M/M/1 Queue

We learned M/M/1 queue in Queueing lectures. For an M/M/1 queue, there is one server with an exponential service rate μ. The arrival rate to the system is $\lambda < \mu$. In addition, the waiting area is infinite. Particularly, we derive that the average number of customers in the **system (both queue and server!!)** is

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{\rho}{1 - \rho}$$

where $\rho = \lambda/\mu$ is the utilization of the server. Note, this L_s is the average number of customers in the entire system, **NOT** the queue (buffer)! By Little's law $L = \lambda W$, we know that the average time a customer spends in the **system** is

$$W_s = \frac{1}{\mu - \lambda}$$

In other words, this W_s already contains the time that a customer spends in the queue as well as the time that customer spends at the server! If you are not sure about the results above, please re-visit slide 46 of the Queueing lecture. Recall that, when we derived this, we built an Markov process model. In that model, our state space is the number of customers in the **system**, **NOT** the **queue**!

Next, let us discuss the average number of customers and average waiting time **in the queue**. Let L_q and W_q be the average number of customers and the average waiting time in the queue, respectively. We know that, the average waiting time in the system, W_s, consists of the average waiting time in the queue (W_q) and the average service time, or

$$W_s = W_q + \text{average service time}$$

We know that the average service time is $1/\mu$, thus,

$$W_q = W_s - \text{average service time} = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu - \lambda)}$$

We can then derive L_q by two ways. We can first apply the Little's law **to the queue**: $L_q = \lambda W_q$, or

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho^2}{1 - \rho}$$

The second way to find this is to realize that L_s is the sum of L_q and the average number of customers at the server. *You may argue that the average number of customers at the service is 1 by intuition.* However, this is not true. To find the correct number, we must realize that the server is

busy with frequency $\rho = \lambda/\mu$ and once the server is busy, there is one customer being served. So, the average number of customers at the server is λ/μ. Therefore,

$$\bar{} \quad \bar{} \quad \lambda \quad \lambda \quad \lambda \quad \lambda^2$$

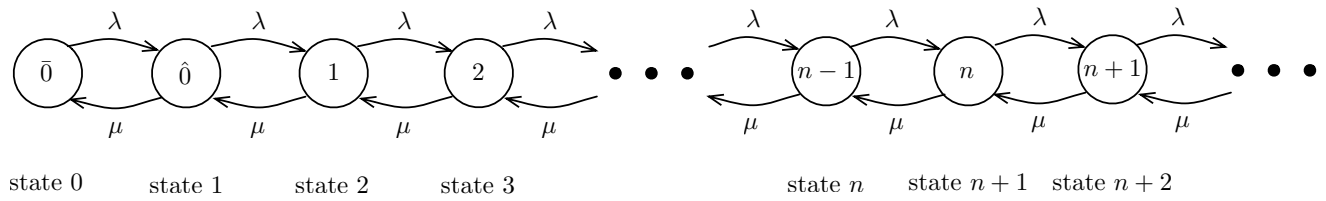

Figure 1: Markov process model for the queue

We define the state as follows:

- State 0 ($\bar{0}$): 0 customer in the queue and 0 customer at the server;
- State 1 ($\hat{0}$): 0 customer in the queue but 1 customer at the server;
- State $i \geq 2$: $i - 1$ customers in the queue.

You may realize that this Markov process for the queue looks very similar to that for the entire system. Particularly, the state structures as well as the transition rates in the two processes are the same, but only the numbers of customers in states differ. So, we can solve the steady state probabilities easily:

$$\mathbf{p}(\text{state } i) = (1-\rho)\rho^i, \qquad i \geq 0$$

Therefore, the average number of customer in the queue is

$$\begin{aligned}
L_q &= 0\mathbf{p}(\text{state } 0) + 0\mathbf{p}(\text{state } 1) + \sum_{i=2}^{\infty}(i-1)\mathbf{p}(\text{state } i) \\
&= \sum_{i=2}^{\infty}(i-1)(1-\rho)\rho^i \\
&= (1-\rho)\rho^2\frac{\mathrm{d}}{\mathrm{d}\rho}\left(\sum_{i=2}^{\infty}\rho^{i-1}\right) = (1-\rho)\rho^2\frac{\mathrm{d}}{\mathrm{d}\rho}\left(\frac{\rho}{1-\rho}\right) \\
&= (1-\rho)\rho^2\frac{1}{(1-\rho)^2} \\
&= \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}
\end{aligned}$$

2.854 / 2.853 Introduction To Manufacturing Systems

Fall 2016