I.J. Education and Management Engineering **2012, 9, 8-13**
Published Online September 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2012.09.02

Available online at http://www.mecs-press.net/ijeme

Approximate Performance analysis for the Double Servers in the Asynchronous Schedule Mechanism of Polling

[a]BAO Li-yong, [b] ZHAO Dong-feng, [c] ZHAO Yi-fan

School of Information Engineering, Yunnan University Kunming, 650091, China

Abstract

With the popularity of network, web is becoming one of the most effective ways in information sharing. However, with the increase of clicking index, a high-function server can hardly meet the increasing demand for service. This thesis provides the schedule strategy in the asynchronous mechanism of multiqueue double servers polling, which is conducted based on theory of polling multi-access and simulation experiments on computer, along with consideration to the load features of Web. It concludes that this strategy enables Web server to take on the features of good extendibility, utility and function.

Index Terms: Web server; multiqueue double servers polling; asynchronous schedule mechanism; approximate analysis Algorithm

1. introduction

With the popularity of network, traditional ways of information exchange have been widely conducted through Internet, and Web is becoming one of the most effective ways in sharing information and obtaining services. However, with the increase of clicking index, the quality of service is declining. The overloading of Web servers and the delayed service because of network congestion have become the predominant problems in Web service[1]. Consequently, a high-function computer can hardly meet the increasing demand for service[2-3]. Considering the situation, an effective solution to the problem lies with the application of Web server cluster technology so as to meet the demands of rapid response, high flexibility, high utility, easy manageability, and high function price ratio.Multiservers polling performance analysis on the multitoken-rings network, multiservers system resources scheduling allocation and load balance has important application prospect. However, every server has its own independent dispatch time that makes the mathematical analysis exceptionally complex [4-5]. At present the main way to solve the problem lies with an approximate analysis method. It is expected to find out the effective analytical method in this research. At present, the mathematical analysis model's establishment is still at the first stage. The analysis focuses on the double servers polling. In general, scheduling of two servers can be divided into synchronous and asynchronous ways. When scheduling mechanism uses the synchronous way, it means that the two servers reach the service node at the same time, and then leave the node after processing the information packets. If it uses asynchronous method, namely, when one

Corresponding author:
E-mail address: [a] lybao@ynu.edu.cn

server reaches the service node alone, the node is serviced by the other server, and then the server leaves the node immediately. Otherwise, it processes all the information packets in the node.

This research established the multiqueue double servers polling asynchronous scheduling model. The model is made up of double servers and N common stations. Approximate analysis is conducted based on queuing theory [6]. Under the stable condition of the model, Markov chain and probability generating function are applied to turn out the results of the mean queue length, the mean cyclic time, and the average time delay of information packets. Simulated experiments are made on the model.

2. The System Model of the Asynchronous Schedule Mechanism

2.1. The Mechanism and Variable Definition of the System Model

Polling services at each station include the following processes:

1) The arrival of the message packets waiting for transmission at each station follows the independent and identical distribution of probability, with the distributional probability generating function , mean value and variance respectively as $A(z), A'(1)=\lambda$ and $\sigma_\lambda^2 = A''(1)+\lambda-\lambda^2$.

2) The time of message packets transmission in the memory the common stations follows an independent and identical distribution of probability, with the distributional probability generating function , mean value and variance respectively as $B(z), B'(1)=\beta$ and $\sigma_\lambda^2 = A''(1)+\lambda-\lambda^2$

3) The transfer time when the server polls from the station to the next station follows an independent and identically distributed probability, with the distributional probability generating function , mean value and variance respectively as $R(z), \gamma = R'(1)$ and $\sigma_\gamma^2 = R''(1)+\gamma-\gamma^2$

4) The system operates on the FCFS rule. There is enough capacity of the memory at each terminal station so that information grouping will not get lost.

2.2. Probability Generating Function of the System Status Variable

$u_i(n)$ refers to the transfer time of the server transfer at the $i+1$ station via this time after finishing its service for the i station.

$\mu_j(u_i)$ refers to the number of information packets enter the j station $(j=1,2,\cdots,N)$ at the time of $u_i(n)$.

$v_i(n)$ refers to the time of the server's transmission service for the information packets at the i station.

$\eta_j(v_i)$ refers to the information packets entering the memory of the j station at the time of $v_i(n)$.

The server polls at the i $(i=1,2,\cdots,N)$ common station at the time of t_n, with the information packets waiting to be transmitted in the i station being $\xi_i(n)$, and the system status variable being $\{\xi_1(n),\xi_2(n),\cdots,\xi_N(n),\xi_c(n)\}$. The server provides transmission service for the i common station under the Gated service policy. After its service for the i station, the server starts to poll at the $i+1$ station via a transfer time of $u_i(n)$, with the information packets entering the j station $(j=1,2,\cdots,N)$ at the time of $u_i(n)$ being $\mu_j(u_i)$. If the memory of the $i+1$ station is not blank, the transmission of information packets is completed via Gated service rules. The server starts the transmission service for the information packets in the $i+1$ common station at the time of t_{n+1}. At this moment, the system status variable is $\{\xi_1(n+1),\xi_2(n+1),\cdots,\xi_N(n+1),\xi_c(n+1)\}$.

The distributed probability of the system status variable is $P[\xi_i(n) = x_i; i = 1, 2, ..., N]$, on stable condition of $\sum_{i=1}^{N} \lambda\beta = N\rho < 1$. The system status can be described with the Markov chain which is non-periodic and ergodic. Joint probability distribution is defined as $\lim_{n\to\infty} p[\xi_i(n) = x_i; i = 1, 2, ..., N] = \pi_i(x_1, x_2, ..., x_i, ..., x_N)$ (with $\rho = \lambda\beta$).

The probability generating function's basic concept already has made the detailed definition in [7]and [8]. The probability generating function of system status is:

$$G_{i+1}(z_1, z_2, \ldots, z_i, \ldots, z_N) = \lim_{n\to\infty} E[\prod_{j=1}^{N} z_j^{\xi_j(n+1)}] = \lim_{n\to\infty} E[\prod_{j=1}^{N} z_j^{\xi_j(n+1)}]$$

$$= R[\prod_{j=1}^{N} A(z_j)]G_i\left(z_1, z_2, ..., z_{i-1}, B\left(\prod_{j=1}^{N} A(z_j)\right), z_{i+1}, ..., z_N\right)$$

$$i = 1, 2, \cdots, N \qquad (1)$$

2.3. The Analysis of Mean Queue Length

Given that $g_i(j)$ is the average number of information packets number at the j station buffer when the i station starts service at the time of t_n, then

$$g_i(j) = \lim_{z_1,\cdots,z_N\to 1} \frac{\partial G_i(z_1, z_2, \cdots, z_N)}{\partial z_j}$$

$$i = 1, 2, \cdots N\ ; \quad j = 1, 2, \cdots N \qquad (2)$$

According to the definition of queue length, the mean queue length is drawn via derived function algorithm as:

$$g_{iL}(i) = \frac{\lambda_k \sum_{k=1}^{N} \gamma_k}{1 - \sum_{k=1}^{N} \hat{\beta}_k \lambda_k} \qquad k = 1,2,\cdots,N$$

$$\hat{\beta} = \frac{\beta}{2} \qquad (3)$$

2.4. The Analysis of the Average Time Delay of Information Packets

The average time delay of information packets refers to the period during which information packets enter the memory of the station and then wait to be served. With regard to the calculation of Note [7], the system's average time delay can be obtained via second-order local derivation of the system status.

Based on the operational mechanism of the system and the theory of queuing, the average time delay can be derived as:

$$\overline{W^*}_{Gi} = \frac{(1+\rho_i)\sum \rho_k}{2\sum \rho_k(1+\rho_k)} \left\{ \frac{\sum R_k''(1)}{\sum \gamma_k} + \right.$$

$$\frac{1}{1-\sum\rho_k}\left[\sum\beta_k^2A_k''(1)+\sum\lambda_kB_k''(1)+\left(\sum\gamma_k+1\right)\sum\rho_k\right.$$

$$\left.+\left(\sum\gamma_k-1\right)\frac{\sum\rho_k^2}{\sum\rho_k}\right]+\sum\gamma_k-\left.\frac{\sum\gamma_k^2}{\sum\gamma}\right\}+\frac{\rho_iA_i''(1)}{2\lambda_i^2}$$

$$i=1,2,\cdots,N \qquad (4)$$

With $\sum=\sum_{k=1}^{n}$. $\hat{\beta}=\frac{\beta}{2}$, $\rho=\lambda\hat{\beta}$

3. Theoretical Calculation, Simulated Experiment and Analysis

3. 1 Theoretical Calculation, Simulated Experiment

Based on theoretical analysis, computer simulation experiment is conducted on the multiqueue double servers polling system under Gated services rule. According to the characteristics of Web service transmission, it is set that the information packets arrival of each station at any time slot width follows the Poisson distribution, with the length of information packets being 10KB, the channel speed being 100Mbps, the inquiry transfer time being 10μs, the service response time being 500μs, and the system's time slot width being 10μs. Consequently, with $\beta_i=50$, $\lambda_i=\lambda$, $\gamma=1$,the system meets the stable condition of $\sum_{i=1}^{N}\rho_i<1$.

Identical parameters are applied in theoretical calculation and simulated experiment apply identical parameters. As illustrated in the Figures from 1 to 4, the features of the system model are analyzed in detail as follows.

3. 2 Analysis of the System Feature

Based on the justification of the rationality of the system model and system analysis, further analyses are made on the system feature:

1) From the system stability condition of $\sum_{i=1}^{N}\rho_i<1$, $0<\rho_i<1$, the findings of theoretical calculation correspond well with those of simulated experiments only when the arrival rate is small. Along with the increase of arrival rate, the error between theoretical value and simulative value increase subsequently, but the change process still has a preferable comparability.

2) Double servers can provide a kind of scalability structure to improve the system's compatibility. In other words, they can support heterogeneous forms, which, in the practical application, can increase the utilization ratio of equipment. All the indexes in Figures 3 and 4 show that that double servers can better support this function. Specially, along with the enhancement of the server's performance, the Mean Queue Length and the Average Time Delay of Information Packets will also be greatly improved even in the high load situation.

4. Conclusion

Each of the servers has its own independent dispatch time which makes exceptionally complex the theoretical analysis of the double servers in the asynchronous schedule mechanism of polling. This thesis provides the strategy in the asynchronous dispatch mechanism of multiqueue double servers polling. Approximate analysis is conducted based on the theory of polling multi-access and simulation experiments on computer, along with consideration to the load features of Web. The theoretical and simulated findings reveal that in the system, when a single server node number increases to double node, with the information packets arrival rate increase even more than the stability condition, the double server system can still operate properly. It is concluded that this strategy enables Web server to take on the features of good extendibility, good utility, and good function.

Acknowledgment

The authors would like to give their sincere thanks to the financial support by the Nation Natural Science Foundation of China(No.F0424104 and No.60362001).

References

[1] V. Cardellini, M. Colajanni, P.S. Yu, “Dynamic load balancing on Web-server systems,” IEEE Internet Computing, vol.3, pp.28–39, 1999.

[2] Luciano Bertini, Julius C.B. Leite, Daniel Mossé, “Power optimization for dynamic configuration in heterogeneous web server clusters,” Journal of Systems and Software vol.83, pp.585-598, April.2010

[3] Seungmin Baek, Hwakyung Rim, Sungchun Kim. “Socket-based RR scheduling scheme for tightly coupled clusters providing single-name images,” Journal of Systems Architecture, Volume 50, Issue 6, pp. 299-308, June. 2004.

[4] Lin Chuang ,YANG Shi qiang, “Performance Analysis of Scheduling Schemes in Multiserver Multiqueue Systems,” Acta Elacta Electronica Sinica, vol.28,pp. 17-20, May. 2000.

[5] Lin Chuang , “Performance Analysis of Request Dispatching and Selecting in Web Server Clusters,” Chinese J Computer, vol.23,pp. 500-508,May. 2000. (in Chinese)

[6] Takagi H. Analysis of Polling Systems[M]. Cambridge, MA: The M.I.T. Press, 1986.

[7] Zhao Dongfeng, Zheng Sumin, “Message Waiting Time Analysis for a Polling System with Gated Service,” Journal of China Institute of Communications, vol. 15, pp. 18-23, February. 1994. (in Chinese)

[8] Zhao Dongfeng, Zheng Sumin, “Analysis of a Polling Model with Exhaustive Service,” Acta. Electronica Sinica, vol.22, pp. 102-107, May.1994. (in Chinese)

[9] LEI YingChun, ZHOU You Ying, and L I J un, “Evaluation on High Performance L52Dispatcher,” Journal of Computer Research and Development, vol.40 ,pp.475-482, May.2003. (in Chinese)

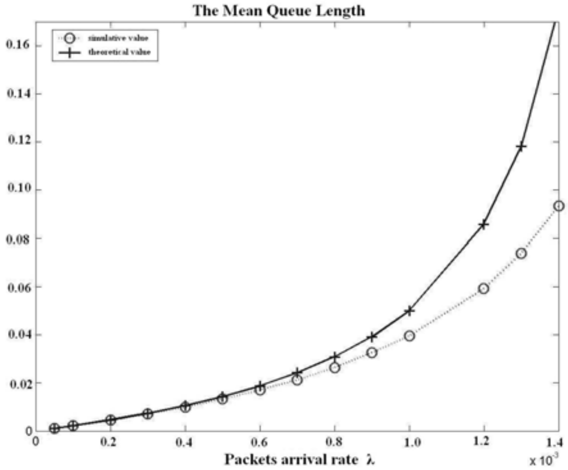

Figure 1. The mean queue length

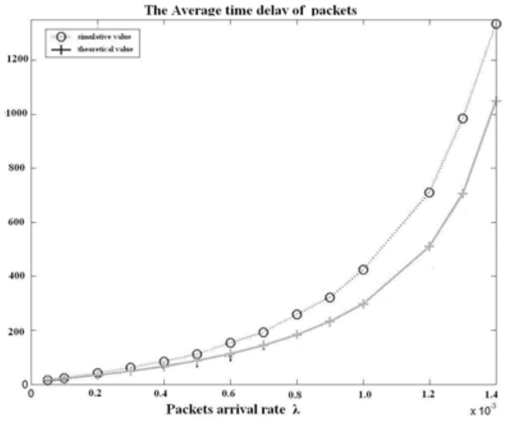

Figure 2. The average time delay of information packets

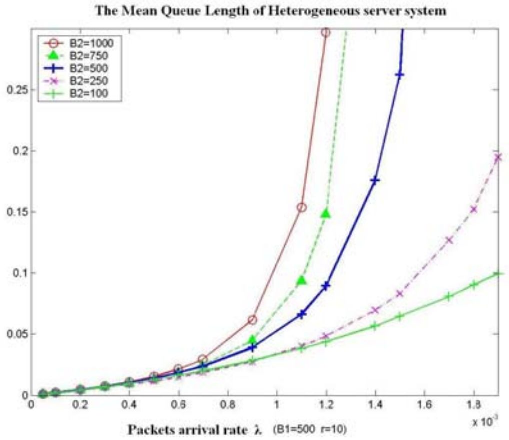

Figure 3. The Mean Queue Length of Heterogeneous server system

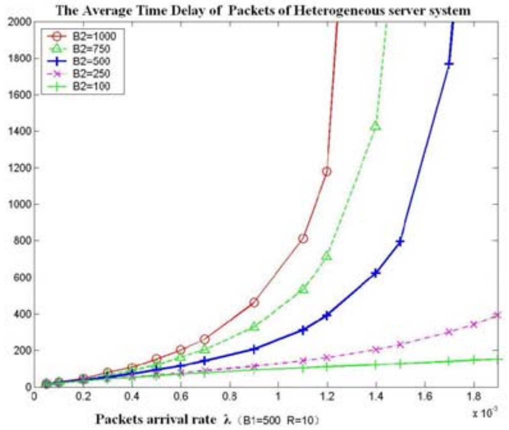

Figure 4. The average time delay of information packets of Heterogeneous server system