早稲田商学第 389 号
2 0 0 1 年 6 月

Performance Modeling a Web-Server Access Operation with Proxy Server Caching Mechanism

Yoshitaka TAKAHASHI

Abstract In an electronic / internet commerce, a tremendous increase in WWW (world wide web) traffic may deteriorate the web-server performance. Performance evaluation of the web-server access operation is requested to guarantee the quality of real-time services. In this article, we propose a web-access operation by using a proxy-server with popularity-degree-based cache. We model the proxy-server caching mechanism via Zipf's law. We then analyze the proposed operation via a modified $M/GI/1$ queueing model. Some numerical examples are provided to show the web-access performance.

key words: electronic commerce, web-access operation, proxy-server, caching mechanism, performance evaluation, queueing analysis.

1 Introduction

A tremendous increase in internet users for WWW (world wide web) services frequently causes insufficient communication facilities and resources. Very recently multimedia contents become more and more common, and the internet traffic monotonically and rapidly increases due to ordinary characters

(texts), codes, voices, and moving pictures. This increasing traffic leads to the deterioration of web-access response times. The response time deterioration should be promptly solved because of some web applications for a real-time electronic commerce. Performance analysis of web-server access operation is requested to guarantee the quality of real-time services.

The web-access response-time deterioration may come from two bottlenecks. One bottleneck is seen in communication channels. The channel speed is too slow and its capacity is insufficient. The other one is seen in web servers. The server processing capacity is also insufficient for some web application traffic. To remove these bottlenecks is essentially needed to solve the performance deterioration problem, but it generally requires much time and expensive cost.

The main goal of this article is i) to solve the web-access performance deterioration by proposing an operation with a proxy server and caching mechanism, and ii) to present the Laplace-Stieltjes transform (LST) formula for the response time distribution. In our operation, we use a popularity-degree-based caching. As in Daigo et al[(2)]. we assume that the popularity-degree frequency follows Zipf's law. We also assume that the inter-arrival time of the web-access requests is exponentially distributed as in Nabe et al[(6)].

Modeling our web-server access operation together with the stochastic assumptions described above then leads to a modified $M/GI/1$ queueing system where the service time is probabilistically zero. However, the standard $M/GI/1$ queueing analysis requires that the service time is always positive; see Refs[(1)(3)(4)(7)]. Here, we will take the approach[(8)(9)] for deriving the conditional waiting time distribution LST, given that an arriving customer requires positive service time and the total waiting time distribution LST by removing the

condition.

The rest of this article is organized as follows. In Section 2, we propose a web-server access operation with popularity-degree-based caching. Section 3 describes our performance modeling. In Section 4. we define the service time in the operation and assume that the service time is independent, and identically distributed (iid) to derive performance measures. Taking the approach as in Takahashi et al[8][9]. enables us to obtain the web-access response time LST formulas by using the modified $M/GI/1$ queueing system. Some numerical examples and comparisons are provided in Section 5. Section 6 finally gives a concluding remark.

2 Web-server access operation

We show our web-access operation with an introduction of a proxy server and caching mechanism. To be more specific,

(1) A client requests contents (an object) to the proxy server.

(2) The proxy server retrieves the contents (via the directory tree who manages the cache object) in itself (the proxy server). If it cannot find the contents, the proxy server makes the client enter the wait-state to transmit the request to the web server. Otherwise (if it finds the contents in the proxy server), the operation procedure goes to (4).

(3) If the proxy server receives the contents (the object) from the web server, it stores (caches) in the proxy server.

(4) The proxy server transmits the contents (the object) to the client.

Assuming the proxy server with caching mechanism described above is expected to reduce the response time at the web server, if we compare the web server operation without proxy server and caching.

There are typical caching methods at the proxy server, LRU (Least-Recently-Used), SIZE, and FREQUENCY; see Nabe et al[6]. for these methods and simulation comparisons. In this article, we adopt another caching method using a popularity degree, described in the following section.

3 Performance modeling

3.1 Web-access request modeling

A web access is reflected by client's favor and the request frequency has a bias. This biased situation implies that polularity does exist in individual contents. Zipf's law is now applied to capture the effect of this popularity degree in the web-access frequency.

Let f_i be the frequency where a web-access request has popularity degree i. Zipf's law leads to

$$f_i=\frac{c}{i^k} \quad (i=1, 2, \cdots, d). \tag{1}$$

for some constant k. Here, c is the normalization factor, given as

$$c=\left\{\sum_{i=1}^{d}\frac{1}{i^k}\right\}^{-1}, \tag{2}$$

and d is the supremum number of contents ($d\leq\infty$).

It should be noted that Zipf's law is far from any uniform distribution and so it has a bias. Zipf's law is well congruent with statistical data of web-access frequency; see e.g. Daigo et al[2].

The inter-arrival time of web-access requests is assumed to be exponentially distributed with parameter λ, as in Nabe et al[6]. The web-access requests with popolarity degree i then form a Poisson process with parameter λ_i ($i=1, 2, \cdots, d$).

3.2 Stochastic behavior of proxy server and web server

The proxy server and web server behave as follows.

(a) The proxy server checks its popularity degree when it receives a client request.

(b) If the popularity degree i is smaller than or equal to cache size, K, the proxy server sends the requested page (contents, object) immediately to the client.

(c) If the popularity degree i is larger than K, the proxy server transmits the page request signal to the web server. The transmission time from the proxy server to the web server is assumed to be zero.

(d) On receiving the page request signal from the proxy server, the web server checks the popularity degree in the web data base and transmits the corresponding page(s) back to the proxy server. The transmission time from the web server to the proxy server is assumed to be independent, and identically distributed (iid). We denote by $T^*(s)$ the Laplace-Stieltjes transform (LST) of the transmission time from the web to the proxy, $T^*(s)=\int_0^\infty e^{-st}dP(T\leq t)$. Let $t^{(n)}$ be the n-th moment of the transmission time. We have $t^{(n)}=(-1)^{(n)}T^{*(n)}(0)$ where $T^{*(n)}(0)$ is the n-th derivative of $T^*(s)$ at $s=0$

Here we have neglected the identification time of the contents version for simplyfing the analysis. This comes from the fact that control information is much less than contents information and the influence on the web-access performance is assumed to be very little. According to this simplifying spirit, we have also neglected the processing times of both web and proxy servers, as well as the transmission time of a page request signal in (*c*).

4 Queueing analysis

The web-access response time is defined as the sojourn time from the client request epoch until the client receives the requested page (contents, object). This section is devoted to the web-access response time analysis by using the queueing theory. Let W [or respectively, R] be the the web-access waiting time [response time] of a customer (client's web-access request). We denote by $W^*(s)$ [$R^*(s)$] the LST of the waiting time [response time] distribution.

We are now in a position to think about the service time of a customer. Here, we define the service time, B, by the elapsed time from the epoch where the proxy server finds a client request until the epoch where the proxy server sends the requested contents (page, object) to the client. Let $B(t)$ be the service time distribution, $B(t)=P(B\leq t)$. We denote by $B^*(s)$ the LST and by $b^{(n)}$ the n-th moment of the service time distribution.

Assuming that the service time is iid, the web-access response time then corresponds to the sum of the waiting time and service time in a modified $M/GI/1$ queueing model with customer arrival rate λ. By *modified,* we mean that the service time can be zero, i.e.,

$$B(0)=\sum_{i=1}^{K} f_i, \quad 0 \leq B(0) \leq 1 \tag{3}$$

$$B^*_+(0)+B(0)=1 \tag{4}$$

where $B^*_+(s)=\int_{0+}^{\infty} e^{-st}dP(B\leq t)$. Namely, our service time distribution is not honest nor proper. Thus, we cannot directly apply the Pollaczek-Khintchine LST formula:

$$W^*(s)=\frac{s(1-\rho)}{s-\lambda+\lambda B^*(s)}, \tag{5}$$

where ρ is the traffic intensity given by

$$\rho=\lambda E(B)=\lambda b. \tag{6}$$

Recall that the standard $M/GI/1$ queueing analysis assumes that the service time distribution is honest or proper ($B(0)=0$ and $B^*_+(0)=1$); see Refs[1][3][4][7]. Thus, as in Takahashi et al[8][9]. we reconsider the thinning arrival process so that the service time distribution is honest (or proper) as follows:

(i) A customer (client request) arrives at the proxy server according to a Poisson process with parameter $\lambda_{eff}=\lambda(1-B(0))$.

(ii) The effective service time B_{eff} is iid with the LST

$$B^*_{eff}(s)=\frac{B^*_+(S)}{1-B(0)}. \tag{7}$$

Here, by definition of our service time and the stochastic assumptions in Section 3, we have

$$B^*(s)=B(0)+(1-B(0))T^*(s), \tag{8}$$

and

$$B^*_{eff}=T^*(s). \tag{9}$$

The effective service time is then seen to correspond to the transmission time from the web to the proxy.

This reconsideration does not change traffic intensity $[\rho=\lambda E(B)=\lambda_{eff}E(B_{eff})]$ but it enables us to obtain the conditional waiting time distribution LST $W^*_{eff}(s)$

as

$$W^*_{eff}(s)=\frac{s(1-\rho)}{s-\lambda(1-B(0))+\lambda B^*_+(s)}$$

$$=\frac{s(1-\rho)}{s-\lambda+\lambda B^*(s)}, \tag{10}$$

given that an arriving customer requires the effective service time. The total waiting time distribution LST $W^*(s)$ then follows:

$$W^*(s)=W^*_{eff}(s), \tag{11}$$

if an arriving customer requiring zero service time has to wait for its service under the first-in-first-out (FIFO) discipline; and

$$W^*(s)=B(0)+(1-B(0))W^*_{eff}(s), \tag{12}$$

otherwise (if an arriving customer requiring zero service time immediately receives its service under the preemptive-resume priority discipline; see Takagi[7]). In the latter case, we may say that the arriving customer with zero service time has the so-called preemptive-resume (PR) priority.

From Eq. (11) or (12) together with Eq. (10), the mean and higher moments of the web-access waiting time can be obtained as $E(W^n)=(-1)^n W^{*(n)}(0)$. For example, if an arriving customer requiring zero service time has to wait under the FIFO discipline, we have

$$E(W)=\frac{\lambda b^{(2)}}{2(1-\rho)}, \tag{13}$$

$$E(W^2)=\frac{\lambda b^{(3)}}{3(1-\rho)}+\frac{[\lambda b^{(2)}]^2}{2(1-\rho)^2}, \tag{14}$$

and otherwise (if an arriving customer requiring zero service time is immediate-

ly served under the PR priority discipline),

$$E(W)=\frac{\lambda(1-B(0))b^{(2)}}{2(1-\rho)}, \tag{15}$$

$$E(W^2)=\frac{\lambda(1-B(0))b^{(3)}}{3(1-\rho)}+\frac{[\rho\sqrt{1-B(0)}\ b^{(2)}]^2}{2(1-\rho)^2}. \tag{16}$$

It should be noted that if the arriving customer requiring zero service time has to wait under the FIFO discipline, the Pollaczek-Khintchine formulas are seen to be still valid even for $B(0)>0$ due to Eqs (5) and (10)—(11).

The response time distribution LST $R^*(s)$ is finally obtained as

$$R^*(s)=W^*_{eff}(s)\,B^*(s), \tag{17}$$

if an arriving customer requiring zero service time has to wait under the FIFO discipline, and

$$R^*(s)=B(0)+(1-B(0))\,W^*_{eff}(s)B^*_{eff}(s), \tag{18}$$

otherwise (under the PR priority discipline).

Thus, recalling Eqs (8) and (9), if an arriving customer requiring zero service time has to wait under the FIFO discipline, we have

$$E(R)=\frac{\lambda(1-B(0))t^{(2)}}{2(1-\rho)}+(1-B(0))E(T), \tag{19}$$

$$E(R^2)=\frac{\lambda(1-B(0))t^{(3)}}{3(1-\rho)}+\frac{[\lambda(1-B(0))t^{(2)}]^2}{2(1-\rho)^2}+\frac{\lambda(1-B(0))^2t^{(2)}}{1-\rho}E(T)$$

$$+(1-B(0))t^{(2)}. \tag{20}$$

Otherwise (under the PR priority discipline), we have

$$E(R)=\frac{\lambda(1-B(0))^2t^{(2)}}{2(1-\rho)}+(1-B(0))E(T), \tag{21}$$

and

$$E(R^2)=\frac{\lambda(1-B(0))^2t^{(3)}}{3(1-\rho)}+\frac{[\lambda(1-B(0))\frac{3}{2}t^{(2)}]^2}{2(1-\rho)^2}+\frac{\lambda(1-B(0))^2t^{(2)}}{1-\rho}E(T)$$

$$+(1-B(0))t^{(2)}. \tag{22}$$

As expectedly, under both the FIFO and PR priority disciplines, our formulas are all reduced to the well-known Pollaczek-Khintchine formulas for the honest service time distribution ($B(0)=0$ and $B^*(s)=B^*_{eff}(s)=T^*(s)$ in this case).

Now, going back to our web-access operation, we recall the assumption under which a client can promptly get the response from the proxy server, when its requesting popularity degree can be seen in the cache, i.e., the latter case where an arriving customer requiring zero service time preempts the current service and leaves the system immediately. From our application point of view, these PR priority (latter case) results [i.e., Eqs (12), (15)−(16), (18), and (21)−(22)] are of more practical importance. However, the FIFO (former case) results still have a potential applicability in the areas of teletraffic theory.

5 Numerical examples

In our numerical examples, we assume that the transmission time from the web to the proxy is geometrically distributed with a mean of two time units. Here, time unity is taken as the page transmission time. We then have

$$T^*(s)=\int_0^\infty e^{-st}dP(T\leq t)=\sum_{i=1}^{\infty}e^{-si}P(T=i)=\sum_{i=1}^{\infty}e^{-si}\left(\frac{1}{2}\right)^i=\frac{1}{2e^s-1},$$

$$E(T)=(-1)T^{*(1)}(0)=2,\ t^{(2)}=E(T^2)=(-1)^2T^{*(2)}(0)=6.$$

We further assume that the popularity-degree frequency follows Zipf's law with parameter $k=1$. The number of contents is limited to 10,000 ($d=$ 10,000). The traffic intensity ρ can be regarded as the channel utilization from the web server to the proxy server for the zero-cache-size case. Recall that the zero-cache-size case corresponds to the no-proxy-server situation.

By using Eq. (21) with Eq. (3), in the figure, we show three graphs of the normalized mean web-access (response) time as a function of cache size, K. Individual graph corresponds to the traffic intensity $\rho=0.4$, 0.6, or 0.8, respectively. By *normalized,* we mean that the web access response time is normalized by the page transmission time. Through these numerical examples, we see that the mean web-access response time can be drastically improved even for a not-so-large cache size.

6 Conclusions

We have proposed a web-server access operation using a proxy server with popularity-degree-based caching mechanism to improve the web-access performance issue arising out of real-time WWW application traffic. The modified $M / GI / 1$ queueing analysis together with Zipf's law has enabled us to obtain the LST formula of the web-access response time distribution. We have analitically seen that Pollaczek-Khintchine's formulas are still valid even for the dishonest service time distribution (where the service time is probabilistically zero), if an arriving customer requiring zero service time has to wait under the FIFO discipline. However, in our web-access operation, an arriving customer requiring zero service time preempts the current service and leaves the system

immediately, i.e., we may say that an arriving customer with zero-service time has the so-called preemptive-resume (PR) priority discipline. Thus, our PR priority results [i.e., Eqs (12), (15)—(16), (18), and (21)—(22)] are of more practical importance. We have numerically seen that the web-access response time can be drastically improved even for a not-so-large cache size.

References

(1) R. B. Cooper: *Introduction to Queueing Theory,* Second Ed., (Elsevier North Holland, New York, 1981).

(2) T. Daigo, K. Ohta, N. Kato, and Y. Nemoto: "A study of effective cache system based on the locality of user behavior," *IEICE Tech. Rep.,* **IN98-161**, pp. 35-42 (1999). (in Japanese)

(3) M. Fujiki and E. Gambe: *Teletraffic Theory,* (Maruzen, Tokyo, 1980). (in Japanese)

(4) K. Kawashima, F. Machihara, Y. Takahashi, and H. Saito: *Fundamentals of the Teletraffic Theory and Multimedia Networks,* (IEICE Press, Tokyo, 1995). (in Japanese)

(5) A. Luotonenn: *Web Proxy Servers,* (Prentice Hall Inc., New Jersey, 1998).

(6) M. Nabe, M. Murata, and H. Miyahara: "Analysis and modeling of WWW traffic characteristics with document caching," *IEICE Trans.,* **J81-B-I**, pp. 325-334 (1998). (in Japanese)

(7) H. Takagi: *Queueing Analysis, A Foundation of Performance Evaluation,* vol.1: Vacation and Priority Systems, Part1 (Elsevier Science Pub., B.V., Amsterdam, 1991).

(8) Y. Takahashi and O. Hashida: "Delay analysis of discrete-time priority queue with structured inputs," *Queueing Systems,* **8**, pp. 149-164 (1991).

(9) Y. Takahashi and S. Shimogawa: "Composite priority single-server queue with structured batch inputs," *Communications in Statistics-Stochastic Models,* **7**, pp. 481-497 (1991).

(10) Y. Takahashi and B. Krishna Kumar: "Pseudo-conservation law for discrete-time multi-queue systems with priority disciplines," *J. Operations Research Soc. Japan,* **38**, pp. 450-466 (1995).

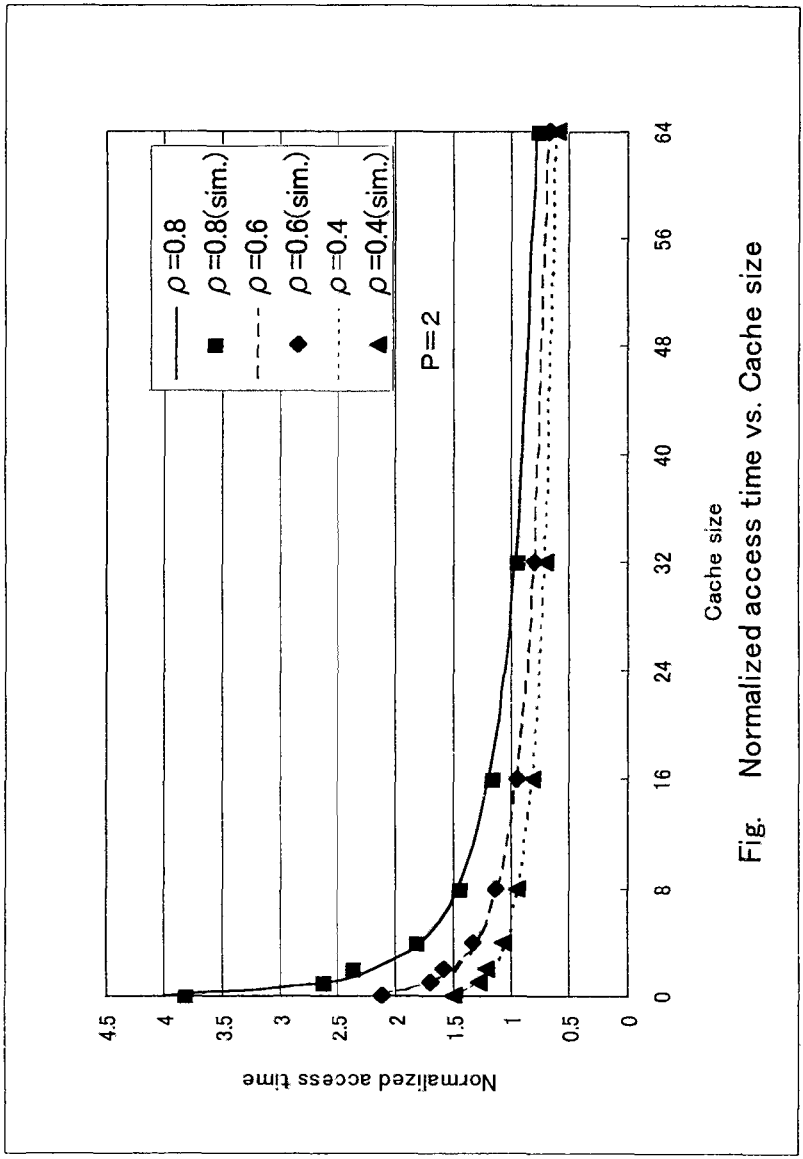

Figure: The mean web-server access (response) time.