*Proceedings of the 7$^{th}$ International Conference on Applied Informatics*
*Eger, Hungary, January 28–31, 2007. Vol. 1. pp. 203–210.*

# A queueing network model to study Proxy Cache Servers*

**Tamás Bérczes, János Sztrik**

Faculty of Informatics, University of Debrecen
Debrecen, Hungary,
e-mail: {tberczes,jsztrik}@inf.unideb.hu

**Abstract**

The primary aim of the present paper is to modify the performance model of Bose and Cheng [1] to a more realistic case when external arrivals are also allowed to the remote Web servers and the Proxy Cache Server is unreliable. We analyze how many parameters affect the performance of a Proxy Cache Server. The main performance and reliability measures are derived, and some numerical calculations are carried out by the help of the MOSEL tool. The numerical results are graphically displayed to illustrate the effect of the non-reliability of the servers on the mean response time.

*Keywords:* Queueing Network, Proxy Cache Server, MOSEL

## 1. Introduction

The last several years have witnessed an explosive growth of the Internet. The explosive use of the Internet and the World Wide Web has caused congested networks and overloaded servers. As a result, the average waiting time for Web page delivery often takes a long time. It is commonly held that the majority of World Wide Web accesses are redundant. From the user's point of view it does not matter whether the requested files are on the firm's computer or on the other side of the world. One of the problems is that the same copy of the file can be claimed by other users at the same time. Because of this situation, identical copies of many files pass through the same network links, resulting in an increased response time. A natural solution to avoid this situation is to store these information, frequently accessed Web pages closer to the requesting users. This can greatly speed up Web page delivery and impose less cost in bandwidth. In general caching can be implemented at browser software; the originating Web sites; and the boundary between the local area network and the Internet. Browser cache are inefficient since they cache for only one user. The caching at the Web sites can improve performance, although

*Research is supported by Hungarian Scientific Research Fund-OTKA K 60698/2006.

the requested files are still subject to delivery through the Internet. Requested documents can be delivered directly from the web server or through a proxy cache server. A PCS has the same functionality as a web server when looked at from the client and the same functionality as a client when looked at from a web server. The primary function of a proxy cache server is to store documents close to the users to avoid retrieving the same document several times over the same connection. It has been suggested that, given the current state of technology, the greatest improvement in response time will come from installing a proxy cache server (PCS) at the boundary between the corporate LAN and the Internet.

In this paper a modification of the performance model of Bose and Cheng [1] is given to deal with a more realistic case when external arrivals are also allowed to the remote Web servers and the Proxy Cache Server is unreliable. For the easier understanding of the basic model and comparisons we follow the structure of the cited work. Furthermore, our aim is to illustrate graphically the effect of the non-reliability of the Proxy Cache Server on the steady-state systems measures. Because of the fact that the state space of the describing Markov chain is very large, it is difficult to calculate the system measures in the traditional way of solving the system of steady-state equations. To simplify this procedure we used the software tool MOSEL (Modeling, Specification and Evaluation Language), see [6], to formulate the model and to obtain the performance measures. By the help of MOSEL we can use various performance tools (like SPNP Stochastic Petri Net Package) to get these characteristics. The results of the tool can graphically be displayed using IGL (Intermediate Graphical Language) which belongs to MOSEL. The organization of the paper is as follows. Section 2 contains the queuing network model to study the dynamics of installing a PCS, the derivation of the main steady state performance measures. Section 3 is devoted to display numerical results graphically. The paper ends with Comments and Conclusion.

## 2. An analytical model of Proxy Cache Server traffic

In this section we briefly describe the queueing network model with the suggested modifications. Using proxy cache server, if any information or file is requested to be downloaded, first it is checked whether the document exists on the proxy cache server. (We denote the probability of this existence by $p$). If the document can be found on the PCS then its copy is immediately transfered to the user. In the opposite case the request will be sent to the remote Web server. After the requested document arrived to the PCS then the copy of it is delivered to the user.

The Proxy Cache Server can fail during the interval $(t, t+dt)$ with probability $\delta dt + o(t)$ if it is idle, and with probability $\gamma dt + o(t)$ if it is busy. If the server fails in busy state, its servicing the interrupted request after it has been repaired. The repair time is exponentially distributed with a finite mean $1/\nu$. If the server is failed all the operations are stopped. All the times involved in the model are assumed to be mutually independent of each other. As it can be seen this systems is rather complicated since it involves two types of failures: busy or idle server

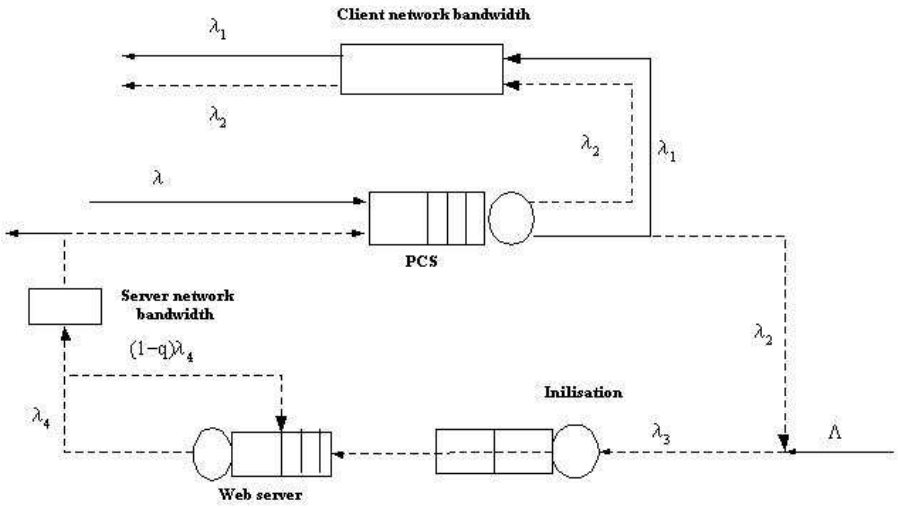


Figure 1: Network Model

state.

Figure 1 illustrates the path of a request in the modified model starting from the user and finishing with the return of the answer to the user. We assume that the requests of the PCS users arrive according to a Poisson process with rate $\lambda$, and the external arrivals at the remote web server form a Poisson process with rate $\Lambda$. The solid line in Figure 1 ($\lambda_1$) represents the traffic when the requested file is available on the PCS and can be delivered directly to the user. The $\lambda_2$ traffic depicted by dotted line, represents those requests which could not be served by the PCS, therefore these requests must be delivered from the remote web server. Naturally the web server serves not only the requests of the studied Proxy Cashe Server but it also serves requests of other external users. Let $\Lambda$ the intensity of these external arrivals. Let $\lambda_3$ denote the intensity of the overall requests arriving to the remote Web server. The overall $\lambda_3$ traffic undergoes the process of initial handshaking to establish a one-time TCP connection [2], [1]. We denote by $I_s$ this initial setup.

According to [1], "The remote Web server performance is characterized by the capacity of its output buffer $B_s$, the static server time $Y_s$, and the dynamic server rate $R_s$." So, the service rate is given by the the equation, where $F$ is the file size:

$$\mu_{Web} = \frac{R_s B_s}{F(Y_s R_s + B_s)} \tag{2.1}$$

The performance of the firm's PCS is characterized by the parameters $B_{xc}$, $Y_{xc}$

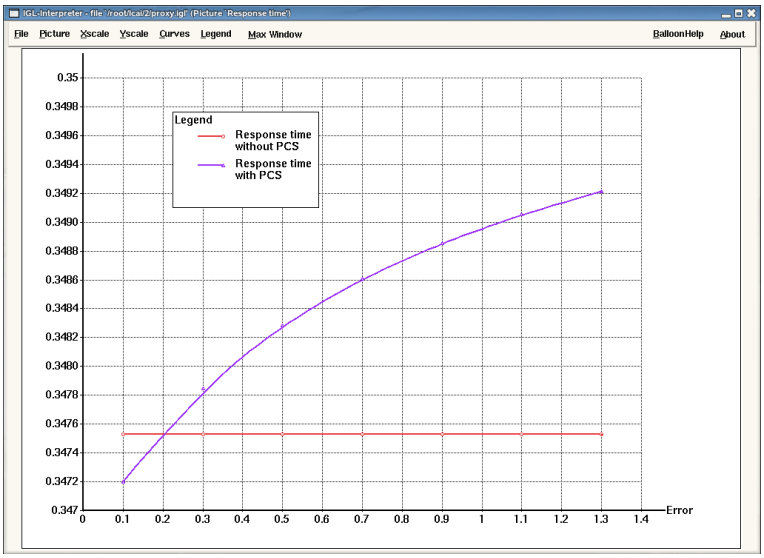


Figure 2: $p = 0.25$, $F = 5000$ bytes, $\lambda = \Lambda = 10$, $\nu = 10$, and $\delta = \gamma$

and $R_{xc}$. The service rate of the Proxy Cache Server is:

$$\mu_{PCS} = \frac{R_{xc}B_{xc}}{F(Y_{xc}R_{xc} + B_{xc})} \tag{2.2}$$

If the size of the requested file is greater then the Web server's output buffer it will start a looping process until the delivery of all requested file's is completed. Let

$$q = \min\left(1, \frac{B_s}{F}\right) \tag{2.3}$$

be the probability that the desired file can be delivered at the first attempt.

## 3. Numerical results

For the numerical explorations the corresponding parameters of Cheng and Bose [1] are used. The value of the other parameters for numerical calculations are: $F = 5000$ bytes, $I_s = I_{xc} = 0.004$ seconds, $B_s = B_{xc} = 2000$ bytes, $Y_s = Y_{xc} = 0.000016$ seconds, $R_s = R_{xc} = 1250$ Mbyte/s, $N_s = 1544$ Kbit/s, and $N_c = 128$ Kbit/s.

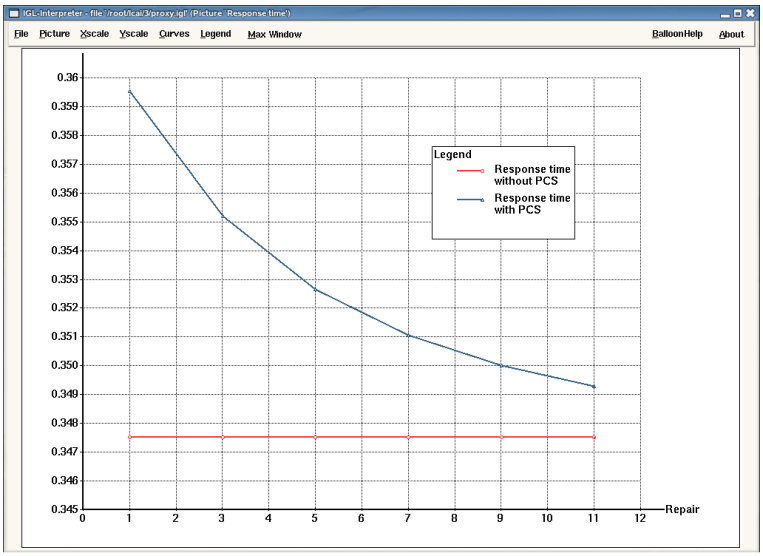


Figure 3: $p = 0.25$, $F = 5000$ bytes, $\lambda = \Lambda = 10$ and $\delta = \gamma = 2$

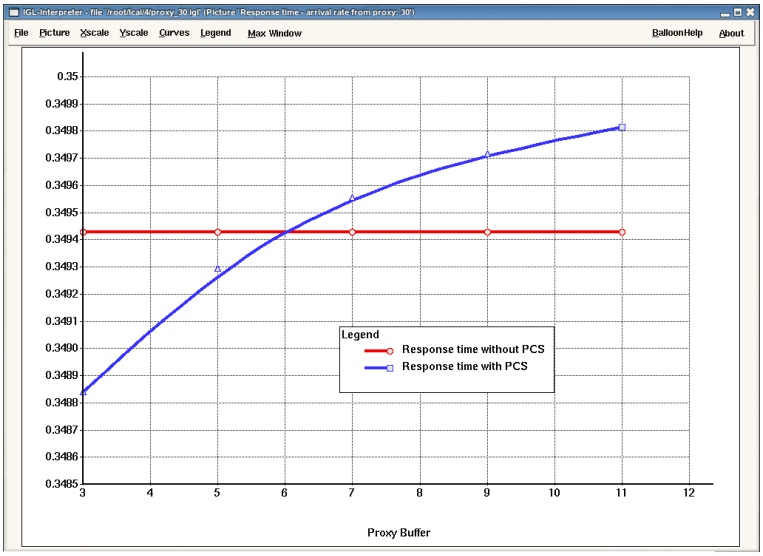


Figure 4: $p = 0.25$, $F = 5000$ bytes, $\lambda = 30, \Lambda = 10$, $\nu = 10$ and $\delta = \gamma = 0.2$

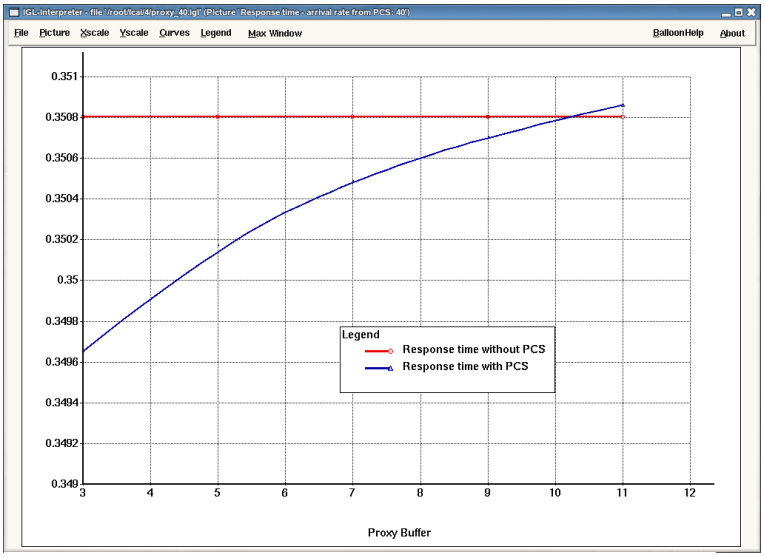


Figure 5: $p = 0.25$, $F = 5000$ bytes, $\lambda = 40, \Lambda = 10$, $\nu = 10$, and $\delta = \gamma = 0.2$

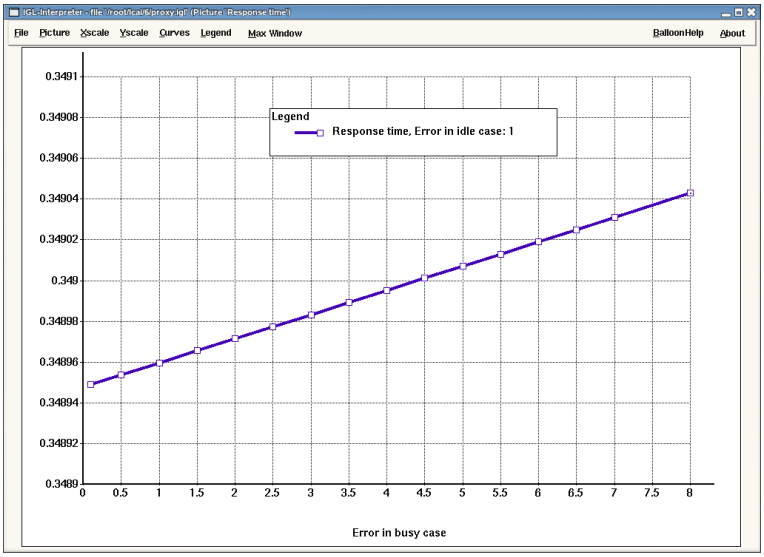


Figure 6: $F = 5000$ bytes, $\Lambda = 10$, $\nu = 10$, and $\delta = 0.2$

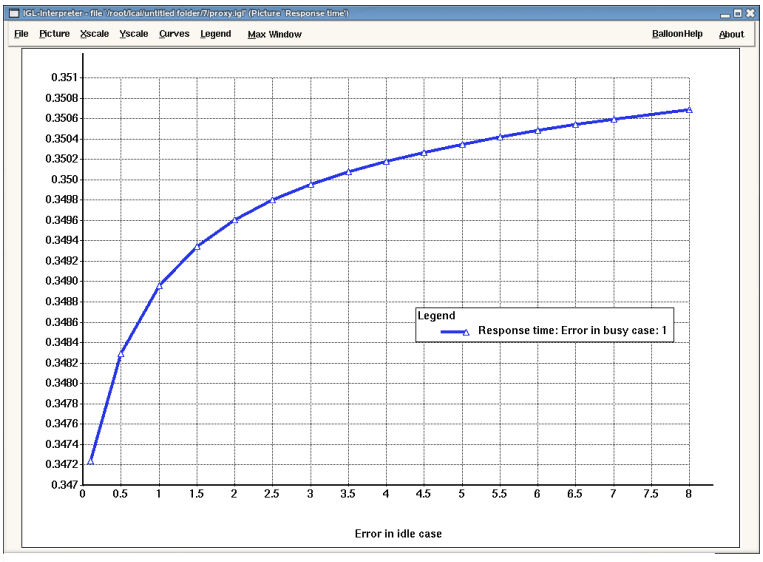


Figure 7: $F = 5000$ bytes, $\Lambda = 10$, $\nu = 10$, and $\gamma = 0.2$

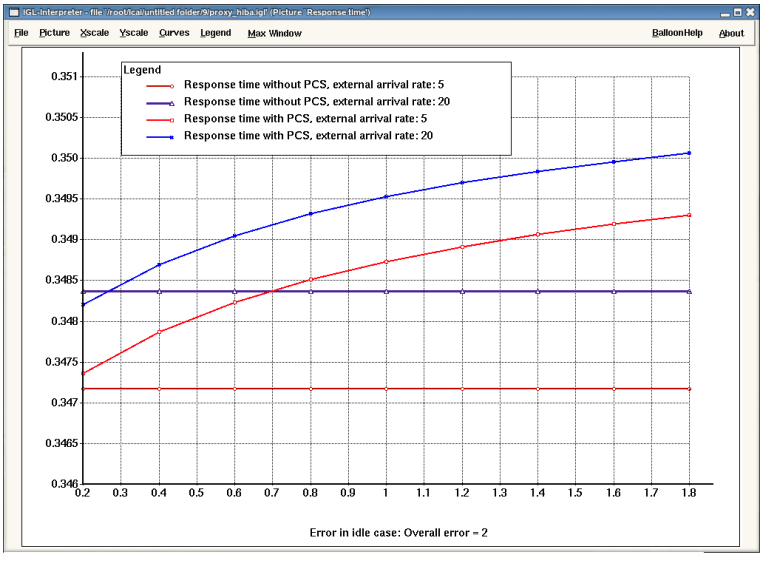


Figure 8: $F = 5000$ bytes, $\Lambda = 10$, $\nu = 10$, and $\gamma = 2 - \delta$

# 4. Conclusions

We modified the queueing network model of Bose and Cheng [1] to a more realistic case when external arrivals are allowed to the remote web server and the Proxy Cache Server is unreliable.

- In Figures 2 we can see the mean response time for the non-reliable system with and without installed a PCS depicted as a function of the server failure intensity. In this case the error intensity in booth case (idle,busy) is equal. As we see the mean response time increases when error rate increasing.
- In Figure 3 we investigate the effect of the mean repair time. As we see, the response time will be smaller as we increase the repair time.
- In Figure 4-5 we depicted the response time as a function of PCS Buffer size. In Figure 4 we use 30 requests/s for arrivals from the PCS. In Figure 5 we use the same parameters, only we use a higher arrival rate from PCS (40 requests/s). When we use smaller arrival rate we get smaller response time. But in this case the response time will be smaller with PCS only when the Buffer size is smaller than 6. When we use higher arrival rate ($\lambda = 40$) the existance of a PCS effect lower response time only when the buffer size is smaller then 10.
- In Figure 6-8 the effect of the failure rate is demonstrated on the response time, in busy and idle server states. In Figure 9 we plot the response time as a function of error mean with idle server state and using the restriction that the overall error is equal with 2.

# References

[1] Bose, I., Cheng, H. K., Performance models of a firms proxy cache server, *Decision Support Systems and Electronic Commerce 29.,* (2000), 45–57.

[2] Slothouber, L. P., A model of Web server performance, *5th International World Wide Web Conference,* Paris, France, (1996).

[3] Lazowska, E. D., Zahorjan, J., Graham, G. S., Sevcik, K. C., Quantitative System Performance, *Prentice Hall,* (1984).

[4] Menasce, D. A., Almeida, V. A. F., Capacity Planning for WebPerformance: Metric, Models, and Methods, *Prentice Hall*, (1998).

[5] Lashinsky, A., Suddenly cache is king the world of net stocks, *Fortune,* (1999), 370–372.

[6] Begain, K., Bolch, G., Herold, H., Practical performance modeling, application of the MOSEL language, *Kluwer Academic Publisher,* Boston, (2001).