

# Modeling and Optimization for the Design of IMS Networks

Nisha Rajagopal and Michael Devetsikiotis  
Department of Electrical and Computer Engineering  
North Carolina State University  
nrjago, mdevets@ncsu.edu

## Abstract

The IP Multimedia Subsystem (IMS) is envisioned as the solution for the next generation multimedia rich communication. Based on an open IP infrastructure, it enables convergence of data, speech, video and mobile network technology. The Session Initiation Protocol (SIP) is the signaling protocol chosen by the 3GPP consortium for IMS. We propose to analyze the IMS network based on the SIP signaling delay and predict performance trends of the network, that allow us to choose parameter values optimally. The paper focuses on the formulation of queuing models for the IMS network and characterization of the SIP server workload, and on a methodology for the design of such networks for optimal performance. Our analysis is based on a careful study of real-life SIP network traffic.

## 1. Introduction

The IP Multimedia Subsystem (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 3GPP, 3GPP2 and IETF standards.

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).

IMS has a signaling and media plane which work separately, unlike Public Switched Telephone Network (PSTN). The signaling plane handles the session control, authorization, security and QoS aspects while the media plane manages media encoding and transport issues. Different protocols in IMS focus on delivering services including presence, instant messaging and push to talk [12].

The Session Initiation Protocol (SIP) [13] lies at the

core of the IMS architecture. SIP is the signaling protocol responsible for VoIP call setup and handling. An IMS based network comprises of one or more SIP servers, user databases known as Home Subscriber Servers (HSS), Application Servers (AS), Media Resource Functions (MRF), PSTN gateways etc [Fig (1)]. 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) [2].

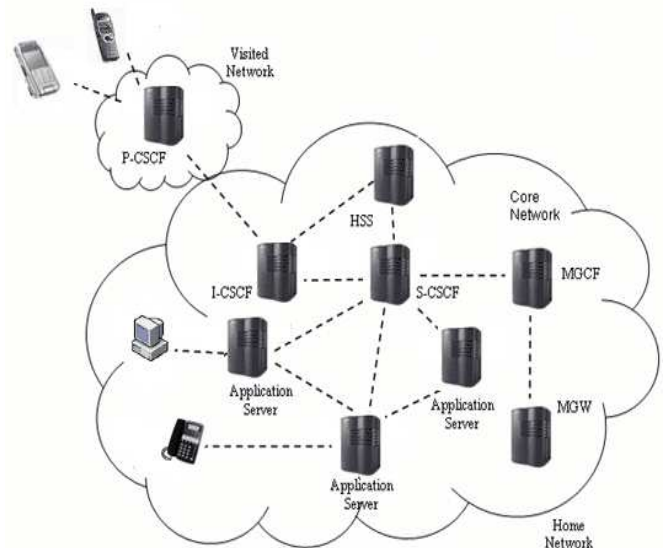


Figure 1. Typical IMS architecture

P-CSCF acts as an outbound/inbound SIP proxy server performing user authentication and verification of correctness of SIP requests. I-CSCF retrieves the user location information for routing purposes and is located at the edge of an administrative domain. S-CSCF acts as a registrar while performing session control and routing services. 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 re-

fer to all these nodes generically as servers in our paper.

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. Thus, we believe that performance evaluation of IMS based networks taking into QoS is essential. Therefore, we propose here a modeling methodology that uses real-life workload characterization, queuing analysis and optimization, in order to provide a systematic way to select system design parameters, such that performance (QoS) will be satisfied while overall system utility is maximized.

The rest of the paper is organized as follows: Section II focuses on existing work related to IMS performance. Section III covers the queuing models and optimization analysis. Section IV presents the case study on SIP server workload. We conclude our paper by summarizing our observations and future work to be done in this area.

## 2. Related Work

Most of IMS-related research so far has emphasized engineering rules, protocol development, compatibility issues and refinements of SIP. However, in order to enable the envisioned advances in internet telephony, the assessment of connection setup delay experienced using SIP needs to be critically evaluated.

ITU-T Recommendation E.721 provide guidelines on network grade of service parameters and target values for circuit-switched services in the evolving ISDN [15]. Recommendation E.721 specifies post selection delays as 3 seconds for local connection using ISDN. With this as a requirement, a study of IMS architectures based on delay sensitivity appears to be a requisite 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 *et al* cover SIP call setup delay based on SIP and H.323 traces from the Surveyor database [5]. The focus is on delay due to UDP loss and assumptions are made about the processing of tasks. Kist *et al* [10] 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.

The SIPstone Benchmark [14] 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. In Zhu [18] analysis is performed of SIP network in IMS from a UMTs perspective under a controlled environment with bottlenecks. The traffic in the

network is assumed to be processed following an M/D/1 model.

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.

In Wu *et al* [17] 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. We focus here on the analysis based on the SIP definitions in [13].

Gurbani *et al* analyze the performance and reliability under varying network parameters of an end-to-end native SIP ecosystem through a hierarchical performance and reliability model [6]. The mean response time for a proxy server is computed based on service time assumptions; for example, the INVITE service rate is fixed at  $0.5 \text{ ms}^{-1}$ . In our paper we analyze 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.

## 3. Proposed Scheme

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  $\lambda$ . The service time distribution of the servers in the network can be geometric, beta, gamma, etc. For the sake of illustration, we assume the server to have an exponentially distributed response time with a mean of  $\rho$  in the following example. The  $K$  servers have infinite queues and do not experience any losses due to buffer overflows. Such a network is modeled as a tandem queueing network of  $K$  nodes of type M/M/1/ $\infty$  [Fig (2)].



Figure 2. IMS queueing network

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 [3]. 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.

The utility function of an IMS network can be modeled as

$$U = V(\lambda) - C - P \quad (1)$$

Here  $U$  is the utility,  $V(\lambda)$  is the revenue earned by serving  $\lambda$  connections,  $C$  is the cost compensation for maintaining  $K$  servers in the network and  $P$  is a penalty function representing revenue losses due to lost connections.

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  $T$ .

The cost compensation  $C$  in (1) is the product of the number of the servers in the network,  $K$ , and the cost per server:

$$C = K * G(\rho) \quad (2)$$

The cost function in (2) is represented as

$$G(\rho) = a\rho - b\rho^2 \quad (3)$$

where  $a$  is the cost of each server based on its service rate  $\rho$  and  $b$  is the sales volume like a discount rate to model the economics of scale. The concavity of 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 [7].

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.,  $\rho \geq \lambda$ . This condition limits the values for  $\rho$  in the range  $\rho \in [\lambda, \frac{a}{2b}]$  (from the quadratic cost function equation (3)).

The penalty function  $P$  in (1) can be computed based on different criteria and constraints. We calculate the penalty for the following cases:

1. The total time for the connection establishment  $T$  exceeds the total average response delay experienced in  $K$  servers in the network.
2. The probability of the average response delay in  $K$  servers exceeds the probability of the session setup  $\epsilon$  being completed within a threshold time limit  $T$ .
3. A penalty of  $c$  being incurred for each lost connection with no constraints.

In a M/M/1/ $\infty$  queueing system, the expected response time is  $\frac{1}{\rho-\lambda}$  and the probability of the response time exceeding a time period  $T$  is given by  $e^{-(\rho-\lambda)T}$ . 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. The total time spent in the network can be approximated as the sum of time spent at each server in the network [11].

Thus, the total average response time experienced in  $K$  servers in the network, i.e., a M/M/K/ $\infty$  queueing network, is  $\frac{K}{\rho-\lambda}$  and the probability of the response time exceeding a time period  $T$  in  $K$  servers is  $K * e^{-(\rho-\lambda)T}$ .

The following sections describe the optimization of the utility functions for the defined scenarios in our example of a M/M/1/ $\infty$  queueing network.

### 3.1. Optimization of Utility Function based on single constraint that $T \geq \frac{K}{(\rho-\lambda)}$

The optimal utility function is

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

where  $\beta$  is the proportionality constant and  $T$  is the total time for connection establishment.

The first order derivative condition to find the maximum  $\rho$  is

$$\frac{\partial U}{\partial \rho} = -aK + 2bK\rho - \frac{K\beta}{(\rho - \lambda)^2} = 0 \quad (4)$$

Since solving for the closed form of  $\rho$  is difficult, we deduce some properties of the solution by applying the conjugate pair theorem from calculus [4]. The 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.

Based on this are the following results:

1. If the cost  $a$  increases, the service rate  $\rho$  decreases.

$$U_{\rho a} = -K \rightarrow \frac{\partial U}{\partial a} < 0$$

2. If the cost  $b$  increases, the service rate  $\rho$  increases.

$$U_{\rho b} = 2K\rho \rightarrow \frac{\partial U}{\partial b} > 0$$

3. If penalty for losing requests  $\beta$  increases, the service rate  $\rho$  decreases.

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

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

4. If the arrival rate  $\lambda$  increases, the service rate  $\rho$  increases.

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

We know that equation (4) is zero for the optimal  $\rho$ . If  $-K(a-2b\rho) < 0$  in equation (4) then  $[-K * \frac{\beta}{(\rho-\lambda)^2}] > 0$ . From this we know that  $U_{\rho\lambda} > 0$  as  $U_{\rho\lambda} = [-K * \frac{\beta}{(\rho-\lambda)^2}] * [\frac{2}{\rho-\lambda}]$ .

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

When the utility function is optimized with respect to the constraint, we get maximum utility when  $\rho = \frac{K}{T} + \lambda$  provided  $(\lambda + \frac{K}{T}) < \frac{a}{2b}$  as shown in Fig 3. We have assumed  $a = 4$ ,  $b = 0.0043$ ,  $\beta = 0.01$ , the number of servers  $K = 5$ ,  $T = 30$  sec, the income  $V = 5000$  and the arrival rate  $\lambda = 250$  arrivals/sec.

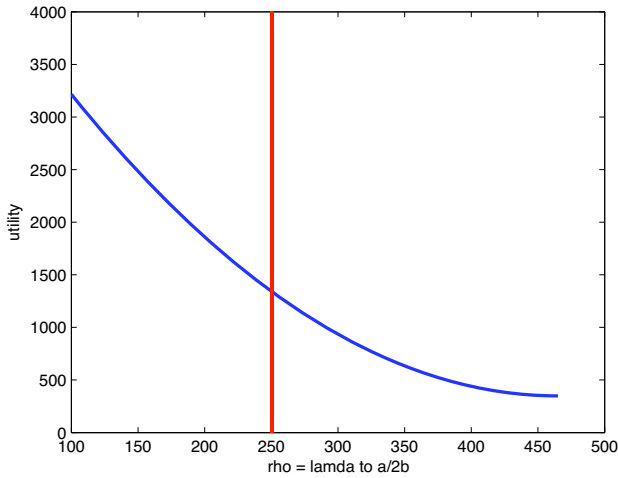


Figure 3. Utility vs. Service rate

### 3.2. Optimization of Utility Function based on constraint that probability of delay exceeds $K * e^{-(\rho-\lambda)T} \leq \epsilon$

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

$$\hat{U} = \max_{\rho} [V - K(a\rho - b\rho^2) - \beta(\epsilon - K * e^{-(\rho-\lambda)T})] \quad (5)$$

Here again  $\beta$  is the proportionality constant,  $\epsilon$  is the probability of completing the connection setup within threshold time limit and  $T$  is the time period.

The first order derivative to compute the maximum  $\rho$  is

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

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

1. As the cost  $a$  increases, the service rate  $\rho$  decreases.

$$U_{\rho a} = -K \rightarrow \frac{\partial U}{\partial a} < 0$$

2. As the cost  $b$  increases, the service rate  $\rho$  increases.

$$U_{\rho b} = 2K\rho \rightarrow \frac{\partial U}{\partial b} > 0$$

3. As the penalty for losing requests  $\beta$  increases, the service rate  $\rho$  decreases.

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

4. As the arrival rate  $\lambda$  increases, the service rate  $\rho$  increases.

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

The first order derivative of  $\hat{U}$  is  $-aK + 2bK\rho - KT\beta * e^{-(\rho-\lambda)T} = 0$  for optimal  $\rho$  from equation (5). Now if  $-K(a - 2b\rho) < 0$ , then  $[-KT\beta * e^{-(\rho-\lambda)T}] > 0$ . That implies  $[-KT\beta * e^{-(\rho-\lambda)T}] * T > 0$ .

The constraint  $K * e^{-(\rho-\lambda)T} \leq \epsilon$  can be simplified to  $\rho \geq \lambda + \frac{\ln K}{T} - \frac{\ln \epsilon}{T}$ .

Upon optimizing the utility function with respect to the constraint, we obtain  $\rho = \lambda + \frac{\ln K}{T} - \frac{\ln \epsilon}{T}$  [Fig (4)] provided  $(\lambda + \frac{\ln K}{T} - \frac{\ln \epsilon}{T}) < \frac{a}{2b}$ . From the above solution, we can note that as  $K$  increases, the service rate  $\rho$  will increase and as  $T$  increases,  $\rho$  will decrease.

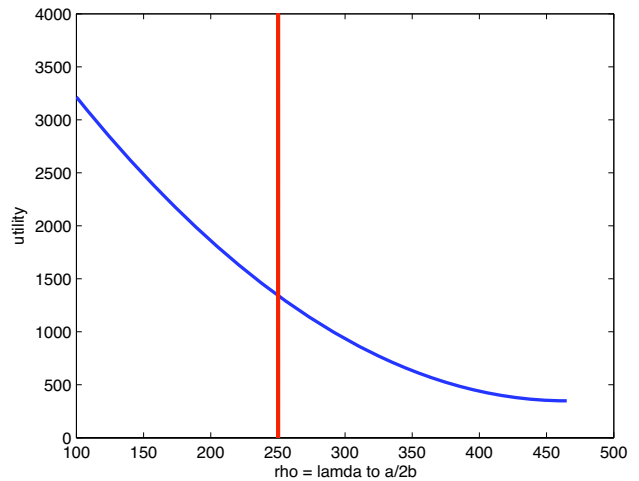


Figure 4. Utility vs. Service rate

We have assumed  $a = 4$ ,  $b = 0.0043$ ,  $\beta = 0.01$ ,  $\epsilon = 0.00005$ , the number of servers  $K = 5$ ,  $T = 30$  sec, the income  $V = 5000$  and the arrival rate  $\lambda = 250$  arrivals/sec to demonstrate the same graphically.

### 3.3. Optimization of Utility Function with no constraints and penalty due to delayed requests

In the last case, we have an unconstrained system. The probability of the processing time experienced in  $K$  servers exceeding time period  $T$  is given by  $K * e^{-(\rho-\lambda)T}$ . From Little's law, we have the number of lost connections is computed as  $\lambda * (K e^{-(\rho-\lambda)T})$ .

The optimal utility function is

$$\hat{U} = \max_{\rho} \left[ V - K(a\rho - b\rho^2) - c\lambda * K e^{-(\rho-\lambda)T} \right] \quad (6)$$

Here  $c$  is the penalty for each lost connection.

The first order derivative for the maximum service rate  $\rho$  is

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

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

1. If the cost  $a$  increases, the service rate  $\rho$  decreases.

$$U_{\rho a} = -K \rightarrow \frac{\partial U}{\partial a} < 0$$

2. If the cost  $b$  increases, the service rate  $\rho$  increases.

$$U_{\rho b} = 2K\rho \rightarrow \frac{\partial U}{\partial b} > 0$$

3. If penalty for losing requests  $c$  increases, the service rate  $\rho$  increases.

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

4. If the arrival rate  $\lambda$  increases, the service rate  $\rho$  increases.

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

Using the above discussed methods, we can analyze and predict trends in different networks. In the sequel, we assume the servers in the network to have a general distribution for their service times with definite first and second moments. We present a case study of collecting real network measurements in order to determine, even if approximately, the actual service distributions, that can be then used to model the end to end service and optimize the performance.

## 4. Case Study: Yahoo SIP Servers

For the M/G/1 analysis, we use the service distribution derived from Yahoo SIP servers. Yahoo Messenger's Voice chat is provided using SIP. We conducted experiments as discussed in this section to collect the SIP signaling messages. We used this data to deduce an appropriate service distribution pattern for Yahoo servers. We compare our findings with the norms specified by the ITU recommendation G.114 [16]. Again, we emphasize here the general methodology, rather than the actual numbers collected, since these can change with different implementations over time and across different providers.

### 4.1. Experimental Setup

All experiments were performed using the latest Yahoo Messenger version. 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 [8] was used to monitor network traffic. All experiments were performed between August and October, 2005.

### 4.2. Methodology

The calls were established between the two machines and the Ethereal dumps 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 [Fig (5)].

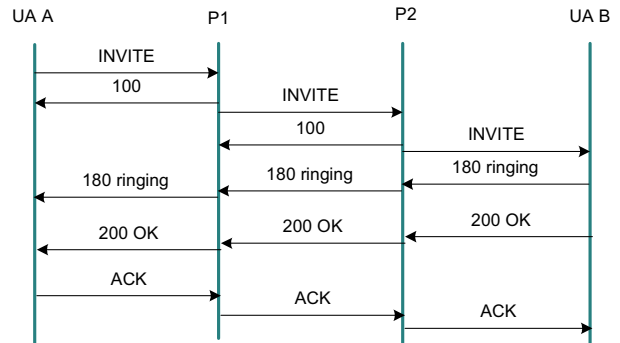


Figure 5. Typical SIP call flow setup

Figure (6) represents the SIP connection between source and destination. The Yahoo proxy 1 and proxy 2 are part of the Yahoo cloud. The messages involved in the 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 [1].

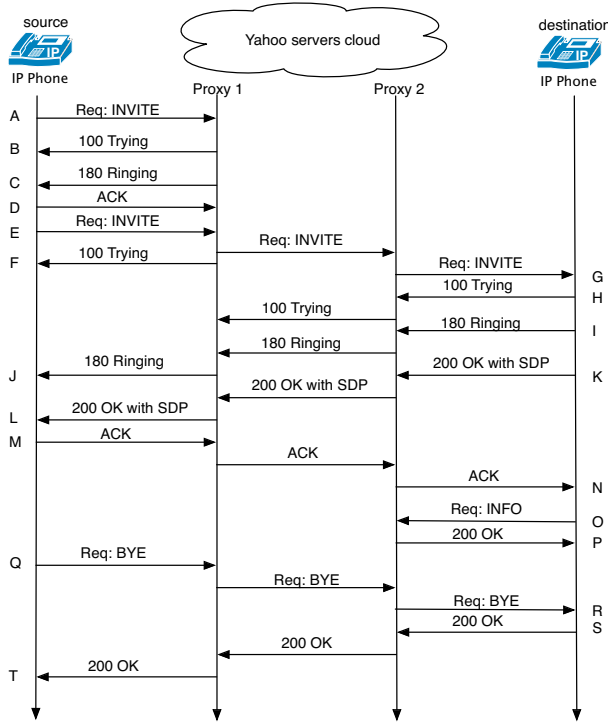


Figure 6. Yahoo call flow

We characterized the SIP server workload based on the response times for unauthorized INVITE, authorized INVITE, BYE, INFO, ACK. The server responded with 100 Trying, 401 Unauthorized, 100 Trying, 180 Ringing, 200 OK messages. The response time for an unauthorized invite was computed by subtracting A from C. The authorized invite message's response time was measured as the time from the invite message E till the time for the 180 ringing J. The response time for the bye message was the time difference between Q and T. The info message had a response time measured from O to P. The time spent in the Yahoo cloud was the cumulative time spent in all the SIP servers in the network was computed from Fig (6). The total setup time for the connection was the time taken from the first INVITE message A to the ACK message M.

We monitored the SIP traffic over 1000 calls and collected their response times. Arena's Input Analyzer [9] was used to fit the collected data to a distribution pattern. A log-normal distribution with the mean of 0.0288 and a standard deviation of 0.0545 was found from the collected data. The

distribution had -4.308 and 1.233 as its shape parameters. The total setup time had an average of 2.54 seconds which satisfies the ITU recommendation E.721 [15].

### 4.3. Analysis

In an M/G/1 queueing system, the average response time can be computed using the Pollaczek - Khintchine mean value formula [11]

$$T_{avg} = \bar{x} + \lambda \bar{x} * \frac{(1 + C_b^2)}{2(1 - \lambda \bar{x})} \quad (7)$$

where  $\bar{x}$  is the mean service time, i.e.,  $\frac{1}{\rho}$  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 with  $T \geq K * T_{avg}$ .

### 4.4. Optimization of Utility Function based on single constraint that $T \geq K * T_{avg}$

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

$$\hat{U} = \max_{\bar{x}} [V - K(a\bar{x} - b\bar{x}^2) - \beta(T - K * T_{avg})] \quad (8)$$

$$\hat{U} = \max_{\bar{x}} \left[ V - K(a\bar{x} - b\bar{x}^2) - \beta \left( T - K * \left\{ \bar{x} + \frac{\lambda \bar{x} * (1 + C_b^2)}{2(1 - \lambda \bar{x})} \right\} \right) \right] \quad (9)$$

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

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

Using the conjugate theorem we obtain the following analysis:

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

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

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

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

3. If the penalty for losing requests  $\beta$  increases, the service rate  $\bar{x}$  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  $\lambda$  increases, the service rate  $\bar{x}$  increases.

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

The service time distribution probability was computed from the SIP traffic collected over Yahoo SIP servers and was defined as a lognormal distribution

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

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

$$f(x) = \frac{0.323}{x} e^{-0.328[\ln(x)+4.308]^2}$$

## 5. Conclusion

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 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. We obtained a lognormal SIP server workload from data collected over Yahoo SIP servers.

Future work will focus on further data collection from different SIP servers and validation of performance evaluation trends. Additional work includes capacity planning of IMS networks based on performance models.

## References

- [1] S. A. Baset and H. Schulzrinne. An analysis of the skype peer-to-peer internet telephony protocol, 2004.
- [2] G. Camarillo, Miguel-Angel, and Garcia-Martin. *The 3G IP Multimedia Subsystem (IMS) : Merging the Internet and the Cellular Worlds*. John Wiley and Sons, August 2004.
- [3] C. Courcoubetis and R. Weber. *Pricing Communication Networks: Economics, Technology and Modelling*. John Wiley and Sons, 2003.
- [4] K. Currier. *Comparative Statics Analysis in Economics*. World Scientific Publishing Company, August 2000.
- [5] T. Evers and H. Schulzrinne. Predicting internet telephony call setup delay. In *IPTel 2000, First IP Telephony Workshop*, Berlin, Germany, 2000.
- [6] 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.
- [7] 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.
- [8] <http://www.ethereal.com/>. Ethereal: A network protocol analyzer.
- [9] D. Kelton, R. Sadowski, and D. Sturrock. *Simulation with ARENA*. McGraw Hill, 2003.
- [10] A. Kist and R. 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.
- [11] L. Kleinrock. *Queueing Systems, Vol. 1: Theory*. Wiley Interscience, New York, 1975.
- [12] 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.
- [13] 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.
- [14] H. Schulzrinne, S. Narayanan, J. Lennox, and M. Doyle. SIPstone - benchmarking SIP server performance. <http://www.sipstone.org>, April 2002.
- [15] I. T. Union. Network grade of service parameters and target values for circuit-switched services in the evolving ISDN. Recommendation E.721, Telecommunication Standardization Sector of ITU, 1992.
- [16] I. T. Union. General characteristics of international telephone connections and international telephone circuits: One-way transmission time. Recommendation G.114, Telecommunication Standardization Sector of ITU, February 1996.
- [17] 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.
- [18] B. Zhu. Analysis of SIP in UMTS IP multimedia subsystem. Master's thesis, Computer Engineering, North Carolina State University, 2003.