ABSTRACT

RAJAGOPAL, NISHA. Modeling and Performance Prediction of IP Multimedia Subsystem Networks. (Under the direction of Professor Michael Devetsikiotis).

IP Multimedia Subsystem (IMS) is envisioned as the solution for the next generation multimedia rich communication. Based on an open IP infrastructure, IMS enables access independent convergence of data, speech, video and mobile network technologies. Session Initiation Protocol (SIP) is the signaling protocol of choice for IMS. In this thesis, we propose service models for IMS networks with utility functions. These service models act as control mechanisms to optimize network properties.

We analyze the IMS network based on the SIP signaling delay and predict performance trends of the network. Our focus is on the formulation of queueing models for the IMS network and characterization of the SIP server workload. This approach of theoretical evaluation combined with realistic performance characterization can be used for designing IMS networks with optimal performance. Our analysis is based on a careful study of real-life SIP network traffic.

Modeling and Performance Prediction of IP Multimedia Subsystem Networks

by

NISHA RAJAGOPAL

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Computer Science

Raleigh

2006

Approved By:

Dr. Michael Devetsikiotis
Chair of Advisory Committee

Dr. Yannis Viniotis

Dr. Khaled Harfoush

To my loving family…

Biography

Nisha Rajagopal was born in Calicut, India on June 5th, 1981. She holds a Bachelors degree in Computer Science, awarded in June 2003, from Vishveswariya Technological University, India. Nisha joined North Carolina State University in August 2003 to pursue her Master of Science in Computer Networking. Her research is in the area of performance evaluation of IP Multimedia Subsystem.

Acknowledgements

First and foremost, I thank my advisor Dr. Micheal Devetsikiotis. His constant moral and intellectual support with never ending patience to my countless questions no matter how silly, have helped me in many ways throughout my Masters. Without our Saturday morning meetings, my research would never have progressed. I express my sincere thanks to Dr. Yannis Viniotis for his constructive comments and guidance. My research would not have been in IP Multimedia Subsystem without his timely help. I thank Dr. Khaled Harfoush for his insightful comments.

I would like to express my eternal gratitude to my parents for their love and affection. Their support and absolute faith in my abilities have been a huge motivation. Without them, my Masters would not have been worth the effort. I would like to thank Indira and Prasanna Dhore for being my family away from home.

I am thankful to Ruben for being a part of my life. He has been my pillar of strength during testing times. I thank all my friends for listening to my frustrations and half-baked theories. Their presence has made my life more meaningful. I thank all my colleagues at Ericsson IPI for their good wishes and encouragement. A special thanks to Nandagopal Ancha for taking time to read my thesis. Finally, I would like to thank all the people I met at NCSU for making my Masters so memorable.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The telecommunications world is being revolutionized with the emergence of next generation networking (NGN). Next-generation networks with IP based cores allows convergence of different network architectures in both wired and wireless worlds. The frontrunners of this convergence include SIP (Session Initiation Protocol) and IMS (IP Multimedia System) making it possible for triple play services (voice,data and video) to flow over the same network.

The term Next Generation Network (NGN) is used to describe an integrated, open network architecture that provides voice, data and multimedia services over the same network. The concept of integrated services on a single network is not new. Both Broadband ISDN and ATM deliver integrated services. The fact that NGN is a packet-based QoS enabled network providing both telecommunication and data services over different access technologies, sets it apart, making it new and appealing.

Most services today are tightly coupled with a specific transport network and signaling protocol. NGN separates service-related functions from the underlying transport-related technologies. This independence between service and transport layer makes the underlying technology invisible to the user regardless of where in a multi-service, multi-protocol, multi-vendor environment the user resides [35]. This concept of nomadicity provides seamless communication between fixed and mobile users.

The most important feature of NGN is the converged, QoS aware infrastructure. All information is transmitted as packets that are labeled according to their type (data, voice etc.) and are handled differently based on their QoS criteria. NGN ensures transparent service availability throughout the network by virtualizing services and provisioning seamless access to critical information. NGN provides ubiquitous access in a converged wireless and wireline network by decoupling transport technologies and services.

Lee and Knight [26] explain the difference between NGN and the Internet since both use IP as one important protocol. One main difference is that NGN does not restrict service delivery to best effort. NGN is a secure and trustworthy network supporting various services to meet the user's dynamic requirements. It also allows the integration of traditional telephone networks with data communications.

International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) Study Group 13 defined an NGN in Recommendation Y.2001 [18] as: A packet-based network able to provide services including telecommunication services and able to make use of multiple broadband, QoS-enabled transport technologies and in which service-related functions are independent from underlying transport-related technologies. It offers unrestricted access to users by different service providers. It supports generalized mobility which will allow consistent and ubiquitous provision of services to users.

Recommendation Y.2001 further characterizes the NGN by the following fundamental aspects:

- Packet-based transfer
- Separation of control functions among bearer capabilities, call/session, and application/service
- Decoupling of service provision from transport, and provision of open interfaces
- Support for a wide range of services, applications, and mechanisms based on service building blocks (including real-time/streaming/non-real-time and multimedia services)
- Broadband capabilities with end-to-end quality of service (QoS)
- Interworking with legacy networks via open interfaces
- Generalized mobility
- Unrestricted access by users to different service providers

- A variety of identification schemes which can be resolved to IP addresses for the purposes of routing in IP networks
- Unified service characteristics for the same service as perceived by the user
- Converged services between fixed and mobile networks
- Independence of service-related functions from underlying transport technologies
- Support of multiple last mile technologies
- Compliance with all regulatory requirements concerning emergency communications, security, privacy etc.

IP Multimedia Subsystem (IMS) is referred as the heart of NGN. IMS is the next generation IP based infrastructure enabling convergence of data, speech, video and mobile network technology. It is the envisioned solution that will provide new multimedia rich communication services by mixing telecom and data on an access independent IP based architecture, defined in 3rd Generation Partnership Project (3GPP), 3rd Generation Partnership Project 2 (3GPP2) and Internet Engineering Task Force (IETF) standards. The IMS architecture [Figure 1.1] will be used by service providers in NGN networks to offer network controlled multimedia services.

The aim of IMS is to provide all the services, current and future, that the Internet provides with roaming facilities. To achieve these goals, IMS supports peer-to-peer IP communications between existing technology standards while providing a framework for inter-operability of voice and data services for both fixed (POTS, ISDN) and mobile users (802.11, GSM, CDMA, UMTS). It provides session control, connection control and an application services framework with both subscriber and services data, while allowing interoperability of these converged services between subscribers. IMS truly merges the Internet with the cellular world; it uses cellular technologies to provide ubiquitous access and Internet technologies to provide appealing services [1].

IMS essentially replaces the control infrastructure in the traditional circuit-switched telephone network, separating services from the underlying networks that carry them. IMS has a signaling and media plane which work separately, unlike PSTN. The signaling plane handles the session control, authorization, security and QoS aspects while media planes

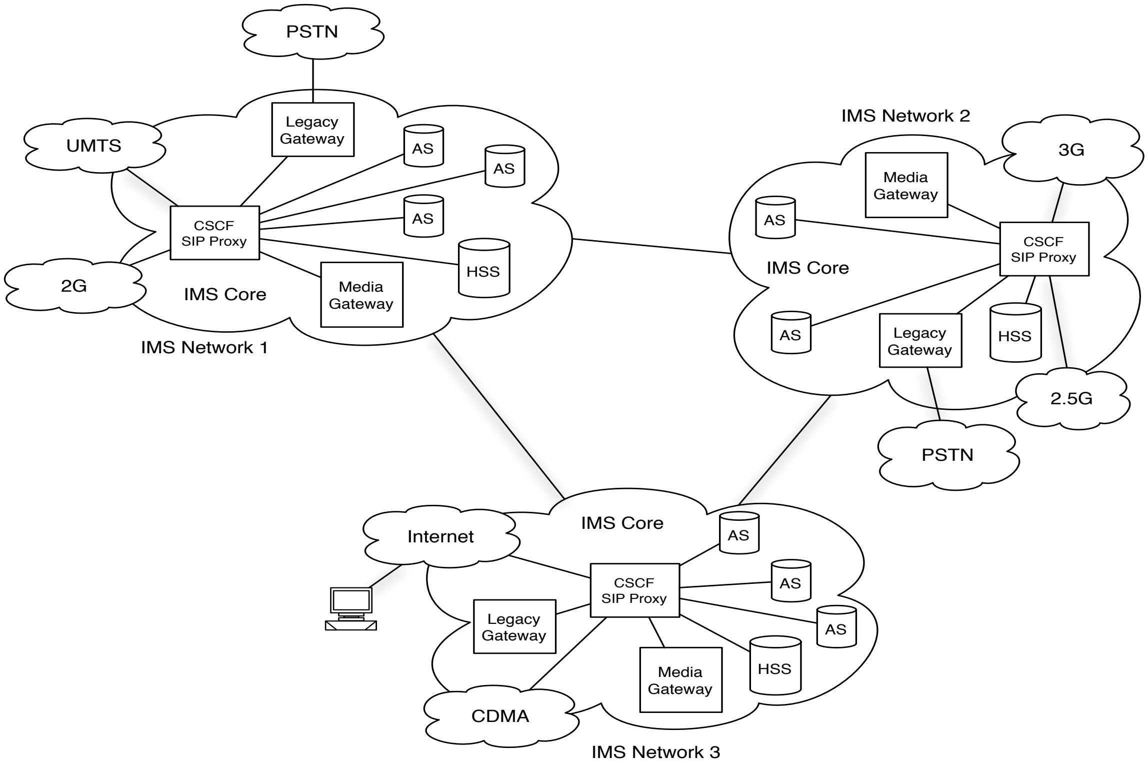

Figure 1.1: IMS overview: NGN Convergence

manages the media encoding and transport issues. IMS enables services such as text messaging, voice mail and file sharing to reside on application servers anywhere and be delivered by multiple wired and wireless service providers.

1.1 Motivation

The emergence of new technologies has led to the birth of new applications with VoIP, IPTV, Presence etc. This convergence has increased the demand for services with new performance characteristics; a proposition that decides the future of service provider's business.

Wong and Verma [36] discuss the advantages and disadvantages of providing new IP multimedia service capabilites as opposed to basic IP connectivity from a subscriber, network operator and third party application vendors perspective. The basic IP connectivity provided by plain vanilla Internet service provider compels the subscribers to use third-party providers for IP multimedia services and applications.

While the network operators may lose potential revenue by not providing basic IP multimedia services like voice-over-IP call capability, it would be difficult to predict which services would be profitable and popular with the customers. So the network operator provides flexibility by allowing third party applications while focusing on their core competency of providing connectivity. A drawback would be lack of reliability, quality issues caused by careless reuse by subscribers of IP multimedia applications, leading to loss of customers and revenue.

Today most of our services are offered by both cable television and telecommunication operators. Service providers are struggling with the opportunities and challenges that the convergence presents while striving to gain a foothold in the market segment. It is assumed that an integrated solution with a combination of services will bolster customer satisfaction to encourage loyalty and discourage churn. Interoperability between different service providers is not considered as a design target.

As service providers build out their 3G, broadband and converged networks, they are moving to a business model that makes the quick introduction of new services while exploiting the full revenue potential of these new services and having the flexibility of scaling easily. Service providers must now focus on management of the network resources to optimize the performance of the services they deliver.

Expectations of the sophisticated customers from the service providers has multiplied with the new technologies, requiring complicated and customized solutions for their satisfaction. Service providers need to monitor and manage network traffic to optimize performance in order to meet the guaranteed levels of service promised to their customers. Effective monitoring of network performance and analysis of network data to correlate end-to-end service performance is essential for the thorough understanding of network behavior [6].

Deploying new applications in a service provider environment creates rigorous demands on the network infrastructure and the network operations staff. Service providers are often targets of easy purchase of new technology products from application vendors. Scalability, flexibility and integration of diverse applications is critical to support large and complex networks while reducing the operational cost of the network.

The challenges in offering these new services in light of today's highly competitive environment with increased customer focus is, not only seamless transition between disparate network topologies but also the flexibility to embrace new service deployment technologies such as the IP Media Subsystem (IMS). Such issues are mostly associated with determining the right business model, back-end support and economic environment rather than technology. Service providers need a model that makes quick introduction of new services of primary importance, while fully exploiting their full revenue potential and having the flexibility to accommodate easy scaling of networks. For example, using the right billing platform to address a variety of subscriber demographics or having the appropriate subscriber density to financially justify the introduction of a new service are a few factors that affect decisions to offer IP multimedia services.

Our research focuses on methodologies to evaluate IMS based architecture from the viewpoint of service providers. We define service models for the IMS network with utility functions and optimize the desired properties like the service rate to achieve maximum revenue generation. Such an economic-theoretic performance evaluation provides flexibility in the use of network resources [7]. This significantly simplifies the mathematical computations without reducing the qualitative applicability of results. The service models with utility functions also serve as decentralized control mechanisms to optimize the network properties while studying revenue generation trends in networks.

1.2 Research Overview

In our service models, we abstract the entire IMS network into a tandem queueing model. Classical queueing theory is used to obtain insights on the performance of the network. As the QoS in IMS for any connection is negotiated during the session setup phase, we illustrate formal methods of modeling the network from an end-to-end session setup delay perspective. We demonstrate variations in the modeling by introducing constraints in the network parameters that the service provider might wish to apply.

The network performance is analyzed over different network scenarios and the maximum utility is determined by optimizing the server service rate. Results show trends in the network performance which can be used to strategize the QoS guarantees of the network. Results remain consistent over different scenarios with different constraints.

IMS uses SIP for establishing connections between users. We explore the SIP signaling messages and propose a methodology to derive the approximate workload for a SIP server. The methodology described could be applied to any server to obtain its workload. Case-studies using popular publicly accessible SIP clients were conducted to characterize the server workload and to gain insights into the structure of SIP servers. This characterization can be inputted into network testing benchmarks and utilized for predicting network performance.

Using simulation, we identify the nature of dependence of the utility with the network parameters. We perform capacity planning by achieving a favorable number of servers with optimized utilization. This approach to evaluate the network is extremely useful for planning and efficiently structuring the network resources.

1.3 Contribution of this Thesis

Today with IMS becoming a fast reality, service providers are competing to offer IP multimedia services. Service providers are streamlining their operations and are focusing on efficient network management to facilitate an improved revenue stream by efficient use of network resources. The resulting saving can be used to fund new and improved service options.

Capacity planning, monitoring network performance and analyzing the relevant data gives the service provider the ability to direct network performance effectively into network actions. An ideal performance management solution is platform independent and extensible while providing integrated, total network coverage. The ideal solution must allow service providers to monitor ongoing physical network performance, analyze its data to correlate end-to-end service performance, and finally, to take action based on a complete understanding of network behavior [6].

Predicting trends of performance is an important phase in the planning and implementation of the network. Its value is more prominent in networks with strict service guarantees. We focus more on the approach of abstraction of IMS network with realistic network workload. Such an analysis should be useful for service providers like Sprint or Verizon who would like to setup their network with efficient utilization and maximum revenues. Deciding on parameters like number of servers, placement, service level and quality guarantees are important. For example if Sprint requires to build an IMS network using

Tekelec SIP servers, it would need a modeling tool to plan and design the network. It can use our approach and analyze the network for quick trends on anticipated performance levels to decide on design parameters.

Such a depth of research is crucial if the service provider must employ service level agreements (SLAs). The SLAs give the service providers a better understanding of the expectations of the customer. Providers who follow such a pattern of planning and performance analysis can consistently perform well and improve network infrastructure to accommodate their customer's evolving needs.

The IMS architecture requires the service providers to change the conceptual and physical make-up of their network architecture and service models. They are required to add new infrastructure like application servers, gateways, soft-switches etc. before being able to provide services. It is unlikely that any service provider will make a leap into IMS architecture without planning and testing while considering the network's new features and eventually migrating to services. It is critical for providers to minimize their risks and maximize the return on their current infrastructure investments while evolving to IMS. Our proposal helps service providers take the best approach while purchasing IMS components. Planning and analysis as we suggest offer service providers the incremental steps for a reasonable deployment of IMS with return of investment.

Chapter 2

Background

IMS is likely to become the norm for all broadband access, forming the foundation of emerging next generation networks. The different protocols in IMS focus on delivering different services including presence, instant messaging and push to talk [30]. IMS builds on IETF protocols like SIP, SDP, DIAMETER, MGCP etc. to create a robust and complete multi-media system with operational profiles providing support for operator control, charging and billing, and security.

The IMS upon which NGN is based uses the Session Initiation Protocol (SIP) to tailor a telephony-oriented signaling network over the IP cloud. It is the new wireless/wired control plane replacing the earlier telco's SS7 signaling. An IMS network consists of many SIP proxy servers that mediate customer connections and access to network resources. Each user has an associated home network and can roam across wired and wireless networks. IMS also has a policy engine and authentication, authorization and accounting (AAA) server for operator control and security.

IMS-based services provide person-to-person and person-to-content communications of voice, text, pictures and videos in a variety of modes in highly personalized and controlled way. Integration with data and voice networks with interoperability and roaming facilities makes IMS a key enabler for fixed-mobile convergence [28].

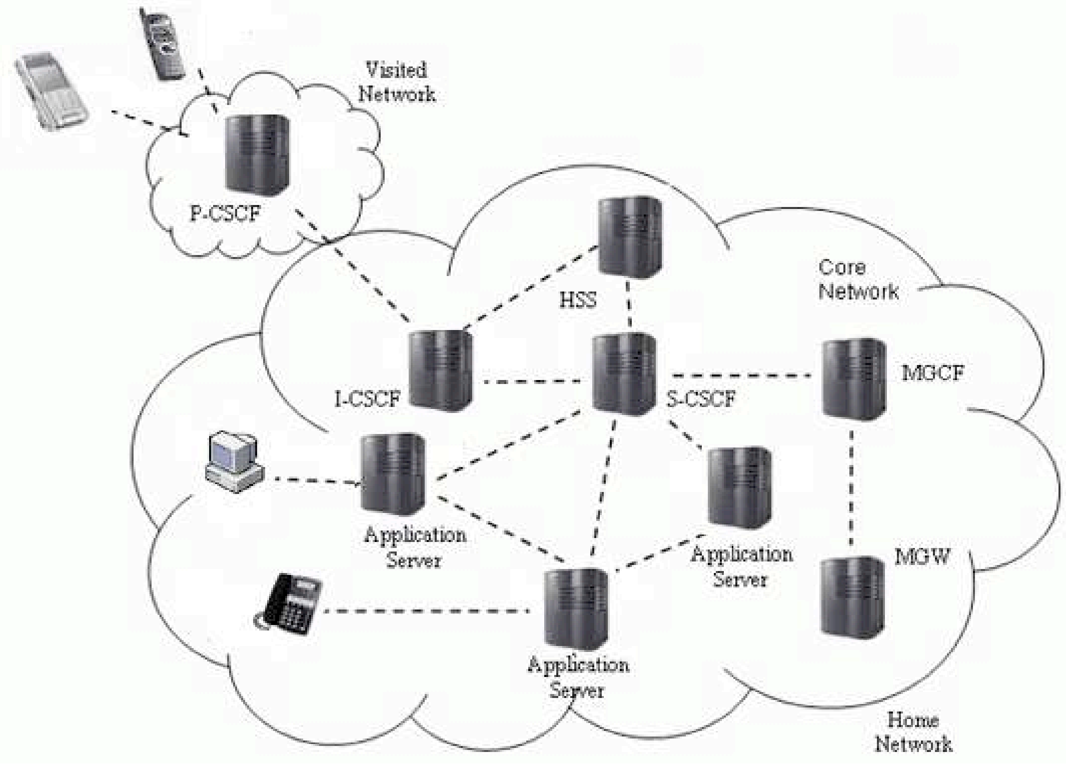

Figure 2.1: Typical IMS architecture

2.1 IMS Architecture

The IMS architecture is a collection of functional components linked by standardized interfaces [Figure 2.1]. An IMS based network comprises of many SIP servers, user databases known as Home Subscriber Servers (HSS), Application Servers (AS), Media Resource Functions (MRF), Public Switched Telephone Network (PSTN) gateways etc to mediate all customer connections and access to network resources [30].

Access Network

The user can connect to an IMS network using fixed access, mobile access or wireless access methods that use the standard Internet Protocol (IP). Direct IMS terminals like mobile phones, PDAs, computers etc. can register directly into an IMS network using a SIP client running over IPv6 or IPv4. PSTN, H.323 and other circuit switched networks with non IMS-compatible VoIP systems can connect through gateways.

Core Network

The core network facilitates standards-based communication to application services. The components in the core network are described below:

Call Session Control Function

The Call Session Control Function (CSCF) is the session routing point in the IMS core network. CSCFs are dynamically associated, service-independent and standardized access points. It distributes incoming calls to the application services and handles initial subscriber authentication. The CSCF uses SIP messages to pass the call event to the service and adds additional header information to maintain control of the call. IMS has a user-focused approach for delivery of services. It allows users to access personal services through CSCFs.

The CSCF component uses IMS Service Control (ISC) interface to intercept call signaling and pass it to application services for processing. The ISC interface compares these filters with the incoming SIP message and using the information stored in HSS determines which application services should be invoked and in which sequence. IMS uses a Session Border Controller component to route calls through the firewall to the CSCF component within the IMS network. CSCFs are discussed in detail in Section 2.2.

Application Server

This is a SIP entity that hosts and executes services. IMS allows reuse of functions for fast service creation and delivery. Multiple services like telephony and messaging can be hosted by a single Application Server. Collating services in one Application Server reduces the workload of CSCF in the control layer.

PSTN Gateways

PSTN Gateways provide interfaces to circuit switched networks, allowing IMS terminals to make and receive calls to and from PSTN. A PSTN gateway can be decomposed into MGCF, SGW and MGW functions.

- **Media Gateway Control Function (MGCF)**

  This is the central node of PSTN gateway that routes calls from or to legacy platforms like SS7/TDM channels or PSTN/PLMN networks. All calls routed through these media gateways enter the IMS core network as SIP/RTP media streams. RTP streams are routed directly between media servers, gateways and endpoints. The MGCF controls the resources in MGW.

- **Signaling Gateway (SGW)**

  This interfaces the signaling plane of the PSTN network or circuit switched networks.

- **Media Gateway (MGW)**

  MGW interfaces the media plane of the PSTN and circuit switched networks.

Media Servers

Media Resource Function (MRF) provides the source of media in the home network. The signaling plane is referred as Media Resource Function Controller (MRFC) and the media plane is the Media Resource Function Processor (MRFP). The MRFC communicates with an application service and directs a separate media server to handle the media stream. It controls the resources in the MRFP. The MRFP implements all media-related functions like playing or mixing media.

Home Subscriber Server

The Home Subscriber Server (HSS) provides a central repository for subscriber information. The HSS stores all subscriber information required to establish sessions between users and provide services to subscribers. The HSS stores subscriber registration, preferences, user profile information, location and service information. Networks with a single HSS do need an Subscriber Location Functions (SLF). The SLF maps the user's address to HSSs. A node that queries the SLF, with user address as input, obtains the HSS that contains all the user information.

Charging Entities

Charging entities are used to control billing information for subscribers using network services. The interfaces used by an application service for charging entities are maintained in the IMS network. IMS uses Diameter messages for offline (Rf) and online (Ro) charging entities.

Different functional components are used for providing different services. For example, to transport media, IMS uses Call Session Control Function (CSCF), Session Border Controller (SBC), IMS Service Control Interface (ISC), Media Gateway Control Function (MGCF), Media Resources Function (MRF), Media Resource Function Controller (MRFC), Home Subscriber Server (HSS) and Charging Entities.

2.2 IMS and SIP

IMS has a number of different protocols with each serving a particular function. This allows for modularity, flexibility, simplicity and extensibility when building such a diverse, robust and complex network infrastructure. Signaling protocols are meant for the end points involved in the call and the network routers treat these signaling packets as any regular data, ignoring any semantics implied by them.

The Session Initiation Protocol (SIP) [32] lies at the core of the IMS architecture. SIP is the real-time communication protocol responsible for VoIP in IMS. SIP has been expanded to support video and instant-messaging applications. Based on SIP, IMS defines standard control plane interfaces for creating new applications. SIP is designed to perform basic call-control tasks, such as session call set up and tear down and signaling for features such as call hold, caller ID, conferencing and call transferring.

SIP is an open standard that is well-supported throughout the carrier and vendor community. SIP implementation is popular due its compatibility with other standards like Voice XML. This interoperability allows developers to focus on both service logic and application services and create efficient services. Also the inherent SIP framework allows service providers to quickly introduce extensions like security, compression etc. to meet the evolving network and service needs [29].

The SIP servers are the essential nodes of IMS. They are collectively referred as Call/Session Control Functions (CSCF) which are further categorized as Proxy-CSCF

(P-CSCF), Interrogating-CSCF (I-CSCF) and Serving-CSCF (S-CSCF) [4].

- **Proxy-CSCF**

  P-CSCF is the first point of contact for the IMS terminal. It acts as an outbound/inbound SIP proxy server performing user authentication and verification of correctness of SIP requests. P-CSCF is assigned during the registration and is on the path of all signaling messages. Authentication using IPsec, compression, QoS policy control and charging are performed by the P-CSCF.

- **Interrogating-CSCF**

  I-CSCF is a SIP proxy located at the edge of an administrative domain. It queries the HSS using the DIAMETER Cx and Dx interfaces and retrieves the user location information for routing purposes. I-CSCF's IP address is published in the DNS of the domain for usage by the P-CSCF in a visited domain or a S-CSCF in a foreign domain as an entry point for all SIP packets to that domain.

- **Serving-CSCF**

  S-CSCF is the central node of the signaling plane. Its a SIP server that acts as a registrar while performing session control and routing services. S-CSCF is always located in the home network. It decides to which application server(s) the SIP message will be forwarded to, in order to provide the services and enforce the policy of the network operator.

Each IMS connection has an initial setup phase during which the service parameters of the connection are negotiated between the source, destination and the intermediate CSCFs. We refer to all these components generically as "servers".

2.2.1 SIP Protocol Details

Sequencing, inter-media synchronization, payload identification and frame indication are services provided by a transport protocol. The Real-time Transport Protocol (RTP) provides services needed for generic real time media like voice and video. The Real Time Control Protocol (RTCP), which is a subpart of RTP provides QoS feedback. Receivers provide feedback on the quality of the reception so that RTP sources can use that information to adjust the data rate [34]. A session is referred as a single RTP session carrying a single media type.

A session description describes the capabilities of a session and contains information required to join the session. Session Description Protocol (SDP) is used to describe multimedia sessions. In multimedia sessions, the information includes the IP address and the port number where the media is accpeted and the codecs to be used to encode the media.

SIP makes minimal assumptions about the underlying transport protocol. It can be directly used with any datagram or stream protocol, provided that either a whole SIP message gets delivered or no message is delivered. SIP can thus be used with UDP or TCP in the Internet, and with X.25, ATM, IPX or PPP.

A SIP connection is initiated by issuing an INVITE request, and terminated by issuing a BYE request. SIP has several messages. For example, INVITE message invites a user to a call and establishes a new connection while a BYE message terminates a connection between two users. OPTIONS message solicits information about capabilities, but does not set up a connection and STATUS message informs another server about the progress of signaling actions that it has requested. ACK is used for reliable message exchanges for invitations. CANCEL terminates a search for a user while REGISTER conveys information about a users location to a SIP server.

SIP is a request-response protocol. When a SIP endpoint is contacted with a request, it sends back a response to the sender. The response has a response code and a response message. There are 6 different classes of response codes that fall from 100 to 600. Responses in the 100 category denoted as 1xx indicate call progress and are provisional. They are always followed by other responses indicating the final outcome of the request like 2xx for success, 3xx for redirection, 4xx, 5xx and 6xx for client, server and global failures respectively. All responses are confirmed with an ACK request to ensure reliability.

SIP is a text-based protocol unlike other signaling protocols like Q.931 and H.323. All parameters in SIP headers have a corresponding value with a single level of sub parameters. SIP identifies users using SIP Uniform Resource Identifiers (URI), which are similar to email addresses. SIP URIs consist of a user name and a domain name.

All SIP messages start with a `request line` in requests and a `status line` in responses as in Figure 2.2. This is followed by the SIP header fields of the format *name:value*. The status line contains the protocol version (SIP/2.0) and the status of the transaction.

```
Request-Line: INVITE sip:timetest123@68.142.233.183:5061;transport=tcp SIP/2.0
Message Header
  From: timetest345<sip:timetest345@68.142.233.183:5061>;
     tag=5d00690-0-13bb-12060-55d2f8bb-12060
  To: <sip:timetest123@68.142.233.183:5061>
  Call-ID: 5d0ad18-0-13bb-12060-2d2405ea-12060@68.142.233.183
  CSeq: 1 INVITE
  Max-Forwards: 70
  Via: SIP/2.0/TCP 10.10.10.5:5051;branch=z9hG4bK-12060-46678cd-70054828
  Contact: <sip:timetest345@127.0.0.1:5051;transport=tcp>
  Content-Length: 0
```

Figure 2.2: SIP Message

The request line consists of a method name, the Request-URI and the protocol version. The method name indicates the purpose of the request and the Request-URI contains the destination of the request.

SIP messages contain a set of mandatory header fields as in Figure 2.2. They are `To`, `From`, `CSeq`, `Call-ID`, `Max-Forwards` and `Via`.

- The header field `To` contains the URI of the destination. The `tag` parameter serves as a mechanism to identify a dialog, which is the combination of the Call-ID along with two tags, one from each participant in the dialog.

- The `From` header fields contains the URI of the sender,

- The `Cseq` header contains the sequence number and a method name. They are used to match requests and responses.

- The `Call-ID` provides a unique identifier for a SIP session.

- The `Max-Forwards` header is used to avoid routing loops. Every server that handles a request decrements its value by one, and if it reaches zero the request is discarded.

- The `Via` header field keeps track of all the proxies a request has traversed. These `Via` entries are used to route the responses.

The message body is separated from the headers by an empty line. A set of header fields provide information about the body like its length, its format, and how it is handled.

Message bodies are transmitted end to end. So the proxies do not parse the message body to route the message. SIP agents can choose to encrypt the contents of the message body for security purposes. We monitor SIP sessions and analyze the QoS parameters in the SIP messages in Chapter 5.

2.3 SIP and Quality of Service

3GPP stresses that the IMS architecture be viewed as a collection of functions linked by standardized interfaces. Quality of service is one of the main reasons for the emergence of IMS. IMS ensures QoS by session layer negotiation with resources granted at transport layer. It provides correlated accounting or charging among service, session and transport layers under operator control.

SIP establishes connections between two or more endpoints in an IP network. The QoS parameters for the session are negotiated during the setup. Thus, performance evaluation of IMS based networks from a QoS perspective is essential. We propose a modeling methodology that uses real-life workload characterization, queueing analysis and optimization to provide a systematic way to select system design parameters, such that the QoS criteria is satisfied and the overall system utility is maximized.

Chapter 3

Related Work

Most of IMS-related research has been focused on the engineering rules, protocol development, compatibility issues and refinements of SIP. Performance evaluation from an end-to-end perspective and capacity planning are areas that haven't been given the much needed attention. With the entire telecommunication industry working towards the convergence, envisioning new IP multimedia services with QoS, the assessment of setup delay using SIP needs to be critically evaluated.

ITU-T Recommendation E.721 provides guidelines on network grade of service parameters and target values for circuit-switched services in the evolving ISDN [17]. Recommendation E.721 specifies post selection delays as 3 seconds for local connection using ISDN. With this as a requirement, a study of IMS architecture based on delay sensitivity is a must for QoS guarantees.

Analysis of performance metrics based on network parameters like number of servers, arrival rates of traffic into the network and service rates can ensure efficient resource utilization while providing the promised QoS. Capacity planning of IMS based networks will be simplified if effective trends of performance can predicted quickly.

Eyers and Schulzrinne [10] cover SIP call setup delay based on SIP and H.323 traces from the Surveyor database. The Surveyor project provided the continuous delay and loss statistics for UDP packets between selected cities. The focus is on delay due to UDP loss and assumptions are made about the processing times of tasks. Kist and Harris [22] present signaling delays in 3GPP with emphasis on DNS lookups. They assume the queueing delays to be less than 5 ms based on current web server implementations and assume the SIP servers to have exponential service distribution.

Haase *et al* propose a unified mobility manager (UMM) for effective inter-working of wireless networks and voice over IP networks. The UMM reduces performance degradation by combining UMTS Home Location Register (HLR) and SIP proxy functionality in one logical entity. The focus is on integration of SIP in UMTS with PSTN [14]. Curcio and Lundan [8] study SIP call setup time in 3G networks and compare it with call setup time over LAN networks. The effect of SIP calls over lossy channels with restricted bandwidth is also studied.

Fathi *et al* evaluate SIP session setup delay with various underlying protocols like transport control protocol (TCP), user datagram protocol (UDP), radio link protocol (RLP) as a function of the frame error rate (FER) [11]. The processing delays and queueing delays through the different SIP servers are not considered in the analysis. Kueh *et al* study the performance of SIP-based session setup over the S-UMTS with consideration to the larger propagation delay over the satellite as well as the impact of the UMTS radio interface under different channel conditions [24].

The service availability of VoIP in the current Internet was studied by Jiang and Schulzrinne [19]. Several metrics of availability like call success probability, time loss below a threshold, distribution of network outages and the call abortion probability induced by the outages were analyzed.

The SIPstone Benchmark [33] by Schulzrinne *et al* attempts to measure request handling capacity of SIP servers. It is useful as a tool for dimensioning and provisioning of the SIP network. SIPstone is designed as a benchmark with a single standardized implementation and workload. The workload is simple and exercises a breadth of system components associated with this environment like concurrent initiation of calls by multiple users, random call arrivals and the forwarding of requests. Such an approach is helpful from a SIP server developer's point of view and not from the service provider.

Chebbo and Wilson [5] describe a flexible modeling tool developed in Fujitsu Labs for a mobile operator network using SIP as the control protocol. The tool is implemented in Excel and incorporates a traffic model, a SIP server model, a functional model, and a system model based on the estimated user activity. The effect of moving functionalities between the different IMS entities is also studied using the tool by comparison of the bandwidth used by SIP servers to support different traffic models. The SIP servers are modeled as

M/M/1 queue with exponential service times. Our models are based on a realistic workload with delay based parameters.

In Zhu [38] analysis is performed of a SIP network in IMS from a UMTS perspective under a controlled environment with bottlenecks. The traffic in the network is assumed to be following an M/D/1 model.

Banerjee *et al* perform a case study for a heterogeneous network with IP backbone having a combination of UMTS and WLAN access networks, to evaluate the performance of SIP-based mobility management. The numerical analysis has been performed using a heterogeneous queueing network with M/M/1 queue models for CSCFs and a priority based M/G/1 model for the destination server [2].

IMS network throughput is dependent on the capacity, capability and efficiency of the SIP server to a large extent. Better implementation of the SIP server with higher processor speed and more memory will enhance the performance of the network significantly. The response times of these servers directly effect the call setup delay. We attempt to characterize the SIP server workload more realistically by monitoring actual traffic and deducing a service distribution pattern. This effort was motivated by the complete lack of literature related to realistic SIP server workload and the SIP message processing times.

In Wu *et al* [37], SIP performance is evaluated for SIP-T which is an effort to provide the integration of legacy telephone signaling into SIP messages through encapsulation and translation. Evaluation is based on QoS parameters and system performance of MGC (Media Gateway Controller) in carrier class VoIP network. Our analysis is based on the SIP definitions in RFC 3261 [32].

Gurbani *et al* analyze the performance and reliability of an end-to-end native SIP ecosystem under varying network parameters through a hierarchical performance and reliability model [13]. The mean response time for a proxy server is computed based on service time assumptions; for example, the INVITE service rate is fixed at 0.5 ms^{-1}. Our research analyzes the IMS architecture's performance from an end to end delay perspective. We propose to view the entire IMS architecture as a open feed forward tandem queueing network and predict trends in performance based on our analysis.

End-to-end QoS monitoring and distributed monitoring with link degradation detections for IMS have become popular recently. QoS monitoring is essential for tracking the ongoing QoS, for comparison of monitored QoS against the expected performance, for detection of possible QoS degradation, and then tuning network resources accordingly to sustain the delivery of the guranteed QoS.

In an end-to-end QoS monitoring approach, only the end-to-end QoS between the sender and receiver of a real-time flow is monitored. In a QoS distribution monitoring approach, the QoS distribution experienced by the flow in different network segments is monitored in addition to the end-to-end QoS. Tham *et al* describe the challenges involved in providing QoS distribution monitoring and propose solutions that can be directly adopted in monitoring many QoS parameters like throughput, loss rate etc [20].

Raty *et al* discuss the feasible placements of network analyzing and monitoring tool to survey the protocol communication in IMS. The placement of the tool can be either at the endpoints or at the common point(s) in the communication path between the endpoints. The network analyzing and monitoring tool can survey the network for QoS attributes once the SIP connections are setup. Protocol information of SIP, SDP, UDP, RTP, RTCP can be collected at the endpoints [31]. Using a subset of the methodology we describe for SIP workload characterization, the end-to-end real-time traffic and the QoS parameters negotiated can be easily monitored. We discuss this in Chapter 5. In what follows, we describe our service models for IMS architecture based networks.

Chapter 4

Service Models for IMS

Our research focuses on developing formal methods for optimizing the utilization of network resources while maximizing the revenues generated. We propose service models for planning and performance prediction which can help service providers built an IMS network successfully. Planning is a crucial phase in the deployment of any network. It is critical for service providers to model and dimension their network, evaluate and predict trends in the network before actually deploying it.

4.1 Theoretical Framework for Performance Measurements

With IMS becoming a reality, the market is buzzing with different vendors offering their IMS solution for the next generation. IMS is an architecture that standardizes functionalities and does not have any specifications on the nodes. Implementors are free to combine two or more functions into a single node or can split a function into two or more nodes. This flexibility in the specifications has led each vendor to develop their own IMS solution.

Today's market is competing with many network equipment supplier's IMS solutions like Nokia's fixed mobile convergence, Motorola's global applications management architecture, Cisco's service exchange solutions for IMS and similar convergence solutions from Alcatel, Ericsson, Lucent and Tekelec. Service providers are the easy targets as they approach deployment of a complete IMS network. With so many choices to choose from, the service providers need a little direction and guidance. They need some quick performance evaluations and insights into the network trends with different vendor solutions before investing. This is crucial with IMS as it offers excellent QoS with all of its multimedia services. Service providers must know their capacity and network constraints for satisfying the SLA criteria. They must also manage the network resources efficiently to get most return of their investment. Our research aims to provide the service providers easy and incremental steps for a reasonable deployment of a full IMS network with maximum returns.

Most multimedia sessions are delay sensitive and require fast setups. IMS is the next generation solution for a complete fixed-mobile convergence between data and circuit switched networks like PSTN. It is expected to replace the traditional and dedicated circuit switched plain old telephone service (POTS). Thus, IMS is required to prove its competence by meeting the grade of service parameters set for circuit-swiched services in evolving ISDN. ITU-T recommendation E.721 [17] specifies a post dial delay of 3 seconds for a local connection, 5 seconds for a toll connection within the country and 8 seconds for international connections.

We study the IMS network from an end-to-end delay perspective. The focus is on SIP signaling delay during which the connection between the end-points with the desired QoS is setup. We denote the network as a utility function and perform queueing analysis. The utility function is then optimized to achieve the maximum revenue. The analysis helps us predict trends in the performance of the network. We also propose a methodology to characterize the workload of SIP servers in the network, discussed in Chapter 5.

4.2 Utility Function for IMS networks

Economics is generally used to study revenue generation in networks. They also serve as decentralized control mechanisms to optimize the network properties. Such an economic-theoretic performance evaluation provides flexibility in the use of network resources [7]. This significantly simplifies the mathematical computations without reducing the qualitative applicability of results. In our modeling we define utility functions for the IMS network and optimize the service rate to achieve maximum revenue generation.

Optimum network performance can be achieved by economic-theoretic network evaluations. An economic performance evaluation helps the service provider analyze the flexibility and the ability of the network to adapt and customize various services. Such an economic perspective views the network and its customers as a whole and defines the system performance to include the value the customers obtain from using the network services.

The service provider offers a variety of services and gains revenues based on the customer subscription to those services. Profit is gained when the revenues earned exceeds the cost of operating the network. The provider also faces losses with outages in the network due to which the promised QoS criteria is not satisfied and the customers are churned. Satisfying the QoS requirements is crucial in IMS networks. The session setup must be fast for quicker connections and must be completed within a specified time period. We capture these factors with our utility function for IMS networks.

We model the utility function of an IMS network as

$$U = V(\lambda) - C - (\Phi_{expected} - \Phi_{actual}) \tag{4.1}$$

Here U is the utility, $V(\lambda)$ is the revenue earned by serving λ connections, C is the cost compensation for maintaining K servers in the network and $(\Phi_{expected} - \Phi_{actual})$ is a penalty function representing revenue losses due to lost connections. $\Phi_{expected}$ is the desired target time to complete the session setup and Φ_{actual} is the actual response time within which the session was completed.

Each connection involves a session setup delay during which the service parameters are mutually agreed upon by the source and destination. We assume connections to be lost if the session setup delay for that connection exceeds a threshold time limit of $\Phi_{expected}$.

V(λ) is the income earned by the service provider by serving λ number of connections. We assume this income to be constant to model a flat charge tariff on the customer for the usage of the network resources. Dynamic models with different pricing schemes can also be modeled similarly. For example, the revenue generated could be dependent on other variable network parameters like the number of servers, the capacity of the server or the service rate of the server etc.

The penalty function $(\Phi_{expected} - \Phi_{actual})$ in (4.1) is computed based on different criteria and constraints. We evaluate the penalty function for the following scenarios:

1. Total average time for session establishment experienced in K tandem servers within time limit τ

2. Probability of session setup over K *i.i.d.* servers being completed within threshold time limit τ
3. Penalty of c being incurred for each lost connection

These scenarios could be used as constraints or to calculate the penalty function.

Cost compensation C in (4.1) should model the variance in the costs for supporting the network along the expansion path. We assume that the cost of a single server is a function of its service rate. The number of servers in the network are linearly dependent on their service rates. So the cost compensation is the product of the number of the servers in the network, K, and the cost function $G(\mu)$:

$$C = K * G(\mu) \tag{4.2}$$

The cost function in (4.2) is represented as

$$G(\mu) = a\mu - b\mu^2 \tag{4.3}$$

where a is the cost of each server based on its service rate μ and b is the sales volume like a discount rate to model the economics of scale. The concavity of the server cost with respect to its service rate is captured by this quadratic equation. The values of parameters a and b are assumed to be externally provided corresponding to the current infrastructure costs [15]. Using the first derivative of the cost function, the upper bound for the cost function is obtained. The maximum value for the cost function is at $\frac{a}{2b}$.

The mean service rate of the server is assumed to be greater than or equal to the mean arrival rate to model fast connection setups, i.e., $\mu \geq \lambda$. This condition limits the values for μ in the range $\mu \in [\lambda, \frac{a}{2b}]$. This is derived from the quadratic cost function in (4.3).

In what follows, we propose a tandem M/G/1 queueing network as the service model for IMS network. We optimize the utility function for a maximum service rate for the efficient usage of the network resources while getting maximum returns.

4.3 M/G/1 Queueing Model

Consider an IMS based network with K identical servers serving a population of size N. The users initiate a connection to the network as a poisson process with an intensity of λ. The service time distribution of the servers in the network can be geometric, beta, gamma, etc. We assume the service time distribution to be general with definite first and second moments. The K servers have infinite queues and do not experience any losses due to buffer overflows. We abstract such a network as a feed-forward tandem queueing network with K servers of type M/G/1/∞.

For example, in Figure 4.1 we have an IMS network with P-CSCF, I-CSCF and S-CSCF represented as a queueing network. All signaling messages during the connection setup have to traverse through these three CSCF sequentially. This can be viewed as a feed-forward tandem path. The processing time and the queueing delay experienced in the CSCFs and any other components that may be present in the IMS network are aptly captured by this feed-forward tandem M/G/1 queueing model.

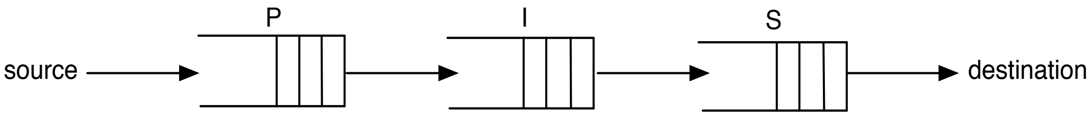

Figure 4.1: IMS queueing network

4.4 Analysis of M/G/1

In an M/G/1 queueing system, the average response time can be computed using the Pollaczek - Khintchine mean value formula [23]

$$T_{avg} = \bar{x} + \lambda\bar{x}^2 * \frac{(1 + C_b^2)}{2(1 - \lambda\bar{x})} \tag{4.4}$$

where $\bar{x}$ is the mean service time, i.e. $\frac{1}{\mu}$ and C_b^2 is the coefficient of variation for the service time.

We evaluate this M/G/1 queueing network based on the earlier defined scenario that the total time to establish the session τ exceeds the total average delay experienced in the K servers in the network, $\tau \geq K * T_{avg}$.

The optimal utility function is written in terms of $\bar{x}$ as

$$\hat{U} = max_{\bar{x}} \left[V - K(a\bar{x} - b\bar{x}^2) - \beta(\tau - K * T_{avg})\right] \tag{4.5}$$

$$\hat{U} = max_{\bar{x}} \left[V - K(a\bar{x} - b\bar{x}^2) - \beta\left(\tau - K * \left\{\bar{x} + \frac{\lambda\bar{x}^2 * (1 + C_b^2)}{2(1 - \lambda\bar{x})}\right\}\right)\right] \tag{4.6}$$

For ease in calculation we assume $M = (1 + C_b^2)$ as a predetermined constant.

The first order derivative is

$$\frac{\partial U}{\partial \bar{x}} = -aK + 2bK\bar{x} + K\beta + \frac{K\beta\lambda\bar{x}M}{(1 - \lambda\bar{x})} + \frac{K\beta\lambda^2\bar{x}^2M}{2(1 - \lambda\bar{x})^2} = 0$$

Since solving for the closed form of $\bar{x}$ is difficult, we deduce some properties of the solution by applying the conjugate pair theorem from calculus [9]. The conjugate pair theorem states that for a maximization problem max_x F(x,a), the derivative $\frac{\partial x^*}{\partial a}$ and the cross partial F_{xa} both have the same sign. Using the conjugate theorem we obtain the following analysis:

1. If the cost a increases, the service time $\bar{x}$ decreases.

$$U_{\bar{x}a} = -K \rightarrow \frac{\partial U^*}{\partial a} < 0$$

2. If the cost b increases, the service time $\bar{x}$ increases.

$$U_{\bar{x}b} = 2K\bar{x} \rightarrow \frac{\partial U^*}{\partial b} > 0$$

3. If the penalty for losing requests β increases, the service time $\bar{x}$ increases.

$$U_{\bar{x}\beta} = K + \frac{K\lambda\bar{x}M}{(1 - \lambda\bar{x})} + \frac{K\beta\lambda^2\bar{x}^2M}{2(1 - \lambda\bar{x})^2} \rightarrow \frac{\partial U^*}{\partial \beta} > 0$$

4. If the arrival rate λ increases, the service time $\bar{x}$ increases.

$$U_{\bar{x}\lambda} = \frac{K\beta\bar{x}M}{(1 - \lambda\bar{x})} + \frac{2K\beta\lambda\bar{x}^2M}{(1 - \lambda\bar{x})^2} + \frac{K\beta\lambda\bar{x}^3M}{(1 - \lambda\bar{x})^3} \rightarrow \frac{\partial U^*}{\partial \lambda} > 0$$

Now, a general service time distribution is difficult to solve if its not in a closed form. Although, in some special cases we can still use M/G/1 analysis. Analysis of penalty scenarios 2 and 3 require calculation of the tail probabilities of the queueing distribution which is hard to solve. Such cases can be solved numerically. For example, certain advanced techniques are described in [27], [12] that use the measured samples to get approximate values. However, solving using these methods is beyond the scope of this thesis.

Instead, for the sake of illustration of the remaining cases mentioned in the penalty scenarios, we assume the servers to have exponentially distributed service time with a mean of μ. So the network is now viewed as a M/M/1/∞ queueing model.

4.5 Illustration using M/M/1

In a M/M/1/∞ queueing system, the expected response time is $\frac{1}{\mu-\lambda}$ and the probability of the response time exceeding a time period T is given by $e^{-(\mu-\lambda)T}$. From Jackson's theorem and Kleinrock's independence approximation, a system of tandem queues can be effectively decomposed into an independent set of M/M/1 queues. So the total time spent in the network can be approximated as the sum of time spent at each server in the network [23].

We assume that all sessions to be established are between IMS compatible terminals. So in our modeling the total number of servers in the network and the number of servers involved in the call setup are the same. For modeling sessions between non IMS end-points, the number of servers involved in the computation of the average response delay can be considered as a subset of the total number of servers present in the network.

Another assumption is that the servers in the network are independent and identically distributed (*i.i.d.*) that is each server has the same probability distribution and are mutually independent. An Erlang distribution can be used to model the penalty function for the end-to-end (tandem) tail probability.

Thus, the total average response time experienced in K tandem servers in the network is $\frac{K}{\mu-\lambda}$. The time period T is assumed to be a fragment of the total threshold time limit τ. So, the probability of the session setup over K *i.i.d.* servers being completed within threshold time limit τ while exceeding a time period T in each server is $K * e^{-(\mu-\lambda)T}$.

Based on the above assumptions, the following sections describe the optimization of the utility functions for the scenarios previously defined using a M/M/1/∞ queueing network.

4.5.1 Utility Function with total time for session setup

The optimal utility function is

$$\hat{U} = max_{\mu}\left[V - K(a\mu - b\mu^2) - \beta\left\{\tau - \frac{K}{\mu - \lambda}\right\}\right]$$

where β is the proportionality constant and τ is the total time limit for session establishment. The first order derivative condition to find the maximum μ is

$$\frac{\partial U}{\partial \mu} = -aK + 2bK\mu - \frac{K\beta}{(\mu - \lambda)^2} = 0 \tag{4.7}$$

We deduce some properties of the solution by applying the conjugate pair theorem.

1. If the cost a increases, the service rate μ decreases.

$$U_{\mu a} = -K \rightarrow \frac{\partial U^*}{\partial a} < 0$$

2. If the cost b increases, the service rate μ increases.

$$U_{\mu b} = 2K\mu \rightarrow \frac{\partial U^*}{\partial b} > 0$$

3. If penalty for losing requests β increases, the service rate μ decreases.

$$U_{\mu\beta} = -\frac{K}{(\mu - \lambda)^2} \rightarrow \frac{\partial U^*}{\partial \beta} < 0$$

   $U_{\mu\beta}$ is positive as $\mu > \lambda$ and $K > 0$.

4. If the arrival rate λ increases, the service rate μ increases.

$$U_{\mu\lambda} = -\frac{2K\beta}{(\mu - \lambda)^3} \rightarrow \frac{\partial U^*}{\partial \lambda} > 0$$

   We know that equation (4.7) is zero for the optimal μ. If $-K(a - 2b\mu) < 0$ in equation (4.7) then $[-K * \frac{\beta}{(\mu-\lambda)^2}] > 0$. From this we know that $U_{\mu\lambda} > 0$ as $U_{\mu\lambda} = [-K * \frac{\beta}{(\mu-\lambda)^2}] * [\frac{2}{\mu-\lambda}]$.

The above trends become intuitive if the constraint $\tau \geq \frac{K}{(\mu-\lambda)}$ is simplified to $\mu \geq \lambda + \frac{K}{\tau}$. We also notice that as the number of servers in the network K increases, the service rate μ will increase.

As a closed form for optimum $\hat{U}$ is difficult, we use $\mathcal{T} \geq \frac{K}{(\mu-\lambda)}$ as a hard constraint on $U = V(\lambda) - C$ and obtain optimum utility when $\mu^* = \frac{K}{\tau} + \lambda$ provided $(\lambda + \frac{K}{\tau}) < \frac{a}{2b}$ as in Figure 4.2. We use $\beta = 0.01$, the number of servers $K = 5$, $\tau = 30$ sec, the income $V = 5000$, the arrival rate $\lambda = 250$ arrivals/sec and obtain values for $a = 4$ and $b = 0.0043$ from [15].

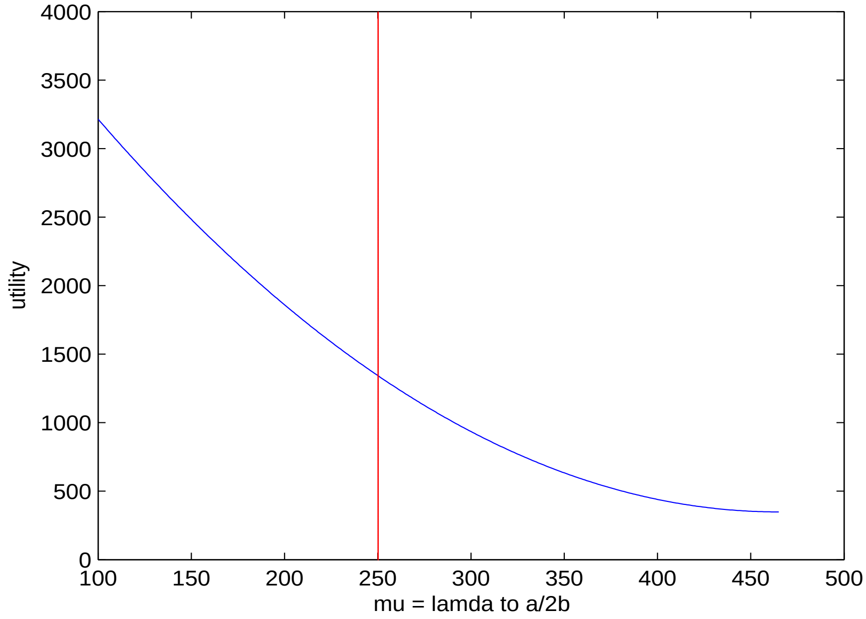

Figure 4.2: Utility vs. Service rate for M/M/1/∞ model, with constraint $\tau \geq \frac{K}{(\mu-\lambda)}$ where K=5 and τ=30

4.5.2 Utility Function with probability of session setup

For the second case, the optimal utility function is given as

$$\hat{U} = max_{\mu}\big[V - K(a\mu - b\mu^2) - \beta(\epsilon - K * e^{-(\mu-\lambda)T})\big] \tag{4.8}$$

Here again β is the proportionality constant, ϵ is the probability of completing the session setup within threshold time limit τ.

The first order derivative to compute the maximum μ is

$$\frac{\partial U}{\partial \mu} = -aK + 2bK\mu - KT\beta * e^{-(\mu-\lambda)T} = 0$$

We perform the following analysis of the solution based on the conjugate pair theorem:

1. As the cost a increases, the service rate μ decreases.

$$U_{\mu a} = -K \rightarrow \frac{\partial U^*}{\partial a} < 0$$

2. As the cost b increases, the service rate μ increases.

$$U_{\mu b} = 2K\mu \rightarrow \frac{\partial U^*}{\partial b} > 0$$

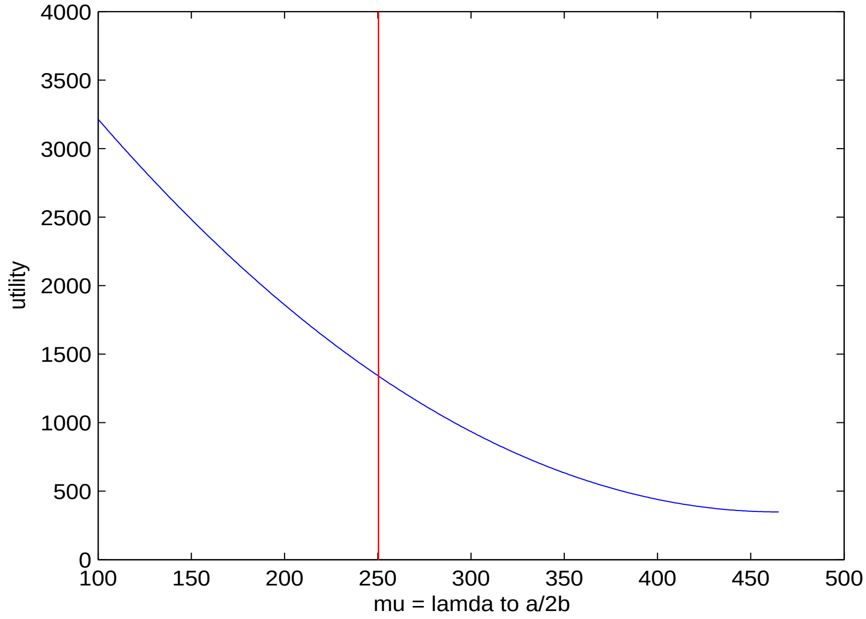

Figure 4.3: Utility vs. Service rate for M/M/1/∞ model, with constraint $K * e^{-(\mu-\lambda)T} \leq \epsilon$ where K=5 and ϵ=0.00005

3. As the penalty for losing requests β increases, the service rate μ decreases.

$$U_{\mu\beta} = -KT * e^{-(\mu-\lambda)T} \rightarrow \frac{\partial U^*}{\partial \beta} < 0$$

4. As the arrival rate λ increases, the service rate μ increases.

$$U_{\mu\lambda} = -KT^2\beta * e^{-(\mu-\lambda)T} \rightarrow \frac{\partial U^*}{\partial \lambda} > 0$$

The first order derivative of $\hat{U}$ is $-aK + 2bK\mu - KT\beta * e^{-(\mu-\lambda)T} = 0$ for optimal μ from equation (4.8). Now if $-K(a - 2b\mu) < 0$, then $[-KT\beta * e^{-(\mu-\lambda)T}] > 0$. That implies $[-KT\beta * e^{-(\mu-\lambda)T}] * T > 0$.

The constraint $K * e^{-(\mu-\lambda)T} \leq \epsilon$ can be simplified to $\mu \geq \lambda + \frac{\ln K}{T} - \frac{\ln \epsilon}{T}$. Upon using $K * e^{-(\mu-\lambda)T} \leq \epsilon$ as a hard constraint on $U = V(\lambda) - C$, we obtain $\mu^* = \lambda + \frac{\ln K}{T} - \frac{\ln \epsilon}{T}$ [Figure 4.3] provided $\left(\lambda + \frac{\ln K}{T} - \frac{\ln \epsilon}{T}\right) < \frac{a}{2b}$. From the above solution, we note that as K increases, the service rate μ will increase and as T increases, μ will decrease.

We use $a = 4$, $b = 0.0043$, $\beta = 0.01$, $\epsilon = 0.00005$, the number of servers $K = 5$, $T = 30$ sec, the income $V = 5000$ and the arrival rate $\lambda = 250$ arrivals/sec for computing the optimum $\hat{U}$ in Figure 4.3.

4.5.3 Utility Function with lost connections

In this last scenario, we have no constraints on the queueing system. A penalty of c is incurred for each session setup that exceeds the threshold time limit τ. The probability of the connection setup time experienced in each server exceeding the time period T for K servers is given by $K * e^{-(\mu-\lambda)T}$. Using Little's law, we have the number of lost connections as $\lambda * (Ke^{-(\mu-\lambda)T})$.

The optimal utility function is

$$\hat{U} = max_{\mu} \left[V - K(a\mu - b\mu^2) - c\lambda * Ke^{-(\mu-\lambda)T} \right] \tag{4.9}$$

Here c is the penalty for each lost connection.

The first order derivative for the maximum service rate μ is

$$\frac{\partial U}{\partial \mu} = -aK + 2bK\mu + KT\beta\lambda * e^{-(\mu-\lambda)T} = 0$$

We infer the following properties of the solution based on the conjugate pair theorem:

1. If the cost a increases, the service rate μ decreases.

$$U_{\mu a} = -K \rightarrow \frac{\partial U^*}{\partial a} < 0$$

2. If the cost b increases, the service rate μ increases.

$$U_{\mu b} = 2K\mu \rightarrow \frac{\partial U^*}{\partial b} > 0$$

3. If penalty for losing requests c increases, the service rate μ increases.

$$U_{\mu\beta} = KT\lambda * e^{-(\mu-\lambda)T} \rightarrow \frac{\partial U^*}{\partial \beta} > 0$$

4. If the arrival rate λ increases, the service rate μ increases.

$$U_{\mu\lambda} = KT\beta * e^{-(\mu-\lambda)T} + KT^2\beta\lambda * e^{-(\mu-\lambda)T} \rightarrow \frac{\partial U^*}{\partial \lambda} > 0$$

We observe that the trends in the network performance remain similar across different scenarios with different constraints. The service rate μ increases if the arrival rate λ is increased. As the cost of each server a increases the value of μ decreases while increasing the sales volume b causes μ to increase. The maximum service rate of the servers for optimal utilization of the network is heavily influenced by the arrival rate λ, the number of servers in the network K and the total time for the session establishment τ.

Using the above discussed methods, we can analyze and predict trends in different networks for different constraints and penalties. In the later sections, we present a methodology to characterize the workload of the servers in real networks to determine, even if approximately, the actual service distributions. This derived workload is used to simulate the end-to-end service to understand the dependence of utility function on the network parameters like the number of servers in the network.

Chapter 5

Characterization of SIP Workload

Capacity planning and performance modeling are essential to understand the complex integration of data and communication services. Questions like how will the service work, what are the bounds and cost of the service, how well will the service work are important in planning. Workload characterization of the servers provides information about the network performance. It tells us how the network will function, how the demand for services will be satisfied, how robust and stable the network will be, how fast will the connections be setup etc. Service providers must know the nature of the network and what services the network can offer to deploy service level agreements (SLAs). Thus, it is important to have a realistic estimate of the workload as it heavily influences the capabilities of the network.

We propose an innovative methodology to obtain an approximate workload of the servers. This study has been motivated by the lack of realistic workload characterization in the literature. Most often, servers are assumed to have constant service times. Such assumptions affect the capacity planning negatively. We performed case-studies of collecting real network measurements in order to determine, even if approximately, the actual service distributions of the servers. This derived workload is used in our simulations to model end-to-end service and identify the dependence of utility with network parameters. This is discussed in Chapter 6.

IMS network depends on the SIP server for the network throughput. SIP is an application-layer protocol. So a better implementation of SIP server with higher processor speed and more memory will provide better response times, better capacity and better capabilities. This will enhance the performance of the network significantly. Our focus is on the methodology rather than the actual numbers collected, as the numbers can change with different implementations over time and across different providers.

5.1 Proposed Characterization Methodology

In our methodology, we monitored real traffic to derive the workload characteristics of the server. Based on the observed traffic, we formulated an approximate service time distribution which provided insights into the network capabilites. We conducted experiments using different SIP clients and used Ethereal to monitor the SIP signaling messages. The approximate response times were deduced from the observed traffic and used to derive the workload.

Today most of the instant messengers like Yahoo, Gizmo, MSN use SIP signaling to provide voice chat. Skype uses SIP to connect to the gateway service for their SkypeIn and SkypeOut service. We performed our study on Yahoo's Voice chat messenger and Sipphone's Gizmo messenger. The study captures the differences between a multi-server and single server network. In what follows, we discuss the details of the methodology and provide inferences from the derived service distribution. We analyze SIP packet contents to obtain information about the session capabilities.

5.2 SIP Packet Flow

Figure 5.1 represents an INVITE request initiated by Yahoo messenger to establish a SIP voice chat. We notice from the traces that Yahoo messenger uses SIP signaling over TCP.

The capabilities of the connection can be obtained from analyzing the SIP and SDP packets. An unauthorized INVITE request receives a 401 Unauthorized response as shown in Figure 5.2. The 401 Unauthorized response includes a challenge in the `WWW-Authenticate` header field that the sender must answer.

In Yahoo, an SDP packet describing the session is sent only after the sender has been authorized. An authorized INVITE request is given in Figure 5.3. The lines with (m) describe the codecs supported for audio. Audio is received on port 8918. The (a)

```
Request-Line: INVITE sip:timetest345@206.190.50.163:5061;transport=tcp SIP/2.0
Message Header
  From: timetest123<sip:timetest123@206.190.50.163:5061>;
    tag=6861ba0-0-13bb-43617-47746061-43617
  To: <sip:timetest345@206.190.50.163:5061
  Call-ID: 1dc9758-0-13bb-43617-7dd4f3e9-43617@206.190.50.163
  CSeq: 1 INVITE
  User-Agent: Yahoo Voice,1.2
  Need-Token: Yes
  Max-Forwards: 70
  Via: SIP/2.0/TCP 10.10.10.2:5051;branch=z9hG4bK-43617-10734b35-2ed3d69f
  Contact: <sip:timetest123@127.0.0.1:5051;transport=tcp>
  Content-Length: 0
```

Figure 5.1: Unauthorized Yahoo SIP INVITE Request

```
Status-Line: SIP/2.0 401 Unauthorized
Message Header
 From: timetest345<sip:timetest345@206.190.50.163:5061>;
   tag=5d00690-0-13bb-12060-55d2f8bb-12060
 To: <sip:timetest123@206.190.50.163:5061>;
   tag=b7e98e44-1b9e-441b65a5-2c2be343-12da2ca4
 Call-ID: 5d0ad18-0-13bb-12060-2d2405ea-12060@206.190.50.163
 CSeq: 1 INVITE
 Token: PEPkR8tXc6nBY0rgrh2kqw
             --\002anHQBeImkiMfkSRhBbUA_68l0v3ZKC.J2qZJBPnGIZxvbfA-
 Y-Token: PEPkR8tXc6nBY0rgrh2kqw
             --\002anHQBeImkiMfkSRhBbUA_68l0v3ZKC.J2qZJBPnGIZxvbfA-
 Via: SIP/2.0/TCP 10.10.10.2:5051;branch=z9hG4bK-12060-46678cd-70054828
 WWW-Authenticate: Digest realm="sip.yahoo.com",
     nonce="PEPkR8tXc6nBY0rgrh2kqw
             --\002anHQBeImkiMfkSRhBbUA_68l0v3ZKC.J2qZJBPnGIZxvbfA-",
     algorithm=MD5
 Content-Length: 0
```

Figure 5.2: 401 Unauthorized Yahoo SIP Response

lines describe the media attribute conditions required for the connection to be setup. For example, `Media Attribute (a):sendrecv` indicates that the streams are bidirectional. We notice from the RTP packet that Yahoo messenger uses SPEEX codec as the media payload.

Gizmo uses UDP as the transport protocol and its INVITE request with session description is given in Figure 5.4.

Using the `Via` header field in the INVITE request received at the destination, we obtain the number of proxies the message has traversed through which is 1 in the case of

```
Request-Line: INVITE sip:timetest123@206.190.50.163:5061;transport=tcp SIP/2.0
Message Header
 From: timetest345<sip:timetest345@206.190.50.163:5061>;
  tag=5d00690-0-13bb-12060-55d2f8bb-12060
 To: <sip:timetest123@206.190.50.163:5061>
 Call-ID: 5d0ad18-0-13bb-12060-2d2405ea-12060@206.190.50.16
 CSeq: 2 INVITE
 User-Agent: Yahoo Voice,1.2
 Max-Forwards: 70
 Via: SIP/2.0/TCP 10.10.10.2:5051;branch=z9hG4bK-12060-4667935-64025691
 Contact: <sip:timetest345@127.0.0.1:5051;transport=tcp>
 Authorization: Digest username="timetest345",
     realm="sip.yahoo.com",
     nonce="PEPkR8tXc6nBY0rgrh2kqw
               --002anHQBeImkiMfkSRhBbUA_68l0v3ZKC.J2qZJBPnGIZxvbfA-",
     uri="sip:timetest123@206.190.50.163:5061;transport=tcp",
     response="ccc0a74c8601a794163ab
 Content-Type: application/SDP
 Content-Length: 463
Message body
 Session Description Protocol Version (v): 0
 Owner/Creator, Session Id (o): - 3351634981 3351634981 IN IP4 127.0.0.
 Session Name (s): Yahoo Voice,1.2
 Connection Information (c): IN IP4 10.10.10.2
 Bandwidth Information (b): AS:80000
 Time Description, active time (t): 0 0
 Media Description, name and address (m): audio 8918 RTP/AVP 100 98 0 8 97
 Media Attribute (a): rtpmap:100 SPEEX/16000
 Media Attribute (a): rtpmap:98 ILBC/8000
 Media Attribute (a): rtpmap:0 PCMU/8000
 Media Attribute (a): rtpmap:8 PCMA/8000
 Media Attribute (a): rtpmap:97 SPEEX/8000
 Media Attribute (a): sendrecv
```

Figure 5.3: Authorized Yahoo SIP/SDP INVITE Request

Gizmo connections. The `Allow` header field lists all the methods supported. We obtain from the `Content-Type` field that the body of the message describes the SDP session description.

The initial part of the message provides session level information, beyond which the media level information and the attribute details are provided. Here, we notice that the session allows audio on port 5004 with RTP/AVP. Codecs like ITU-T G.711 PCMU, ITU-T G.711 PCMA, GSM 06.10 and other media formats are supported. The media description line (m) also specifies the preference order of the codecs to be used. The media attribute line (a) specifies the preconditions that are required for the session of the desired capabilities to be established. Here the media attributes provide the acceptable rate of sampling for each codec. For example, `Media Attribute (a): rtpmap:3 GSM/8000` means GSM codecs sampled at the rate 8000 are only accepted by the session. QoS attributes can be satisfied using such preconditions.

```
Request-Line: INVITE sip:15479327867@10.10.10.3:64064 SIP/2.0
Message Header
  Record-Route: <sip:timetest123@198.65.166.131;ftag=3b215939;lr=on>
  Via: SIP/2.0/UDP 198.65.166.131;
     branch=z9hG4bK7e13.1a39dcd4.0
  Via: SIP/2.0/UDP 10.10.10.2:64064;
     branch=z9hG4bK-d87543-62667511706a245f-1--d87543-;rport=64064
  Max-Forwards: 69
  Contact: <sip:15479327867@10.10.10.2:64064>
  To: <sip:timetest123@proxy01.sipphone.com:5060>
  From: <sip:15479327867@proxy01.sipphone.com>;tag=3b215939
  Call-ID: bf2de14bab44a408@ZGV2ZXRzMS5lY2V3MmsubmNzdS5lZHU.
  CSeq: 1 INVITE
  Allow: INVITE, ACK, CANCEL, OPTIONS, BYE, REFER, INFO, NOTIFY, MESSAGE
  Content-Type: application/sdp
  User-Agent: WinGizmo (Gizmo-s2n1)
  Content-Length: 435
  CQBM: 248
  RemoteIP: 10.10.10.2
  P-hint: usrloc routed
Message body
  Session Description Protocol Version (v): 0
  Owner/Creator, Session Id (o): GizmoProject 1370765871 1 IN IP4 10.10.10.2
  Session Name (s): GizmoAudioSession
  Connection Information (c): IN IP4 10.10.10.2
  Time Description, active time (t): 0 0
  Media Description, name and address (m): audio 19491
       RTP/AVP 97 103 100 101 0 8 102 3 106 117 13
  Media Attribute (a): rtpmap:97 IPCMWB/16000
  Media Attribute (a): rtpmap:103 ISAC/16000
  Media Attribute (a): rtpmap:100 EG711U/8000
  Media Attribute (a): rtpmap:101 EG711A/8000
  Media Attribute (a): rtpmap:0 PCMU/8000
  Media Attribute (a): rtpmap:8 PCMA/8000
  Media Attribute (a): rtpmap:102 iLBC/8000
  Media Attribute (a): rtpmap:3 GSM/8000
  Media Attribute (a): rtpmap:106 telephone-event/8000
  Media Attribute (a): rtpmap:117 red/8000
  Media Attribute (a): rtpmap:13 CN/8000
```

Figure 5.4: SIP/SDP Gizmo INVITE Request

We notice that Gizmo lacks support to G.726 and G.729 codecs. This might influence the service on low-bandwidth SIP sessions and PSTN calls. Codec 103 i.e. ISAC/16000 is a high quality codec which has been given high preference. This explains the reason for superior voice quality in Gizmo. Such inferences based on the SIP session can easily be used in the deployment of SLAs by the service providers.

We notice from the traces that Gizmo uses the ISAC codec for its RTP packet payload as in Figure 5.5. RTCP packets including senders report and source descriptors (SDES) are sent between the source and destination as shown in Figure 5.6. They describe the amount of data sent so far and the RTP sampling timestamp etc.

For multimedia service, QoS satisfaction is critical. End-to-end QoS monitoring

```
Stream setup by SDP (frame 53)
 10.. .... = Version: RFC 1889 Version (2)
 ..0. .... = Padding: False
 ...0 .... = Extension: False
 .... 0000 = Contributing source identifiers count: 0
 1... .... = Marker: True
 .110 0111 = Payload type: ISAC (103)
 Sequence number: 11017
 Timestamp: 2353220837
 Synchronization Source identifier: 3075819743
 Payload: 8D76A518EC0763A679AD20579644972AA212186BBC7F2923...
```

Figure 5.5: Gizmo RTP packet

can be performed by verifying the session description provided by SDP. RTCP mesages provide QoS feedback to the sender and can be used to assure consistent quality.

```
Stream setup by SDP (frame 53)
10.. .... = Version: RFC 1889 Version (2)
..0. .... = Padding: False
...0 0001 = Reception report count: 1
Packet type: Sender Report (200)
Length: 12
Sender SSRC: 3075819743
Timestamp, MSW: 3351563746
Timestamp, LSW: 2070646683
RTP timestamp: 2353253153
Sender's packet count: 31
Sender's octet count: 3085
Source 1
  Identifier: 459758815
  SSRC contents
     Fraction lost: 0 / 256
     Cumulative number of packets lost: 0
  Extended highest sequence number received: 15176
     Sequence number cycles count: 0
     Highest sequence number received: 15176
  Interarrival jitter: 2629678
  Last SR timestamp: 529
  Delay since last SR timestamp: 3454171482
```

Figure 5.6: Gizmo RTCP Sender's Report

5.3 Yahoo Case Study

Yahoo Messenger's Voice chat is provided using SIP. We conducted experiments to collect the SIP signaling messages. We used this data to deduce an appropriate service distribution pattern for Yahoo servers. We compared our findings with the norms specified by the ITU recommendation E.721 [17]. Again, we emphasize here on the general methodology, rather than the actual numbers collected, as they are likely to change with advancements in server technology.

5.3.1 Experimental Setup

All experiments were performed using the latest version of Yahoo Messenger with voice. The messenger was installed on two Windows machines. One machine was a Pentium III, 930 MHz with 256 MB RAM running Windows 2000, and the other machine was a Pentium 4, 1.48 GHz with 256 MB RAM with Windows XP. Each machine had a 10/100 Mb/s Ethernet card and was connected to a 100 Mb/s network. Both machines were with public IP addresses. Ethereal [16] was used to monitor network traffic. All experiments were performed between August and October, 2005.

5.3.2 Methodology

The calls were established between the two machines and the Ethereal traces were collected to derive the service times of the SIP servers. A SIP connection's setup typically comprises of INVITE, 100 Trying, 180 Ringing, 200 OK, and ACK messages [Figure 5.7].

Figure 5.8 represents the SIP connection between source and destination using the Yahoo messenger. The Yahoo proxy 1 and proxy 2 are part of the Yahoo cloud. The messages involved in a session have been labeled chronologically from A to T. The round trip times from source to proxy 1 and from destination to proxy 2 were recorded using the ping command. The corresponding propagation delays were subtracted from the Ethereal traces to provide the response time of the servers. A similar approach is used in the analysis of the peer to peer protocol Skype in [3].

We characterized the SIP server workload based on the response times for unauthorized INVITE, authorized INVITE, BYE, INFO and ACK. The server responded with 100 Trying followed by a 401 Unauthorized for an unauthorized INVITE request. 100 Trying followed by a 180 Ringing were the responses for an authorized INVITE request. Both BYE and INFO requests are responded with 200 OK messages. The ACK request does not have any response and is used for reliability. We propose the following steps to determine the response times.

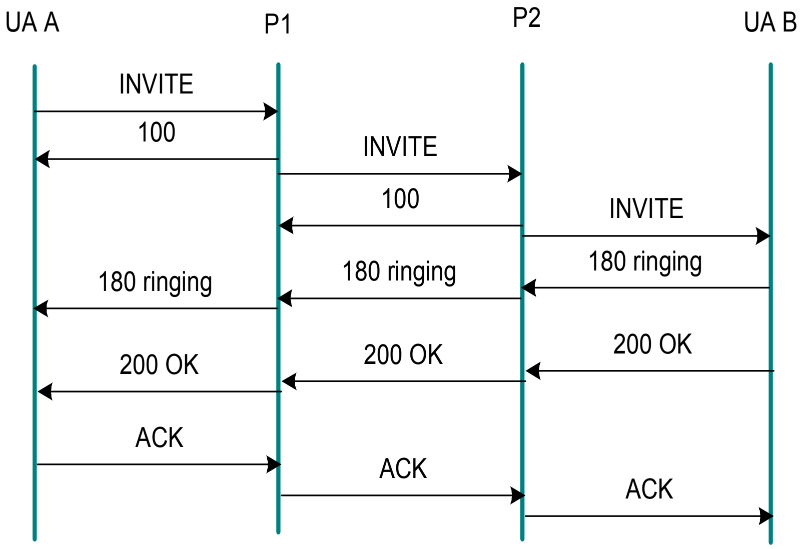

Figure 5.7: Typical SIP call flow setup

- The response time for an unauthorized invite is computed by subtracting A from C.
- The authorized invite message's response time is measured as the time from the invite message E till the time for the 180 ringing J.
- The response time for the bye message is the time difference between Q and T.
- The info message has a response time measured from O to P.
- The total setup time for the connection is the time taken from the first INVITE request A to the ACK message M.

The time spent in the Yahoo cloud is the cumulative time spent in all the SIP servers in the network is computed from Figure 5.8. All computations were performed on the values obtained at the source to avoid the complications of determining the clock differences between the two end-points.

We monitored the SIP traffic over 1000 calls and collected their response times. The total setup time had an average of 2.61 seconds which satisfies the ITU recommendation E.721 [17]. We fitted the calculated values to different distributions to obtain the server workload. We also derived the distribution for the total SIP setup time. Arena's Input

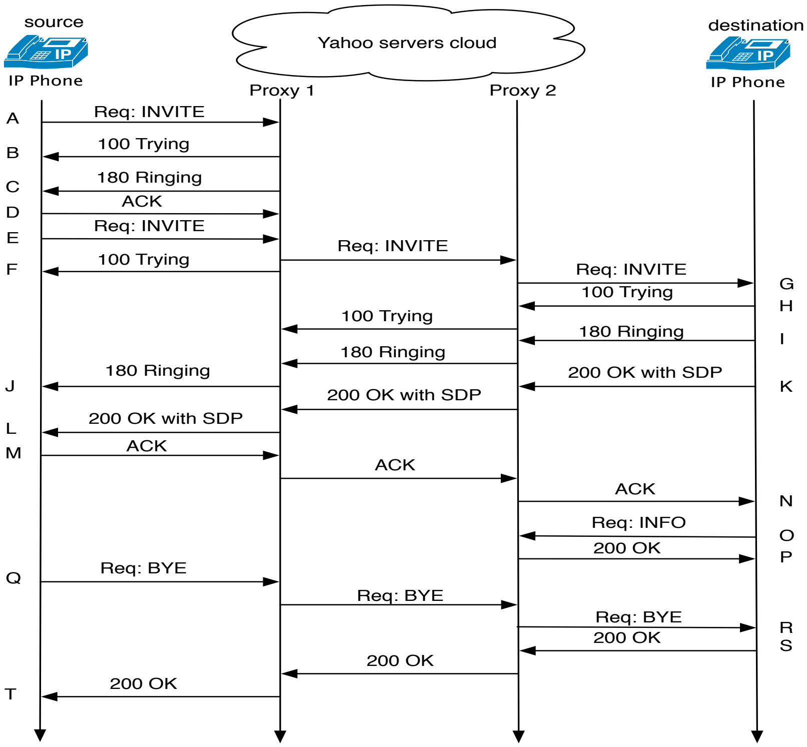

Figure 5.8: Yahoo call flow

Analyzer [21] was used to analyze the collected data. A lognormal distribution with the mean of 0.0261 and a standard deviation of 0.0476 was found to have a better fit from the mean square error value [Figure 5.9].

The square error is a measure of the quality of the distribution's match to the data. The square error value is the average of the squares of the differences between the relative frequencies of the observations in a range and the relative frequency for the fitted distribution function over that data range. A larger square error value implies that the fitted distribution is a poor match to the actual data. The square error values for the collected data are given in Table 5.1. Based on the square error values, we choose the distribution with the least square error value as the best fit.

Arena also performs Chi-square and Kolmogorov-Smirnov (K-S) goodness-of-fit hypothesis tests [25]. These are standard distribution tests that can be used to assess whether the fitted theoritical distribution is a good fit to the data. The K-S and Chi square

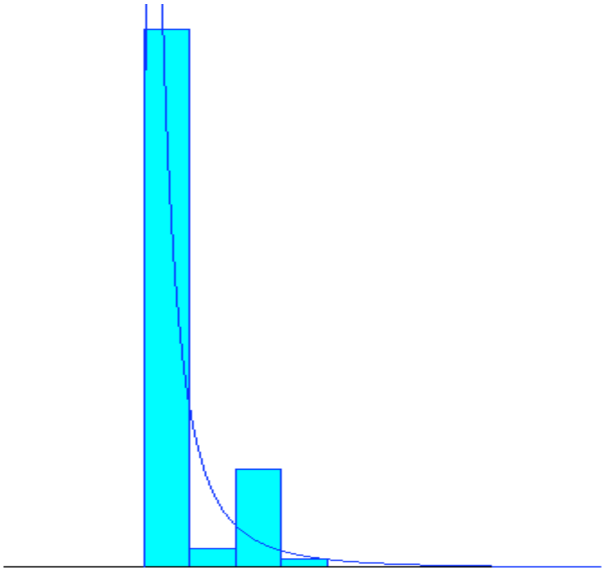

Figure 5.9: Yahoo SIP server fitted distribution

goodness-of-fit hypothesis test values for the collected data are given in Table 5.2. Larger p-values indicate a better fit.

As with any statistical hypothesis test, a high p-value does not constitute as proof of a good fit, just the lack of evidence against the fit. Its popularly agreed upon that there is no exact or precise approach to fit or choose the best fitting distribution. Different statistical tests like K-S, Chi Square, Anderson - Darling tests etc. might rank distributions differently. Changes in the presentation of data like changing the number of histogram cells, specifying bounds on the data might reorder the distributions. We explored SAS statistical software and Matlab for the analysis of the collected data. Neither tools could provide any consistent or additional insights about the characteristics of the measured data. Our emphasis is on the methodology to derive the workload of the server rather than the fit. Accuracy of the fit can be further investigated in future work.

Therefore, in this thesis, the workload of the Yahoo server is modeled as a lognormal distribution with a mean of 0.0261 and a standard deviation of 0.0476. The distribution had -4.3 and 1.2 as its shape parameters. A lognormal distribution is defined as

$$f(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{(ln(x)-\rho)^2}{2\sigma^2}}$$

Table 5.1: Square Error Values for Yahoo SIP server workload

Function	**Sq Error**
Lognormal	0.0211
Beta	0.0252
Exponential	0.0385
Erlang	0.0385
Gamma	0.0403
Normal	0.284
Triangular	0.634
Uniform	0.654
Weibull	1.01

Table 5.2: Goodness of fit tests for Yahoo SIP Server Workload

	Chi Square Test	**Kolmogorov-Smirnov Test**
Number of intervals	9	
Degrees of freedom	6	
Test Statistic	1.79e+003	0.209
Corresponding p-value	< 0.005	< 0.01

Based on the collected values, we characterize the SIP server workload as

$$f(x) = \frac{0.33}{x} e^{-0.34[ln(x)+4.3]^2}$$

Using Arena's input analyzer, the distribution for the total setup time was obtained. The total setup time distribution is the sum of 2.16 and a lognormal distribution with a mean of 0.444 and a standard deviation of 0.105. The shape parameters of the lognormal distribution are -0.83 and 0.23. Thus, the distribution of the total setup time is

$$f(x) = 2.16 + \frac{1}{0.57x} e^{\frac{-[ln(x)+0.83]^2}{0.10}}$$

This derived distribution of the session setup time is used in the simulations and is discussed in Chapter 6.

5.4 Gizmo Case Study

Gizmo Messenger is a SIP based internet phone with instant messaging offered by Sipphone. We collected SIP signaling messages using the network monitoring tool "Ethereal". The experiments were performed with the same setup used for Yahoo messenger as discussed in the previous section. The latest version of Gizmo software was installed on two Windows machines. All data was collected between October and December 2005.

5.4.1 Methodology

The calls were established between the two machines and the Ethereal traces were collated. Gizmo supports all its SIP calls using a single proxy server provided by Sipphone. A SIP call flow using a single SIP proxy server is described in Figure 5.10.

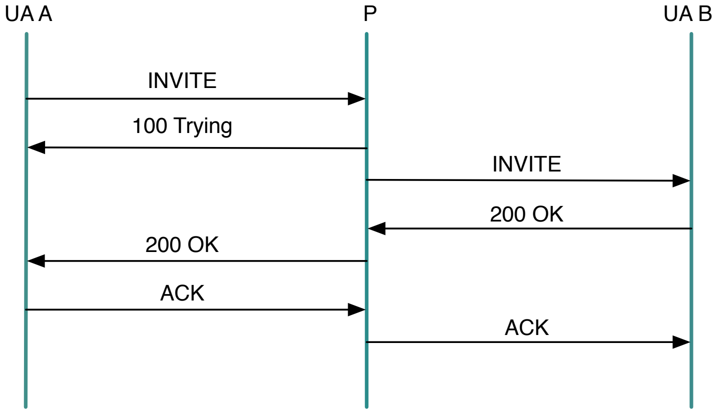

Figure 5.10: SIP call flow with single proxy

Figure 5.11 represents the call flow of a SIP connection between source and destination using the Gizmo messenger. The messages in the session have been labeled chronologically from A to M. The round trip times from source to proxy and from destination to proxy were obtained using the ping command. The corresponding propagation delays are deducted from the Ethereal traces to provide the response time of the server.

The SIP workload is derived from the response times for INVITE, BYE and ACK. The server responded with a 100 Trying message followed by a 180 Ringing for an INVITE request while a 200 OK message is received for the BYE request. The total setup time is the time difference from the INVITE request A to the ACK message M as shown in Figure 5.11. Clock differences between the end-points can cause complications in the computations. We avoided this by determining response times using the time values obtained on the same

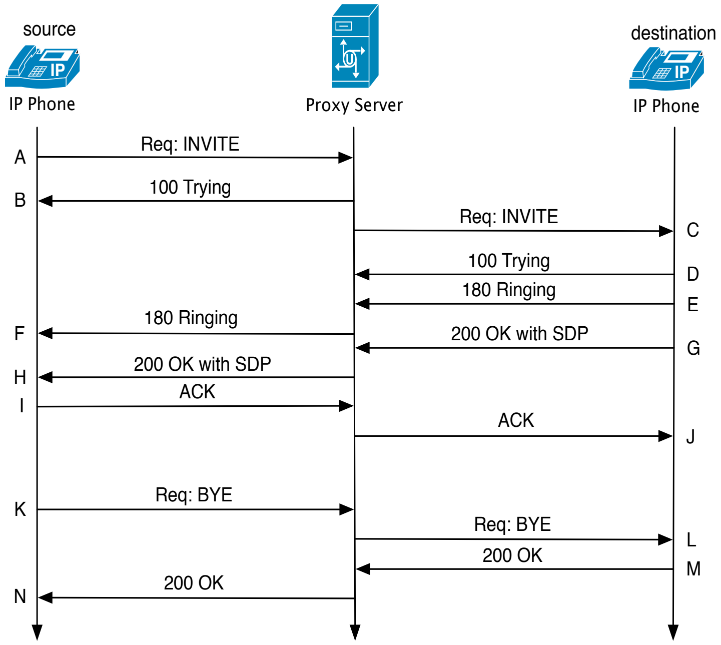

Figure 5.11: Gizmo call flow setup

end-point. The following steps are used to determine the Gizmo server workload :

- The BYE request received the response 200 OK within the time difference from J to M.

- The response times of a BYE message and its response is determined by adding the time difference between J and K with the time difference between L and M.

- The response time of the OK message and the ACK message is the obtained by adding the time difference between F and G with the time difference between H and I.

- The response time for an INVITE request is measured from A to B.

- The response time for the responses of the INVITE message is determined by adding the time difference from B to C with time difference from D to E.

The propagation delays have been subtracted from all time values used in the computations to obtain a realistic response time of the message at the server.

We monitored the SIP traffic over 1000 calls and collected their response times. The total setup time had an average of 6.68 seconds. Arena's Input Analyzer was used to fit the collected values to a distribution as described in Section 5.3. A scaled beta distribution with shape parameters 0.846 and 14.5 was found to have a better fit from the square error values [Figure 5.12]. The square error values are given in Table 5.3.

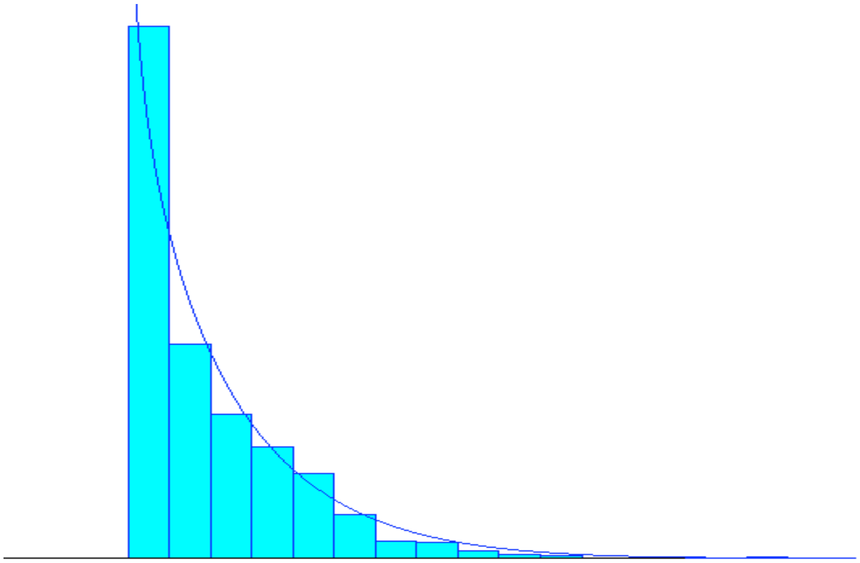

Figure 5.12: Gizmo SIP server fitted distribution

Table 5.3: Square Error Values for Gizmo SIP server workload

Function	**Sq Error**
Beta	0.00538
Erlang	0.0104
Exponential	0.0104
Lognormal	0.0326
Gamma	0.0398
Weibull	0.0488
Normal	0.107
Triangular	0.212
Uniform	0.238

The K-S and Chi square goodness-of-fit hypothesis test values for the collected data are given in Table 5.4. As discussed in the previous section, our focus is on the methodology and not on the accuracy of the fit.

Table 5.4: Goodness of fit tests for Gizmo SIP Server Workload

	Chi Square Test	**Kolmogorov-Smirnov Test**
Number of intervals	28	
Degrees of freedom	25	
Test Statistic	379	0.1
Corresponding p-value	< 0.005	< 0.01

The workload of the Gizmo SIP server is approximated to a scaled beta distribution. The Beta distribution is defined as

$$f(x) = \frac{x^{\beta-1} * (1-x)^{\alpha-1}}{B(\beta,\alpha)}$$

where $B(\beta,\alpha)$ is a beta function, defined by

$$B(\beta,\alpha) = \int_0^1 t^{\beta-1}(1-t)^{\alpha-1}\,dt$$

Based on the collected values, we characterize the SIP server workload as

$$f(x) = 4 * \frac{x^{0.846-1} * (1-x)^{14.5-1}}{B(0.846, 14.5)}$$

Using Arena's input analyzer, the distribution for the total setup time was obtained. The total setup time distribution is the sum of 1 and a lognormal distribution with a mean of 6.43 and a standard deviation of 12.4. The shape parameters of the lognormal distribution are 1.08 and 1.24. Thus, the distribution of the total setup time in Gizmo is modeled as:

$$f(x) = 1 + \frac{1}{3.12x} e^{\frac{-[ln(x)-1.08]^2}{3.10}}$$

This distribution is used in the simulations and is discussed in Chapter 6.

5.5 Inferences

We have derived the SIP server workload for both the Yahoo messenger and Gizmo Project. We notice from the traces that Yahoo messenger permits only authorized users to establish SIP sessions while Gizmo does not require authentication. We also observe from the SDP messages that Gizmo supports more audio codecs than Yahoo messenger.

We note from collected data that the setup time distributions for both Yahoo and Gizmo messengers are lognormal distributions which are heavy tail in nature. Complex networks have unexpected properties and irregularities which leads to heavy tails in the servers in the network. Heavy tails imply large level of fluctuations in the standard deviation of the degree of distribution and are limited only by finite size of the systems.

Complex and dynamic network interactions requires the network to be robust causing heavy tails. We observe from the collected data that the total average setup time in Yahoo is 2.61 seconds while the total setup time in Gizmo is around 6.68 seconds.

As known from the P-K mean formula, the steady state average delay d in a M/G/1 system [23] is affected by both the variance $Var(s)$ and the expected mean service time $E(s)$ as in (5.1).

$$d = \frac{\lambda\{Var(s) + [E(s)]^2\}}{2[1 - \lambda E(s)]} \tag{5.1}$$

Specifically, the Yahoo SIP server has a lognormal service time with $E(s) = 0.026$ and standard deviation $\sqrt{Var(s)} = 0.047$. So the variance of Yahoo SIP server is 0.002. The Gizmo server has a scaled beta distribution with shape parameters 0.846 and 14.5. The variance and mean of a beta distribution are

$$Var[x] = \frac{\beta\alpha}{(\beta + \alpha)^2(\beta + \alpha + 1)}$$

$$E(s) = \frac{\beta}{\beta + \alpha}$$

The variance and mean of a scaled distribution are $Var[Cx] = C^2 Var[x]$ and $E[Cx] = CE[x]$. This yields the Gizmo server a variance of 0.051 and a mean of 0.22.

From this analysis, we note that the Gizmo server in comparison to the Yahoo server has a higher variance and a higher mean. A large mean implies that the congestion will be large. High variances have larger positive values, so the server will be tied up for a long time causing the queue to build up. Thus, the Gizmo messenger having only a single SIP server to handle all the SIP connections will develop a large queue soon and get congested. This explains the reason for such a large total setup time observed in Gizmo. So, if Gizmo employs more servers, the total setup time will be reduced as the network load will be distributed between the servers. This load balancing will reduce congestion and offer more availability.

The steady state delay for the Gizmo server is calculated as 0.064 from (5.1) while the steady state delay for Yahoo server is 0.001. A high delay implies more congestion and large queueing delays. Hence we infer that the Gizmo servers get congested faster than the Yahoo servers. We can safely summarize that networks having servers with smaller variances, means and steady state delays are more stable and can establish connections quickly.

It is known that the average total time spent in a system is the sum of the average time spent in service $E(s)$ and the average time spent in queue d.

$$W = d + E(s)$$

Thus, the total time spent by a message in a Gizmo server is 0.284 seconds and around 0.028 seconds in a Yahoo server.

From our models and measurements, we propose a first order planning formula,

$$K \leq \frac{T_{SLA}}{W}$$

where K is the number of servers, T_{SLA} is the threshold time to complete a connection and W is the the total time spent in the server. Using the first order planning formula, we obtain the upper bounds for the number of servers in the network for a given T_{SLA}.

In this chapter, we proposed a methodology to characterize the workload of SIP servers. We understood the qualities of the server by analyzing its service time distribution. A practical method for end-to-end QoS monitoring in the IMS network was demonstrated. In what follows, we examine the dependencies of the utility function on the number of servers in the network.

Chapter 6

Simulation of SIP Servers

In today's highly competitive environment, offering new services with seamless transition among disparate network topologies and flexibility to deploy new technologies requires a right service model. Service providers need a model to quickly introduce new services and exploit the full revenue potential of those new services. We proposed a utility based service model for IMS which can help in planning and performance prediction in Chapter 4.

The network utilization and revenue generation are heavily influenced by the service models. Thus, it is important to understand the internal dependencies of the service models and how variations in the parameters will effect the overall utility. We presented a first order planning formula in last chapter to provision some maximum bounds on the number of servers based on the T_{SLA}. We now investigate the sensitivity and dependence of the utility with respect to the number of severs in the network using simulation. Here, we utilize the total setup time distribution derived from real traffic in our simulations and examine the tendencies displayed by the utility model while varying the number of servers.

The simulations were performed using Arena. We use the setup time distribution of both Yahoo servers and Gizmo servers and have two scenarios for the simulation. The input traffic is varied in both the scenarios to observe the sensitivity of the other parameters to the input traffic. The setup times for the connections are determined from the simulation and are used to calculate their respective utility functions.

Scenario I

We model the simulation to represent SIP call sessions between the source and destination routed through the service provider's network [Figure 6.1]. The simulation model for this scenario is in Figure 6.2. The sessions to the network are initiated as a Poisson arrival process with a mean arrival rate of 1 session per second. The network follows Yahoo's total setup time distribution to complete a session, i.e.,

$$f(x) = 2.16 + \frac{1}{0.57x} e^{\frac{-[ln(x)+0.83]^2}{0.10}}$$

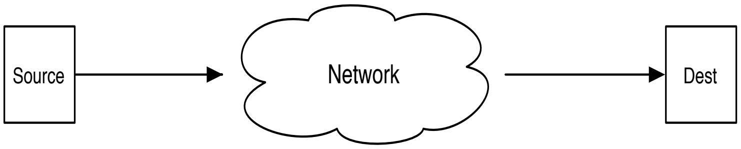

Figure 6.1: SIP network to simulate

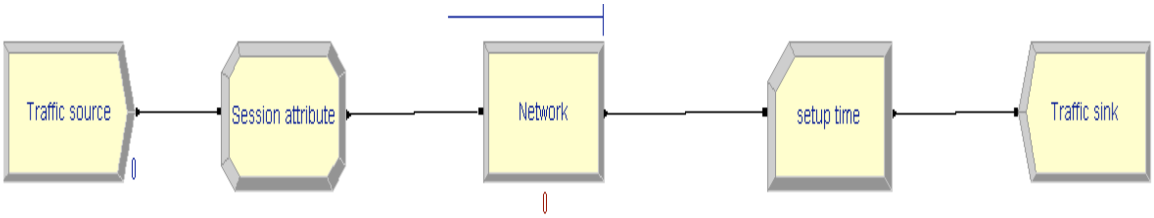

Figure 6.2: Simulation setup

The number of servers in the network is varied and the setup times for the connections are computed. The simulation is run for 5 hours each time the number of servers is changed.

The average session setup values obtained from the simulation for each case are in Table 6.1(a). We note from the values in Table 6.1(a) that variation in the number of servers influences the time consumed in the setup. The setup time did not change significantly for the given traffic once the number of servers approached 11. The utility function is calculated

Table 6.1: Simulated Setup Times

(a) Yahoo Setup Times

Total setup time (sec)	**Number of Servers**
5581.58	1
3959.71	2
5.08	3
2.93	4
2.69	5
2.62	6
2.61	7
2.605	8
2.604	9
2.6039	10
2.6038	11
2.6038	12
2.6038	13
2.6038	14
2.6038	15
2.6038	16
2.6038	17
2.6038	18
2.6038	19
2.6038	20

(b) Gizmo Setup Times

Total setup time (sec)	**Number of Servers**
7689.2	1
6583.9	2
5155.5	3
4286.82	4
2939.38	5
1738.79	6
621.16	7
21.7	8
10.60	9
8.8	10
7.86	11
7.60	12
7.54	13
7.42	14
7.41	15
7.39	16
7.39	17
7.39	18
7.39	19
7.39	20

for each simulated setup time. Figure 6.3 represents the dependence of the utility function on the number of servers. The utility decreases with the increase in the number of servers and is in an optimal range when there are favorable number of servers in the network.

The same simulation is repeated using the setup time distribution of Gizmo server, i.e.,

$$f(x) = 1 + \frac{1}{3.12x} e^{\frac{-[ln(x)-1.08]^2}{3.10}}$$

Table 6.1(b) represents the average setup times using Gizmo. In Gizmo results, we notice the same trend as seen in Yahoo. The setup time did not change significantly after the number of servers reached around 16.

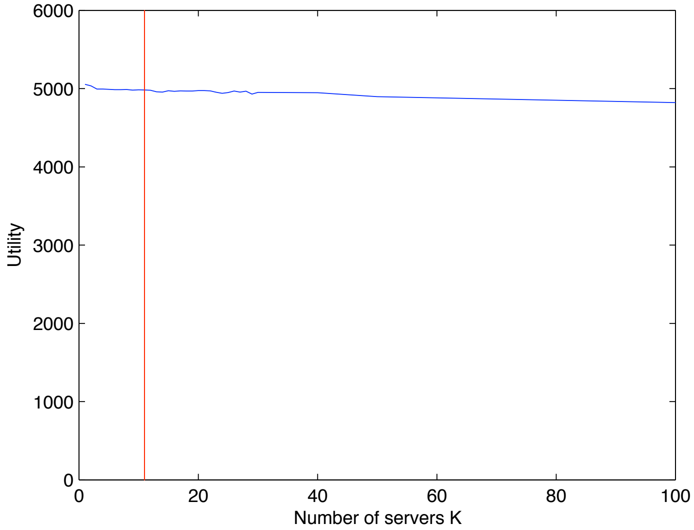

Figure 6.3: Utility vs Number of Servers

Scenario II

In this simulation, the sessions to the network are initiated as a Poisson arrival process with a mean arrival rate of 2 sessions per second. The network follows Yahoo's total setup time distribution to complete a session, i.e.,

$$f(x) = 2.16 + \frac{1}{0.57x} e^{\frac{-[ln(x)+0.83]^2}{0.10}}$$

Figure 6.2 represents the simulation model. The simulation is run for 5 hours each time the number of servers are changed and the average total setup time is determined. The setup times calculated using Yahoo are in Table 6.2.

From Table 6.2, we observe that the setup time does not change much once the number of servers reaches 15.

Results and Inferences

From the above simulations, we infer that the setup times for the connections do not vary once a favorable number of servers have been reached for the given traffic. This trend remains consistent across different servers with different input traffic. Such favorable values for the number of servers are useful in capacity planning. The simulation helps to

Table 6.2: Simulated Yahoo Setup Times

Total setup time (sec)	**Number of Servers**
14581.8	1
11408	2
774.29	5
2.949	8
2.693	10
2.631	12
2.609	15
2.605	18
2.604	20
2.604	21
2.604	22
2.604	25

identify the dependency between utility and other network parameters such as the number of servers. It provides insights on the number of servers in the network for effective completion of setups within the given target times for a given input traffic.

Chapter 7

Conclusion

The evolution of IMS has provided new multimedia rich communication services by collaborating telecom with data on access independent IP core network. This convergence of data, voice and video has increased the demand for services with new performance characteristics. The expectations of the consumers from the service providers has increased with the emergence of new technologies. In our research, we have focused on analyzing this network migration from a service provider's perspective.

As service providers deploy their IMS networks, they are moving to a service model that has the ability to introduce new services while guaranteeing maximum revenues. We introduced a service model for IMS in Chapter 4. In our modeling, the entire IMS network was viewed as an open feed forward tandem queueing network. Components in the network were typically modeled as a M/G/1 system. Classical queueing theory was used to obtain insights on the performance of the network. We illustrated formal modeling methods to capture the end-to-end SIP signaling delay during which QoS for the session was negotiated.

We defined the service model for IMS network with utility functions and optimized the desired properties like the service rate to achieve maximum revenue generation. This technique of performing economic-theoretic evaluation simplifies the mathematical computations without reducing the qualitative applicability of results. The service models with utility functions served as decentralized control mechanisms to optimize the network properties while studying revenue generation trends in networks.

We demonstrated variations in the modeling by introducing constraints in the network parameters that the service provider might wish to apply. We evaluated the network performance based on a model under the constraint on the total time to establish the session exceeding the total average response delay experienced in the servers. The network performance was discussed under the constraint of the probability of the session setup being completed within a threshold time provided by the service provider in the service agreements exceeding the probability of the average response delay in the servers. We considered the scenario of having penalties for each session that was unable to be completed within the threshold time and analyzed the trends in the network performance.

We observed that the performance trends of the network remain consistent across different scenarios under different constraints. The maximum service rate of the servers for optimal utilization of the network was heavily influenced by the service rate, the number of servers in the network and the total time for the connection establishment. Such formal methods of predicting the performance trends in the network can help service providers in capacity planning and to strategize the QoS guarantees of the network.

We presented a methodology to characterize the workload of SIP servers in the network in Chapter 5. Case-studies using popular publicly accessible SIP clients were conducted to characterize the server workload, even if approximately. We obtained a lognormal distribution as the SIP server workload from data collected over Yahoo SIP servers and a scaled beta workload for the Gizmo servers. The analysis of this realistic server service distribution provided insights into the structure of SIP servers. This characterization can be inputted into network testing benchmarks and utilized for predicting network performance.

IMS uses SIP for establishing connections between users. A practical method for end-to-end QoS monitoring in the IMS network was demonstrated. We explored SIP signaling messages used by popular voice chat messengers and analyzed the session capabilities through analysis of SDP and RTCP messages.

We performed capacity planning by providing a first order planning formula to determine the maximum bounds on the number of the servers in the network for a given service agrrement. A favorable number of servers with optimized utilization was obtained through simulation and the dependency of the utility function on the number of servers in the network was examined. This approach to evaluate the network is extremely useful for planning and efficiently structuring the network resources.

In this thesis, we presented queueing models for the signaling part of IMS networks. We demonstrated the maximization of generic utility functions by optimizing the server service rate and analyzed the trends of network performance. An approach to characterize the workload of a SIP server that can be combined with the modeling and optimization procedures was described.

Deployment of IMS architecture requires a rigorous change in the service provider's existing network infrastructure. Service providers need to monitor and optimize the network performance in order to meet the guarantee levels of service promised to their customers. Service providers must model and dimension their network to evaluate and predict trends in the network before actually deploying it. We have shown in our thesis an approach of theoretical analysis combined with realistic performance characterization that offer service providers incremental steps for a reasonable deployment of IMS with return of investment.

We have reported some of the ideas presented in this dissertation in the following publications:

- Nisha Rajagopal, Michael Devetsikiotis, "Modeling and Optimization for the Design of IMS Networks", to be published in the Proceedings of the 39th Annual Simulation Symposium, 2006 (ANSS-39 2006).

- Vladica Stanisic, Nisha Rajagopal, Yannis Viniotis, Michael Devetsikiotis, Dan Cude, "Engineering Rules for IMS Signaling Networks", OPNETWORK 2005 Conference, Washington D.C, August 2005.

Future work can focus on further data collection from different SIP servers and validation of performance evaluation trends. Enhancements can be performed by verification of the accuracy of the derived service distribution. Additional work includes introduction of dynamic and robust pricing schemes for IMS networks based on the performance models.

Bibliography

[1] 3GPP TS 23.221 V5.4.0 (2002-03). Architectural Requirements (Release 5).

[2] N. Banerjee, W. Wu, K. Basu, and S. K. Das. Analysis of SIP-based mobility management in 4G wireless networks. *Computer Communications*, 27(8):697–707, 2004.

[3] S. A. Baset and H. Schulzrinne. An analysis of the skype peer-to-peer internel telephony protocol, 2004.

[4] G. Camarillo and M. A. Garcia-Martin. *The 3G IP Multimedia Subsystem (IMS) : Merging the Internet and the Cellular Worlds*. John Wiley and Sons, August 2004.

[5] H. Chebbo and M. Wilson. Traffic and load modelling of an IP mobile network. In *4th International Conference on 3G Mobile Communication Technologies*, 2003.

[6] International Engineering Consortium. Performance management for next-generation networks.

[7] C. Courcoubetis and R. Weber. *Pricing Communication Networks: Economics, Technology and Modelling*. John Wiley and Sons, 2003.

[8] I. D. D. Curcio and M. Lundan. SIP call setup delay in 3G networks. In *Seventh International Symposium on Computers and Communications*, 2002.

[9] K. M. Currier. *Comparative Statics Analysis in Economics*. World Scientific Publishing Company, August 2000.

[10] T. Eyers and H. Schulzrinne. Predicting internet telephony call setup delay. In *IPTel 2000, First IP Telephony Workshop*, Berlin, Germany, 2000.

[11] H. Fathi, S. Chakraborty, and R. Prasad. Optimization of VoIP session setup delay over wireless links using SIP. In *Globecom*, 2004.

[12] S. S. Gokhale and R. E. Mullen. From test count to code coverage using the lognormal failure rate. In *15th International Symposium on Software Reliability Engineering*, 2004.

[13] V. K. Gurbani, L. Jagadeesan, and V. B. Mendiratta. Characterizing session initiation protocol (SIP) network performance and reliability. In *International Service Availability Symposium*, April 2005.

[14] O. Haase, K. Murakami, and T.F. La Porta. Unified mobility manager - enabling efficient SIP/UMTS mobile network control. *IEEE Wireless Communications Magazine*, Aug. 2003.

[15] K. Hosanagar, R. Krishnan, M. Smith, and J. Chuang. Optimal pricing of content delivery network (CDN) services. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 7*, page 70205.1, Washington, DC, USA, 2004.

[16] http://www.ethereal.com/. Ethereal: A network protocol analyzer.

[17] ITU-T. Network grade of service parameters and target values for circuit-switched services in the evolving ISDN. Recommendation E.721, 1992.

[18] ITU-T. General overview of NGN. Recommendation Y.2001, Dec 2004.

[19] W. Jiang and H. Schulzrinne. Assessment of VoIP service availability in the current internet, 2003.

[20] Y. Jiang, C. K. Tham, and C. C. Ko. Challenges and approaches in providing QoS monitoring. In *International Journal of Network Management*, 2000.

[21] D. Kelton, R. Sadowski, and D. Sturrock. *Simulation with ARENA*. McGraw Hill, 2003.

[22] A.A. Kist and R.J. Harris. SIP signalling delay in 3GPP. In *Proceedings of Sixth International Symposium on Communications Interworking of IFIP - Interworking 2002*, Perth, Australia, October 13-16 2002.

[23] L. Kleinrock. *Queueing Systems, Vol. 1: Theory.* Wiley Interscience, New York, 1975.

[24] V. Y. H. Kueh, R. Tafazolli, and B. G. Evans. SIP-based session establishment over an integrated satellite-terrestrial 3G network.

[25] A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis.* McGraw Hill, 2000.

[26] C. S. Lee and D. Knight. Realization of the next-generation network. *IEEE Communications Magazine*, October 2005.

[27] R. E. Mullen. The lognormal distribution of software failure rates: Application to software reliability growth modeling. Proc. 9th International Symposium on Software Reliability Engineering, 1988.

[28] Ericsson White Paper. IMS – IP Multimedia Subsystem, The value of using the IMS architecture, 2004.

[29] IP Unity White Paper. IP Multimedia Subsystem - IMS, 2005.

[30] M. Poikselka, G. Mayer, H. Khartabil, and A. Niemi. *The IMS : IP Multimedia Concepts and Services in the Mobile Domain.* John Wiley and Sons, June 2004.

[31] T. Raty, J. Sankala, and M. Sihvonen. Network traffic analyzing and monitoring locations in the IP multimedia subsystem. In *31st EUROMICRO Conference on Software Engineering and Advanced Applications*, 2005.

[32] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnson, J. Peterson, R. Sparks, M. Handley, and E. Schooler. The session initiation protocol (SIP) RFC 3261. Internet Engineering Task Force, 2001.

[33] H. Schulzrinne, S. Narayanan, J. Lennox, and M. Doyle. SIPstone - benchmarking SIP server performance. http://www.sipstone.org, April 2002.

[34] H. Schulzrinne and J. Rosenberg. Internet telephony: architecture and protocols – an IETF perspective. *Computer Networks and ISDN Systems*, 31(3):237–255, Feb. 1999.

[35] P. Tomsu. The next generation network: Separating myth from reality. *Mobility Loop*, 2004.

[36] K. D. Wong and V. K. Varma. Supporting real-time IP multimedia services in UMTS. *IEEE Communications Magazine*, 2003.

[37] J. S. Wu and P. Y. Wang. The performance analysis of SIP-T signaling system in carrier class VoIP network. In *Proceedings of the 17th International Conference on Advanced Information Networking and Applications (AINA'03)*, Washington, DC, USA, 2003.

[38] B. Zhu. Analysis of SIP in UMTS IP multimedia subsystem. Master's thesis, Computer Engineering, North Carolina State University, 2003.