

EFFECT OF SERVICE DISCIPLINES ON WAITING TIME IN $M/M/(1,n)$ SYSTEMS

Robert N. Will

Center for Enterprise Modernization, The MITRE Corporation
7515 Colshire Dr., McLean, VA 22102-7508

Abstract: Performance tuning often involves consideration of adding another server to the system. Owing to changes in technology, the additional server may be of a service rate different from that of the existing servers. Moreover, with heterogeneous servers, consideration must be given to what service discipline should be employed. An analysis is provided of the effect on waiting time using two service disciplines (1) weighted-priority and (2) priority to the faster server. Examples of $M/M/(1,1)$ and $M/M/(1,2)$ systems are used to illustrate the effects, and a discussion of the results is presented.

I. INTRODUCTION

Performance tuning sometimes requires the incorporation of one or more additional servers in a system. These new servers, usually a newer model, may have a faster service rate μ_1 than the service rate μ_2 of the existing servers in the system. The problem discussed in this paper is what service discipline will give the shortest waiting time in such a heterogeneous system, given that some time may have to be spent in a router. Two service disciplines are considered that give some priority to the faster server: (1) a weighted-priority (WP) service discipline in which the probability that a job will go to a server of a particular server type is equal to the fraction of available servers of that type weighted by the service rates, as shown in the Markov lattice for the general $M/M/(m,n)$ system (m faster servers and n slower servers) in Figs. 1 and 2, where the transition coefficients $r_{\alpha\beta}$ and $s_{\alpha\beta}$ in the Markov lattice are given for the WP service discipline by

$$r_{\alpha\beta} = \frac{(m-\alpha)\mu_1}{(m-\alpha)\mu_1 + (n-\beta)\mu_2},$$

$$s_{\alpha\beta} = \frac{(n-\beta)\mu_2}{(m-\alpha)\mu_1 + (n-\beta)\mu_2}$$

(note the edge conditions $s_{mj}=1$, $r_{mj}=0$, $s_{in}=0$, and $r_{in}=1$) and (2) a faster-server (FS) service discipline in which a job is always sent to one of the faster servers if one is available, otherwise to one of the slower servers. It

can be seen that the FS case is a special case of the WP case with the transition coefficients s_{ij} and r_{ij} defined as follows, assuming the type-1 servers are the faster servers:

$$s_{ij} = 0 \text{ for } i < m \text{ and } s_{mj} = 1$$

$$r_{ij} = 1 \text{ for } i < m \text{ and } r_{mj} = 0.$$

The solution¹ to the general $M/M/(m,n)$ system has been treated, but in this paper only the $m=1$ case is considered, i.e., there is only one server of the faster type. Two cases are treated here, the $n=1$ and $n=2$ cases.

It should be remarked that a system utilization ρ is defined for systems in which all the servers have the same service rates; but this becomes ambiguous when the service rates are not all the same since utilization will depend on the particular service discipline employed. As can be shown¹ from the expressions for the mean queue length and the mean number of jobs in the system, the mean number of jobs being served depends on the service discipline being used, so a different quantity for the utilization would have to be derived for each type of service discipline and for each type of server, which would be rather complicated functions of the μ_i . Accordingly, it seems preferable to avoid use of the term "utilization" when treating heterogeneous server systems and to use specific service rates $m_i = \mu_i/\lambda$. instead.

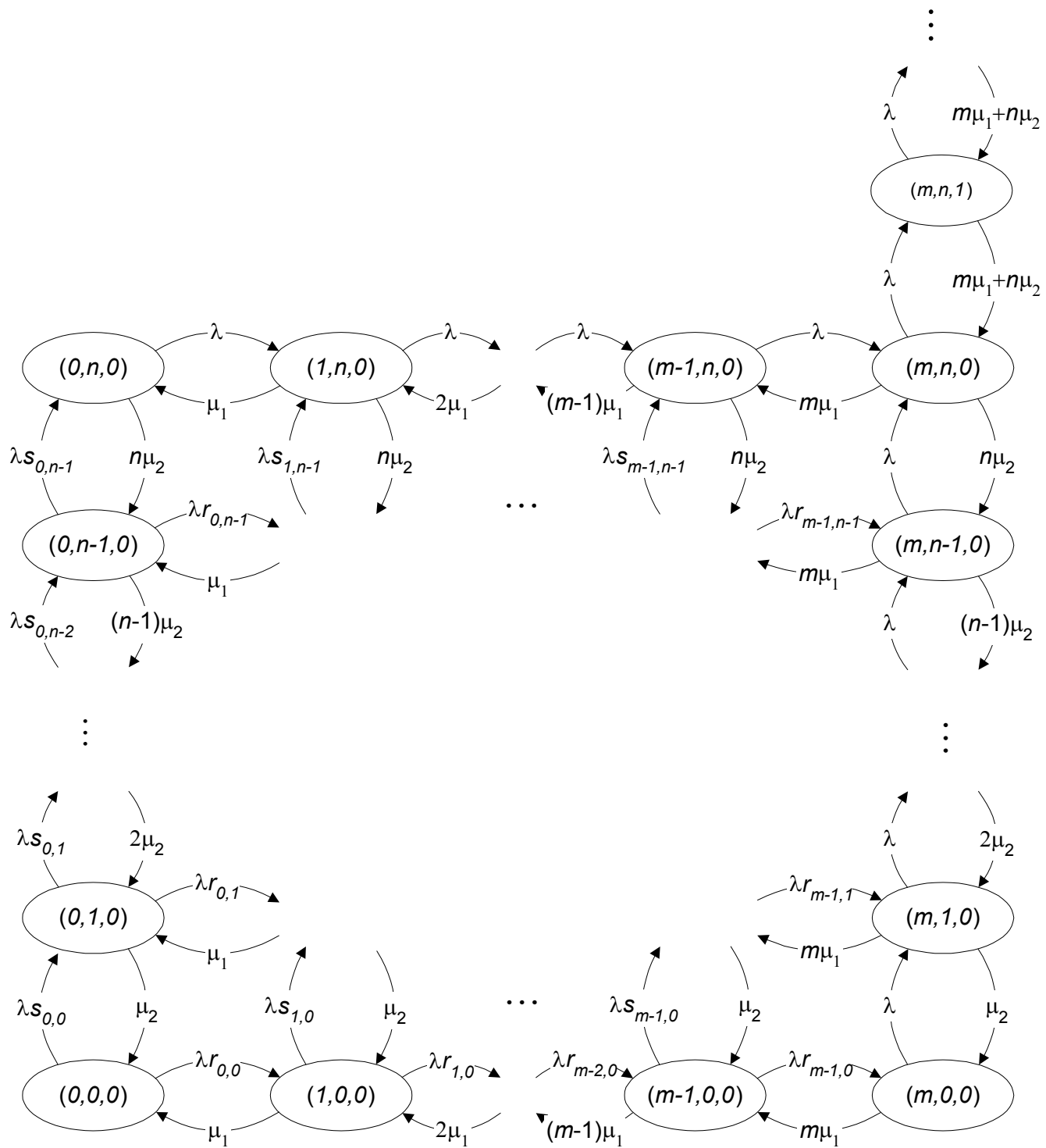


Figure 1. Exterior of the Markov lattice for the $M/M/(m,n)$ system of two-server-types.

II. CASE (1,1)

This is the case of one fast server and one slow server. The complete solution¹ for all the state

probabilities and mean queue lengths for this case for both WP and FS service disciplines has been worked out, but only the mean number of jobs in the system $L^{(1,1)}$ is needed here:

$$L_{FS}^{(1,1)} = \sum_{k=0} k p(k,0) + \sum_{k=0} (k+1) p(k,1)$$

$$= \frac{(m_1 + m_2)^3 (m_2 + 1)}{(m_1 + m_2 - 1)[m_1 m_2 (m_1 + m_2 - 1)(m_1 + m_2 + 2) + (m_1 + m_2)^2 (m_2 + 1)]}$$

$$L_{WP}^{(1,1)} = p(1,0,0) + p(0,1,0) + 2p(1,1,0) + \sum_{n=1}^{\infty} (n+2) p(1,1,n)$$

$$= \frac{(m_1 + m_2)^2 (m_1 + m_2 + 2m_1 m_2)}{(m_1 + m_2 - 1)[(m_1 + m_2)^2 (m_1 m_2 + 1) + 3m_1 m_2 (m_1 + m_2) - 2m_1 m_2]}$$

The state probability $p(i,j,k)$ is defined as the probability that the system has i jobs in type-1 servers, j jobs in type-2 servers, and k jobs in the queue; and specific service rates $m_1 = \mu_1/\lambda$, $m_2 = \mu_2/\lambda$ are used to simplify the expressions. Note that in the case of equal service rates, $\mu_1 = \mu_2$, the usual results² for the $M/M/2$ case are recovered, where $\rho = 1/(m_1 + m_2) = \lambda/2\mu$:

$$L^{(2)} = \frac{2\rho}{1 - \rho^2}.$$

By Little's formula,² the waiting times are directly proportional to the number of jobs in the system. Figure 3 shows the ratio of and Fig. 4 the difference between the WP and FS waiting times for the (1,1) system as a function of the ratio of the service rates for

several values of specific service rate of the slower server.

III. CASE (1,2)

This is the case of one fast server of service rate μ_1 and two slow servers of service rate μ_2 . With the aid of sum rules³ for the state probabilities, the complete solution¹ for all the state probabilities and mean queue lengths for this case for both WP and FS service disciplines has also been worked out. The solution of the state probabilities is rather more complicated than in the (1,1) case. For convenience in simplifying the expressions, the quantities M and N are defined as $M = m_1 + m_2$, $N = m_1 + 2m_2$. The mean numbers of jobs in the system for FS and WP service disciplines become, respectively,

$$L_{FS}^{(1,2)} = [-m_1^2 - 2m_2^2 + 2m_1^3 + 3m_1^2 m_2 + 2m_1 m_2^2 + 2m_2^3 + 2m_2 N(3m_1^2 + 7m_1 m_2 + 7m_2^2) + 2m_2 N^2(m_1^2 + 5m_1 m_2 + 5m_2^2) + 2m_2^2 M N^3] / [(N-1)\Delta_{FS}]$$

and

$$L_{WP}^{(1,2)} = \{MN^2 + N^2(M^2 + MN + 3m_1 m_2) + 2m_2[M^2 N^2 + 3m_1(M^2 N + N^3 + m_2^2(2m_1 + m_2))]\} + 2m_1 m_2 N[N^3 + m_2(11m_1^2 + 23m_1 m_2 + 8m_2^2)] + 6m_1 m_2^2 M^2 N^2 \} / \Delta_{WP}$$

$$+ \frac{MN + 3m_1 m_2(3m_2 + 2m_1) + 4m_1 m_2^2(2m_1 + m_2)}{(1 - 1/N)\Delta_{WP}},$$

where the last term in $L_{WP}^{(1,2)}$ is the WP mean queue length and the Δ 's in the denominators are

$$\begin{aligned}\Delta_{WP} = & MN(m_1^2 + 2m_2^2) \\ & + m_2[8M^4 - m_1m_2(7m_1^2 + 6m_1m_2 - 2m_2^2)] \\ & + 2m_1m_2N(m_1^3 + 12m_1^2m_2 + 26m_1m_2^2 + 15m_2^3) \\ & + 2m_1m_2^2N^2(6m_1^2 + 12m_1m_2 + 5m_2^2) \\ & + 2m_1m_2^2M^2N^3\end{aligned}$$

and

$$\begin{aligned}\Delta_{FS} = & m_1^2 + 2m_2^2 + m_2(5MN - 9m_1m_2) \\ & + 2m_2(N^3 + 3m_1m_2^2) \\ & + 2m_2^2[4M^3 + m_1m_2(2M + N)] \\ & + 2m_1m_2^2MN^2.\end{aligned}$$

Note that in the case of equal service rates, $\mu_1 = \mu_2$, the usual results² for the $M/M/3$ case are recovered, where $\rho = 1/(m_1+2m_2) = \lambda/3\mu$:

$$L^{(3)} = \frac{3\rho(2+2\rho-\rho^2)}{(1-\rho)(2+4\rho+3\rho^2)}.$$

IV. DISCUSSION OF RESULTS

A. Qualitative results

The similarities in the curves in Figs. 3 and 5 and in Figs. 4 and 6 suggest general functional forms that likely obtain for larger numbers of the slower server. Of course, to obtain explicit results for one fast server and three or more of the slower servers, the equations for the state probabilities must be solved for both WP and FS cases, and this has been done numerically¹ for some cases but not algebraically. The examples for (1,1) and (1,2) systems treated here provide a good indication of what results might be expected.

The ratio W_{WP} / W_{FS} is unity for $\alpha = 1$, since all servers have the same service rates and there is no difference in the service disciplines. As the ratio of the service rates $\alpha = m_1/m_2$ increases for fixed m_2 , both W_{WP} and W_{FS} must decrease to zero since the service rate of the faster server is increasing and jobs are taking less time to process using either service discipline. The

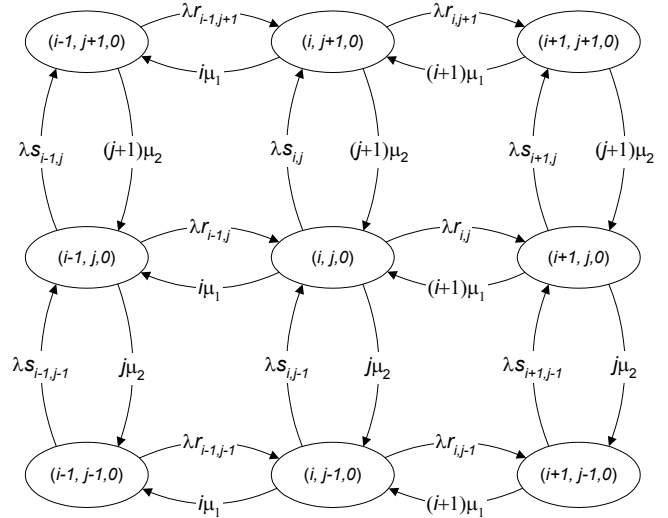


Figure 2. Interior of the Markov lattice, $(0 < i < m)$ and $(0 < j < n)$, for the $M/M/(m,n)$ system of two server types.

ratio of the waiting times W_{WP} / W_{FS} increases with α since W_{FS} is decreasing faster than W_{WP} is decreasing owing to more jobs being serviced by the faster server in the FS service discipline. However, this ratio does not increase without bound but approaches an asymptotic limit; specifically, W_{WP} / W_{FS} has a least upper bound of $[1+(n+1)m_2]/(1+m_2)$, where n is the number of type-2, i.e., slower, servers in the system.

The difference in the waiting times vanishes at $\alpha=1$ since all servers have the same service rates and there is no difference in the service disciplines. This difference is positive for $\alpha>1$ since the WP service discipline will not take advantage of the faster server as well as the FS will. This difference will reach a maximum since it must eventually go to zero as α increases; specifically, $m_1(W_{WP} - W_{FS})/n \rightarrow 1$ as $\alpha \rightarrow \infty$.

B. Quantitative results

From Figs. 3 and 5, it is seen that the ratio of the waiting times increases rapidly for small values of α . As μ_1 increases up to $5\mu_2$, for $m_2 = 1.0$, W_{WP} / W_{FS} increases by 23% and 44% for (1,1) and (1,2) cases, respectively, showing the advantage of the FS service discipline with the one additional server. As discussed above, this advantage increases as the service rate of the faster server increases relative to that of the slower server, but the rate of increase in this advantage is decreasing. It is clear from the sets of curves in Figs. 3 and 5 that the advantage in the use of the FS service discipline is greater for larger specific service rates of the slower server, i.e., for a larger service rate μ_2 or

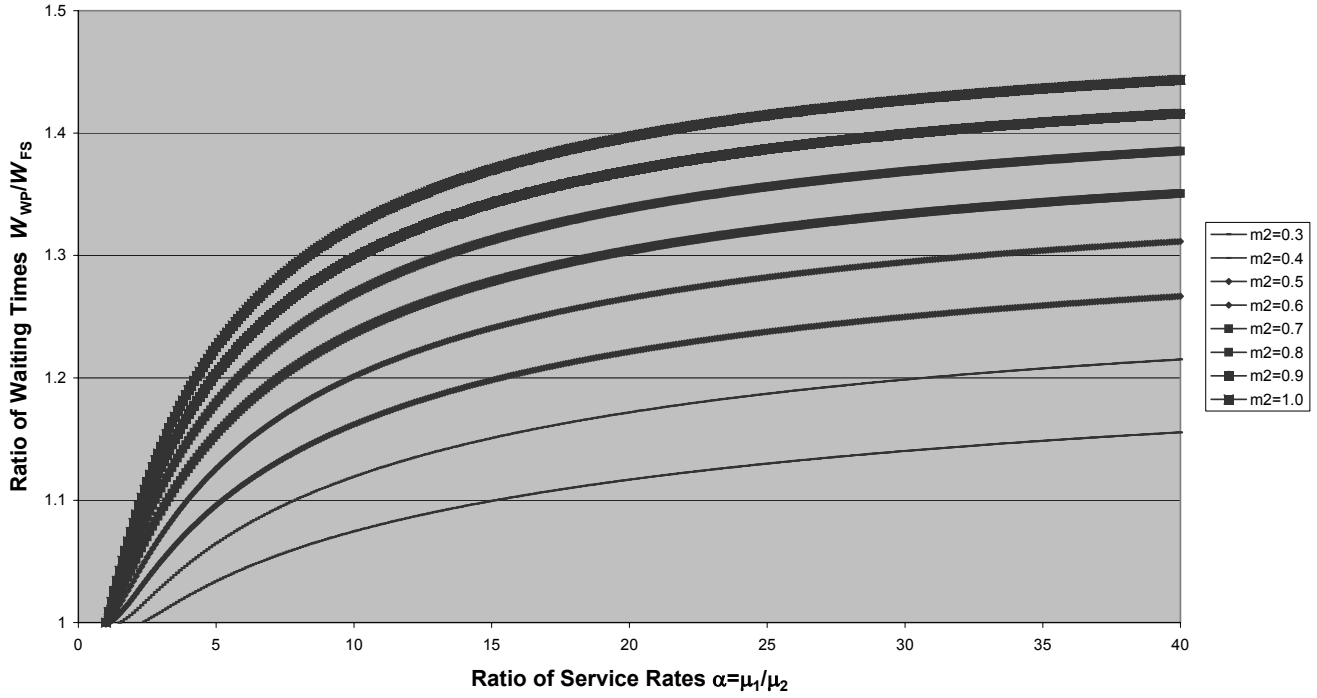


Figure 3. Ratio of waiting times W_{WP}/W_{FS} for (1,1) system as a function of the ratio $\alpha = \mu_1/\mu_2$ of the service rates for several values of the specific service rate of the slower server.

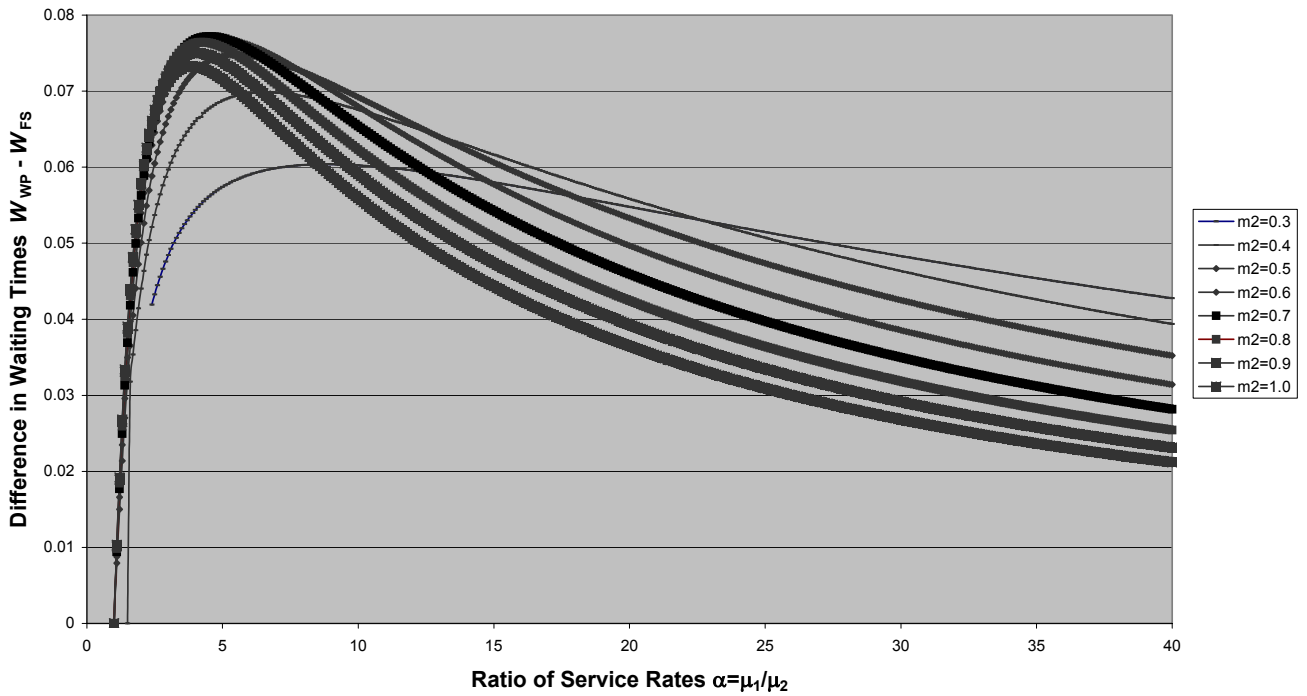


Figure 4. Difference in waiting times $W_{WP} - W_{FS}$ for (1,1) system as a function of the ratio $\alpha = \mu_1/\mu_2$ of the service rates for several values of specific service rate of the slower server and for unit arrival rate λ .

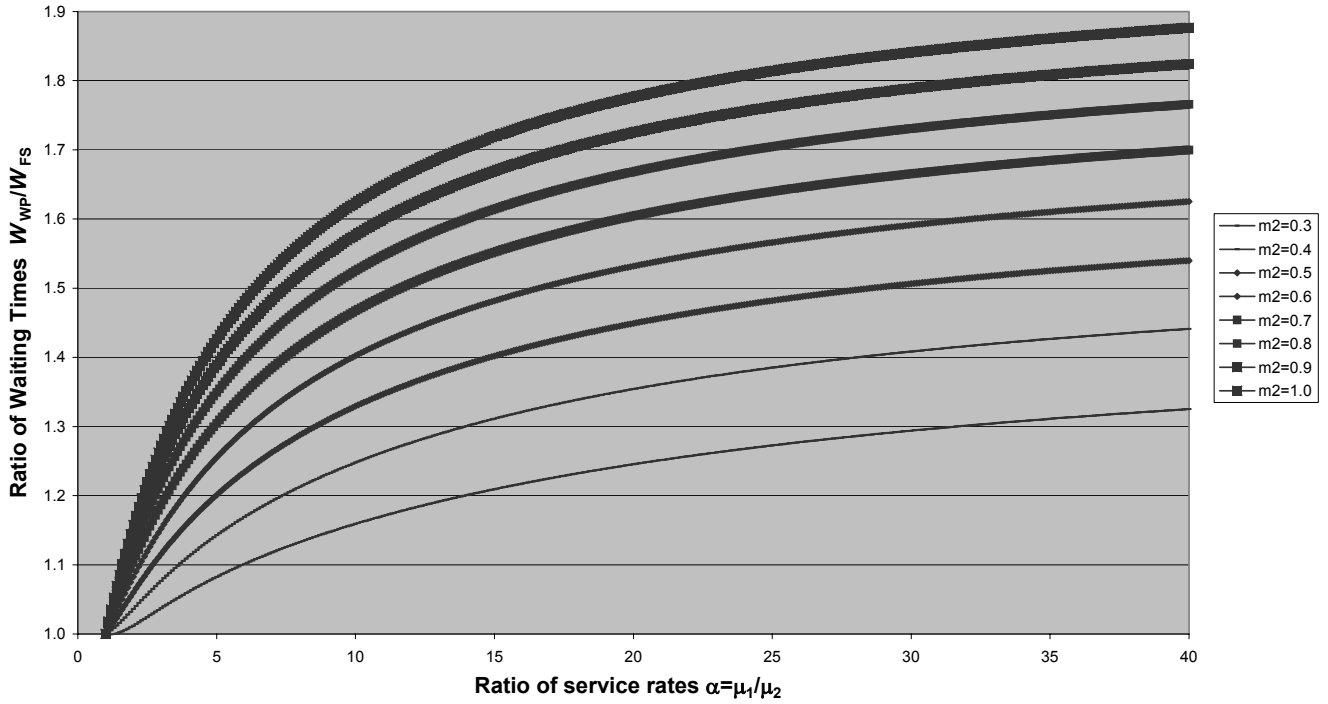


Figure 5. Ratio of waiting times W_{WP}/W_{FS} for (1,2) system as a function of the ratio $\alpha=\mu_1/\mu_2$ of the service rates for several values of the specific service rate of the slower server.

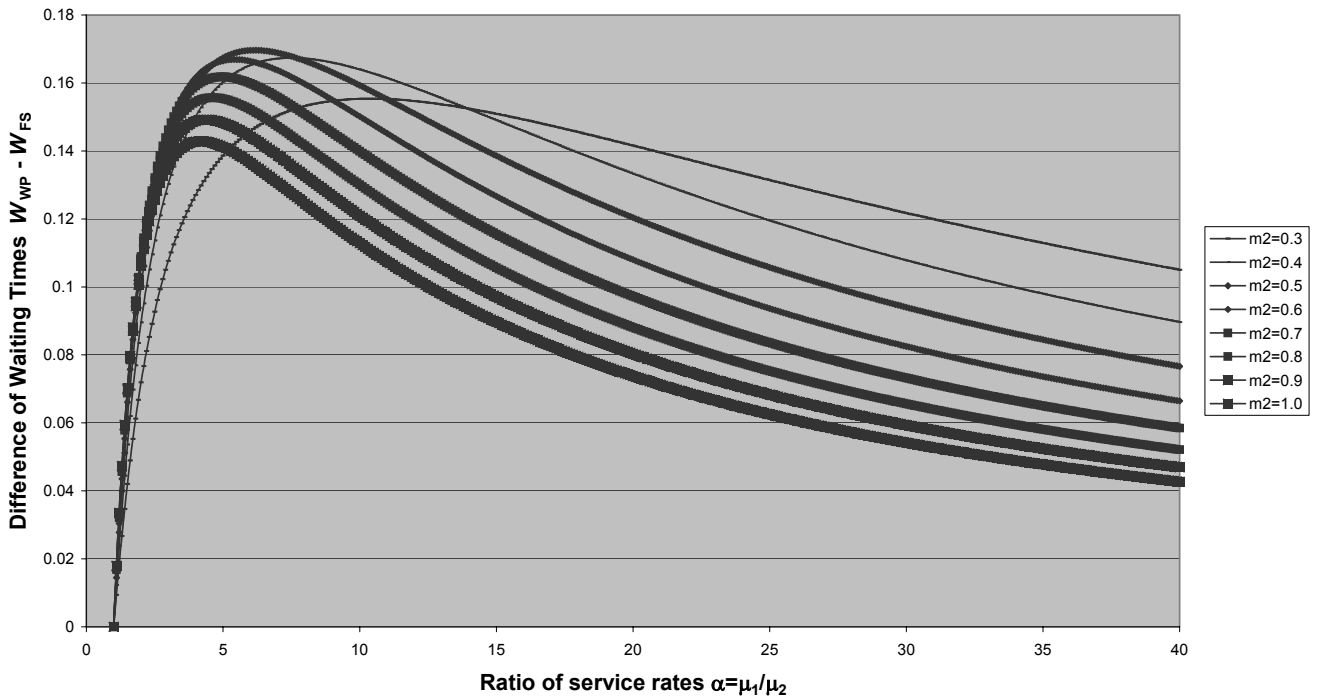


Figure 6. Difference in waiting times $W_{WP}-W_{FS}$ for (1,2) system as a function of the ratio $\alpha=\mu_1/\mu_2$ of the service rates for several values of specific service rate of the slower server and for unit arrival rate λ .

smaller arrival rate λ . This should be taken into account in deciding whether to use the FS service discipline.

The difference curves, Figs. 4 and 6, show more structure and are of special interest since the difference must be greater than the difference in the delays needed to route the jobs through the router. These curves are generated from the difference in the mean number of jobs in the system, $L_{WP} - L_{FS}$, with unit arrival rate assumed. Use is made of Little's formula to convert the mean number of jobs in the system to total waiting time, $W=L/\lambda$. Note, however, that doubling the arrival rate λ does not simply halve the difference in waiting times since $L_{WP} - L_{FS}$ is also a function of the arrival rate.

From Figs. 4 and 6 it is seen that the difference in waiting times rises to a peak for small values of α , in the range $5 \leq \alpha \leq 10$. This means that if the difference in the routing delays is significant, then it may not be an advantage to use FS instead of WP service discipline if the faster server is very much faster than the slower server(s). That is, even though the waiting time (not counting router delay) is shorter with FS, the additional delay caused by the routing may result in a longer rather than shorter waiting time than with WP and nullify the advantage.

Note also from Figs. 4 and 6 that the curve with the highest peak in the waiting time difference occurs for specific service rate of the slower server $m_2 = 0.7$ for the (1,1) case and for $m_2 = 0.5$ for the (1,2) case; that

is, the peak does not increase or decrease monotonically with m_2 but is maximized for a particular value of m_2 .

The objective in performance tuning is to minimize the total waiting time W , while bearing in mind the cost implications of the changes to the system. In performance tuning, both the number of slower servers and their service rate μ_2 are given, and only μ_1 can vary. So the tuning will involve only one curve on a chart such as in Figs. 3 – 6. Increasing μ_1 will certainly decrease the waiting time W . The question is whether to use the FS or WP service discipline, either of which will require a router. In order to use the FS service discipline instead of the WP service discipline, one must use a faster server with service rate μ_1 such that μ_1 is large enough to produce as large a ratio W_{WP} / W_{FS} as possible while also maintaining the difference $W_{WP} - W_{FS}$ greater than the difference in the router delays and also considering the cost effectiveness of the total solution.

References

1. [Will03] R. N. Will, "Solution of $M/M/(1,1)$ and $M/M/(1,2)$ systems with three service disciplines," report (unpublished).
2. [Gross98] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory* (Wiley, New York, 1998).
3. [Will03] R. N. Will, "Four sum rules for state probabilities for $M/M/(m,n)$ systems with two server types and three service disciplines," report (unpublished).