Informal derivation of queueing time for M/G/1

CS162

April 3, 2016

An M/G/1 queue is a system with exponentially distributed ("memoryless") arrivals, but with a general distribution of departures (servicing rate). We characterize the service rate with a variable **C**, the squared coefficient of variation. Important values of **C** include $\mathbf{C} = 0$ (deterministic service rate), $\mathbf{C} = 1$ (memoryless service rate), and $\mathbf{C} = 1.5$ (the value of C we use for disk response time questions on exams, for some reason).

First, let's consider the **squared coefficient of variation** of the service time, which we assign the letter **C**. The coefficient of variation is the ratio of the standard deviation to the mean. Recall that the standard deviation is the square root of the variance, and the variance of a random variable like T_{ser} is the expected value of $T_{ser}^2 - \mu^2$.

$$\sigma = \sqrt{\mathbb{E}\left[T_{ser}^2\right] - \mathbb{E}\left[T_{ser}\right]^2} \tag{1}$$

$$\mu = \mathbb{E}\left[T_{ser}\right] \tag{2}$$

So, the *squared* coefficient of variation is:

$$\mathbf{C} = \left(\frac{\sigma}{\mu}\right)^2 \tag{3}$$

$$= \left(\frac{\sqrt{\mathbb{E}\left[T_{ser}^2\right] - \mathbb{E}\left[T_{ser}\right]^2}}{\mathbb{E}\left[T_{ser}\right]}\right)^2 \tag{4}$$

$$= \frac{\mathbb{E}\left[T_{ser}^2\right] - \mathbb{E}\left[T_{ser}\right]^2}{\mathbb{E}\left[T_{ser}\right]^2} \tag{5}$$

We'll also keep in mind that $\mathbf{C} + 1$ is a useful quantity:

$$\mathbf{C} + 1 = \frac{\mathbb{E}\left[T_{ser}^2\right] - \mathbb{E}\left[T_{ser}\right]^2 + \mathbb{E}\left[T_{ser}\right]^2}{\mathbb{E}\left[T_{ser}\right]^2} \tag{6}$$

$$= \frac{\mathbb{E}\left[T_{ser}^2\right]}{\mathbb{E}\left[T_{ser}\right]^2} \tag{7}$$

Now, our goal is to find $\mathbb{E}\left[T_q\right]$, which is **how long a job expects to wait in the queue** before being serviced. Intuitively, we can think of the queueing time as a sum of two components: (1) if there is currently a job being serviced, then we need to wait for it to finish, and (2) we need to wait behind all the other jobs in line to finish being serviced.

Recall that u (our system utilization) tells us the proportion of time that the system is busy servicing a request. If the system is currently servicing a request when our job arrives, then we can expect the currently running job to be 50% complete on average. Additionally, we'll define $\mathbb{E}\left[L_q\right]$ to be the expected length of the queue. Let's define $\mathbb{E}\left[T_q\right]$ using these two components:

$$\mathbb{E}\left[T_q\right] = \frac{1}{2} u\, \mathbb{E}\left[T_{ser}\right] + \mathbb{E}\left[L_q\right] \mathbb{E}\left[T_{ser}\right] \tag{8}$$

We can express u as $\lambda \mathbb{E}[T_{ser}]$, and we can express $\mathbb{E}[L_q]$ as $\lambda \mathbb{E}[T_q]$ (Little's Law). However, it's not true in general that $\mathbb{E}[AB] = \mathbb{E}[A]\mathbb{E}[B]$ (if A and B are not independent events). In this case, we actually get $\mathbb{E}\left[T_{ser}^2\right]$ instead of $\mathbb{E}[T_{ser}]^2$ in our expression:

$$\mathbb{E}[T_q] = \frac{1}{2}\lambda \mathbb{E}\left[T_{ser}^2\right] + \lambda \mathbb{E}[T_q]\mathbb{E}[T_{ser}] \tag{9}$$

Now, we can use our expression for $C+1$ from before to transform $\mathbb{E}\left[T_{ser}^2\right]$ into an expression involving only $\mathbb{E}[T_{ser}]$:

$$\mathbb{E}[T_q] = \frac{1}{2}\lambda(C+1)\mathbb{E}[T_{ser}]^2 + \lambda \mathbb{E}[T_q]\mathbb{E}[T_{ser}] \tag{10}$$

Finally, we just rearrange our equation to isolate the $\mathbb{E}[T_q]$, while keeping in mind that $u = \lambda \mathbb{E}[T_{ser}]$:

$$\mathbb{E}[T_q](1 - \lambda \mathbb{E}[T_{ser}]) = \frac{1}{2}\lambda(C+1)\mathbb{E}[T_{ser}]^2 \tag{11}$$

$$\mathbb{E}[T_q](1-u) = \frac{1}{2}u(C+1)\mathbb{E}[T_{ser}] \tag{12}$$

$$\mathbb{E}[T_q] = \frac{C+1}{2} \cdot \frac{u}{1-u} \cdot \mathbb{E}[T_{ser}] \tag{13}$$

This is our final equation for $\mathbb{E}[T_q]$, which is exactly the one presented in lecture.