DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

# DATA TRANSFER OVER INTERNET PROTOCOL

by
Bilge NASUH

October, 2008
İZMİR

# DATA TRANSFER OVER INTERNET PROTOCOL

**A Thesis Submitted to the**
**Graduate School of Natural And Applied Sciences of Dokuz Eylül University**
**In Partial Fulfillment of the Requirements for the Degree of Master of Science**
**in Electrical and Electronics Engineering**

**by**
**Bilge NASUH**

**October, 2008**
**İZMİR**

## ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor Asst. Prof. Dr. Olcay AKAY for his continuous support, technical guidance and encouragement throughout this research. Also, I would like to thank my parents and sister for their never ending support and understanding.

**BILGE NASUH**

# DATA TRANSFER OVER INTERNET PROTOCOL

## BILGE NASUH

Session Initiation Protocol (SIP) is an application layer protocol which is generally used in “voice over internet protocol” (VoIP) technology. This protocol is utilized in order to establish, modify and terminate variable kinds of sessions such as VoIP, videoconferencing or data. SIP is suggested as the signalling protocol of 3rd generation networks which increases its popularity and application area.

Current trends in the telecommunication industry favour VoIP technology. In the future, this technology may substitute for “public switched telephony network” (PSTN). Since PSTN network has been tuned for performance and evolved to become highly reliable with individual switches experiencing only a few seconds of downtime per year, the performance of VoIP network should be closer to the performance of PSTN in order to realize substitution of PSTN network. At that point, characterization of the performance of VoIP network becomes essential, which is directly related to the performance analysis of the SIP.

In this thesis, performance analysis is performed using an analytical model of SIP messaging with respect to varying arrival rates and service rates. The SIP messaging is modeled as a queuing network and performance analysis is based on this network model. The main performance criteria used in the analysis are total number of customers in the system and the mean waiting time. Real time SIP messaging data are obtained by using real server and clients systems and it is modeled as a known probability density function (PDF) by analytical analysis. This known PDF is used for realizing queuing network performance analysis. As a conclusion, the performance analysis of real time SIP messaging is realized.

**Keywords:** Session Initiation Protocol, Queuing analysis

# INTERNET PROTOKOLÜ ÜZERINDEN VERI TRANSFERI

## BILGE NASUH

SIP (Session Initiation Protocol), “voice over internet protocol” (VoIP) teknolojisinde kullanılan bir uygulama katmanı protokolüdür. Bu protokol VoIP, video-konferans ya da veri aktarımı oturumlarının kurulması, değiştirilmesi ve bitirilmesinde kullanılır. SIP protokolünün 3. nesil ağların sinyalleşme protokolü olarak seçilmiş olması, popülaritesini ve uygulama alanını artırmış durumdadır.

Telekomünikasyon endüstrisindeki güncel eğilimler, VoIP teknolojisini desteklemektedir. Gelecekte bu teknolojinin “public switched telephony network” (PSTN) sisteminin yerini alacağı öngörülmektedir. Fakat PSTN ağları, yüksek güvenilirliğe ve performansa erişmek için yıllardır gelişmekte ve santrallerin kayıpları yılda birkaç saniyeyi geçmemektedir. PSTN ağlarının yerini alabilmek için VoIP ağlarının bu performansa yaklaşması gerekmektedir. Bu noktada, VoIP ağlarının SIP protokolünün performans analizine doğrudan bağlı olan performans karakteristiğinin incelenmesi önem kazanmaktadır.

Bu araştırmada performans analizleri SIP mesajlaşmasının değişen geliş oranları ve servis oranlarina göre analitik olarak incelenmesine dayanmaktadir. SIP mesajlaşması bir kuyruklama ağı olarak modellenmekte ve performans analizi bu model taban alınarak gerçekleştirilmektedir. İncelemede kullanılan temel performans kıstasları ortalama müşteri sayısı ve ortalama bekleme zamanıdır. Gerçek zamanlı SIP mesajlaşması verileri gerçek servis sağlayıcı ve müşteri sistemleri kurularak gerçekleştirilmiştir ve analitik inceleme sonucunda bilinen olasılık yoğunluk fonksiyonu (PDF) olarak modellenmiştir. Bu bilinen PDF, kuyruklama ağının performans analizinde kullanılmış ve sonuç olarak, gerçek zamanlı SIP mesajlaşmasının performans analizi gerçekleştirilmiştir.

**Anahtar sözcükler**: Session Initiation Protocol, Kuyruklama analizi

## CONTENTS

# CHAPTER ONE
# INTRODUCTION

## 1.1 Introduction

Telecommunication industry relieved from the hegemony of the public switched telephone network (PSTN) in the last 10-15 years. PSTN was invented to be used as a circuit-switched analog fixed-line telephony system and it evolved over a century to become highly reliable. The down rate of a PSTN switch is almost 99.999 %. In circuit-switched systems, a dedicated path is established from source to destination and all communication data are sent through this path. In time, digital transmission took place of analog transmission and efficiency of the network increased with this evolution.

The main disadvantage of the circuit-switched system is that all bandwidth is dedicated to one transmission and can not be reassigned for another one. The bursty and non-uniform nature of the data traffic causes the communication path to be idle or overloaded. Hence, dedicated circuit-switched systems cause a decrease in efficiency and the associated waste of bandwidth.

A packet-switched network is designed at the end of the 1960s as a solution to the waste of bandwidth and efficiency problems. In that system, the source splits the information to packets that are actually discrete blocks of data. Packets are routed independently according to the current capacity of the network by making root balancing. Since there is not a dedicated link as in the circuit-switched network, data links are shared with other transmissions and routing mechanism is determined dynamically according to the network resources.

Next generation networking is an alternative to the PSTN system. Next Generation Networks (NGN) are packet-based networks able to provide telecommunication services to users and able to make use of multiple broadband, QoS-enabled transport technologies. In NGN, service-related functions are independent of the underlying transport-related technologies (ITU-T, 2001).

Packet based transfer is one of the characteristics of the NGN systems. It provides both an important bandwidth and price advantage in comparison to PSTN systems. Many industry pundits claim that packet-switched voice will displace circuit-switched voice in the long term. However, in order to realize this transition, NGN systems should ensure an approximate degree of reliability and performance as today's PSTN technology.

The Internet Protocol Multimedia Subsystem (IMS) is a standardized NGN architecture defined by European Telecommunications Standards Institute (ETSI) and the 3rd Generation Partnership Project (3GPP) for implementing IP based telephony and multimedia services. IMS defines a complete architecture and framework that enables the convergence of voice, video, data and mobile network technology over an IP-based infrastructure. It is a part of the Universal Mobile Telecommunications System (UMTS) as standardization for 3G mobile phone systems. In November 2000, Session Initiation Protocol (SIP) was accepted as a 3GPP signaling protocol and became a permanent element of the IMS architecture for IP based streaming multimedia services in cellular systems. At the transition point from the PSTN to NGN network topology, performance issue is one of the most important comparison criteria. The service provider should promise guaranteed levels of service close to PSTN. Performance evaluation from end-to-end perspective is not an area that is deeply interesting. SIP plays an important role in performance analysis of the IMS since it is the signaling protocol of the IMS.

SIP is an application-layer control protocol that can establish, modify and terminate multimedia sessions or calls. These multimedia sessions include multimedia conferences, distance learning, Internet telephony and similar applications (Camarillo, 2002). It is originally submitted as an Internet draft to Internet Engineering Task Force (IETF) in February 1996 by Henning Schulzrinne. In February 1999, it reached to the proposed standard level and was published as RFC 2543. An expanded version of the protocol was published in June 2002 as RFC 3261 (Rosenberg et. al., 2002). The simulation programs used in this thesis and performance analysis are based on this latest version of the SIP.

Schulzrinne et.al, (2002) evaluate and benchmark the performance of SIP servers in the SIPStone project. In that project, a workload for SIP requests is generated and request handling capacity of SIP servers is evaluated. It describes the performance measurement through the point of view of a server designer.

Yin, (2005) investigates SIP proxies and benchmarks for SIP. An existing SIP implementation (SIP Express Router) is modified to allow the benchmarks to measure the performance of the protocol. The results of the benchmark suite present the profiling performance (in terms of cycles) and the architectural characteristics of the SIP. Like the SipStone project, it describes the performance measurement through the point of view of a server designer.

Fathi et.al, (2004) evaluate session setup of SIP and deal with various underlying protocols (transport control protocol (TCP), user datagram protocol (UDP), radio link protocol (RLP)) as a function of the frame error rate (FER).

De Marco et.al, (2005) analyze the traffic load generated by the SIP with an analytical technique to simulate SIP Finite State Machine (FSM) in an IP network by means of the theory of queuing networks. FSM of the SIP protocol is modeled with closed queuing network and performance characteristic of the model is evaluated according to the probability of session drop criteria.

Gurbani et.al, (2005) propose analytical open feedforward queuing models for the performance analysis of a SIP network and use these models to analyze the performance of a SIP network with respect to varying arrival rates, service rates and network delays. As a second contribution, SIP network reliability is also evaluated in this paper.

Zhu, (2003) analyzes the SIP operation in IP Multimedia Subsystem of UMTS network with respect to two aspects: bottleneck and delay. The bottleneck analysis was based on the investigation of the detailed call set up and release procedures. The delay analysis is firstly performed by describing the SIP signaling traffic with M/M/1 notation. After that, the delay in each node is calculated and the waiting time distribution is obtained. Then, the traffic is assumed to be M/D/1 and the delay in each node is evaluated.

Garetto et.al, (2001) describe an analytical model for the estimation of the performance of TCP connections. The model is based on the description of the behavior of TCP-Tahoe in terms of a closed queuing network. The assessment of the accuracy of the analytical model is based on gathering data using the simulation results of NS-2 package.

Wu and Wang, (2003) analyze the queue size, the mean of queuing delay, and the variance of queuing delay of SIP-T signaling system using embedded Markov chains in an M/G/1 queueing model. The theoretical estimates are compared with the simulation results.

Stuckmann, (2003) analyzes traffic engineering concepts for cellular packet radio network with QoS support. After evaluation of the characteristic arrival and service time distributions in GPRS systems by simulation, the mean values of these distributions are used for the analysis of an M/M/n queuing system.

Rajagopal, (2006) analyzes the IMS network based on the SIP signaling delay and predicts performance trends of the network based on an economic-theoretical approach. The focus of this work is on the formulation of queuing models for the IMS network and on the characterization of the SIP server workload.

## 1.2 Outline of the Thesis

In Chapter 2, SIP infrastructure and SIP messaging are presented. Protocol syntax and signaling procedures of SIP are described. A basic SIP call is given as an example in order to explain SIP messaging.

Chapter 3 is devoted to queuing theory. The relevant topics from the queuing theory and basic notations employed to describe queuing models are given. Different queuing models classified according to queuing and service discipline are explained. Performance measures used to evaluate queues are described.

In Chapter 4, firstly SIP messaging is modeled as a network of queues. Then, performance analysis of SIP messaging queues is performed according to M/M/1 queuing model. Collected real time data of SIP messaging are analyzed and Chi-square goodness-of-fit tests are performed in order to model the real time data of SIP messaging as a known probability density function (PDF). The modeled PDF is used to perform performance analysis of M/G/1 queuing model of SIP messages.

Finally, in Chapter 5, we present our conclusions and possibilities for future study.

# CHAPTER TWO
# GENERAL OVERVIEW OF SESSION INITIATION PROTOCOL

## 2.1 A Brief History of Session Initiation Protocol (SIP)

SIP was originally submitted as an Internet draft to IETF in February 1996 by Mark Handley and Eve Schooler to describe the methods of the session establishment between users. This first version was named SIPv1 and this abbreviation was meant as Session Invitation Protocol. It was text based and used UDP as the transport protocol. Also in February 1996, Henning Schulzrinne submitted an Internet draft to the IETF as the Simple Conference Invitation Protocol. This protocol's purpose was also establishing sessions between users. It was text based and used TCP as the transport protocol. In December 1996, IETF merged these two protocols and named them Session Initiation Protocol (SIP). SIP was approved in March 1999 as RFC 2543. In June 2002, RFC 3261 was approved as the latest SIP standard (Rosenberg et. al., 2002).

### *2.1.1 SIP Functionality*

SIP is an application-layer control (signaling) protocol for creating, modifying, and terminating sessions with one or more participants. These sessions include Internet telephone calls, multimedia distribution, and multimedia conferences (Camarillo, 2002).

A session is a state of connection between two or more devices. When it is established, all devices can communicate with each other and exchange data. Participants of a session could change their location and may be addressable by multiple names. They may communicate in several different media. In the mid-call, they could change the session properties like audio codec or could open a new session with another participant. SIP is designed in order to handle such different conditions through the establishment, modification and termination of a session.

SIP supports following abilities to the participants of a session:

- establishing connection
- adding parties
- transfering sessions
- changing session parameters
- terminating multimedia communications
- User location: determination of the end system
- User capabilities: determination of the media and parameters
- User availability: determination of the willingness for communications
- Call setup: "ringing", setting call parameters at the called and calling party
- invoking services

SIP does not depend on the transport protocols and the type of the sessions; i.e. transport protocol could be UDP, TCP, SCTP etc. and sessions could be voice, data, video etc.

SIP is a part of the IETF multimedia architecture and it works in a parallel manner with other protocols to provide complete service delivery to the users. This architecture includes different protocols such as:

- Real-Time Transport Protocol (RTP) for transporting real-time data like audio, video etc. (RFC 1889, 1996).
- Real-Time Streaming Protocol (RTSP) for setting up and controlling on-demand media streams, (RFC 2326, 1998).

§ Media Gateway Control Protocol (MGCP) and Megaco (also known as H.248) are signaling and call control protocols used within a distributed VoIP system for controlling media gateways (RFC 3435, 2003).

§ Session Description Protocol (SDP) for describing multimedia sessions (RFC 4566, 2006).

§ Session Announcement Protocol (SAP) for announcing multicast sessions (RFC 2964, 2000).

In Figure 2.1, SIP protocol is described within the TCP/IP Protocol hierarchy.

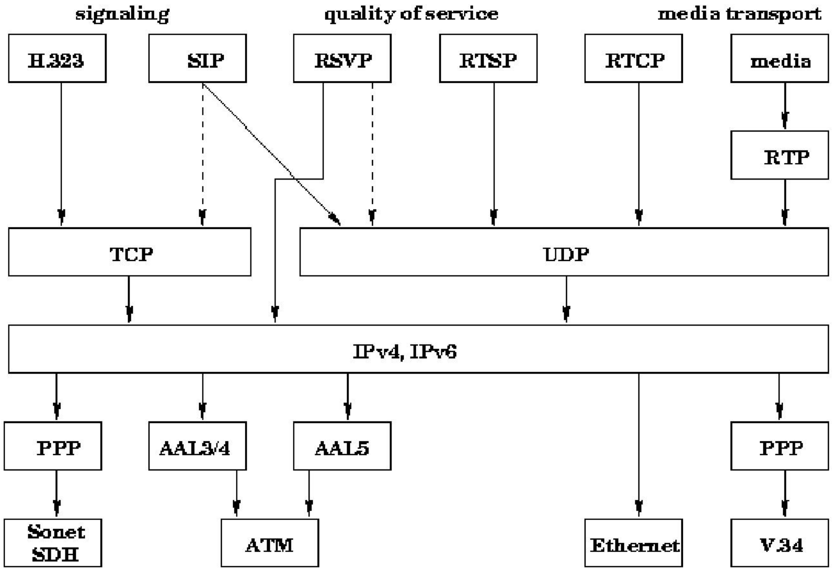


Figure 2.1 SIP Protocol in the TCP/IP architecture.

### *2.1.2 SIP Applications*

SIP could be used in setting up VoIP calls, setting up multimedia conferences, event notification (subscribe/notify) and for instant messaging.

## 2.2 SIP Infrastructure

SIP has several elements that describe the architectural behaviour of the protocol.

§Call: A call refers to some communication between peers.

§Dialog: A dialog is established by SIP messages as a peer-to-peer SIP relationship between two user agents.

§Request: A SIP message sent from a client to a server, for the purpose of invoking a particular operation.

§Response: A SIP message sent from a server to a client, for indicating the status of a request sent from the client to the server.

§SIP Transaction: A SIP transaction occurs between a client and a server and includes all messages from the first request sent from the client to the server up to a final response sent from the server to the client.

§User Agent (UA): A logical entity that interacts with a user and usually has an interface towards the user. IP Phone, PC and conference bridges are the example of SIP user agents. In a transition, it can act as both a user agent client and user agent server.

§User Agent Client (UAC): A logical entity that creates a new request.

§User Agent Server (UAS): A logical entity that generates a response to a SIP request. The response can be accepting, rejecting or redirecting the request.

§Redirect Server: Redirect servers help locate SIP UAs by providing alternative locations where the user can be reachable.

§Proxy Server: A proxy server primarily plays the role of routing call requests.

§Registrar: A registrar is a server that accepts REGISTER requests and places the information it receives in those requests into the location service for the domain it handles. Namely, it maintains mappings from names to addresses.

### *2.2.1 Description of the Protocol Syntax*

SIP is a text-based protocol similar to Hypertext Transfer Protocol (HTTP/1.1) and uses the UTF-8 character encoding. Like HTTP, SIP is a request/response protocol. Requests are SIP messages from a client to a server and Responses are SIP messages from a server to a client.

Every SIP message consists of three main parts:

§ method (in requests) or status code (in responses),

§ message header,

§ message body (description of session type in SDP, plain text, html etc.).

#### *2.2.1.1 Request.*

SIP requests have a Request-Line for a start-line. A Request-Line contains a method name, a Request-URI, and the protocol version.

Six methods are specified in RFC 3261:

- §REGISTER is used to register contact information of a client.
- §INVITE indicates that a user or server is being invited to participate in a session.
- §ACK is used to confirm that the client has received a final response to an INVITE request.
- §CANCEL is used to end a pending INVITE request.
- §BYE is generated to indicate other one to end the call.
- §OPTIONS is used for querying servers about their capabilities.

Additional methods could be defined in SIP extension RFCs.

#### *2.2.1.2 Responses.*

SIP responses have a Status-Line which consists of the protocol version followed by a numeric Status-Code and its associated textual phrase as their start-line.

Status-Code of the responses is described in RFC 3261:

- §1xx: Provisional responses indicate that request was received, continuing to process the request;
- §Typically, 100 (Trying) message implies that proxy server starts processing an INVITE and 180 (Ringing) implies that callee has received the request.

- 2xx: Success response indicates that the action was successfully received, understood, and accepted. Typically, 200 (OK) messages imply that callee hook off the phone.
- 3xx: Redirection response indicates that further action needs to be taken in order to complete the request and callee's new location information is sent to the caller.
- 4xx: Client Error response indicates that the request contains bad syntax or can not be fulfilled at this server.
- 5xx: Server Error response indicates that the server failed to fulfill an apparently valid request.
- 6xx: Global Failure response indicates that the request cannot be fulfilled at any server.

#### *2.2.1.3 Message Header.*

Message Header contains related information of SIP message such as host address, the destination address, call sequence number etc.

Obligatory header files to be included in all headers are Via, To, From, CSeq and Call-ID.

Definitions of these headers are given bellow:

- To identifies the address of the callee.
- From indicates the address of the caller.
- Via indicates the path that the request has traversed so far namely the routing information of the message.

§CSeq contains the sequence number and is incremented for each message in the dialog.

§Call-ID identifies a particular identification number for a call.

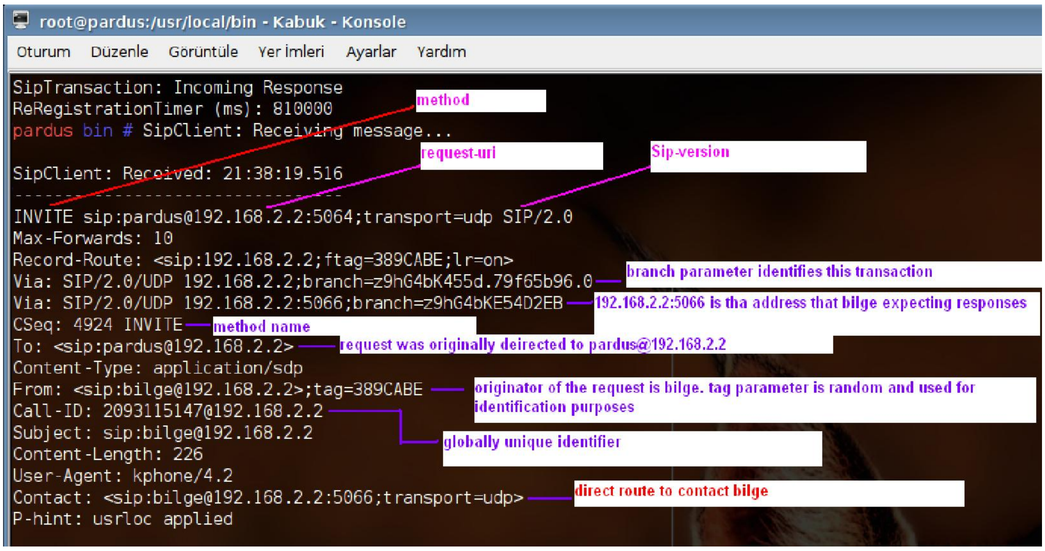
§Contact indicates the information about current location of the user.

§Content-Length indicates the length of the message body.

§Max-Forwards identify the maximum number of network nodes that the message should go through.

§Proxy-Authenticate is used to support a proxy authentication operation.

In Figure 2.2, SIP headers are explained in an SIP INVITE message.



```
SipTransaction: Incoming Response
ReRegistrationTimer (ms): 810000
pardus bin # SipClient: Receiving message...

SipClient: Received: 21:38:19.516
---------------------------------
INVITE sip:pardus@192.168.2.2:5064;transport=udp SIP/2.0
Max-Forwards: 10
Record-Route: <sip:192.168.2.2;ftag=389CABE;lr=on>
Via: SIP/2.0/UDP 192.168.2.2;branch=z9hG4bK455d.79f65b96.0
Via: SIP/2.0/UDP 192.168.2.2:5066;branch=z9hG4bKE54D2EB
CSeq: 4924 INVITE
To: <sip:pardus@192.168.2.2>
Content-Type: application/sdp
From: <sip:bilge@192.168.2.2>;tag=389CABE
Call-ID: 2093115147@192.168.2.2
Subject: sip:bilge@192.168.2.2
Content-Length: 226
User-Agent: kphone/4.2
Contact: <sip:bilge@192.168.2.2:5066;transport=udp>
P-hint: usrloc applied
```

method

request-uri

Sip-version

branch parameter identifies this transaction

192.168.2.2:5066 is tha address that bilge expecting responses

method name

request was originally deirected to pardus@192.168.2.2

originator of the request is bilge. tag parameter is random and used for identification purposes

globally unique identifier

direct route to contact bilge

Figure 2.2 An example of SIP request message. SIP Headers are explained in white boxes.

Message Body includes data that describe the multimedia properties of a session. This multimedia information is described in the message body using SDP (Session Description Protocol). Actually SIP is independent from this protocol; however, it is a general convention to use SDP. SDP message generally contains the name of the session and its purpose, the time during which the session will be active and the information needed to build a session, such as media type, transport etc.

Here is an example of a message body:

v=0 // *Protocol version*

o=mhandley 29739 7272939 IN IP4 126.5.4.3 // *owner/creator and session identifier*

s=SIP Call // *session name*

t=3149328700 0 // *time the session is active*

c=IN IP4 135.180.130.88 // *connection information*

m=audio 49210 RTP/AVP 0 12 // *media and transport address*

m=video 3227 RTP/AVP 31 *//media and transport address*

a=rtpmap:31 LPC/8000

## 2.3 Basic SIP Messaging

A basic SIP operation in a dialog is demonstrated in Figure 2.3.

In order to initiate a transaction, caller (UAC) sends an INVITE message to the callee (UAS).

The proxy server receives the INVITE request from a UAC and responses a 100 (Trying) message, which implies that proxy server starts processing an INVITE.

When the UAS receives the INVITE request, it replies with a 180 (Ringing) response to the server which implies that the callee has received the request.

When the UAS decides to accept the request, it sends a 200 (OK) response. This message is also forwarded by the proxy server to the UAC.

The UAC sends an ACK request message to the UAS directly, because at this time, the UAC must have obtained the location of the UAS from the information in previous responses. This information is specified in the contact header.

The two parties start the session following the agreement in the INVITE message body.

When the UAS closes the session, it sends a BYE message to the UAC and session is finished.

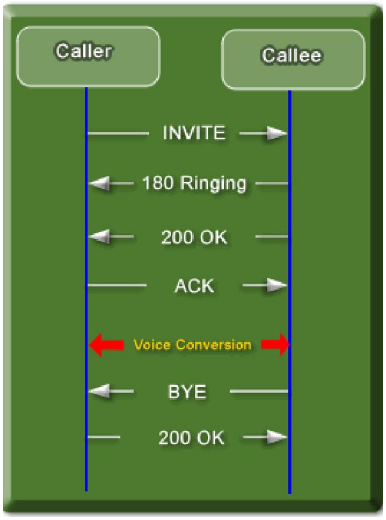


Figure 2.3 A basic SIP message flow.

# CHAPTER THREE
# QUEUING THEORY

Queuing theory is considered as a branch of applied probability theory that is mainly interested in generating an analytical model of customers needing service and uses that model to predict queue lengths and waiting times  (Willig,1999).

Important application areas of queuing models are communication networks, computer systems, production systems, transportation and stocking systems. The subject of queuing theory can be defined as the mechanism that describes sharing of expensive resources such as communication lines among a community of users (Kleinrock, 1975).

A queuing system is modeled by a server and population of customers which arrive at random times to the server and wait for service. After completion of the service, customers leave the server. Server can only serve a limited number of customers at a time. If a new customer arrives when the server is busy, customer enters a *waiting line* and waits until the service facility becomes available. Queue is constituted by the customers that are waiting for service. Waiting time in a queue demonstrates how long a customer has to wait between arrival time at the server and the time the server actually starts the service. Queuing theory mainly deals with the study of waiting lines, server and customers. Also within the scope of queuing theory is the case where several service centers are arranged in a network and a single customer can walk through this network at a specific path, visiting several service centers.

Customers waiting at the checkouts in a supermarket could be an example of a queuing system. Queuing theory could be considered as a guide to determine the required number of checkouts by analyzing waiting time of the customers during peak hours.

Data communication channels could be given as another example of queuing system by modeling package transmission over one switch to the next. In that system, packages arrive in a switch randomly. In the case that the arrival rates are greater than the switch capacity, new incoming packages are queued in order to prevent packages from being lost. In this example, queuing theory helps us in answering some performance questions such as determining the delay at the switches or determining the optimum size of the buffer for preventing packages from being lost.

The goal of the queuing theory is to determine some parameters like the mean waiting time in the queue, the mean system response time, mean utilization of the service facility, distribution of the number of customers in the queue, distribution of the number of customers in the system and so forth. These questions are mainly investigated in a stochastic scenario, where the interarrival times of the customers or the service times are assumed to be random.

Queuing systems could be classified according to various parameters like distribution of the interarrival and service times, number of servers in the system, the size of the waiting line (infinite or finite) and the service discipline.

Before investigating the classification of a queuing system, it is essential to explain some parameters that specify a queue.

Customers from some population arrive at the system at random arrival times. Interarrival times of the customers are assumed to be independent and have a common probability distribution. Costumer arrival rate is denoted by “$\lambda$” In many practical situations, customers arrive according to a Poisson stream (i.e. exponential interarrival times). Customers may arrive one by one, or in batches.

Service times of the server are independent and identically distributed (iid), and they are independent of the interarrival times. For example, service times can be deterministic or exponentially distributed. Service rate of the server is denoted by “**μ”**.

The service capacity is the number of the servers helping the customers in the system. There may be a single server or a group of servers.

The waiting room specifies the limitations with respect to the number of customers in the system. Sometimes, only a limited number of waiting spaces are available so customers that arrive when there is no room are turned away. Such customers are called blocked. The waiting room capacity specifies this turn ratio.

According to Kendall notation, a queuing system could be characterized briefly by using a four-part code system a/b/m/K. The first letter (a) specifies the interarrival time distribution and the second one (b) denotes the service time distribution. For example, M is used for exponential distribution (M denotes its memoryless characteristic), G is used for a general distribution and D is used for deterministic arrival and service times. The third letter (m) specifies the number of servers. Some examples of this notation are M/M/1, M/M/c, M/G/1, M/G/c, G/M/1 and M/D/1. The addition of a fourth letter (K) to the notation ensures the description of the waiting room size.

The queue or service discipline specifies the order in which customers are selected from the queue and allowed into service. Some common queuing disciplines could be summarized as follows:

*FIFO (First in, First out)*: A customer that finds the service center busy goes to the end of the queue.

*LIFO (Last in, First out)*: A customer that finds the service center busy proceeds immediately to the head of the queue. She will be served next, given that no further customers arrive.

*Random Service:* The customers in the queue are served in random order.

*Round Robin:* Every customer gets a time slice. If her service is not completed, she will re-enter the queue.

*Priority Disciplines:* Every customer has a (static or dynamic) priority; the server always selects the customers with the highest priority. This scheme can use preemption or not.

Performance measures used in the queuing systems could be divided into separate articles:

§ Distribution of the waiting time and the sojourn time of a customer. The sojourn time is the waiting time plus the service time.

§ Distribution of the number of customers in the system (including or excluding the one or those in service).

§ Distribution of the amount of work in the system. That is the sum of service times of the waiting customers and the residual service time of the customer in service.

§ Distribution of the busy period of the server. This is a period of time during which the server is working continuously.

§ The utilization gives the fraction of time that the server is busy.

§ The mean response time T is the mean time a customer spends in the system, i.e. waiting in the queue and being serviced.

In this thesis, some of these performance measurement criteria are employed in the analysis of SIP messaging. The M/M/1 and M/G/1 queuing systems are examined and performance analysis is performed based on these two models.

The simplest queuing system, the M/M/1 system (with FIFO service) consists of a single server and an infinite waiting line. The customer interarrival times are iid and exponentially distributed. The customer service times are also iid and exponentially distributed. The state of the system can be summarized in a single variable, namely the number of customers in the system. One of the most important properties of that system is its memoryless characteristic. This property provides the opportunity to observe the number of customers in the system at any time.

The other queueing system, which is the M/G/1 system, also consists of a single server, an infinite waiting line and iid customer interarrival times that are exponentially distributed. Although the customer service times are iid, the distribution of the customer service times could be Erlang, hyper-exponential, geometric, etc. The memoryless characteristic of M/M/1 queue is not valid for that model. Embedded Markov chains are used in order to determine performance measurement characteristics.

Above, systems are modeled as a single standalone queuing system. However, most real systems are better represented as a network of queues. In such networks, the departures from some queues become the arrivals to other queues. An obvious example is the Internet, where we can model each outgoing link of each router as a single queuing system, and where an end-to-end path traverses a multitude of intermediate routers. In a queuing network, a customer finishing service in a service facility immediately proceeds to another service facility or it leaves the system.

One basic classification of queuing networks is the distinction between *open* and *closed* queuing networks. In an open network, new customers may arrive from outside the system (coming from a conceptually infinite population) and leave the system later on. In a closed queuing network, the number of customers is fixed and customers neither enter nor leave the system. In that system, customer may enter or leave the system in any queue. The system may contain loopbacks and splitting points, where a customer has several possibilities for selecting the next queue. In the latter case each possibility is often assigned a fixed probability for a customer taking this branch.

In this chapter, it is intended to give a general overview about queuing theory and basic queuing systems. Necessary formulas and system parameters will be given in detail in Chapter 4 as they are needed in explaining the simulation applications.

# CHAPTER FOUR
# QUEUING ANALYSIS OF SIP MESSAGING

In this chapter of the thesis, firstly SIP messaging is modelled as a network of queues. Each SIP message is represented by a basic queue model and any basic SIP messaging is represented by a queuing network. Performance analysis of SIP messaging is realized according to the M/M/1 queuing model. Secondly, a SIP messaging traffic is run in real time and servicing times of the messages are obtained**.** By analyzing service time data, an estimated probability density function (PDF) is derived and queuing model is determined according to this PDF. Lastly, performance analysis of the traffic is performed by using this derived model.

## 4.1 SIP Messaging Model as a Network of Queues

Network of queues is a collection of two or more queues that are bound to each other. In a network of queues, a customer enters a queue and after its job is completed in that queue, it passes to a tandem queue in order to be served for another job. After all sequential jobs are finished, customer leaves the queuing system.

Basic SIP messaging in Figure 4.1 can be modelled as an open feed-forward network of queues. Each queue represents a SIP message serving station and queuing network is composed of these sequential queues. The model is generated by employing a proxy server. As an assumption, the delay between the messages is omitted. In addition, 100 (Trying) and ACK messages are also omitted.

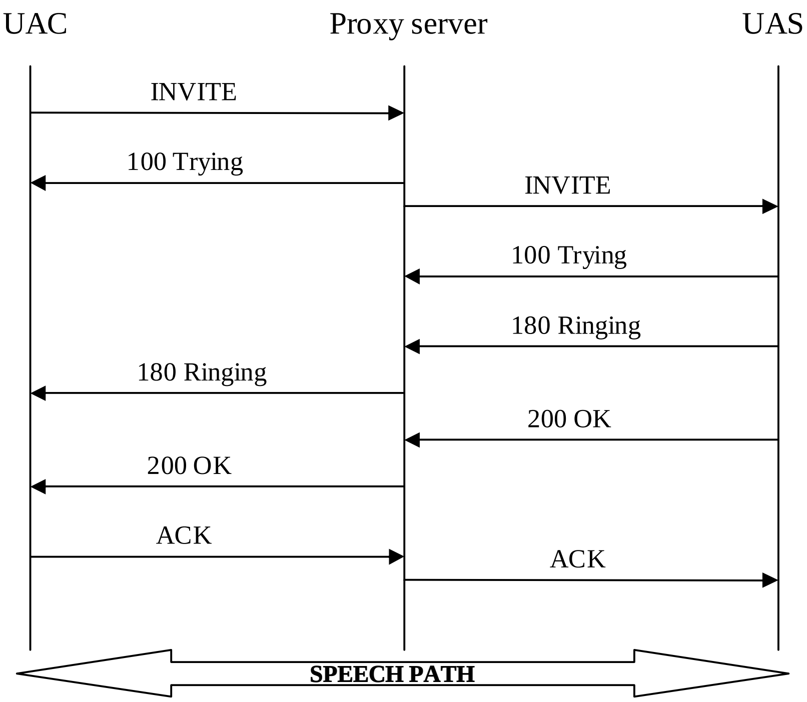


Figure 4.1 Basic SIP messaging.

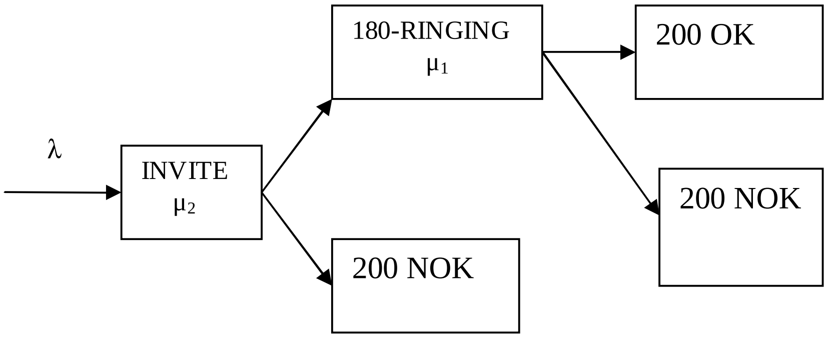


Figure 4.2 Open feed-forward network of queues model of SIP messaging.

The model is constructed in Figure 4.2 as a network of queues. SIP messaging is begun by an INVITE message and the first queue represents serving the INVITE message. Arrival rate to this queue is λ which represents the arrival rate to the network of queues system. Service rate of the INVITE message is represented by μ. By convention, 90 % of the coming SIP invite requests are continuing with 180 (Ringing) message and 10 % of the INVITE requests are continuing with non-2XX final response that corresponds to the failure and end of the messaging. 70 % of the 180 (Ringing) messages are followed by the 200 (OK) messages. This condition corresponds to the success of the communication. UAS answers the INVITE request and a speech path is established between UAS and UAC. 30 % of the 180 (Ringing) messages are followed by non-2XX final response that points to the failure in the communication and end of the session.

At first, performance analysis of the basic SIP messaging is performed according to M/M/1 queuing model. In the analysis of the system, mean waiting time and mean number of customers in the system are used as the performance criteria.

## 4.2 M/M/1 Queuing Model

M/M/1 queuing model is a system in which customers arrive according to a Poisson process of rate λ. Interarrival times are iid exponential random variables with mean 1/μ. If it is assumed that the queuing model of the SIP messaging network of queues is an M/M/1 system, the queuing model has memoryless property since service times and arrival times are exponential random variables. Memoryless (Markov) property of the system implies that the time until the next departure is independent of the time already spent in service. Thus, the past history of the system is irrelevant as far as the properties of the future state. Therefore, the state of the system can be summarized in a single variable, namely the number of customers in the system (N (t)) (Kleinrock, 1975).

In all queuing models, the parameter ρ indicates the server utilization of the system. It is defined by

$$\rho = \frac{\lambda}{\mu} . \tag{4.1}$$

The mean number of customers in the system is given by the following formula

$$E[N]= \sum_{j=0}^{\infty} j * P[N(t) = j] = \frac{\rho}{1-\rho}, \text{ and} \tag{4.2}$$

the mean waiting time in the queue is represented by

$$E[W]= \frac{\lambda}{1-\rho}. \tag{4.3}$$

### 4.3 Network of Queues by Using M/M/1 Queuing Model

According to the Jackson's theorem (Kleinrock, 1975), M/M/1 network of queues are analyzed by using the following formulas and abbreviations:

$\lambda$: Total arrival rate to the system

$\lambda_i$: Arrival rate to the $i^{th}$ node

$p_{ji}$: The probability that a customer leaving node i proceeds to node j

K: Number of the nodes traversed in the queuing system

$\rho_{k:}$ Utilization of the $k^{th}$ node

$\mu_{k\,:}$ Service rate of the $k^{th}$ node

The arrival rate to the $i^{th}$ node can be expressed by

$$\lambda_i = \sum_{j=1}^{K} p_{ji} * \lambda_j \quad \text{where } \lambda_1 = \lambda. \tag{4.4}$$

Service utilization of the the k$^{th}$ node is described by

$$\rho_k = \frac{\lambda_k}{\mu_k} . \qquad (4.5)$$

The mean number of customers in the M/M/1 network of queues is given by

$$E[N] = \sum_{k=1}^{K} \frac{\rho_k}{1-\rho_k}, \text{ and} \qquad (4.6)$$

the mean waiting time in M/M/1 network of queues is given by the following formula

$$E[W] = \sum_{k=1}^{K} \frac{\lambda_k}{1-\rho_k} . \qquad (4.7)$$

In order to analyze the mean number of customers in our SIP messaging network, Jackson's theorem is applied to the M/M/1 SIP messaging network of queues and analysis is repeated for changing arrival rates and service rates.

By setting the arrival rate as 0.3 ms$^{-1}$, mean number of customers versus changing service rate is obtained as in Figure 4.3. In that figure, it is seen that, as expected, mean number of customers in the system decreases as the service rate is increased. However, increasing service rate beyond 0.7 ms$^{-1}$ does not affect mean number of customers significantly. Since increasing service rate causes an increase in the cost of the system, there is no need to increase service rate more than 0.7 ms$^{-1}$.

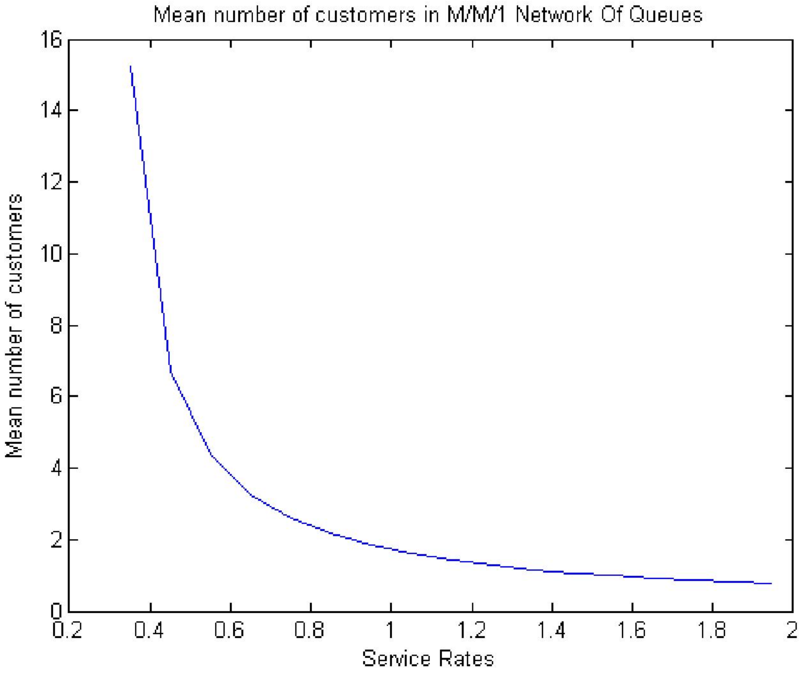


Figure 4.3 Mean number of customers in M/M/1 network of queues versus changing service rate.

By setting the service rate as 0.5 ms$^{-1}$, mean number of customers versus changing arrival rate is obtained as in Figure 4.4. In that figure, it is seen that the mean number of customers in the system increases as the arrival rate is increased. However, increasing arrival rates beyond 0.35 ms$^{-1}$ causes a dramatic accumulation in the mean number of customers in the system. This accumulation creates to an increase in the queue length which is not a desired result. Because of this reason, 0.35 ms$^{-1}$ seems to be an optimal service rate value for our system.

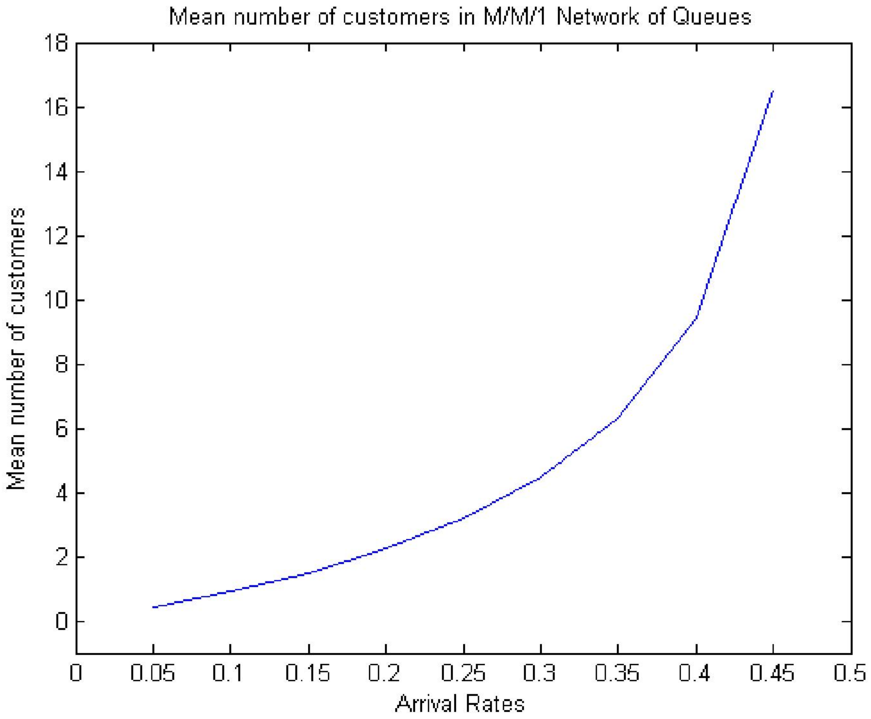


Figure 4.4 Mean number of customers in M/M/1 network of queues versus changing arrival rate.

By setting the arrival rate as 0.3 ms$^{-1}$, mean waiting time in the system versus changing service rate is obtained as in Figure 4.5. Observing this graph, we can say that mean waiting time decreases with the increasing service rate, and for service rates grater than 0.6 ms$^{-1}$ mean waiting time stays approximately constant.

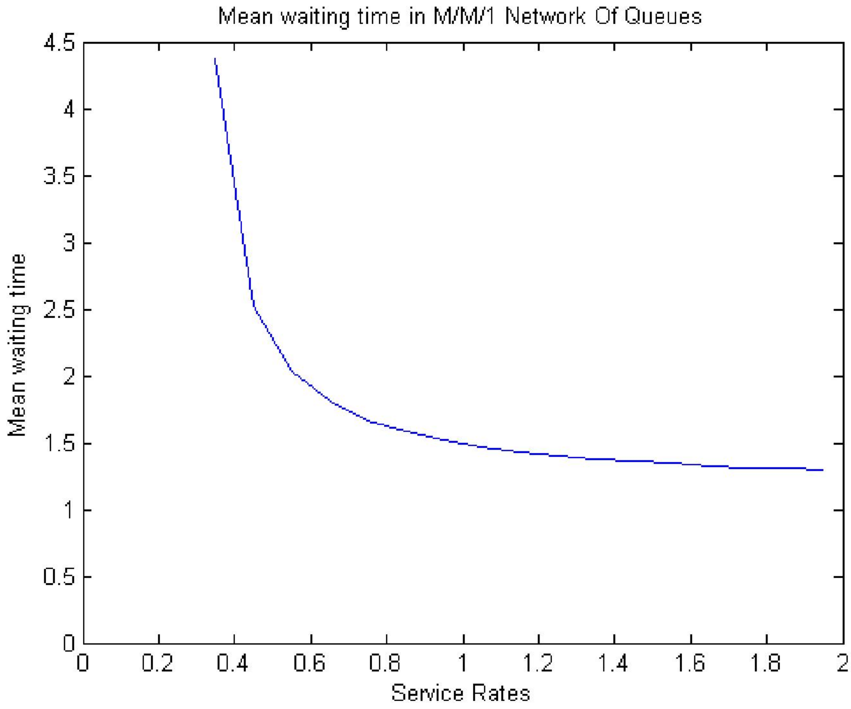


Figure 4.5 Mean waiting time in M/M/1 network of queues versus changing service rate.

By setting the service rate as 0.5 ms$^{-1}$, mean waiting time in the system versus arrival rate is obtained as in Figure 4.6. From this figure, we can deduce the same conclusion that the mean number of customer as 0.35 ms$^{-1}$ is an optimal service rate value for our system.

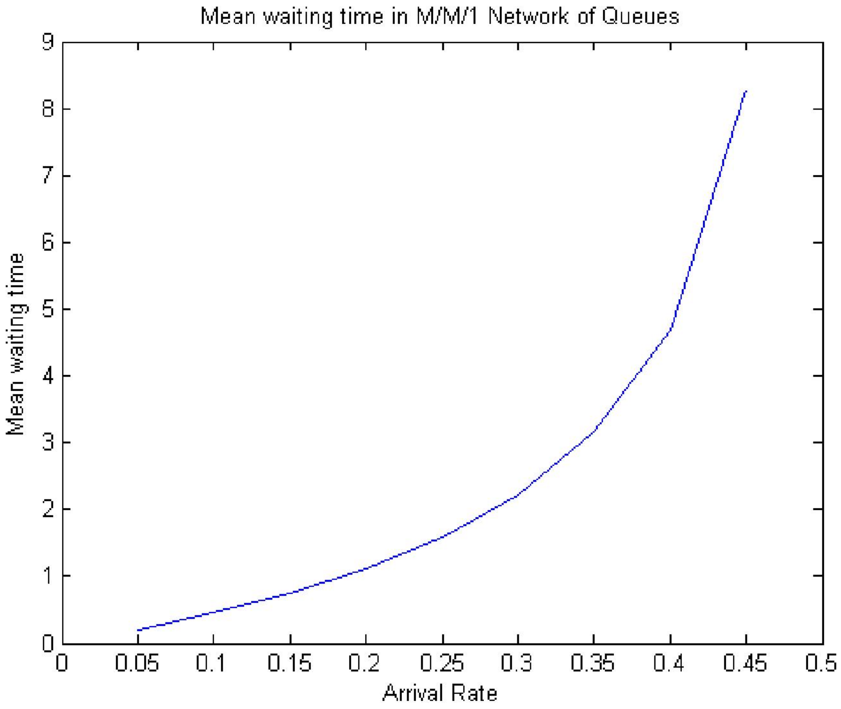


Figure 4.6 Mean waiting time in M/M/1 network of queues versus changing arrival rate.

## 4.4 Real Time Data Analysis of SIP Messaging

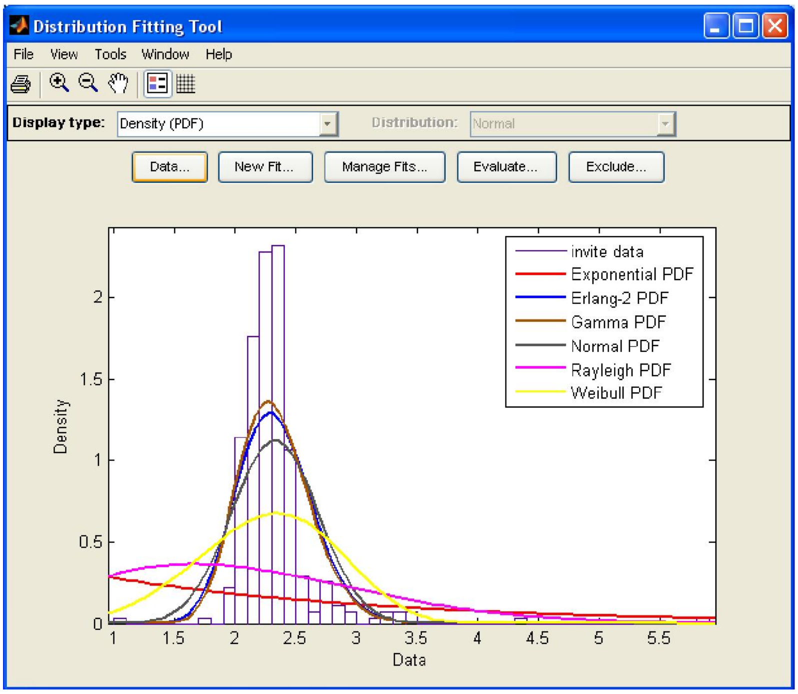
In order to determine the queuing model of the SIP traffic, it is essential to determine the distribution of the service time. All experiments are performed using SIP Express router as proxy server. Kphone is used as UAC and UAS. SIP sessions type is VoIP. The operating system used in computers is Pardus 2008. One computer used in the experiments is Intel Centrino Dual Core T5250 1.7 GHz 2GB RAM and the other one is Intel Centrino Dual Core T7250 2 GHz 2GB RAM.

We would like to emphasize that the general methodology used in deriving the distribution of the service time is more relevant than the actual data collected during the experiments since these actual data numbers are dependent upon the configuration of the servers.

Basic SIP calls are generated between two machines and traces are collected using Wireshark tools. A basic SIP call consists of INVITE, 100 (Trying), 180 (Ringing), 200 (OK), ACK and non-2XX final response. In order to provide simplicity, ACK and 100 (Trying) messages are omitted while forming basic SIP messaging queuing model.

Service times are obtained by examining Wireshark tool’s data. It is intended to obtain the service time distribution by using real time data. In order to analyze Wireshark tool’s data, Dfittool (Distribution Fitting Tool) of the MATLAB is used. Dfittool is a graphical user interface for displaying PDF and cumulative distribution function (CDF) which are fitted to actual data. By inputting the data to Dfittool, estimated mean and variance values of the fitting distribution are also acquired.

In the analysis of the service times of the messages, we upload the obtained Wireshark tool’s data to Dfittool and estimate PDF of service times. Erlang-2, gamma, exponential, normal, Rayleigh and Weibull distributions are used as possible fitting distributions. We obtain mean and variance values of the fitting distribution of the service times and corresponding graphics of these distributions. Figure 4.7 is one example of the output of the distribution fitting tool. In that figure, acquired Wireshark tool’s data for the INVITE messaging are displayed as histograms and PDF of the fitting distribution is plotted over the histogram. In Figure 4.8, output of the Dfittool for the fitting CDF distribution of the INVITE message service times is represented. Mean and variance parameters of the fitting distribution of the INVITE message are shown in Table 4.1**.**

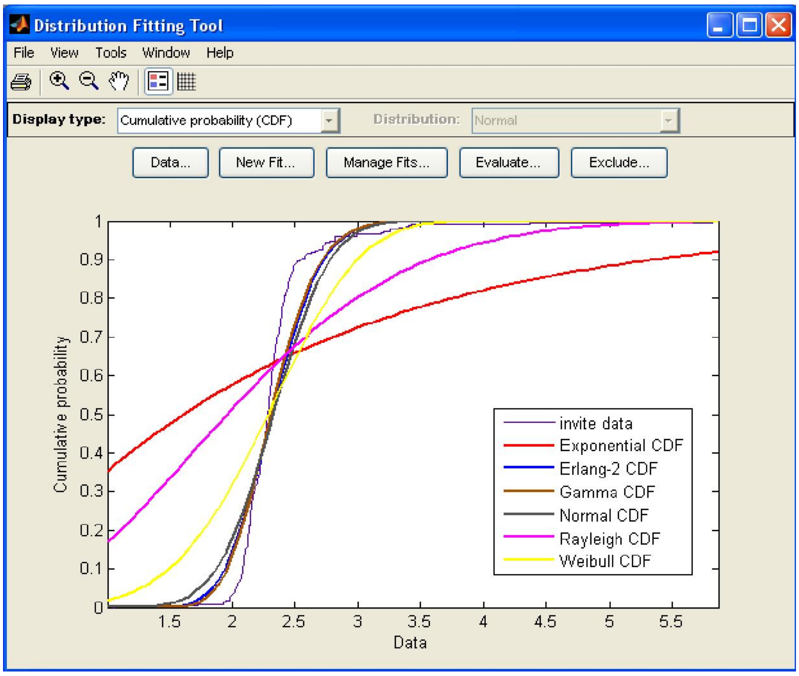


Distribution Fitting Tool

Display type: Density (PDF)

Distribution: Normal

Figure 4.7 Dfittool output estimated PDFs of the INVITE message service times.



Distribution Fitting Tool

Display type: Cumulative probability (CDF)

Distribution: Normal

Figure 4.8 Dfittool output estimated CDFs of the INVITE message service times.

Table 4.1 Mean and variance values of the estimated PDFs of INVITE message as determined by Dfittool.

| **Estimated CDF** | **Mean (ms$^{-1}$)** | **Variance (ms$^{-1}$)** |
|---|---|---|
| Exponential | 2.32632 | 5.41177 |
| Erlang-2 | 2.32632 | 0.0962687 |
| Gamma | 2.3246 | 0.089136 |
| Normal | 2.32632 | 0.12576 |
| Rayleigh | 2.08538 | 1.18827 |
| Weibull | 2.26374 | 0.333571 |

Dfittool estimated PDF and CDF plots for the 180 (Ringing), 200 (OK) and non-2XX messages could be found in Appendix A.1 at the end of the thesis.

Mean and variance values of the estimated PDFs determined by Dfittool for the 180 (Ringing), 200 (OK) and non-2XX messages could be found in Appendix A.2 at the end of the thesis.

After obtaining estimated values of the fitting distributions, it is essential to determine the optimally fitted distribution to the service times. For that purpose, goodness-of-fit tests are used.

### 4.5 Goodness-of-Fit Tests

Goodness-of-fit of a statistical model describes how well the data are modelled by a distribution (Bock, R. and Krischer, W. 1998). Measures of goodness-of-fit typically summarize the discrepancy between observed values and the values expected under the model in question. Since we have observed values of the data, we should compare them with expected values of the investigated model and determine how well tha data are modelled by the selected distribution. For that purpose, we prefer to use chi-square goodness-of-fit test.

### *4.5.1 Chi-square goodness-of-fit test*

Following steps are carried out in order to perform chi-square goodness-of-fit test:

§ Collection of the data obtained as Wireshark tool's results are denoted by $X_1,\ldots,$ Xn.

A histogram is created by selecting *m* number of intervals called bins denoted by $[e_j,e_{j+1})$ where $e_1<\ldots<e_{m+1}$, $e_1$ and $e_{m+1}$ satisfy

$$e_1 \leq \min X_i \text{ and } \max X_i \leq e_{m+1}, \qquad (4.8)$$

$$p_j = P(e_j \leq X_i < e_{j+1}). \qquad (4.9)$$

The histogram count denoted by $H_j$ can be described as

$$\frac{H_j}{n} \approx p_j . \qquad (4.10)$$

An example of the histogram acquired from INVITE data is plotted in Figure 4.9.

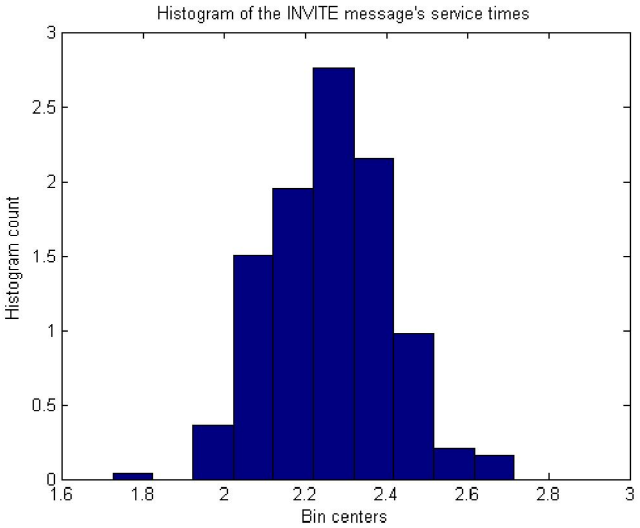


Figure 4.9 Histogram of the INVITE message service times.

Histograms for the 180 (Ringing), 200 (OK) and non-2XX final responses can be found in Appendix A.3 at the end of the thesis.

§ A static denoted as *Z* is determined by the following formula using histogram count

$$Z=\sum_{j=1}^{m}\frac{\left[H_j - np\right]^2}{np_j}\ . \qquad (4.11)$$

§ A candidate density thought as a good fit to the data is selected. This candidate is named as the *hypothesis*. A threshold, $z_\alpha$, is determined by using the candidate PDF (hypothesis) via

$$F_z(z_\alpha)=1-\alpha \qquad (4.12)$$

where $\alpha$ is usually taken as a small number such as 0.05.

The value of $z_\alpha$ could be easily found by using MATLAB command *chi2inv (1- α,k)* where k is the number of degrees of freedom of Z. The value of "k" for one estimated parameter could be computed by the formula k = m-1-1 where *m* is the number of bins. For two estimated parameters, "k" could be computed by the formula k = m-1-2.

§ If $Z \le z_\alpha$, hypothesis is well suited to the data and it is a good fit of data.
If $Z > z_\alpha$, hypothesis is rejected and a new hypothesis should be selected.

These chi-square goodness-of-fit test steps are applied to the PDFs obtained by the Dfittool. Z and $z_\alpha$ values of the INVITE messaging distributions are summarized in Table 4.2. Tables for the 180 (Ringing), 200 (OK) and non-2XX final responses are shown in Appendix A.4 at the end of the thesis.

Table 4.2 Calculated Z and $z_\alpha$ values of the INVITE message service time distribution and chi-square goodness-of-fit test results.

| DISTRIBUTION NAME | Z | $Z_\alpha$ | RESULT |
|---|---|---|---|
| Exponential | 2.49E+03 | 15.5073 | NOT REASONABLE |
| Erlang-2 | 10.6337 | 14.0671 | REASONABLE |
| Gamma | 10.6408 | 14.0671 | REASONABLE |
| Normal | 10.3296 | 14.0671 | REASONABLE |
| Rayleigh | 1.06E+03 | 15.5073 | NOT REASONABLE |
| Weibull | 36.8167 | 14.0671 | NOT REASONABLE |

By analyzing the results in Table 4.2**,** we can say that the suitably fitted distribution could be gamma, Erlang-2 or normal distribution. All of these distributions satisfy the condition $Z \leq z_\alpha$.

At this point, we can say that the service time distribution of the INVITE, 180 (Ringing), 200 (OK) and non-2XX messages could be modelled by gamma, Erlang-2 and normal distributions. Thus, the real time data of the SIP messaging show that the network of SIP messaging queues should be modelled as M/G/1. The analysis of the M/G/1 queues is more complex than the M/M/1 queues and requires Laplace-Stieltjes transformation of the PDF. We choose Erlang-2 distribution as the service time distribution PDF due to its analytical advantages.

The estimated μ and ρ parameter values of Erlang-2 distribution are listed in Table 4.3 for INVITE, 180 (Ringing), 200 (OK) and non-2XX final responses.

Table 4.3 Arrival Rate, Service rate and utilization numerical values of the SIP messages with Erlang-2 distribution.

| MESSAGE | ARRIVAL RATE λ ($ms^{-1}$) | SERVICE RATE μ ($ms^{-1}$) | UTILIZATION ρ |
|---|---|---|---|
| INVITE | 0.3 | 0.8562 | 0.35 |
| 180 (RINGING) | 0.24 | 1.1716 | 0.21 |
| 200 (OK) | 0.216 | 3.175 | 0.07 |
| NON-2XX | 0.06 | 1.905 | 0.032 |

### 4.6 M/G/1 Queuing Model

In the M/G/1 queuing model, customers arrive according to a Poisson process. However, service rates have an arbitrary distribution whose mean and standard deviation are known. Service times are independent and iid random variables with a general PDF, $f_x(x)$.

The number of customers N(t) in an M/G/1 system is a continuous-time random process. Memoryless property of the M/M/1 system is no longer valid for the M/G/1 system. It can be shown that the sequence $N_j$ is a discrete-time Markov chain and it represents the steady state probability mass function (PMF) of the system at arbitrary time instants. Thus we can find steady state of N(t) if we can find the steady state PMF for the chain $N_j$ .

By using the embedded Markov chains approach, probability generating function of the number of customers N(t) could be obtained by the Pollaczek-Khinchin transform equation

$$G_N(z)=\frac{(1-\rho)(1-z)\beta(\lambda(1-z))}{\beta(\lambda(1-z))-z} \qquad (4.13)$$

where β(s) is the Laplace-Stieltjes transform of the service time distribution.

Since we choose Erlang-2 distribution as the service time distribution, β(s) in the Pollaczek-Khinchin formula should be Laplace-Stieltjes transform of the Erlang-2 distribution.

Laplace-Stieltjes transform of the Erlang-2 distribution is given by the following equation (Kleinrock, 1975)

$$\beta(s) = \left(\frac{\mu}{\mu + s}\right)^2 . \tag{4.14}$$

By substituting equation (4.14) in (4.13), we obtain

$$G_N(z) = \frac{(1-\rho)(\frac{\mu}{\mu+\lambda-\lambda z})^2(1-z)}{(\frac{\mu}{\mu+\lambda-\lambda z})^2 - z} . \tag{4.15}$$

By simplifying (4.15), we reach at the expressions

$$G_N(z) = \frac{(1-\rho)\mu^2(1-z)}{\mu^2 - z(\mu+\lambda+\lambda z)^2}, \text{ and} \tag{4.16}$$

$$G_N(z) = \frac{(1-\rho)(1-z)}{1-z(1+\rho(1-z)/2)^2}. \tag{4.17}$$

Therefore, we find the probability generating function of the number of customers of M/E2/1 queueing system as

$$G_N(z) = \frac{1-\rho}{1-\rho z-\rho^2 z(1-z)/4}. \tag{4.18}$$

By substituting ρ values of the INVITE, 180 (Ringing), 200 (OK) and non-2XX messages in (4.18), we can obtain numeric form of the probability generating function.

If we carry out this operation for INVITE message, we find that $G_N(z)$ for INVITE message as

$$G_N(z) = \frac{0.65}{1 - 0.38z + 0.03z^2}. \qquad (4.19)$$

Then, by obtaining partial fraction expansion of $G_N(z)$, we can decompose (4.19) as follows

$$G_N(z) = \frac{0.65}{z - 3.73} - \frac{0.65}{z - 8.94}. \qquad (4.20)$$

After taking inverse Laplace transformation of the above partial fraction expansion of $G_N(z)$, we obtain service time PDFs of SIP messages in a $M/E_2/1$ queueing system. Obtained service time distributions are listed in Table 4.4.

Table 4.4 Service time distributions of SIP messages according to Erlang-2 distribution.

| **MESSAGE** | **DISTRIBUTION** |
|---|---|
| INVITE | $f_x$ (x)= $1.11(0.27)^t$-$0.47(0.11)^t$ |
| 180 (RINGING) | $f_x$ (x)= $1.7(0.14)^t$-$0.91(0.077)^t$ |
| NON-2XX | $f_x$ (x)= $4.04(0.0183)^t$-$3.075(0.014)^t$ |
| 200 (OK) | $f_x$ (x)= $2.59(0.0433)^t$-$1.66(0.028)^t$ |

### 4.7 Network of Queues by Using M/G/1 Queuing Model

By using Jackson’s theorem, we can say that mean number of customers in the M/G/1 queueing network is the sum of the mean number of customers in individual queues.

Abbreviations used below could be summarized as:

$N_i(t)$ : mean number of customers in the i$^{th}$ queue,

N(t) : total number of customers in the system,

N: number of the queueing station.

Mean number of customers in the i$^{th}$ queue are calculated by using the obtained PDF distributions of each queueing station (SIP message queues). Last-come-first-served (LCSF) model servers are assumed in order to ensure Jackson’s theorem is valid for this queuing model.

Total number of customers in the system can be calculated by the formula

$$E[N]=\sum_{i=1}^{N} N_i(t). \tag{4.21}$$

By changing the arrival rates and the service rates of the queues, we find mean number of customers in M/E$_2$/1 SIP messaging queueing model with respect to different service and arrival rates.

By setting the arrival rate as 0.3 ms$^{-1}$, mean waiting time in the system versus changing service rate is obtained as in Figure 4.10. Based on the curve in this figure, we can say that mean number of customers increases dramatically when service rates are chosen between 0.6 ms$^{-1}$ and 1.3 ms$^{-1}$. Therefore, using a service rate within this range causes an accumulation in the length of queues which is not a desired result. According to the graphic, setting service rate as 0.4 ms$^{-1}$ is a reasonable choice for the successful performance of M/G/1 network of queues.

Setting the service rate as 0.4 ms$^{-1}$, mean waiting time in the system versus changing service rate is obtained as in Figure 4.11. Based on this figure, we can say that setting arrival rate greater than 0.5 ms$^{-1}$ causes a sharp increase in the mean number of customers in the system.

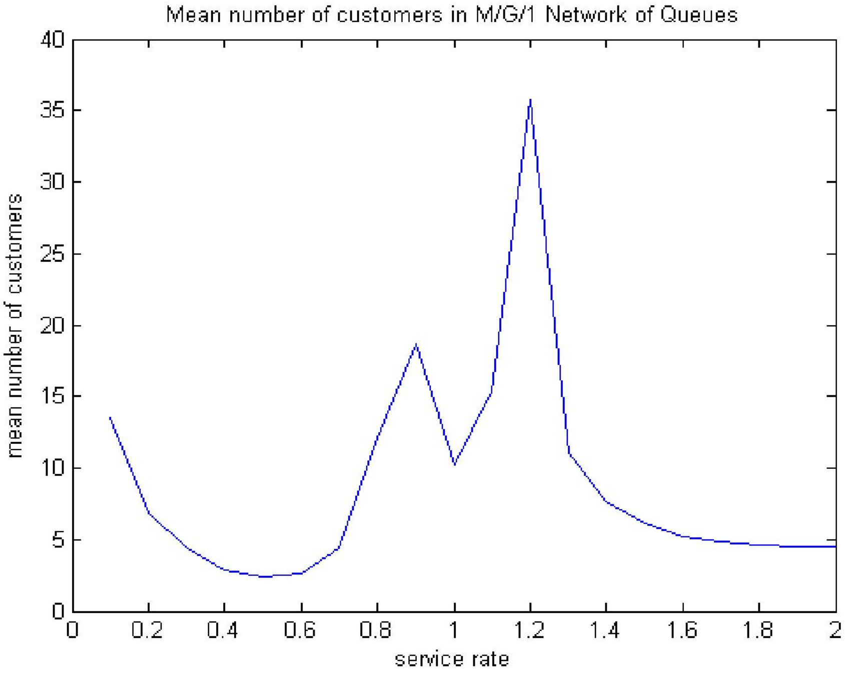


Figure 4.10 Mean number of customers in M/G/1 network of queues of SIP messaging versus changing service rates.

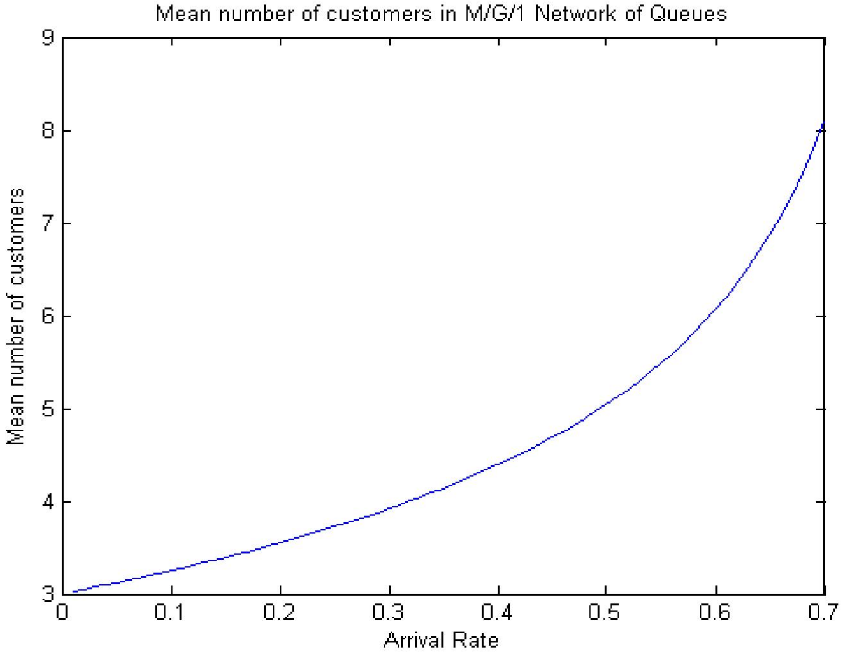


Figure 4.11 Mean number of customers in M/G/1 network of queues of SIP messaging versus changing arrival rates.

# CHAPTER FIVE
# CONCLUSIONS AND FUTURE STUDY

In this thesis, SIP messaging was modelled as a network of queues. At first, this model was analyzed according to the M/M/1 queuing model. The results of this analysis show that increasing service rate beyond 0.7 ms$^{-1}$ does not affect mean number of customers. Hence, setting service rate as 0.7 ms$^{-1}$ is reasonable for M/M/1 queueing model of SIP messages. It was also obtained that mean number of customers with arrival rate beyond 0.6 ms$^{-1}$ is approximately constant.

By analyzing real time SIP messaging, real time data values of INVITE, 180 (Ringing), 200 (OK) and non-2XX messages were obtained. These data were input to Dfittool in order to obtain estimated mean and variance parameters for various distributions. These estimated distributions were tested via Chi-square goodness-of-fit test. Three reasonable fits for PDF were obtained as a result of this test. Because of its analytical simplicity, Erlang-2 distribution was chosen as the most suitable fit.

Real time SIP messaging analyses show that M/G/1 queueing model should be chosen for SIP message network of queues. Analysis of M/G/1 network of queues was realized by using the Erlang-2 PDF with mean values obtained from Dfittool.

Real time SIP messaging data are obtained in real time in an environment composed of real servers and clients. SIP sessions type are preferred as VoIP. M/G/1 network of queues are analyzed based on the datas gathered in these conditions. This analysis shows that choosing service rate between 0.6 ms$^{-1}$ and 1.3 ms$^{-1}$ caused an accumulation in the number of customers in the system. This is not a desired result since accumulation of the number of customers indicates a performance problem. Hence, setting service rate as 0.6 ms$^{-1}$ for M/G/1 SIP network of queues is preferable from the performance point of view. By realizing the same analysis for arrival rates, we obtain that setting arrival rate as 0.5 ms$^{-1}$ is reasonable to prevent sharp increases in the number of customers in the system.

Performance measurement of SIP is essential for the future of NGN telecommunication systems. Queueing analysis is one of the methods of performing performance analysis of SIP systems. In this thesis, we realized M/M/1 and M/G/1 queuing models with Erlang-2 distribution analysis of SIP messaging. For further study, it could be possible to perform analysis of M/G/1 SIP queuing network using different distributions and performance measures.

## REFERENCES

ITU-T. *General overview of NGN (2004)*, Retrieved June 2004, from https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200706sf1.html

Camarillo, G. (2002). *SIP demistified*. NY: McGraw-Hill.

Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M. & Schooler, E. (June 2002). *SIP: Session initiation protocol, RFC 3261*, Retrieved June 2002, from http://www.ietf.org/rfc/rfc3261.txt.

Schulzrinne, H., Narayanan, S., Lennox & J., Doyle, M. (2002). *SIPstone-benchmarking SIP server performance*, Retrieved March 10, 2002, from http://www1.cs.columbia.edu/~library/TR-repository/reports/reports-2002/cucs-005-02.pdf

Yin, J. (2005). Session initiation protocol benchmark suite. *Master thesis, Computer Engineering, Delft University of Technology*.

Fathi, H., Chakraborty, S., & Prasad, R. (2004). Optimization of VoIP session setup delay over wireless links using SIP. *Globecom*, 742-752.

De Marco, G., Iacovani, G., & Barolli, L. (2005). A technique to analyse session initiation protocol traffic. *11th International Conference on Parallel and Distributed Systems*, 595-599.

Gurbani, V. K., Jagadeesan, L. & Mendiratta, V. B. (2005). Characterizing session initiation protocol (SIP) network performance and reliability. *International Service Availability Symposium*, 196-211.

Zhu, B., Analysis of SIP in UMTS IP Multimedia Subsytem (2003). *Master thesis, Computer Engineering, North Carolina State University*.

Garetto, M., Lo Cigno, R., Meo, M., & Ajmone Marsan, M. (2001). A detailed and accurate closed queueing network model of many interacting TCP flows. *Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies,* 1706-1715.

Wu, J. S., & Wang, P.Y. (2003). The performance analysis of SIP-T signalling system in carrier class VoIP network. *Proceedings of the 17th International Conference on Advanced Information and Applications*, 39-40.

Stuckmann, P. (2003).Traffic Engineering concepts for cellular packet radio networks with quality of service support. *Ph.D thesis, Rheinisch-Westf˙alischen Technischen Hochschule Aachen zur Erlangun.*

Rajagopal, N. (2006). Modelling and performance prediction of IP multimedia subsytem Networks. *Master thesis, Computer Sciences, North Carolina State University*.

Adan, I., & Resing,J. (2002). *Queueing theory. Lecture notes,* Retrived February 21, 2002, from http://www.win.tue.nl/~iadan/queueing.pdf

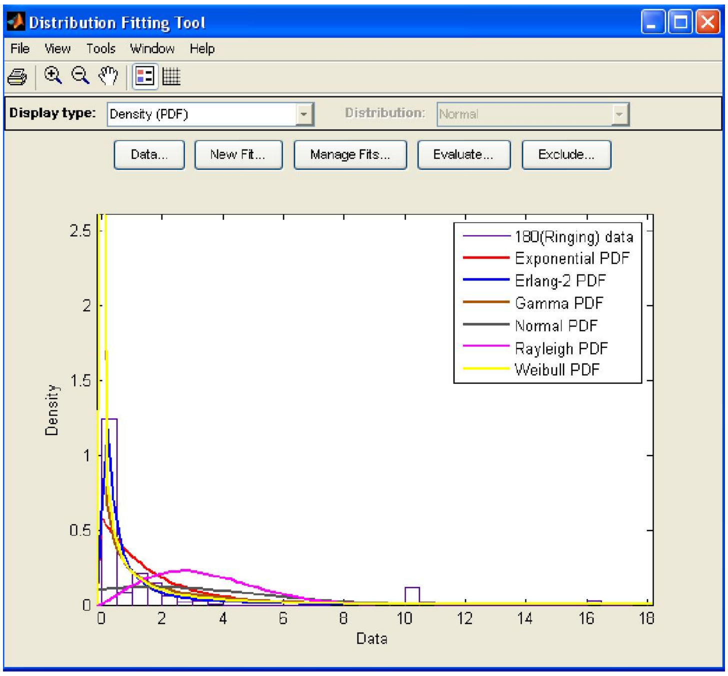
Willig, A. (1999). *A short introduction to queueing theory. Lecture's notes,* Retrieved July 21, 1999, from http://www.tkn.tu-berlin.de/curricula/ws0203/ue-kn/qt.pdf

Kleinrock, L. (1975). *Queueing systems, Vol. 1: Theory*. NY:Wiley Interscience.

Bock, R.K., & Krischer, W. (1998). *Data analysis briefbook*. Sweden:Springer

# APPENDIX A

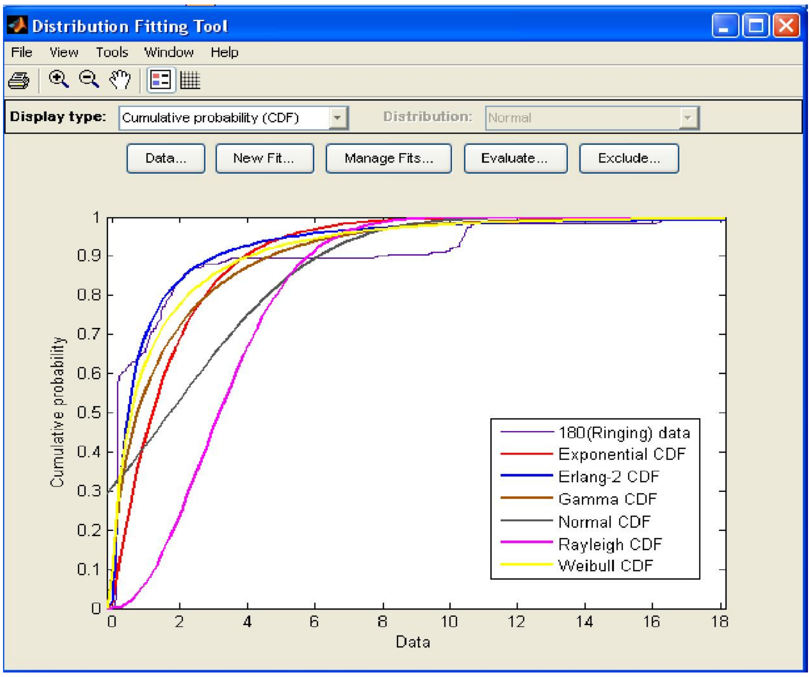
## A.1 Dfittool Estimated PDF and CDF Plots for the 180 (Ringing), 200 (Ok) and Non-2xx Messages



Display type: Density (PDF)

Distribution: Normal

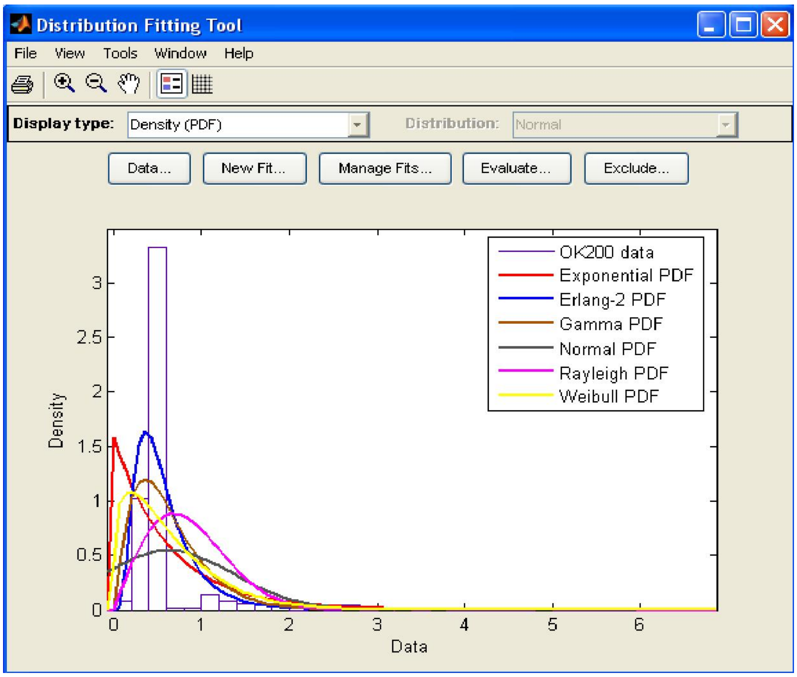
Figure A.1.1 Dfittool output estimated PDFs of the 180 (Ringing) message service times.



Display type: Cumulative probability (CDF)

Distribution: Normal

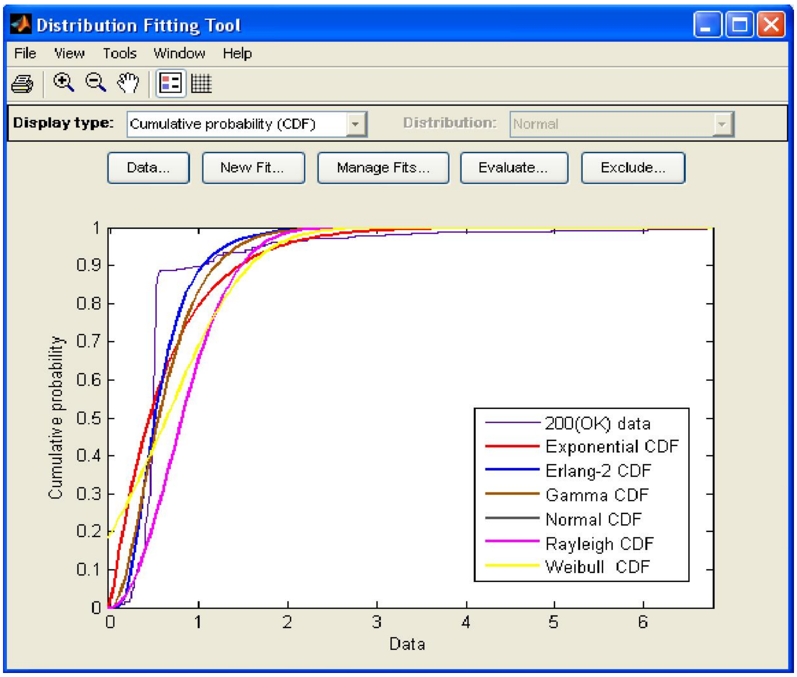
Figure A.1.2 Dfittool output estimated CDFs of the 180(Ringing) message service times.



Display type: Density (PDF)

Distribution: Normal

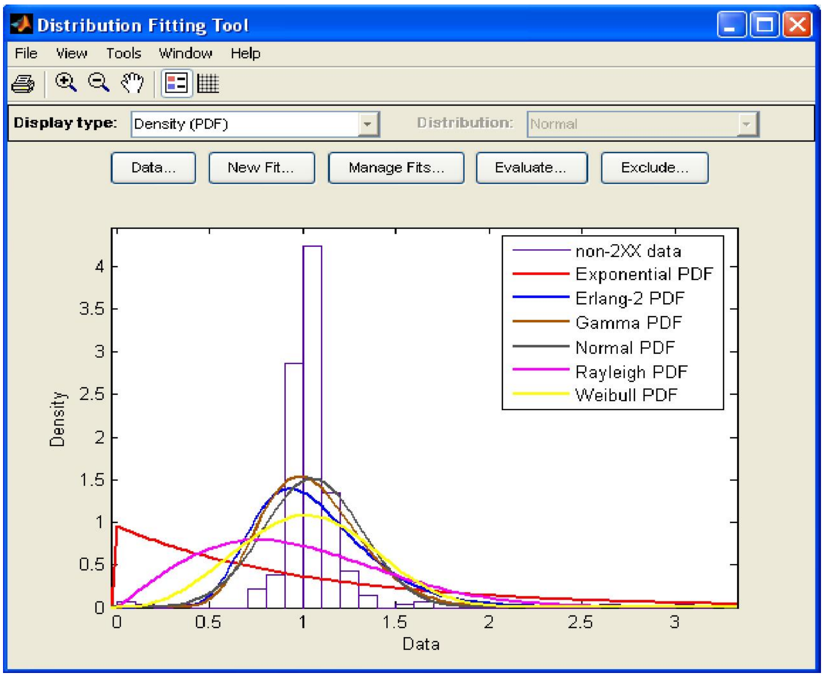
Figure A.1.3 Dfittool output estimated PDFs of the 200(OK) message service times.



Display type: Cumulative probability (CDF)

Distribution: Normal

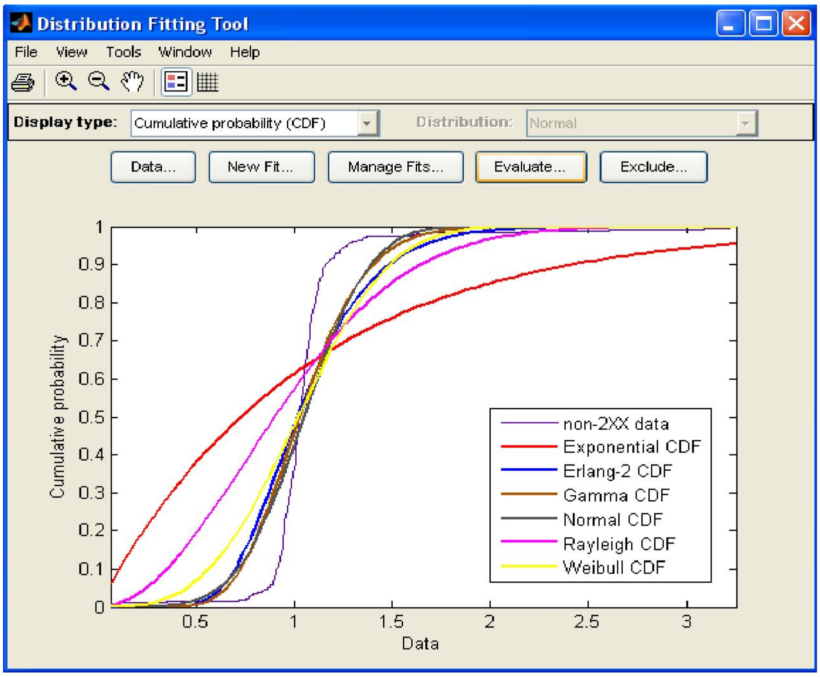
Figure A.1.4 Dfittool output estimated CDFs of the 200(OK) message service times.



Display type: Density (PDF)

Distribution: Normal

Figure A.1.5 Dfittool output estimated PDFs of the non-2XX message service times.



Display type: Cumulative probability (CDF)

Distribution: Normal

Figure A.1.6 Dfittool output estimated CDFs of the non-2XX message service times.

## A.2 Mean and Variance Values of the Estimated PDFs Determined by Dfittool for the 180 (Ringing), 200 (OK) and Non-2xx Messages

Table A.2.1 Mean and variance values of the estimated PDFs of 180(Ringing) message as determined by Dfittool.

| Estimated CDF | Mean ($ms^{-1}$) | Variance ($ms^{-1}$) |
|---|---|---|
| Exponential | 1.70691 | 2.91353 |
| Erlang-2 | 1.38158 | 15.9542 |
| Gamma | 1.70691 | 6.0614 |
| Normal | 1.70691 | 11.6542 |
| Rayleigh | 3.37949 | 3.12065 |
| Weibull | 1.51648 | 7.05927 |

Table A.2.2 Mean and variance values of the estimated PDFs of 200(OK) message as determined by Dfittool.

| Estimated CDF | Mean (ms$^{-1}$) | Variance (ms$^{-1}$) |
|---|---|---|
| Exponential | 0.63 | 0.3969 |
| Erlang-2 | 0.63 | 0.135303 |
| Gamma | 0.63 | 0.172162 |
| Normal | 0.63 | 0.540547 |
| Rayleigh | 0.857124 | 0.200738 |
| Weibull | 0.641324 | 0.268589 |

Table A.2.3 Mean and variance values of the estimated PDFs of non-2XX message as determined by Dfittool.

| Estimated CDF | Mean (ms$^{-1}$) | Variance (ms$^{-1}$) |
|---|---|---|
| Exponential | 1.05 | 1.1025 |
| Erlang-2 | 1.05 | 0.102003 |
| Gamma | 1.05 | 0.0704765 |
| Normal | 1.05 | 0.069478 |
| Rayleigh | 0.95931 | 0.251456 |
| Weibull | 1.0253 | 0.124341 |

## A.3 Histograms for The 180 (Ringing), 200 (OK) and Non-2xx Messages

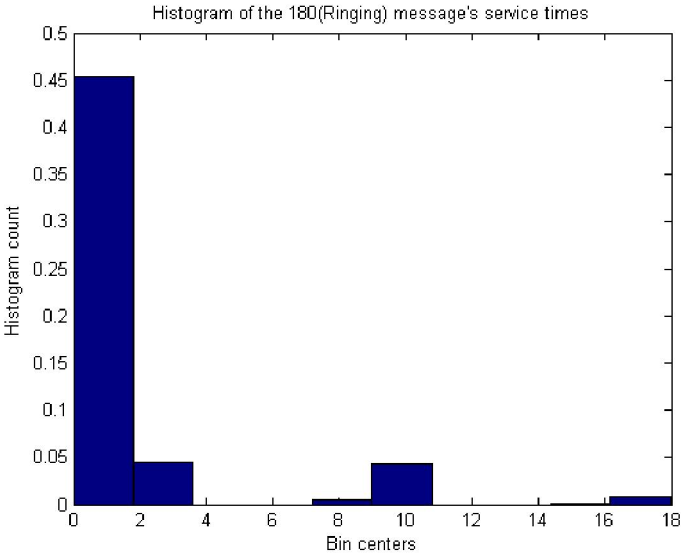


Figure A.3.1 Histogram of the 180(Ringing) message service times.

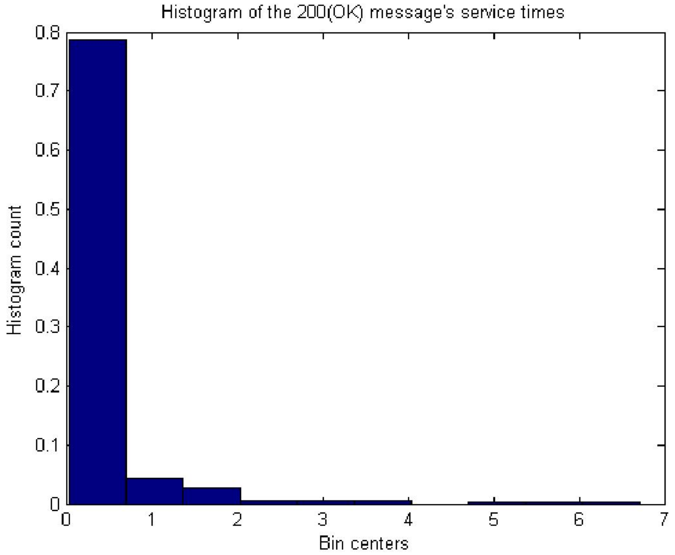


Figure A.3.2 Histogram of the 200(OK)message service times.

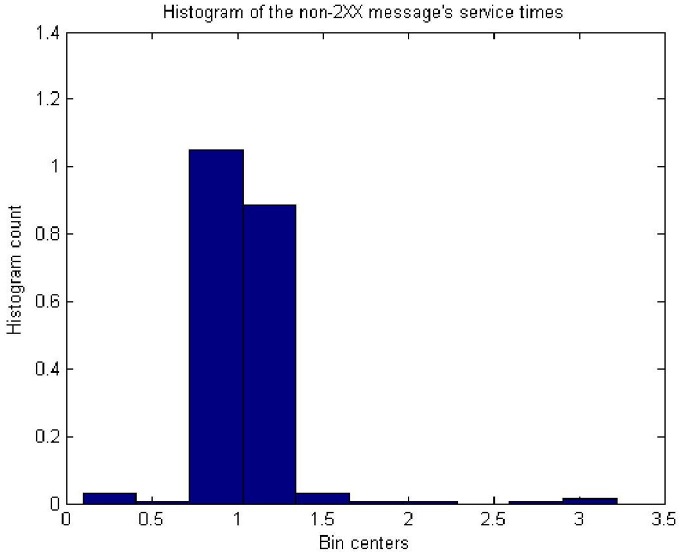


Figure A.3.3 Histogram of the non-2XX message service times.

## A.4 Calculated Z and Z $_\alpha$ Values for the 180 (Ringing), 200 (OK) And Non-2xx Messages

Table A.4.1 Calculated Z and z $_\alpha$ values of the 180(Ringing) message service time distribution and Chi-square goodness-of-fit test results.

| DISTRIBUTION NAME | Z | $Z_\alpha$ | RESULT |
|---|---|---|---|
| Exponential | 1.39E+03 | 22.362 | NOT REASONABLE |
| Erlang-2 | 19.2763 | 21.0261 | REASONABLE |
| Gamma | 19.2767 | 21.0261 | REASONABLE |
| Normal | 19.0552 | 21.0261 | REASONABLE |
| Rayleigh | 5.39E+02 | 22.362 | NOT REASONABLE |
| Weibull | 30.6798 | 21.0261 | NOT REASONABLE |

Table A.4.2 Calculated Z and z $_\alpha$ values of the 200(OK) message service time distribution and Chi-square goodness-of-fit test results.

| DISTRIBUTION NAME | Z | $Z_\alpha$ | RESULT |
|---|---|---|---|
| Exponential | 3.96E+05 | 27.6882 | NOT REASONABLE |
| Erlang-2 | 25.7691 | 26.217 | REASONABLE |
| Gamma | 359.2535 | 26.217 | NOT REASONABLE |
| Normal | 134.2993 | 26.217 | NOT REASONABLE |
| Rayleigh | 4.67E+02 | 15.5073 | NOT REASONABLE |
| Weibull | 44.6183 | 27.6882 | NOT REASONABLE |

Table A.4.3 Calculated Z and z $_\alpha$ values of the non-2XX message service time distribution and Chi-square goodness-of-fit test results.

| DISTRIBUTION NAME | Z | $Z_\alpha$ | RESULT |
|---|---|---|---|
| Exponential | 2.00E+03 | 27.6882 | NOT REASONABLE |
| Erlang-2 | 20.2506 | 26.217 | REASONABLE |
| Gamma | 20.2312 | 26.217 | REASONABLE |
| Normal | 25.9619 | 26.217 | REASONABLE |
| Rayleigh | 8.19E+02 | 15.5073 | NOT REASONABLE |
| Weibull | 83.898 | 27.6882 | NOT REASONABLE |