CS 547 Lecture 12: The M/M/1 Queue

Daniel Myers

The M/M/1 queue is the classic, canonical queueing model. By itself, it usually isn't the right model for most computer systems, but studying it will develop the analysis techniques we'll use for more flexible models.

The three-part notation is the preferred way of describing the parameters of an open queueing model. The first letter refers to the distribution of the interarrival times, the second letter to the distribution of the service times, and the final value is the number of servers. The letter M refers to a *memoryless* (or *Markovian*) distribution, that is, to the exponential distribution.[1]

Therefore, the M/M/1 queue is a model with exponentially distributed interarrival times – which implies that the arrivals are Poisson – exponentially distributed service times, and a single server.

The Tagged Customer Method

To analyze the M/M/1 queue, we'll make use of the *tagged customer method*, or, as I like to call it, "the method of who's-in-front-of-me."

Consider an arbitrarily chosen customer just arriving to the queue. We'll "tag" this customer and follow it through the queue, adding up all the delays that it encounters. The average total time the customer needs to move through the queue, receive its service, and exit is simply the average of all the indvidual delay sources it encounters on its trip through the system.

A newly arriving tagged customer has to wait for three sources of delay:

- the residual service time of the customer in service, if the queue is occupied at the arrival instant
- the time for any customers that are waiting in the queue but not being served at the arrival instant
- the time for the tagged customer to get its own service

The PASTA property and the memoryless property of the exponential provide the key to analyzing these delay sources. The arrivals to the M/M/1 queue are Poisson, so the average state of the queue at the instant of an arrival is simply the long-run average state of the queue.

Therefore, the probability that the queue is occupied at an arrival instant is simply U, the utilization, and the average number of customers waiting but not being served at the arrival instant is $\overline{Q} - U$.

On average, each customer receives a service time of $\overline{s}$. Therefore, the expected time required to serve all the customers waiting in the queue at an arrival instant is $(\overline{Q} - U)\overline{s}$.

Because of the memoryless property of the exponential service times, the expected time for a customer in service to finish is simply $\overline{s}$, regardless of how long the customer has already been in service. Therefore,

[1] Why not use E for the exponential? That letter is reserved for the Erlang distribution.

the expected time waiting due to a customer in service at an arrival instant is $U\overline{s}$, where U comes from the probability that the server is busy at the arrival instant.

Finally, the tagged customer requires an average of $\overline{s}$ for its own service.

Deriving the Average Residence Time

Adding all three of the average delays gives an equation for the average residence time in the system.

$$\begin{aligned}\overline{R} &= U\overline{s} + (\overline{Q} - U)\overline{s} + \overline{s} \\ &= \overline{Q}\overline{s} + \overline{s}\end{aligned}$$

Now use Little's result and the Utilization Law to remove $\overline{Q}$.

$$\begin{aligned}\overline{R} &= \lambda\overline{R}\overline{s} + \overline{s} \\ &= U\overline{R} + \overline{s}\end{aligned}$$

Solving for $\overline{R}$ gives the result.

$$\overline{R} = \frac{\overline{s}}{1 - U}$$

We can immediately derive two corollary results. The average queue length can be found by using Little's result and the Utilization Law.

$$\overline{Q} = \frac{U}{1 - U}$$

The average waiting time is given by $\overline{W} = \overline{R} - \overline{s}$.

$$\overline{W} = \frac{\overline{s}U}{1 - U}$$

Behavior of the Residence Time Equation

Figure 1 shows a plot of $\overline{R}$ as a function of U when $\overline{s} = 1$.

Residence time increases very rapidly at utilizations beyond 80%. This leads to one of the most important and counterintuitive design insights in queueing theory.

Extra capacity is the price of low latencies. To achieve low residence times, you must allow the system to occasionally become idle.

Many people assume that expensive machines must be run at 100% utilization to justify their cost, or that low utilizations are a sign of waste in a system. In reality, some amount of idle time is necessary for good performance. In practice, 70% utilization is considered a good operating level.

One word of warning, though. Not all systems need minimal latency. The design process usually requires trading off between several factors, including latency, cost, and reliability. Analytic models aid designers by providing performance measures for each possible system configuration.

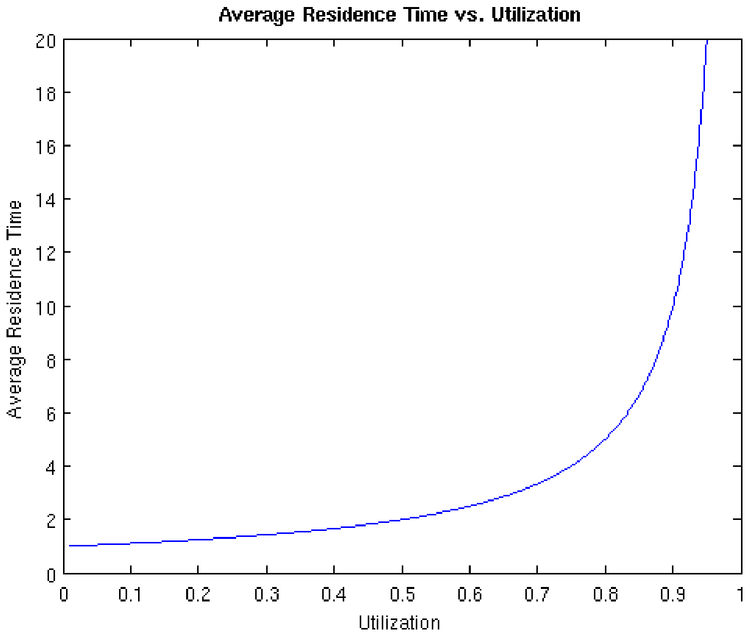

Figure 1: Residence time vs. utilizations for the M/M/1 queue

The Queue Length Distribution

Let N be a random variable denoting the number in the queue at a random moment in time. The queue length distribution, $P[N = k]$, is the probability of having k customers in the queue, including the one in service. Because of PASTA, this is equal to the probability of finding k in the queue at an arrival instant.

First, two observations.

$$\begin{aligned} &P[N = k] = P[N > k-1] - P[N > k] \\ &P[N > 0] = U \end{aligned}$$

Note that $P[N > k-1]$ is equal to the fraction of time that position k is occupied (if there are more than $k-1$ customers in the queue, there must be someone in position k). Therefore, we can apply Little's result just to position k to derive an expression for $P[N > k-1]$.

$$P[N > k-1] = \lambda_k \, \overline{s}_k$$

Here, λ_k denotes the throughput at position k and $\overline{s}_k$ denotes the average time a customer spends at k.

There are two ways a customer can reach position k.

- with probability $P[N = k-1]$, there are exactly $k-1$ customers in the queue at an arrival instant, so the new customer arrives directly to position k and waits for the residual life of the customer in service
- with probability $P[N \geq k]$, there are k or more customers in the queue at the arrival instant, so the new customer arrives to a position greater than k, then waits until it advances into position k

In both cases, the expected time spent at position k is simply $\overline{s}$.

Combining the two cases with Little's result,

$$\begin{aligned}
P[N > k-1] &= \lambda\,(P[N = k-1] + P[N \geq k])\,\overline{s} \\
&= \lambda\,P[N \geq k-1]\,\overline{s} \\
&= \lambda\,P[N > k-2]\,\overline{s} \\
&= U\,P[N > k-2]
\end{aligned}$$

We now have a recursive definition of $P[N > k-1]$ in terms of $P[N > k-2]$. The base case of the recursion is $P[N > 0] = U$. Simplifying yields

$$P[N > k-1] = U^k$$

Now, use this formula to evaluate the probability of having exactly k customers in the queue.

$$\begin{aligned}
P[N = k] &= P[N > k-1] - P[N > k] \\
&= U^k - U^{k+1} \\
&= U^k(1-U)
\end{aligned}$$

To verify the correctness of this formula, let's use it to calculate $\overline{Q}$, the expected number in the queue.

$$\begin{aligned}
\overline{Q} &= \sum_{k=0}^{\infty} k\,P[N = k] \\
&= \sum_{k=0}^{\infty} k\,U^k(1-U) \\
&= (1-U)\frac{U}{(1-U)^2} \\
&= \frac{U}{1-U}
\end{aligned}$$

The expected value calculation recovers the previous formula for $\overline{Q}$, exactly as it should.