PERFORMANCE AND ANALYSIS OF TRANSFER CONTROL PROTOCOL OVER VOICE OVER WIRELESS LOCAL AREA NETWORK

A Thesis

Presented to

the Faculty of the College of Science and Technology

Morehead State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Rajendra Patil

August 2008

Accepted by the faculty of the College of Science and Technology, Morehead State University, in partial fulfillment of the requirements for the Master of Science degree.

Ahmad Zargari
Director of Thesis

Master's Committee: Ahmad Zargari, Chair

Robert E. Frank

8-11-2008
Date

PERFORMANCE AND ANALYSIS OF TRANSFER CONTROL PROTOCOL OVER VOICE OVER WIRELESS LOCAL AREA NETWORK

Rajendra Patil, M.S.
Morehead State University, 2008

Director of Thesis: Ahmad Zargari

Transfer control Protocol/Internet Protocol is standard network protocol stack for seamless communication on the Internet. There are significant packet losses due to transmission errors, which results in degradation in TCP performance. The implementation of VoIP faces problems like latency and jitter; but the main cause of the packet loss is congestion. There are many challenging questions ahead of deploying the VoIP over existing network. A detailed simulation approach for deploying VoIP successfully is offered. The simulation tool used in this paper is the OPNET network simulator. OPNET has gained considerable popularity in academia as it is being offered free of charge to academic institutions. For introducing a new network service such as VoWLAN, one has to characterize first the nature of its traffic, QoS requirements, and any additional components or devices. The number of VoIP calls that can be sustained by an existing network while satisfying QoS requirements of all network services and leaving adequate capacity for future growth.

As a case study, the simulation approach was applied on a typical network of a small enterprise. The paper presents a detailed description of simulation models for network topology and elements using OPNET. The paper describes modeling and representation of background and VoIP traffic, as well as various simulation configurations. Moreover, many design and engineering issues pertaining to the deployment of VoIP are discussed. These issues include characteristics of VoIP traffic and QoS requirements, VoIP flow and call distribution, defining future growth capacity, and measurement and impact of background traffic.

Accepted by: ______________________, Chair

____

____

ACKNOWLEDGEMENTS

First and foremost I offer my sincerest gratitude to my advisor, Dr. Ahmad Zargari, who has supported me throughout my thesis with his patience and knowledge whilst allowing me the room to work in my own way. I attribute the level of my Masters degree to his encouragement and effort and without him this thesis, too, would not have been completed or written. One simply could not wish for a better or friendlier advisor. In my thesis, I have applied a significant knowledge what I acquired in the two years of my Masters in Industrial Technology. A very special thanks goes out to Dr. You for her guidance in understanding the various concepts of Transfer Control Protocol over Wireless Network. I would also like to thank Dr. Frank for agreeing to be on my thesis committee despite his extremely busy schedule. I must also acknowledge Dr. Frank for his suggestions.

Dr. Zargari, yet again has offered much advice and insight throughout my work on Video Conference over Internet Protocol. In my daily work I have been blessed with a friendly and cheerful group of fellow students. The Department of Industrial Engineering Technology has provided the support and equipment I have needed to produce and complete my thesis. Finally, I thank my parents for supporting me throughout all my studies at Morehead State University, fulfilling my dream of acquiring the quality education in United States of America.

Table of Contents

Chapter 1

Introduction

The number of users accessing the internet using wireless technology is growing very fast; it is going to increase day by day. As the internet has become part of daily life, it is changing the way we communicate and search for information. Wireless introduces a different environment from the one in the fixed networks, with the limitation of the physical layer. A wireless mobile internet means any communication task like email, web browsing, internet radio that is possible with the wired internet connection is equally possible with a wireless internet connection (Mark Sportack 2005).

TCP/IP stands for Transmission Control Protocol/Internet Protocol. TCP is the most popular protocol on the internet. It works satisfactorily for the reliable delivery of data. Most of the popular applications like The World-Wide Web, File Transfer Protocol and email requires reliable delivery over the network. TCP was designed mainly to perform well on fixed networks, where the main functionality is to utilize the whole bandwidth and avoid overloading the network. But nowadays users want all the applications to run on wireless networks. Optimizing TCP for the wireless environment has been the major area of research for the last few years. One of the primary reasons for the widespread use of TCP on the internet is its inbuilt algorithms for congestion control and avoidance (Mark Sportack 2005). Over the years, the internet community has incorporated new schemes into the TCP suite to make these protocols more robust to congestion. (Bakshi 1997).

Packets on the internet may get lost either due to congestion or due to corruption by the underlying physical medium. Given the low bit error rates of wired links, almost all losses are related to congestion. TCP's reaction to losses is based on this very observation (Bikram 1996). A TCP connection is a bidirectional flow controlled, reliable stream of data between two end points, which is identified by an IP- address; TCP uses a sliding window flow control. This window sets the limit on the amount of data to be sent without acknowledgement from the receiver. The result is called "Acknowledgement clock" (ACK clock) where the window is constant and the timing for every packet is measured by the reception of the ACK of the previous packet; the window size is then adjusted according to the ACK signal received. TCP reacts by setting a slow-start threshold to half the value of congestion window, subsequently decreasing the congestion window to one, and entering the slow-start phase. This measure would appear severe, but works well because cutting the window size and limiting the amount of unacknowledged data on the network is the most effective way of dealing with congestion. In addition to the above measures, the timeout value is doubled upon each consecutive packet loss. Only upon receipt of an acknowledgement for a non- retransmitted packet is the timeout value recomputed (Parsa 1999).

These days a massive deployment of Videoconference over Internet protocol (VoIP) is taking place over data networks. Most of these networks are Ethernet-based and running IP protocol. Many network managers are finding it very attractive and cost effective to merge and unify voice and data networks into one. It is easier to run, manage, and maintain. However, one has to keep in mind that IP networks are best-effort networks that were designed for non-real time applications. On the other hand, VoIP requires timely packet delivery with low latency, jitter, packet loss, and sufficient bandwidth. To achieve this goal, an efficient deployment of VoIP must ensure that these real-time traffic requirements can be guaranteed over new or existing IP networks.

When deploying a new network service such as VoIP over existing networks, many network architects, managers, planners, designers, and engineers are faced with common strategic and sometimes challenging questions. What are the QoS requirements for VoIP? How will the new VoIP load impact the QoS for currently running network services and applications? Will my existing network support VoIP and satisfy the standardized QoS requirements? If so, how many VoIP calls can the network support before upgrading prematurely any part of the existing network hardware?

These challenging questions have led to the development of some commercial tools for testing the performance of multimedia applications in data networks. A list of the available commercial tools that support VoIP is listed in (B. Karacali, M. Bearden, L. Denby, J. Meloche and D.T. Stott, 2002; B. Karacali, L. Denby and J. Melche, 2004). For the most part, these tools use two common approaches in assessing the deployment of VoIP into the existing network. One approach is based on first performing network measurements and then predicting the network readiness for supporting VoIP. The prediction of the network readiness is based on assessing the health of network elements. The second approach is based on injecting real VoIP traffic into the existing network and measuring the resulting delay, jitter, and loss.

Other than the cost associated with commercial tools, none of the commercial tools offers a comprehensive approach for successful VoIP deployment. In particular, none give any prediction for the total number of calls that can be supported by the network taking into account important design and engineering factors. These factors include VoIP flow and call distribution, future growth capacity, performance thresholds, impact of VoIP on existing network services and applications, and impact background traffic on VoIP. This paper attempts to address those important factors and lay out a comprehensive methodology for a successful deployment of any multimedia application such as VoIP and video-conferencing. However, the paper focuses on VoIP as the new service of interest to be deployed. The paper also contains many useful engineering and design guidelines, and discusses many practical issues pertaining to the deployment of VoIP. These issues include characteristics of VoIP traffic and QoS requirements, VoIP flow and call distribution, defining future growth capacity, and measurement and impact of background traffic. TCP assumes that each packet loss is solely due to congestion; however, in a wireless network packet losses may be unrelated to congestion.

Review of TCP/IP:

TCP/IP is actually a collection of protocols that govern the way data travels from one machine to another across networks. A protocol is designed to perform a specific function. These protocols are divided into two categories based on their functions, one focuses on the processing and handling data from applications called TCP suite. The other is network oriented and designed to accommodate the transmission and receipt of application data across a network called IP suite. TCP/IP enables different types of computers and devices to use networks and contact each other and share the information (Charles L. Hedrick 1987).

TCP/IP protocols work together to break the data into small pieces that can be efficiently handled by the network known as segments and the process is called segmentation. Then these segments are wrapped in a data structure known as packets. A packet has all the information that a network needs for the delivery of the data to its destination and then acknowledge delivery. It verifies the receipt of the data on the other end of the transmission and reconstructs the data in its original form.

Packetization is the communication by two or more computers sending and receiving individual packets of data. It was first demonstrated when two computers where connected with a 15 foot cable at the University of California – Los Angeles. At first the data was sent one bit at a time without the benefits of having first been segmented for transmission. This was the first simple test that changed the way for technologies that no one could have predicted.

In the TCP/IP suite, TCP and IP are the most important parts but there are other protocols that we have to take into consideration. Some of these other protocols are: File Transfer Protocol (FTP), Hypertext Transfer Protocol (HTTP), Simple Mail Transfer Protocol (SMTP), Post Office Protocol Version 3 (POP3) and Telecommunication Network (TELNET). FTP enables the user to send and receive files over the Internet. HTTP is the application layer protocol used to communicate between Web servers and Web browsers. SMTP is the application layer protocol used to move messages from mail server to mail server on the Internet. POP3 is the application layer protocol used to communicate mail servers and mail software on client computers. TELNET enables the users on one computer to log into other computers on the Internet.

Significance Of Study

The demand for continuous network connectivity exists for various wired and wireless integrated networks. Since TCP is the standard network protocol stack for communication on the internet, its use over the network is a certainty because it allows seamless integration with the fixed infrastructure. The significance of VoIP is underlined by the increasing demand for higher quality of service from consumers. VoIP is not a traditional phone call, but a technology that offers yet another substitute for in-person voice. Traditional voice technology is dependent upon privately-owned networks; the Internet however is non-proprietary, part of the public domain is the main distinguishing factor between VoIP and traditional voice technology. Internet can ride across any privately owned networks, as well as across a number of other technologies. The study will help network researchers and designers to determine quickly and easily how well VoIP will perform on a network prior to deployment. Prior to the purchase and deployment of VoIP equipment, it is possible to predict the number of VoIP calls that can be sustained by the network while satisfying QoS requirements of all existing and new network services and leaving enough capacity for future growth.

Statement Of Problem

TCP is a popular transport protocol used in present-day internet. When packet losses occur, TCP assumes that the packet losses are due to congestion, and responds by reducing its congestion window. When a TCP connection traverses a wireless link, a significant fraction of packet losses may occur due to transmission errors. TCP responds to such losses also by reducing the congestion window. This results in unnecessary degradation in TCP performance. VoIP implementations face problems dealing with latency and jitter. The principal cause of packet loss is congestion, which can sometimes be managed or avoided. Effective software programming translates into effective management of calls—in terms of routing of the calls through the least congested paths—and clarity of voice.

Definitions Of Terms

- Client - A device (or application) that initiates a request for a connection with a server.
- Device - A network entity that is capable of sending and receiving packets of information and has a unique device address. A device can act as both a client and a server within a given context or across multiple contexts. For example, a device can service a number of clients (as a server) while being a client to another server (Xylomenos 1999).
- Origin Server - The server on which a given resource resides or is to be created. Often referred to as a web server or an Hypertext Transfer Protocol (HTTP) server.
- Proxy - An intermediary program that acts as both a server and a client for the purpose of making requests on behalf of other clients.
- Router - An intermediary mechanism that determines the path taken by IP packets.
- Server - A device (or application) that passively waits for connection requests from one or more clients. A server may accept or reject a connection request from a client.
- Terminal - A device providing the user with user agent capabilities, including the ability to request and receive information. Also called a mobile terminal or mobile station.

- User - A user is a person who interacts with a user agent to view, hear, or otherwise use a resource.
- User Agent - A user agent is any software or device that interprets Wireless Markup Language (WML), WMLScript, Wireless Telephony Application Interface (WTAI) or other resources. This may include textual browsers, voice browsers, search engines, etc.
- WebServer - The same as Origin Server.
- Delay - It is the time taken by a packet to reach its destination starting from the time it leaves its source.
- Jitter - A fluctuation in a transmission signal.
- Latency – It is a time delay between the moment something is initiated, and the moment one of its effects begins or becomes detectable.
- TCP - TCP/IP stands for Transmission Control Protocol/Internet Protocol; the suite of communications protocols used to connect hosts on the Internet.
- VoWLAN - The union of Wireless Local Area Network (WLAN) and Internet Protocol (IP) telephony technologies is Voice over WLAN (VoWLAN), which enables voice communications throughout a network served by a WLAN.

Chapter 2

Historical Overview

The primary reason for developing TCP/IP was the internet. They were developed together and TCP/IP was use to provide the mechanism for implementing the Internet. The TCP/IP protocols were first developed by the United State Defense Advanced Research Projects Agency (DARPA). At the start, ARPAnet was used for the number of protocols that has been modified from existing technologies. All the technologies had limitations in concept or the capacity when used on the ARPAnet. So the developers of the new network realized that using these technologies on ARPAnet will lead to problems (H. Gilbert 1995)..

In 1974, a new set of core protocols for ARPAnet was proposed and the official name for the set of protocols was TCP/IP Internet Protocol Suite. In March 1982, the US Department of Defense made TCP/IP the standard for all military computers. In 1983, sites connected to ARPAnet were supposed to switch to TCP/IP, which further enhanced the scope and importance of ARPAnet. During the same time, private regional service providers and government agencies like National Science Foundation were building their own networks. All these networks used TCP/IP as there native protocols (Andrew Anderson 1996).

The government realized the importance of ARPANET for uses other than military information and they created MILNET (Military Network) along side it. MILNET was limited for only military use but there were gateways created so the two networks could communicate. Finally, ARPANET evolved into what is now known as the Internet, in today's standards. ARPANET finally was out of use in 1990 and was finally turned off.

Review of Literature

These days a massive deployment of Voice over Wireless Lan (VoWLAN) is taking place over data networks. Most of these networks are Wireless LAN-based and running IP protocol. Many network managers are finding it very attractive and cost effective to merge and unify voice and data networks into one. It is easier to run, manage, and maintain (B. Karacali, L. Denby and J. Melche, 2004; B. Karacali, M. Bearden, L. Denby, J. Meloche and D.T. Stott, 2002). However, one has to keep in mind that IP networks are best-effort networks that were designed for non-real time applications. On the other hand, VoWLAN requires timely packet delivery with low latency, jitter, packet loss, and sufficient bandwidth. To achieve this goal, an efficient deployment of VoWLAN must ensure these real-time traffic requirements can be guaranteed over new or existing IP networks.

When deploying a new network service such as VoWLAN over existing network, many network architects, managers, planners, designers, and engineers are faced with common strategic and sometimes challenging questions. What are the QoS requirements for VoWLAN? How will the new VoWLAN load impact the QoS for currently running network services and applications? Will my existing network support VoWLAN and satisfy the standardized QoS requirements? If so, how many

VoWLAN calls can the network support before upgrading prematurely any part of the existing network hardware?

These challenging questions have led to the development of some commercial tools for testing the performance of multimedia applications in data networks. For the most part, these tools use two common approaches in assessing the deployment of VoWLAN into the existing network. One approach is based on first performing network measurements and then predicting the network readiness for supporting VoWLAN. The prediction of the network readiness is based on assessing the health of network elements. The second approach is based on injecting real VoWLAN traffic into existing network and measuring the resulting delay, jitter, and loss.

Other than the cost associated with the commercial tools, none of the commercial tools offer a comprehensive approach for successful VoWLAN deployment. In particular, none give any prediction for the total number of calls that can be supported by the network taking into account important design and engineering factors. These factors include VoWLAN flow and call distribution, future growth capacity, performance thresholds, impact of VoWLAN on existing network services and applications, and impact background traffic on VoWLAN.

Related work

The work in this article is related to the performance analysis of different architectures for distributed IR systems, especially focused on the networking issues. The main articles in this area using different networking approaches are described briefly below.

Using a small collection, Burkowski (1990) described a simulation study which measures the retrieval performance of a distributed IR system. In his experiments, a high performance Wi-Fi is used to connect a group of user workstations with a cluster of servers, although the network times are not considered in the simulation model.

Tomasic and Garcia-Molina (1993) studied the performance of several parallel query processing strategies using various options for the organization of the inverted index, simulating a small group of hosts interconnected through a local area network (Wi-Fi). They represented a shared access network using a global Wi-Fi queue and included a fixed overhead for each network transmission.

Coevreur et al. (1994) analyzed the performance of searching large text collections (more than 100GB data) on parallel systems. They used simulation models to investigate three different hardware architectures (a mainframe system, a collection of RISC processors and a special purpose machine architecture) and search algorithms. For the second architecture, a Wireless LAN network interconnected the system, although the defined model did not include the simulation of the communications.

Cahoon and McKinley (1996) described the result of simulated experiments on the distributed INQUERY architecture. They used the observed behaviour of a mono-server implementation to estimate the performance figures for a distributed implementation. In their simulation experiments the network is represented as the sender's and receiver's overhead and the network latency. The sender's and receiver's overhead is the processing time to read and write a message on the network, and it was measured empirically. The network latency is the amount of time the message spends on the network and depends on the size of the message and the bandwidth of the network.

Ribeiro-Neto and Barbosa (1998) used a simple analytical model coupled with a small simulator to study how query performance is affected by different parameters (e.g. network speed) in distributed digital libraries. In their work, the network is represented as part of the analytical model including a parameter that corresponds to the average time to transfer 1 byte from one host to another.

Lu and McKinley (2000) used the same simulator as Cahoon & McKinley, (1996) to analyze the effects of partial collection replication, and to improve the performance on a collection of 1TB.

Some other articles studied the performance of different architectures for distributed IR systems operating with real implementations. Martin, Macleod, and Nordin (1986) described the design of a distributed information system for full text retrieval, developed on a network of PC's interconnected by PC Network. The authors did not provide any details about the network technology used, although they described briefly the used communication protocol.

Lin and Zhou (1993) implemented a distributed IR system on a network of workstations, showing large speedup improvements due to parallelization. The authors used a Wireless LAN network to interconnect a cluster of DEC5000 workstations.

Finally, in Cacheda et al. (2004) and Cacheda Plachouras et al. (2005), the performance of a distributed, replicated and clustered system simulating a very large Web collection such as SPIRIT (Jones et al., 2002) was performed. The simulated network represented a shared access local area network operating at 100 Mbps, following the model of Tomasic and Garcia-Molina (1993). Two main bottlenecks were identified in the distributed and replicated IR systems: the brokers and the network. The load on the brokers was mainly due to the number of local answer sets to be merged. The network bottleneck was due to the high number of query servers and the continuous data interchange with the brokers, especially in a replicated Infrared (IR) system.

The network model

In previous work (Cacheda et al., 2004 and Cacheda, Plachouras, et al., 2005), the simulated distributed IR system assumed a single Wi-Fi that was represented by a single First Come First Serve (FCFS) infinite length queue. This Wi-Fi managed all the messages sent by the brokers to the query servers and the answers from the query servers to the brokers. Cacheda et al., 2004 and Cacheda, Plachouras, et al stated that the service time for a request was calculated by the following equation: $LANOverhead+RequestLength\times(LANBandwitdh/8)^{-1}\times1000$.

The parameters used in the simulation of the network are described in Table 1. The *RequestLength* parameter depends on the type of message sent. If a query is sent to the query servers, the value of the *QuerySize* parameter will be used. If the local answer set for query q_i is sent from query server j to the broker, then the length of the packet will be: $tr_{i,j} \times DocAnswerSetSize$.

Table 1.

Parameters for the network model

Parameter	**Value**	**Description**
LANOverhead	0.1 ms	Network overhead for each packet sent
LANBandwidth	100 Mbps	Network speed (in bits per second)
QuerySize	100 bytes	Number of bytes sent from the broker to the query servers for each query request
DocAnswerSetSize	8 bytes	Number of bytes per document sent in the local answer sets (document id and document ranking)

This initial network model has certain limitations that reduce the capacities of the simulated IR systems. The network model represents a simple shared access network, where all the hosts of the distributed IR system (query servers and brokers) are included. This system is a relatively unrealistic environment, as it does not take into account certain physical restrictions of LANs.

LANs have two physical restrictions: the number of nodes that can be connected to the network and the maximum length of each segment. By definition, in a shared access network the maximum segment length is the maximum length for the network itself, while in a switched network, the maximum segment length is the maximum length from each node to the switch. For example, according to the IEEE 802.3 specifications for a shared network using a bus topology at 10 Mbps in 10BASE5, the maximum segment length is 500 m, with a maximum of 100 nodes per segment. In a switched network 10BASET the maximum segment length is 100 m and the maximum number of nodes is determined by the number of interfaces of the switch (Spurgeon, 2000). These limitations, especially the maximum number of nodes per segment, were not taken into account in designing the previous network model, which limits its capacity to represent a real environment.

Moreover, all network technologies have a certain overhead that is inherent to the operation of the network. This overhead consists of the increase in the amount of information sent via the network, through the incorporation of the headers of different protocols, which finally results in a reduction in the effective bandwidth and a subsequent increase in transmission time.

In previously proposed network models, this overhead is represented approximately by the *LANOverhead* parameter, with a constant value of 0.1 ms, following the model described by Tomasic and Garcia-Molina (1993). Although this is a good approximation for a generic case of sending a packet via the network, it is necessary to take into account the specific sizes of the headers used in the communication protocol of the distributed IR system, and in particular to consider IP fragmentation for packets that exceed a given size.

With the aim of improving the limitations of the previously proposed network model, this paper has defined a new network model based on Switched Wi-Fi model for a distributed IR system (Fig 1), equivalent to a switched network FastEthernet 100BASE-T at 100 Mbps. This new model represents a switched local area network (compared to a shared access network), where the hosts are interconnected via one or more switches. Depending on the number of hosts to be interconnected, one or more switches are used, assuming that each switch has a capacity for 64 hosts. Moreover, the overhead estimation is carried out exhaustively, taking into account the different headers of the communication protocols, IP fragmentation, and even the propagation delay. The design of this new network model has also extended the capacity to represent multicast traffic, compared to the previous model that only allowed unicast messages to be used amongst the different hosts.

In more details, the design of the switched Wi-Fi is shown in the figure 1 below. Each host connected to the network has two first come first service (FCFS) infinite length queues (one for sending packets and another for receiving packets), compared to the previous model, which used a global FCFS infinite length queue for the entire network. This model makes it possible to represent a switched Wi-Fi as two independent origin hosts may communicate with two independent destination hosts without having to have shared access to the network.

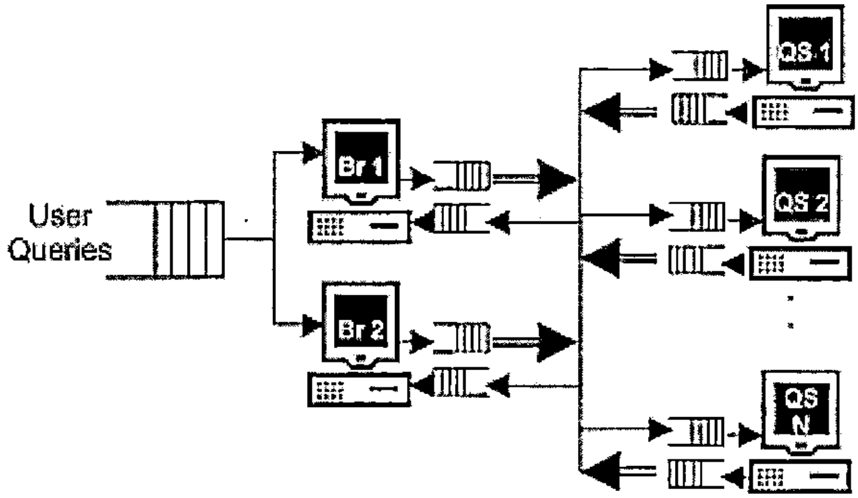


Fig 1. Switched Wi-Fi model for a distributed IR system (Br: broker, QS: query server).

The send queue represents the output buffer at the output interface of the transmitter, while the reception queue represents the entry buffer at the receiver interface. To send a packet, this is introduced in the send queue of the origin host and then passes directly to the reception queue of the destination host.

The output of a packet from the send queue to the reception queue is carried out when the packet is the first element in the queue, if and only if the receiver is not receiving another packet. In the reception queue, each packet has an assigned service time (before being delivered to the host), equivalent to the transmission time via the network, and it is calculated as it will be shown below.

Furthermore, the ability to manage multicast traffic has been incorporated in the newly implemented network model. This is achieved by permitting, the direct communication from the send queue of the sender to multiple entry queues, considering the time associated with the sending of a single packet. The advantages of messages of this type will be analyzed in detail in the next section.

When estimating the transmission time of a packet via the network, consideration has been given to the headers of the different protocols used in the communication and the propagation delay of the signal via the transmission system. The communication protocol used by the distributed system between the query servers and the brokers is assumed to be based on a non-connection oriented service, taking into account the volume of interconnected hosts and the characteristics of the exchange of data. Therefore, it is considered that the transport layer uses the UDP protocol (with a header size of 8 bytes), the network level uses IP (with a standard header of 20 bytes) and for the connection level, the Wireless LAN protocol was considered (with a header of 26 bytes). This network model also considers IP fragmentation according to the mechanism described in RFC 791 (Postel, 1981). Put briefly, if the size of a message at network level is greater than the Maximum Transfer Unit (MTU) corresponding to the connection level (1500 bytes in the case of Wireless LAN), the original message is divided into fragments of acceptable size for the MTU. These fragments are only reconstructed when they reach the final destination.

Moreover, propagation delay represents the time that an electromagnetic signal takes to circulate from one end of the segment to the other. This time is usually measured in bits and mainly depends on the size of the segment and the type of used transmission method. In a FastEthernet 100BASE-T network, the maximum level for propagation delay is 512 bits (corresponding to 5.12 ns) and equivalent to a maximum segment size approximately 200 m (Spurgeon, 2000). In the simulated model, it is assumed as a uniform distribution of the machines, hence an intermediary value of 256 bits was chosen (corresponding to 2.56 ns), equivalent to a mean segment size (approximately 100 m).

In these experiments, two computers were connected via a crossed cable representing a connection through a switch (the delay introduced by a switch is negligible). The direct connection using a switch was not possible, as the available switches had insufficient buffers which caused packet losses. A transmitter process was installed in one computer and a receiver process in the other. The transmitter process was responsible for generating multiple consecutive packets (from 10 to 1025), which were received by the receiver process, responsible for measuring transmission times.

In an initial experiment, communication was evaluated via Wireless LAN at 10 Mbps connecting two Ultra Sparc 1 (128MB RAM and one 167 MHz processor) with 10BASE-T network cards and sending messages of 1000, 2000, 3000 and 4000 bytes, measuring the corresponding send times.

Existing network

Fig. 2 illustrates a typical network topology for a small enterprise residing in a high-rise building. The network is Wireless LAN-based and has two Layer-2 Wireless LAN switches connected by a router. The router is Cisco 2621, and the switches are 3Com Superstack 3300. Switch 1 connects Floors 1 and 2 and two servers; while Switch 2 connects Floor 3 and four servers. Each floor's Wi-Fi is basically a shared Wireless LAN connecting employee PCs with workgroup and printer servers. The network makes use of Virtual LAN (VLANs) in order to isolate broadcast and multicast traffic. A total of five LANs exist. All VLANs are port based. Switch 1 is configured such that it has three VLANs. VLAN1 includes the database and file servers. VLAN2 includes Floor 1. VLAN3 includes Floor2. On the other hand, Switch 2 is configured to have two VLANs. VLAN4 includes the servers for E-mail, HTTP, Web and cache proxy, and firewall. VLAN5 includes Floor 3. All the links are switched Wireless LAN 100 Mbps full duplex except for the links for Floors 1–3 which are shared Wireless LAN 100 Mbps half duplex.

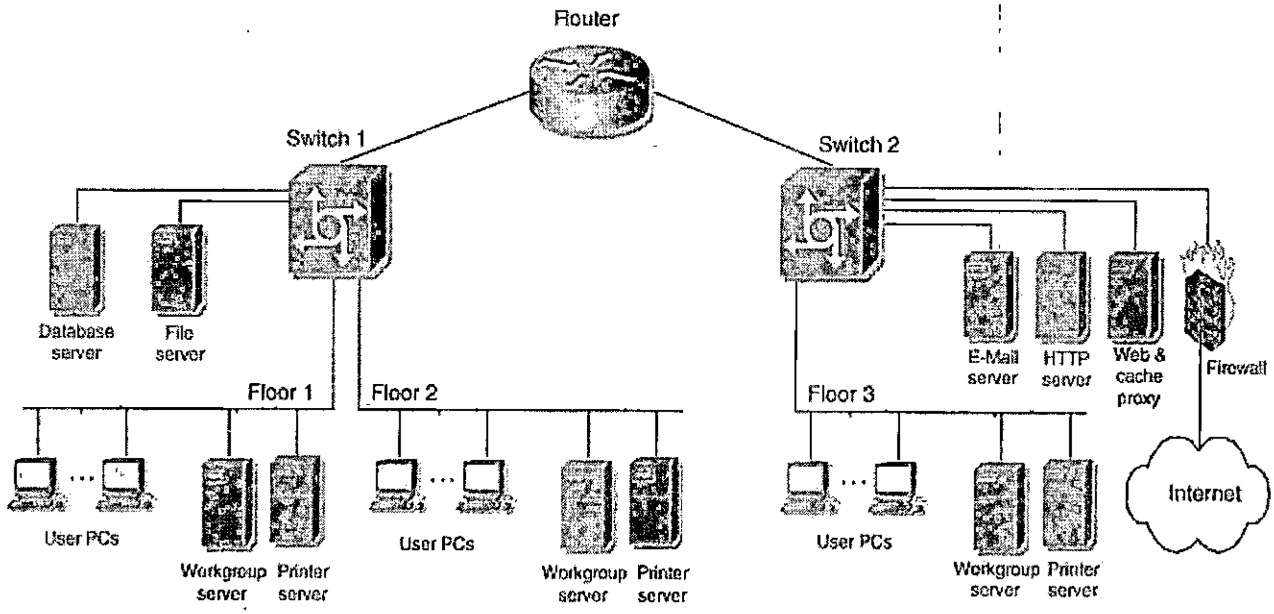


Fig. 2. Logical diagram of a small enterprise.

Chapter 3

Methodology:

In this paper, the popular OPNET simulation tool can be leveraged to assess the readiness of existing data networks to support video-conferencing. OPNET modeler does not have built-in features to support videoconferencing or deployment of real-time services. In the literature, there exists no known simulation approach on how to deploy a popular real-time network service such as videoconferencing. OPNET has gained considerable popularity in academia as it is being offered free of charge to academic institutions. That has given OPNET an edge over DES NS2 in both the market place and academia. Another reason to choose OPNET is the fact that OPNET contains a vast amount of models of commercially available network elements, and has various real-life network configuration capabilities. This makes the simulation of a real-life network environment close to reality. Other features of OPNET include GUI interface, comprehensive library of network protocols and models, source code for all models, graphical results and statistics, etc.

In previously related work Salah and Alkhoraidly (2006), an analytic approach based on the principles of queuing networks was presented to determine approximately the number of video sessions an existing data network can support. In sharp contrast to Salah (2006), this paper primarily focuses on showing how to deploy successfull videoconferencing using OPNET modeling and simulation. The simulation configuration and setup for videoconferencing are considerably different when considering the deployment of both voice and video calls simultaneously.This paper considers two types of traffic (traffic of fixed video packet sizes as that used for the analytic approach as described by Salah (2006) and also traffic of variable video packet sizes measured from well-known traffic traces. This paper discusses in great detail the simulation, configuration, setup, and generation of traffic for videoconferencing. Such information can be extremely useful for network researchers and engineers who are interested in deploying videoconferencing. The paper also gives in-depth analysis and interpretations of OPNET simulation results.

VoWLAN traffic characteristics, requirements, and assumptions

For introducing a new network service such as VoWLAN, one has to characterize first the nature of its traffic, QoS requirements, and any additional components or devices. For simplicity, we assume a point-to-point conversation for all VoWLAN calls with no call conferencing. For deploying VoWLAN, a gatekeeper or CallManager node has to be added to the network (B Goode, 2002; P. Mehta and S. Udani, 2001; W. Jiang and H. Schulzrinne, 2001). The *gatekeeper* node handles signaling for establishing, terminating, and authorizing connections of all VoWLAN calls. Also a VoWLAN *gateway* is required to handle external calls. A VoWLAN *gateway* is responsible for converting VoWLAN calls to and from the Public Switched Telephone Network (PSTN). As an engineering and design issue, the placement of these nodes in the network becomes crucial. Other hardware requirements include a VoWLAN client terminal, which can be a separate VoWLAN device, i.e. IP phones, or a typical PC or workstation that is VoWLAN-enabled. A

VoWLAN-enabled workstation runs VoWLAN software such as IP SoftPhones (B. Duysburgh, S. Vanhastel, B. DeVreese, C. Petrisor and P. Demeester, 2002; W. Jiang, K. Koguchi and H. Schulzrinne, 2003).

Fig. 3 identifies the end-to-end VoWLAN components from sender to receiver (A. Markopoulou, F. Tobagi and M. Karam, 2003). The first component is the *encoder* which periodically samples the original voice signal and assigns a fixed number of bits to each sample, creating a constant bit rate stream. The traditional sample-based encoder G.711 uses Pulse Code Modulation (PCM) to generate 8-bit samples every 0.125 ms, leading to a data rate of 64 kbps. The packetizer follows the *encoder* and encapsulates a certain number of speech samples into packets and adds the RTP, UDP, IP, and Wireless LAN headers. The voice packets travel through the data network. An important component at the receiving end, is the *playback buffer* whose purpose is to absorb variations or jitter in delay and provide a smooth playout. Then packets are delivered to the *depacketizer* and eventually to the *decoder* which reconstructs the original voice signal.

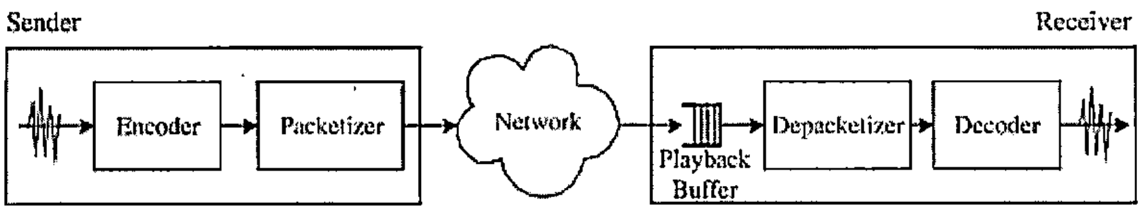


Fig. 3. VoWLAN end-to-end components.

The widely adopted recommendations of H.323, G.711, and G.714 standards for VoWLAN QoS requirements are followed in the paper. Table 1 compares some commonly used ITU-T standard codecs and the amount of one-way delay that they impose. To account for upper limits and to meet desirable quality requirements according to ITU recommendation P.800, we will adopt G.711u codec standards for the required delay and bandwidth. G.711u yields around 4.4 MOS rating. MOS, *Mean Opinion Score*, is a commonly used VoWLAN performance metric given in a scale of 1–5, with 5 being the best (A. Takahasi and H. Yoshino, 2004; L. Sun and E.C. Ifeachor, 2003). However, with little compromise to quality, it is possible to implement different ITU-T codecs that yield much less required bandwidth per call and relatively, higher but, acceptable end-to-end delay. This can be accomplished by applying compression, silence suppression, packet loss concealment, queue management techniques, and encapsulating more than one voice packet into a single Wireless LAN frame (A. Markopoulou, F. Tobagi and M. Karam, 2003; B Goode, 2002; J. Walker, J. Hicks, 2002; W. Jiang and H. Schulzrinne, 2002).

Table 1.

Common ITU-T codecs and their defaults

Codec	**Data rate (kbps)**	**Datagram size (ms)**	**A/D Conversion delay (ms)**	**Combined bandwidth (bi-directional) (kbps)**
G.711u	64.0	20	1.0	180.80
G.711a	64.0	20	1.0	180.80
G.729	8.0	20	25.0	68.80
G.723.1 (MPMLQ)	6.3	30	67.5	47.80
G.723.1 (ACELP)	5.3	30	67.5	45.80

End-to-end delay for a single voice packet

Fig. 3 illustrates the sources of delay for a typical voice packet. The end-to-end delay is sometimes referred to by M2E or Mouth-to-Ear delay (W. Jiang, K. Koguchi and H. Schulzrinne, 2003). G.714 imposes a maximum total one-way packet delay of 150 ms end-to-end for VoWLAN applications. In (J.H James, B. Chen and L. Garrison, 2004), a delay of up to 200 ms was considered to be acceptable. This delay can be broken down into at least three different contributing components, which are as follows (i) encoding, compression and packetization delay at the sender (ii) propagation, transmission and queuing delay in the network and (iii) buffering, decompression, depacketization, decoding, and playback delay at the receiver.

Bandwidth for a single call

The required bandwidth for a single call one direction is 64 kbps. G.711 codec samples 20 ms of voice per packet. Therefore, 50 such packets need to be transmitted per second. Each packet contains 160 voice samples in order to give 8000 samples per second. Each packet is sent in one Wireless LAN frame. With every packet of size 160 bytes, headers of additional protocol layers are added. These headers include RTP+UDP+IP+Wireless LAN with preamble of sizes 12+8+20+26, respectively. Therefore, a total of 226 bytes, or 1808 bits, needs to be transmitted 50 times per second, or 90.4 kbps, in one direction. For both directions, the required bandwidth for a single call is 100 pps or 180.8 kbps assuming a symmetric flow.

Other assumptions

Throughout the analysis and work, the voice calls are assumed to be symmetric and no voice conferencing is implemented. The signaling traffic generated by the *gatekeeper* is also ignored. The analysis and design are based on the worst-case scenario for VoWLAN call traffic. The signaling traffic involving the *gatekeeper* is mostly generated prior to the establishment of the voice call and when the call is finished. This traffic is relatively small compared to the actual voice call traffic. In general, the *gatekeeper* generates no or very limited signaling traffic throughout the duration of the VoWLAN call for an already established on-going call (B Goode, 2002).

In this paper, no QoS mechanisms will be implemented that can enhance the quality of packet delivery in IP networks. A myriad of QoS standards are available and can be enabled for network elements. QoS standards may include IEEE 802.1p/Q, the IETF's RSVP, and DiffServ. Analysis of implementation cost, complexity, management, and benefit must be weighed carefully before adopting such QoS standards. These standards can be recommended when the cost for upgrading some network elements is high and the network resources are scarce and heavily loaded.

Perform network measurements

In order to characterize the existing network traffic load, utilization, and flow, network measurements have to be performed. This is a crucial step as it can potentially affect results to be used in analytical study and simulation. There are a number of tools available commercially and non-commercially to perform network measurements.

Network measurements must be performed for network elements such as routers, switches, and links. Numerous types of measurements and statistics can be obtained using measurement tools. As a minimum, traffic rates in bits per second (bps) and packets per second (pps) must be measured for links directly connected to routers and switches. To get adequate assessment, network measurements have to be taken over a long period of time, at least a 24-h period. Sometimes it is desirable to take measurements over several days or a week.

One has to consider the worst-case scenario for network load or utilization in order to ensure good QoS at all times including peak hours. The peak hour is different from one network to another and it depends totally on the nature of business and the services provided by the network. Table 2 shows a summary of peak-hour utilization for traffic of links in both directions connected to the router and the two switches of the network topology of Fig. 1 (Salah and Alkhoraidly, 2006). These measured results will be used in the analysis and simulation study.

Table 2.

Worst-case network measurements

Link	**Bit rate (Mbps)**	**Packet rate (pps)**	**Utilization (%)**
Router⇔Switch 1	9.44	812	9.44
Router⇔Switch 2	9.99	869	9.99
Switch 1⇔Floor 1	3.05	283	6.1
Switch 1⇔Floor 2	3.19	268	6.38
Switch 1⇔File Server	1.89	153	1.89
Switch 1⇔DB Server	2.19	172	2.19
Switch 2⇔Floor 3	3.73	312	7.46
Switch 2⇔Email Server	2.12	191	2.12
Switch 2⇔HTTP Server	1.86	161	1.86
Switch 2⇔Firewall	2.11	180	2.11
Switch 2⇔Proxy	1.97	176	1.97

Upfront network assessment and modifications

In this step, the existing networks are accessed and determine, based on the existing traffic load and the requirements of the new service to be deployed, if any immediate modifications are necessary. Immediate modifications to the network may include adding and placing new servers or devices, upgrading PCs, and re-dimensioning heavily utilized links. As a good upgrade rule, topology changes need to be kept to a minimum and should not be made unless it is necessary and justifiable. Over-engineering the network and premature upgrades are costly and considered poor design practices.

Based on the existing traffic load discussed in design steps in the previous section, all the links connecting the router and the switches and links connecting the servers and the switches are underutilized. If any of the links were heavily utilized, e.g. 30–50%, the network engineer should decide to re-dimension the link to 1-Gbps link at this stage. As for shared links of Floors 1–3, the replacement or re-dimensioning of these links must be decided on carefully. At first, it looks cost effective not to replace the shared-Wireless LAN Wi-Fi for each floor with a switched Wi-Fi. However, shared Wireless LAN scales poorly. More importantly, shared Wireless LAN offers zero QoS and are not recommended for real-time and delay-sensitive applications as it introduces excessive and variable latency under heavy loads and when subjected to intense bursty traffic (S. Riley and R. Breyer, 2000). In order to consistently maintain the VoWLAN QoS, a switched fast full-duplex Wireless LAN Wi-Fi becomes necessary.

Based on the hardware requirement for deploying VoWLAN, two new nodes have to be added to the existing network: a VoWLAN *gateway* and a *gatekeeper*. As a network design issue, an appropriate node placement is required for these two nodes. Since most of the users reside on Floors 1 and 2 are connected directly to Switch 1, connecting the *gatekeeper* to Switch 1 is practical in order to keep the traffic local. For the VoWLAN *gateway*, it is connected to Switch 2 in order to balance the projected load on both switches. Also, it is more reliable and fault-tolerant not to connect both nodes to the same switch in order to eliminate problems that stem from a single point of failure. For example, if Switch 2 fails, only external calls to and from the network will be affected. It is proper to include the gatekeeper to be a member of VLAN1 of Switch 1 which includes the database and file servers.

This isolates the gatekeeper from multicast and broadcast traffic of Floors 1 and 2. In addition, the *gatekeeper* can access locally the database and file servers to record and log phone calls. On the other hand, we create a separate VLAN for the *gateway* in order to isolate the *gateway* from multicast and broadcast traffic of Floor 3 and the servers of switch 2. Therefore, the network has now a total of six VLANs. Fig. 4 shows the new network topology after the incorporation of necessary VoWLAN components. As shown, two new *gateway* and *gatekeeper* nodes for VoWLAN were added and the three shared Wireless LANs were replaced by 100 Mbps switched Wireless LAN LANs.

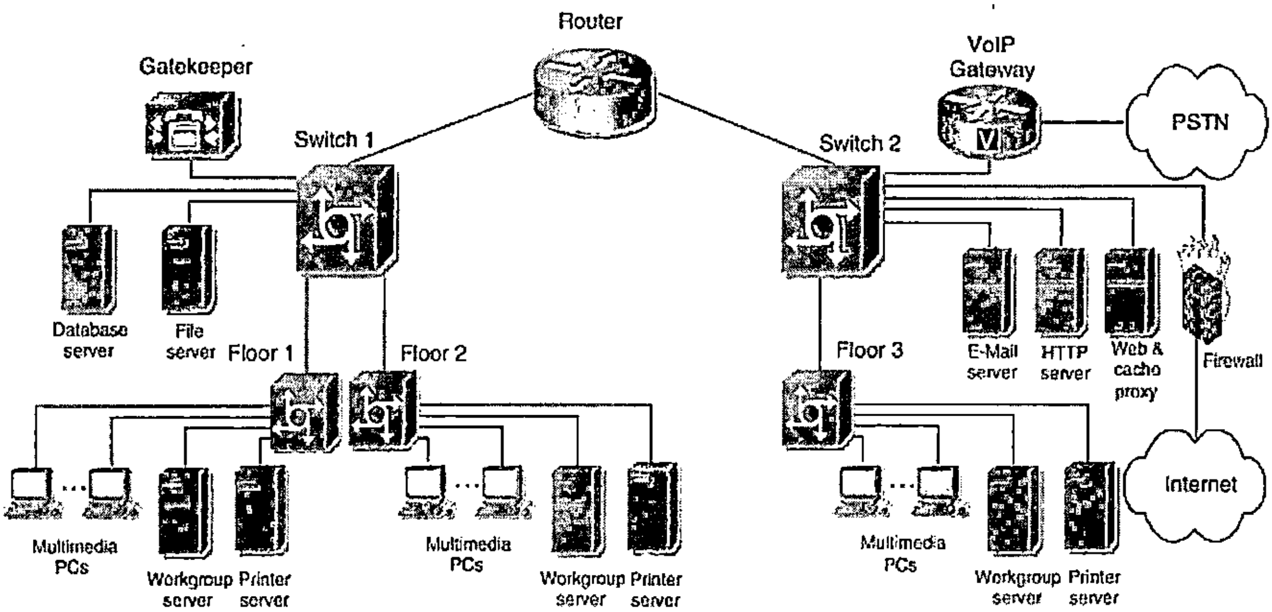


Fig. 4. Network topology with VoWLAN components.

Analysis

VoIP is bounded by two important metrics. First is the available bandwidth. Second is the end-to-end delay. The actual number of VoIP calls that the network can sustain and support is bounded by those two metrics. Depending on the network under study, either the available bandwidth or delay can be the key dominant factor in determining the number of calls that can be supported.

Bandwidth bottleneck analysis

Bandwidth bottleneck analysis is an important step to identify the network element, whether it is a node or a link that puts a limit on how many VoIP calls can be supported by the existing network. For any path that has N network nodes and links, the bottleneck network element is the node or link that has the minimum available bandwidth. According to (R. Prasad, C. Dovrolis,, M. Murray and K.C. Claffy, 2003), this minimum available bandwidth is defined as follows

$$A = \min_{i=1,\ldots,N} A_i,$$

and

$$A_i = (1 - u_i)C_i,$$

where C_i is the capacity of network element i and u_i is its current utilization. The capacity C_i is the maximum possible transfer or processing rate (Fig. 5).

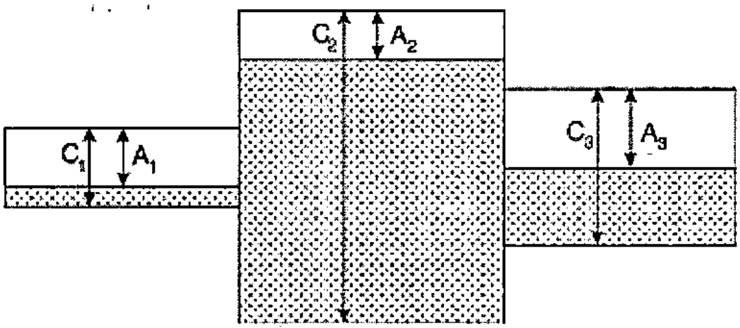


Fig. 5. Bandwidth bottleneck for a path of three network elements.

Therefore, the theoretical maximum number of calls that can be supported by a network element E_i can be expressed in terms of A_i as

$$\text{MaxCalls}_i = \frac{A_i(1 - \text{growth}_i)}{\text{CallBW}}, \qquad (1)$$

where $growth_i$ is the growth factor of network element E_i, and takes a value from 0 to 1. *CallBW* is the VoIP bandwidth for a single call imposed on E_i. As previously discussed in design section, the bandwidth for one direction is given as 50 pps or 90.4 kbps. In order to find the bottleneck network element that limits the total number of VoIP calls, one has to compute the maximum number of calls that can be supported by each network element, as in Eq. (1), and the percentage of VoIP traffic flow passing by this element. The percentage of VoIP traffic flow for E_i, denoted as $flow_i$, can be found by examining the distribution of the calls. The total number of VoIP calls that can be supported by a network can be expressed as:

$$\mathrm{TotalCallsSupported} = \min_{i=1,\ldots,N} \left(\frac{\mathrm{MaxCalls}_i}{\mathrm{flow}_i} \right). \tag{2}$$

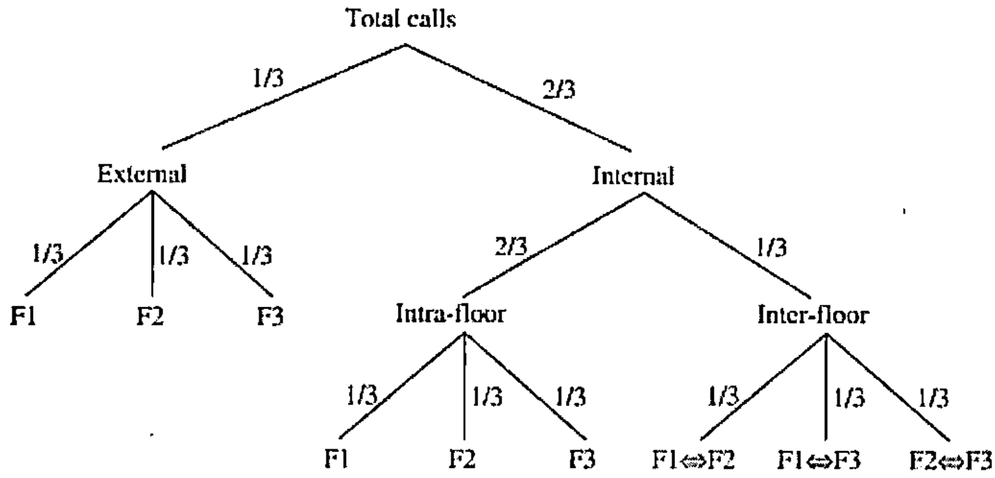
Let us for the sake of illustration compute the $\mathrm{MaxCalls}_i$ and flow_i supported by the Router, Switch 1, and uplink from Switch 2 to the Router. Table 3 shows the maximum calls that can be supported by those network elements. For the network example, $growth_i$ is chosen to be 25% for all network elements. u_i is determined by Table 2. C_i, for the router and the switch is usually given by the product datasheets. According to Cisco systems and 3Com networking, the capacity C_i for the router or the switch, is 25,000 pps and 1.3 M pps, respectively. flow_i is computed by examining the probability tree for call distribution.

Table 3.

Maximum VoIP calls support for few network elements

Network Element	C_i	u_i (%)	*CallBW*	$flow_i$	$MaxCalls_i$
Router	25,000 pps	6.72	100 pps	5/9	174
Switch 1	1.3 Mpps	0.13	100 pps	14/27	9737
Uplink from Switch 2 to Router	100 Mbps	9.99	90.4 kbps	16/27	746

Table 3 shows the MaxCalls_i for only three network elements. In order to find the actual calls that the network can sustain, i.e. TotalCallsSupported of Eq. (2), flow_i and MaxCalls_i have to be computed for all network elements. This can be automated by implementing the equations using MATLAB, and therefore these values can be computed quickly. When computing the MaxCalls_i for all network elements, it turns out that the router is the bottleneck element. Hence, TotalCallsSupported is 313 VoIP calls.



Probability tree describing the VoWLAN call distribution.

For the sake of illustration, u_i and flow$_i$ are computed. u_i can be computed by Table 2. For example, the utilization for the router is the total incoming traffic (or received traffic) into the router divided by the router's capacity. According to Table 2, this yields to (812+869)/25,000=6.72%. flow$_i$ can be computed using the probability tree shown above. For the router, flow$_i$ is the percentage of the inter-floor and external calls, which is (2/3)(1/3)+1/3. Similarly, flow$_i$ for Switch 1 and the uplink from Switch 2 to the router would be 14/27 and 16/27, respectively.

For Switch 1, flow$_i$ is the percentage of external calls going out of Floors 1 and 2, plus the percentage of inter-floor calls between Floors 1 and 2, Floors 1 and 3, and Floors 2 and 3. This can be expressed as (1/3){1/3+1/3}+(2/3)(1/3){2/3+(1/3)(1/3)}. Note that the fraction of inter-floor calls between Floors 1 and 2 is 2/3, since the calls pass through the switch twice as they get routed by the router back to Switch 1. For the uplink from Switch 2 to the router, flow$_i$ is the percentage of external calls going out of the three floors plus the percentage of inter-floor calls between Floors 1 and 3 and Floors 2 and 3. This can be simply expressed as (1/3){1/3+1/3+2/3}+(2/3)(1/3){1/3+1/3}. Note that the fraction of the external calls going out of Floor 3 is 2/3 since the calls pass through the link twice as they get routed by the router.

Delay analysis

As defined in previous section for the existing network, the maximum tolerable end-to-end delay for a VoIP packet is 150 ms. The maximum number of VoIP calls that the network can sustain is bounded by this delay. It should be always as certain that the worst-case end-to-end delay for all the calls must be less than 150 ms. It should be kept in mind that our goal is to determine the network capacity for VoIP, i.e. the maximum number of calls that existing network can support while maintaining VoIP QoS. This can be done by adding calls incrementally to the network while monitoring the threshold or bound for VoIP delay. When the end-to-end delay, including network delay, becomes larger than 150 ms, the maximum number of calls can then be known.

As described in the previous section, there are three sources of delay for a VoIP stream: sender, network, and receiver. An equation is given in (M. Karam and F. Tobagi, 2002) to compute the end-to-end delay D for a VoIP flow in one direction from sender to receiver

$$D = D_{pack} + \sum_{h \in Path} (T_h + Q_h + P_h) + D_{play},$$

Where D_{pack} is the delay due to packetization at the source. At the source, there is also D_{enc} and $D_{process}$. D_{enc} is the encoder delay of converting A/D signal into samples. $D_{process}$ is the PC of IP phone processing that includes encapsulation. In G.711, D_{pack} and D_{enc}, are 20 and 1 ms, respectively. Hence, it is appropriate for our analysis to have a fixed delay of 25 ms being introduced at the source, assuming a worst case situation. D_{play} is the playback delay at the receiver, including jitter buffer delay. The jitter delay is at most 2 packets, i.e. 40 ms. If the receiver's delay of $D_{process}$ is added, we obtain a total fixed delay of 45 ms at the receiver. $T_h+Q_h+P_h$ is the sum of delays incurred in the packet network due to transmission, queuing, and propagation going through each hop h in the path from the sender to the receiver. The propagation delay P_h is typically ignored for traffic within a WLAN. For transmission delay T_h and queueing delay Q_h we apply queueing theory. Hence the delay to be introduced by the network, expressed as $\sum_{h \in Path}(T_h + Q_h)$, should not exceed (150–25–45) or 80 ms.

We utilize queueing analysis to approximate and determine the maximum number of calls that the existing network can support while maintaining a delay of less than 80 ms. In order to find the network delay, we utilize the principles of the Jackson theorem for analyzing queueing networks. In particular, we use the approximation method of analyzing queueing networks by decomposition discussed in (K.M. Chandy and C.H. Sauer, 1978). In this method, the arrival rate is assumed to be Poisson and the service times of network elements are exponentially distributed. Analysis by decomposition is summarized in first isolating the queueing network into subsystems, e.g., single queueing node. Next, analyzing each subsystem separately, considering its own network surroundings of arrivals and departures. Then, finding the average delay for each individual queueing subsystem. And finally, aggregating all the delays of queueing subsystems to find the average total end-to-end network delay.

For our analysis, we assume the VoIP traffic to be Poisson. In reality, the inter-arrival time, $1/\lambda$, of VoIP packets is constant, and hence its distribution is deterministic. However, modeling the voice arrival as Poisson gives adequate approximation according to (M. Karam and F. Tobagi, 2002), especially when employing a high number of calls. More importantly, the network element with a non-Poisson arrival rate makes it difficult to approximate the delay and lead to an intractable analytical solution. Furthermore, analysis by decomposition method will be violated if the arrival rate is not Poisson.

Fig. 6 shows queueing models for three network elements of the router, switch and link. The queueing model for the router has two outgoing interfaces: an interface for SW1 and another for SW2. The number of outgoing interfaces for the switches are many and such a number depends on the number of ports for the switch. The switches and the router were modelled as *M/M/1* queues. Ethernet links are modeled as *M/D/1* queues. This is appropriate since the service time for Ethernet links is more of a deterministic than variable. However, the service times of the switches and the router are not deterministic since these are all CPU-based devices.

According to the datasheet found in Cisco systems and 3Com networking, the switches and the router used in Fig. 1 have a somewhat similar design of a store-and-forward buffer pool with a CPU responsible for pointer manipulation to switch or route a packet to different ports. (F. Gebali, 2005) provides comprehensive models of common types of switches and routers. According to (L. Kleinrock (vol. 1), 1975), the average delay for a VoIP packet passing through an *M/M/1* queue is basically $1/(\mu-\lambda)$, and through an *M/D/1* queue is $(1-\lambda/2\mu)/(\mu-\lambda)$, where λ is the mean packet arrival rate and μ is the mean network element service rate. The queueing models in Fig. 6 assume Poisson arrival for both VoIP and background traffic. In (M. Karam and F. Tobagi, 2002), it was concluded that modeling VoIP traffic as Poisson is adequate. However, in practice, background traffic is bursty in nature and characterized as self-similar with long range dependence (W. Leland, M. Taqqu, W Willinger and D. Wilson, 1994). For the analysis and design, using bursty background traffic is not practical. For one thing, under the network of queues being considered an analytical solution becomes intractable when considering non-Poisson arrival. Also, it is important to remember that in order to ensure good QoS at all times, analysis and design are based on the worst-case scenario of network load or utilization, i.e. the peak of aggregate bursts. And thus in a way our analytical approach takes into account the bursty nature of traffic.

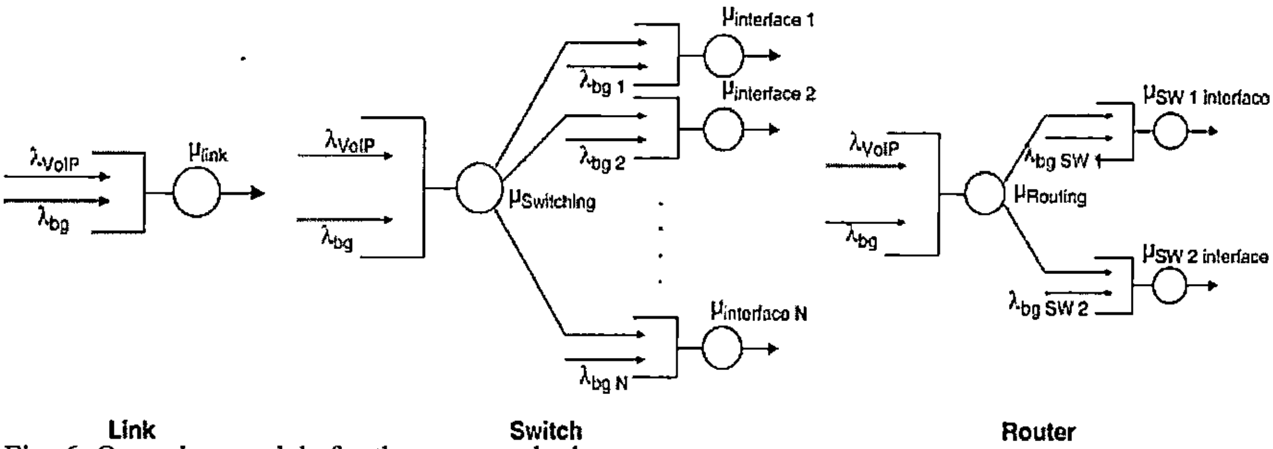


Fig. 6. Queueing models for three network elements.

It is worth noting that the analysis by decomposition of queueing networks in (K.M. Chandy and C.H. Sauer, 1978) assumes exponential service times for all network elements including links. But (R. Suri, 1983) proves that acceptable results with adequate accuracy can be still obtained if the homogeneity of service times of nodes in the queueing network is deviated (R. Suri, 1983) shows that the main system performance is insensitive to violations of the homogeneity of service times. Also, it was noted that when changing the models for links from *M/D/1* to *M/M/1*, a negligible difference was observed. More importantly, as will be demonstrated in this paper with simulation, our analysis gives a good approximation.

The total end-to-end network delay starts from the Ethernet outgoing link of the sender PC or IP phone to the incoming link of receiver PC or IP phone. To illustrate this further, let's compute the end-to-end delay encountered for a single call initiated from Floors 1 to 3. Fig. 7 shows an example of how to compute the network delay. Fig. 7a shows the path of a unidirectional voice traffic flow going from Floors 1 to 3. Fig. 7b shows the corresponding networking queueing model for such a path.

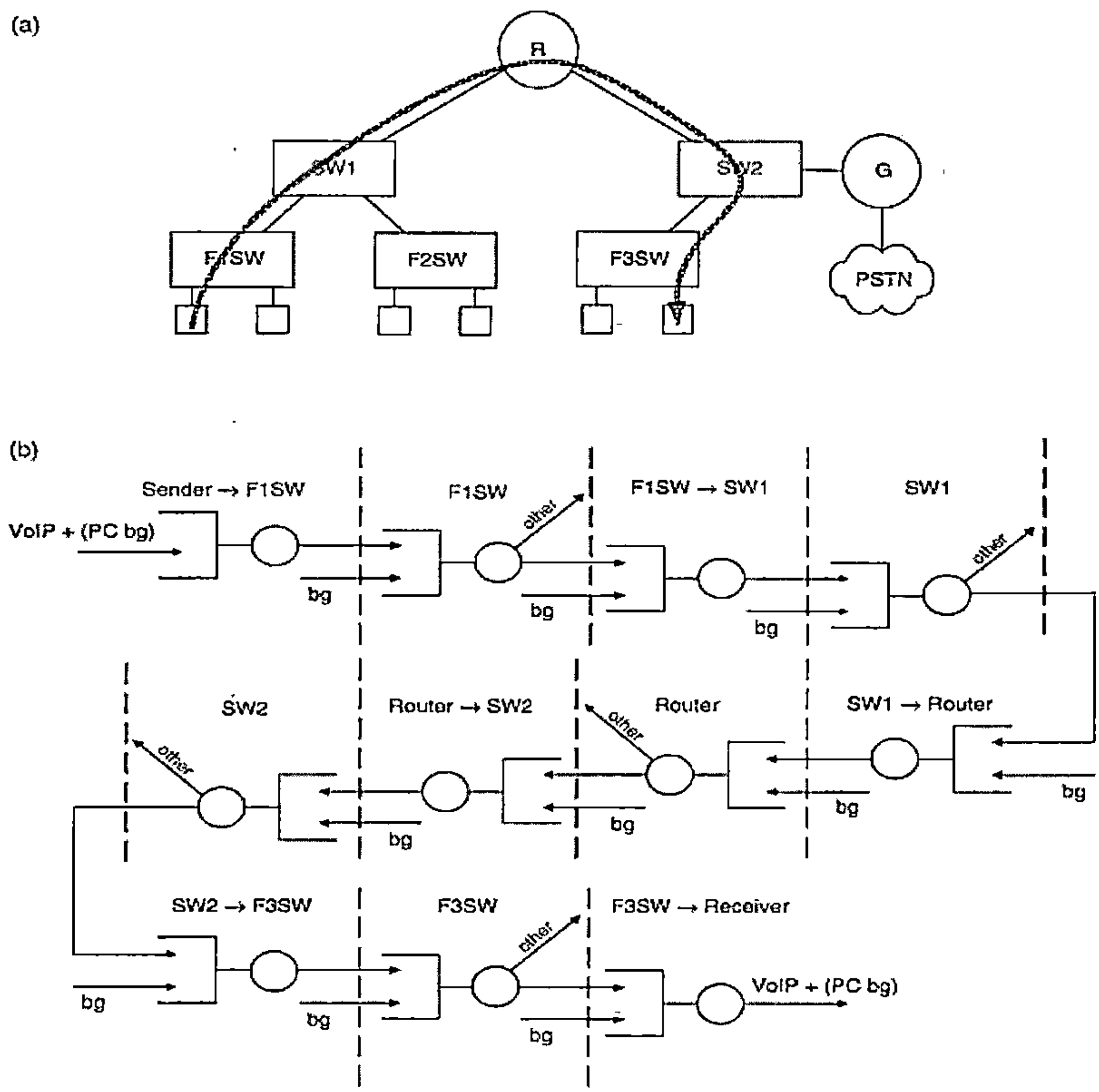


Fig. 7. Computing network delay. (a) Unidirectional voice traffic flow path from Floors 1 to 3. (b) Corresponding network queueing model of the entire path.

For Fig. 7b, in order to compute the end-to-end delay for a single bi-directional VoIP call, the delay must be computed for each network element. It is shown here how to compute the delay for the switches, links, and router. For the switch, whether it is that of intra-floor or inter-floor, $\mu = (1 - 25\%)1.3$ Mpps, where

25% is the growth factor. $\lambda = \lambda_{VoIP} + \lambda_{bg}$, where λ_{VoIP} is the total added new traffic from a single VoIP in pps, and λ_{bg} is the background traffic in pps. For an uplink or downlink, $\mu = (1 - 25\%)100\ \mathrm{Mbps}$, $\lambda = \lambda_{VoIP} + \lambda_{bg}$. Since the service rate is in bps, λ_{VoIP} and λ_{bg} must be expressed in bps.

Table 2 and Table 3 express the bandwidth for background traffic and for a single call in both pps and bps. Similarly for the router, $\mu = (1 - 25\%)25{,}000\ \mathrm{pps}$ and $\lambda = \lambda_{VoIP} + \lambda_{bg}$. Both λ_{VoIP} and λ_{bg} must be expressed in pps. Remember for a single bi-directional VoIP call, λ_{VoIP} at the router and switches for a single call will be equal to 100 pps. However, for the uplink and downlink links, it is 90.4 kbps. One should consider no λ_{bg} for the outgoing link if IP phones are used. For multimedia PCs which equipped with VoIP software, a λ_{bg} of 10% of the total background traffic is utilized in each floor. For a more accurate assessment of PC's λ_{bg}, actual measurement should be taken. For our case study of the small enterprise network, we use multimedia PCs.

The total delay for a single VoIP call of Fig. 7b, can be determined as follows:

$$\begin{aligned} D_{\mathrm{path}} = {} & D_{\mathrm{Sender\text{-}F1SW\ Link}} + D_{\mathrm{F1SW}} + D_{\mathrm{F1SW\text{-}SW1\ Link}} \\ & + D_{\mathrm{SW1}} + D_{\mathrm{SW1\text{-}Router\ Link}} + D_{\mathrm{Router}} \\ & + D_{\mathrm{Router\text{-}SW2\ Link}} + D_{\mathrm{SW2}} + D_{\mathrm{SW2\text{-}F3SW\ Link}} \\ & + D_{\mathrm{F3SW}} + D_{\mathrm{F3SW\text{-}Receiver\ Link}} \end{aligned}$$

Network capacity algorithm

In order to determine the maximum number of calls that can be supported by an existing network while maintaining VoIP delay constraint, the following algorithm was developed that basically determines network capacity in terms of VoIP calls. Calls are added iteratively until the worst-case network delay of 80 ms has reached. The algorithm can be described in the following steps:

(i) Initially, no calls are introduced and the only traffic in the network is the background traffic.

(ii) A new call is added, according to the call distribution described in Probability tree.

(iii) For each network element, $\lambda=\lambda_{VoIP}+\lambda_{bg}$ is computed. λ_{bg} is known for each element; however, λ_{VoIP} can get affected by introducing a new call depending on the call traffic flow, i.e. whether or not the new call flow passes through the network element.

(iv) For each network element, the average delay of a VoIP packet is computed.

(v) The end-to-end delay is computed by summing up all the delays of step (iv) encountered for each possible VoIP flow. This includes all external and internal flows, with internal flows consisting of intra-floor and inter-floor.

(vi) The maximum network delay of all possible flows is determined. If the maximum network delay is less than 80 ms, then the maximum number of calls has not been reached. Therefore, a new call can be added, and hence go to step (ii).

(vii) If not, the maximum delay has been reached. Therefore the number of VoIP calls bounded by the delay is one less than the last call addition.

The above algorithm was implemented using MATLAB and the results for the worst incurred delay are plotted in Fig. 8. It can be observed from the figure that the delay increases sharply when the number of calls goes beyond 310 calls. To be more precise, MATLAB results showed the number of calls that are bounded by the 80 ms delay is 316.

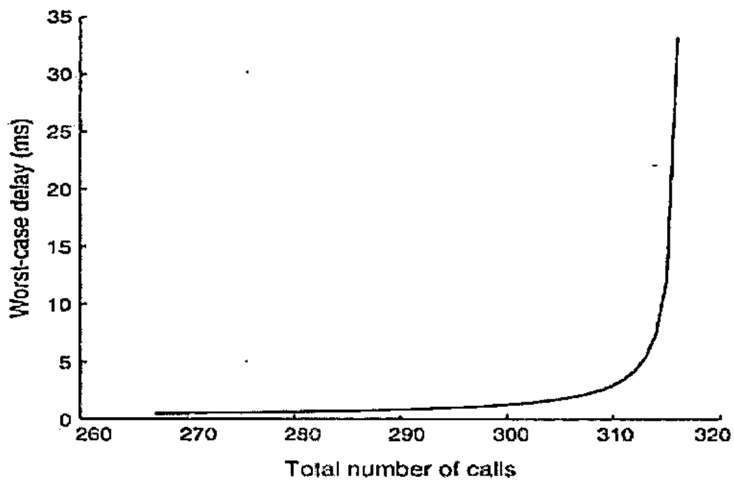


Fig. 8. Worst-case delay vs. number of VoIP calls.

When comparing the number of calls that network can sustain based on bottleneck bandwidth and worst-delay analysis, the number of calls is limited by the available bandwidth more than the delay, though the difference is small. Therefore, we can conclude that the maximum number of calls that can be sustained by the existing network is 313.

Packet loss

A question related to determining the number of calls to be supported by a particular data network is packet loss. VoIP packet loss should be below 1% according to (J.H James, B. Chen and L. Garrison, 2044), and hence packet loss can be a third constraint that plays a key role in determining the number of calls to be supported by a network. In this case, finite queueing systems of *M/M/1/B* and *M/D/1/B*, as opposed to *M/M/1* and *M/D/1*, must be used instead. In a finite queueing system, due to dropping of packets, the flow of one node will affect the flow of another because we have bidirectional flows. Consequently, it ends up with a model of somewhat closed queueing networks with blocking (R. Onvural, 1990). Determining packet loss for this type of network is not a trivial task, and can be only approximated, according to (R. Onvural, 1990; J. Bolot, 1993). Approximation algorithms found in literature for solving closed networking queueing systems are not accurate and does not have a closed form solution. The solution is typically heuristic and it takes a long time to converge (R. Onvural, 1990). Due to lack of closed-form analytical solutions and according to(J. Bolot, 1993), simulation is a more practical approach to study packet loss. In the work presented in this paper, the simulation is use to verify that the packet loss constraint is satisfied with no packet loss.

Chapter 4

Simulation

Over recent years, there has been tremendous growth in wireless communications. This growth also encompasses personal and business computing. The original IEEE 802.11 (B. Karacali, M. Bearden, L. Denby, J. Meloche and D.T. Stott, 2002; B. Karacali, L. Denby and J. Melche, 2004), standard was basically built to support data applications over contention-based access control protocol. As the use of multimedia applications increased it became obvious that WLANs support real-time applications with quality of service (QoS) guarantees the same as their wired counterpart. Recently, many researchers have investigated this issue and proposed several mechanisms to tackle this problem. The focus was on developing adaptive schemes working on top of the existing distributed access control. The most important motivation for this approach is that the widely used wireless adapters are mainly supporting the distributed scheme.

Then, by using simple software, the access control scheme can be adapted to the needs of the network. Further, it was stated in (B Goode, 2002) that distributed medium access control (MAC) with QoS is more flexible and effective than the centralized MAC, as the dominant operational mode in IEEE 802.11 LANs is the distributed coordination function (DCF) mode (P. Mehta and S. Udani, 2001). Also recent research shows that point coordination function (PCF) performs poorly either alone or incorporated with DCF mode. The contribution of this work is focused on analyzing the performance of Internet applications besides video conferencing using the prioritized adapters suggested in (B Goode, 2002). To the best of our knowledge, this work is the first in analyzing and discussing the influence of these adapters on Internet applications. Previous works were general in nature, assuming either general prioritized flows of packets as in (B Goode, 2002) or real-time and non-real time applications without considering the unique traffic characteristics of www or e-mail applications as in (W. Jiang and H. Schulzrinne, 2001).

The object of the simulation is to verify analysis results of supporting VoIP calls. The popular MIL3's OPNET Modeler simulation package,[1] Release 8.0.C was used for simulation. OPNET Modeler contains a vast amount of models of commercially available network elements, and has various real-life network configuration capabilities. This makes the simulation of real-life network environment close to reality. Other features of OPNET include GUI interface, comprehensive library of network protocols and models, source code for all models, graphical results and statistics, etc. More importantly, OPNET has gained considerable popularity in academia as it is being offered free of charge to academic institutions. That has given OPNET an edge over DES NS2 in both market place and academia. This section gives a brief description of the simulation model, configurations, and results.

Modeling the network

A snapshot of the OPNET simulation model for the existing network under study is shown in Fig. 9. The simulation model of the organization network, for the most part, is an exact replica of the real network. In OPNET Modeler, many vendor-specific models are included in the pre-defined component libraries. VoIP *gateway* is modeled as an Ethernet workstation; and the enterprise servers are modeled as Ethernet servers. All network elements have been connected using 100 Base-T links. Fig. 9 shows the described topology. As discussed in the previous section, the *gatekeeper* signaling traffic is ignored and hence modeling such and element and its traffic is not taken into account as we base our study on the worst case situation.

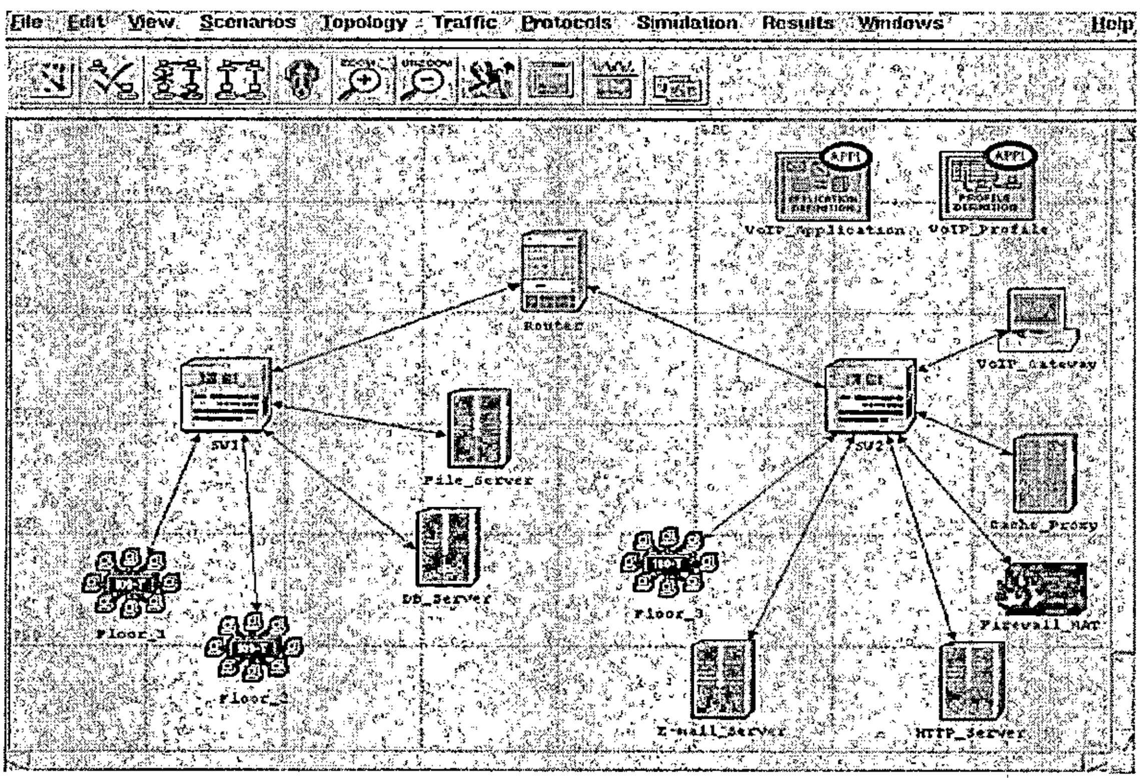


Fig. 9. OPNET simulation model of the organization network.

Floor LANs have been modeled as subnets that enclose an Ethernet switch and three Ethernet workstations used to model the traffic of the LAN users, as shown in Fig. 10. One of these workstations generates the background traffic of the floor while the other two act as parties in VoIP sessions. For example, the Ethernet workstations for Floor 1 are labeled as F1_C1, F1_C2, and F1_C3. F1_C1 is the source for sending VoIP calls. F1_C2 is the sink for receiving VoIP calls. F1_C3 is the sink and source of background traffic.

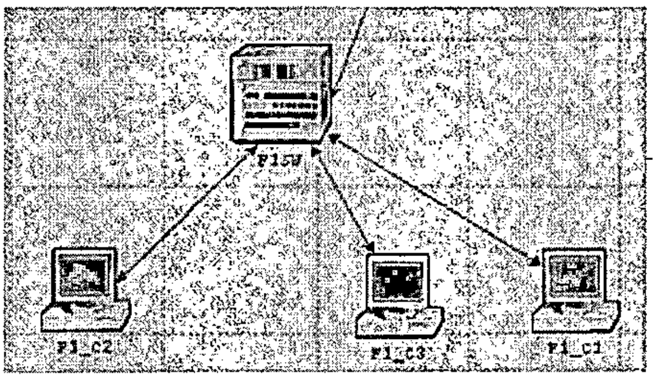


Fig. 10. Floor 1 subnet model.

Various OPNET Modeler configurations were made which included the network VLANs, router, switches, and links. Also, background traffic was incorporated into the network as well as the generation of VoIP traffic. For VoIP traffic generation, a VoIP application and a profile have to be created. OPNET Modeler has a predefined voice application. The VoIP traffic was generated and received by workstations within the floors. The VoIP traffic was generated according to the flow and call distribution discussed in the previous section. We set up OPNET Modeler such that three new VoIP calls are generated every two seconds.

Chapter 5

Simulation results

In this section, the most relevant graphed results for the VoIP traffic volume and delay are reported. The duration of the OPNET simulation was configured to run for 8 min. The generation of background traffic, by default in OPNET, started at 40 s from the start time of the simulation run. The VoIP traffic started at 70 s at which a total of 3 VoIP bi-directional calls are initially added. Then, every 2 s 3 VoIP calls are added. The Simulation stops at 8 min in which a total of $3+((7\times60+58-70)/2)\times3=615$ calls got generated. This should translate into a total of 61,500 packets being generated every second. Note that since the simulation stops at 8 min, the last three calls to be added were at 7 min and 58 s.

Fig. 11 shows the VoIP traffic and the corresponding end-to-end delay as VoIP calls are added every two seconds. Fig. 11a shows the total VoIP traffic that was sent, received, and dropped. Fig. 11b is a zoom-in version of Fig. 11a, focusing on the mismatch region between traffic sent and received. From Fig. 11a, it is clear that the total VoIP traffic generated by the end of simulation run is very close 61,500 pps. In fact, simulation results gave 61,429 pps.

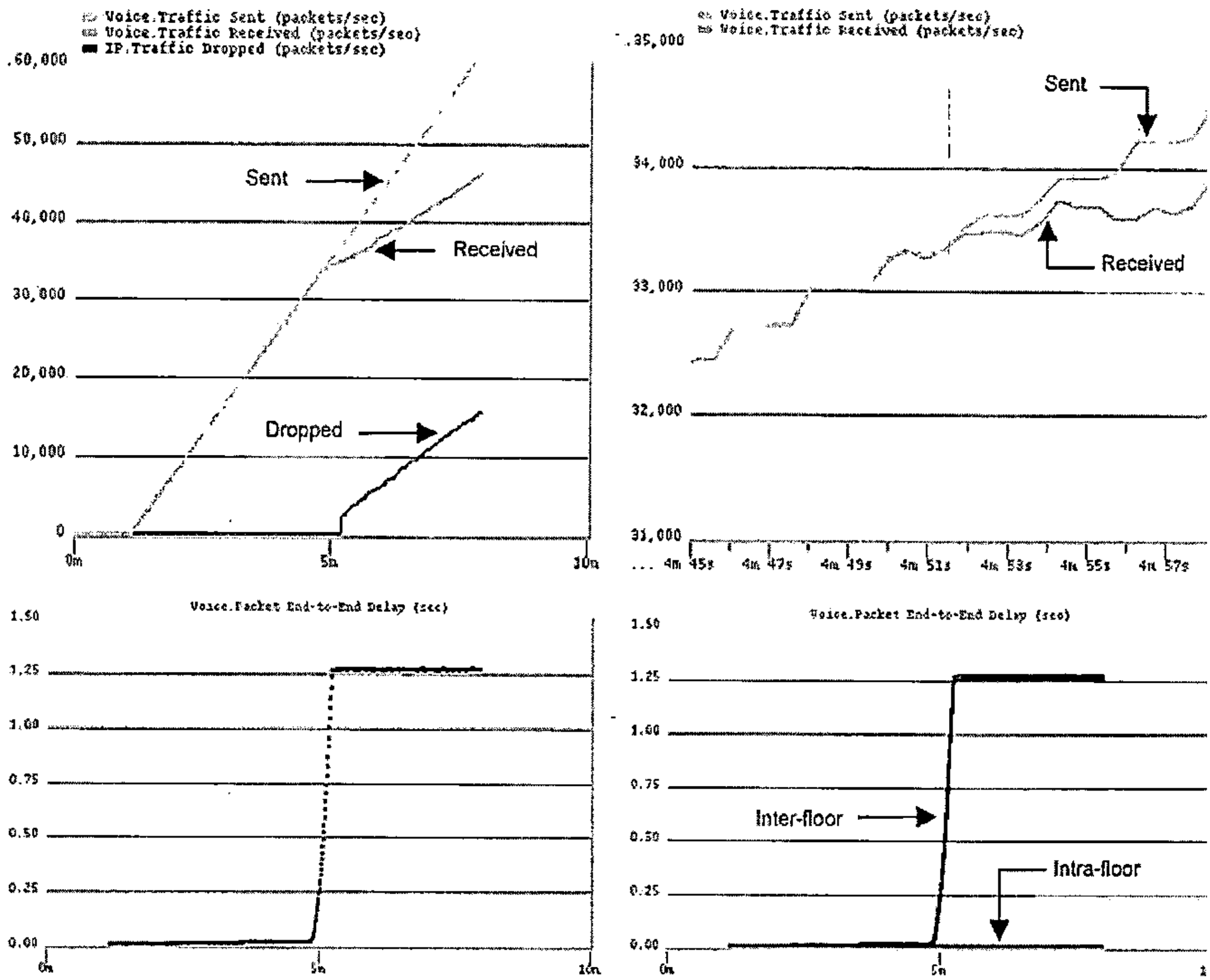


Fig. 11. VoIP traffic and delay.

One can determine the total number of calls that the network can sustain by examining network bandwidth or delay bounds; first investigate the bandwidth bound. Fig. 11a and b show clearly that not all of VoIP packets being sent get received. i.e. there is a mismatch between traffic sent and received. Fig. 11b captures clearly the addition of the three calls every 2 s; and how this addition is repeated in gradual steps of 300 pps. Examining the X-axis of the simulation run time, it is clear that the last successful addition of three calls was at exactly 4 min and 48 s, as seen clearly in Fig. 11b. The next addition, as shown, was at 4 min and 50 s which resulted in a mismatch. For the last successful addition of voice calls, which occurred at 4 min and 48 s, we had a traffic volume (see *Y*-axis) of exactly 33,000 pps or 330 VoIP calls. Also one can arrive at the same number of calls by calculating how many calls have been added until the last successful addition of three calls, i.e. 4 min and 48 s. This yields to $3+((4\times60+48-70)/2)\times3=330$ calls.

Fig. 11c shows the corresponding VoIP end-to-end delay. Remember this delay should not exceed 80 ms, as discussed in previous sections. As depicted, the delay stays less than 80 ms until a simulation time of 4 min and 54 s at which the delay increases sharply. One can then find out the number of VoIP calls that the network can support to satisfy the 80 ms time constraint. The number of calls can be computed as $3+((4\times60+54-70)/2)\times3=339$ calls. Therefore, one can conclude that, based on these simulation results, the number of voice calls to be supported by the network is bounded more by the network bandwidth than the delay. Hence, the number of the VoIP calls that the network can support based on simulation is 330 calls.

The simulation's reported delays shown in Fig. 11c is the maximum values of a bucket of 100 collected values. The OPNET default reported delay configuration is the sample mean of a bucket of 100 collected values. Fig. 11d depicts a different collection mode, in which 'all values' are collected and plotted. Fig. 11d depicts two types of delays. First, the delays of external and inter-floor VoIP packets passing through the router. These are the bigger delays and they resemble the delays of Fig. 11c. Second, the delays of intra-floor VoIP packets that are not passing through the router. These are the smaller delays, in which the majority of these values stay close to 2.5 ms.

When comparing the VoIP end-to-end delay in Fig. 11c with the delay obtained by analysis in Fig. 9, it is shown that the delay in Fig. 11c does not shoot to infinity as that of Fig. 8. In Fig. 11c, the delay stays flat at about 1.25 s. This is so because in the analysis that was modeled the network elements with infinite buffer. Another observation can be made about the dropped VoIP packets in Fig. 11a. It is seen that the dropping of VoIP packets occurs after the mismatch of the sent and received packets. This is due to the fact that CPU processing, especially of the router, gets 100% saturated before the memory buffer of the router gets filled up. It was observed that the memory buffer of the router gets completely full 25 s after the router's CPU utilization reaches 100%.

Simulation accuracy

In order to gain accuracy (with a narrow confidence interval) of our simulation results, following the popular guidelines presented in (A. Law and W. Kelton, 1991; K. Pawlikowski, H. Jeong and J. Lee, 2002), five simulation replications were run by feeding OPNET with different initial seeds. OPNET's pseudo random number generator is based on BSD's algorithm which permits safely, i.e. with no concern of overlapping of random number streams, any integer value to be an initial seed. Five simulation replications were sufficient. Each simulation replication produced very similar graphical results, which when interpreted as done in previous sections, led to the same total number of calls to be supported.

Final simulation run

Based on the simulation results, the existing network can support 330 VoIP calls. In the simulation, the voice calls were added every 2 s and the simulation was not allowed to stabilize for a long-time. Our attention was focused on finding out the number of voice calls that the network can sustain. As a final check to ensure a healthy network and a normal behavior for all network elements, we perform a final simulation run in which 330 voice calls are added, all at once at the start of the simulation, say after 70 s. We let the simulation run execute for a prolonged amount of time, say a good 5 min, to reach a steady state. Then, examine the health of each network element. In the example, this final simulation of 330 voice calls was not successful. The simulation run showed a mismatch between traffic sent and received and a delay of more than 80 ms. However, a successful simulation run of 306 voice calls showed normal and healthy results with no packet loss, average delay of 2.15 ms, and adequate utilization of router and switch CPUs and links. Therefore we can conclude, based on OPNET simulation, that the network can support a total of 306 voice calls.

Pilot deployment

Before embarking on changing any of the network equipment, it is always recommended to build a pilot deployment of VoIP in a test lab to ensure smooth upgrade and transition with minimum disruption of network services. A pilot deployment comes after training of IT staff. A pilot deployment is the place for the network engineers, support and maintenance team to get firsthand experience with VoIP systems and their behavior. During the pilot deployment, the new VoIP devices and equipment are evaluated, configured, tuned, tested, managed, monitored, etc. The whole team needs to get comfortable with how VoIP works, how it mixes with other traffic, how to diagnose and troubleshoot potential problems. Simple VoIP calls can be set up and some benchmark testing can be performed.

Design and engineering decisions

The following network design and engineering decisions can be justified from the analytic and simulation approaches:

1) The existing network, with a reserved growth factor of 25%, can safely support up to 306 calls while meeting the VoIP QoS requirements and having no negative impact on the performance of existing network services or applications.
2) For 306 calls, a network delay of about 2 ms is encountered. To be precise, analysis gave a delay of 1.50 ms, while simulation gave a delay of 2.15 ms.
3) A safety growth factor of 25% is maintained across all network resources.

4) The primary bottleneck of the network is the router. If the enterprise under study is expected to grow in the near future, i.e., more calls are required than 306 calls, the router replacement is a must. The router can be replaced with a popular Layer-3 Ethernet switch, and thus relieving the router from routing inter-floor calls from Floors 1 to 2. Before prematurely changing other network components, one has to find out how many VoIP calls can be sustained by replacing the router. To accomplish this, the design steps and guidelines outlined in this paper must be revisited and re-executed.
5) The network capacity to support VoIP is bounded more by the network throughput than the delay. This is due to the fact the existing network under study is small and does not have a large number of intermediate nodes. The network delay bound can become dominant if we have a large-scale LAN or WAN.

Conclusion

This paper presents a New TCP scheme that attempts to effectively distinguish transmission losses from packet losses on various networks including wireless links. The paper described in great detail how OPNET can be utilized to assess the readiness of existing TCP/IP networks to support desktop videoconferencing. The paper offered extensive interpretations and analysis of simulation results and showed how to draw proper conclusions.

The paper outlined a step-by-step methodology on how VoIP can be deployed successfully. The methodology can help network researchers and designers to determine quickly and easily how well VoIP will perform on a network prior to deployment. Prior to the purchase and deployment of VoIP equipment, it is possible to predict the number of VoIP calls that can be sustained by the network while satisfying QoS requirements of all existing and new network services and leaving enough capacity for future growth. The work presented in this paper can be adopted *easily* for larger and general networks by following the same principles, guidelines, and concepts laid out in this paper. In addition, the paper discussed many design and engineering issues pertaining to the deployment of VoIP. These issues include characteristics of VoIP traffic and QoS requirements, VoIP flow and call distribution, defining future growth capacity, and measurement and impact of background traffic.

It was considered a case study of deploying VoIP in a small enterprise network. The methodology and guidelines outlined in this paper are applied on such a network. Both analysis and simulation were utilized to determine the number of VoIP calls that can be supported for such a network. From results of analysis and simulation, it is apparent that both results are in line and give a close match. Based on the analytic approach, a total of 313 calls can be supported. Based on the simulation approach, a total of 306 calls can be supported. There is only a difference of seven calls. The difference can be contributed to the degree of accuracy between the analytic approach and OPNET simulation. The analytic approach is just an approximation. Also, the difference is linked to the way the OPNET Modeler adds the distribution of the calls. It was found that external and inter-floor calls are added before intra-floor calls. In anyways, to be safe and conservative, one can consider the minimum number of calls of the two approaches.

In this paper, only peer-to-peer voice calls were considered. As a future work, one can consider implementing important VoIP options such as VoIP conferencing and messaging. Also as a future work, one can look into assessing the network support and readiness of deploying other popular real-time network services such as multimedia, video, and web conferencing.

References:

Andrew Anderson, Introduction to TCP/IP-Networks 1996, http://tldp.org/LDP/nag/node8.html

Annex to Recommendation G.729, Coding of Speech at 8 kbit/s using Conjugate Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP), Annex A: Reduced Complexity 8 kbit/s CS-ACELP Speech Codec, ITU, November 1996.

Avaya Inc., Avaya IP voice quality network requirements, http://www1.avaya.com/enterprise/whitepapers, 2001.

A. Law and W. Kelton, *Simulation Modeling and Analysis* (2nd ed), McGraw-Hill, New York (1991).

A. Markopoulou, F. Tobagi and M. Karam, Assessing the quality of voice communications over internet backbones, *IEEE/ACM Transaction on Networking* **11** (2003) (5), pp. 747–760.

A. Takahasi and H. Yoshino, Perceptual QoS Assessment Technologies for VoIP, *IEEE Communications Magazine* **42** (2004) (7), pp. 28–34.

Bakshi, B.; Krishna, P.; Vaidya, N. and Pradhan, D. (1997). *Improving Performance of TCP over Wireless Networks*. In Proceedings of the IEEE 17th ICDCS'97, pages 365-373.

Bakshi, Bikram S.; Krishna, P.; Vaidya, N. H. and Pradhan, D. K. (1996). *Improving Performance of TCP over Wireless Net works.* Technical Report 96-014, Texas A & M University,.

Barroso et al., 2003 L.A. Barroso, J. Dean and U. Hölzle, Web search for a planet: the Google cluster architecture, *IEEE Micro* 23 (2003) (2), pp. 22–28.

Burkowski, 1990 F.J. Burkowski, Retrieval performance of a distributed database utilising a parallel process document server, *Proceedings of the second international symposium on databases in parallel and distributed systems*, ACM Press, New York (1990), pp. 71–79.

B. Duysburgh, S. Vanhastel, B. DeVreese, C. Petrisor and P. Demeester, *On the influence of best-effort network conditions on the perceived speech quality of VoIP connections, Proceedings of IEEE 10th International Conference of Computer Communications and Networks, Scottsdale, AZ* (2001) pp. 334–339.

B Goode, Voice over Internet Protocol (VoIP), *Proceedings of the IEEE* **90** (2002) (9), pp. 1495–1517.

B. Karacali, M. Bearden, L. Denby, J. Meloche and D.T. Stott, *Assessing network readiness for IP telephony, Proceedings of IEEE International Conference on Communications, ICC02* **vol. 4** (2002) pp. 2568–2572.

B. Karacali, L. Denby and J. Melche, *Scalable network assessment for IP telephony, Proceedings of IEEE International Conference on Communication, ICC04, Paris* (2004) pp. 1505–1511.

Cacheda, Carneiro, et al., 2005 F. Cacheda, V. Carneiro, V. Plachouras and I. Ounis, Performance network analysis for distributed information retrieval architectures, *Proceedings of 27th European conference on information retrieval* *research (ECIR'05), Lecture notes on computer science* Vol. 3408, Springer-Verlag, Berlin, Heidelberg (2005), pp. 527–529.

Cacheda et al., 2004 F. Cacheda, V. Plachouras and I. Ounis, Performance analysis of distributed architectures to index one terabyte of text, *Proceedings of 26th European conference on information retrieval research (ECIR'04), Lecture notes on computer science* Vol. 2997, Springer-Verlag, Berlin Heidelberg (2004), pp. 394–408.

Cacheda, Plachouras, et al., 2005 F. Cacheda, V. Plachouras and I. Ounis, A case study of distributed information retrieval architectures to index one terabyte of text, *Information Processing and Management Journal* 41 (2005) (5), pp. 1141–1161.

Cacheda and Viña, 2001 F. Cacheda and A. Viña, Experiences retrieving information in the World Wide Web, *Proceedings of the 6th IEEE symposium on computers and communications*, IEEE Computer Society (2001), pp. 72–79.

Cahoon and McKinley, 1996 B. Cahoon and K.S. McKinley, Performance evaluation of a distributed architecture for information retrieval, *Proceedings of 19th ACM-SIGIR international conference on research and development in information retrieval*, ACM Press, New York (1996), pp. 110–118.

CAIDA,, April 2004.

Charles L. Hedrick, Rutgers University, Introduction to the Internet Protocols 1987, http://www.linuxjunkies.org/network/tcpip/intro2.html

Charles M. Kozierok, The TCP/IP Guide 2005, http://www.tcpipguide.com/free/t_TCPIPOverviewandHistory.htm

Cisco Systems Inc., Cisco 2621 Modular Access Router Security Policy, 2001,

Clarke et al., 2004 Clarke, C., Craswell, N., & Soboroff, I. (2004). Overview of the TREC 2004 Terabyte track. In E. M. Voorhees, & L. P. Buckland (Ed.), *Proceedings of the thirteenth text retrieval conference*. Gaithersburg, MD: NIST Special Publication 500-261.

Coevreur et al., 1994 T.R. Coevreur, R.N. Benzel, S.F. Miller, D.N. Zeitler, D.L. Lee and M. Singhal *et al.*, An analysis of performance and cost factors in searching large text databases using parallel search systems, *Journal of the American Society for Information Science* 45 (1994) (7), pp. 443–464.

Craswell and Hawking, 2002 Craswell, N., & Hawking, D. (2002). Overview of the TREC-2002 Web track. In E. M. Voorhees, & L. P. Buckland (Ed.), *Proceedings of the eleventh text retrieval conference*. Gaithersburg, MD: NIST Special Publication 500-251.

F. Gebali, *Computing Communication Networks: Analysis and Designs* (3rd ed), Northstar Digital Design, Inc (2005).

Gibbons and Chakraborti, 2003 J.D. Gibbons and S. Chakraborti, Nonparametric statistical inference (4th ed.), Marcel Dekker Ltd, New York (2003).

Heaps, 1978 H.S. Heaps, Information retrieval: Computational and theoretical aspects, Academic Press, New York (1978).

Hawking, 1997 D. Hawking, Scalable text retrieval for large digital libraries, *Lecture notes in computer science* 1324 (1997), pp. 127–146.

Hawking and Thistlewaite, 1999 D. Hawking and P. Thistlewaite, Methods for information server selection, *ACM Transactions on Information Systems* 17 (1999) (1), pp. 40–76.

H. Gilbert Introduction to TCP/IP PCLT 1995, http://www.yale.edu/pclt/COMM/TCPIP.HTM

ITU-T Recommendation P.800, Methods for subjective determination of transmission quality, www.itu.in/publications/main_publ/itut.html.

Jeong and Omiecinski, 1995 B. Jeong and E. Omiecinski, Inverted file partitioning schemes in multiple disk systems, *IEEE Transactions on Parallel and Distributed Systems* 6 (1995) (2), pp. 142–153.

Jones et al., 2002 C.B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester and M. van Kreveld *et.al.*, Spatial information retrieval and geographical ontologies an overview of the SPIRIT project, *Proceedings of the 25th ACM-SIGIR conference on research and development in information retrieval*, ACM Press, New York (2002), pp. 387–388.

J. Bolot, *End-to-End Packet Delay and Loss Behavior in the Internet, Proceedings of ACM Conference on Communications, Architectures, Protocols and Applications* (1993) pp. 289–298.

J.H James, B. Chen and L. Garrison, Implementing VoIP: a voice transmission performance progress report, *IEEE Communications Magazine* **42** (2004) (7), pp. 36–41.

J.S. Han, S.J. Ahn and J.W. Chung, Study of delay patterns of weighted voice traffic of end-to-end users on the VoIP network, *International Journal of Network Management* **12** (2002) (5), pp. 271–280.

J. Walker, J. Hicks, Planning for VoIP, NetIQ Corporation white paper, December 2002, http://www.telnetnetworks.ca/products/netIq/whitepapers/planning_for_voip.pdf.

Kenneth Brown, VOIP – Embracing Invention, January 2004, http://www.adti.net/telecom/voipembracinginvention.htm

K.M. Chandy and C.H. Sauer, Approximate methods for analyzing queueing network models of computing systems, *Journal of ACM Computing Surveys* **10** (1978) (3), pp. 281–317.

K. Pawlikowski, H. Jeong and J. Lee, On credibility of simulation studies of telecommunication networks, *IEEE Communications Magazine* **40** (2002) (1), pp. 132–139.

K. Salah and A. Alkhoraidly, An OPNET-based simulation approach for deploying VoIP, 2006.

Lin and Zhou, 1993 Z. Lin and S. Zhou, Parallelizing I/O intensive applications for a workstation cluster: a case study, *ACM SIGARCH Computer Architecture News* 21 (1993) (5), pp. 15–22.

Little, 2001 Little, M.C. (2001). JavaSim user's guide. Public Release 0.3, Version 1.0. University of Newcastle upon Tyne. Retrieved 1 June, 2003, from the World Wide Web: http://javasim.ncl.ac.uk/manual/javasim.pdf.

Lu and McKinley, 2000 Z. Lu and K. McKinley, Partial collection replication versus caching for information retrieval systems, *Proceedings of the 25th ACM-SIGIR conference on research and development in information retrieval*, ACM Press, New York (2000), pp. 248–255.

L. Kleinrock (vol. 1), *Queueing Systems: Theory*, Wiley, New York (1975).

L. Sun and E.C. Ifeachor, Prediction of Perceived Conversational Speech Quatlity and Effects of Playout Buffer Algorithms, Proceedings of International Conference on Communications, ICC'03, Anchorage (2003) pp. 1–6.

Mark Sportack, The ABCs of TCP/IP 2005, http://www.ciscopress.com/articles/article.asp?p=377101&seqNum=1

Martin et al., 1986 P. Martin, I.A. Macleod and B. Nordin, A design of a distributed full text retrieval systems, Proceedings of the 9th ACM-SIGIR conference on research and development in information retrieval, ACM Press, New York (1986), pp. 131–137.

M. Karam and F. Tobagi, Analysis of delay and delay jitter of voice traffic in the Internet, *Computer Networks Magazine* **40** (2002) (6), pp. 711–726.

OPNET Technologies, http://www.mil3.com.

Parsa, C. and Aceves, G. Improving TCP Congestion Control over Internets with Heterogeneous Transmission Media. IEEE International Conference on Network Protocols (ICNP '99), 1999.

Plachouras et al., 2004 Plachouras, V., He, B., & Ounis, I. (2004). University of Glasgow at TREC2004: experiments in Web, Robust and Terabyte tracks with

Terrier. In E. M. Voorhees, & L. P. Buckland (Ed.), *Proceedings of the thirteenth text retrieval conference*. Gaithersburg, MD: NIST Special Publication 500-261.

Postel, 1981 Postel, J. (1981). *Internet protocol*. Request for Comment 791, Internet Engineering Task Force.

P. Mehta and S. Udani, Voice over IP, *IEEE Potentials Magazine* **20** (2001) (4), pp. 36–40.

Recommendation H.323, Packet-based Multimedia Communication Systems, ITU, 1997.

Recommendation G.114, One-Way Transmission Time, ITU, 1996.

Recommendation G.711, Pulse Code Modulation (PCM) of Voice Frequencies, ITU, November 1988.

Recommendation G.723.1, Speech Coders: Dual Rate Speech Coder for Multimedia Communication Transmitting at 5.3 and 6.3 kbit/s, ITU, March 1996.

Recommendation G.726, 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM), ITU, December 1990.

Ribeiro-Neto and Barbosa, 1998 B. Ribeiro-Neto and R. Barbosa, Query performance for tightly coupled distributed digital libraries, *Proceedings of the 3rd ACM conference on digital libraries*, ACM Press, New York (1998), pp. 182–190.

R. Onvural, Survey of closed queueing networks with blocking, *ACM Computing Surveys* **22** (1990) (2), pp. 83–121.

R. Prasad, C. Dovrolis,, M. Murray and K.C. Claffy, Bandwidth estimation: metrics, measurement techniques, and tools, *IEEE Network Magazine* **17** (2003) (6), pp. 27–35.

R. Suri, Robustness of queueing network formulas, *Journal of the ACM* **30** (1983) (3), pp. 564–594.

Spurgeon, 2000 C.E. Spurgeon, Wireless LAN: The definitive guide, O'Reilly & Associates, Sebastopol (2000).

Steve Valentino, VoIP Software, May. 2006, http://ezinearticles.com/?VoIP-Software&id=210099

S. Riley and R. Breyer, *Switched, Fast, and Gigabit Ethernet* (3rd ed), Macmillan Technical Publishing (2000).

TCP/IP Fundamentals for Microsoft Windows 2004, http://technet.microsoft.com/en-us/library/bb726991.aspx

Tmg-emedia, inc. 1999-2007, http://www.historyoftheinternet.com/chap4.html

Tomasic and Garcia-Molina, 1993 A. Tomasic and H. Garcia-Molina, Performance of inverted indices in shared-nothing distributed text document information retrieval systems, *Proceedings of the 2nd international conference on parallel and distributed information systems*, IEEE Computer Society, San Diego, CA (1993), pp. 8–1

W. Jiang and H. Schulzrinne, *Assessment of VoIP Service Availability in the Current Internet, Proceedings of International Workshop on Passive and Active Measurement, PAM2003, San Diago, CA* (2003).

W. Jiang and H. Schulzrinne, *Comparison and Optimization of Packet Loss Repair Methods on VoIP Perceived Quality under Bursty Loss, Proceedings of ACM 12th International Workshop on Network and Operating System Support for Digital Audio and Video, Miami, FL* (2002) pp. 73–81.

W. Jiang, K. Koguchi and H. Schulzrinne, *QoS Evaluation of VoIP End-Points, Proceedings of IEEE International Conference on Communications, ICC'03, Anchorage* (2003) pp. 1917–1921.

W. Jiang and H. Schulzrinne, *Towards Junking the PBX: deploying IP Telephony, Proceedings of ACM 11th International Workshop on Network and Operating System Support for Digital Audio and Video, Port Jefferson, NY* (2001) pp. 177–185.

W. Leland, M. Taqqu, W Willinger and D. Wilson, On the self-similar nature of ethernet traffic, *IEEE/ACM Transaction on Networking* **2** (1994) (1), pp. 1–15.

Xylomenos, G. and Polyzos, G. TCP and UDP Performance over a Wireless LAN. In Proceedings of the IEEE INFO COM, 1999.

3Com Networking Product Guide, April 2004.