate a new INVITE message with more information and send it to the callee again in a way similar to Digest-Authentication. After the additional authentication, the actual call will take place according to the confidentiality preferences of the parties, provided that they match as indicated in Table 9.3.

<table>
<tr><th colspan="3">Anonymity<br>(First digit)</th><th colspan="3">Confidentiality<br>(Middle two digits)</th><th rowspan="6">Reserved<br>(Last digit)</th></tr>
<tr><td rowspan="2">Caller</td><td>0</td><td>No</td><td>00</td><td colspan="2">Either</td></tr>
<tr><td>1</td><td>Yes</td><td>01</td><td colspan="2">Yes</td></tr>
<tr><td rowspan="3">Callee</td><td>0</td><td>Accept all</td><td>10</td><td colspan="2">No</td></tr>
<tr><td rowspan="2">1</td><td rowspan="2">Reject anonymous calls</td><td rowspan="2">11</td><td>Caller</td><td>Reserved</td></tr>
<tr><td>Callee</td><td>More information required for authentication</td></tr>
</table>

**Table 9. 2: Caller/Callee Option**

<table>
<tr><th rowspan="2">Anonymity</th><th colspan="2">Rejected Call</th><th rowspan="2">More Information Required</th><th rowspan="2">Confidentiality</th></tr>
<tr><th>Anonymous Call</th><th>Confidentiality Not Matched</th></tr>
<tr><td rowspan="4">1XXXXXXX</td><td rowspan="4">1XXX1XXX</td><td rowspan="2">X01XX10X</td><td rowspan="4">XXXXX11X</td><td>X00XX00X</td></tr>
<tr><td>X00XX01X</td></tr>
<tr><td rowspan="2">X10XX01X</td><td>X01XX00X</td></tr>
<tr><td>X01XX01X</td></tr>
</table>

**Table 9. 3: Call-Type**

## 9.5.3 Inter-network authentication

Making VoIP calls crossing different networks needs network-to-network authentication, which is performed at the gateway between networks. It is similar to hop-by-hop authentication described above. The gateway will attach the sending network's signature, and then encrypt the message together with signature and wrapped session key using communicating network's public key obtained from X.509

Certificate. The only difference is that this authentication procedure moves to the gateway, which keeps the public/private key pair of the network.

## 9.6 Conclusion

This chapter summarizes and compares various solutions to SIP signaling authentication. After reviewing current proposals, a scheme for authentication is proposed to be executed by each hop: user-to-server, server-to-server or network-to-network. By chaining trust among SIP components across the trusted network, end-to-end authentication is realized. In addition, advanced level of authentication service is offered for users in our scheme through options of anonymity, confidentiality and additional authentication, if required.

## Chapter 10 Conclusions and Future Work

This dissertation has studied the performance of the VoIP networks from three aspects, Quality of Service, price, and security. As a real-time application, VoIP networks are challenged by the legacy voice quality as exhibited by the traditional PSTN. This dissertation has introduced the notion of upper bound of delay and jitter, termed threshold, to characterize the QoS of VoIP traffic. Traffic that suffers a delay/jitter higher than the bound is considered lost and does not constitute effective throughput. This dissertation has presented the impact of the threshold delay/jitter on the resource consumption under different queuing modelswhich have proven to adequately represent traffic for VoIP. The learning from the analysis has been extended and applied to assessing the impact of risks constituted by a number of transportation channels where the risk associated with each channel can be quantified by a known distribution. Finally, this dissertation has also presented an authentication mechanism using which networks can mutually authenticate each other to afford the possibility of authentication across networks.

In the first part of this dissertation, the average delay and the bounded delay have been compared in characterizing both the QoS and network management aspects. Two different queuing models have been evaluated, M/M/1 and M/D/1. For the M/M/1 model, a single-hop network can achieve a hundred percent throughput, if the upper delay bound goes to infinity, while the maximum throughput of the n-hop network can only reach $\rho^n$ %. For the M/D/1 model, a similar result is found in that the throughput performance of the multi-hop network continues to deteriorate as the number of hops increases; the needed capacity of both the single-hop and the multi-

hop networks can be analytically evaluated from the closed form solution presented. The results obtained can be used in scaling resources in a VoIP network for different thresholds of acceptable delays. The notion of the upper bound delay has been also extended to jitter which is, potentially, the largest source of degradation in the quality of voice in VoIP systems. This dissertation has also provided a way to compute the traffic handling capability of a multi-hop resource constrained network under a defined limit of end-to-end jitter. The presented analytical solutions help scale up the resources in the network as the number of hops increases.

The second part of this dissertation has proposed a new pricing scheme based on the cost of lost opportunity vs. the cost of consumption of resources. The notion of a unit price has been developed. Separate results are derived for single-, two- as well as multi-hop networks. It has been observed that the unit price for two-hop traffic is more than four times the unit price for single-hop traffic if the performance were to remain the same. The examples given in this dissertation show that both the ratios of the capacity per each link and the unit price of two-hop network, relative to single-hop network, decrease as the threshold delay increases. A similar pricing solution for multi-hop network is also presented which depends on the number of hops as well as on the relative change in capacity of each hop compared to the single hop case with identical QoS. This dissertation has appropriately priced VoIP services at equitable levels that are consistent with the resources consumed in order to achieve the contracted QoS.

Extending the techniques developed in the previous chapters, this dissertation has evaluated the probability that the cumulative risk of a number of

transportation channels in series is bounded to a specified value. In particular, the risk associated with the flow of containerized traffic in a cascade of channels with diverse risk characteristics has been used as an example. The dissertation has developed a closed form solution in terms of the characteristics of the constituent channels with dissimilar risk characteristics. This approach can be used to shape the risk characteristics of individual channels through additional investment in order to maximize the impact of such investments.

The third part of this dissertation has considered network security solutions in VoIP applications. The potential for attacks on security is large in the VoIP environment. For the network authentication, this dissertation has presented a VoIP networking solution that incorporates network-based authentication as an inherent feature and introduces a range of flexibilities not available in the PSTN. After reviewing current proposals of SIP signaling authentication, a new authentication scheme has been developed based on the chaining of trust among SIP components across the trusted networks, realizing end-to-end authentication. In addition, advanced level of authentication service is offered through options of anonymity, confidentiality and additional authentication, if required.

Future work would potentially include a further generalization of the solutions provided in this dissertation from different standpoints. Additional queuing models such as M/G/1 and G/G/1 can be considered for analysis to improve the optimum channel capacity allocation in a more generalized manner. It is also possible to define the packet size ranges based on actually measured statistics of traffic that traverses the Internet. Further the independence assumption used in this dissertation

can be weakened to include self-similar traffic. For implementing VoIP security, authentication can be done via biometric features such as finger print, eye scan, voice spectrogram, etc. The authentication mechanism presented in this dissertation can be further augmented to address other security issues such as confidentiality and non-repudiation.

## References:


[1] B. Goode, "Voice Over Internet Protocol (VOIP)," Proc. IEEE, vol. 90, Sept 2002, pp. 1495-1517

[2] P.A. Bonenfant and S.M. Leopold, "Trends in the U.S. communications equipment market: a Wall Street perspective," in IEEE Communications Magazine, vol. 44, Feb. 2006, pp. 102 – 108.

[3] Morgan Keegan & Co., Proprietary Equity Research, 2005.

[4] S. Cherry, "Seven myths about voice over IP," in IEEE Spectrum, vol. 42, March 2005, pp. 52 - 57

[5] K. Wallace, Authorized Self-Study Guide Cisco Voice over IP (CVoice), Cisco Press, September 2006

[6] G. Thomsen and Y.Jani, "Internet Telephony: Going Like Crazy," in IEEE spectrum, May 2000, vol. 37, no.5, pp. 52-58.

[7] W. Stallings, Data and Computer Communications, New Jersey, USA: Prentice-Hall, 1997.

[8] M. Maresca, N. Zingirian and P. Baglietto, "Internet protocol support for telephony," Proc. IEEE, vol 92, Sept 2004, pp. 1463-1477

[9] H.M. Chong, H.S. Matthews, "Comparative analysis of traditional telephone and voice-over-Internet protocol (VoIP) systems," Electronics and the Environment, IEEE International Symposium, May 2004, pp. 106 - 111

[10] M.Hamdi, O. Verscheure, J-P. Hubaux, I. Dalgic and P.Wang, "Voice Service Interworking for PSTN and IP Networks," IEEE Commun. Mag, vol 37, issue 5, May 1999, pp. 104-111

[11] W. C. Hardy, QoS Measurement and Evaluation of Telecommunications Quality of Service, West Sussex, England: John Wiley & Sons, 2001

[12] S. Jha and M. Hassan, Engineering Internet QoS. London, UK: Artech House, 2002.

[13] G. Armitage, Quality of Service in IP Networks. Indiana, USA: Macmillan Technical Publishing, 2000.

[14] CT Labs, Inc., "Speech Quality Issues & Measurement Techniques Overview". Revision: 10-23-2000 CJB, 2000. http://www.ct-labs.com/Documents/Speech_Quality_Testing.pdf .

[15] M. Morrow, V. Sharma, T.D. Nadeau, and L. Andersson, "Challenges in Enabling Interprovider Service Quality in the Internet," in IEEE Communications Magazine, vol. 43, 2005, pp. 110 – 111.

[16] Christos Bouras and Afrodite Sevasti, "Pricing QoS over transport networks Internet Research, Volume 14 · Number 2 · 2004 · pp. 167-174

[17] Odlyzko, A. (2001), "Internet pricing and the history of communications", Computer Networks, Vol. 36 No. 5, pp. 493-517.

[18] Qu, Y., Verma, P.K.: 'Notion of cost and quality in telecommunication networks: an abstract approach', IEE Proc.-Commun., April 2005, Vol. 152, pp. 167-171

[19] T.J. Walsh, D.R. Kuhn, “Challenges in securing voice over IP,” in IEEE Security & Privacy Magazine, vol. 3, May-June 2005, pp. 44 – 49.

[20] L. Sun, “Speech Quality Prediction for Voice over Internet Protocol Networks,” PhD Thesis, University of Plymouth, January 2004

[21] H. Schulzrinne, http://www.cs.columbia.edu/~hgs/internet/

[22] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, “RTP: A transport protocol for real-time applications,” IETF RFC 1889, 1996. ftp://ftp.ietf.org/rfc/rfc1889.txt.

[23] J. F. Kurose and K. W. Ross, Computer Networking: a Top-down Approach Featuring the Internet, Prentice Hall, 2006.

[24] “ITU-T Recommendation H.323: Packet-based multimedia communications systems,” International Telecommunication Union, 1997.

[25] J. Rosenberg, H. Schulzrinne, Camarillo, Johnston, Peterson, Sparks, Handley and Schooler, “SIP: Session initiation protocol v.2.0,” IETF RFC 3261, 2002.

[26] M. Arango, A. Dugan, I. Elliott, C. Huitema and S. Pickett, “Media gateway control protocol (MGCP) Version 1.0,” IETF RFC 2705,1999.

[27] N. Greene, M. Ramalho and B. Rosen, “Media gateway control protocol architecture and requirements,” IETF RFC 2805, 2000.

[28] F. Cuervo, N. Greene, A. Rayhan, C. Huitema, B. Rosen and J. Segers, “Megaco Protocol Version 1.0,” IETF RFC 3015, 2000.

[29] D. J. Wright, Voice over Packet Networks, John Wiley & Sons, 2001

[30] H. Schulzrinne, “IP Networks.” http://citeseer.nj.nec.com/schulzrinne00ip.html.

[31] ITU-T Recommendation G.114 ‘One-way transmission time’, 1996

[32] “Stability and Echo,” CCITT Recommendation G.131 , 1988.

[33] M. Hassan, “Internet Telephony: Services, Technical Challenges, and Products”, IEEE Communication Magazine, Apr. 2000.

[34] Y. Bai and M.R. Ito, “QoS Control for Video and Audio Communication in Conventional and Active Networks: Approaches and Comparison”, IEEE Communications Surveys & Tutorials, vol.6, no.1, First Quarter, 2004.

[35] D. De Vleeschauwer, M.J.C. Büchli and A. Van Moffaert, “End-to-End Queuing Delay assessment in Multi-service IP Networks,” Journal of Statistical Computation and Simulation, vol. 72, no. 10, pp.803-824, 2002.

[36] C.J. Sreenan, J.Chen, P.Agrawal, and B. Narendran, “Delay Reduction Techniques for Playout Buffering”, IEEE Transactions on Multimedia, 2(2):88-100, June 2000.

[37] R. Wang and X. Hu, “VoIP Development in China”, Computer, vol.37, issue 9, Sept. 2004, pp. 30 – 37.

[38] Y. Bai and M.R.Ito, “A Study for Providing Better Quality of Service to VoIP Users,” IEEE Proceedings of the 20th International Conference on Advanced Information Networking and Applications (AINA’06), vol. 1, pp. 799-804, 2006.

[39] C. Boutremans and J. Y. Le Boudec, “Adaptive Joint Playout Buffer and FEC Adjustment for Internet Telephony," IEEE INFOCOM’2003, San-Francisco, USA, Apr. 2003.

[40] W. Jiang and H. Schulzrinne, “Comparison and Optimization of Packet Loss Repair Methods on VoIP Perceived Quality under Bursty Loss,” NOSSDAV’02, May 2002.

[41] M.E., Nasr and S.A. Napoleon, “On Improving Voice Quality Degraded by Packet Loss in Data Networks” the 7th AFRICON Conference in Africa, AFRICON, Sept. 2004,vol. 1, pp.51 - 55.

[42] J. H. James, Bing Chen and L. Garrison, “Implementing VoIP: A Voice Transmission Performance Progress Report,” in IEEE Communications Magazine, vol. 42, July 2004, pp. 36-41.

[43] Spirent Communications, “IPWave”, Online Documentation, http://www.spirentcom.com/analysis/product.cfm?WS=13&PR=8.

[44] Agilent Technologies, “Voice Quality Tester (VQT)”, Online Documentation. http://www.home.agilent.com/USeng/nav/-536885778.536882651/pd.html.

[45] L. Sun and E. Ifeachor, “Perceived Speech Quality Prediction for Voice over IPbased Networks,” in Proceedings of IEEE International Conference on Communications ICC’02, (New York, USA), pp. 2573–2577, April 2002.

[46] Agilent Technologies, “VQT Phone Adapter”, Technical Specification. http://cp.literature.agilent.com/litweb/pdf/5968-7723E.pdf .

[47] L. Ding and R. A. Goubran, “Speech Quality Prediction in VoIP Using the Extended E-Model,” IEEE. GLOBECOM, San Francisco, USA, 2003

[48] A.E. Conway, and Y. Zhu, “Analyzing voice-over-IP subjective quality as a function of network QoS: A simulation-based methodology and tool,” in Computer Performance Evaluation Modelling Techniques and Tools, Lecture Notes in Computer Science, vol. 2324, pp. 289-308, Springer, 2002.

[49] H. Avshalom: ‘A Sip of SIP’, Lotus Software-IBM SWG, white paper, Nov. 2003.

[50] “SIP: Protocol Overview”, Radvision white paper, http://www.radvision.com/NR/rdonlyres/51855E82-BD7C-4D9D-AA8A-E822E3F4A81F/0/RADVISIONSIPProtocolOverview.pdf, 2005.

[51] M. Handley and V. Jacobson, “SDP: Session description protocol,” IETF RFC 2327, 1998.

[52] SIP Introduction,http://www.iptel.org/ser/doc/sip_intro/sip_introduction.html

[53] M. Poikselka, A. Niemi, H. Khartabil and G. Mayer, The IMS: IP Multimedia Concepts and Services, 2nd Edition, Wiley, Mar 2006

[54] M. Handley, H. Schulzrinne, E. Schooler and J. Rosenberg, “SIP: session initiation protocol,” IETF RFC 2543, March 1999.

[55] "SIP Server Technical Overview", Radvision white paper, http://www.radvision.com/NR/rdonlyres/0AFA30DF-DAD6-461D-943C-ED33F3E7ABD8/0/SIPServerTechnicalOverviewWhitepaper.pdf, 2004.

[56] J. Roberts, U. Mocci, and J. Virtamo, Broadband Network Teletraffic. Final Report of Action COST 242. Berlin, Springer, 1996.

[57] K. Park and W. Willinger, Self-Similar Network Traffic and Performance Evaluation. New York, USA: John Wiley & Sons, 2000.

[58] http://www.tele.dtu.dk/teletraffic/handbook/telehook.pdf, accessed Feb. 2007.

[59] J. Daigle and J. Langford, "Models for analysis of packet voice communications systems," IEEE Journal on Selected Areas in Communications, vol. SAC-4, no. 6, pp. 847-55, Sept. 1986.

[60] Kendall, D.G. (1951): Some problems in the theory of queues. Journal of Royal Statistical Society, Series B, Vol. 13 (1951): 2, 151–173.

[61] Y. C. Jenq, "Approximations for Packetized Voice Traffic in Statistical Multiplexer," in Proceedings of INFOCOM'84, April 1984.

[62] L. Kleinrock, Communication Nets: Stochastic Message Flow and Delay, McGraw-Hill, New York, 1964.

[63] O. Østerbø, Models for End-to-end Delay in Packet Networks Queuing. R&D report 4/2003

[64] D. P. Bertsekas and R. Gallager, Data Networks, New Jersey, USA: Prentice-Hall, 1992

[65] E.C. Molina, "The Theory of Probabilities Applied to Telephone Trunking Problems," Bell System Tech. J., pp. 69-81, 1922

[66] W. Stallings, High Speed Networks and Internets. New Jersey, USA: Prentice-Hall, 2002.

[67] Cox, D.R.& Isham, V. (1980): Point processes. Chapman and Hall. 1980.

[68] C.D. Crommelin, "Delay probability formulae when the holding times are constant," Post Office Electrical Engineers Journal, 1932, vol. 25, pp. 41–50.

[69] V.B. Iversen, "Exact calculation of waiting time distributions in queueing systems with constant holding times," NTS-4, Fourth Nordic Teletraffic Seminar, Helsinki 1982.

[70] L. Kleinrock, Queueing systems. Vol. II: Computer applications. New York, 1976.

[71] L. Wang and P. K. Verma, "Impact of Bounded Delays on Resource Consumption in VoIP Networks," submitted to the IASTED International Conference Communication Systems, Networks, and Applications, CSNA 2007, Beijing, China, October. 8-10, 2007.

[72] J. Janssen, D. De Vleeschauwer, M. Buchli and G.H. Petit, "Assessing voice quality in packet-based telephony," IEEE Internet Computing, May-June 2002, pp. 48 – 56.

[73] "Clearing the Way for VoIP, An Alternative to Expensive WAN Upgrades", white paper, Gen2 Ventures, 2003

[74] Miyahara, H., Teshigawara, Y. and Hasegawa, T.: 'Delay and Throughput Evaluation of Switching Methods in Computer Communication Networks'. IEEE Trans. Commun., Mar 1978, pp. 337 – 344

[75] F.P. Kelly, Blocking probabilities in large circuit-switched networks, Adv. Appl. Probab., 1986.

[76] V.S. Frost and B. Melamed, "Traffic modeling for telecommunications networks," IEEE Communications Magazine, March 1994, 32, (3), pp. 70 – 81.

[77] H. Heffes and D. M. Lucantoni, A Markov Modulated characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance, IEEE JSAC, Sept. 1986, vol. SAC-4, No. 6, pp. 856-868.

[78] Cao, J., Cleveland, W., Lin, D. and Sun, D.: 'Internet Traffic Tends Toward Poisson and Independent as the Load Increases'. Nonlinear Estimation and Classification, eds. C. Holmes, D. Denison, M. Hansen, B. Yu, and B. Mallick, Springer, New York, 2002.

[79] Sharafeddine, S., Riedl, A., Glasmann, J. and Totzke, J.: 'On traffic characteristics and bandwidth requirements of voice over IP applications'. Proc. 8th IEEE International Symposium on Computers and Communication(ISCC), 2003, pp. 1324 – 1330.

[80] Gardner, M.T., Frost, V.S., Petr, D.W.: 'Using optimization to achieve efficient quality of service in voice over IP networks'. Proc. Int. Conf. Performance, Computing, and Communications, Phoenix,Arizona, April 2003, pp.475-480.

[81] L. Kleinrock, Queueing Systems Volume I: Theory, John Wiley and Sons: 1975.

[82] P.K. Verma, Performance Estimation of Computer Communication Networks: A structured Approach. Computer Science Press, Rockville, MD, USA 1989.

[83] H. Harada and R.Prasad, Simulation and Software Radio for Mobile Communication, Artech House, Boston, MA, 2002.

[84] L. Wang, Y. Tachwali, P. K. Verma and A. K. Ghosh, "Impact of Bounded Delay on Throughput in Multi-hop Wireless Sensor Networks," Journal of Network and Computer Applications, Elsevier, May. 2008, Under Review.

[85] K. VAN DER WAL, M. MANDJES, and H. BASTIAANSEN, "Delay performance analysis of the new Internet services with guaranteed QoS," Proc. IEEE, 1997, 85, (12), pp. 1947-1957.

[86] F. A. Tobagi, "Modeling and performance analysis of multihop packet radio networks," Proceedings of the IEEE, Vol.75 no.1, January 1987, pp.135-155.

[87] M. Mandjes, K. van der Wal, R. Kooij, and H. Bastiaansen, "End-to-end delay models for interactive services on a large-scale IP network," Seventh IFIP workshop on Performance Modelling and Evaluation of ATM Networks: IFIP ATM'99, Antwerp, Belgium, June 28-30, 1999.

[88] D. De Vleeschauwer, G.H. Petit, B. Steyaert, S. Wittevrongel, and H. Bruneel, "Calculation of end-to-end delay quantile in network of M/G/1 queues," IEEE Letters, 2001, vol. 37, 8, (12), pp. 535 – 536.

[89] John F. Shortle and Percy H. Brill, "Analytical Distribution of Waiting Time in the M/{iD}/1 Queue," July 2005, vol. 50, pp. 185-197.

[90] G. Ramamurthy, B. Sengupta, "Delay Analysis of a Packet Voice Multiplexer by the $\sum D_i / D/1$ Queue," IEEE Transactions on Communications, vol. 39, no. 7, pp. 1107-1114, July 1991.

[91] O. Østerbø, "An approximative Method to Calculate the Distribution of End-to-end Delay in Packet Network," R&D report, Feb. 2002.

[92] O. Østerbø, "End-to-end Queuing Models with Priority," R&D report, Mar. 2003

[93] Iversen, V.B. and Staalhagen, L.: 'Waiting time distribution in M/D/1 queuing systems'. Electronics Lett., Dec. 1999, vol. 35, pp. 2184-2185.

[94] H. Takagi, Queueing Analysis, Volume 1: Vacation and Priority Systems, Part 1. Amsterdam, North-Holland, 1991.

[95] P. M. Gopal and B. Kadaba, "A Simulation Study of Network Delay for Packetized Voice," in Proceedings of GLOBECOM'86, Dec. 1986.

[96] J. Hancock, "Jitter Fundamentals," High Frequency Electronics, April 2004.

[97] "Jitter Solutions for Telecom, Enterprise, and Digital Designs," Agilent Technologies, August 2005.

[98] N. Davies, J. Holyer, and P. Thompson, "End-to-end Management of Mixed Applications Across Networks," in IEEE Workshop on Internet Applications, 1999, pp. 12-19.

[99] M. Karol, P. Krishnan, and J. J. Li, "enProtect: Enterprise-Based Network Protection and Performance Improvement," Avaya Labs Research-Technical Report, August 2002.

[100] F. Guilleman and J. W. Roberts, "Jitter and bandwidth enforcement," in Proc.IEEE Globecom'91, Phoenix,AZ, pp. 261-265, 1991.

[101] L. Wang and P. K. Verma, "Impact of Bounded Delays on Resource Consumption in VoIP Networks," in IASTED International Conference Communication Systems, Networks, and Applications, CSNA, Oct. 2007.

[102] M. H. Dahshan and P. K. Verma, "Resource Based Pricing Framework for Integrated Service Networks," Journal of Networks, vol. 2, pp. 36-45, June 2007.

[103] L. Wang and P. K. Verma, "The Notion of Cost and Quality in Packet Switched Networks An Abstract Approach," Journal of Networks, June. 2008, Under Review.

[104] U. Black: 'Voice over IP'. Pretice Hall, 1999

[105] C.R. Johnson, Y. Kogan, Y. Levy, F. Saheban and P. Tarapore, "Voice Reliability: A Service Provider's Perspective," IEEE Commun. Mag., July 2004, pp.48-54

[106] TSB116: 'Voice quality recommendations for IP telephony', 2001

[107] L.A. DaSilva, "Pricing for QoS-enabled networks: a survey", IEEE Communications Surveys & Tutorials, Vol. 3 No. 2, pp. 14-20.

[108] M. Falkner, M. Devetsikiotis and I. Lambadaris, "An overview of pricing concepts for broadband IP networks", IEEE Communications Surveys & Tutorials, Vol. 3 No. 2, pp. 2-13.

[109] Qu, Y., Verma, P.K.: 'Notion of cost and quality in telecommunication networks: an abstract approach', IEE Proc.-Commun., April 2005, Vol. 152, pp. 167-171.

[110] L. Zhen, L. Wynter, and C. Xia: 'Usage-Based Versus Flat Pricing for E-Business Services with Differentiated QoS'. IEEE International Conference on E-Commerce - CEC '03, June 2003, pp. 355-362.

[111] Fishburn, P. C. and Odlyzko, A. M.: 'Dynamic Behavior of Differential Pricing and Quality of Service Options for the Internet'. 1st International Conference on Information and Computation Economies, Charleston, South Carolina, USA, 1998.

[112] L. Wang and P. K. Verma, "Cumulative Impact of Inhomogeneous Channels on Risk," Under Review.

[113] John R. Harrald, Hugh W. Stephens, Johann Rene vanDorp, A Framework for Sustainable Port Security, Journal of Homeland Security and Emergency Management, Vol. 1, Issue 2, 2004.

[114] H. W. Stephens, Barriers to Port Security, Journal of Security Administration, Vol. 12, No. 2, pp. 29-40.

[115] Stephen Flynn, The Edge of Disaster, Random House, New York, 2007.

[116] J. M. Lewis, S. Lakshmivarahan, and S. K. Dhall, Dynamic data assimilation: a least squares approach. Cambridge: Cambridge university press, 2006.

[117] G. A. Kivman, Sequential parameter estimation for stochastic systems, Nonlinear Processes in Geophysics, vol. 10, pp. 253-259, 2003.

[118] H. J. Nussbaumer, Fast Fourier Transform and Convolution Algorithms. Berlin Heidelberg: Springer-Verlag, 1990.

[119] E. Grosswald and Samuel Kotz, An integrated lack of memory characterization of the exponential distribution, Annals of the Institute of Statistical Mathematics, vol. 33, pp. 205-214, Springer Netherlands, 1981

[120] L. Wang and P. K. Verma, "A Network Based Authentication Scheme for VoIP," IEEE International Conference for Communication Technology ICCT'06, Guilin, China, Nov. 2006.

[121] Franks, J. et al., "HTTP authentication: Basic and Digest Access Authentication", IETF RFC 2617, June 1999.

[122] S. Kent and R. Atkinson, IP Encapsulating Security Payload (ESP), November 1998. RFC 2406.

[123] Z. Anwar, W. Yurcik, R. E. Johnson, M. Hafiz, and R. H. Campbell,, "Multiple Design Patterns for Voice over IP (VoIP) Security," in Proc. 25th IEEE International Performance Computing and Communications Conference (IPCCC) , Phoenix, Arizona, USA, pp. 485-492, April. 10-12, 2006.

[124] Dierks, T. and C. Allen, "The TLS Protocol Version 1.0", IETF RFC 2246, January 1999.

[125] Housley, R. et al.,"InternetX.509 Public Key Infrastructure: Certificate and CRL Profile", IETF RFC 3280, April 2002.

[126] S. Salsano, L. Veltri, and D. Papalilo, “SIP security issues: The SIP authentication procedure and its processing load,” IEEE Network, vol. 16, no. 6, pp. 38–44, Nov.–Dec. 2002.

[127] Ramsdell B., “SMIME Version 3 Message Specification”, IETF RFC 2633, June 1999.

[128] J. Orrblad, Alternatives to MIKEY/SRTP to secure VoIP, Master Thesis, KTH, Stockholm, March 2005.

[129] J. Bilien, E. Eliasson, J. Orrblad, and J-O. Vatn, “Secure VoIP: call establishment and media protection,” 2nd Workshop on Securing Voice over IP, Washington DC, June 2005.

[130] K. Ono and S. Tachimoto, “SIP signaling security for end-to-end communication,” in Proc. 9th IEEE Asia-Pacific Conf. Commun., Penang, Malaysia, pp. 1042–1046, 2003.

[131] W. Stallings, Cryptography and Network Security. Englewood Cliffs, NJ: Prentice-Hall, 2005.