Tutorial 7 – M/M/m systems

Ioannis Glaropoulos

November 12, 2012

1 Exercise 7.2

The performance of a system with one processor and another with two processors will be compared. Let the inter-arrival times of jobs be exponentially distributed with parameter λ. We consider, first, the system with one processor. The service time of the jobs is exponentially distributed with a mean of 0.5sec.

1. *For an average response time of 2.5sec (the total time for a job in the system) how many jobs per second can be handled?*
2. *For an increase of λ with 10% how much will the response time increase?*
3. *Calculate the average waiting time, the average number of customers in the server and the utilization of the server. What is the probability of the server being empty?*

Let us now compare this system with a system of two cheaper processors, each with a mean service time of 1 sec.

1. *How many jobs can now be handled per second with a mean response time of 2.5 sec?*
2. *For an increase of λ with 10% how much will the response time increase?*
3. *Calculate the average waiting time, the average number of customers in the server and the utilization of the server. What is the probability of both of the servers being busy?*

Solution: We consider first the system with one processor. Clearly, since we are not given any information regarding the buffer data, and we are also told that there is waiting time, we can safely conclude that the buffer is **infinite**.

As a result we are dealing with a typical M/M/1 system, and we will derive the answers from the M/M/1 formulas. The average number of jobs in the M/M/1 system is given as

$$\overline{N} = \frac{\rho}{1-\rho} = \frac{\lambda/2}{1-\lambda/2} = \frac{\lambda}{2-\lambda}$$

since $\mu = \frac{1}{E[T]} = \frac{1}{0.5} = 2$. Applying the LITTLE's formula, we get

$$\overline{T}_{\text{system}} = \frac{\overline{N}}{\lambda} = \frac{1}{2-\lambda} \rightarrow \lambda = 1.6$$

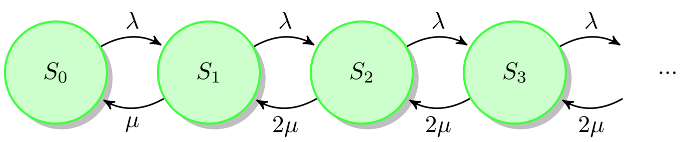

Figure 1: System diagram for the second model of exercise 7.2

since $\overline{T}_{\text{system}} = 2.5$. Assume now that we increase the arrival rate by 10%. The new rate will be $\hat{\lambda} = 1.1 \cdot 1.6 = 1.76$. This will change the offered load of the system to $\hat{\rho} = \frac{\hat{\lambda}}{\mu} = \frac{1.76}{2} = 0.88$. Then the response time will be given in a similar way:

$$\hat{T}_{\text{system}} = \frac{1}{\mu - \hat{\lambda}} = \frac{1}{2 - 1.76} = 4.1666\text{sec}$$

So the increase is $\frac{4.166-2.5}{2.5} = 66.66\%$

For the initial ($\lambda = 1.6$) system the average waiting time is $\overline{W} = \overline{T}_{\text{system}} - E[T] = \overline{T}_{\text{system}} - \frac{1}{\mu} = 2\text{sec}$, the mean number of customers in the system is $\overline{N} = \rho/(1-\rho) = 0.8/0.2 = 4$, and the probability of empty server is $1 - \rho = 0.2$. The utilization is of course ρ and is equal to the mean number of customers in the server.

We consider now the second system. This one has 2 processors with $\mu = 1$, so half of the service rate. The offered load for this system is $\rho = \lambda/1 = \lambda$. The state diagram is given in Fig 1. We could use the formulas for the M/M/2 system. However, since we only have 2 servers, we can solve it analytically with the Balance Equations.

$$\begin{array}{l}
\lambda P_0 = \mu P_1 \rightarrow P_1 = \rho P_0 \\
\lambda P_1 = 2\mu P_2 \rightarrow P_2 = \frac{1}{2}\rho P_1 = \frac{\rho^2}{2} P_0 \\
\lambda P_2 = 2\mu P_3 \rightarrow P_3 = \frac{1}{2}\rho P_2 = \frac{\rho^3}{4} P_0 \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
\lambda P_{k-1} = \mu P_k \rightarrow P_k = \frac{\rho^k}{2^{k-1}} P_0
\end{array}$$

From the above equation set and the normalization equation we derive

$$P_0 \left(1 + \rho \sum_{k=1}^{\infty} \frac{\rho^{k-1}}{2^{k-1}} \right) = 1 \rightarrow P_0 = \frac{2-\rho}{2+\rho}.$$

Then, we calculate the mean number of customers in the system, from the state distribution

$$\overline{N} = \sum_{k=1}^{\infty} k P_k = \sum_{k=1}^{\infty} k \frac{\rho^k}{2^{k-1}} \frac{2-\rho}{2+\rho} = \ldots = \frac{4\rho}{(2+\rho)(2-\rho)} = \frac{4\lambda}{(2+\lambda)(2-\lambda)}.$$

Then, based on LITTLE we compute the average system time:

$$\overline{T}_{\text{system}} = \frac{\overline{N}}{\lambda} = \frac{4}{(2+\lambda)(2-\lambda)} \rightarrow \lambda = \sqrt{\frac{6}{2.5}}$$

We increase, now the λ by 10%, so $\hat{\lambda} = 1.704$. The average system time will, then be

$$\hat{T}_{\text{system}} = \frac{4}{(2 + 1.704)(2 - 1.704)} = ... = 3.65\text{sec}$$

So the increase is $\frac{3.65-2.5}{2.5} = ...$ The average waiting time will be given through the series:

$$\overline{W} = \sum_{k=2}^{\infty} P_k = 1 - P_0 - P_1 = ...$$

but, simply, can be given as $\overline{T}_{\text{system}} - \frac{1}{\mu} = 1.5$ sec.

The average number of customers in the servers (together) is: $\overline{N}_{\text{server}} = P_1 + 2 \cdot (1 - P_0 - P_1) = ...$

Utilization of an ARBITRARY server: $U = 0 \cdot P_0 + \frac{1}{2}P_1 + 1 \cdot (1 - P_0 - P_1)$, since at state S_1 a server is utilized with probability 1/2. Alternatively, we can divide the offered load, which is also the actual load, by the number of servers.

The probability that both servers are busy is: $1 - P_0 - P_1$.

2 Exercise 7.3

Customers arrive to an M/M/3 system with intensity 1.6 arrivals per time unit. The average service time is 1.25 time units. The service strategy is FCFS. Consider a customer that finds all 3 servers busy and 2 customers waiting in the queue. Calculate the probability that this customer has to wait longer than 1.25 time units.

Solution: Check the solution in the solution manual.

3 Exercise 7.6

Consider a pure delay system where customers arrive according to a Poisson process with intensity $\lambda = 3$. The service time is exponentially distributed with mean value 1/3. The queuing discipline is FCFS. There are two servers in the system, and one of them is always available. The other one starts service when the queue length would become two (so that it immediately becomes one). If there are no more customers in the queue, the server which becomes idle first is closed (and stays closed until the queue length becomes two again). Let us denote the state of the system with (i,j) where i is the total number of customers in the system and j is the number of open servers. Give the Kendall notation of the system and draw the system diagram.

Solution: In this problem the state space is given. S_{ij} denotes the state where the total number of customers [in the server(s) and in the queue] is i and the number of active servers is j. Given the state space we draw the system diagram in Fig. 2. The system has 2 servers, infinite queue, Poisson arrivals and exponential service times, so the Kendall notation is: M/M/2, although it is not a typical case, as you see in the diagram.

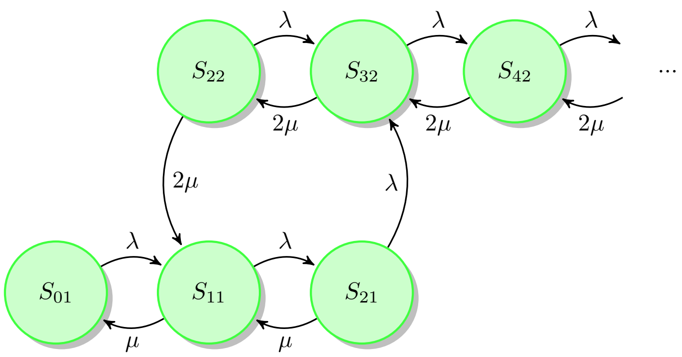

Figure 2: System diagram for exercise 7.6. Notice that the system transits from S_{21} to S_{32}, since this new arrival would make the queue length become 2, so the second server is activated, and the queue length remains one, while both servers, now, serve a customer.