1995

Large Deviations and the Generalized Processor Sharing Scheduling: Upper and Lower Bounds Part I: Two-Queue Systems

Zhi-Li Zhang
University of Massachusetts - Amherst

Large Deviations and the Generalized Processor Sharing Scheduling: Upper and Lower Bounds Part I: Two-Queue Systems

Zhi-Li Zhang
Computer Science Department
University of Massachusetts at Amherst
Amherst, MA 01003, USA
Email: zhzhang@gaia.cs.umass.edu

October 1994; Revised June 1995, September 1995

UMASS CMPSCI Technical Report UM-CS-95-96

Abstract

We prove asymptotic upper and lower bounds on the asymptotic decay rate of per-session queue length tail distributions for a single constant service rate server queue shared by multiple sessions with the *generalized processor sharing* (GPS) scheduling discipline. The simpler case of a GPS system with only two queues needs special attention, as under this case, it is shown that the upper bounds and lower bounds match, thus yielding exact bounds. This result is established in this part (Part I) of the paper. The general case is much more complicated, and is treated separately in Part II of the paper [42], where tight upper and lower bound results are proved by examining the dynamics of bandwidth sharing nature of GPS scheduling. The proofs use sample-path large deviation principle and are based on some recent large deviation results for a single queue with a constant service rate server. These results have implications in call admission control for high-speed communication networks.

1 Introduction

In the future high speed digital networks, *e.g.*, ATM networks or future integrated services Internet, an important open and challenging issue is how to effectively and efficiently manage network resources, by means of call admission control, bandwidth allocation and packet/cell scheduling, to support a variety of applications including voice, video and datagram traffic with diverse traffic characteristics and *quality of service* (QoS) requirements. This issue has been studied extensively from both theoretical and practical point of view (see [1, 2] for some recent theoretical effort). One solution for dealing with the diversity of traffic characteristic and QoS requirements is to provide different QoS service classes with dedicated queues shared only by sources in the same class. More sophisticated scheduling mechanism other than the simple First-In First-Out (FIFO) service discipline is needed to provide both protection and bandwidth sharing among service classes.

For this purpose, the Generalized Process Sharing (GPS) service discipline [34, 33] was proposed recently and is recommended for use in future integrated services packet networks [11, 36]. GPS is a work-conserving scheduling

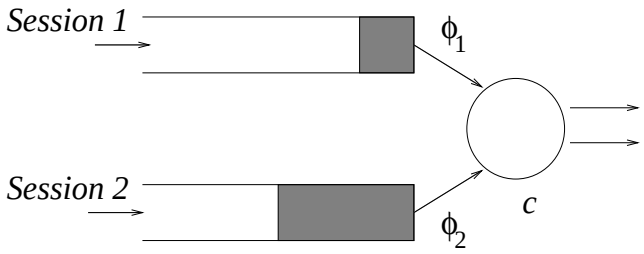

Figure 1: A two-queue GPS system.

discipline, it assumes a fluid source model where source traffic is treated as infinitely divisible fluid[1]. Consider n sessions sharing a GPS server with rate c, each session with its own queue (see Figure 1 for a GPS server with two queues). Associated with the sessions are parameters $\{\phi_i\}_{1\le i\le n}$ (called *GPS assignment*) which determine the minimum sharing of bandwidth of each session. Each session is guaranteed a minimum service rate of $g_i = \frac{\phi_i}{\sum_{j=1}^{n}\phi_j}c$. More generally, if the set of sessions with queued packets at time t is $\mathcal{B}(t) \subseteq \{1,\ldots,n\}$, the session $i \in \mathcal{B}(t)$ receives service at rate $\frac{\phi_i}{\sum_{j\in\mathcal{B}(t)}\phi_j}$ at time t.

The performance of GPS has been studied in both deterministic [34, 35, 33] and stochastic setting [41, 44] using the so-called bounding approach [30]. For sources conforming to certain general bounding source models [12] and [40], upper bounds on the interested metrics such as loss or delay are derived. In the deterministic case, Parekh and Gallager [34, 35, 33] show that the upper bounds are attainable in the worst-case. In the stochastic setting, how tight the upper bounds are is still an open question.

In this paper, we study the asymptotic behavior of the GPS system by applying the theory of large deviation. In particular, we are interested in deriving upper and lower bounds on the asymptotic decay rate of the queue length tail distribution of each session. We consider a *discrete-time fluid model*, by which we mean that arrival and service happen at discrete-time slot indexed by integers, but arrival and service are in the form of *fluid*, *i.e.*, they are infinite divisible.

In Part I of the paper, we look at a two-queue GPS system. Under this special case, bandwidth sharing of the two sessions in the system can be easily captured. Let the GPS assignment $\{\phi_i\}_{i=1,2}$ for the two sessions is such that $0 \le \phi_1, \phi_2 \le 1$ and $\phi_1 + \phi_2 = 1$. Whenever both sessions are busy (*i.e.*, both queues are no empty), then each session i gets exactly ϕ_i share of the total bandwidth c. But, if one session is not busy (hence its queue is empty), then the residual bandwidth not consumed by this session is taken over by the other session if its queue is not empty. Due to this simplicity in bandwidth sharing, the upper bound and lower bounds we obtain are exactly the same for the two-queue GPS system.

In part II of the paper [42], we consider a general GPS system with more than two queues. Due to the complexity of the bandwidth sharing mechanism in the general GPS system, the upper and lower bounds we obtain do not match exactly, but have similar form, indicative of their tightness.

[1] Hence there is no notion of "packet" in this fluid traffic model [34, 33]. For practical implementation, a packetized version of GPS, called *packet-by-packet* GPS (PGPS), is designed that closely approximates the behavior of the ideal fluid GPS with the error term bounded by the transmission time of the largest packet [34, 33]. Thus results about the ideal fluid GPS can be easily adapted to the packetized version of GPS or any its variations by properly taking the error terms into consideration [34, 33, 41]. Equivalent forms of GPS and PGPS are also proposed in [15], where the packetized version PGPS is known as *Weighted Fair Queueing* (WFQ) and the ideal fluid GPS as *bit-by-bit* WFQ. The name, *Fluid Fair Queueing* (FFQ), is also used in the literature [24] for GPS.

Study of asymptotic behavior of queueing systems has its implication in call admission control with QoS guarantees for the future high-speed networks. The theory of effective bandwidths (see [2], *e.g.*, [26, 25, 22, 27, 21, 29, 39, 6, 23, 18, 31]) developed in recent years exploits this asymptotics to provide a simple theoretical call admission control scheme for networks represented by a single server with a shared queue. This scheme is *asymptotically optimal*. For networks employing GPS service discipline, a theoretical admission control framework is laid out in [43] for various network service models based on the results in [44]. Optimal and sub-optimal call admission control schemes are designed using the stochastic envelope process model [6] and the theory of effective bandwidths. Although the upper bounds obtained in this paper are tighter than those in [44], they are generally impossible to compute effectively. Hence they are mostly of theoretical interests. In [32], approximation methods are used to obtained tight bounds for the GPS system.

The result for the two-queue GPS system is first stated in [16] and proved under weaker assumptions than ours. However, due to their resort to a Loynes-type argument [28] [3], their lower bound argument is somewhat less convincing and rigorous. In contrast, we argue directly with the stationary version of the processes. To obtain the lower bound, we apply the sample-path large deviation principle [14, 7] which requires stronger assumptions on the arrival processes. Our results for the GPS system with more than two queues are more general than theirs, as we exploit the bandwidth sharing dynamics in more details.

The rest of Part I of the paper is organized as follows. Section 2 briefly reviews the large deviation principle and state several results regarding discrete-Time G/ D/ 1/ ∞ queueing systems which will be used later. In section 3 we state and prove the upper and lower bounds for the two-queue GPS system. Section 4 concludes Part I of the paper.

2 Large Deviations and Discrete-Time G/ D/ 1/ ∞ Queueing Systems

In this section, we briefly review some concepts and results from large deviation theory on the real-line $I\!R$ that are needed in this paper [4] and its application to performance analysis of discrete-time, single-server G/ D/ 1/ ∞ queueing systems.

The large deviation principle (LDP) on $I\!R$ characterizes the limiting behavior of a sequence of probability measures $\{\mu_n, n = 1, 2, \ldots\}$ on $I\!R$. We say a function I from $I\!R$ to $[0, \infty]$ is a good rate function if all the level sets $\{y \in I\!R : I(y) \leq x\}$, $x \in [0, \infty)$, are compact.

Definition 1 *A sequence of probability measures $\{\mu_n, n = 1, 2, \ldots\}$ on $I\!R$ satisfies the large deviation principle with a good rate function I if,*

[2] For an excellent survey on the theory of effective bandwidths, see [9].

[3] Note that when applying Loynes' Theorem [28], stationarity of both arrival and service processes are required. Although by applying Loynes' Theorem to the whole GPS system [16], we know that the queue length distribution for each session, $Q^{i,m}$, converges to a stationary process Q^i in distribution, but it is *not* necessarily true that $Q^{i,m}$ converges *monotonically* to Q^i in distribution, as would be the case *if* Loynes' Theorem could be applied to the individual queue directly. Due to this technical difficulty, when passing from the non-stationary regime to the stationary regime, great caution should be exercised. The first equality involving $\liminf_{B\to\infty} \frac{1}{B} \log$ in the lower bound proof (p. 11) of [16], where Q^1 and B are simultaneously replaced by $Q^{1,m}$ and $m\beta$, is dubious.

[4] For simplicity, we do not state the results in their most general form. [14, 20, 37] are good sources for reference on the subject. For application of large deviation theory in communication networks, see the excellent survey paper [38].

Upper Bound: *For every closed set F,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(F) - \inf_{x\in F} I(x); \tag{1}$$

Lower Bound: *For every open set G,*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \geq - \inf_{x\in G} I(x). \tag{2}$$

Let $\{X(n), n = 1, 2, \ldots\}$ be a sequence of random variables on $I\!R$ and $\{\mu_n, n = 1, 2, \ldots\}$ be corresponding sequence of probability measures of the random variables. If $\{\mu_n, n = 1, 2, \ldots\}$ satisfies the large deviation principle with a good rate function I, we also say $\{X(n), n = 1, 2, \ldots\}$ satisfies the large deviation principle with a good rate function I.

For any $\theta \in I\!R$, define

$$\Lambda_n(\theta) = \log E[e^{\theta X_n}].$$

Note that $\Lambda_n(\theta)$ is the logarithmic moment generating function of X_n.

Theorem 1 (Gärtner-Ellis Theorem) *[20] Let $\{X_n, n = 1, 2, \ldots\}$ be a sequence of random variables on $I\!R$. If $\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n}\Lambda_n(n\theta)$ exists (as a finite number) and is differentiable for all $\theta \in I\!R$, then $\{X_n, n = 1, 2, \ldots\}$ satisfies the large deviation principle with the good rate function $\Lambda^*(x)$ defined as follows:*

$$\Lambda^*(x) = \sup_{\theta\in I\!R}\{\theta x - \Lambda(\theta)\}.$$

$\Lambda^*(x)$ is called the Legendre-Fenchel transform (or convex conjugate) of $\Lambda(\theta)$. Note that under the assumption of the theorem, $\Lambda(\theta)$ is a strictly convex function.

Definition 1 to a space of functions on R and this leads to the sample path large deviation principle on $I\!R$. Let $D([0,1], I\!R)$ denote the space of right continuous and left limit functions from $[0,1]$ to $I\!R$ equipped with the supremum norm topology, *i.e.*, $\|f\| = \sup_{\{0\leq t\leq t\}} |f(t)|$, for $f \in D([0,1], I\!R)$. We say a sequence of probability measures μ_n on $D([0,1], I\!R)$ satisfies the *sample path large deviation principle* with a good rate function $I(\phi)$ if $I(\phi)$ is a function from $D([0,1], I\!R)$ to $[0,\infty]$ with compact level sets and the upper bound (1) and the lower bound (2) hold for any closed and open sets in $D([0,1], I\!R)$, respectively.

Let $\{Y_n, n = 1, 2, \ldots\}$ be a set of random variables on $I\!R$. Define the partial sum process $Z_n = \sum_{i=1}^n Y_i/n$, $n = 1, 2, \ldots$. Assume that $\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E e^{n\theta Z_n}$ exists (as a finite number) and is differentiable for all $\theta \in I\!R$ and denote its Legendre-Fenchel transform as $\Lambda^*(x)$. From Gärtner-Ellis Theorem, $\{Z_n, n = 1, 2, \ldots\}$ satisfies the large deviation principle with the good rate function Λ^*.

Now for $n = 1, 2, \ldots$, define a sequence of scaled partial sum processes of $\{Y_n, n = 1, 2, \ldots\}$ on $[0,1]$:

$$Z^{(n)}(t) = \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} Y_i, \quad 0 \leq t \leq 1.$$

Clearly, $Z^{(n)} \in D([0,1], I\!R)$. Let $\mu^{(n)}$ be the probability measure of $Z^{(n)}$. In [13], conditions are established under which $\{\mu^{(n)}, n = 1, 2, \ldots\}$ satisfies the sample large deviation deviation principle and the good rate function $I(\phi)$ is identified with the following form: for any $\phi \in D([0,1], I\!R)$,

$$I(\phi) = \begin{cases} \int_0^1 \Lambda^*(\phi'(t))dt, & \text{if } \phi \in AC_0([0,1], I\!R), \\ \infty, & \text{otherwise} \end{cases}$$

where $AC_0([0,1], I\!R)$ is the space of absolutely continuous functions from $[0,1]$ to $I\!R$ with $\phi(0) = 0$, and $\phi'(t)$ is the derivative of $\phi(t)$ at t.

Following [7], we say the sequence of probability measures $\{\mu^{(n)}, n = 1, 2, \ldots\}$ (or $\{Z_n, n = 1, 2, \ldots\}$) satisfies the sample path large deviation principle with respect to Λ.

Large deviation theory has been widely applied in queueing theory to study the tail probabilities of various queueing behaviors (see, *e.g.*, [3, 5, 6, 10, 19, 23, 39, 4]). Of particular relevance to us are the results on the discrete-time G/ D/ $1/\infty$ queueing system. The presentation below follows primarily the formulation in [6].

We describe the arrival process to a discrete-time G/ D/ $1/\infty$ queueing system by a sequence of bounded, nonnegative random variables on $I\!R$, $\{a(t), t \in I\!N\}$, where $I\!N$ is the set of nonnegative integers. In other words, at time t, the amount of arrivals to the queue is $a(t)$. For any $\tau = 0, 1, 2, \ldots$ and any $t \in I\!N$, $t > \tau$, define $A(\tau, t) = \sum_{s=\tau}^{t-1} a(s)$, the number of arrivals during the time interval $[\tau, t)$. Also let $A(\tau, \tau) = 0$. We call A the cumulative arrival process. We make the following assumptions on the arrival process $\{a(t), t = 0, 1, 2, \ldots\}$ [7].

(A1) The arrival process $\{a(t), t = 0, 1, 2, \ldots\}$ is ergodic and stationary.

(A2) For any $\theta \in I\!R$,

$$\Lambda_A(\theta) = \lim_{t\to\infty} \frac{1}{t} \log E e^{\theta A(0,t)} < \infty \tag{3}$$

and is differentiable.

(A3) $\{a(t), t = 0, 1, 2, \ldots\}$ is adapted to a filtration $\{\mathcal{F}_t^A, t \in I\!N\}$ with the following property: for any $\theta \in I\!R$, there exists a function $\Gamma_A(\theta)$, $0 \le \Gamma_A(\theta) < \infty$ such that for any $s = 0, 1, 2, \ldots$, $t \in I\!N$,

$$\Lambda_A(\theta)t - \Gamma_A(\theta) \le \log E(e^{\theta A(s,t+s)} | \mathcal{F}_s^A) \le \Lambda_A(\theta)t + \Gamma_A(\theta) \ \ a.s. \tag{4}$$

Note that (A3) implies (3) by taking $s = 0$ in (4). To emphasize (A2), we list it separately. Examples of random processes that satisfy (A1), (A2) and (A3) can be found in [8].

By Gärtner-Ellis Theorem, (A1) and (A2) imply that $\{A(0,t)/t, t \in I\!N\}$ satisfies the large deviation principle with the rate function

$$\Lambda_A^*(x) = \sup_{\theta \in I\!R} \{\theta x - \Lambda_A(\theta)\}.$$

Moreover, if (A3) is also satisfied, then $\{a(t), t = 0, 1, 2, \ldots\}$ satisfies the sample path large deviation principle [7]. More precisely, for $t = 1, 2, \ldots$, define the scaled process

$$A^{(t)}(u) = \frac{1}{t} A(0, \lfloor tu \rfloor), \quad 0 \le u \le 1. \tag{5}$$

Let $\mu^{(t)}$ be the distribution of $A^{(t)}(u)$. Then $\{\mu^{(t)}, t \in I\!N\}$ satisfies the sample path large deviation principle with the rate function $I_A(\phi)$ defined as follows:

$$I_A(\phi) = \begin{cases} \int_0^1 \Lambda_A^*(\phi'(u))du, & \text{if } \phi \in AC_0([0,1], I\!R), \\ \infty, & \text{otherwise.} \end{cases}$$

Let c be the rate of the server in the G/ D/ $1/\infty$ system. Assume that the system starts with an empty queue at time 0. Denote the backlog at time $t \in I\!N$ (or the queue length at time t) by $Q(t)$. Then $Q(0) = 0$. Moreover, by Lindley's equation, for $t = 0, 1, 2, \ldots$,

$$Q(t+1) = \max\{Q(t) + a(t) - c, 0\}. \tag{6}$$

Expanding (6) recursively, we have

$$Q(t) = \max_{0 \le \tau \le t} \{A(\tau, t) - c(t-\tau)\} \tag{7}$$

where τ takes only integer values. Throughout the paper, whenever a discrete-time system is considered, all time indices are integers.

A necessary and sufficient condition for the G/ D/ 1/ ∞ queueing system to be stable is that the average arrival rate is less than the service rate, *i.e.*, $Ea(0) < c$. Under this stability condition, by Loynes' Theorem [28], assuming that the system starts with an empty queue at time 0, the distribution of $Q(t)$ increases monotonically to a stationary distribution $Q(\infty)$ as $t \to \infty$ and $Q(\infty) < \infty$ almost surely (*a.s.*).

Given that the assumptions (A1) and (A2) on the arrival process and the above stability condition are satisfied, it has been proven (see, *e.g.* [6]) that for any $x \ge 0$,

$$\lim_{x \to \infty} \frac{1}{x} \log Pr\{Q(\infty) > x\} = -\theta^* \tag{8}$$

where θ^* is the unique solution to the equation $\Lambda_A(\theta) = \theta c$ or $\theta^* = \sup\{\theta \in I\!R : \Lambda_A(\theta) < c\theta\}$.

In other words, for any $\theta > 0$,

$$\lim_{x \to \infty} \frac{1}{x} \log Pr\{Q(\infty) > x\} \le -\theta \text{ iff } \Lambda_A(\theta) < c\theta. \tag{9}$$

Define $\alpha_A(\theta) = \Lambda_A(\theta)/\theta$. $\alpha_A(\theta)$ is called the *effective bandwidth* of the arrival process $\{a(t), t = 0, 1, 2, \ldots\}$ or the corresponding cumulative arrival process A.

For any $t \in I\!N$, let $S(0,t) = \sum_{\tau=0}^{t-1} b(\tau)$, where $b(\tau)$ is the number of departures at time τ. Thus $\{S(0,t), t \in I\!N\}$ is the (cumulative) departure process. Using the sample path large deviation principle, it is proved in [7] (see also [17]) that the $\{S(0,t)/t, t \in I\!N\}$ satisfies the large deviation principle with the rate function

$$\Lambda_D^*(\alpha) = \begin{cases} \Lambda_A^*(\alpha) & \text{if } \alpha \le c \\ \infty & \text{otherwise.} \end{cases} \tag{10}$$

Thus

$$\Lambda_D(\theta) = \sup_{\alpha \in I\!R} \{\theta\alpha - \Lambda_D^*(\alpha)\} = \begin{cases} \Lambda_A(\theta) & \text{if } 0 \le \theta \le \tilde{\theta} \\ \theta c - \tilde{\theta} c + \Lambda_A(\tilde{\theta}) & \text{if } \theta > \tilde{\theta} \end{cases} \tag{11}$$

where $\tilde{\theta}$ is such that $\Lambda_A'(\tilde{\theta}) = c$, *i.e.*, $\Lambda_A^*(c) = c\tilde{\theta} - \Lambda_A(\tilde{\theta})$.

Therefore, $\alpha_D(\theta) = \Lambda_D(\theta)/\theta$ is the effective bandwidth of the departure process.

We remark that the condition $\theta \le \tilde{\theta}$ is equivalent to $\alpha_A^*(\theta) \le c$ where

$$\alpha_A^*(\theta) = \arg \sup_{\alpha \in I\!R} \{\alpha\theta - \Lambda_A^*(\alpha)\} \text{ or } \alpha_A^*(\theta) = \Lambda_A'(\theta). \tag{12}$$

In [17], $\alpha_A^*(\theta)$ is called the *decoupling bandwidth* of the arrival process $\{a(t), t \in I\!N\}$.

For reasons that will be clear later, we are primarily interested in *stationary* G/D/1/∞ queueing system. More specifically, we assume the backlog process of the system has reached its steady state, thus having the same distribution as $Q(\infty)$. We study the system at time 0 and look backward in time. Since the arrival process is stationary, this will have no effect on the assumptions (A1), (A2) and (A3). However, for easy reference, we re-state them from this point of view.

(A1′) The arrival process $\{a(-t), t \in I\!N\}$ is ergodic and stationary.

(A2′) For any $\theta \in I\!R$,

$$\lim_{t\to\infty} \frac{1}{t} \log E e^{\theta A(-t,0)} = \Lambda_A(\theta)$$

exists and $\Lambda_A(\theta)$ is differential.

(A3′) $\{a(-t), t \in I\!N\}$ is adapted to a filtration $\{\mathcal{F}^A_{-t}, t \in I\!N\}$ with the following property: for any $\theta \in I\!R$, there exists a function $\Gamma_A(\theta)$, $0 \le \Gamma_A(\theta) < \infty$ such that for any $s, t \in I\!N$,

$$\Lambda_A(\theta)t - \Gamma_A(\theta) \le \log E(e^{\theta A(-t-s,-s)} | \mathcal{F}^A_{-t-s}) \le \Lambda_A(\theta)t + \Gamma_A(\theta) \ \ a.s.$$

As $Q(0)$ has the same distribution as $Q(\infty)$, from (8), it can be proved that for any positive $\theta < \theta^*$,

$$E e^{\theta Q(0)} < \infty. \tag{13}$$

The following lemmas are instrumental in proving the main theorem of the paper regarding the GPS system, the proofs of which are included in the appendix.

Lemma 2 *Assume $Ea(0) < c$, then for any positive $\theta < \theta^*$,*

$$\lim_{t\to\infty} \frac{1}{t} \log E[e^{\theta(Q(-t)+A(-t,0))}] = \Lambda_A(\theta). \tag{14}$$

Let $x \wedge y = \min\{x, y\}$. For any $t \in I\!N$, define

$$D(-t) = [Q(-t) + A(-t, 0)] \wedge ct. \tag{15}$$

Lemma 3 *Assume $Ea(0) < c$. Then for any $\alpha \in I\!R$,*

$$\lim_{t\to\infty} \frac{1}{t} \log Pr\{D(-t)/t \ge \alpha\} = -\inf_{x\ge\alpha} \Lambda^*_D(x) \tag{16}$$

where

$$\Lambda^*_D(\alpha) = \begin{cases} \Lambda^*_A(\alpha) & \text{if } \alpha \le c \\ \infty & \text{otherwise.} \end{cases} \tag{17}$$

Moreover, let $\tilde{\theta}$ be such that $\Lambda'_A(\tilde{\theta}) = c$. Then for any $\theta \ge 0$,

$$\lim_{t\to\infty} \frac{1}{t} \log E[e^{\theta D(-t)}] = \Lambda_D(\theta) \tag{18}$$

where

$$\Lambda_D(\theta) = \sup_{\alpha \ge Ea(0)} \{\theta\alpha - \Lambda^*_D(\alpha)\} = \begin{cases} \Lambda_A(\theta) & \text{if } 0 \le \theta \le \tilde{\theta} \\ \theta c - \tilde{\theta} c + \Lambda_A(\tilde{\theta}) & \text{if } \theta > \tilde{\theta}. \end{cases} \tag{19}$$

Remark 4 *From the convexity of $\Lambda^*_A(\theta)$ and the definition of $\tilde{\theta}$, $\Lambda^*_A(c) = \tilde{\theta} c - \Lambda_A(\tilde{\theta})$. Thus $\theta c - \tilde{\theta} c + \Lambda_A(\tilde{\theta}) = \theta c - \Lambda^*_A(c)$, and the condition $\theta \le \tilde{\theta}$ is equivalent to $\alpha^*_A(\theta) \le c$ where $\alpha^*_A(\theta)$ is the decoupling bandwidth of A defined in (12).*

For any $t = 1, 2, \ldots$, define the scaled process $A^{(t)}(s) = \frac{1}{t}A(-[ts], 0)$, $0 \le s \le 1$. Let

$$B(-t) = \min_{0 \le \tau \le t}\{A(-\tau, 0) + c(t-\tau)\} = t \min_{0 \le s \le 1}\{A^{(t)}(s) + c(1-s)\}. \tag{20}$$

Lemma 5 *$\{B(-t)/t, t \in I\!N\}$ satisfies the large deviation principle with the rate function $\Lambda_B^*(x)$ defined as follows: If $Ea(0) < c$, then*

$$\Lambda_B^*(x) = \begin{cases} \Lambda_A^*(x) & \text{if } x \le c \\ \infty & \text{otherwise} \end{cases} \tag{21}$$

and if $Ea(0) \ge c$, then

$$\Lambda_B^*(x) = \begin{cases} 0 & \text{if } x = c \\ \infty & \text{otherwise.} \end{cases} \tag{22}$$

3 Discrete-Time, Two-Queue GPS Systems

In this section, we consider a two-queue GPS system (Figure 1). Let c be the service rate of the GPS server and $\{\phi_i\}_{i=1,2}$ the GPS assignment for the two sessions sharing the GPS server such that $\phi_1 + \phi_2 = 1$ and $\phi_i \ge 0$, $i = 1, 2$. For any time t, let $a_i(t)$ denote the amount of arrival from session i to queue i at time t, and for any time interval $[\tau, t)$, let $A(\tau, t) = \sum_{s=\tau}^{t-1} a_i(s)$ denote the total amount of arrival during $[\tau, t)$. Similarly, let $b_i(t)$ denote the amount of service session i received at time t and $S_i(\tau, t) = \sum_{s=\tau}^{t-1} b_i(s)$ the total amount of service session i received during $[\tau, t)$. The backlog of queue i at time t is denoted by $Q_i(t)$. From the definition of GPS scheduling, if session i is busy throughout $[\tau, t)$ (*i.e.*, $Q_i(s) \ne 0$ for $s \in [\tau, t)$), then $S_i(\tau, t) \ge \phi_i c(t-\tau)$. In other words, session i is guaranteed a service rate of $\phi_i c$ whenever it is busy.

Given that the arrival processes $\{a_i(t)\}$, $i = 1, 2$, are stationary, and that the stability condition, $Ea_1(0) + Ea_2(0) < c$, is satisfied, the two-queue GPS system is stable. In particular, the queue length process $Q_i(t)$ tends to a finite random variable Q_i *a.s.*, as $t \to \infty$. In the following exposition, we consider the *stationary* two-queue GPS system, *i.e.*, the system has reached its steady state. In particular, we assume the queue length distribution Q_i of each queue has reached its steady state at time 0 (hence it has the same distribution as Q_i). We examine the system at time 0 and look backward in time. To derive upper and lower bounds for the two-queue GPS system, we make the following assumptions on the arrival processes [5]: for $i = 1, 2$,

(A1′) The arrival process $\{a_i(-t), t \in I\!N\}$ is ergodic and stationary.

(A2′) For any $\theta \in I\!R$,

$$\lim_{t \to \infty} \frac{1}{t} \log E e^{\theta A_i(-t, 0)} = \Lambda_{A_i}(\theta)$$

exists and $\Lambda_{A_i}(\theta)$ is differential.

(A3′) $\{a_i(-t), t \in I\!N\}$ is adapted to a filtration $\{\mathcal{F}_{-t}^{A_i}, t \in I\!N\}$ with the following property: for any $\theta \in I\!R$, there exists a function $\Gamma_{A_i}(\theta)$, $0 \le \Gamma_{A_i}(\theta) < \infty$ such that for any $s, t \in I\!N$,

$$\Lambda_{A_i}(\theta)t - \Gamma_{A_i}(\theta) \le \log E(e^{\theta A_i(-t-s, -s)} | \mathcal{F}_{-t-s}^{A_i}) \le \Lambda_{A_i}(\theta)t + \Gamma_{A_i}(\theta) \text{ a.s.}$$

[5] The time index used reflects the point of view of looking backward in time. Recall that the set of assumptions (A1′), (A2′) and (A3′) is equivalent to (A1), (A2) and (A3).

Under this set of assumptions, we have that

Theorem 6 (*cf.* [16]) *Suppose that $\{a_i(-t), t \in IN\}$, $i = 1, 2$, are independent and satisfy (A1′), (A2′) and (A3′). Moreover, assume that the stability condition $Ea_1(0) + Ea_2(0) < c$ is satisfied. Let $\alpha_{A_i}(\theta) = \Lambda_{A_i}(\theta)/\theta$ be the effective bandwidth of the arrival process $\{a_i(t), t \in IN\}$. Define*

$$\alpha_{D_i}(\theta) = \begin{cases} \alpha_{A_i}(\theta) & \text{if } Ea_i(0) < \phi_i c \text{ and } \theta \le \tilde{\theta}_i \\ \phi_i c - \frac{1}{\theta}(\tilde{\theta}_i \phi_i c - \Lambda_{A_i}(\tilde{\theta}_i)) & \text{if } Ea_i(0) < \phi_i c \text{ and } \theta > \tilde{\theta}_i \\ \phi_i c & \text{if } Ea_i(0) \ge \phi_i c \end{cases}$$

where $\tilde{\theta}_i$ is such that $\Lambda'_{A_i}(\tilde{\theta}_i) = \phi_i c$.
Then, for $i = 1, 2$,

$$\lim_{x \to \infty} \frac{1}{x} \log Pr\{Q_i > x\} = -\theta_i^* \tag{23}$$

where θ_i^* is the unique solution to the following equation:

$$\alpha_{A_i}(\theta) + \alpha_{D_j}(\theta) = c$$

with $j = 2$ if $i = 1$ and $j = 1$ if $i = 2$.
In other words,

$$\lim_{x \to \infty} \frac{1}{x} \log Pr\{Q_i > x\} \le -\theta \quad \textit{iff} \quad \alpha_{A_i}(\theta) + \alpha_{D_j}(\theta) < c. \tag{24}$$

Remark: This theorem is first stated in [16] under weaker assumptions and proved there using an argument based on Loynes' construction [28]. However, to make the proof completely rigorous, the issue of stationarity of the departure process from each queue needs to be addressed when applying Loynes' construction. In this paper, instead of dealing with such technicality, we argue directly using the stationary version of the queue length process Q_i, $i = 1, 2$.

Without loss of generality, we prove (23) for queue 1. Then (24) becomes

$$\lim_{x \to \infty} \frac{1}{x} \log Pr\{Q_1(0) > x\} = -\theta_1^* \tag{25}$$

where θ_1^* is the unique solution to the equation $\alpha_{A_1}(\theta) + \alpha_{D_2}(\theta) = c$.

The proof of (25) is divided into two parts:

Upper Bound:

if $\alpha_{A_1}(\theta) + \alpha_{D_2}(\theta) < c$, then

$$\limsup_{x \to \infty} \log Pr\{Q_1(0) > x\} \le -\theta \tag{26}$$

thus

$$\limsup_{x \to \infty} \log Pr\{Q_1(0) > x\} \le -\theta_1^*.$$

Lower Bound:

$$\liminf_{x \to \infty} \log Pr\{Q_1(0) > x\} \ge -\theta_1^*. \tag{27}$$

We prove (26) in § 3.1 and (27) in § 3.2.

3.1 Proof of the Upper Bound

First observe that

$$Q_1(0) = \max_{t \in I\!N}\{A_1(-t,0) - S_1(-t,0)\} \tag{28}$$

where as $S_1(-t,0) = Q_1(-t) + A_1(-t,0) - Q_1(0)$, the maximum is attained whenever $Q_1(-t) = 0$.

In particular, choose t be such that queue 1 is not empty at any time between $-t$ and 0, *i.e*, $Q_1(-s) > 0$ for any integer s, $0 < s < t$. In other words, session 1 is busy throughout $[-t, 0)$. Hence, $S_1(-t,0) \geq \phi_1 ct$. But $S_1(-t,0) = ct - S_2(-t,0)$, we have

$$S_1(-t,0) = [ct - S_2(-t,0)] \vee \phi_1 ct$$

where $x \vee y = \max\{x, y\}$.

Therefore, (28) becomes

$$\begin{aligned} Q_1(0) &= \max_{t \in I\!N}\{A_1(-t,0) - [ct - S_2(-t,0)] \vee \phi_1 ct\} \\ &= \max_{t \in I\!N}\{A_1(-t,0) - S_2(-t,0) \wedge \phi_2 ct - ct\} \end{aligned} \tag{29}$$

where $x \wedge y = \min\{x, y\}$ and the last equality follows as $\phi_1 ct = ct - \phi_2 ct$.

Case 1: $Ea_2(0) < \phi_2 c$.

Consider a single queue system where the session 2 is serviced exclusively by a server of service rate $\phi_2 c$. In other words, the arrival process to the system is $\{a_2(t), t \in I\!N\}$. The assumption $Ea_2(0) < \phi_2 c$ ensures this single queue system is stable. Assume the system has already reached steady state, let $\bar{S}_2(-t,0)$ denote the amount of the service session 2 received in $(-t,0)$, and $\bar{Q}_2(-t)$ the backlog in the queue at time $-t$.

By the definition of GPS scheduling, whenever session 2 is busy (*i.e.*, queue 2 of the two-queue GPS system is not empty), the rate of service received by session 2 is at least $\phi_2 c$ in the two-queue GPS system. Hence, $\bar{Q}_2(-t) \geq Q_2(t)$ for any $t \in N$. Therefore,

$$\begin{aligned} S_2(-t,0) &= Q_2(-t) + A_2(-t,0) - Q_2(0) \\ &\leq Q_2(-t) + A_2(-t,0) \\ &\leq \bar{Q}_2(-t) + A_2(-t,0). \end{aligned} \tag{30}$$

From (29), we have

$$Q_1(0) \leq \max_{t \in I\!N}\{A_1(-t,0) + [\bar{Q}_2(-t) + A_2(-t,0)] \wedge \phi_2 ct - ct\}. \tag{31}$$

Let $\bar{D}_2(-t) = [\bar{Q}_2(-t) + A_2(-t,0)] \wedge \phi_2 ct$. From Lemma 2, for any $\theta > 0$,

$$\lim_{t \to \infty} \frac{1}{t} \log E e^{\theta \bar{D}_2(-t)} = \Lambda_{\bar{D}_2}(\theta) \tag{32}$$

where

$$\Lambda_{\bar{D}_2}(\theta) = \begin{cases} \Lambda_{A_2}(\theta) & \text{if } \theta \leq \tilde{\theta}_2 \\ \phi_2 c\theta - \phi_2 c\tilde{\theta}_2 + \Lambda_{A_2}(\tilde{\theta}_2) & \text{otherwise} \end{cases}$$

and $\tilde{\theta}_2$ is such that $\Lambda_2'(\tilde{\theta}_2) = \phi_2 c$.

This, together with the fact that

$$\lim_{t\to\infty}\frac{1}{t}\log Ee^{\theta A_1(-t,0)}=\Lambda_{A_1}(\theta) \tag{33}$$

yields that for any $\epsilon>0$, there exists a t_ϵ such that for any $t\geq t_\epsilon$, $t\in IN$,

$$Ee^{\theta A_1(-t,0)}\leq e^{(\Lambda_{A_1}(\theta)+\epsilon)t} \tag{34}$$

and

$$Ee^{\theta \bar{D}_2(-t,0)}\leq e^{(\Lambda_{\bar{D}_2}(\theta)+\epsilon)t}. \tag{35}$$

Now from (31), we have

$$\begin{aligned} Ee^{\theta Q_1(0)} &\leq \sum_{t\in IN} Ee^{\theta(A_1(-t,0)+\bar{D}_2(-t)-ct)} \\ &\leq C_\epsilon+\sum_{t\geq t_\epsilon} e^{t(\Lambda_{A_1}(\theta)+\epsilon)}e^{t(\Lambda_{\bar{D}_2}(\theta)+\epsilon)}e^{-tc\theta} \end{aligned} \tag{36}$$

where the last equality follows from (34) and (35). C_ϵ is a constant that depends on ϵ.
Note that if $e^{(\Lambda_{A_1}(\theta)+\Lambda_{\bar{D}_2}(\theta)+2\epsilon-c\theta)}<1$, then

$$\sum_{t\geq t_\epsilon} e^{t(\Lambda_{A_1}(\theta)+\epsilon)}e^{t(\Lambda_{\bar{D}_2}(\theta)+\epsilon)}e^{-tc\theta}=\frac{e^{t_\epsilon(\Lambda_{A_1}(\theta)+\Lambda_{\bar{D}_2}(\theta)+2\epsilon-c\theta)}}{1-e^{(\Lambda_{A_1}(\theta)+\Lambda_{\bar{D}_2}(\theta)+2\epsilon-c\theta)}}.$$

Therefore $Ee^{\theta Q_1(0)}<\infty$ if $e^{(\Lambda_{A_1}(\theta)+\Lambda_{\bar{D}_2}(\theta)+2\epsilon-c\theta)}<1$ or $\Lambda_{A_1}(\theta)+\Lambda_{\bar{D}_2}(\theta)+2\epsilon-c\theta<0$. By Chebyshev's Inequality, for any $x\geq 0$,

$$Pr\{Q_1(0)>x\}\leq e^{-\theta x}Ee^{\theta Q_1(0)}.$$

Thus if $\Lambda_{A_1}(\theta)+\Lambda_{\bar{D}_2}(\theta)+2\epsilon-c\theta<0$, then

$$\limsup_{x\to\infty}\frac{1}{x}\log Pr\{Q_1(0)>x\}\leq -\theta.$$

Taking $\epsilon\to 0$, and noting that $\alpha_{D_2}(\theta)=\alpha_{\bar{D}_2}(\theta)=\Lambda_{\bar{D}_2}(\theta)/\theta$, we have (26).
Case 2: $Ea_2(0)\geq\phi_2 c$.
This case is easier to prove. First note that from (29), we have

$$Q_1(0)\leq\max_{t\in IN}\{A_1(-t,0)-\phi_1 ct\}.$$

Hence, for any $\theta>0$,

$$Ee^{\theta Q_1(0)}\leq\sum_{t\in N}Ee^{\theta(A_1(-t,0)-\phi_1 ct)}.$$

Using (33), we can show that for any $\epsilon>0$, if $\Lambda_{A_1}(\theta)+\epsilon-\phi_1 c\theta<0$, then $Ee^{\theta Q_1(0)}$ is finite. Therefore, by Chebeshev's Inequality, if $\Lambda_{A_1}(\theta)+\epsilon-c\theta<0$,

$$\limsup_{x\to\infty}\frac{1}{x}\log Pr\{Q_1(0)>x\}\leq -\theta.$$

Taking $\epsilon\to 0$, and noting that $\alpha_{D_2}(\theta)=\phi_2 c$, we have (26). This completes the proof of the upper bound.

3.2 Proof of the Lower Bound

For any $t \in I\!N$, let $Q(t)$ denote the aggregate backlog of the two queues at time t, *i.e.*, $Q(t) = Q_1(t) + Q_2(t)$. Since GPS is work-conserving, we have that, at time $t = 0$,

$$Q(0) = \max_{t \in I\!N}\{A_1(-t,0) + A_2(-t,0) - ct\}. \tag{37}$$

Note that if the maximum in the above equation is attained at t, then $[-t, 0)$ is contained in the system busy period of the two-queue GPS system starting at $-t$, *i.e.*, $Q(-t) = 0$, but for any τ, $0 < \tau < t$, $Q(-\tau) > 0$. Note also that $Q(-t) = 0$ implies that $Q_1(-t) = Q_2(-t) = 0$.

Applying (29) to queue 2, we have

$$Q_2(0) = \max_{\tau \in I\!N}\{A_2(-\tau,0) + S_1(-\tau,0) \wedge \phi_1 c\tau - c\tau\}. \tag{38}$$

Again if the maximum is attained at τ, then $(-\tau, 0)$ is contained in the session 2 busy period starting at $-\tau$. We observe that if t maximize (37) and τ maximizes (38), we must have $\tau \leq t$. Therefore,

$$\begin{aligned}
Q_1(0) &= Q(0) - Q_2(0) \\
&= \max_{t \in I\!N}\{A_1(-t,0) + A_2(-t,0) - ct\} - Q_2(0) \\
&= \max_{t \in I\!N}\{A_1(-t,0) + A_2(-t,0) - ct \\
&\qquad - \max_{0 \leq \tau \leq t}\{A_2(-\tau,0) + S_1(-\tau,0) \wedge \phi_1 c\tau - c\tau\}\} \\
&\geq \max_{t \in I\!N}\{A_1(-t,0) + A_2(-t,0) - ct - \max_{0 \leq \tau \leq t}\{A_2(-\tau,0) - \phi_2 c\tau\}\} \\
&= \max_{t \in I\!N}\{A_1(-t,0) + \min_{0 \leq \tau \leq t}\{A_2(-t,-\tau) + \phi_2 c\tau\} - ct\}
\end{aligned}$$

where $A_2(-t,-\tau) = A_2(-t,0) - A_2(-\tau,0) = \sum_{s=-t}^{-\tau-1} a_2(s)$.

Since the arrival process, $\{a_2(-t), t \in I\!N\}$, is stationary, we can replace $A_2(-t,-\tau)$ by $A_2(-(t-\tau),0)$ without changing the associated probability distribution. Hence

$$Q_1(0) \geq \max_{t \in I\!N}\{A_1(-t,0) + \min_{0 \leq \tau \leq t}\{A_2(-(t-\tau),0) + \phi_2 c\tau\} - ct\}$$

In other words, for any $t \in I\!N$,

$$\begin{aligned}
Q_1(0) &\geq A_1(-t,0) + \min_{0 \leq \tau \leq t}\{A_2(-(t-\tau),0) + \phi_2 c\tau\} - ct \\
&= A_1(-t,0) + \min_{0 \leq \tau \leq t}\{A_2(-\tau,0) + \phi_2 c(t-\tau)\} - ct.
\end{aligned} \tag{39}$$

For any $x \geq 0$, let $t = \lfloor \frac{x}{\beta} \rfloor$ where $\beta > 0$ is a constant fixed temporarily. Then

$$\begin{aligned}
\liminf_{x \to \infty} \frac{1}{x} \log Pr\{Q_1(0) > x\} &= \frac{1}{\beta} \liminf_{t \in I\!N} \frac{1}{t} \log Pr\{Q_1(0) > \beta t\} \\
&\geq \frac{1}{\beta} \liminf_{t \in I\!N} \frac{1}{t} \log Pr\{A_1(-t,0) + \min_{0 \leq \tau \leq t}\{A_2(-\tau,0) + \phi_2 c(t-\tau)\} - ct > \beta t\} \\
&\geq \frac{1}{\beta} \liminf_{t \in I\!N} \frac{1}{t} \log Pr\{A_1(-t,0)/t + \bar{B}_2(-t)/t > c + \beta\}
\end{aligned} \tag{40}$$

where $\bar{B}_2(-t) = \min_{0\le\tau\le t}\{A_2(-\tau,0) + \phi_2 c(t-\tau)\}$.

From Lemma 5, $\{\bar{B}_2(-t)/t, t \in I\!N\}$ satisfies the LDP with the rate function $\Lambda^*_{\bar{B}_2}(x)$ as defined in Lemma 5. Moreover, $\{A_1(-t,0)/t, t \in I\!N\}$ also satisfies the LDP with the rate function $\Lambda^*_{A_2}(x)$. Hence, by the Contraction Principle (see, [14]), we have

$$\begin{aligned}&\liminf_{x\to\infty}\frac{1}{x}\log Pr\{Q_1(0) > x\}\\ &\quad\ge -\frac{1}{\beta}\inf_{\{(\alpha_1,\alpha_2):\alpha_1+\alpha_2=c+\beta\}}\{\Lambda^*_{A_1}(\alpha_1)+\Lambda^*_{\bar{B}_2}(\alpha_2)\}.\end{aligned}$$

As $\beta > 0$ is arbitrary,

$$\begin{aligned}&\liminf_{x\to\infty}\frac{1}{x}\log Pr\{Q_1(0) > x\}\\ &\quad\ge -\inf_{\beta>0}\inf_{\{(\alpha_1,\alpha_2):\alpha_1+\alpha_2=c+\beta\}}\left\{\frac{\Lambda^*_{A_1}(\alpha_1)+\Lambda^*_{\bar{B}_2}(\alpha_2)}{\beta}\right\}\\ &\quad= -\inf_{\{(\alpha_1,\alpha_2):\alpha_1+\alpha_2>c\}}\left\{\frac{\Lambda^*_{A_1}(\alpha_1)+\Lambda^*_{\bar{B}_2}(\alpha_2)}{\alpha_1+\alpha_2-c}\right\}\end{aligned}$$

We claim that

$$\inf_{\{(\alpha_1,\alpha_2):\alpha_1+\alpha_2>c\}}\left\{\frac{\Lambda^*_{A_1}(\alpha_1)+\Lambda^*_{\bar{B}_2}(\alpha_2)}{\alpha_1+\alpha_2-c}\right\} = \theta^*_1, \tag{41}$$

then (27) holds.

For any $\alpha \in I\!R$, define

$$I(\alpha) = \inf_{\left\{\begin{array}{c}\alpha_1,\alpha_2\in I\!R\\ \alpha_1+\alpha_2=\alpha\end{array}\right\}}\{\Lambda^*_{A_1}(\alpha_1)+\Lambda^*_{\bar{B}_2}(\alpha_2)\}$$

and let $I^*(\theta)$ be the Legendre-Fenchel transform of $I(\alpha)$, *i.e.*, $I^*(\theta) = \sup_{\alpha\in I\!R}\{\alpha\theta - I(\alpha)\}$. It is easy to see that $I^*(\theta) = \Lambda_{A_1}(\theta) + \Lambda_{\bar{B}_2}(\theta)$ where $\Lambda_{\bar{B}_2}(\theta) = \sup_{\alpha\in I\!R}\{\alpha\theta - \Lambda^*_{\bar{B}_2}(\alpha)\}$. In particular, let $\tilde{\theta}_2$ be such that $\Lambda'_{A_2}(\tilde{\theta}_2) = \phi_2 c$. Then for $Ea_2(0) < c$,

$$\Lambda_{\bar{B}_2}(\theta) = \left\{\begin{array}{ll}\Lambda_{A_2}(\theta) & \text{if } \theta \le \tilde{\theta}_2\\ \phi_2 c\theta - \phi_2 c\tilde{\theta}_2 + \Lambda_{A_2}(\tilde{\theta}_2) & \text{otherwise}\end{array}\right.$$

and for $Ea_2(0) \ge c$, $\Lambda_{\bar{B}_2}(\theta) = \phi_2 c\theta$. From the definition of $\alpha_{D_2}(\theta)$, we see that in either case, $\alpha_{D_2}(\theta) = \frac{\Lambda_{\bar{B}_2}(\theta)}{\theta}$. Clearly $\theta^*_1 = \sup_{\theta\in I\!R}\{\alpha_{A_1}(\theta) + \alpha_{D_2}(\theta) \le c\} = \sup_{\theta\in I\!R}\{I^*(\theta) \le c\theta\}$. To show (41), we note that

$$\inf_{\{(\alpha_1,\alpha_2):\alpha_1+\alpha_2>c\}}\left\{\frac{\Lambda^*_{A_1}(\alpha_1)+\Lambda^*_{\bar{B}_2}(\alpha_2)}{\alpha_1+\alpha_2-c}\right\} = \inf_{\alpha>c}\left\{\frac{I(\alpha)}{\alpha-c}\right\}. \tag{42}$$

Then, for any θ such that $I^*(\theta) \le c\theta$, $I(\alpha) \ge \theta\alpha - I^*(\theta) \ge \theta(\alpha - c)$. Hence,

$$\inf_{\alpha>c}\left\{\frac{I(\alpha)}{\alpha-c}\right\} \ge \theta.$$

Since the above inequality is true for any θ such that $I^*(\theta) \leq c\theta$, we have

$$\inf_{\alpha>c}\left\{\frac{I(\alpha)}{\alpha-c}\right\} \geq \theta_1^*.$$

Now let $\alpha^* = I^{*'}(\theta_1^*) = \Lambda'_{A_1}(\theta_1^*) + \Lambda'_{B_2}(\theta_1^*)$, then $I(\alpha^*) = \alpha^*\theta_1^* - I^*(\theta_1^*) > 0$. But, from the definition of θ_1^*, we have $I^*(\theta_1^*) = \theta_1^* c$, therefore

$$\inf_{\alpha>c}\left\{\frac{I(\alpha)}{\alpha-c}\right\} \leq \frac{I(\alpha^*)}{\alpha^*-c} = \theta_1^*.$$

Hence (41) holds. This completes the proof of the lower bound.

4 Conclusion

In this part of the paper, we prove an exact bound on the asymptotic decay rate of the queue length tail distribution for the two-queue GPS system. Our proof uses the sample-path large deviation principle, thus avoiding the subtle technical pitfall faced by the work of [16]. However, a stronger technical assumption on the arrival processes are needed. Upper and lower bounds for the general multiple-queue GPS system is presented separately in the second part of the paper.

We have looked at the discrete-time GPS system. The results of the paper may be extended to the continuous-time model by imposing appropriate conditions (corresponding to (A1), (A2) and (A3)) on the continuous-time arrival processes. Then the arguments of this paper can be applied to pass from the discrete case to the continuous case (*cf.*, the proof of Theorem 5.1.19 in [14]). Methods, for instance, employed in [23, 4], may also be used to establish results for the continuous-time GPS system.

The paper deals only with the large buffer asymptotics under the GPS scheduling. Another future direction is to study the asymptotical behavior of the GPS scheduling with a large number of sources *à la* the methods of [37, 5].

Acknowledgement I am indebted to Prof. Richard Ellis for teaching me *Probability Theory* and *Large Deviation Theory* and to my advisor Don Towsley for encouragement and many helpful discussions.

A Proofs of the Lemmas in Section 2

Proof of Lemma 2:

$$\begin{aligned} & E\left[e^{\theta(Q(-t)+A(-t,0))}\right] \\ &= E\left[E\left[e^{\theta(Q(-t)+A(-t,0))}|\mathcal{F}_{-t}^A\right]\right] \\ &= E\left[e^{\theta Q(-t)}E\left[e^{\theta A(-t,0)}|\mathcal{F}_{-t}^A\right]\right] \end{aligned}$$

From (A2′),

$$e^{\Lambda_A(\theta)t-\Gamma_A(\theta)} \leq E\left[e^{\theta A(-t,0)}|\mathcal{F}_{-t}^A\right] \leq e^{\Lambda_A(\theta)t+\Gamma_A(\theta)}. \tag{43}$$

Thus,

$$e^{\Lambda_A(\theta)t-\Gamma_A(\theta)}E\left[e^{\theta Q(-t)}\right] \leq E\left[e^{\theta(Q(-t)+A(-t,0))}\right] \leq e^{\Lambda_A(\theta)t+\Gamma_A(\theta)}E\left[e^{\theta Q(-t)}\right]. \tag{44}$$

Since we assume that the queue is in its steady state, for any t, $Q(-t) =_{st} Q(0)$, i.e., $Q(-t)$ and $Q(0)$ have the same stationary distribution, *i.e.*, that of $Q(\infty)$ in (8). (13) implies that $E[e^{\theta Q(-t)}] < \infty$ for any $\theta < \theta^*$. This together with (44) yields (14). ∎

Proof of Lemma 3: The lower bound parts of (16) and (18), namely, for any $\alpha \in I\!R$,

$$\liminf_{t\to\infty} \frac{1}{t} \log Pr\{D(-t)/t \geq \alpha\} \geq - \inf_{x \geq \alpha} \Lambda_D^*(x) \tag{45}$$

and

$$\liminf_{t\to\infty} \frac{1}{t} \log E[e^{\theta D(-t)}] \geq \Lambda_D(\theta) \tag{46}$$

are easy to prove. For any $t \in I\!N$, as $Q(-t) \geq 0$,

$$D(-t)/t \geq [A(-t,0)/t] \wedge c. \tag{47}$$

Since $\{A(-t,0)/t, t \in I\!N\}$ satisfies the large deviation principle with the rate function $\Lambda_A^*(\alpha)$ and $x \wedge c$ is a continuous function in x, by the Contraction Principle (see, e.g., Theorem 4.2.1 in [14]), it is easy to see that $\{(A(-t,0)/t) \wedge c, t \in I\!N\}$ also satisfies the large deviation principle with the rate function $\Lambda_D^*(\alpha)$. In particular, for any $\alpha \in I\!R$,

$$\begin{aligned}
&\liminf_{t\to\infty} \frac{1}{t} \log Pr\{[A(-t,0)/t] \wedge c \geq \alpha\} \\
&\quad \geq \liminf_{t\to\infty} \frac{1}{t} \log Pr\{[A(-t,0)/t] \wedge c > \alpha\} \\
&\quad \geq - \inf_{x > \alpha} \Lambda_D^*(x).
\end{aligned}$$

If $\alpha > c$, clearly $\inf_{x>\alpha} \Lambda_D^*(x) = \inf_{x\geq\alpha} \Lambda_D^*(x)$. Otherwise, $\inf_{x>\alpha} \Lambda_D^*(x) = \inf_{x\geq\alpha} \Lambda_D^*(x)$ follows from the continuity of $\Lambda_A^*(\alpha)$. Therefore, (45) holds.

To prove (46), note that for any $\theta \geq 0$, $\theta([A(-t,0)/t] \wedge c)$ is bounded above by θc. Applying Varadhan's Integral Lemma (see, e.g., Theorem 4.3.1 in [14]) yields that

$$\lim_{t\to\infty} \frac{1}{t} \log E[e^{\theta(A(-t,0)\wedge ct)}] = \sup_{\alpha\in I\!R} \{\theta\alpha - \Lambda_D^*(\alpha)\} \geq \sup_{\alpha \geq Ea(0)} \{\theta\alpha - \Lambda_D^*(\alpha)\} = \Lambda_D(\theta).$$

For the upper bound parts, we first prove that for any $\alpha \in I\!R$,

$$\limsup_{t\to\infty} \frac{1}{t} \log Pr\{D(-t)/t \geq \alpha\} \leq - \inf_{x \geq \alpha} \Lambda_D^*(x). \tag{48}$$

Case 1: $\alpha > c$. Then clearly $Pr\{D(-t)/t \geq \alpha\} = 0$. But in this case, $\inf_{x\geq\alpha} \Lambda_D^*(x) = \Lambda_D^*(\alpha) = \infty$, hence (48) holds trivially.

Case 2: $\alpha \leq m$. As $Pr\{D(-t)/t \geq \alpha\} \leq 1$ and $\Lambda_D^*(m) = \Lambda_A^*(m) = 0$ is the infimum of $\Lambda_A^*(\alpha)$ over $I\!R$, (48) holds trivially as well.

Case 3: $m < \alpha \leq c$, where $m = Ea(0)$. By Chebyshev's Inequality, for any $\theta \geq 0$,

$$\begin{aligned}
Pr\{D(-t)/t \geq \alpha\} &= Pr\{Q(-t) + A(-t,0) \geq \alpha t\} \\
&\leq e^{-\theta\alpha t} E[e^{\theta(Q(-t)+A(-t,0))}].
\end{aligned}$$

For any $0 < \theta < \theta^*$, by Lemma 2,

$$\begin{aligned}\limsup_{t\to\infty}\frac{1}{t}\log Pr\{D(-t)/t \geq \alpha\} &\leq -\theta\alpha + \limsup_{t\to\infty}\frac{1}{t}\log E[e^{\theta(Q(-t)+A(-t,0))}] \\ &= -\theta\alpha + \Lambda_A(\theta). \end{aligned} \tag{49}$$

Since $0 < \theta < \theta^*$ is arbitrary, we have

$$\limsup_{t\to\infty}\frac{1}{t}\log Pr\{D(-t)/t \geq \alpha\} \leq -\sup_{0<\theta<\theta^*}\{\theta\alpha + \Lambda_A(\theta)\}.$$

We claim that for $\alpha \in (m, c]$,

$$\sup_{0<\theta<\theta^*}\{\theta\alpha - \Lambda_A(\theta)\} = \Lambda_A^*(\alpha) = \Lambda_D^*(\alpha). \tag{50}$$

First note that by continuity of $\Lambda_A(\theta)$, $\sup_{0<\theta<\theta^*}\{\theta\alpha - \Lambda_A(\theta)\} = \sup_{0\leq\theta\leq\theta^*}\{\theta\alpha - \Lambda_A(\theta)\}$. By strict convexity of $\Lambda_A^*(\alpha)$, $\Lambda_A^*(\alpha) = \alpha\theta_\alpha - \Lambda_A(\theta_\alpha)$ if and only if $\alpha = \Lambda_A'(\theta_\alpha)$. Since $\Lambda_A'(\theta)$ is increasing and $\Lambda_A'(0) = m$ and $\Lambda_A'(\tilde{\theta}) = c$, we have that for $\alpha \in (m, c]$, $0 \leq \theta_\alpha \leq \tilde{\theta}$. As $\Lambda_A^*(c) > \Lambda_A^*(m) = 0$, $\Lambda_A(\tilde{\theta}) = \tilde{\theta}c - \Lambda_A^*(c) < \tilde{\theta}c$, hence $\tilde{\theta} < \theta^* = \sup\{\theta : \Lambda_A(\theta) < c\theta\}$. Therefore, $0 \leq \theta_\alpha \leq \theta^*$. This proves the claim (50).

We now proceed to prove that

$$\limsup_{t\to\infty}\frac{1}{t}\log E[e^{\theta D(-t)}] \leq \Lambda_D(\theta) \tag{51}$$

.

Since $E[e^{\theta D(-t)}] = \int_{-\infty}^{\infty} Pr\{e^{\theta D(-t)} \geq y\}dy$, by a change of variable, $y = e^{\theta\alpha t}$, we have that

$$\begin{aligned}E[e^{\theta D(-t)}] &= \int_{-\infty}^{\infty} Pr\{D(-t) \geq \alpha t\}\theta t e^{\theta\alpha t}d\alpha \\ &= \int_{-\infty}^{m} Pr\{D(-t) \geq \alpha t\}\theta t e^{\theta\alpha t}d\alpha + \int_{m^+}^{c} Pr\{D(-t) \geq \alpha t\}\theta t e^{\theta\alpha t}d\alpha \\ &+ \int_{c^+}^{\infty} Pr\{D(-t) \geq \alpha t\}\theta t e^{\theta\alpha t}d\alpha.\end{aligned}$$

As $Pr\{D(-t) \geq \alpha t\} = 0$ for $\alpha > c$, the last integral vanishes. Let $Z_1(t)$ denote the first integral, and $Z_2(t)$ denote the second. Then

$$\begin{aligned}\limsup_{t\to\infty}\frac{1}{t}\log E[e^{\theta D(-t)}] &= \limsup_{t\to\infty}\frac{1}{t}\log(Z_1(t) + Z_2(t)) \\ &= \max\left\{\limsup_{t\to\infty}\frac{1}{t}\log Z_1(t), \limsup_{t\to\infty}\frac{1}{t}\log Z_2(t)\right\}. \end{aligned} \tag{52}$$

We will show that

$$\limsup_{t\to\infty}\frac{1}{t}\log Z_1(t) \leq \theta m - \Lambda_D^*(m) \tag{53}$$

and

$$\limsup_{t\to\infty}\frac{1}{t}\log Z_2(t) \leq \sup_{\alpha\geq m}\{\theta\alpha - \Lambda_D^*(\alpha)\}. \tag{54}$$

Then, from (52), we have

$$\begin{aligned}\limsup_{t\to\infty}\frac{1}{t}\log E[e^{\theta D(-t)}] &\leq \max\left\{\theta m-\Lambda_D^*(m),\sup_{\alpha\geq m}\{\theta\alpha-\Lambda_D^*(\alpha)\}\right\}\\ &= \sup_{\alpha\geq m}\{\theta\alpha-\Lambda_D^*(\alpha)\}=\Lambda_D(\theta).\end{aligned}$$

To prove (53), note that $Pr\{D(-t)\geq\alpha t\}\leq 1$, thus

$$\limsup_{t\to\infty}\frac{1}{t}\log Z_1(t)\leq\limsup_{t\to\infty}\frac{1}{t}\log\int_{-\infty}^{m}\theta te^{\theta\alpha t}d\alpha=\theta m=\theta m-\Lambda_D^*(m)$$

as $\Lambda_D^*(m)=0$.

To prove (54), we consider two cases $0\leq\theta\leq\tilde{\theta}$ and $\theta>\tilde{\theta}$ separately.

Case 1: $0\leq\theta\leq\tilde{\theta}$.

$$\begin{aligned}&\limsup_{t\to\infty}\frac{1}{t}\log Z_2(t)\\ &\quad=\limsup_{t\to\infty}\frac{1}{t}\log\int_{m^+}^{c}Pr\{Q(-t)+A(-t,0)\geq\alpha t\}\theta te^{\theta\alpha t}d\alpha\\ &\quad\leq\limsup_{t\to\infty}\frac{1}{t}\log\int_{m^+}^{c}e^{-\theta\alpha t}E[e^{\theta(Q(-t)+A(-t,0))}]\theta te^{\theta\alpha t}d\alpha\\ &\quad=\limsup_{t\to\infty}\frac{1}{t}\log(\theta t(c-m))+\limsup_{t\to\infty}\frac{1}{t}\log E[e^{\theta(Q(-t)+A(-t,0))}]\\ &\quad=\Lambda_A(\theta)=\Lambda_D(\theta)\end{aligned}\tag{55}$$

where (55) follows from Chebyshev's Inequality and (55) from Lemma 2.

Case 2: $\theta>\tilde{\theta}$.

$$\begin{aligned}&\limsup_{t\to\infty}\frac{1}{t}\log Z_2(t)\\ &\quad=\limsup_{t\to\infty}\frac{1}{t}\log\int_{m^+}^{c}Pr\{Q(-t)+A(-t,0)\geq\alpha t\}\theta te^{\theta\alpha t}d\alpha\\ &\quad\leq\limsup_{t\to\infty}\frac{1}{t}\log\int_{m^+}^{c}e^{-\tilde{\theta}\alpha t}E[e^{\tilde{\theta}(Q(-t)+A(-t,0))}]\theta te^{\theta\alpha t}d\alpha\\ &\quad=\limsup_{t\to\infty}\frac{1}{t}\log E[e^{\theta(Q(-t)+A(-t,0))}]+\limsup_{t\to\infty}\frac{1}{t}\log(\theta t\int_{m^+}^{c}e^{(\theta-\tilde{\theta})\alpha t}d\alpha)\\ &\quad=\Lambda_A(\tilde{\theta})+(\theta-\tilde{\theta})c\\ &\quad=\Lambda_D(\theta)\end{aligned}$$

where Chebyshev's Inequality and Lemma 2 are used again in the derivation. ∎

Proof of Lemma 5:

The Upper Bound:

It suffices to prove that for any F, $F = [-\alpha, \infty)$, or $F = (-\infty, \alpha]$, where $\alpha \in I\!R$,

$$\limsup_{t\to\infty} \frac{1}{t} \log Pr\{B(-t)/t \in F\} \leq -\inf_{x\in F} \Lambda_B^*(x) \tag{56}$$

We first prove (56) for $F = [\alpha, \infty)$.

if $\alpha > c$, then $Pr\{B(-t)/t \geq \alpha\} = 0$, thus

$$\limsup_{t\to\infty} \frac{1}{t} \log Pr\{B(-t)/t \geq \alpha\} = -\inf_{x\geq\alpha} \Lambda_B^*(x).$$

If $\alpha \leq c$, as $Pr\{B(-t)/t \geq \alpha\} \leq Pr\{A(-t,0)/t \geq \alpha\}$, using the fact that $\{A(-t,0)/t\}$ satisfies the LDP with the rate function $\Lambda_A^*(x)$, we have

$$\limsup_{t\to\infty} \frac{1}{t} \log Pr\{B(-t)/t \geq \alpha\} \leq \limsup_{t\to\infty} \frac{1}{t} \log Pr\{A(-t,0)/t \geq \alpha\} \leq -\inf_{x\geq\alpha} \Lambda_A^*(x).$$

Note that if $Ea(0) \geq c$, then $\alpha \leq c$ implies that $\alpha \leq Ea(0)$, hence $\inf_{x\geq\alpha} \Lambda_A^*(x) = 0 = \inf_{x\geq\alpha} \Lambda_B^*(x)$. If $Ea(0) < c$, then clearly, $\inf_{x\geq\alpha} \Lambda_A^*(x) = \inf_{x\geq\alpha} \Lambda_B^*(x)$. In either case, (56) holds for $F = [\alpha, \infty)$.

For $F = (-\infty, \alpha]$, we consider the cases $Ea(0) < c$ and $Ea(0) \geq c$ separately.

In the case that $Ea(0) \geq c$, note that for any $\alpha \in I\!R$, by the definition of $\Lambda_B(x)$, $\inf_{x\in F} \Lambda_B(x) = \inf_{x\leq\alpha} \Lambda_B(x) = 0$. But on the other hand, $Pr\{B(-t)/t \in F\} = Pr\{B(-t)/t \leq \alpha\} \leq 1$. Hence the upper bound (56) holds trivially in this case.

The case that $Ea(0) < c$ is a little harder. Note first that if $\alpha \geq Ea(0)$, then $\inf_{x\leq\alpha} \Lambda_B(x) = \Lambda_B(Ea(0)) = 0$. Again as $Pr\{B(-t)/t \leq \alpha\} \leq 1$, the upper bound (56) holds trivially when $\alpha \geq Ea(0)$.

If $\alpha < 0$, then $\Lambda_B^*(\alpha) = \Lambda_A^*(\alpha) = \infty$. But $Pr\{B(-t)/t \leq \alpha\} = 0$, hence (56) also holds trivially when $\alpha < 0$.

We now consider $0 \leq \alpha < Ea(0)$. Since $Ea(0) < c$, $\alpha < c$. Observe that

$$\begin{aligned} Pr\{B(-t)/t \leq \alpha\} &= Pr\{\min_{0\leq s\leq 1}\{A^{(t)}(s) + c(1-s)\} \leq \alpha\} \\ &= Pr\{\min_{0\leq s\leq 1}\{A^{(t)}(s) - \alpha s + (c-\alpha)(1-s)\} \leq 0\} \\ &\leq Pr\{\min_{0\leq s\leq 1}\{A^{(t)}(s) - \alpha s \leq 0\}\} \end{aligned}$$

Since $\{A^{(t)}(s), 0 \leq s \leq 1, t \in I\!N\}$ satisfies the sample path large deviation principle, we have that

$$\begin{aligned} &\limsup_{t\to\infty} \frac{1}{t} \log Pr\{B(-t)/t \leq \alpha\} \\ &\quad\leq \limsup_{t\to\infty} \frac{1}{t} \log Pr\{\min_{0\leq s\leq 1}\{A^{(t)}(s) - \alpha s \leq 0\}\} \\ &\quad\leq -\inf_{\{\phi\,:\,\min_{0\leq s\leq 1}\{\phi(s)-\alpha s\}\leq 0\}} \int_0^1 \Lambda_A^*(\phi'(s))ds \end{aligned}$$

where ϕ is an absolutely continuous function on $[0,1]$ and $\phi(0) = 0$.

For any absolutely continuous function ϕ on $[0,1]$ such that $\min_{0\le s\le 1}\{\phi(s)-\alpha s\}\le 0$ and $\phi(0)=0$. Let s_0, $0\le s_0\le 1$, be such that $\phi(s_0)\le \alpha s_0$. Since Λ_A^* is nonnegative and convex, we have

$$\int_0^1 \Lambda_A^*(\phi'(s))ds \ge \int_0^{s_0} \Lambda_A^*(\phi'(s))ds \ge \Lambda_A^*(\int_0^{s_0}\phi'(s)ds) = \Lambda_A^*(\phi(s_0)).$$

Since $0\le \alpha < Ea(0)$ and $\phi(s_0)\le \alpha s_0\le \alpha$, by convexity of Λ_A^*, we have $\Lambda_A^*(\phi(s_0))\ge \Lambda_A^*(\alpha s_0)\ge \Lambda_A^*(\alpha)$. Therefore,

$$\inf_{\{\phi:\,\min_{0\le s\le 1}\{\phi(s)-\alpha s\}\le 0\}}\int_0^1 \Lambda_A^*(\phi'(s))ds \ge \Lambda_A^*(\alpha).$$

This completes the proof of the upper bound (56) for $F=(-\infty,\alpha]$.

The Lower Bound:

For any $\alpha\in I\!R$, let $G_\delta=(\alpha-\delta,\alpha+\delta)$ denote a neighborhood of α of radius $\delta>0$. It suffices to prove that

$$\liminf_{t\to\infty}\frac{1}{t}\log Pr\{B(-t)/t\in G_\delta\}\ge -\Lambda_B^*(\alpha). \tag{57}$$

First note that if $c<\alpha$, then by the definition of Λ_B^*, $\Lambda_B^*(\alpha)=\infty$, hence the lower bound (57) holds trivially. Moreover, if $\alpha<c$ but $Ea(0)\ge c$, then again $\Lambda_B^*(\alpha)=\infty$, (57) holds also trivially in this case.

Therefore we assume that $\alpha\le c$ if $Ea(0)<c$, and $\alpha=c$ if $Ea(0)\ge c$. Let δ' be any real number such that $0<\delta'<\delta$. Hence $\alpha-\delta<\alpha-\delta'<c$. As $G_\delta\supset[\alpha-\delta',\alpha+\delta']$,

$$\begin{aligned} Pr\{B(-t)/t\in G_\delta\} &> Pr\{B(-t)/t\in[\alpha-\delta',\alpha+\delta']\}\\ &= Pr\{\alpha-\delta'\le \min_{0\le s\le 1}\{A^{(t)}(s)+c(1-s)\}\le \alpha+\delta'\} \end{aligned} \tag{58}$$

The left condition on $\min_{0\le s\le 1}\{A^{(t)}(s)+c(1-s)\}$ in (58) can be tightened as follows:

$$\begin{aligned} &\left\{\min_{0\le s\le 1}\{A^{(t)}(s)+c(1-s)\}\ge \alpha-\delta'\right\}\\ &\quad\supseteq \{A^{(t)}(s)\ge(\alpha-\delta')s, 0\le s\le 1\}\cap\{c(1-s)\ge(\alpha-\delta')(1-s), 0\le s\le 1\}\\ &\quad= \{A^{(t)}(s)\ge(\alpha-\delta')s, 0\le s\le 1\} \end{aligned} \tag{59}$$

where last equality holds as $\alpha-\delta'<c$.

To deal with the right condition on $\min_{0\le s\le 1}\{A^{(t)}(s)+c(1-s)\}$ in (58), we consider the cases $Ea(0)\ge c$ and $Ea(0)<c$ separately.

If $Ea(0)\ge c$, then $\alpha=c$. As $\min_{0\le s\le 1}\{A^{(t)}(s)+c(1-s)\}\le c=\alpha<\alpha+\delta'$, $\min_{0\le s\le 1}\{A^{(t)}(s)+c(1-s)\}\le\alpha+\delta'$ always holds. Therefore, from (58) and (59), we have

$$Pr\{B(-t)/t\in G_\delta\}> Pr\{A^{(t)}(s)\ge(\alpha-\delta')s, 0\le s\le 1\}$$

Now from the fact that $\{A^{(t)}(s), 0\le s\le 1, t\in I\!N\}$ satisfies the sample path large deviation principle, we have

$$\begin{aligned} &\liminf_{t\to\infty}\frac{1}{t}\log Pr\{B(-t)/t\in G_\delta\}\\ &\quad\liminf_{t\to\infty}\frac{1}{t}\log Pr\{A^{(t)}(s)\ge(\alpha-\delta')s, 0\le s\le 1\}\\ &\quad\ge -\inf_{\{\phi:\,\phi(s)\ge(\alpha-\delta')s,\,0\le s\le 1\}}\int_0^1\Lambda_A^*(\phi'(s))ds \end{aligned}$$

where ϕ is an absolutely continuous function from $[0,1]$ to $I\!R$ and $\phi(0)=0$.

Clearly $\inf_{\{\phi:\,\phi(s)\geq(\alpha-\delta')s,\,0\leq s\leq 1\}}\int_0^1\Lambda_A^*(\phi'(s))ds\geq 0$. Moreover, if $\phi(s)=Ea(0)s>(al-\delta')s, 0\leq s\leq 1$, then $\int_0^1\Lambda_A^*(\phi'(s))ds=\int_0^1\Lambda_A^*(Ea(0))ds=0$. Therefore, $\inf_{\{\phi:\,\phi(s)\geq(\alpha-\delta')s,\,0\leq s\leq 1\}}\int_0^1\Lambda_A^*(\phi'(s))ds=0$. But in this case, $\Lambda_B^*(\alpha)=\Lambda_B^*(c)=0$. Therefore, (57) holds.

Now we consider the case that $Ea(0)<c$. Since $\min_{0\leq s\leq 1}\{A^{(t)}(s)+c(1-s)\}\leq A^{(t)}(1)$,

$$\left\{\min_{0\leq s\leq 1}\{A^{(t)}(s)+c(1-s)\}\leq\alpha+\delta'\right\}\subseteq\{A^{(t)}(1)\leq\alpha+\delta'\}. \tag{60}$$

Hence, combining (59) and (60) with (58), and recalling that $\{A^{(t)}(s),0\leq s\leq 1,t\in I\!N\}$ satisfies the sample path large deviation principle, we have

$$\begin{aligned}
&\liminf_{t\to\infty}\frac{1}{t}\log Pr\{B(-t)/t\in G_\delta\}\\
&\quad>\liminf_{t\to\infty}\frac{1}{t}\log Pr\{A^{(t)}(s)\geq(\alpha-\delta')s,0\leq s\leq 1,A^{(t)}(1)\leq\alpha+\delta'\}\\
&\quad\geq-\inf_{\left\{\phi:\ \begin{smallmatrix}\phi(s)\geq(\alpha-\delta')s,\,0\leq s\leq 1\\ \phi(1)\leq\alpha+\delta'\end{smallmatrix}\right\}}\int_0^1\Lambda_A^*(\phi'(s))ds
\end{aligned} \tag{61}$$

where ϕ is an absolutely continuous function on $[0,1]$ and $\phi(0)=0$.

Note that (61) holds for any δ' such that $0<\delta'<\delta$. We claim that

$$\lim_{\delta'\to 0}\inf_{\left\{\phi:\ \begin{smallmatrix}\phi(s)\geq(\alpha-\delta')s,\,0\leq s\leq 1\\ \phi(1)\leq\alpha+\delta'\end{smallmatrix}\right\}}\int_0^1\Lambda_A^*(\phi'(s))ds=\Lambda_B^*(\alpha). \tag{62}$$

Therefore (57) holds in this case too.

We now prove the claim (62).

If $\alpha=Ea(0)$, then $\Lambda_B^*(\alpha)=\Lambda_A^*(\alpha)=\Lambda_A^*(Ea(0))=0$. On the other hand, the left-hand side of (62) is always nonnegative, and it attains zero with $\phi(s)=Ea(0)s, 0\leq s\leq 1$. Therefore (62) holds.

If $\alpha>Ea(0)$, then for sufficiently small $\delta'>0$, $\alpha-\delta'>Ea(0)$. Observe that for any absolute continuous function ϕ on $[0,1]$ such that $\phi(s)\geq(\alpha-\delta')s, 0\leq s\leq 1$ and $\phi(0)=0,\phi(1)\leq\alpha+\delta'$, by the convexity of Λ_A^*, we have

$$\int_0^1\Lambda_A^*(\phi'(s))ds\geq\Lambda_A^*(\int_0^1\phi'(s)ds)=\Lambda_A^*(\phi(1))\geq\Lambda_A^*(\alpha-\delta').$$

On the other hand, for $\phi(s)=(\alpha-\delta')s, 0\leq s\leq 1$, $\int_0^1\Lambda_A^*(\phi'(s))ds=\Lambda_A^*(\alpha-\delta')$. Therefore,

$$\inf_{\left\{\phi:\ \begin{smallmatrix}\phi(s)\geq(\alpha-\delta')s,\,0\leq s\leq 1\\ \phi(1)\leq\alpha+\delta'\end{smallmatrix}\right\}}\int_0^1\Lambda_A^*(\phi'(s))ds=\Lambda_A^*(\alpha-\delta')=\Lambda_B^*(\alpha-\delta')$$

Take $\delta'\to 0$, we have (62).

If $\alpha<Ea(0)$, (62) can be similarly proved by using the fact that for sufficiently small $\delta'>0$, $\alpha+\delta'<Ea(0)$. ∎

References

[1] "Advances in the Fundamentals of Networking – Part I: Bridging Fundamental Theory and Networking". *IEEE Journal on Selected Area in Communications*, Vol. 13, No. 6, August 1995.

[2] "Advances in the Fundamentals of Networking – Part II: Bridging Fundamental Theory and Networking". *IEEE Journal on Selected Area in Communications*, Vol. 13, No. 7, September 1995.

[3] V. Anantharam, "How Large Delays Build up in a GI/ G/ 1 Queue", *Queueing Systems*, Vol. 5, pp. 345-368, 1988.

[4] D. Bertsimas, I. Ch. Paschalidis and John N. Tsitsiklis, "On the Large Deviations Behavior of Acyclic Networks of G/ G/ 1 Queues", *Preprint*, 1994.

[5] D. D. Botvich, T. J. Corcoran, N. G. Duffield and P. Farrell, "Economies of Scale in Long and Short Buffers of Large Multiplexers", *Preprint*, 1995.

[6] C. S. Chang, "Stability, Queue Length and Delay of Deterministic and Stochastic Queueing Networks", *IEEE Transaction on Automatic Control*, May 1994.

[7] C. S. Chang, "Sample Path Large Deviation and Intree Network", To appear in *Queueing Systems*, 1994.

[8] C. S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin, "Effective Bandwidth and Fast Simulation of ATM Intree Networks", *Performance Evaluation*, Vol. 20, pp. 45-66, 1994.

[9] C.-S. Chang and J. A. Thomas, "Effective Bandwidth in High-Speed Digital Networks", *IEEE Journal on Selected Area in Communications*, Vol. 13, No. 6, pp. 1091-1101, August 1995.

[10] C.-S. Chang and T. Zajic, "Effective Bandwidths of Departure Processes from Queues with Time Varying Capacities", In *Proceedings of IEEE INFOCOM'95*, pp. 1001-1009, 1995.

[11] D. Clark, S. Shenker and L. Zhang, "Supporting Real-Time Applications in an Integrated Service Packet Network: Architecture and Mechanism", In *Proceedings of ACM SIGCOMM'92*, pp. 14-26, 1992.

[12] R. L. Cruz, "A Calculus for Network Delay, Part I: Network Elements in Isolation", *IEEE Transaction on Information Theory*, Vol. 37, No. 1, Jan. 1991, pp. 114-131.

[13] A. Dembo and T. Zajic, "Large Deviations:from empirical mean and measure to partial sums process Techniques", *Preprint*, 1994.

[14] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Jones and Bartlett Publishers, 1993.

[15] A. Demers, S. Keshav and S. Shenker, "Analysis and Simulation of a Fair Queueing Algorithm", *Journal of Internetworking: Research and Experience*, 1, pp. 3-26, 1990. Also in *Proceedings of ACM SIGCOMM '89*, pp. 3-12.

[16] G. de Veciana and G. Kesidis, "Bandwidth Allocation for Multiple Qualities of Service Using Generalized Processor Sharing", Revised Version, *Preprint*. An earlier version appeared in *Proceedings of IEEE GLOBECOM'94*, 1994.

[17] G. de Veciana, C. Courcoubetis and J. Walrand, "Decoupling Bandwidths for Networks: a Decomposition Approach to Resource Management", In *Proceedings of IEEE INFOCOM'94*, 1994.

[18] G. de Veciana and J. Walrand, "Effective Bandwidths: Call Admission , Traffic Policing and Filtering for ATM Networks", Submitted to *IEEE/ACM Transactions on Networking*, 1993.

[19] N. G. Duffield, "Exponential Upper Bounds for Queues with Markovian Arrivals", *Preprint*, 1993.

[20] R. S. Ellis, *Entropy, Large Deviations and Statistical Mechanics.* New York, Springer-Verlag, 1985.

[21] A. Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks", *IEEE/ACM Trans. on Networking*, Vol. 1, No. 3, June 1993, pp. 329-357.

[22] R. J. Gibbens and P. J. Hunt, "Effective Bandwidth for Multi-Type UAS Channel", *QUESTA*, No. 9, pp. 17-28, 1991.

[23] P. W. Glynn and W. Whitt, "Logarithmic Asymptotics for Steady-State Tail Probabilities in a Single Server Queue", *J. Appl. Prob.*, Vol. 31, 1994.

[24] S. J. Golestani, "Network Delay Analysis of a Class of Fair Queueing Algorithms", *IEEE Journal on Selected Area in Communications*, Vol. 13, No. 6, pp. 1057-1070, August 1995.

[25] R. Guérin, H. Ahmadi and M. Naghshineh, "Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 7, Sept.,1991, pp. 968-981.

[26] J. Hui, "Resource Allocation for Broadband Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 6, No. 9, Dec. 1988, pp.1598-1608.

[27] F. P. Kelly, "Effective Bandwidths for Multi-Class Queues", *QUESTA*, Vol. 9, pp. 5-16, 1991.

[28] R. M. Loynes, "The Stability of a Queue with Non-Independent Inter-Arrival and Service Times", *Process. Cambridge Philos. Soc.*, Vol. 58, pp.497-520, 1962.

[29] G. Kesidis, J. Walrand and C. S. Chang, "Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources", *IEEE/ACM Trans. Networking*, Aug. 1993.

[30] J. F .Kurose, "Open Issues and Challenges in Providing Quality of Service Guarantees in High-Speed Networks", *ACM Computer Communication Review*, pp. 6-13, Jan. 1993.

[31] Z. Liu, P. Nain and D. Towsley, "Exponential Bounds for a Class of Stochastic Processes with Application to Call Admission Control in Networks", to appear in the Proc. of the 33rd *Conference on Decision and Control (CDC'93)*, February, 1994.

[32] F. Lo Presti, Z.-L. Zhang and D. Towsley, "Bounds, Approximations and Applications for a Two-Queue GPS System", To appear in *Proceedings of IEEE INFOCOM'96.* See also *Technical Report*, Computer Science Department, University of Massachusetts, July 1995.

[33] A. K. Parekh, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks", Ph.D Thesis, Department of Electrical Engineering and Computer Science, MIT, February 1992.

[34] A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single Node Case", *IEEE/ACM Transaction on Networking*, Vol. 1, No. 3, pp. 344-357, June 1993.

[35] A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Multiple Node Case", *IEEE/ACM Transaction on Networking*, No. 2, Vol. 2, pp. 137-150, April 1994.

[36] S. Shenker, D. Clark and L. Zhang, "A Scheduling Service Model and a Scheduling Architecture for an Integrated Services Packet Network", *Preprint*, 1993.

[37] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis*. New York, Chapman and Hall, 1995.

[38] A. Weiss, "An introduction to Large Deviations for Communication Networks", *IEEE Journal on Selected Area in Communications*, Vol. 13, No. 6, pp. 938-952, August 1995.

[39] W. Whitt, "Tail Probabilities with Statistical Multiplexing and Effective Bandwidths in Multi-Class Queues", *Telecommunication Systems*, No. 2, 1993, pp. 71-107.

[40] O. Yaron and M. Sidi, "Performance and Stability of Communication Networks via Robust Exponential Bounds", *IEEE/ACM Trans. on Networking*, Vol. 1, No. 3, pp. 372-385, 1993.

[41] O. Yaron and M. Sidi, "Generalized Processor Sharing Networks with Exponentially Bounded Burstiness Arrivals", In *Proceedings of IEEE INFOCOM '94*, June 1994.

[42] Z.-L. Zhang, "Large Deviations and the Generalized Processor Sharing Scheduling Discipline: Upper and Lower Bounds, Part II: Multiple-Queue Systems", *Technical Report* UM-CS-95-97, Computer Science Department, University of Massachusetts, Oct., 1995. Available via FTP from `gaia.cs.umass.edu` in pub/Zhan95:TR95-97.ps.Z.

[43] Z.-L. Zhang, Z. Liu, J. Kurose and D. Towsley, "Call Admission Control Schemes under the Generalized Processor Sharing Scheduling Discipline", *Technical Report* UM-CS-95-52, Computer Science Department, University of Massachusetts, March 1995. Available via FTP from `gaia.cs.umass.edu` in pub/Zhan95:TR95-52.ps.Z. Submitted to *Telecommunication Systems*.

[44] Z.-L. Zhang, D. Towsley and J. Kurose, "Statistical Analysis of Generalized Processor Sharing Scheduling Discipline", *IEEE Journal on Selected Area in Communications*, Vol. 13, No. 6, pp. 1071-1080, August 1995. See also *Technical Report* UM-CS-95-10, Computer Science Department, University of Massachusetts, February 1995. Available via FTP from `gaia.cs.umass.edu` in pub/Zhan95:TR95-10.ps.Z.