An Analysis of the Preemptive Repeat Queueing Discipline

Tony Field

August 30, 2006

Abstract

An analysis of two variants of preemptive repeat (or preemptive repeat) queueing discipline is presented: one in which the preempted customer's service time distribution is re-sampled when the customer rejoins service, and one where the service time at each subsequent visit is the same. The two variants are *not* the same, even for the exponential distribution.

1 Preemptive Repeat Different

With exponentially-distributed service times, the preemptive repeat policy is identical to the preemptive resume policy, because the residual service time distribution is the same as that of the original service time.

Consider the problem of determining the distribution of the number of visits made to the server. At each visit the customer is preempted if the time to the next arrival is smaller than the time to completion.

Because the service demand at each visit is recomputed (resampling) the successive service times are independent. Let N the number of visits made to the server by a customer. Because of the independence assumption the probability that a customer will be preempted during their service time is simply $\mu/(\lambda+\mu)$ where λ and μ are the arrival rate and service rate respectively. N thus has a geometric distribution with

$$\begin{aligned} P(N=n) &= \left(\frac{\lambda}{\lambda+\mu}\right)^{n-1} \frac{\mu}{\lambda+\mu} \\ &= \frac{\mu}{\lambda}\left(\frac{\lambda}{\lambda+\mu}\right)^{n} \end{aligned}$$

The mean number of visits is $E(N) = (\lambda+\mu)/\mu$ and the variance is $V(N) = (\lambda+\mu)\lambda/\mu^2$.

2 Preemptive Repeat Identical

If the service time distribution is sampled once for each customer and the identical time re-used at each visit to the server, then the successive service times seen by the server are no longer independent.

Consider again the problem of determining the distribution of the number of visits made to the server and the specific case of exponentially-distributed inter-arrival and service times, with rates λ and μ respectively. Let S denote the service time random variable and N the number of visits made to the server by a customer. Let I denote the inter-arrival times of new customers. Because the inter-arrival times are independent of the service times the probability that a customer whose service time is s is preempted is $P(I \leq s) = 1 - e^{-\lambda s}$. The number of visits made by a customer whose service time is t is thus geometrically distributed with parameter $e^{-\lambda t}$:

$$P(N = n \mid t) = \left(1 - e^{-\lambda t}\right)^{n-1} \; e^{-\lambda t}$$

Now, deconditioning on t,

$$\begin{aligned} P(N = n) &= \int_0^\infty P(N = n \mid t)\, \mu e^{-\mu t} dt \\ &= \int_0^\infty \mu e^{-(\lambda+\mu)t} \left(1 - e^{-\lambda t}\right)^{n-1} dt \end{aligned}$$

for $n \geq 1$. This is messy, but the integral can be reduced to a sum, thus:

$$\begin{aligned} P(N = n) &= \int_0^\infty \mu e^{-(\lambda+\mu)t} \left(1 - e^{-\lambda t}\right)^{n-1} dt \\ &= \int_0^\infty \mu e^{-(\lambda+\mu)t} \left(\sum_{i=0}^{n-1} (-1)^i \left(\begin{array}{c} n-1 \\ i \end{array} \right) e^{-i\lambda t} \right) dt \\ &= \sum_{i=0}^{n-1} (-1)^i \left(\begin{array}{c} n-1 \\ i \end{array} \right) \mu \int_0^\infty e^{-((i+1)\lambda+\mu)} dt \\ &= \sum_{i=0}^{n-1} (-1)^i \left(\begin{array}{c} n-1 \\ i \end{array} \right) \frac{\mu}{(i+1)\lambda + \mu} \end{aligned}$$

2.1 Mean Number of Server Visits

The mean number of visits to the server is given by

$$\begin{aligned} E(N) &= \sum_{n=1}^{\infty} n \int_0^\infty \mu e^{-(\lambda+\mu)t} \left(1 - e^{-\lambda t}\right)^{n-1} dt \\ &= \int_0^\infty \mu e^{-(\lambda+\mu)t} \sum_{n=1}^{\infty} n \left(1 - e^{-\lambda t}\right)^{n-1} dt \end{aligned}$$

Let u be any function of t with $a \leq t \leq b$ for some a and b and assume that $0 \leq u < 1, a \leq t \leq b$. Then

$$\begin{aligned}\sum_{n=1}^{\infty} nu^{n-1} &= \sum_{n=1}^{\infty} \frac{dt}{du}\frac{d}{dt}u^n \\ &= \frac{dt}{du}\frac{d}{dt}\left(\sum_{n=1}^{\infty} u^n\right) \\ &= \frac{1}{(1-u)^{-2}}\end{aligned}$$

Similarly, note also that

$$\sum_{n=1}^{\infty} nu^n = \frac{u}{(1-u)^2}$$

which will be useful later on. Thus

$$\sum_{n=1}^{\infty} n\left(1-e^{-\lambda t}\right)^{n-1} = e^{2\lambda t}$$

and

$$\begin{aligned}E(N) &= \int_0^{\infty} \mu e^{-(\lambda+\mu)t} e^{2\lambda t} \\ &= \frac{\mu}{\mu-\lambda}\end{aligned}$$

2.2 Variance of the Number of Server Visits

Putting $u = 1 - e^{-\lambda t}$, the second moment of N is given by

$$\begin{aligned}E(N^2) &= \sum_{n=1}^{\infty} n^2 \int_0^{\infty} \mu e^{-(\lambda+\mu)t} u^{n-1} dt \\ &= \mu \int_0^{\infty} e^{-(\lambda+\mu)t}\left(\sum_{n=1}^{\infty} n^2 u^{n-1}\right) dt\end{aligned}$$

The summation can be simplified thus:

$$\begin{aligned}\sum_{n=1}^{\infty} n^2 u^{n-1} &= \sum_{n=1}^{\infty} n(nu^{n-1}) \\ &= \sum_{n=1}^{\infty} n\frac{d}{du}u^n \\ &= \frac{d}{du}\sum_{n=1}^{\infty} nu^n\end{aligned}$$

$$
\begin{aligned}
&= \frac{d}{du}\frac{u}{(1-u)^2} \\
&= \frac{1+u}{(1-u)^3}
\end{aligned}
$$

Expanding u, $E(N^2)$ therefore simplifies to

$$
\begin{aligned}
E(N^2) &= \mu \int_0^\infty e^{-(\lambda+\mu)t} \left(\frac{2 - e^{-\lambda t}}{e^{-3\lambda t}} \right) dt \\
&= 2\mu \int_0^\infty e^{-(\mu-2\lambda)t} dt - \mu \int_0^\infty e^{-(\mu-\lambda)t} dt \\
&= -\left[\frac{2\mu}{\mu - 2\lambda} e^{-(\mu-2\lambda)t} \right]_0^\infty + \left[\frac{\mu}{\mu-\lambda} e^{-(\mu-\lambda)t} \right]_0^\infty \\
&= \frac{2\mu}{\mu-2\lambda} - \frac{\mu}{\mu-\lambda} \\
&= \frac{\mu^2}{(\mu-\lambda)(\mu-2\lambda)}
\end{aligned}
$$

Thus we obtain,

$$
\begin{aligned}
V(N) &= E(N^2) - E(N)^2 \\
&= \frac{\mu^2}{(\mu-\lambda)(\mu-2\lambda)} - \left(\frac{\mu}{\mu-\lambda} \right)^2 \\
&= \frac{\lambda\mu^2}{(\mu-\lambda)^2(\mu-2\lambda)}
\end{aligned}
$$

The most curious aspect of this result is that the variance of the number of visits is undefined when $\mu \leq 2\lambda$. Compare this with the usual stability condition for the M/M/1 queue, for example. Thus, for $\lambda < \mu \leq 2\lambda$ the variance of the number of server visits is infinite, even though the mean is finite, and the service process has many of the characteristics of a heavy-tailed distribution. Simulating this queue as μ approaches 2λ from above is notoriously difficult as the system is subject to rare events.

2.3 Dilated Service Time

Because a customer may make several visits to the server, each with the same service demand, the total time spent at the server will, on average, be greater than the initial demand.

The mean total time spent at the server comprises the sum of a number of inter-arrival times and the final service demand, which is satisfied exactly once without interruption. The probability that a customer makes n visits to the server, $n > 0$, is known from above. We will call the first $n-1$ interrupted visits "failed visits".

Each inter-arrival time is conditioned on the fact that it is smaller than the service demand, otherwise the customer would have completed service. If t is the service demand and I_t is the random variable denoting inter-arrival times smaller than t, the density function for I_t is that of a "normalised" exponential distribution, given by:

$$f_t(x) = \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda t}}$$

Let T be the total time spent at the server during its failed visits, i.e. prior to the customer's final visit, when its service demand is finally satisfied. The mean total time spent at the server is then $E(T) + \mu^{-1}$.

To compute $E(T)$ we will first derive $L(s)$, the Laplace Transform of $f_T(x)$, the density function of T. Then,

$$E(T) = -\left[\frac{d}{ds}L(s)\right]_{s=0}$$

The distribution of T is made up of a convolution of normalised exponentials, in the above sense, weighted by the visit count probability distribution, and deconditioned on the service demand.

Beginning with the normalised exponentials, the Laplace Transform of $f_t(x)$ above is given by

$$\begin{aligned} L_t(s) &= \int_0^t e^{-sx} f_t(x) dx \\ &= \int_0^t e^{-sx} \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda t}} \\ &= \frac{\lambda}{\lambda + s} \frac{1 - e^{-(\lambda+s)t}}{1 - e^{-\lambda t}} \end{aligned}$$

Putting it all together, we obtain:

$$\begin{aligned} L(s) &= \int_0^\infty \mu e^{-\mu t} \sum_{n=1}^{\infty} \left(P(N = n)\, L_t(s)^{n-1}\right) dt \\ &= \int_0^\infty \mu e^{-\mu t} \sum_{n=1}^{\infty} \left(\left(1 - e^{-\lambda t}\right)^{n-1}\ e^{-\lambda t} \left(\frac{\lambda}{\lambda + s} \frac{1 - e^{-(\lambda+s)t}}{1 - e^{-\lambda t}} \right)^{n-1} \right) dt \\ &= \mu \int_0^\infty e^{-(\lambda+\mu)t} \sum_{n=0}^{\infty} \left(\frac{\lambda(1 - e^{-(\lambda+s)t})}{\lambda + s} \right)^n dt \\ &= \mu \int_0^\infty e^{-(\lambda+\mu)t} \left(\frac{\lambda + s}{s + \lambda e^{-(\lambda+s)t}} \right) dt \end{aligned}$$

In principle, it is possible to invert this numerically to derive $f_T(x)$, but here we will focus on deriving the mean, $E(T)$. Differentiating $L(s)$ with respect to

s, we obtain

$$\begin{aligned}\frac{d}{ds}L(s) &= \frac{d}{ds}\left[\mu\int_0^\infty e^{-(\lambda+\mu)t}\left(\frac{\lambda+s}{s+\lambda e^{-(\lambda+s)t}}\right)dt\right] \\ &= \mu\int_0^\infty e^{-(\lambda+\mu)t}\left(\frac{s+\lambda e^{-(\lambda+s)t}-(\lambda+s)(1-\lambda te^{-(\lambda+s)t})}{\left(s+\lambda e^{-(\lambda+s)t}\right)^2}\right)dt\end{aligned}$$

Setting $s=0$ we obtain

$$\begin{aligned}E(T) &= -\left(\mu\int_0^\infty e^{-(\lambda+\mu)t}\frac{\lambda e^{-\lambda t}-\lambda(1-\lambda te^{-\lambda t})}{\left(\lambda e^{-\lambda t}\right)^2}dt\right) \\ &= -\left(\frac{\mu}{\lambda}\int_0^\infty e^{-(\lambda+\mu)t}\left(e^{\lambda t}-e^{2\lambda t}+\lambda te^{\lambda t}\right)\right)dt \\ &= \frac{\mu}{\lambda}\left(\frac{1}{\mu-\lambda}-\frac{1}{\mu}-\frac{\lambda}{\mu^2}\right) \\ &= \frac{1}{\mu-\lambda}-\frac{1}{\mu}\end{aligned}$$

The total time spent at the server *excluding the queueing time* is thus $1/(\mu-\lambda)$. Extraordinarily, this is the same as the waiting time in an M/M/1 FIFO queue!