A Web Server Performance Analysis Based on Queuing Models

-Course Project for IK1611 (Dimensioning of Communication Systems)

Sun Gang, Liu Lu-an

Master Program in Internetworking, School of Information and Communication Technology

Royal Institute of Technology (KTH)

gangs@kth.se

laliu@kth.se

***Abstract* — This paper presents a detailed web server performance analysis for a company whose website has performance problems – their customers have been experiencing long response times and high rejection rates. In order to analyze and solve the problem, two queuing models are suggested: M/M/1/L and IPP/M/1/L. Relevant parameters for both queuing models are estimated, and a simulation for each model is performed with the corresponding parameters having been estimated. Comparing the results to reality, IPP/M/1/L is more suitable for modelling the web server. In the end of the paper, a suggestion on how to lower the response time and the rejection rate is proposed.**

I. INTRODUCTION

A company founded by some graduates from Royal Institute of Technology in Stockholm has a web site. The main idea of the company is to provide reviews for different commercial products over the Internet. The types of the reviews span over everything from cars through travel resorts to movies, music and games. The main income of the company is generated by advertisements on its web site. The data are provided through a web site which can be visited for free. The web server will collect the associated files from the corresponding database and provide the users with the information found at a request for a certain type of review.

A report shows that the average number of visitors per day is continuously growing ever since the company was founded. The business idea of the company relies on a high availability of the web site, so it is essential to keep the web site highly functional. However, the web site has been found to have performance problems. The customers who have visited the web site complains that they have been experiencing long response times and high rejection rates. The stuffs of the company tried to figure the problem out by themselves, but they finally failed due to the lack of relevant knowledge within the field of queuing systems. Therefore, they had to turn to us, a network consulting company for help.

This paper utilizes MATLAB [1] to make an intensive analysis on the performance of the web server based on queuing models. Two different queuing systems are first suggested, and then relevant system parameters for each of them are estimated based on the log files of the web server. After that, a simulation for each queuing model is performed, and one of the two queuing models is finally chosen by comparing the results of the simulations to reality. Based on the chosen queuing model, a solution is finally proposed.

II. PERFORMANCE ANALYSIS USING QUEUING MODELS

We use two queuing models to present the performance analysis: M/M/1/L and IPP/M/1/L. The software MATLAB is chosen to be used as the computer-aided tool for statistics and simulation because it is very convenient to make calculations and plots in MATLAB. 12 log files of the server are provided, which can be grouped as four equally probable typical days: d1 = (v1, w1, v2), d2 = (w1, v3, w3), d3 = (v4, v5, v6), and d4 = (v7, v8, v9). Each log file represents an approximately 8-hour record. However, the log files corresponding to d4 are cut in order to maintain a reasonable file size. In each log file, a column contains data for one request that has arrived to the system. The first row of a column records the arrival time and the second row of the column records the departure time. If a request is rejected, the second row of the column will be marked with zero. The time unit is minute, and thus the unit for arrival rates and service rates is min^{-1}.

With the exception of the 12 log files, a small file called "cpu_load" is provided to estimate the service rate. This file contains CPU loads for a number of different arrival rates.

A. M/M/1/L Analysis

For the M/M/1/L queuing model, there are three parameters to estimate: arrival rate λ, service rate μ, and system size L. Since M/M/1/L model has a limited queue, we are also supposed to estimate λ_{eff} for each λ because they are different.

1) *Parameter λ and λ_{eff}*: For each log file, the number of columns represents the total number of requests within the total time. The total time for each log file is determined in MATLAB by:

$$X\,(length(X),\ 1) - X\,(1,\ 1)$$

X represents the name of the log file. The number of effective requests within the total time is the number of requests whose departure times are more than zero. We used a simple MATLAB function "**amount.m**" to calculate the number of effective requests for each log file. All the source codes of the MATLAB functions mentioned in this paper can be found in the Appendix. λ and λ_{eff} are then calculated by the following expressions:

$$\lambda = Total\ Requests\ /\ Total\ Time$$

And

$$\lambda_{eff} = Effective\ Requests\ /\ Total\ Time$$

The results are shown in Table Ⅰ.

TABLE Ⅰ
M/M/1/L PARAMETER ESTIMATION RESULTS-**LAMBDA**

Log File	Measured Parameters				
	Total Requests	**Effective Requests**	**Total Time**	**λ**	**λ_{eff}**
V_1	4843	4843	499.6178	9.6934	9.6934
W_1	7467	7467	499.6210	14.9453	14.9453
V_2	9965	9965	499.7487	19.9400	19.9400
W_2	12307	12307	499.7384	24.6269	24.6269
V_3	14720	14720	499.8853	29.4468	29.4468
W_3	17612	17612	499.8864	35.2320	35.2320
V_4	25315	25315	499.9144	50.6387	50.6387
V_5	37362	37362	499.7975	74.7543	74.7543
V_6	49830	49830	499.9079	99.6784	99.6784
V_7	45335	41844	298.8640	151.6911	140.0102
V_8	58368	42529	298.8953	195.2791	142.2873
V_9	73848	42676	298.5784	247.3320	142.9306

2) *Parameter μ:* By Little's Theorem, we know that:

$$\rho = \lambda_{eff} / \mu$$

As a result, it is easy to obtain that:

$$\mu = \lambda_{eff} / \rho$$

In the file "cpu_load", 10 λ_{eff} and 10 corresponding ρ for each λ_{eff} are provided. Therefore, we can obtain the value of μ by calculating the mean value of λ_{eff} and ρ and then using Little's Theorem to obtain the average μ. The results are shown in Table Ⅱ. We also noticed that μ is independent of λ_{eff}.

TABLE Ⅱ
M/M/1/L PARAMETER ESTIMATION RESULTS-**MU**

Column	Measured Parameter		
	λ_{eff}	**ρ**	**μ**
1	12.3450	0.0876	140.9247
2	17.4440	0.1233	141.4761
3	23.0120	0.1630	141.1779
4	27.1450	0.1921	141.3066
5	33.2120	0.2356	140.9677
6	40.8970	0.2891	141.4632
7	46.4230	0.3289	141.1462
8	53.1210	0.3759	141.3168
9	62.2230	0.4400	141.4159
10	73.9870	0.5243	141.1158
Mean	38.9809	0.2760	141.2454

3) *Parameter L:* When estimating parameter L for M/M/1/L model, it is critical to find some special requests whose departure times are equal to zero, which indicates that those requests are rejected by the system since both the server and the queues are full. We found that V8 (2983, 2) = 0 and V8 (2983, 1) = 15.0137. As a result, for the log file V8, the queue is full at the time 15.0137, and the number of requests in the system at this time is equal to L. We also found for the next column 2984, V8 (2984, 2) > 0 and thus V8 (2984, 1) < V8 (2983, 1). What's more, since V8 (3202, 1) = 15.0128 < V8 (2983, 1) = 15.0137 < V8 (3203, 1) = 15.1785, and there are 18 requests whose departure time are equal to zero from column 2984 to column 3202 in the log file V8, the number of requests in the system at the time 15.0137 is equal to:

$$(3202 - 2984 + 1) - 18 = 201$$

And L is also equal to 201.

B. IPP/M/1/L Analysis

For the IPP/M/1/L queuing model, there are four parameters to estimate: arrival process state transition rate α, arrival rate at state 1 β, service rate μ, and system size L. The last two parameters for the IPP/M/1/L queuing model are the same as those for the M/M/1/L queuing model because they are not dependent on the arrival process.

1) *Parameter α and β:* The difference between the IPP/M/1/L queuing model and the M/M/1/L queuing mode is that the arrival process for the IPP/M/1/l queuing model is a modified Poisson process called "Interrupted Poisson Process". The Markov chain below describes an IPP:

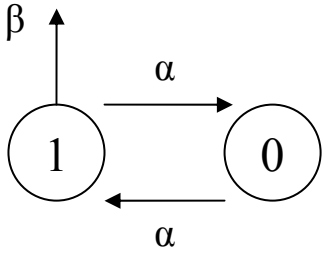

Fig. 1 IPP Markov Chain

When the process is in the state 1, it sends requests with intensity β to the system. When the process is in the state 0, no requests are sent. An IPP arrival process has the following Laplace transform:

$$B^*(s) = \frac{\beta(s+2\alpha)}{s^2 + (\beta+2\alpha)s + \beta\alpha}$$

In order to estimate α and β, we need to find two independent equations which contain α and β as two unknown quantities. The following two independent equations can be illustrated in terms of the properties of Laplace transform:

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i = -B^{*'}(s)\Big|_{s=0} = \frac{2}{\beta}$$

$$E(X^2) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = B^{*''}(s)\Big|_{s=0} = \frac{8\alpha + 2\beta}{\alpha\beta^2}$$

Where $\{x^1, x^2, \ldots, x^n\}$ is the set of measured interarrival times.

We use MATLAB functions "**inter.m**" and "**inters.m**" to calculate E(X) and E(X^2) for each log file. Another MATLAB function "**ippv.m**" is used to calculate the value of α and β based on the results of the MATLAB functions

"**inter.m**" and "**inters.m**". The results are shown in Table Ⅲ.

TABLE Ⅲ
IPP/M/1/L PARAMETER ESTIMATION RESULTS-**ALPHA** AND **BETA**

Log File	Measured Parameter			
	E(*X*)	**E(X^2)**	**α**	**β**
V_1	0.1032	0.0298	12.1958	19.3828
W_1	0.0669	0.0139	13.6211	29.8867
V_2	0.0502	0.0088	13.3017	39.8760
W_2	0.0406	0.0064	13.3040	49.2498
V_3	0.0340	0.0051	12.2074	58.8895
W_3	0.0284	0.0038	13.2406	70.4600
V_4	0.0197	0.0021	14.5073	101.2733
V_5	0.0134	0.0014	12.6365	149.5046
V_6	0.0100	8.7735×10^{-4}	14.8398	199.3527
V_7	0.0066	5.0271×10^{-4}	15.8553	303.3755
V_8	0.0051	3.7453×10^{-4}	15.9718	389.2177
V_9	0.0041	2.7939×10^{-4}	16.4664	492.8627

2) *Parameter μ and L:* As stated above, these two parameters are the same as those of the M/M/1/L queuing model, so μ is equal to 141.2454 and L is equal to 201.

III. COMPARISON AND SELECTION

Before we make simulations based on the system parameters calculated from the previous section, we should first calculate some performance measures for the web server. For a general queuing system, the most three important performance measures are the average number of requests in system, the average system response time, and the request rejection rate. We used simple MATLAB functions "**avT.m**" and "**avN.m**" to calculate the average system response time and the average number of requests in the system for each log file. As for the request rejection rate, by definition:

R = (Total Requests- Effective Requests) / Total Requests

The results are shown in Table Ⅳ.

TABLE Ⅳ
SYSTEM PERFORMANCE MEASURE ESTIMATION RESULTS-T, $\overline{N}$ AND R

Log File	Measured Parameters		
	T	$\overline{N}$	***R***
V_1	0.0081	0.0781	0
W_1	0.0084	0.1258	0
V_2	0.0093	0.1845	0
W_2	0.0100	0.2471	0
V_3	0.0109	0.3221	0
W_3	0.0118	0.4148	0
V_4	0.0167	0.8474	0
V_5	0.0332	2.4804	0
V_6	0.0877	8.7426	0
V_7	0.9773	136.8305	0.0770
V_8	1.2559	178.6963	0.2714
V_9	1.3009	185.9440	0.4221

The MATLAB function "**MM1L.m**" is used to simulate the web server based on the M/M/1/L queuing model and the MATLAB function "**IPP1L.m**" is used to simulate the web server based on the IPP/M/1/L queuing model. Note that the common argument "endtime" is equal to *X (Length(X), 1)*, which is the arrival time for the last request in each log file. Besides, another two MATLAB functions "**avMM1L.m**" and "**avIPPM1L.m**" are used to run the simulation with relevant parameters of each log file for 100 times and calculate the mean values for the outputs. The results of the simulations are shown in Table Ⅴ (M/M/1/L queuing model) and Table Ⅵ (IPP/M/1/L queuing model).

TABLE Ⅴ
M/M/1/L PERFORMANCE MEASURE SIMULATION RESULTS-T, $\overline{N}$ AND R

Log File	Measured Parameters		
	T	$\overline{N}$	***R***
V_1	0.0076	0.0733	0
W_1	0.0079	0.1179	0
V_2	0.0082	0.1637	0
W_2	0.0086	0.2113	0
V_3	0.0090	0.2636	0
W_3	0.0094	0.3330	0
V_4	0.0110	0.5588	0
V_5	0.0150	1.1218	0
V_6	0.0240	2.3918	0
V_7	1.2843	181.8711	0.0664
V_8	1.3929	197.1578	0.2732
V_9	1.4060	199.0307	0.4263

TABLE Ⅵ
IPP/M/1/L PERFORMANCE MEASURE SIMULATION RESULTS-T, $\overline{N}$ AND R

Log File	Measured Parameters		
	T	$\overline{N}$	***R***
V_1	0.0081	0.0788	0
W_1	0.0088	0.1307	0
V_2	0.0095	0.1900	0
W_2	0.0103	0.2529	0
V_3	0.0112	0.3314	0
W_3	0.0126	0.4445	0
V_4	0.0176	0.8909	0
V_5	0.0362	2.7022	0
V_6	0.0817	8.1561	0
V_7	0.9755	137.1253	0.0717
V_8	1.2825	181.2690	0.2723
V_9	1.3339	188.9113	0.4231

Comparing the results of Table Ⅴ and Table Ⅵ to the results in Table Ⅳ, it is obvious that the results of Table Ⅵ is much closer to the results to in Table Ⅳ than that of Table Ⅴ. As a result, IPP/M/1/L queuing model is more accurate for modelling this web server and thereby we will use IPP/M/1/L queuing model in the next section to propose our solution.

IV. SOLUTION

The duration between a HTTP request sent by a customer and successful receipt of the web contents is referred to as the HTTP response time. A research [2] shows that a HTTP response time longer than 10 seconds is not acceptable for a customer. However, when estimating HTTP response time, one should always take its three parts into account: Processing delay, Transmission delay and Propagation delay. The processing delay is also known as the system response time, as those mentioned in the previous section. We assumed that the processing delay takes up about 20% of the total network delay in case that the web server is not busy. We also assumed

that a rejection rate under 0.1% is acceptable. Therefore, a system with the following performance measures will be satisfying:

$$T \le 20\% \times 10s = 2s \quad And \quad R \le 0.1\%$$

We notice that in Table Ⅳ, only log file V6 – V9 do not satisfy the requirements. By looking at the corresponding arrival rate in Table Ⅰ, we found that log file V9 has the highest average arrival rate λ = 247.3320. If a system can handle an average arrival rate as high as that of the log file V9 satisfyingly, it will also perform well for log file V6, V7 and V8. In addition, we also should allow for a 20% increase of traffic load each year. If we want to do a 5-year analysis, the average arrival rate in the fifth year will be:

$$\lambda = 247.3320 \times (1 + 20\%)^4 = 512.8676$$

Since IPP/M/1/L queuing model is more suitable able for modelling this system, we are supposed to calculate the corresponding α and β. Since we know that:

$$E(X) = \frac{1}{\lambda} = \frac{2}{\beta}$$

Accordingly, we are able to derive the following formula:

$$\beta = 2\lambda$$

Substitute that λ = 512.8676 into the formula

$$\beta = 2\lambda = 2 \times 512.8676 = 1025.7352$$

It is very difficult to estimate $E(X^2)$ in an accurately analytical way. However, we can do a linear approximation instead to get its value. We define that:

$$K = \frac{E(X^2)}{E^2(X)} \text{ And } N = \frac{\beta^x}{\beta^y}$$

For log file V4 and W2, N_1 = 101.2733 / 49.2498 = 2.0563. Meanwhile, by using MATLAB function "**inter.m**" and "**inters.m**", K_{V4} / K_{W2} = 1.4257. For log file V7 and V5, N_2 = 303.3755 / 149.5026 = 2.0292. Similarly, K_{V7} / K_{V5} = 1.4613. Since for β = 1025.7352 and log file V9, N_3 = 1025.3752 / 492.8627 = 2.0804. Since $N_1 \approx N_2 \approx N_3$, and $K_{V4} / K_{W2} \approx K_{V7} / K_{V5}$, we can suppose that $K_\beta / K_{V9} \approx K_{V4} / K_{W2} \approx K_{V7} / K_{V5}$. What's more, since N_3 = 2.0804 is a little closer to N_1 = 2.0563 than N_2 = 2.0292, we let $K_\beta / K_{V9} = K_{V4} / K_{W2}$ = 1.4257 approximately. We can then calculate that K_{V9} = 16.9646, so K_β = 16.9646 × 1.4257 = 24.1861. $E^2(X)$ = (1 / λ) ^{2} = (1 / 512.8676) ^{2} = 3.8018 × 10^{-6}, and thus $E(X^2)$ = 24.1861 ×3.8018 × 10^{-6} = 9.1952 × 10^{-5}. We have obtained the formula in the previous section stating that:

$$E(X^2) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = B^{*''}(s)\Big|_{s=0} = \frac{8\alpha + 2\beta}{\alpha\beta^2}$$

And we can use it to calculate that α = 23.1161.

If we want to lower the rejection rate, we can increase the queue size Q = L – 1, increase the service rate μ, or increase the number of servers C. However, as the queue size increases, the system response time also increases, which is against another goal – to lower the system response time. If we want to increase the service rate μ, we must change the old CPU into a new one, but how fast should the new CPU be? We know that for the M/M/1/L queuing system, the rejection rate is very small when λ < μ. As a result, the minimum value of the service rate μ for the new CPU should be 512.8676, which is about 3.6310 times larger than the previous CPU. It is very difficult to find such a CPU in today's market, and a single server system is not scalable, so we should increase the number of servers, which is to buy some parallel servers. How many parallel servers do we need? It is clear that we ought to buy another three parallel servers at least to make λ < μ, but is it enough to satisfy the requirement mentioned above? In order to test it out, we make two simulations based on the IPP/M/C/L queuing system, for C = 4 and C = 5. Two MATLAB functions "**IPPM4L.m**" and "**IPPM5L.m**" are used to simulate the corresponding queuing system and another two MATLAB functions "**avIPPM4L.m**" and "**avIPPM5L.m**" are used to run the corresponding simulation function for 100 times and calculate the average value of the outputs. We use the "endtime" of the log file V9 here. The results are shown in Table Ⅶ. Notice that the system size is equal to 204 for the first one and 205 for the second one.

TABLE Ⅶ

IPP/M/C/L PERFORMANCE MEASURE SIMULATION RESULTS-T, $\overline{N}$ AND R

C	Measured Parameters		
	T	$\overline{N}$	***R***
4	0.1404	70.5146	0.0177
5	0.0510	26.1697	0.0003

From the results of Table Ⅶ, we can see that the IPP/M/4/L system cannot fulfil our requirements but the IPP/M/5/L system can, so we should buy another four parallel servers to establish a IPP/M/5/L queuing system.

V. CONCLUSION

This paper aims at finding a suitable queuing model for the server of a company's website which has performance problems. Two queuing models are suggested: M/M/1L and IPP/M/1/L. In Section Ⅱ, relevant parameters for each queuing model is estimated and in Section Ⅲ, a simulation based on the parameters found in Section Ⅱ for each queuing model is performed. The results show that the IPP/M/1/L queuing model appears to be more accurate and thus should be selected. In Section Ⅳ, a discussion on how to decrease the system response time and request rejection rate is made. Finally, we proposed that the company should buy some parallel servers. Simulations for both the IPP/M/4/L and IPP/M/5/L queuing systems are performed. The results show that the IPP/M/4/L queuing system cannot fulfil the requirement, so the company should buy four extra parallel severs, which can grant a five-year web server stability.

REFERENCES

[1] MathWorks website. [Online]. Available: http://www.mathworks.com/products/new_products/latest_features.html?s_cid=HP_RH_2007b, last visited Feb. 25th

[2] J. Nielsen, "Response Times: The Three Important Limits," *Papesr and Essays by Jakob Nielson*. [Online]. Available: http://www.useit.com/papers/responsetime.html