Sojourn Times in a Priority Queue with Preemption

Tony Field
Department of Computing
Imperial College London
London SW7 2AZ

December 21, 2005

Abstract

The Laplace transform of the sojourn time density of tasks of given priority class in a preemptive multi-class priority queueing system is derived. The analysis is a straightforward extension to that previously published for a queue with probabilistic overtaking. The density functions can be recovered by standard Laplace transform inversion techniques. Some results are presented.

1 The Problem

Consider a priority queueing system with P priority classes where the highest priority is 1. Tasks of class $c, 1 \leq c \leq P$ arrive according to a Poisson process with rate λ_c and join a FIFO queue for that class. There is a single server and the service times are exponentially distributed with mean μ^{-1}. The server services all tasks of class c before servicing those of classes $c' > c, 1 \leq c, c' \leq P$. In this formulation of the problem the task in service is assumed to be preempted by the arrival of a higher-priority task, but not by the arrival of a task of equal or lower priority.

We wish to determine the sojourn time density of tasks of a specified priority class. Note that the cumulative distribution function can also be derived straightforwardly.

2 Analysis

Consider tasks of some chosen class $1 \leq c \leq P$ and let T denote the sojourn time random variable of a task of class c. Tasks with priority higher than c arrive at rate $\lambda = \sum_{i=1}^{c-1} \lambda_i$ and those with lower or equal priority at rate $\lambda' = \sum_{i=c}^{P} \lambda_i$. Consider a special monitored task of class c and let the random

variable N denote the total number of tasks 'ahead' of the monitored task in priority order, i.e. the total number of tasks (classes $c' \leq c$) that will be serviced before it. This includes tasks of higher priority and other tasks of class c that arrived in the queue before the monitored task. Define for $n \geq 0$,

$$F_n(t) = \Pr\{T \leq t \mid N = n\}$$

Assume that the monitored task arrives at time 0 and consider the intervals $[0, h)$ and $[h, t + h)$. During the interval h there will be an arrival of a task of class $i < c$ (higher priority) with probability $\lambda h + o(h)$, an arrival of class $i \geq c$ (lower priority) with probability $\lambda' h + o(h)$ and a completion with probability $\mu h + o(h)$; all other combinations of two or more events in the same interval occur with probability $o(h)$.

Using the fact that the forward recurrence time of an exponential random variable is an identical exponential random variable,

$$\begin{aligned} F_n(t+h) &= (1 - (\lambda + \lambda' + \mu)h)\, F_n(t) \\ &+ \lambda h\, F_{n+1}(t) \\ &+ \lambda' h\, F_n(t) \\ &+ \mu h\, F_{n-1}(t) \end{aligned}$$

for $n > 0$. Note that the terms involving λ' are superfluous since the monitored task cannot be influenced in any way by tasks of lower or equal priority. However, they are included for completeness.

Let $f_n(t) = \frac{dF_n(t)}{dt}$ be the probability density function of the sojourn time. Rearranging the above and letting $h \to 0$,

$$f_n(t) = -(\lambda + \mu)\, F_n(t) + \lambda\, F_{n+1}(t) + \mu\, F_{n-1}(t) \tag{1}$$

for $n > 0$. The monitored task is in service when $n = 0$. However, the arrival of a task of higher priority will cause the task to be preempted. Thus,

$$F_0(t+h) = (1 - (\lambda + \lambda' + \mu)h)\, F_0(t) + \lambda h\, F_1(t) + \lambda' h\, F_0(t) + \mu h$$

from which we obtain

$$f_0(t) = -(\lambda + \mu)\, F_0(t) + \lambda\, F_1(t) + \mu \tag{2}$$

Let $L_n(t)$ be the Laplace transform of $f_n(t)$,

$$L_n(s) = \int_{t=0}^{\infty} e^{-st} f(t)\, dt$$

Now, using integration by parts, observe that, provided $f(0) = 0$,

$$\begin{aligned} L_n(s) &= \int_{t=0}^{\infty} e^{-st} f(t)\, dt \\ &= \left[e^{-st} \int f(t)\, dt \right]_0^{\infty} + s \int_0^{\infty} \left(\int f(t)\, dt \right) e^{-st} dt \\ &= sL'_n(s) \end{aligned}$$

where $L'_n(s)$ is the Laplace transform of $F_n(t)$. Thus, from equation (1),

$$(s+\lambda+\mu)L_n(s) = \lambda\, L_{n+1}(s) + \mu\, L_{n-1}(s) \tag{3}$$

for $n > 0$. When $n = 0$ we obtain similarly from equation (2),

$$(s+\lambda+\mu)L_0(s) = \lambda\, L_1(s) + \mu \tag{4}$$

Now define the generating function

$$G(z) = \sum_{n=0}^{\infty} z^n\, L_n(s)$$

By multiplying equation (3) by z^n and summing over $n > 0$,

$$\begin{aligned}(s+\lambda+\mu)\left(\sum_{n=1}^{\infty} z^n\, L_n(s)\right) &= \lambda z^{-1}\left(\sum_{n=1}^{\infty} z^{n+1}\, L_{n+1}(s)\right) \\ &+ \mu z\left(\sum_{n=1}^{\infty} z^{n-1}\, L_{n-1}(s)\right)\end{aligned}$$

i.e.

$$(s+\lambda+\mu)[\, G(z) - L_0(s)\,] = \lambda z^{-1}[\, G(z) - L_0(s) - z\, L_1(s)\,] + \mu z\, G(z)$$

Rearranging, we obtain

$$G(z) = \frac{(-(s+\lambda+\mu)\, z + \lambda)\, L_0(s) + \lambda z\, L_1(s)}{\mu z^2 - (s+\lambda+\mu)\, z + \lambda} \tag{5}$$

There are two unknowns: $L_0(s)$ and $L_1(s)$ and we have already one equation (equation (4)) relating the two. A second equation can be derived using the analyticity of $G(z)$ inside the unit disk. The denominator above vanishes at $z = r$ and $z = r'$ where r, r' are the roots of the quadratic equation $Q(x) = 0$, with

$$Q(x) = \mu x^2 - (s+\lambda+\mu)\, x + \lambda$$

For $s > 0$ the smaller root (r say) satisfies $0 \le r < 1$ since $Q(x) \ge 0$ for all $x < 0$ and since $Q(1) = -s$. A condition on the analyticity of $G(z)$ is that the numerator in equation (5) must therefore vanish similarly when $x = r$. In this case,

$$(s+\lambda+\mu)r = \mu r^2 + \lambda$$

The numerator then becomes

$$\begin{aligned}(-(\mu r^2+\lambda)+\lambda)\, L_0(s) + \lambda\, r\, L_1(s) &= \lambda r\, L_1(s) - \mu r^2\, L_0(s) \\ &= 0\end{aligned}$$

Thus, $L_1(s) = \lambda^{-1}\mu r\, L_0(s)$. Substituting into equation (4) above, we obtain

$$(s + \lambda + \mu)\, L_0(s) \quad = \quad \mu r\, L_0(s) + \mu$$

whereupon

$$L_0(s) = \frac{\mu}{s + \lambda + \mu(1 - r)}$$

$L_1(s)$ then follows.

Now let M be the random variable denoting the number of tasks ahead of a class c task at its arrival instant in the queue. Specifically, M is the total population of queues of class $c' \leq c$ at the arrival instant. Note that this includes tasks of class c that are already awaiting service. Let $p_m = P(M = m)$ be the equilibrium probability that $M = m$.

The Laplace transform of the sojourn time density of tasks of priority class c can now be written

$$L(s) = \sum_{n=0}^{\infty} p_n\, L_n(s)$$

where $L_n(s)$ is as above. Because the service discipline is preemptive, an arriving task of class $c' \leq c$ is guaranteed to be serviced before tasks of class $c'' > c$, even if there is a task of class c'' in service at its arrival instant. Thus, from the random observer property, M has the same distribution as that of the queue length in an M/M/1 queue whose arrival rate is $\lambda^+ = \lambda + \lambda_c = \sum_{i=1}^{c} \lambda_i$ and whose service rate is μ. Crucially, in the preemptive case, all tasks of priority class $c'' > c$ can be ignored.

Defining $\rho = \lambda^+/\mu$ we thus have $p_n = \rho^n(1 - \rho)$ and we obtain for $L(s)$,

$$\begin{aligned} L(s) &= \sum_{i=0}^{\infty} \rho^n (1 - \rho)\, L_n(s) \\ &= (1 - \rho)\, G(\rho) \\ &= (1 - \rho)\, \frac{(-(s + \lambda + \mu)\, \rho + \lambda)\, L_0(s) + \lambda\, \rho\, L_1(s)}{\mu\rho^2 - (s + \lambda + \mu)\, \rho + \lambda} \end{aligned}$$

where

$$\begin{aligned} L_0(s) &= \frac{\mu}{s + \lambda + \mu(1 - r)} \\ L_1(s) &= \frac{\mu^2 r}{\lambda(s + \lambda + \mu(1 - r))} \\ r &= \frac{s + \mu + \lambda - \sqrt{(s + \mu + \lambda)^2 - 4\mu\lambda}}{2\mu} \end{aligned}$$

Note that the (unconditional) Laplace transform of the cumulative distribution function of the sojourn time for a task of class c is given by $L(s)/s$.

3 Results

The Laplace transform $L(s)$ can be inverted using standard numerical algorithms. Figure 1 shows the sojourn time densities for tasks of classes 2, 4 and 6, in a priority queue with six priority classes, where class i has arrival rate $\lambda_i = i + 1$ and where the common service rate is $\mu = 30$ for all tasks.

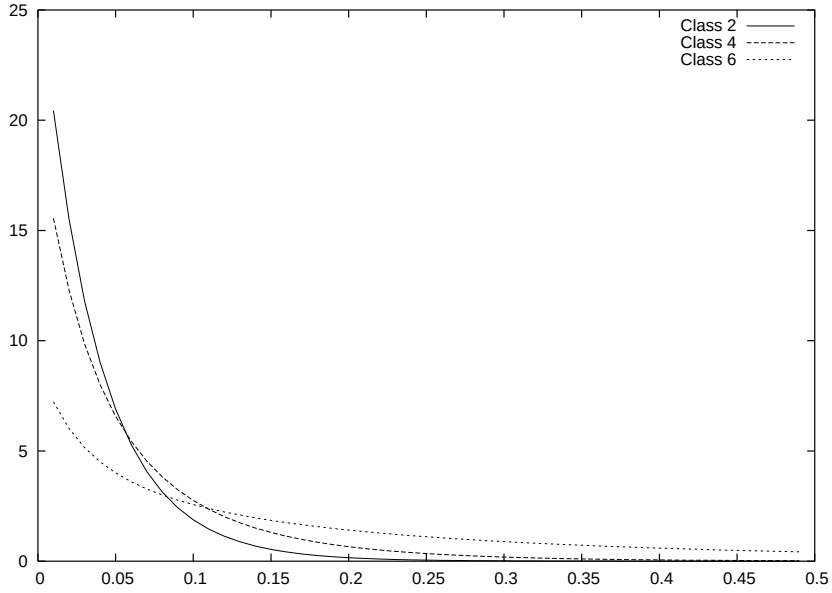

Figure 1: Sample densitites from a priority queue with six size classes