A SINGLE SERVER PRIORITY QUEUE WITH SERVER FAILURES AND QUEUE FLUSHING

Don Towsley and Satish K. Tripathi

COINS Technical Report 89-16

February 27, 1989

A Single Server Priority Queue with Server Failures and Queue Flushing

Don Towsley[*]and Satish K. Tripathi [†]

February 27, 1989

Abstract

We consider a single server queue serving two classes of customers according to a preeemptive resume head of the line priority discipline. The server is prone to failures and at the time that they occur, all customers are flushed out of the system. The system is analyzed under the assumption of a bulk arrival Poisson arrival processe, exponential service times, general repair times and exponential interfailure times.

1 Introduction

In this paper we study the behavior of a single server that serves two classes of customers. Each customer has a priority assigned to it according to its class and a preeemptive resume discipline is used to serve all customers. If we label the classes $k = 1, 2$, then we assume that class 1 customers have higher priority. Customers receive an amount of service that is exponentially distributed with a mean that depends on the customer class. The server is subject to failures. At the time of a failure, all customers in the system are flushed out. The time between failures is assumed to be an exponential r.v. and the repair time is allowed to have a general distribution. Last, customers arrive according to a compound Poisson process whose parameters depend on whether the server is up or down.

[*] Department of Computer & Information Science, Univ. Massachusetts, Amherst, MA 01003. The work of this author was supported by ONR under contract N00014-87-K-0796.

[†] Department of Computer Science, University of Maryland, College Park, MD 20742

This work differs from earlier work in two ways. First, the typical models of systems where servers suffer failures usually assume that customers remain in the system during the failures [4,1]. The only exception known to the authors is a found in a report by Finkel and Woodside [3]. That report considers a special case of our model when there is only one customer class, no bulk arrivals, and negligible repair times. Second, most of the work on priority systems assumes that the arrival streams for the different priorities are independent of each other. Consequently our work breaks new ground in this direction.

The paper is organized in the following manner. Section 2 formulates the problem when there is only one customer class. A complete solution of this problem is included in that section. Section 3 uses these results to solve the original two priority class problem. We conclude the paper with two applications in Section 4 and a summary of the results in Section 5.

2 Flush out model

In this section we first consider the case where there is only one class of customers and where all customers are served in a first come first serve manner. We assume that jobs arrive to this queue in batches of size D^u and D^d according to Poisson processes with parameter λ^u and λ^d during the times that the server is up and down respectively. We further assume that the batch sizes form a sequence of independent random variables (r.v.'s) with distributions $\alpha_k^u = P[D^u = k]$, $\alpha_k^d = P[D^d = k]$, $k = 1, \cdots$ and probability generating functions (p.g.f.s) $D^u(z) = E[z^{D^u}]$, $D^d(z) = E[z^{D^d}]$. Job service times are assumed to be independent and identically distributed exponential random variables with mean $1/\mu$. Last, the server fails according to a Poisson process with parameter γ. When a failure occurs all jobs are lost from the system. The failure is of duration R with cumulative distribution $F_R(t) = P[R \leq t]$, Laplace transform $R(s) = E[e^{-sR}]$ and mean $1/\beta$. We introduce Y to denote the number of arrivals during a failure period. It has distribution $\alpha_i^r = P[Y = i]$ with (p.g.f.) $Y(z) = E[z^Y] = R(\lambda^d(1 - D^d(z)))$. A special case of this problem has been studied by Finkel and Woodside [3] where jobs are assumed to arrive singly and failure intervals are of zero length.

Let U denote the state of the server, $U = 0$ if the server is down, and $U = 1$ otherwise.

Because the combined failure repair process is an alternating renewal process, we can write

$$\Pr[U = i] = \begin{cases} \gamma/(\gamma+\beta), & i = 0, \\ \beta/(\gamma+\beta), & i = 1. \end{cases} \tag{1}$$

We are interested in the statistics of N, the number of customers in the system. We will obtain the p.g.f. for the distribution of N, $N(z) = E[z^N]$ when the system is in equilibrium. This is most easily done by conditioning on the value of U. In the case that the server is down, we have

$$E[z^N | U = 0] = [1 - R(\lambda_d(1 - D^d(z)))]\beta. \tag{2}$$

In order to determine the behavior of the system when the server is operational, we study the behavior of a semi-Markov process (SMP) imbedded at points in time immediately after arrivals and service completions that occur while the server is up, and after each failure and repair. In order to simplify the analysis, we shall assume that completion of fictitious customers can occur while the server is up and the queue is empty. The behavior of this system is modeled as a Markov chain with state (L_t) where L_t denotes the number of jobs in the system after the t-th event, $t = 0, 1, \cdots$, when the server is operational and takes value $0'$ when the server is down. We are interested in the stationary behavior of L_t which we denote as $L = \lim_{t\to\infty} L_t$. Let $p_i = P[L = i]$, $i = 0', 0, 1, \cdots$ denote the stationary distribution of L. When the system is ergodic, these probabilities satisfy

$$p_{0'} = \gamma(1 - p_{0'})/\sigma, \tag{3}$$

$$p_0 = \mu(p_0 + p_1)/\sigma + p_{0'}\alpha'_0, \tag{4}$$

$$p_i = \mu p_{i+1}/\sigma + \lambda^u/\sigma \sum_{k=0}^{i} p_k \alpha_{i-k} + p_{0'}\alpha'_i, \quad i = 1, \cdots \tag{5}$$

where $\sigma = \lambda^u + \mu + \gamma$.

If we let $L^u(z)$ denote $\sum_{i=0}^{\infty} p_i z^i$, then multiplying the left and right hand sides of eq. (5) by z^i and summing over $i = 1, \cdots$ yields

$$L^u(z) = \frac{(L^u(z) - p_0)\mu}{z\sigma} + \lambda^u L^u(z) D^u(z)/\sigma + p_{0'}Y(z) + p_0\mu/\sigma \tag{6}$$

which can be manipulated to obtain

$$L^u(z) = \frac{p_0(z-1)\mu + p_{0'}Y(z)z\sigma}{\sigma z - \mu - \lambda^u D^u(z) z}. \tag{7}$$

Equation (3) can be used to obtain the following expression for $p_{0'}$

$$p_{0'} = \frac{\gamma}{\sigma + \gamma}. \tag{8}$$

¿From the theory of SMPs [5] we know that the joint probability that $N = i$ and $U = 1$ is

$$\Pr[N = i, U = 1] = \frac{p_i/\sigma}{p_{0'}/\beta + \sum_{l=0}^{\infty} p_l/\sigma}, \quad i = 0', 0, 1, \cdots \tag{9}$$

We define a new partial generating function $N^u(z) = \sum_{i=0}^{\infty} \Pr[N = i, U = 1] z^i$. Standard techniques yield

$$N^u(z) = \frac{\gamma Y(z) z \beta/(\beta + \gamma) + \Pr[N = 0, U = 1]\mu(z-1)}{\sigma z - \mu - \lambda^u D^u(z) z} \tag{10}$$

and we can express

$$\begin{aligned} E[z^N | U = 1] &= N^u(z)/\Pr[U = 1], \\ &= \frac{\gamma Y(z) z + \Pr[N = 0, U = 1]\mu(z-1)(\gamma+\beta)/\beta}{\sigma z - \mu - \lambda^u D^u(z) z}. \end{aligned} \tag{11}$$

The probability $\Pr[N = 0, U = 1]$ is determined by recognizing that $L^u(z)$ is an analytic function within $|z| < 1$. Consequently, the numerator takes on value 0 whenever the denominator does in that region. We now show that the denominator has exactly one real root within $[0, 1)$ whenever $\gamma > 0$.

Let $A(z) = \sigma z - \mu - \lambda^u D^u(z) z$. It is easy to show that $A(0) = -\mu$, $A(1) = \gamma$ and that $d^2 A(z)/dz^2 \leq 0$ in the region $0 \leq z < 1$. Consequently, there exists one real root within $[0, 1)$.

In the case that jobs arrive singly, the root, z^*, is

$$z^* = \frac{(\lambda^u + \gamma + \mu) - \sqrt{(\lambda^u + \gamma + \mu)^2 - 4\lambda^u \mu}}{2\lambda^u}. \tag{12}$$

Finally,

$$\Pr[N = 0, U = 1] = \frac{\gamma Y(z^*) z^* \beta}{(1 - z^*)\mu(\gamma + \beta)}. \tag{13}$$

We can now express the p.g.f. of the distribution for N as

$$N(z) = \Pr[U = 0] E[z^N | U = 0] + \Pr[U = 1] E[z^N | U = 1],$$

$$= \quad [1 - R(\lambda^d(1 - D^d(z)))]\beta\gamma/(\gamma + \beta)$$

$$+ \frac{\gamma\beta[zY(z)/(\gamma + \beta) + (z - 1)Y(z^*)z^*/(1 - z^*)]}{(\gamma + \beta)[\sigma z - \mu - \lambda^u D^u(z)z]}. \tag{14}$$

with mean

$$\begin{aligned} E[N] &= \frac{\lambda^d\gamma E[R^2]E[D^f]}{2(\gamma + \beta)E[R]} \\ &+ \frac{Y(z^*)z^*\beta}{\gamma\mu(\gamma + \beta)(1 - z^*)} + \frac{\beta}{\gamma(\gamma + \beta)}[\gamma + \lambda^u + \gamma E[Y] + \beta E[D^u] - \sigma]. \end{aligned} \tag{15}$$

Three other performance measures of interest to us are the system throughput, the probability that a customer is lost, and the distribution for the number of customers that are flushed out of the system at the time of a failure. The system throughput, η, is

$$\eta = (\beta/(\gamma + \beta) - Pr[N = 0, U = 1])\mu \tag{16}$$

and the probability of customer loss, q, is

$$q = 1 - \frac{\eta(\beta + \gamma)}{\beta\lambda^u E[D^u] + \gamma\lambda^d E[D^d]} \tag{17}$$

Last, if C denotes the number of customers lost at the time of a failure, then it has p.g.f. $C(z)$ given by

$$C(z) = E[z^N | U = 1] \tag{18}$$

with average

$$E[C] = \frac{Y(z^*)z^*}{\gamma\mu(1 - z^*)} + \frac{\gamma + \lambda^u + \gamma E[Y] + \beta E[D^u] - \sigma}{\gamma} \tag{19}$$

We conclude this section with the observation that the case $\gamma = 0$ corresponds to a bulk arrival $M^D/M/1$ queue for which the p.g.f. of the queue length distribution is (see [7] for details)

$$N(z) = \frac{\mu(z - 1)(1 - \lambda^u E[D^u]/\mu)}{\lambda^u z(1 - D^u(z)) + \mu(z - 1)}.$$

3 Flush out model with priorities

Consider a system with two classes of customers, $i = 1, 2$, where class 1 customers receive preemptive priority over class 2 customers. During the time that the server is up, customers arrive in bulks where D_i^u denotes the number of customers of class i contained within the bulk. Let $\alpha_{i,j}^u = P[D_1^u = i, D_2^u = j]$, $i,j = 0,1,\cdots$ be the joint probability distribution of the bulk sizes and let $D^u(w,z) = E[w^{D_1^u} z^{D_2^u}]$ denote the p.g.f. of this distribution. We assume that the arrival of these bulks is described by a time invariant Poisson process with parameter λ^u. A similar process operates during the down periods except that the arrival rate is λ^d and the bulk sizes are D_i^d having distribution $\alpha_{i,j}^d$ and p.g.f. $D^d(w,z) = E[w^{D_1^d} z^{D_2^d}]$. Service times for class i customers are exponentially distributed with mean $1/\mu_i$, i=1,2. Last, failures are generated by a Poisson process with rate γ. When a failure occurs, all jobs are lost from the system and the failure lasts for R units of time with distribution $F_R(t) = P[R \leq t]$ and Laplace transform $R(s)$. We let Y_i denote the number of customers of class i ($i = 1,2$) that arrive by the end of a failure period. These r.v.'s have distribution $\alpha_{i,j}^f = \Pr[Y_1 = i, Y_2 = j]$, $0 \leq i,j$ and p.g.f. $Y(w,z) = R(\lambda^d(1 - D^d(w,z)))$. We again define the r.v. U to denote whether the server is up ($U = 1$) or down ($U = 0$) with probabilities given in equation (1).

We treat the case $\gamma > 0$ first in great detail. We then conclude the section with the main results for the case $\gamma = 0$.

We are interested in the statistics of N_1 and N_2, the number of customers of the high and low priority classes respectively present in the system. As in section 2, we obtain the p.g.f. of the joint distribution by conditioning on the value of U. In the case that $U = 0$, this conditional p.g.f. is

$$E[w^{N_1} z^{N_2} | U = 0] = [1 - R(\lambda^d(1 - D^d(w,z)))]\beta. \tag{20}$$

In order to obtain the p.g.f. given that the server is up, we again study the Markov process imbedded at the points in time immediately after events (except arrivals during the failure period). The state of this system is $(L_{1,i}, L_{2,i}, L_{3,i})$ after the i-th event where $L_{1,i}$ and $L_{2,i}$ denote the number of each priority class and $L_{3,i}$ denotes whether the server is up or not, $L_{3,i} = 0$ if down and $L_{3,i} = 1$ if up. Note that whenever $L_{3,i} = 0$, then $L_{1,i} = L_{2,i} = 0$. We are interested in the stationary behavior of $(L_{1,i}, L_{2,i}, L_{3,i})$ which we denote as $(L_1, L_2, L_3) = \lim_{i\to\infty}(L_{1,i}, L_{2,i}, L_{3,i})$. This system is ergodic for non-negative values of λ_i^u and μ_i, $i = 1,2$ whenever $\gamma > 0$. Let $p_{i,j} = P[L_1 = i, L_2 = j, L_3 = 1], i,j = 0,1,\cdots$ and

$p_{0'} = P[L_3 = 0]$. When the system is ergodic, the stationary probabilities satisfy

$$p_{0'} = \gamma(1 - p_{0'})/\sigma, \tag{21}$$

$$p_{0,0} = [\mu_1(p_{0,0} + p_{1,0}) + \mu_2(p_{0,0} + p_{0,1})]/\sigma + \alpha'_{0,0}p_{0'}, \tag{22}$$

$$p_{0,j} = [\mu_2 p_{0,j+1} + \mu_1(p_{0,j} + p_{1,j}) + \lambda^u \sum_{l=0}^{j} \alpha^u_{0,l} p_{0,j-l}]/\sigma + \alpha'_{0,j}p_{0'}, \quad j \geq 1 \tag{23}$$

$$p_{i,j} = [\lambda^u \sum_{l_2=0}^{j} \sum_{l_1=0}^{i} \alpha^u_{l_1,l_2} p_{i-l_1,j-l_2} + \mu_1 p_{i+1,j} + \mu_2 p_{i,j}]/\sigma + \alpha'_{i,j}p_{0'}, \quad 0 < i; 0 \leq j \tag{24}$$

where $\sigma = \lambda^u + \mu_1 + \mu_2 + \gamma$.

We define the following partial generating functions $L_i^u(z) = \sum_{j=0}^{\infty} p_{i,j}z^j$ and $L^u(w,z) = \sum_{i=0}^{\infty} L_i(z)w^i$. Multiplying both sides of equation (23) by z^j, summing over j, adding in equation (22), and solving for $L_0^u(z)$ yields

$$L_0^u(z)[(\sigma - \mu_1)z - \mu_2 - \lambda^u D_0^u(z)z] = z\mu_1 L_1^u(z) + p_{0,0}\mu_2(z-1) + zY_0(z)\sigma p_{0'}. \tag{25}$$

Multiplying both sides of equation (24) by z^j and summing over j yields

$$(\sigma - \mu_2)L_i^u(z) = \mu_1 L_{i+1}^u(z) + \lambda^u \sum_{l=0}^{i} D_l^u(z)L_{i-l}^u(z) + \sigma p_{0'}Y_i(z) \quad i > 0 \tag{26}$$

Here $D_l^u(z) = E[z^{D_2^u} | D_1^u = l]$ and $Y_l(z) = E[z^{Y_2} | Y_1 = l]$, $l = 0, 1, \cdots$. Finally, multiplying both sides of equations (25) and (26) by w^i, summing and performing some algebra yields

$$L^u(w,z) = \frac{L_0^u(z)[zw(\mu_1 - \mu_2) + w\mu_2 - z\mu_1] + p_{0,0}w(z-1)\mu_2 + zwY(w,z)p_{0'}\sigma}{[zw(\sigma - \mu_2) - z\mu_1 - \lambda^u zwD^u(w,z)]}. \tag{27}$$

Equation (21) can be used to obtain $p_{0'} = \gamma/(\sigma + \gamma)$.

We define $N(w,z) = E[w^{N_1}z^{N_2}]$. Following the same line of reasoning displayed in section 2, we are able to obtain the following expression for $N(w,z)$,

$$N(w,z) = \frac{\gamma\beta(1 - R(\lambda^d(1 - D^d(w,z))))}{\gamma + \beta} + \frac{[zw(\mu_1 - \mu_2) + \mu_2 w - \mu_1 z]N_0^u(z)}{zw(\sigma - \mu_2) - z\mu_1 - \lambda^u zwD^u(w,z)}$$
$$+ \frac{zw\gamma\beta Y(w,z)/(\gamma + \beta) + \Pr[N_1 = 0, N_2 = 0, U = 1]w(z-1)\mu_2)}{zw(\sigma - \mu_2) - z\mu_1 - \lambda^u zwD^u(w,z)} \tag{28}$$

where $N_0^u(z) = \sum_{i=0}^{\infty} \Pr[N_1 = 0, N_2 = i, U = 1]z^i$. We are left with the task of obtaining expressions for $\Pr[N_1 = 0, N_2 = 0, U = 1]$ and $N_0^u(z)$.

We focus on the derivation of an expression for the conditional p.g.f., $E[z^{N_2}|U = 1, N_1 = 0]$. Once we obtain an expression for this quantity we can use the relation

$$N_0^u(z) = E[z^{N_2}|U = 1, N_1 = 0]\Pr[U = 1, N_1 = 0] \tag{29}$$

to obtain $N_0^u(z)$.

The probability $\Pr[N_1 = 0, U = 1]$ is obtained by recognizing that the high priority customers do not observe low priority customers. Thus we can use the results from the preceding section to obtain this probability,

$$\Pr[N_1 = 0, U = 1] = \gamma z_1^*(1 - z_1^*)\mu_1, \tag{30}$$

where z_1^* is the root of the function $(\lambda^u + \gamma + \mu_1)z - \mu_1 - \lambda^u D^u(z, 1)z$ lying within $[0, 1)$.

An expression for $E[z^{N_2}|U = 1, N_1 = 0]$ is obtained by studying a modified system where all class 1 busy periods are removed and replaced by bulk arrivals of the class 2 customers that arrive during those busy periods. Figure 1 illustrates this modification. Figure 1a illustrates the original system and Figure 1b illustrates the modified system with class 1 busy periods replaced by arrivals. The numbers associated with the arrivals identify the classes of the arriving customers. This transformation results in a single class system whose behavior was analyzed in the preceding section. We shall use a ' to denote the parameters of the model of the modified system, e.g., γ' is the failure rate for the modified system.

Not all of the class 1 busy periods are replaced by bulk arrivals in the modified system. This replacement occurs only if the busy period completes before the server fails. Let us consider what happens to those class 1 busy periods that do not complete before the server fails.

We distinguish between two types of class 1 busy periods. Type-1 busy periods are those initiated while the server is up due to the arrival of a class 1 customer and type-2 busy periods are those initiated by class 1 customers that arrived during a failure period. Let P_f^1 and P_f^2 denote the probabilities that these busy periods do not complete before the server fails. We focus on the first of these probabilities. Let B denote the busy period of a bulk arrival $M^{[D_2]}/M/1$ queue without failures that serves the high priority jobs. Denote the probability density function and Laplace transform for B as $b(t)$ and $B(s)$ respectively.

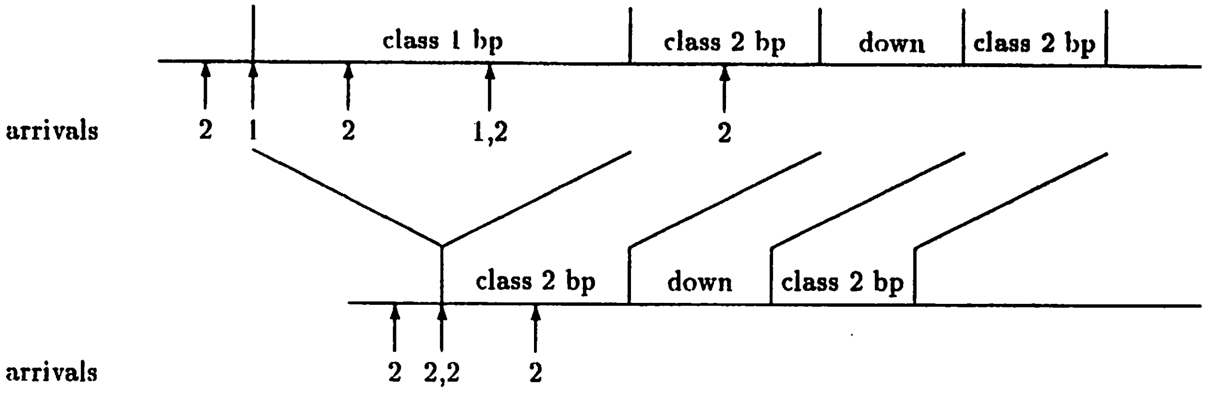

Figure 1: Original System and Modified System.

Then P_f^1 can be expressed as

$$\begin{aligned} P_f^1 &= 1 - \int_0^\infty b(t)\, e^{-\gamma t} dt \\ &= 1 - B(\gamma). \end{aligned} \tag{31}$$

In a similar manner we derive the following expression for P_f^2,

$$P_f^2 = 1 - R(\lambda^d(1 - B(\gamma))) \tag{32}$$

Each failure that occurs in the original system while the system 1) is empty, 2) contains only class 2 jobs, or 3) is in a type-1 busy period, *is represented as a distinct failure in the modified system.* Each failure that occurs during a type-2 busy period is merged with the failure period that preceded it. Thus the failure rate in the modified system is $\gamma' = \gamma + \lambda^u(1 - D^u(0,1))P_f^1$ and the average failure period duration is $1/\beta' = (1/\beta)/(1 - P_f^2)$.

We now focus on the bulk size of class 2 customers in the modified system. Specifically, we derive expressions for $D^{u\prime}(z)$ and $Y'(z)$. This analysis requires that we study the class 1 busy periods in the original system that do not contain failures. These can be shown to exhibit the same statistical behavior as the busy periods of a $M^{[D_1]}/M/1$ queue with arrival rate λ^u, bulk size D_1^u, and service rate $\mu_1 + \gamma$. $D^{u'}$ corresponds to the number of class 2 customers that either arrive alone or during a class 1 busy period. It can be shown that the p.g.f for the distribution of $D^{u'}$ satisifies the following functional equation,

$$D^{u'}(z) = D^u((\mu_1 + \gamma)/[\mu_1 + \gamma + \lambda^u - \lambda^u D^u(D^{u'}(z), z)], z) \tag{33}$$

and the p.g.f. for the number of arrivals at the end of a failure period in the modified system is

$$Y'(z) = R(\lambda^d(1 - D^d(W(z), z))). \tag{34}$$

The remaining parameter is $\mu' = \mu_2$.

The conditional pgf $E[z^{N_2}|N_1 = 0, U = 1]$ is obtained from equation 11 by replacing the unprimed parameters with the primed parameters derived above and then solving for p_0' as described in the text following that equation. The resulting expression obtained for $E[z^{N_2}|N_1 = 0, U = 1]$ is

$$E[z^{N_2}|N_1 = 0, U = 1] = \frac{\Pr[N_2 = 0, U = 1|N_1 = 0]\mu_2(z - 1)(\gamma' + \beta')/\beta' + z\gamma' Y'(z)}{(\lambda^u + \mu_2 + \gamma)z - \mu_2 - \lambda D^{u'}(z)z} \tag{35}$$

where

$$\Pr[N_2 = 0, U = 1|N_1 = 0] = \frac{\gamma' + z_2^* \beta' Y'(z_2^*)}{(1 - z_2^*)\mu_2(\gamma' + \beta')} \tag{36}$$

and z_2^* is the root of the denominator of the expression in equation (35).

We are now left with the task of obtaining $\Pr[N_1 = 0, N_2 = 0, U = 1]$. This probability is

$$\begin{aligned} \Pr[N_1 = 0, N_2 = 0, U = 1] &= \Pr[N_2 = 0|N_1 = 0, U = 1] \ \Pr[N_1 = 0, U = 1], \\ &= E[z^{N_2}|N_1 = 0, U = 1]|_{z=0} \ \Pr[N_1 = 0, U = 1], \\ &= (\gamma' + \beta')\Pr[N_2 = 0, U = 1|N_1 = 0] \ \Pr[N_1 = 0, U = 1]/\beta' \\ &= \frac{\gamma z_1^*(1 - z_1^*)\mu_1(\gamma' + z_2^* \beta' Y'(z_2^*))}{\beta'(1 - z_2^*)\mu_2}. \qquad (37) \end{aligned}$$

The last two probabilities are given by equations (30) and (36).

We denote the throughputs of each class as η_i, $i = 1, 2$. they are

$$\eta_1 = (\beta/(\beta + \gamma) - \Pr[N_1 = 0, U = 1])\mu_1, \tag{38}$$

$$\eta_2 = (\Pr[N_1 = 0, U = 1] - \Pr[N_1 = 0, N_2 = 0, U = 1])\mu_2. \tag{39}$$

The probability of customer loss, q_i, for class $i = 1,2$ is

$$q_i = 1 - \frac{\eta_i(\beta + \gamma)}{\beta\lambda^u E[D_i^u] + \gamma\lambda^d E[D_i^d]}, \quad i = 1,2. \tag{40}$$

Last, if C_i denotes the number of customers lost of each customer class when the server fails, then the joint distribution has p.g.f. $C(w,z) = E[w^{C_1} z^{C_2}]$ given by

$$C(w,z) = E[w^{N_1} z^{N_2} | U = 1]. \tag{41}$$

The moment generating properties of the p.g.f. can be used to obtain the moments of these r.v.s.

We conclude this section with a description of the results for the case $\gamma = 0$. The system is modeled as a continuous time Markov chain with state $(L_{1,t}, L_{2,t})$ at time $t \geq 0$ where $L_{1,t}$ and $L_{2,t}$ denote the number of each priority class in the system. We are interested in the stationary behavior of $(L_{1,t}, L_{2,t})$ which we denote as $(L_1, L_2) = \lim_{t\to\infty}(L_{1,t}, L_{2,t})$. SInce there are no down periods, we will omit the superscript ^{u} from λ^u, D_i^u, etc... This system is ergodic for non-negative values of λ, $E[D_i]$, and μ_i, $i = 1,2$ such that $\lambda \sum_{i=1}^2 E[D_i]\mu_i < 1$. Let $p_{i,j} = P[L_1 = i, L_2 = j]$, $i,j = 0,1,\cdots$. When the system is ergodic, the stationary probabilities satisfy

$$\lambda p_{0,0} = \mu_1(p_{0,0} + p_{1,0}) + \mu_2(p_{0,0} + p_{0,1}), \tag{42}$$

$$(\lambda + \mu_2)p_{0,j} = \mu_2 p_{0,j+1} + \mu_1(p_{0,j} + p_{1,j}) + \lambda \sum_{l=0}^{j} \alpha_{0,l} p_{0,j-l}, \quad i \geq 1 \tag{43}$$

$$(\lambda + \mu_1)p_{i,j} = \lambda \sum_{l_2=0}^{j} \sum_{l_1=0}^{i} \alpha_{l_1,l_2} p_{i-l_1,j-l_2} + \mu_1 p_{i+1,j} + \mu_2 p_{i,j}, \quad 0 < i; 0 \leq j. \tag{44}$$

If we define $N(w,z) = E[w^{L_1} z^{L_2}]$ and $N_0(z) = \sum_{i=0}^{\infty} p_{0,i} z^i$, then standard techniques yield

$$N(w,z) = \frac{[zw(\mu_1 - \mu_2) + \mu_2 w - \mu_1 z]N_0(z) + p_{0,0}w(z-1)\mu_2)}{zw(\lambda + \mu_1) - z\mu_1 - \lambda zw D(w,z)}. \tag{45}$$

Following the same procedure used earlier for the case $\gamma > 0$ we obtain the following expression for $N_0(z)$,

$$N_0(z) = \frac{\mu_2(z-1)(1 - \lambda E[D']/\mu_2)}{(\lambda + \mu_2)z - \mu_2 - \lambda D'(z)z} \tag{46}$$

where

$$D'(z) = D(\mu_1/[\mu_1 + \lambda - \lambda D(D'(z), z)], z), \tag{47}$$

and $E[D'] = (E[D_1]\mu_2/(\mu_2 - \lambda E[D_2])$.

The moment generating properties of the p.g.f. can be used to obtain various statistics of L_1 and L_2.

4 Applications

We describe two applications of our model to the evaluation of performance of distributed computer systems. The first application is to the problem of evaluating the performance of a processor which handles requests that arrive over the network as well as requests that originate at that processor. The second application is to the problem of evaluating a ***buddy*** protocol for fault recovery in a distributed system.

4.1 Remote Execution of Requests in a Distributed System

We consider a distributed system consisiting of M processors connected by a network. We assume that jobs arrive at each node either from the outside world or from some other processor. In either case, the arrival process is assumed to be Poisson with rate λ_1 for the external arrivals and λ_2 for the network arrivals. All jobs require an amount of service that is exponentially distributed with rate μ_2. Last the processor is required in order to handle the communications required to transfer the remote job over the network. In this case the service requirements are assumed to be exponential random variables with rate μ_1. Last, communications is given preemptive priority over job execution.

We have described a two priority queue in which the arrivals of the two job classes are correlated. Thus the results found at the end of the last section apply here. Our parameters are $\lambda = \lambda_1 + \lambda_2$, $D(w, z) = (\lambda_1 + \lambda_2 w)z/\lambda$, μ_1 and μ_2. The p.g.f. for the joint queue length distribution is given by equation (45) where $N_0(z)$ is given in equation (46) and $D'(z)$ is

$$D'(z) = \frac{\mu_1 + \lambda + \lambda_2^2 z/\lambda}{2\lambda_1}\left[1 - \left(1 - \frac{4z\lambda_1[\lambda_1\mu_1 + \lambda_2 z(\mu_1 + \lambda - \lambda_2 z)]}{\lambda[\lambda_2^2 z/\lambda - \mu_1 - \lambda]^2}\right)^{1/2}\right].$$

The moment generating properies of the p.g.f. can be used to obtain the moments of the queue lengths.

4.2 The Buddy Algorithm

In this section we apply the model to evaluate a simple buddy algorithm that can be applied to a distributed computer system so as to recover from faults. The algorithm operates in the following manner. Whenever a job arrives to a processor, a second processor is selected to store a copy of the job. If the first processor fails before the job completes, then it is the responsibility of the second processor (*buddy*) to execute the job. At the time that a processor fails, it relinquishes control of all jobs within its queue and they are flushed out. Thus the model developed in this paper can be used to model the behavior of the processor.

The processor must perform two activities for each job, select a buddy node and transfer the job to it, and execute the job. The first activity has higher priority than the second activity. Let jobs arrive according to a Poisson arrival process and require an exponential amount of service with mean $1/\mu_1$ to select a buddy and transfer the job, and an exponential amount of time with mean $1/\mu_2$ to execute the job. We assume the time between failures to be exponentially distributed with mean $1/\gamma$ and general service time R. Last, we assume that jobs are routed to a different processor when a processor is down.

The results from the previous section are directly applicable. The p.g.f. for the joint queue length distribution is

$$N(w,z) = \frac{[zw(\mu_1 - \mu_2) + \mu_2 w - \mu_1 z]N_0^u(z)}{zw(\lambda + \mu_1 + \gamma) - z\mu_1 - \lambda z^2 w^2} + \frac{zw\gamma\beta/(\gamma + \beta) + \Pr[N_1 = 0, N_2 = 0, U = 1]w(z-1)\mu_2}{zw(\lambda + \mu_1 + \gamma) - z\mu_1 - \lambda z^2 w^2} \tag{48}$$

where

$$N_0(z) = \frac{\gamma z_1^*(1 - z_1^*)\mu_1[(\gamma' + z_2^*\beta)(z-1) + z\gamma'\beta(1 - z_2^*)]}{\beta(1 - z_2^*)[(\lambda + \gamma + \mu_2)z - \mu_2 - \lambda D'(z)]} \tag{49}$$

$$\gamma' = \gamma + \lambda(1 - B(\gamma)),$$

$$B(s) = \frac{\mu_1 + \gamma + \lambda + s - [(\mu_1 + \gamma + \lambda + s)^2 - 4(\mu_1 + \gamma)\lambda]^{1/2}}{2\lambda},$$

$$D'(z) = -\frac{\mu_1 + \lambda + \gamma}{2\lambda}\left[1 - \left(1 - \frac{4\lambda z(\mu_1 + \gamma)}{(\mu_1 + \lambda + \gamma)^2}\right)^{1/2}\right],$$

$$z_1^* = \frac{(\lambda + \gamma + \mu_1) - \sqrt{(\lambda + \gamma + \mu_1)^2 - 4\lambda\mu_1}}{2\lambda},$$

$$\Pr[N_1 = 0, N_2 = 0, U = 1] = \frac{\gamma z_1^*(1 - z_1^*)\mu_1(\gamma' + z_2^*\beta)}{\beta(1 - z_2^*)\mu_2}.$$

Here z_2^* is the root of the denominator of equation (49) that lies within $[0, 1)$. The function $B(s)$ is the Laplace transform of the busy period distribution for an M/M/1 queue with arrival rate λ and service rate $\mu_1 + \gamma$. The moment generating properties of the p.g.f. allows us to obtain the moments of the queue lengths.

5 Summary

In this paper we obtained expressions for the p.g.f.'s of the marginal queue length distributions for each priority class in a single server queue where the server is subject to random failures. This system is interesting because 1) there is a correlation between the arrival processes of the two customer classes and 2) all customers disappear from the system at the time of a server failure. Furthermore, the arrival processes are turned off during the times required to repair the server. Further work is required to generalize these results to 1) more general service times, and 2) more than two priority classes.

References

[1] B. Avi-Itzhak and P. Naor, "Some Queueing Problems with the Service Sation Subject to Breakdowns," *Operations Research*, vol. 11, (1963), pp. 303-320.

[2] L. Kleinrock, *Queueing Systems*, Vol. 1, John Wiley, 1975.

[3] D. Finkel and C.M. Woodside, "A single server queue with server failures and queue flushing," Bucknell Univ. Computer Science Dept. Technical Report 88-4, May 1988.

[4] D.P. Gaver, "A Waiting Line with Interrupted Service, Including Priorities," *J. Royal Statistical Soc.*, Series B24, (1962), pp. 73-90.

[5] R.A. Howard, *Dynamic Probabilistic Systems, Vol. II2: Semi-Markov and Decision Processes*, John Wiley, New York, 1971.

[6] A.N. Kolmogorov and S.V. Fomin, *Introductory real analysis*, Dover, 1975.

[7] D.D. Yao. "Some results for the queues $M^X/M/c$ and $GI^X/G/c$", *Operations Research Letters*, vol. 4, (1985), pp. 79-83.