CS 547 Lectures 26 and 27: Priority Queueing

Daniel Myers

Multi-Class Models

To this point, we've only considered models with one customers class, where all jobs arrive according to the same Poisson process and have the same average service time. Now, we'll generalize the basic M/M/1 and M/G/1 models to account for multiple classes of customers, with each class having its own distinct arrival rate and average service time.

We'll derive most of these results for only two customer classes – extensions to more than two classes are straightforward. Class 1 customers arrive at rate λ_1 and require an average service time of $\overline{s}_1$. Class 2 customers arrive at rate λ_2 and receive an average service time of $\overline{s}_2$.

Little's Result for Multiple Classes

Little's result applies to each class individually. For example, the average number of class 1 customers in the queue, including one in service, is

$$\overline{Q}_1 = \lambda_1 \overline{R}_1$$

The fraction of time the server is busy with a class 1 customer is

$$U_1 = \lambda_1 \overline{s}_1$$

We can make a measurement-based argument for the validity of these results. Suppose that we measured the system for some large time period T. During the measurement period, we observed C_1 completions of class 1 customers. The total service time accumulated by all class 1 customers during the period was D_1. The throughput of class 1 customers is simply

$$\lambda_1 = \frac{C_1}{T}$$

The average service time required for a class 1 customer is the total accumulated service time for all class 1 customers, divided by the number of class 1 customers (assuming the measurement period is large and statistically representative).

$$\overline{s}_1 = \frac{D_1}{C_1}$$

Finally, the fraction of time the server is busy serving a class 1 customer is

$$\begin{aligned} U_1 &= \frac{D_1}{T} \\ &= \frac{C_1}{T}\frac{D_1}{C_1} \\ &= \lambda_1 \overline{s}_1 \end{aligned}$$

Two Classes with Equal Priorities

For our first multi-class model, let's consider a queue with two classes, exponentially distributed service times, and *equal* priorities. The two classes have different arrival rates and average service times, but customers are still served in FCFS order.

When a class 1 customer arrives, it must wait for three sources of delay.

- any class 1 customers already in the queue, possibly including one in service
- any class 2 customers already in the queue, possibly including one in service
- its own class 1 service time

Combining these three cases, we have

$$\overline{R}_1 = \overline{Q}_1\overline{s}_1 + \overline{Q}_2\overline{s}_2 + \overline{s}_1$$

Similarly, a class 2 customer must wait for any class 1 or class 2 customers already in the queue, plus a class 2 service time.

$$\overline{R}_2 = \overline{Q}_1\overline{s}_1 + \overline{Q}_2\overline{s}_2 + \overline{s}_2$$

Using Little's result to simplify $\overline{Q}_1$ and $\overline{Q}_2$ yields a system of two linear equations with two unknowns.

$$\overline{R}_1 = U_1\overline{R}_1 + U_2\overline{R}_2 + \overline{s}_1$$
$$\overline{R}_2 = U_1\overline{R}_1 + U_2\overline{R}_2 + \overline{s}_2$$

The system can be solved algebraically or numerically to find the values of $\overline{R}_1$ and $\overline{R}_2$.

Two Classes with Preemptive Priority

Now, consider the case where class 1 has *preemptive priority* over class 2. If a class 1 customer arrives to find a class 2 customer in service, the class 2 customer is preempted so the class 1 customer can run immediately.

A class 1 customer needs to wait for other class 1 customers already in the queue, possibly including one in service, but it never needs to wait for any class 2 customers.

$$\begin{aligned}\overline{R}_1 &= \overline{Q}_1\overline{s}_1 + \overline{s}_1 \\ &= \frac{\overline{s}_1}{1 - U_1}\end{aligned}$$

From the persepective of class 1 customers, this system behaves just like an M/M/1 queue.

A class 2 customer must wait for any class 1 and class 2 customers that are already in the queue when it arrives, *and* any class 1 customers that arrive during its residence period. If the residence period of the class 2 customer is $\overline{R}_2$, we expect $\lambda_1\overline{R}_2$ class 1 customers to arrive before the tagged class 2 customer can leave the queue. Each class 1 arrival requires an average service time of $\overline{s}_1$.

$$\begin{aligned}\overline{R}_2 &= \overline{Q}_1\overline{s}_1 + \overline{Q}_2\overline{s}_2 + \lambda_1\overline{R}_2\overline{s}_1 + \overline{s}_2 \\ &= \frac{\overline{s}_2 + \overline{Q}_1\overline{s}_1}{1 - U_1 - U_2}\end{aligned}$$

Note that a class 2 customer can be interrupted multiple times before it finally receives all of its service and departs.

Two Classes with Non-Preemptive Priority

If class 1 has *non-preemptive* priority over class 2, then a class 2 customer cannot be preempted once it enters service. Class 1 customers still have priority over any class 2 customers that are waiting but not being served.

A class 1 customer now has two sources of waiting time: time for any other class 1 customers in the queue, and time for any class 2 customer that may be in service at the arrival instant. By the PASTA property, the probability of finding a class 2 customer in service at an arrival instant is U_2. Service times are exponentially dsitributed, so the expected residual service time is simply $\overline{s}_2$. After waiting, the class 1 customer receives its own service.

$$\overline{R}_1 = U_2\overline{s}_2 + \overline{Q}_1\overline{s}_1 + \overline{s}_1$$

A class 2 customer still has to wait for any customers of either class that are in the queue when it arrives. It also needs to wait for any class 1 customers that arrive, but only until it enters service, after which point it can't be preempted. A class 2 customer spends an average of $\overline{R}_2 - \overline{s}_2$ waiting but not receiving service. Therefore, we expect $\lambda_1(\overline{R}_2 - \overline{s}_2)$ class 1 customers to arrive before the class 2 customer enters service.

$$\overline{R}_2 = \overline{Q}_1\overline{s}_1 + \overline{Q}_2\overline{s}_2 + \lambda_1(\overline{R}_2 - \overline{s}_2)\overline{s}_1 + \overline{s}_2$$

More Than Two Classes

What if we have preemptive priority and more than two classes?

Consider a class k customer, where classes $1 \ldots k-1$ all have priority over class k. The class k customer must wait for

- any customers of equal or higher priority already in the queue
- any customers of strictly higher priority that arrive during the class k customer's residence period
- its own service time

Combining all three cases:

$$\overline{R}_k = \sum_{c=1}^{k} \overline{Q}_c\overline{s}_c + \sum_{c=1}^{k-1} \lambda_c\overline{R}_k\overline{s}_c + \overline{s}_k$$

The first summation represents the waiting time due to customers of classes 1 through k that are already in the queue when a class k customer arrives. The second summation accounts for all of the customers of classes 1 through $k-1$ that arrive while the class k customer is in the queue. The final term is the class k customer's own average service time.

It's easy to modify this result to use non-preemptive priority.

General Service Times

These results assume exponentially distributed service times for all classes. To incorporate general service times into priority queueing models, we need to adapt our equations to include the expected residual life of the customer currently in service at an arrival instant.

We'll analyze a two-class system with general service times and preemptive priority.

Let class 1 and class 2 have squared coefficients of variation of c_1^2 and c_2^2, respectively.

When a class 1 customer arrives to the system, it must wait for

- the residual service time of a class 1 customer in service
- any class 1 customers waiting but not begin served
- its own service time

The probabiltiy of finding a class 1 customer in service is U_1. By Little's result, the expected number of class 1 customers waiting but not being served is $\overline{Q}_1 - U_1$. The expected residual service time of a class 1 customer is $\frac{\overline{s}_1}{2}(1 + c_1^2)$. Combining all three terms,

$$\overline{R}_1 = U_1 \frac{\overline{s}_1}{2}(1 + c_1^2) + (\overline{Q}_1 - U_1)\overline{s}_1 + \overline{s}_1$$

This equation is very similar to the single class M/G/1 residence time equation, as we'd expect. To obtain the final equation, use Little's result and the Utilization Law to solve for $\overline{R}_1$.

A class 2 customer must wait for

- the residual time of a class 1 customer, if one is in service
- the residual time of a class 2 customer, if one is in service
- any class 1 customers waiting but not being served
- any class 2 customers waiting but not being served
- any class 1 customers that arrive while the class 2 customer is in residence
- its own service time

Combining all of the delay sources,

$$\overline{R}_2 = U_1 \frac{\overline{s}_1}{2}(1 + c_1^2) + (\overline{Q}_1 - U_1)\overline{s}_1 + U_2 \frac{\overline{s}_2}{2}(1 + c_2^2) + (\overline{Q}_2 - U_2)\overline{s}_2 + \lambda_1 \overline{R}_2 \overline{s}_1 + \overline{s}_2$$