Priority M/G/1 Queue with and without Preemption

Version 2.3: Jun 9, 2005 10:50 am GMT-5

Bart Sinclair

Abstract

(Blank Abstract)

Priority M/G/1 Queue with and without Preemption

1 M/G/1 Queue with Priority

Another useful extension to the classical M/G/1 queue is the priority M/G/1 queue. The workload for the queue consists of two or more classes of jobs. Each job belongs to a single class. Jobs receive service in an order that is based on a predetermined priority ranking among the classes of jobs. Each class of job may have its own, distinct arrival rate (with exponentially distributed interarrival times) and service time distribution.

The objective is to determine the average waiting time for jobs of each class and the average number of jobs of each class in the queue. These will depend on whether the queueing discipline is **preemptive** or **non-preemptive**. A non-preemptive queueing discipline requires a job that begins service to complete its service without interruption. In a preemptive priority queue, if a job arriving at the queue finds a job of lower priority in service, the arriving job preempts the job being served and begins service immediately. A preempted job will resume service, at the point at which its service was suspended, as soon as there are no higher priority jobs remaining in the queue. We call a preemptive scheduling policy with this latter property **preemptive resume**.

Notation:

- X_i = service time random variable for priority i jobs
- R_i = residual service time random variable for priority i jobs
- λ_i = arrival rate for priority i jobs (Poisson arrivals)
- $\rho_i = \lambda_i \overline{X_i}$ = utilization of the server by jobs of priority i
- $\sigma_i = \sum_{j=1}^{i} \rho_j$ = utilization of the server by jobs of priority 1 to i
- W_i = random variable for the time a job of priority i spends waiting from arrival until service begins

*http://creativecommons.org/licenses/by/1.0

- N_i = random variable for the number of jobs in queue i (not counting jobs for which service has already begun)
- T_i = random variable for the time a job of priority i spends in the system from arrival until the completion of service

$$W_i = W_i^q + W_i^a + W_i^r$$

where

- W_i^q = random variable for the delay seen by an arriving class i job due to other jobs in the system which have not yet begun service
- W_i^a = random variable for the delay seen by an arriving class i job due to jobs which arrive after it but before it begins service
- W_i^r = random variable for the delay seen by an arriving class i job due to jobs whose service is in progress (either actually in service or preempted) at the time of the arrival

We want to know $\overline{W_i} = \overline{W_i^q} + \overline{W_i^a} + \overline{W_i^r}$.

The first two terms depend only on the priority of the arriving job and not on whether queueing is preemptive or non-preemptive.

$$\overline{W_i^q} = \sum_{j=1}^{i} \left(\overline{N_j X_j}\right) = \sum_{j=1}^{i} \left(\lambda_j \overline{W_j X_j}\right) = \sum_{j=1}^{i} \left(\rho_j \overline{W_j}\right)$$

$$\overline{W_i^a} = \sum_{j=1}^{i-1} \left(\lambda_j \overline{W_i X_j}\right) = \sum_{j=1}^{i-1} \left(\rho_j \overline{W_i}\right)$$

Substituting these two expressions gives

$$\overline{W_i} = \sum_{j=1}^{i} \left(\rho_j \overline{W_j}\right) + \sum_{j=1}^{i-1} \left(\rho_j \overline{W_i}\right) + \overline{W_i^r}$$

which we can manipulate into two slightly different forms:

$$\overline{W_i}\left(1 - \sigma_{i-1}\right) = \sum_{j=1}^{i} \left(\rho_j \overline{W_j}\right) + \overline{W_i^r} \tag{1}$$

$$\overline{W_i}\left(1 - \sigma_i\right) = \sum_{j=1}^{i-1} \left(\rho_j \overline{W_j}\right) + \overline{W_i^r} \tag{2}$$

(we define σ_0 to be 0).

From Equation 1, we get (for $i \geq 2$)

$$\overline{W_{i-1}}\left(1 - \sigma_{i-2}\right) = \sum_{j=1}^{i-1} \left(\rho_j \overline{W_j}\right) + \overline{W_{i-1}^r}$$

$$\Rightarrow \sum_{j=1}^{i-1} \left(\rho_j \overline{W_j}\right) = \overline{W_{i-1}}\left(1 - \sigma_{i-2}\right) - \overline{W_{i-1}^r}$$

Substituting the in Equation 2 gives

$$\overline{W_i}\left(1-\sigma_i\right)=\overline{W_{i-1}}\left(1-\sigma_{i-2}\right)-\overline{W_{i-1}^r}+\overline{W_i^r}$$

$$\overline{W_i}=\frac{\overline{W_{i-1}}\left(1-\sigma_{i-2}\right)+\overline{W_i^r}-\overline{W_{i-1}^r}}{1-\sigma_i}$$

Now we need to consider the effects of what happens when a job first arrives.

1.1 No Preemption

If an arrival never preempts a job in service, the time that a class i job expects to wait on the job in service must be the same as the time that a class $i-1$ job expects to wait: $\overline{W_i^r}=\overline{W_{i-1}^r}$. Hence,

$$\overline{W_i}=\frac{\overline{W_{i-1}}\left(1-\sigma_{i-2}\right)}{1-\sigma_i} \tag{3}$$

for all $i\geq 2$. For $i=1$,

$$\overline{W_1}=\overline{W_1^r}+\overline{N_1X_1}=\overline{W_1^r}+\lambda_1\overline{W_1X_1}=\overline{W_1^r}+\rho_1\overline{X_1}=\frac{\overline{W_1^r}}{1-\rho_1}$$

With this as a starting point, we can solve the recurrence (Equation 3).

$$\overline{W_2}=\frac{\overline{W_1}\left(1-\sigma_0\right)}{1-\sigma_2}=\frac{\overline{W_1^r}}{\left(1-\sigma_2\right)\left(1-\sigma_1\right)}$$

$$\overline{W_3}=\frac{\overline{W_1}\left(1-\sigma_1\right)}{\left(1-\sigma_3\right)\left(1-\sigma_2\right)\left(1-\sigma_1\right)}=\frac{\overline{W_1^r}}{\left(1-\sigma_3\right)\left(1-\sigma_2\right)}$$

and in general

$$\overline{W_i}=\frac{\overline{W_1^r}}{\left(1-\sigma_i\right)\left(1-\sigma_{i-1}\right)}$$

The expected time an arrival waits on a class i job in service at the time of arrival is the probability of a job seeing a class i job in service at its arrival times the expected residual service time for a class i job. Therefore,

$$\overline{W_1^r}=\sum_{i=1}^{M}\left(\rho_i\overline{R_i}\right)=\sum_{i=1}^{M}\left(\lambda_i\overline{X_i}\frac{\overline{X_i^2}}{2\overline{X_i}}\right)=\sum_{i=1}^{M}\left(\frac{\lambda_i\overline{X_i^2}}{2}\right)$$

where M is the number of classes. Then

$$\overline{W_i}=\frac{\sum_{j=1}^{M}\left(\frac{\lambda_j\overline{X_j^2}}{2}\right)}{\left(1-\sigma_i\right)\left(1-\sigma_{i-1}\right)} \tag{4}$$

$$\overline{T_i}=\frac{\sum_{j=1}^{M}\left(\frac{\lambda_j\overline{X_j^2}}{2}\right)}{\left(1-\sigma_i\right)\left(1-\sigma_{i-1}\right)}+\overline{X_i} \tag{5}$$

1.2 Preemption

Let C_i be the random variable for the time it takes a class i job to complete service (that is, the time between starting service and finishing, including the time that the job is preempted). Then during an average service completion interval for a class i job, $\lambda_j \overline{C_i}$ class j jobs arrive, and hence

$$\overline{C_i} = \overline{X_i} + \sum_{j=1}^{i-1} \left(\lambda_j \overline{C_i X_j}\right) = \overline{X_i} + \sigma_{j-1}\overline{C_i}$$

$$\Rightarrow \overline{C_i} = \frac{\overline{X_i}}{1-\sigma_{i-1}}$$

The probability that a class j job service is in progress (either a class j job is in service or is preempted) is just the fraction of time such a job is in the system, or $\lambda_j \overline{C_j}$, since only one class j job can be in service or preempted at a time. Therefore,

$$\Pr\left[\text{class j service in progress at an arrival}\right] = \lambda_j \frac{\overline{X_j}}{1-\sigma_{j-1}} = \frac{\rho_j}{1-\sigma_{j-1}}$$

and

$$\overline{W_i^r} = \sum_{j=1}^{i} \left(\frac{\rho_j \overline{R_j}}{1-\sigma_{j-1}}\right)$$

From this we see that

$$\overline{W_i^r} - \overline{W_{i-1}^r} = \frac{\rho_i \overline{R_i}}{1-\sigma_{i-1}}$$

Substituting this into the previous expression for $\overline{W_i}$ leads to the following simple recurrence equation:

$$\overline{W_i} = \frac{\left(1-\sigma_{i-2}\right)\overline{W_{i-1}} + \frac{\rho_i \overline{R_i}}{1-\sigma_{i-1}}}{1-\sigma_i} = \frac{\left(1-\sigma_{i-1}\right)\left(1-\sigma_{i-2}\right)\overline{W_{i-1}} + \rho_i \overline{R_i}}{\left(1-\sigma_i\right)\left(1-\sigma_{i-1}\right)}$$

Solving the recurrence:

$$\overline{W_1} = \rho_1 \overline{R_1} + \lambda_1 \overline{W_1 X_1} = \frac{\rho_1 \overline{R_1}}{1-\sigma_1}$$

$$\overline{W_2} = \frac{\left(1-\sigma_1\right)\left(1-\sigma_0\right)\overline{W_1} + \rho_2 \overline{R_2}}{\left(1-\sigma_2\right)\left(1-\sigma_1\right)} = \frac{\left(1-\sigma_1\right)1\frac{\rho_1 \overline{R_1}}{1-\sigma_1} + \rho_2 \overline{R_2}}{\left(1-\sigma_2\right)\left(1-\sigma_1\right)} = \frac{\rho_1 \overline{R_1} + \rho_2 \overline{R_2}}{\left(1-\sigma_2\right)\left(1-\sigma_1\right)}$$

and in general,

$$\overline{W_i} = \frac{\sum_{j=1}^{i}\left(\rho_j \overline{R_j}\right)}{\left(1-\sigma_i\right)\left(1-\sigma_{i-1}\right)}$$

which can be verified by substitution into the recurrence equation.

$$\begin{aligned} \overline{W_i} &= \frac{(1-\sigma_{i-1})(1-\sigma_{i-2})\overline{W_{i-1}} + \rho_i \overline{R_i}}{(1-\sigma_i)(1-\sigma_{i-1})} \\ &= \frac{(1-\sigma_{i-1})(1-\sigma_{i-2})\frac{\sum_{j=1}^{i}\left(\rho_j \overline{R_j}\right)}{(1-\sigma_{i-1})(1-\sigma_{i-2})} + \rho_i \overline{R_i}}{(1-\sigma_i)(1-\sigma_{i-1})} \\ &= \frac{\sum_{j=1}^{i}\left(\rho_j \overline{R_j}\right) + \rho_i \overline{R_i}}{(1-\sigma_i)(1-\sigma_{i-1})} \end{aligned} \tag{6}$$

Finally, the response time for a class i job in a preemptive resume system is the waiting time plus the completion (not execution!) time:

$$\overline{T_i} = \frac{\sum_{j=1}^{i}\left(\rho_j \overline{R_j}\right)}{\left(1-\sigma_i\right)\left(1-\sigma_{i-1}\right)} + \frac{\overline{X_i}}{1-\sigma_{i-1}}$$

We can rewrite the equations for waiting time and response time in terms of the first and second principal moments of the service time distribution:

$$\overline{W_i} = \frac{\sum_{j=1}^{i} \left(\frac{\lambda_j \overline{X_j{}^2}}{2} \right)}{(1 - \sigma_i)(1 - \sigma_{i-1})} \tag{7}$$

$$\overline{T_i} = \frac{\sum_{j=1}^{i} \left(\frac{\lambda_j \overline{X_j{}^2}}{2} \right)}{(1 - \sigma_i)(1 - \sigma_{i-1})} + \frac{\overline{X_i}}{1 - \sigma_{i-1}} \tag{8}$$