Annales Univ. Sci. Budapest., Sect. Comp. 25 (2005) 153-169

THE IMPACT OF MULTIMEDIA TRAFFIC ON THE PERFORMANCE OF PROXY CACHE SERVER

T. Bérczes (Debrecen, Hungary)
J. Sztrik (Debrecen, Hungary)
C.S. Kim (Wonju, Korea)

Abstract. An open Jackson-type multi-class queuing network model is proposed to study the impact of multimedia traffic on the overall response times to Web requests for cases with and without a proxy cache server (PCS). 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. Numerical results showed that an increase of multimedia traffic percentage significantly impact the response times for both cases with and without a PCS. Several numerical examples illustrate the effect of arrival, external arrival rate, multimedia and non-multimedia file sizes on the mean response times.

1. Introduction

One of the major reasons the popularity of the Web is the ability to access multimedia content including sound, image and video files linked together in the form of hypermedia, resulting in a significant increase of multimedia traffic. The advent of the Web has prompted new standards and protocols (e.g. MIDI, MP3, MBone, VRML, etc.) for handling multimedia. Multimedia traffic is characterized by comparatively much larger file sizes and thus contributes to more network congestion. The focus of recent researches is to examine the

The research is partially supported by KOSEF-Hungarian Academy of Sciences Bilateral Scientific Cooperation (grant KOSEF F01-2004-000-100510-0), 2004.

performance of a network containing PCS in the presence of richer multimedia content.

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. For the easier understanding of the basic model and comparisons we follow the structure of the cited work. In Section 2 we construct a multi-class queuing network model to study the dynamics of installing a PCS in the presence of multimedia content. Overall response-time formulas are developed for both the case with and without a PCS. In Section 3 numerical experiments are conducted to examine the response-time behavior of the PCS with respect to various parameters of the model. Concluding remarks can be found in Section 4.

2. An analytical model of PCS involving multimedia traffic

In this section we briefly describe the mathematical model with the sug-

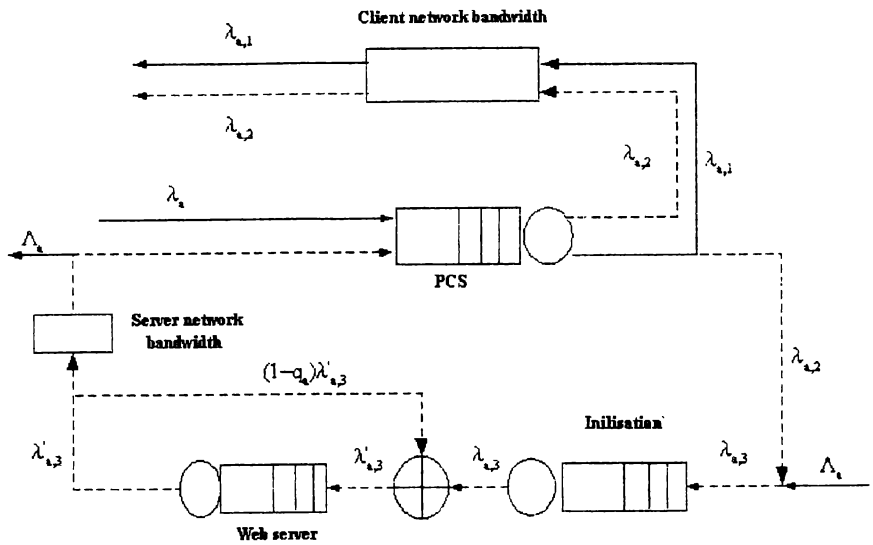

Figure 1.

gested modifications. If information or file is requested to be downloaded then first it is checked whether the document exists on the proxy cache server. (We denote the probability of this existence by p_a in case of a multimedia file, and by p_b otherwise) If the document can be found on the PCS then its copy is immediately transferred to the user. In the opposite case the request is transferred to the remote web server. After the requested document arrived to the PCS then the copy of it is delivered to the user.

The advantage of a PCS depends on several factors. These factors are: the probability of the "cache hit rate" of the PCS, the speed of the PCS, the bandwidth of the firm's network connection, the speed of the remote web server and the bandwidth of the remote network of the web server [1], [2].

Fig. 1 illustrates from start to finish the fulfillment of requests for multimedia files, denoted by subscript a. The notations used in this model are collected in Table 1.

We assume that the requests of the PCS users arrive according to a Poisson process with rate of λ_a, and the exogenous arrivals at the remote web server form a Poisson process too, with rate Λ_a.

Let F_a the average multimedia file size. We define $\lambda_{a,1}$, $\lambda_{a,2}$ and $\lambda_{a,3}$, such that:

$$(1) \qquad \lambda_{a,1} = p_a * \lambda_a \quad \text{and} \quad \lambda_{a,2} = (1 - p_a) * \lambda_a \quad \text{and} \quad \lambda_{a,3} = \Lambda_a + \lambda_{a,2}.$$

The solid line in Fig. 1 represents the $\lambda_{a,1}$ traffic. That means, the requested file is available on the PCS and can be delivered directly to the user. The $\lambda_{a,2}$ traffic depicted by dotted line, represents those requests which could not be served by the PCS, because the desired document is not on PCS, therefore these requests must be delivered from the remote web server. The $\lambda_{a,2}$ traffic must establish a one-time TCP connection at first [9], [1], [2]. We denote I_s this initial setup. Naturally the web server serves not only the requests of the studied PCS, but it also serves requests of other external users.

Therefore, we assume that requests arrivals at the web server form a Poisson process with rate $\lambda_{a,3}$. 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 dynamic server rate R_s." The performance of the firm's PCS is characterized by the same parameters B_{xc}, Y_{xc} and R_{xc}.

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

$$(2) \qquad q_a = \min\left(1, \frac{B_s}{F_a}\right),$$

the probability means that the desired file can delivered for the first attempt. According to the conditions of equilibrium and flow balance theory of queuing networks

$$q_a * \lambda'_{a,3} = \lambda_{a,3}. \tag{3}$$

The response time for multimedia traffic (depicted in Fig. 1) is denoted by T_a^{xc}. The process for non-multimedia files are the same. Then we denoted the response time for non-multimedia traffic by T_b^{xc}. Then we get

$$\begin{aligned} T_a^{xc} = & \frac{1}{\frac{1}{I_{xc}} - (\lambda_a + \lambda_b)} + \\ & + p_a * \left\{ \frac{\frac{F_a}{B_{xc}} * \left(Y_{xc} + \frac{B_{xc}}{R_{xc}}\right)}{1 - \sum_{j=a}^{b} \lambda_{j,1} \frac{F_j}{B_{xc}} \left(Y_{xc} + \frac{B_{xc}}{R_{xc}}\right)} + \frac{F_a}{N_c} \right\} + \\ & + (1 - p_a) * \left\{ \frac{1}{\frac{1}{I_s} - (\lambda_{a,3} + \lambda_{b,3})} + \frac{\frac{F_a}{B_s} * \left(Y_s + \frac{B_s}{R_s}\right)}{1 - \sum_{j=a}^{b} \frac{\lambda_{j,3}}{q_j} \frac{F_j}{B_s} \left(Y_s + \frac{B_s}{R_s}\right)} + \frac{F_a}{N_s} + \right. \\ & \left. + \frac{\frac{F_a}{B_{xc}} * \left(Y_{xc} + \frac{B_{xc}}{R_{xc}}\right)}{1 - \sum_{j=a}^{b} \lambda_{j,2} \frac{F_j}{B_{xc}} \left(Y_{xc} + \frac{B_{xc}}{R_{xc}}\right)} + \frac{F_a}{N_c} \right\}, \end{aligned} \tag{4}$$

and

$$\begin{aligned} T_b^{xc} = & \frac{1}{\frac{1}{I_{xc}} - (\lambda_a + \lambda_b)} + \\ & + p_b * \left\{ \frac{\frac{F_b}{B_{xc}} * \left(Y_{xc} + \frac{B_{xc}}{R_{xc}}\right)}{1 - \sum_{j=a}^{b} \lambda_{j,1} \frac{F_j}{B_{xc}} \left(Y_{xc} + \frac{B_{xc}}{R_{xc}}\right)} + \frac{F_b}{N_c} \right\} + \\ & + (1 - p_b) * \left\{ \frac{1}{\frac{1}{I_s} - (\lambda_{a,3} + \lambda_{b,3})} + \frac{\frac{F_b}{B_s} * \left(Y_s + \frac{B_s}{R_s}\right)}{1 - \sum_{j=a}^{b} \frac{\lambda_{j,3}}{q_j} \frac{F_j}{B_s} \left(Y_s + \frac{B_s}{R_s}\right)} + \frac{F_b}{N_s} + \right. \end{aligned}$$

$$(5) \qquad + \frac{\frac{F_b}{B_{xc}} * \left(Y_{xc} + \frac{B_{xc}}{R_{xc}}\right)}{1 - \sum_{j=a}^{b} \lambda_{j,2} \frac{F_j}{B_{xc}} \left(Y_{xc} + \frac{B_{xc}}{R_{xc}}\right)} + \frac{F_b}{N_c} \Bigg\}.$$

Then the overall response time is

$$(6) \qquad T_{xc} = \frac{\lambda_a}{\lambda_a + \lambda_b} * T_a^{xc} + \frac{\lambda_b}{\lambda_a + \lambda_b} * T_b^{xc}.$$

These equations is follow from the multi-class mean value analyzes specified in [5]. The response time T_a^{xc} consists of three terms.

The first term is the time to check whether the requested file is on the PCS or not. This is derived from the waiting time in an $M/M/1$ queuing system where the arrival is Poisson process with $\lambda_a + \lambda_b$ rate and the service rate is $1/I_s$.

The second term is the response time in case the requested document exists on the PCS, which probability is p_a. The first item in this term is the waiting time of the multi-class queuing system on the PCS which follows from [5], where the numerator

$$\frac{F_a}{B_{xc}} * \left(Y_{xc} + \frac{B_{xc}}{R_{xc}}\right)$$

is the "service demand". The second item in the second term is the travelling time when the requested file goes through the client network bandwidth.

The third term is the response time when the requested file does not exist on the PCS. That event's probability is $(1 - p_a)$. This term consists of three terms, too. The first item is the initialization time of TCP connection between the PCS and the remote web server. The second item is the waiting time of the queuing system on the remote web server, where $\lambda_{j,3}/q_a = \lambda'_{j,3}$ and F_a/N_s is the expected time transferring the requested documents on the server network bandwidth. The third term is the waiting time of the PCS when the copy of the requested document is transferred to the user.

Eq. (4) represents the response time for non-multimedia traffic. When there is no PCS, the overall response time, T, is given by the same logic for Eqs. (7)-(9):

$$(7) \qquad T_a = \frac{1}{\frac{1}{I_s} - ((\lambda_a + \Lambda_a) + (\lambda_b + \Lambda_b))} +$$

$$+\frac{\frac{F_a}{B_s} * \left(Y_s + \frac{B_s}{R_s}\right)}{1 - \sum\limits_{j=a}^{b} \frac{\lambda_j + \Lambda_j}{q_j} \frac{F_j}{B_s} \left(Y_s + \frac{B_s}{R_s}\right)} + \frac{F_a}{N_s} + \frac{F_a}{N_c}$$

and

$$T_b = \frac{1}{\frac{1}{I_s} - ((\lambda_a + \Lambda_a) + (\lambda_b + \Lambda_b))} + \tag{8}$$

$$+\frac{\frac{F_b}{B_s} * \left(Y_s + \frac{B_s}{R_s}\right)}{1 - \sum\limits_{j=a}^{b} \frac{\lambda_j + \Lambda_j}{q_j} \frac{F_j}{B_s} \left(Y_s + \frac{B_s}{R_s}\right)} + \frac{F_b}{N_s} + \frac{F_b}{N_c}.$$

Then the overall response time without a PCS is

$$T = \frac{\lambda_a}{\lambda_a + \lambda_b} * T_a + \frac{\lambda_b}{\lambda_a + \lambda_b} * T_b. \tag{9}$$

Examining Eqs. (4)-(9) could see there are same case when the response times go to infinity. These cases occur when the denominator is zero. Let $\lambda_b/\lambda_a = m$ the proportion of non-multimedia and multimedia requests. So, when one of the given equations below is realized then the overall response time will go to infinity.

$$\lambda = \frac{1}{I^{xc}},$$

$$\lambda_{a,1} = \frac{B_{xc}R_{xc}}{(F_a + mF_b)(Y_{xc}R_{xc} + B_{xc})},$$

$$\lambda_{a,2} = \frac{B_{xc}R_{xc}}{(F_a + mF_b)(Y_{xc}R_{xc} + B_{xc})},$$

$$\lambda_{a,3} + \lambda_{b,3} = \frac{1}{I_s},$$

$$\lambda_{a,3} = \frac{q_a q_b B_s R_s}{(q_b F_a + m q_a F_b)(Y_s R_s + B_s)},$$

$$\lambda + \Lambda = \frac{1}{I^s},$$

$$\lambda_a + \Lambda_a = \frac{q_a q_b B_s R_s}{(q_b F_a + m q_a F_b)(Y_s R_s + B_s)}.$$

Now we give a value for $\lambda_{\max}$ and for $(\lambda + \Lambda)_{\max}$ using $\lambda_{a,1}, \lambda_{a,2} < \lambda$ and $\lambda_{a,3}, \lambda_a + \Lambda_a < \lambda + \Lambda$. The given value will not be the best limit, but that value could be count using only the server parameters.

$$\lambda_{\max} = \min\left(\frac{1}{I_{xc}}, \frac{B_{xc}R_{xc}}{(F_a + F_b)(Y_{xc}R_{xc} + B_{xc})}\right)$$

and

$$(\lambda + \Lambda)_{\max} = \min\left(\frac{1}{I_s}, \frac{q_a q_b B_s R_s}{(q_b F_a + q_a F_b)(Y_s R_s + B_s)}\right).$$

The first equation uses only the solutions for PCS and the second equation uses the solutions for remote web server.

3. Numerical results

For the numerical explorations the corresponding parameters of Cheng and Bose [1], [2] are used. The file sizes are class 0 and class 1 file taken from [1], [2] and [6] ($F_a = 7000$ bytes and $F_b = 1000$ bytes). The value of other parameters for numerical explorations were: $I_s = I_{xc} = 0.004$ 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.

In all Figures the dotted lines plot the case with a PCS and the normal line depicts the case without a PCS.

3.1. Effect of arrival rate

In Fig. 2 the response time is depicted as a function of arrival rate. In this Figure the percentage of multimedia files is 10% and the external rate is 100 requests/s. The cache hit rate in the case of both multimedia and non-multimedia files is 0.25. When λ is smaller than 60 requests/s the response time is larger with a PCS than without a PCS. When the arrival rate is greater than 60 the response time is smaller when a PCS is installed. In Fig. 3 we used the same parameters, only the multimedia percentage was 20%. In this case the existence of the PCS results a smaller response time when $\lambda > 25$. When we use a higher cache hit rate for multimedia files (Fig. 4, $p_a = 0.4$) the efficiency of PCS is clear. In this case the response time with PCS will be smaller than the response time without a PCS for any value of the arrival rate. In the other hand, when we use a smaller cache hit rate for multimedia files (Fig. 5, $p_a = 0.1$) the response time with a PCS will be smaller than without

a PCS only when the arrival rate is greater than 60 requests/s. So, we can see that the performance of a PCS depends on a high scale of the firms behavior, but when the requests/s from the firm is greater than 60, then there is enough a small cache hit rate for multimedia files to access a smaller response time.

3.2. Effect of external arrival rate

Now we investigate the effect of external arrival rate. In Fig. 6 the arrival rate from the PCS is 10 requests/s, the percentage of multimedia files is 30% and the cache hit rate for both multimedia and non-multimedia files is 0.25. We can see that the PCS will be efficient when the external arrivals are greater than 70 requests/s. In Fig. 7 we modified only the arrivals from the PCS to 30 requests/s. In this situation it is enough to have 50 external requests/s that the response time with PCS will be smaller than without a PCS. When the cache hit rate for non-multimedia files is larger ($p_b = 0.5$) (Fig. 8) then the response time with a PCS will be smaller, independently of the number of external arrivals. From the observation of Fig. 6-8 we can find that in general the response time with and without a PCS increase when the external arrival rate is increased. When the arrival rate of the studied firm is modest (10 requests/s) then the PCS's benefit will be visible when the external arrival rates are bigger than 70 requests/s. Increasing the multimedia traffic percentage the existence of a PCS will be more pronounced. For example, in Fig. 7 (30% multimedia traffic) it is enough to have 50 external requests/s to realize a small advantage.

3.3. Effect of multimedia and non-multimedia file size

Fig. 9-11 depict the overall response time as a function of multimedia file size and Fig. 12 shows the effect of non-multimedia file size. In Fig. 9 we use a small cache hit rate for multimedia files ($p_a = 0.1$). The cache hit rate p_b for non-multimedia files remains 0.25. The percentage of multimedia files is 20%. The arrivals from the PCS is 30 requests/s and the external arrivals $\Lambda = 100$ requests/s. Now, we can see that with these parameters the file size has no considerable effect on the response time. Increasing the multimedia cache hit

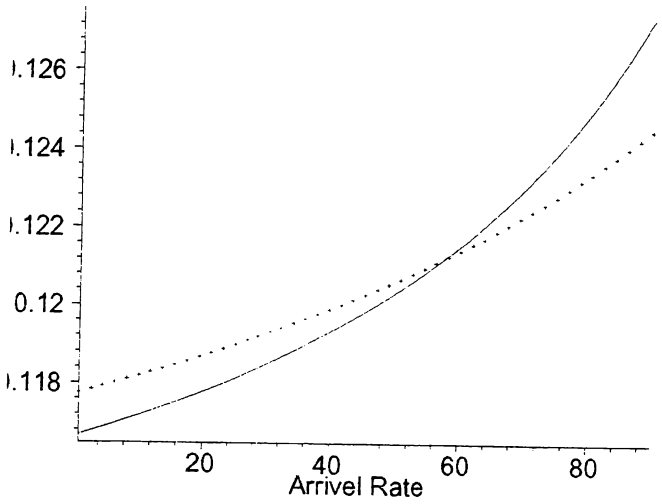

Figure 2. 10% multimedia, $\Lambda = 100$, $p_a = p_b = 0.25$, $F_a = 7000$ bytes, $F_b = 1000$ bytes

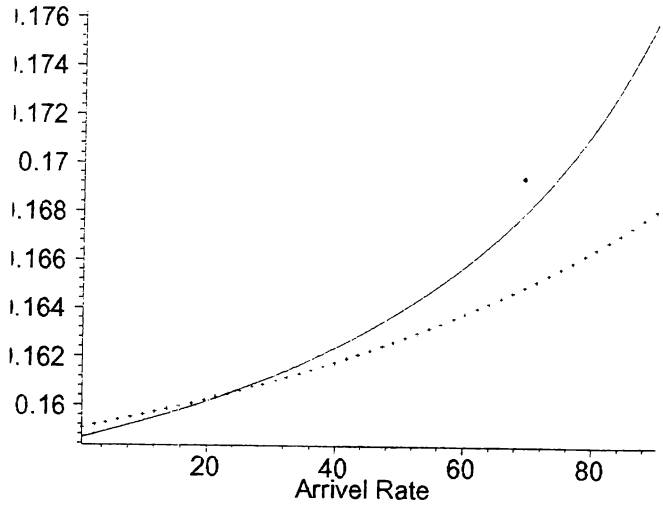

Figure 3. 20% multimedia, $\Lambda = 100$, $p_a = p_b = 0.25$, $F_a = 7000$ bytes, $F_b = 1000$ bytes

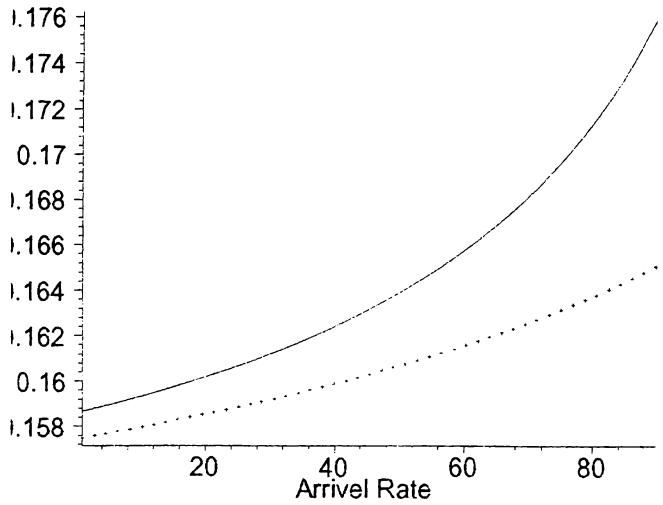

Figure 4. 20% multimedia, $\Lambda = 100$, $p_a = 0.4$, $p_b = 0.25$, $F_a = 7000$ bytes, $F_b = 1000$ bytes

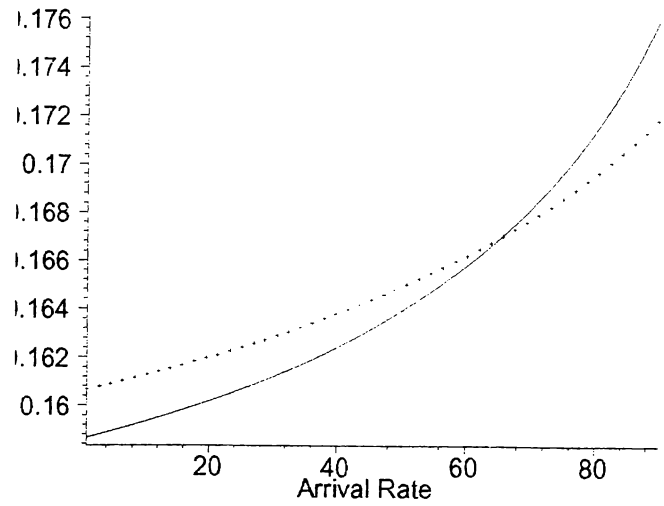

Figure 5. 20% multimedia, $\Lambda = 100$, $p_a = 0.1$, $p_b = 0.25$, $F_a = 7000$ bytes, $F_b = 1000$ bytes

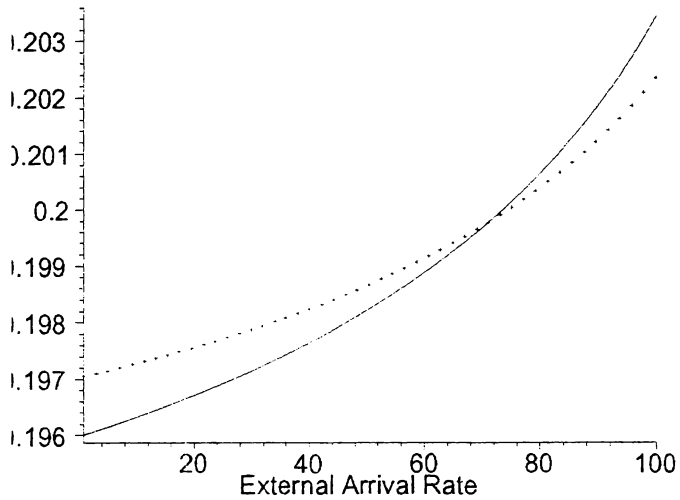

Figure 6. 30% multimedia, $\lambda = 10$, $p_a = p_b = 0.25$, $F_a = 7000$ bytes, $F_b = 1000$ bytes

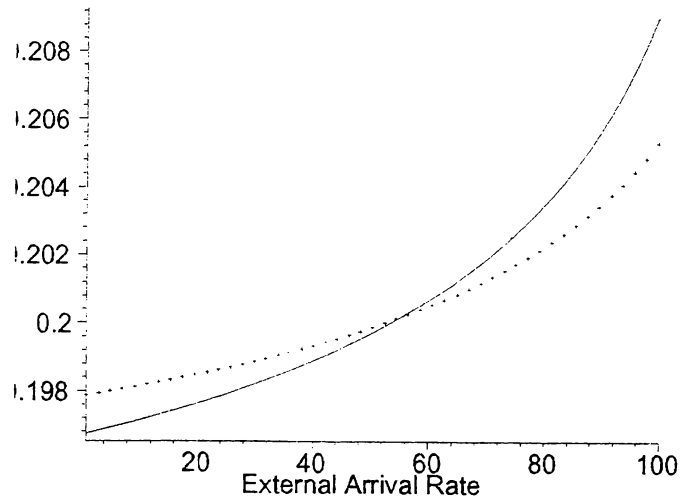

Figure 7. 30% multimedia, $\lambda = 30$, $p_a = p_b = 0.25$, $F_a = 7000$ bytes, $F_b = 1000$ bytes

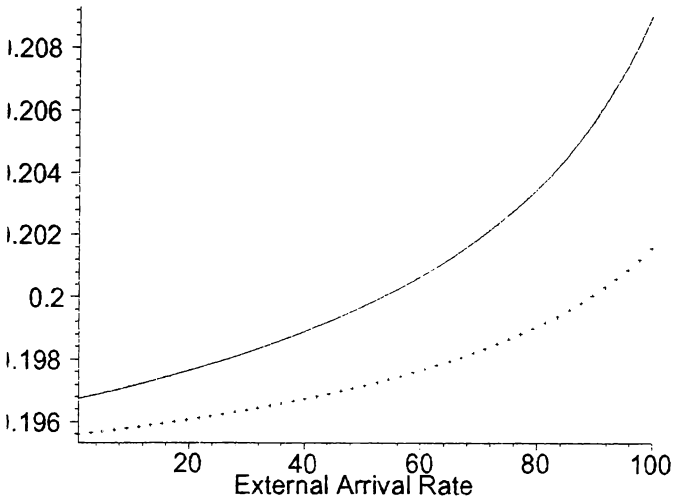

Figure 8. $\lambda = 30$, $p_a = 0.25$, $p_b = 0.5$, $F_a = 7000$ bytes, $F_b = 1000$ bytes

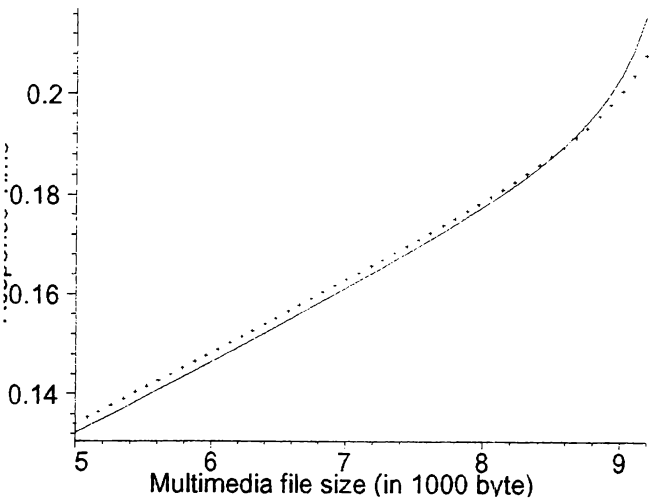

Figure 9. 20% multimedia, $\lambda = 30$, $\Lambda = 100$, $p_a = 0.1$, $p_b = 0.25$, $F_b = 1000$ bytes

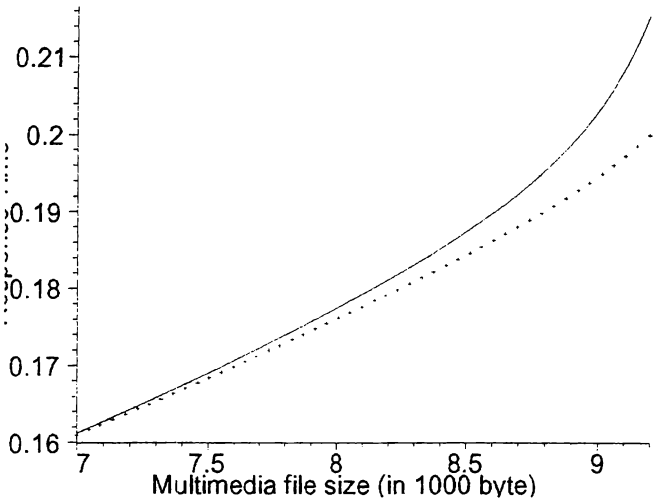

Figure 10. 20% multimedia, $\lambda = 30$, $\Lambda = 100$, $p_a = 0.25$, $p_b = 0.25$, $F_b = 1000$ bytes

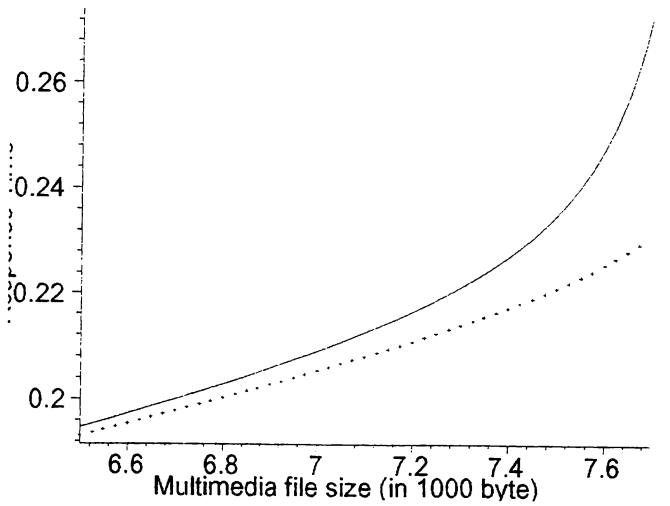

Figure 11. 30% multimedia, $\lambda = 30$, $\Lambda = 100$, $p_a = 0.25$, $p_b = 0.25$, $F_b = 1000$ bytes

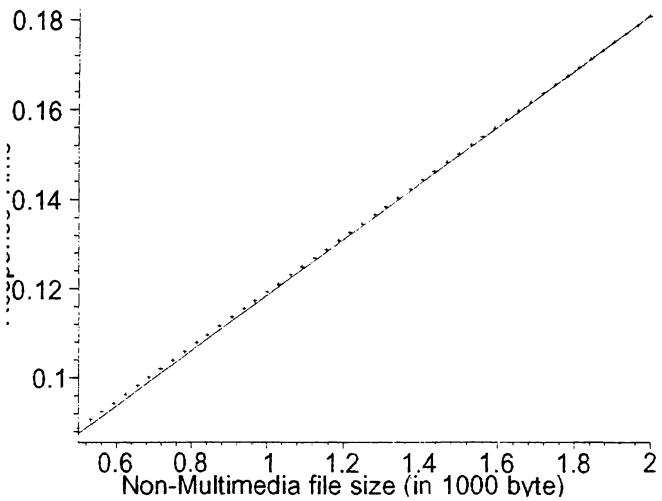

Figure 12. 10% multimedia, $\lambda = 30$, $\Lambda = 100$, $p_a = 0.25$, $p_b = 0.25$, $F_a = 7000$ bytes

rate ($p_a = 0.25$), the response time with a PCS is smaller than without a PCS, when the multimedia file size is more than 7500 bytes (Fig. 10). When we increase the percentage of multimedia files from 20% to 30% it is enough that the multimedia file size be 6000 bytes to access a smaller response time with PCS than without a PCS (Fig. 11).

Examining Fig. 12, we can see that there is no considerable effect of non-multimedia file size when the percentage of the multimedia files is 10%.

So, the multimedia file size has no considerable effect when the cache hit rate is higher or the percentage of multimedia file size is at least 30%. In these cases increasing the multimedia file size the response time with a PCS is smaller than without a PCS.

Table 1. Notations

λ_a : arrival rate of multimedia files,

λ_b : arrival rate of non – multimedia files,

Λ_a : external arrival rate of multimedia files,

Λ_b : external arrival rate of non – multimedia files,

F_a : average file size of multimedia files (in bytes),

F_b : average file size of non – multimedia files (in bytes),

p_a : cache hit rate for multimedia files,

p_b : cache hit rate for non – multimedia files,

B_{xc} : PCS output buffer (in bytes),

I_{xc} : lookup time of the PCS (in seconds),

Y_{xc} : static server time of the PCS (in seconds),

R_{xc} : dynamic server time of the PCS (in bytes/second),

N_c : client network bandwidth (in bits/second),

B_s : Web output buffer (in bytes),

I_s : lookup time of the Web server (in seconds),

Y_s : static server time of the Web sever (in seconds),

R_s : dynamic server time of the Web server (in bytes/second),

N_s : server network bandwidth (in bits/second).

4. Conclusion

We modified the multi-class queuing network model of Bose and Cheng [1] to a more realistic case when external arrivals are allowed to the remote web server. To examine this model we conduct numerical experiments adapted to realistic realistic parameters. In general, when the arrival rate of requests increases, then the response times increase as well regardless the existence of PCS. But in contrast with [1] when external arrivals are allowed to the remote web server, the PCS is beneficial with a low percentage of multimedia traffic and a low multimedia cache hit rate. When we use a high percentage of multimedia content and a high arrival rate, then the response time gap is more significant between the cases with and without a PCS.

To compare the two models we examined the effect of the external arrival rate. With low external arrival rate installing a PCS results to a higher response time. Increasing the external arrival rate, the difference between response

time with and without a PCS is smaller and smaller, then this difference vanished and the existence of a PCS results a lower response time. Using a low percentage of multimedia files and a low arrival rates from the firm we can access a slight benefit installing a PCS when the external arrivals are high. However, a slight improvement of non-multimedia cache hit rate speeds up the Web access.

Examining our numerical results it is clear that allowing external arrivals we get a more realistic model. To decide whether to install a PCS or not in order to speed up the Web access will be easier by using the demonstrated numerical results.

References

[1] **Cheng H.K. and Bose I.,** Performance models of a proxy cache server. The impact of multimedia traffic, *European Journal of Operational Research,* **154** (2004), 218-229.

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

[3] *CacheFlow White Papers* (available from http://cacheflow.com/technology/wp), CACHEFLOW INC., 1999.

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

[5] **Lazowska E.D., Zahorjan J., Graham G.S. and Sevcik K.C.,** *Quantitative system performance,* Prentice Hall, 1984.

[6] **Menasce D.A. and Almeida V.A.F.,** *Capacity planning for web performance: Metric, models and methods,* Prentice Hall, 1998.

[7] **Rubenstein R., Hersch H.M. and Ledgard H.F.,** *The human factor: Designing computer systems for people,* Digital Press, Burlington, MA, 1984.

[8] **Zhao J.L. and Kumar A.,** Data management issues for large scale, distributed workflow systems on the internet., *ACM SIGMIS Data Base,* **29** (4), 22-32.

[9] **Slothouber L.P.,** A model of Web server performance. *5th Int. World Wide Web Conf., Paris, France, (1996).*

(Received February 9, 2005)

T. Bérczes and J. Sztrik
Department of
Informatics Systems and Networks
University of Debrecen
H-4010 Debrecen, P.O.B. 12
Hungary
tbeczes@inf.unideb.hu
jsztrik@inf.unideb.hu

C.S. Kim
Sangji University
Wonju, Korea
dowoo@sangji.ac.kr