CS 547 Lecture 15: The Paradox of Residual Life

Daniel Myers

The M/G/1 Queue

We're now ready to begin our analysis of the M/G/1 queue. This is a queueing system with Poisson arrivals, a single server, and a *general* service time distribution. By the end of the next lecture, we'll see that it's possible to predict the average residence time in such a queue using only the mean and variance of the service times.

Recall the master tagged customer equation, which we previously used to analyze the M/M/1 and M/D/1 queues, where we'll use $\overline{z}$ to represent the expected residual service time of a customer currently in service at the arrival instance.

$$\overline{R} = U\overline{z} + (\overline{Q} - U)\overline{s} + \overline{s}$$

We've previously seen that identifying a value of $\overline{z}$ would allow us to solve this equation and obtain an expression for the average residence time, $\overline{R}$. Therefore, our goal is to understand the behavior of the residual service time, so that we can derive an expression for $\overline{z}$.

Residuals vs. Service Times

Suppose we have an M/G/1 queue with servive time distribution $f_S(x)$. Let Y be a random variable denoting the length of the service time interrupted by a new arrival. Y has a distribution, which we'll call $f_Y(x)$.

In general, these two distributions are not the same. That is, $f_S(x) \neq f_Y(x)$. Here's an example to illustrate this point.

Suppose we have a queue with only two possible service times, such that 5/6 of all jobs receive a service time of 1 and 1/6 of all jobs have a service time of 5. What is the probability that a randomly arriving customer arrives while a job of length 5 is running?

Suppose we observe some very large number N of jobs. The probability that random customer finds a job of length 5 in service can be represented as

$$\frac{\text{Total service time taken up by jobs of length 5}}{\text{Total service time taken up by all jobs}}$$

Substituting in the values

$$\frac{N \cdot \frac{1}{6} \cdot 5}{N \cdot \frac{5}{6} \cdot 1 + N \cdot \frac{1}{6} \cdot 5} = \frac{1}{2}$$

It turns out that half of the queue's total service time is dedicated to jobs of length 5, even though they make up only 1/6 of all jobs.

This is the paradox of resdiual life: a newly arriving customer is more likely to experience a long residual service time, because customers are more likely to arrive during long service periods.[1]

[1] This isn't a true paradox, because it isn't a violation of logic. It's just somewhat counterintuitive.

The Mass-Weighted Distribution

To actually calculate the expected residual service time $\overline{z}$, we need to find the probability distribution governing the length of an interrupted service period, $f_Y(x)$.

In our previous example, we calculated the probability of interrupting a job of length 5 by reasoning about the fraction of total load on the server dedicated to serving jobs of length 5. We can generalize this concept of "fraction of load" to arbitrary distributions using the *mass-weighted distribution function*.

$$f_X^{mass-weight} = \frac{x f_X(x)}{\overline{X}}$$

The mass-weighted distribution identifies the fraction of total output dedicated to jobs of size x.

Applying this idea to the service time distribution allows us to find an expression for $f_Y(x)$.

$$\begin{aligned}
f_Y(x) &= \text{Prob. a newly arriving job finds a job of length } x \text{ in service} \\
&= \text{Fraction of total server load due to jobs of length } x \\
&= f_S^{mass-weight}(x) \\
&= \frac{x f_S(x)}{\overline{s}}
\end{aligned}$$

In the next lecture, we'll use this fact to derive the actual value of $\overline{z}$ and a formula for $\overline{R}$ in the M/G/1 queue.