The Korean Journal of Applied Statistics (2009)
22(3), 541–551

Stochastic Upper Bound for the Stationary Queue Lengths of GPS Servers

Sunggon Kim[1]

[1]Department of Information Statistics, Gyeongsang National University

(Received February 2009; accepted March 2009)

Abstract

Generalized processor sharing(GPS) service policy is a scheduling algorithm to allocate the bandwidth of a queueing system with multi-class input traffic. In a queueing system with single-class traffic, the stationary queue length becomes larger stochastically when the bandwidth (*i.e.* the service rate) of the system decreases. For a given GPS server, we consider the similar problem to this. We define the monotonicity for the head of the line processor sharing(HLPS) servers in which the units in the heads of the queues are served simultaneously and the bandwidth allocated to each queue are determined by the numbers of units in the queues. GPS is a type of monotonic HLPS. We obtain the HLPS server whose queue length of a class stochastically bounds upper that of corresponding class in the given monotonic HLPS server for all classes. The queue lengths process of all classes in the obtained HLPS server has the stationary distribution of product form. When the given monotonic HLPS server is GPS server, we obtain the explicit form of the stationary queue lengths distribution of the bounding HLPS server. Numerical result shows how tight the stochastic bound is.

Keywords: Scheduling, generalized processor sharing, head of the line processor sharing, monotonicity.

1. Introduction

Generalized processor sharing(GPS) service policy is a scheduling algorithm for multi-class input traffic. Parekh and Gallager (1993) proposed the algorithm to allocate the bandwidth of the system to each class. The unit, which is located in the head of the queue of a class, is served at the rate of the bandwidth allocated to the class. The backlogged queues share the full bandwidth according to the preassigned ratios. Thus, the total bandwidth allocated to all of the queues is equal to the bandwidth of the system when there is at least one unit in the system, *i.e.* GPS is work-conserving.

When the GPS system has two classes with Poisson arrivals and the service requirement of a unit is exponentially distributed, the distribution of the queue length has been obtained by Cohen (1988) and Fayolle and Iasnogorodski (1979). However, for other cases, the previous studies by Adan *et al.* (2001), Borst and Zwart (2003), Borst *et al.* (2003) and Brandt and Brandt (1998) show that it

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2006-003-C00061).

[1]Associate Professor, Department of Information Statistics and RINS, Gyeongsang National University, Jinju 660-701, Republic of Korea. E-mail: sgkim@gnu.ac.kr

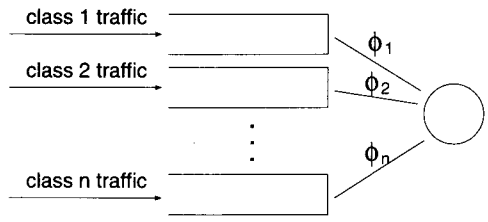


Figure 2.1. GPS scheduling

is difficult to analyze the performance mathematically. For the asymptotic behavior of the queue length distribution in large buffer regimes, there have been many studied including by de Veciana and Kesidis (1996), Zhang (1998), Dupuis and Ramanan (1998) and Bertsimas *et al.* (1999)

Bonald and Proutière (2004) have defined the monotonicity for the stochastic networks in which each node serves the units in its queue according to the processor sharing policy. For such networks, they proposed a method to find two balanced networks whose number of units in each node stochastically bound upper and lower, respectively, that of given monotonic network. In this paper, we consider the same problem for the head of the line processor sharing(HLPS) servers. HLPS is a scheduling algorithm for the server with multiple queues. Under HLPS, the units in the head of the queues are served simultaneously. The bandwidth allocated to each queue is determined by the numbers of units in the queues. Clearly, GPS is a type of HPLS. In GPS, the bandwidth allocated to each queue is determined by the backloggedness of the queues, in stead of the numbers of units in the queues. The HLPS server with n queues can be treated as a stochastic network with n nodes if we assume that each queue of the HLPS server is severed by n separated servers. However, the monotonicity defined by Bonald and Proutiére is not adaptable to the HLPS servers because the units in each queue are not served by the processor sharing policy. We modify their definition to propose the monotonicity for HLPS servers, and show that by the method of Bonald and Proutière (2004), it can be found the balanced HLPS server whose number of units in each queue stochastically bounds upper that of a given monotonic HPLS server.

For a given GPS server, we obtain the explicit stationary queue lengths distribution of the balanced HLPS server bounding upper the GPS server stochastically. We do some numerical studies to check how tight the stochastic bound is.

2. Generalized Processor Sharing

We consider a queueing system with GPS service policy as shown in Figure 2.1. The number of classes is denoted by n and each class has its own logical or physical queue. The full bandwidth of the system is given by C and ϕ_i is the service ratio of class i, for $i = 1, 2, \ldots, n$, where $\sum_{i=1}^{n} \phi_i = 1$. The units in the heads of queues are served simultaneously, and the others are waiting to be served. We assume that each unit is served infinite-divisibly, *i.e.* it is treated as if it is fluid when served. Let $c_i = \phi_i C$ and $S_i(\tau, t)$ be the amount of class i units served in an interval (τ, t). Then, the GPS system is described by the following two properties (Parekh and Gallager, 1993):

- It is work-conseving, *i.e.* no bandwidth is allocated to a class with empty queue, and if there is at least one class with backlogged queue, then the total allocated bandwidth is C.
- If the queue of class i is continuously backlogged in the interval (τ, t), then for any j,

$$\frac{S_i(\tau,t)}{S_j(\tau,t)} \geq \frac{c_i}{c_j}.$$

Let $s_i(t)$ be the service rate of class i at time t, which is defined as

$$s_i(t) = \lim_{u \to t+} \frac{S_i(t,u)}{u-t}.$$

Under the above assumption, if the queue of a class i is backlogged at time t, we can easily derive

$$s_i(t) = \frac{c_i\, C}{\sum\limits_{k \in \mathcal{S}_b(t)} c_k},$$

where $\mathcal{S}_b(t)$ is the set of classes with backlogged queues at time t. Clearly, for class i with empty queue at time t,

$$s_i(t) = 0.$$

Above two equations imply that the service rate of a queue is determined by the queue lengths of all classes, more specifically, the backlog states of all queues. Let the n-dimensional vector $L^X(t)$ denotes the queue lengths, *i.e.* $L_i^X(t)$, $i = 1, 2, \ldots, n$ denotes the number of units in the class i's queue, or queue i. When $L^X(t) = \mathbf{x}$, the two equations in the above say that the service rate or the bandwidth allocated to the queue i by the server X at the time t is given by

$$\begin{aligned} r_i(\mathbf{x}) &= 0, && \text{for } x_i = 0, \\ r_i(\mathbf{x}) &= \frac{c_i\, C}{\sum\limits_{\{k; x_k > 0\}} c_k}, && \text{for } x_i > 0. \end{aligned} \tag{2.1}$$

From the second equation in the above, we can see that the bandwidth allocated to a backlogged queue is not less than c_i. We call c_i, $i = 1, 2, \ldots, n$, the guaranteed bandwidth of the class i because the class i traffic has the minimum service rate c_i when the queue i is backlogged. Note that the full bandwidth of the system, C is equal to $\sum_{i=1}^n c_i$. Since $\phi_i = c_i/C$ for $i = 1, 2, \ldots, n$, the guaranteed bandwidths $(c_1, c_2, \ldots, c_n)$ uniquely define the full bandwidth of the GPS system and its service ratios.

For the input process to the GPS server, we assume that the units of each class arrive according to Poisson process. Let λ_i be the arrival rate of class i units, for $i = 1, 2, \ldots, n$. For different classes, their arrival processes are assume to be independent. We also assume that the service requirement of class i units are exponentially distributed with mean μ_i, $i = 1, 2, \ldots, n$. The service requirements of different units are also independent. The input load of the class i traffic is given by

$$\rho_i = \lambda_i\, \mu_i, \quad i = 1, 2, \ldots, n.$$

Then, the total load of the system is equal to $\sum_{i=1}^n \rho_i$. We assume that $\sum_{i=1}^n \rho_i < C$. Then, all of the queue lengths processes are stable.

3. Monotonic Head of the Line Processor Sharing Servers

GPS scheduling is a type of HLPS server, which is a scheduling policy for a server with multiple queues. Under HLPS policy, the units in the head of the queues are served simultaneously. The

bandwidth allocated to each queue are determined by the numbers of units in the queues. To say more clearly, we consider a HLPS server Y with n queues. Let $L^Y(t)$ be the n-dimensional vector which denotes the numbers of units in the queues. Suppose that for all n-dimensional nonnegative integer valued vector $\mathbf{y}$, the bandwidth allocated to the queue i when $L^Y(t) = \mathbf{y}$ is given by $r_i(\mathbf{y})$, $i = 1, 2, \ldots, n$. Then, we call the n-dimensional vector $r(\mathbf{y}) = (r_1(\mathbf{y}), r_2(\mathbf{y}), \ldots, r_n(\mathbf{y}))$ the bandwidths vector of the HLPS server Y. In what follows, $\mathbf{x}$ and $\mathbf{y}$ denote n-dimensional nonnegative integer valued vectors.

Bonald and Proutiére (2004) have defined the monotonicity for the stochastic networks in which each node serves the units in its queue according to the processor sharing policy. The HLPS server with n queues can be treated as a stochastic network with n nodes if we assume that the n queues of the HLPS server are severed by n separated servers with service rates $r_1(\cdot), r_2(\cdot), \ldots, r_n(\cdot)$, respectively. However, the monotonicity defined by Bonald and Proutiére is not adaptable to the HLPS servers because the units in each queue is not served by the processor sharing policy. We modify their definition to propose the monotonicity for the HLPS servers. For two n-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x} \leq \mathbf{y}$ denotes that $x_i \leq y_i$ for all i. Then, in the HLPS servers, the monotonicity is defined as follows:

Definition 3.1. *A HLPS server Y with the bandwidths vector $r(\cdot)$ is monotonic iff for all $\mathbf{x} \leq \mathbf{y}$,*

$$r_i(\mathbf{x}) \geq r_i(\mathbf{y}), \quad \forall\ i \text{ such that } x_i > 0.$$

It can be easily checked from Equation (2.1) that a GPS server is monotonic. For the monotonic HLPS servers Y_a and Y_b with the same number of the n-queues, the n-dimensional vectors $L^a(t)$ and $L^b(t)$ denote the numbers of units in the queues of the severs Y_a and Y_b, respectively, and let $r^a(\cdot)$ and $r^b(\cdot)$ be the bandwidths vectors of Y_a and Y_b, respectively. Suppose that

$$r^a(\mathbf{y}) \geq r^b(\mathbf{y}), \quad \forall\ \mathbf{y}$$

and that the server Y_a is monotonic. Then, for any state $\mathbf{x} \leq \mathbf{y}$, it follows from the monotonicity that

$$r^a(\mathbf{x}) \geq r^a(\mathbf{y}).$$

Then, the condition on the service rates $r^a(\cdot)$ and $r^b(\cdot)$ gives that

$$r^a(\mathbf{x}) \geq r^b(\mathbf{y}).$$

If both the servers are given the same sample input process and they are empty at time 0, then the above inequality says that

$$L^a(t) \leq L^b(t), \quad t \geq 0.$$

For any given sample path, the above is valid. When the input processes to the two servers follows the same law, we obtain the following theorem:

Theorem 3.1. *Suppose that two HLPS servers Y_a and Y_b have the bandwidths vectors $r^a(\cdot)$ and $r^b(\cdot)$, respectively, and that all of the queues of them are empty at time 0. If the server Y_a is monotonic and the input processes to the severs Y_a and Y_b follow the same law, then*

$$r^a(\mathbf{y}) \geq r^b(\mathbf{y}), \quad \forall\ \mathbf{y} \Longrightarrow L^a(t) \leq_{st} L^b(t), \quad t \geq 0.$$

4. Stochastic Upper Bound of the Stationary Queue Lengths

In this section, for a monotonic HLPS server X, we find a HLPS server Y such that the stationary queue lengths distribution has the product form, and that

$$L^X(t) \leq_{st} L^Y(t), \quad t \geq 0,$$

if the input processes to the servers X and Y have the same law. To find such a HLPS server Y, we adopt the method by Bonald and Proutiére (2004). Using the method, for a given stochastic network of nodes adopting the processor sharing policy, they found the stochastic network whose queue lengths processes of all nodes stochastically bound upper those of the given network for all time.

Using the bandwidths vector $r(\cdot)$ of the server X, we define a balance function $\Phi(\cdot)$ recursively as follows:

$$\begin{aligned} \Phi(\mathbf{0}_n) &= 1, \\ \Phi(\mathbf{x}) &= \max_{i:x_i>0} \frac{\Phi(\mathbf{x}-e_i)}{r_i(\mathbf{x})}, \end{aligned} \tag{4.1}$$

where $\mathbf{0}_n$ is the n-dimensional zero vector and e_i is the unit vector with i^{th} element being equal to 1 and others all zero. The value of $\Phi(\mathbf{x})$ for all possible $\mathbf{x}$ can be computed applying the above equation in a recursive manner. It can be easily checked that $\Phi(\mathbf{x})$ is well defined. We define Y as a HLPS server with service rate $\hat{r}(\cdot)$ given by

$$\begin{aligned} \hat{r}_i(\mathbf{x}) &= 0, && \text{for } x_i = 0, \\ \hat{r}_i(\mathbf{x}) &= \frac{\Phi(\mathbf{x}-e_i)}{\Phi(\mathbf{x})}, && \text{for } x_i > 0. \end{aligned} \tag{4.2}$$

Let us show that $\forall\, \mathbf{x}$,

$$r_i(\mathbf{x}) \geq \hat{r}_i(\mathbf{x}), \quad i = 1, 2, \ldots, n. \tag{4.3}$$

It follows from Equation (4.1) that

$$r_i(\mathbf{x}) \geq \frac{\Phi(\mathbf{x}-e_i)}{\Phi(\mathbf{x})}, \quad i = 1, 2, \ldots, n.$$

The latter term in the above inequality is equal to $\hat{r}_i(\mathbf{x})$ by definition, which gives the Inequality (4.3). Since the server X is a monotonic HLPS server, we have

$$L^X(t) \leq_{st} L^Y(t), \quad t \geq 0, \tag{4.4}$$

due to Theorem 3.1 when the input processes to the servers X and Y have the same law. At any time, the number of units in a queue of Y bounds upper that in the corresponding queue of X stochastically. For the stationary queue lengths, the same result also hold.

The service rates $\{\hat{r}_1(\cdot), \hat{r}_2(\cdot), \ldots, \hat{r}_n(\cdot)\}$ are Φ-balanced, *i.e.* it satisfies that for $j = 1, 2, \ldots, n$,

$$\Phi(\mathbf{x})\, \hat{r}_j(\mathbf{x}) = \Phi(T_{jk}\mathbf{x})\, \hat{r}_k(T_{jk}\mathbf{x}), \quad \text{for } \mathbf{x} \text{ such that } x_j > 0,$$

where $T_{jk}\mathbf{x}$ means the vector $\mathbf{x} - e_j + e_k$. The above equation follows directly from Equation (4.2). For the details of the Φ-balance, refer to Serfozo (1999). Since the HLPS server Y has the

bandwidths vector which is Φ-balanced, the stationary distribution $\hat{\pi}(\mathbf{x})$ of $\{L^Y(t), t \geq 0\}$ is given by the following product form:

$$\hat{\pi}(\mathbf{x}) = \hat{\pi}(\mathbf{0})\, \Phi(\mathbf{x})\, \Pi_{i=1}^{n} \rho_i^{x_i},$$

where the values of $\hat{\pi}(\mathbf{0})$ is obtained by summing $\Phi(\mathbf{x})\, \Pi_{i=1}^{n} \rho_i^{x_i}$ over all possible $\mathbf{x}$.

Now, we consider the case that X is a GPS server. Without loss of the generality, we assume $c_1 \leq c_2 \leq \cdots \leq c_n$ in what follows. Since $r_i(k\, e_i) = C, k = 1, 2, \ldots$, it follows from Equation (4.1) that $\Phi(k\, e_i) = \Phi((k-1)e_i)/C, k = 1, 2, \ldots$, which gives that for $i = 1, 2, \ldots, n$,

$$\Phi(k\, e_i) = \frac{1}{C^k}, \quad k = 1, 2, \ldots. \tag{4.5}$$

It is shown in Appendix that for $\mathbf{x}$ being the sum of different e_i's

$$\Phi\left(\sum_{j=1}^{m} e_{i_j}\right) = \prod_{j=1}^{m} \left(\frac{\sum_{k=j}^{m} c_{i_k}}{c_{i_j}\, C}\right), \tag{4.6}$$

where $e_{i_1}, e_{i_2} \ldots, e_{i_m}$ are the elementary vectors with an order such that $c_{i_1} \leq c_{i_2} \leq \cdots \leq c_{i_m}$, and that for $\mathbf{x} \neq \mathbf{0}$,

$$\Phi(\mathbf{x}) = \Phi\left(\sum_{j \in B(\mathbf{x})} e_j\right) \prod_{j \in B(\mathbf{x})} \left(\frac{\sum_{k \in B(\mathbf{x})} c_k}{c_j\, C}\right)^{x_j - 1}, \tag{4.7}$$

where $B(\mathbf{x})$ is the set $\{j; x_j > 0\}$.

Using the above, we rewrite the stationary distribution $\hat{\pi}(\mathbf{x})$ as follows:

$$\hat{\pi}(\mathbf{x}) = \hat{\pi}(\mathbf{0})\, \Phi\left(\sum_{j \in B(\mathbf{x})} e_j\right) \prod_{j \in B(\mathbf{x})} \rho_j \left(\frac{\rho_j \sum_{k \in B(\mathbf{x})} c_k}{c_j\, C}\right)^{x_j - 1}. \tag{4.8}$$

The value of $\hat{\pi}(\mathbf{0})$ can be obtained by summing $\hat{\pi}(\mathbf{x})$ over all $\mathbf{x}$, whose value is equal to 1. Let S_q be the set of all n-dimensional vectors of nonnegative integers, *i.e.* the set of all possible queue lengths states. Then, it follows from the above equation that

$$\hat{\pi}(\mathbf{0}) \sum_{\mathbf{x} \in S_q} \Phi\left(\sum_{j \in B(\mathbf{x})} e_j\right) \prod_{j \in B(\mathbf{x})} \rho_j \left(\frac{\rho_j \sum_{k \in B(\mathbf{x})} c_k}{c_j\, C}\right)^{x_j - 1} = 1. \tag{4.9}$$

To obtain the more refined form of $\hat{\pi}(\mathbf{0})$, we denote by S_b the set of all n-dimensional vectors $\mathbf{x}$ such that x_i is 0 or 1. An n-dimensional vector $\mathbf{x}$ is in S_b iff $\mathbf{x}$ has the form $\mathbf{x} = e_{i_1} + \cdots + e_{i_m}$ for an $m\,(m > 0)$, where all e_{i_k}'s are different and all i_k's are not larger than n. Then, we have from the above that

$$\hat{\pi}(\mathbf{0}) \left\{1 + \sum_{\mathbf{x} \in S_b} \Phi(\mathbf{x}) \prod_{j \in B(\mathbf{x})} \left(\frac{\rho_j c_j C}{c_j C - \rho_j \sum_{k \in B(\mathbf{x})} c_k}\right)\right\} = 1. \tag{4.10}$$

Note that the number of elements in S_b is $2^n - 1$. Thus, the value of $\hat{\pi}(\mathbf{0})$ is calculated without an approximation, which is the advantage of using the above equation in calculating $\hat{\pi}(\mathbf{0})$ over using the Equation (4.9).

The above equation says that $\hat{\pi}(\mathbf{x})$ exists iff $\rho_i < c_i$, $\forall i$. In other words, the stationary distribution of Y exists iff the input load of class i units is less than the guaranteed bandwidth c_i for all i. This sufficient and necessary condition for the server Y to have the stationary queue lengths distribution is somewhat restrictive as compared with the corresponding condition for X which is $\sum_{i=1}^{n} \rho_i < C$. This means that we cannot find always such a server Y whose stationary queue lengths bound upper those of a given GPS server X. However, it is inevitable that the condition for Y is more restrictive than that for X because the stationary queue length of a queue i in Y is stochastically larger than or equal to that in X for all i.

A direct consequence of the Equation (4.8) is that we can obtain the upper bound of the probability $\Pr\{L_i^X \geq x_i, i = 1, 2, \ldots, n\}$, where L_i^X is the stationary queue length of class i unit in the server X. It follows from the Equation (4.4) that for a given $\mathbf{x}$,

$$\Pr\left\{L_i^X \geq x_i, i = 1, 2, \ldots, n\right\} \leq \Pr\left\{L_i^Y \geq x_i, i = 1, 2, \ldots, n\right\},$$

where L_i^Y is the stationary queue length of class i unit in the server Y. Then, the Equation (4.8) gives

$$\Pr\left\{L_i^X \geq x_i, i = 1, 2, \ldots, n\right\} \leq \sum_{\mathbf{y}: y_i \geq x_i, \forall i} \hat{\pi}(\mathbf{0})\, \Phi\left(\sum_{j \in B(\mathbf{y})} e_j\right) \prod_{j \in B(\mathbf{y})} \rho_j \left(\frac{\rho_j \sum\limits_{k \in B(\mathbf{y})} c_k}{c_j\, C}\right)^{y_j - 1}.$$

When all of x_i's are positive, the $B(\mathbf{y})$ in the above inequality is equal to $\{1, 2, \ldots, n\}$. In this case, the above inequality can be rewritten as a simpler form, which is given by the following theorem:

Theorem 4.1. *Let $L_i^X, i = 1, 2, \ldots, n$, is the stationary queue length of class i unit in the server X. Then, for positive integer x_i, $i = 1, 2, \ldots, n$,*

$$\Pr\left\{L_i^X \geq x_i, i = 1, 2, \ldots, n\right\} \leq \hat{\pi}(\mathbf{0})\, \Phi\left(\sum_{j=1}^{n} e_j\right) \prod_{j=1}^{n} \frac{\rho_j c_j}{c_j - \rho_j} \left(\frac{\rho_j}{c_j}\right)^{x_j - 1}. \tag{4.11}$$

5. Special Case

In this section, we consider the GPS server X with two classes and $\rho_i < c_i$ for $i = 1, 2$. Since $r_1(x, 0) = C$, $x = 1, 2, \ldots$, it follows from Equation (4.5) that

$$\Phi(x_1, 0) = \frac{1}{C^{x_1}}, \quad x_1 = 1, 2, \ldots.$$

In the same manner, we also have

$$\Phi(0, x_2) = \frac{1}{C^{x_2}}, \quad x_2 = 1, 2, \ldots.$$

For $x_1 > 0$ and $x_2 > 0$, the Equation (4.7) gives

$$\Phi(x_1, x_2) = \frac{1}{c_1^{x_1} c_2^{x_2 - 1} C}.$$

From the Equation (4.10), we have

$$\hat{\pi}(0,0)\left(\Phi(0,0) + \Phi(1,0)\frac{\rho_1 C}{C - \rho_1} + \Phi(0,1)\frac{\rho_2 C}{C - \rho_2} + \Phi(1,1)\frac{\rho_1 c_1}{c_1 - \rho_1}\frac{\rho_2 c_2}{c_2 - \rho_2}\right) = 1.$$

From the above equations and $\Phi(0,0) = 1$, the value of $\hat{\pi}(0,0)$ is given by

$$\hat{\pi}(0,0) = \left(1 + \frac{\rho_1}{C-\rho_1} + \frac{\rho_2}{C-\rho_2} + \frac{\rho_1\rho_2 c_2}{(c_1-\rho_1)(c_2-\rho_2)C}\right)^{-1}.$$

For x_1 and x_2 being positive, we have

$$\hat{\pi}(x_1,0) = \hat{\pi}(0,0)\left(\frac{\rho_1}{C}\right)^{x_1},$$
$$\hat{\pi}(0,x_2) = \hat{\pi}(0,0)\left(\frac{\rho_2}{C}\right)^{x_2},$$
$$\hat{\pi}(x_1,x_2) = \hat{\pi}(0,0)\frac{c_2}{C}\left(\frac{\rho_1}{c_1}\right)^{x_1}\left(\frac{\rho_2}{c_2}\right)^{x_2}.$$

Then, the above equations give the following bound for the stationary queue lengths of X:

$$\Pr\left\{L_1^X \geq x_1, L_2^X \geq x_2\right\} \leq \hat{\pi}(0,0)\frac{\rho_1\rho_2 c_2}{(c_1-\rho_1)(c_2-\rho_2)C}\left(\frac{\rho_1}{c_1}\right)^{x_1-1}\left(\frac{\rho_2}{c_2}\right)^{x_2-1}, \quad x_1 > 0,\ x_2 > 0.$$

which also can be obtained from Theorem 4.1.

6. Numerical Results

In this section, we consider some numerical studies to check how good the result obtained in Section 4 is when it is applied to obtain some upper bounds of interest. GPS system with three classes is simulated. The arrival rate of each class is equally set to be 0.5, *i.e.* $\lambda_i = 0.5$, $i = 1,2,3$. The service requirement of each unit does not depend on the class which the unit belongs to, and it is exponentially distributed with mean 1. Thus, $\rho_i = 0.5$, $i = 1,2,3$. We consider the following three cases with different sets of the guaranteed bandwidths:

Case 1: $c_1 = 1.0$, $c_2 = 1.0$, $c_3 = 1.0$.

Case 2: $c_1 = 0.8$, $c_2 = 1.0$, $c_3 = 1.2$.

Case 3: $c_1 = 0.6$, $c_2 = 1.0$, $c_3 = 1.4$.

In case 1 we treat GPS system with the equal guaranteed bandwidths, in case 2 GPS system with the different guaranteed bandwidths, and in case 3 GPS system with the very different guaranteed bandwidths.

Figure 6.1 shows the true value of the probability that the sum of the stationary queue lengths is larger than given thresholds for the three cases. Let $Z(t)$ be the sum of the queue lengths at time t, *i.e.* $Z(t) = L_1(t) + L_2(t) + L_3(t)$, where $L_i(t), i = 1,2,3$, is the queue length of the class i units in its queue at time t. It is clear that $\{Z(t), t \geq 0\}$ is a birth-and-death process. The increasing rate at any state is given by the sum of the arrival rates λ_i's, which is equal to 1.5. The work-conserving property of GPS server says that total bandwidth of the sever is equal to 3 when there is at least one unit in any queue. This implies that the decreasing rates of $\{Z(t), t \geq 0\}$ at all positive states are the same as 3 and 0 at the zero state. These increasing and decreasing rates are all the same for the three different cases. Since the birth-and-death process is uniquely characterized by the increasing and decreasing rates at each state, the stationary distributions of $\{Z(t), t \geq 0\}$ for the three cases are the same. We can easily obtain the stationary distribution, given by

$$\Pr\{Z > k\} = (0.5)^{k+1},$$

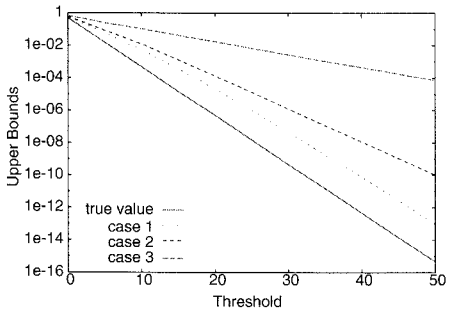
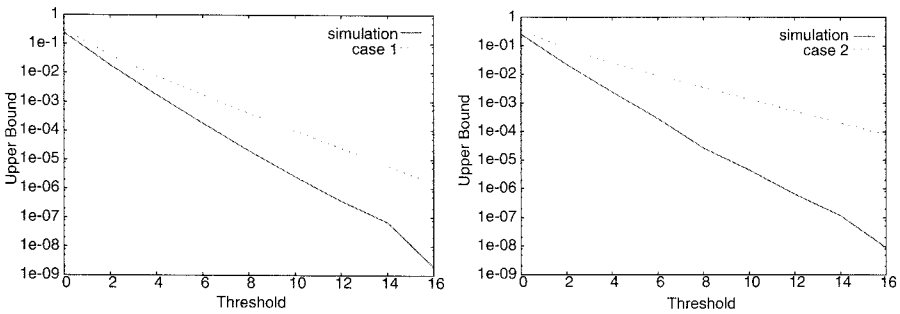
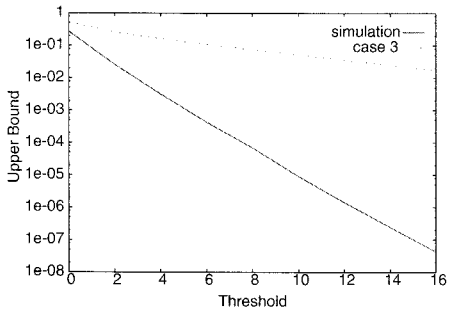


Figure 6.1. True value and the upper bounds on the probability of the sum of the stationary queue lengths being larger than various thresholds for the three different cases.



(a) $c_1 = 1.0,\ c_2 = 1.0,\ c_3 = 1.0$ (b) $c_1 = 0.8,\ c_2 = 1.0,\ c_3 = 1.2$



(c) $c_1 = 0.6,\ c_2 = 1.0,\ c_3 = 1.4$

Figure 6.2. Simulation result and the upper bounds on the probability of the stationary queue length of the first class being larger than various thresholds for the three different cases. The guaranteed bandwidths to each queue vary with the different cases.

where Z is the sum of the stationary queue lengths. True value in Figure 6.1 is obtained from the above equation. The upper bounds for the three case in Figure 6.1 are obtained from the Equation (4.8). In the figure we can see that for the case 1, the Equation (4.8) gives a relatively good upper bound. The ratio of the upper bound to the true value is about 100 or less. However, for the case 3, the ratio is very huge. Thus, in this case, the upper bound by the Equation (4.8) is very worse.

When we are interested in the stationary distribution of $\{L_i(t), t \geq 0\}$, we have no explicit formula for the distribution. Thus, we should rely on the simulation. Figure 6.2 gives the simulation results on the probability $\Pr\{L_1 > x\}$, $x = 0, 2, \ldots, 16$, where L_1 is the stationary queue length of the class

1 units. The corresponding upper bounds, for the three cases, are obtained from the Equation (4.8). We can see the similar behavior of the upper bounds to the above. For the case 1, the Equation (4.8) gives a relatively good upper bound, while for the case 3, the upper bound by the Equation (4.8) is very worse.

Appendix

In Appendix, we give the derivation of (4.6) first, and then (4.7). Equation (4.5) says that $\Phi(e_i) = 1/C, i = 1, 2, \ldots, n$. Thus, the Equation (4.6) is valid when $\mathbf{x} = e_i$ for an i. Suppose that the Equation (4.6) is valid for all $\mathbf{x}$ which is the sum of $m - 1 (1 < m \leq n)$ different e_i's. Let $\mathbf{x} = e_{i_1} + \cdots + e_{i_m}$ with $c_{i_1} \leq c_{i_2} \leq \cdots \leq c_{i_m}$. From the Equation (4.6), we can see that $\Phi(\mathbf{x} - e_{i_j})$, $j = 1, 2, \ldots, m$, has its maximum value at $j = 1$, *i.e.*

$$\max_{j=1,2,\ldots,m} \Phi(\mathbf{x} - e_{i_j}) = \prod_{j=2}^{m} \left(\frac{\sum_{k=j}^{m} c_{i_k}}{c_{i_j} C} \right).$$

Moreover, $r_{i_j}(\mathbf{x}) = c_{i_j} C / \sum_{j=1}^{m} c_{i_j}$ has its minimum value at $j = 1$. Thus, it follows from the Equation (4.1) that

$$\Phi(\mathbf{x}) = \prod_{j=1}^{m} \left(\frac{\sum_{k=j}^{m} c_{i_k}}{c_{i_j} C} \right).$$

Now, by induction, we have shown that the Equation (4.6) is valid for all $m = 1, 2, \ldots, n$.

We also show the Equation (4.7) for $\mathbf{x} \neq \mathbf{0}$ by induction. When $\mathbf{x}$ is the sum of different e_i's, $x_j = 1$ for all $j \in B(\mathbf{x})$. Thus, the Equation (4.7) is valid by definition in this case. For a given $\mathbf{x}$ such that at least one of x_i is larger than 1, we assume that for all $\mathbf{y} \in \{\mathbf{y}; \mathbf{y} \leq \mathbf{x} \text{ and } \mathbf{y} \neq \mathbf{x}\}$, $\Phi(\mathbf{y})$ is given by the Equation (4.7). For i with $x_i > 0$, $r_i(\mathbf{x})$ is given by $c_i C / \sum_{k \in B(\mathbf{x})} c_k$. Then, we have that for i such that $x_i > 0$,

$$\frac{\Phi(\mathbf{x} - e_i)}{r_i(\mathbf{x})} = \left(\frac{\sum_{k \in B(\mathbf{x})} c_k}{c_i C} \right) \Phi \left(\sum_{j \in B(T_i \mathbf{x})} e_j \right) \prod_{j \in B(T_i \mathbf{x})} \left(\frac{\sum_{k \in B(T_i \mathbf{x})} c_k}{c_j C} \right)^{(T_i \mathbf{x})_j - 1}. \tag{A.1}$$

where $T_i \mathbf{x} = \mathbf{x} - e_i$ and its j^{th} element is denoted by $(T_i \mathbf{x})_j$. If $x_i > 1$, then $B(T_i \mathbf{x}) = B(\mathbf{x})$ and $(T_i \mathbf{x})_j = x_j$ for all $j \neq i$. Then, the above equation is rewritten as

$$\frac{\Phi(\mathbf{x} - e_i)}{r_i(\mathbf{x})} = \Phi \left(\sum_{j \in B(\mathbf{x})} e_j \right) \prod_{j \in B(\mathbf{x})} \left(\frac{\sum_{k \in B(\mathbf{x})} c_k}{c_j C} \right)^{x_j - 1}, \quad i \text{ such that } x_i > 1. \tag{A.2}$$

The right hand side of the above equation does not depend on i. If $x_i = 1$, then the difference of the sets $B(\mathbf{x})$ and $B(T_i \mathbf{x})$ is $\{i\}$ and $(T_i \mathbf{x})_j = x_j$ for all $j \neq i$. Thus, we have

$$\prod_{j \in B(T_i \mathbf{x})} \left(\frac{\sum_{k \in B(T_i \mathbf{x})} c_k}{c_j C} \right)^{(T_i \mathbf{x})_j - 1} = \prod_{j \in B(\mathbf{x})} \left(\frac{\sum_{k \in B(T_i \mathbf{x})} c_k}{c_j C} \right)^{x_j - 1}.$$

Since $\sum_{k\in B(T_i\mathbf{x})} c_k$ is less than $\sum_{k\in B(\mathbf{x})} c_k$, the above equation gives

$$\prod_{j\in B(T_i\mathbf{x})} \left(\frac{\sum_{k\in B(T_i\mathbf{x})} c_k}{c_j C} \right)^{(T_i\mathbf{x})_j - 1} \leq \prod_{j\in B(\mathbf{x})} \left(\frac{\sum_{k\in B(\mathbf{x})} c_k}{c_j C} \right)^{x_j - 1}.$$

Moreover, it follows from the Equation (4.1) that

$$\left(\frac{\sum_{k\in B(\mathbf{x})} c_k}{c_i C} \right) \Phi\left(\sum_{j\in B(T_i\mathbf{x})} e_j \right) \leq \Phi\left(\sum_{j\in B(\mathbf{x})} e_j \right).$$

Applying the above two equations to the Equation (A.1) yields

$$\frac{\Phi(\mathbf{x}-e_i)}{r_i(\mathbf{x})} \leq \Phi\left(\sum_{j\in B(\mathbf{x})} e_j \right) \prod_{j\in B(T_i\mathbf{x})} \left(\frac{\sum_{k\in B(\mathbf{x})} c_k}{c_j C} \right)^{x_j - 1}, \quad i \text{ such that } x_i = 1. \tag{A.3}$$

Since $\Phi(\mathbf{x})$ is given by $\max_{i;x_i>0} \Phi(\mathbf{x}-e_i)/r_i(\mathbf{x})$ and $\mathbf{x}$ is such a vector that at least one of x_i is larger than 1, the Equations (A.2) and (A.3) completes the proof.

References

Adan, I. J. B. F., Boxma, O. J. and Resing, J. A. C. (2001). Queueing models with multiple waiting lines, *Queueing Systems*, **37**, 65–98.

Bertsimas, D., Paschalidis, I. C. and Tsitsiklis, J. N. (1999). Large deviations analysis of the generalized processor sharing policy, *Queueing Systems*, **32**, 319–349.

Bonald, T. and Proutière, A. (2004). On stochastic bounds for monotonic processor sharing networks, *Queueing Systems*, **47**, 81–106.

Borst, S., Mandjes, M. and van Uitert, M. (2003). Generalized processor sharing queues with heterogeneous traffic classes, *Advances in Applied Probability*, **35**, 806–845.

Borst, S. and Zwart, B. (2003). A reduced-peak equivalence for queues with a mixture of light-tailed and heavy-tailed input flows, *Advances in Applied Probability*, **35**, 793–805.

Brandt, A. and Bradnt, M. (1998). On the sojourn times for many-queue head-of-the-line processor-sharing systems with permanent customers, *Mathematical Methods of Operations Research*, **47**, 181–220.

Cohen, J. W. (1988). Boundary value problems in queueing theory, *Queueing Systems*, **3**, 97–128.

de Veciana, G. and Kesidis, G. (1996). Bandwidth allocation for multiple qualities of service using generalized processor sharing, *IEEE Transactions on Information Theory*, **42**, 268–272.

Dupuis, P. and Ramanan, K. (1998). A Skorokhod problem formulation and large deviation analysis of a processor sharing model, *Queueing Systems*, **28**, 109–124.

Fayolle, G. and Iasnogorodski, R. (1979). Two coupled processors: The reduction to a Riemann-Hilbert problem, *Probability Theory and Related Fields*, **47**, 325–351.

Parekh, A. K. and Gallager, R. G. (1993). A generalized processor sharing approach to flow control in integrated services networks: The single node case, *IEEE/ACM Transactions on Networking*, **1**, 344–357.

Serfozo, R. (1999). *Introduction to Stochastic Networks*, Springer, New York.

Zhang, Z. L. (1998). Large deviations and the generalized processor sharing scheduling for a multiple-queue system, *Queueing Systems*, **28**, 349–376.