Department of Computer Science Technical Reports

Department of Computer Science

1989

On the Analysis of the Tail Queue Length and Waiting Time Distributions of a GI/GI/c Queue

John S. Sadowsky

Wojciech Szpankowski
Purdue University, spa@cs.purdue.edu

Report Number:
89-937

ON THE ANALYSIS OF THE TAIL QUEUE LENGTH AND WAITING TIME DISTRIBUTIONS OF A GI/GI/c QUEUE

John S. Sadowsky
Wojciech Szpankowski

CSD-TR-937
November 1989
(Revised March 1990)

ON THE ANALYSIS OF THE TAIL QUEUE LENGTH AND WAITING TIME DISTRIBUTIONS OF A GI/GI/c QUEUE

John S. SADOWSKY
School of Electrical Engineering
Purdue University
West Lafayette, IN 47097

Wojciech SZPANKOWSKI*
Department of Computer Science
Purdue University
West Lafayette, IN 47097

This paper deals with the tail behavior of the stationary queue length and waiting time distributions for a GI/GI/c queueing system. There are $c \geq 1$ servers and each server may have a different service time distribution. Our results extend well known classical results for the GI/GI/1 queue. Presentation of the key constructions and ideas of this new proof is the main goal of this paper. We also apply these results to obtain limiting distributions of the maximum queue length and waiting time distributions for the case of $c \geq 1$ servers.

1. INTRODUCTION AND SUMMARY

Consider a stable GI/GI/1 queueing system. Let $\{A_k\}$ denote the i.i.d. inter-arrival times and let $\{B_k\}$ denote the i.i.d. service times. $A^*(\alpha)$ and $B^*(\alpha)$ shall denote respectively the inter-arrival time and service time distribution Laplace-Stieltjes transforms. Q_k and W_k shall denote the queue length and waiting time observed by the k'th job upon it's arrival. Define

$$\Lambda(\eta) \;=\; \log(A^*(\eta)) \;+\; \log(B^*(-\eta)). \tag{1.1}$$

If there is a positive solution of the *characteristic equation* $\Lambda(\theta) = 0$ (which is necessarily unique because of the convexity of $\Lambda(\eta)$), and if $\Lambda'(\theta) < \infty$, then Feller [1] proved that under stationary operation

$$\mathcal{P}(\, Q_k \geq n \,) \;\sim\; K_Q\, \omega^n \tag{1.2}$$

where $\omega = A^*(\theta)$, and for FIFO (first in - first out) queueing discipline

$$\mathcal{P}(\, W_k \geq w \,) \;\sim\; K_W\, e^{-\theta w}, \tag{1.3}$$

where K_Q and K_W are constants.

The classical approach by which one can obtain these results is presented by Feller [1]. (There are also several refinements, such as [2].) This approach is based on the FIFO waiting time relationship

* This research was supported in part by AFOSR (90-0107), in part by NSF (CCR-8900305) and in part bv the National Librarv of Medicine (R01 LM05118).

$$W_k = (W_{k-1} + B_k - A_k)^+ \tag{1.4}$$

where $x^+ = x$ if $x \ge 0$ and $= 0$ if $x < 0$. From (1.4) it follows that

$$\mathcal{P}(W_k \ge w) = \mathcal{P}\Big(\max_{k \ge 1} \sum_{k' = 1}^{k} (B_{k'} - A_{k'}) \ge w \Big). \tag{1.5}$$

Feller obtains (1.3) by using an exponential change in distribution of the form $dF_A^{(\theta)}(a) = e^{-\theta a}\, dF_A(a) / A^*(\theta)$ and $dF_B^{(\theta)}(a) = e^{\theta b}\, dF_B(b) / B^*(-\theta)$.

In works such as [1] and [2], the change in distribution is actually applied to analyze the maximum of a random walk, that is, the right side of (1.5), and hence, this approach is not really a direct analysis of the queueing problem. Moreover, in [3] Kingman presents a very convincing argument indicating that any approach based on (1.4) and (1.5) can not be extended to the case of $c > 1$ servers.

An alternative approach is to use the analytical methods associated with phase-type distributions (abbreviated PH). In [4], Takahashi obtains these results for the PH/PH/c queue with homogeneous service time distributions, that is, all c servers have the same service time distribution. In [5], Neuts and Takahashi present the extension to the GI/PH/c with heterogeneous service time distributions. As in the GI/GI/1 case, the results of [4] and [5] can be expressed in terms of the solution of a characteristic equation of the form $\Lambda(\theta) = 0$ with $\Lambda(\eta)$ being appropriately defined in terms of the inter-arrival time and service time Laplace transforms. For example, in [4] the function is

$$\Lambda(\eta) = \log(A^*(\eta)) + \log(B^*(-\eta/c)). \tag{1.6}$$

The analysis presented in this paper is similar to classical approach because our basic tool is an exponential change in distribution. However, we do not employ (1.4) or (1.5), so Kingman's argument against generalization does not apply. Instead, we apply change in distribution directly to the queueing problem. Our analysis is based on a natural regeneration of the queue in terms of busy/idle cycles and the basic idea is to consider the behavior of a busy periods that produce large maximum queue lengths, or large maximum waiting times.

In comparison to the work of Neuts and Takahashi, our approach has certain advantages. First, our results for the GI/GI/c queue are somewhat more general. We obtain (1.2) and (1.3) under the essentially the full generality classical $c = 1$ result; $\Lambda(\theta) = 0$ and $\Lambda'(\theta) < \infty$. We also obtain weaker "logarithmic limits" which hold under essentially no hypothesis (other than the i.i.d. inter-arrival time and service time assumption). But generality is not the only advantage. The exponential change in distribution provides the means to efficiently and accurately evaluate the actual probabilities $\mathcal{P}(Q_k \ge n)$ and $\mathcal{P}(W_k \ge w)$, for finite n and w. This can be done using the Monte Carlo technique of *importance sampling* (or "quick simulation") as in [6], [7] and [8]. Reference [8] gives an asymptotic analysis of the importance sampling estimator's variance for the case of estimating $\mathcal{P}(\overline{Q} \ge n)$ where $\overline{Q}$ is the maximum queue length over a busy period. It is demonstrated that the exponential change in distribution as an importance sampling simulation distribution has a strong asymptotic optimality property called *asymptotic efficiency*. While we shall not do this here, the analysis of [8] can be applied to the constructions of this paper to obtain asymptotically efficient importance sampling simulation for estimating $\mathcal{P}(Q_k \geq n)$ and $\mathcal{P}(W_k \geq w)$. Finally, a third advantage of our approach is that one can apply the exponential twisting techniques for Markov additive processes (see [9] and references therein) to extend our analysis to the case of stationary Markov distributions for both inter-arrival time and service time processes. This extension is in principle straightforward, at least in the case of the logarithmic limits.

This paper is organized as follows. Section 2 presents the decomposition of the queueing problem into busy/idle cycles and then presents the main results. Cycle based analysis is not new [6], [8]. However, for the case of c > 1 servers we use a sequence of dependent cycles that form a Markov chain. Section 3 then presents the main ideas and constructions used to prove our results for queue length. Section 4 briefly presents the parallel approach for waiting time analysis. Due to space limitation, we make no attempt to provide complete proofs. Instead, our goal here is to simply present the main ideas behind these proofs in an illustrative fashion. Section 5 presents some sample numerical results obtained by applying our constructions to the importance sampling method.

2. CYCLE BASED ANALYSIS AND STATEMENT OF MAIN RESULTS

We consider a single queue with $c \geq 1$ servers. A_k shall denote the inter-arrival time between the k-1'th and k'th jobs. The sequence $\{A_k: k = 1,2,.. \}$ is i.i.d. with distribution function $F_A(a)$. $B_j^{(i)}$ shall denote the service time required for the j'th job processed by the i'th server. For each fixed i, $\{B_j^{(i)}: j = 1,2,.. \}$ is an i.i.d. sequence with distribution function $F_i(b)$. The sequences $\{A_k\}$ and $\{B_j^{(i)}\}$, $i = 1,..,c$, are independent. Throughout this paper we shall refer to k as the "arrival index", i as the "server index" and, for each fixed i, j is the "service index." We assume throughout that $F_A(0) < 1$ and $F_i(0) < 1$ for each $i = 1,..,c$. We also assume that the service time distributions are "spread out," that is, some convolution power has an absolutely continuous component. This "spread out" condition is required for application of certain renewal theory results in the proof, but it will not be discussed further.

We define a *busy period* to be a contiguous time interval during which *all* servers are continuously busy. A busy period begins at the instant that a job arrives to find exactly c–1 servers busy, and terminates at the first instant that one of the servers becomes idle. Conversely, an *idle period* is contiguous time interval during which at each instant there is at least one idle server. Notice that this definition differs from the conventional busy/idle cycle definition in which a busy period has at least one server busy and an idle period has all servers idle. An advantage of the conventional definition is that successive cycles are independent. However, in the case of multiple servers, the disadvantage is that idle time during busy periods significantly complicates the analysis. In order to distinguish our definition from the conventional one, we shall refer to our busy/idle cycle as a *c-cycle.*

A c-cycle begins with the arrival of a job that finds exactly c–1 busy servers and one idle server. Define $\tilde{B}_m^{(i)}$ to be the residual service time (that is, the service time remaining) for the job being processed by the i'th server at the instant that the (m+1)'th c-cycle begins. (Precisely one of the $\tilde{B}_m^{(i)}$'s is zero.) For compact notation we write $\tilde{\mathbf{B}}_m = (\tilde{B}_m^{(1)}, .. ,\tilde{B}_m^{(c)})$. Let $\mathbf{C}_m$ denote the m'th c-cycle; that is, $\mathbf{C}_m$ is a random element that includes all of the service times and inter-arrival times that occur during the m'th cycle. In particular $\tilde{\mathbf{B}}_m$ is determined by $\mathbf{C}_m$. It is evident that $\{\mathbf{C}_m\}$ is a Markov chain, and furthermore,

$$\mathcal{P}(\mathbf{C}_{m+1} \in \cdot \mid \mathbf{C}_m, \mathbf{C}_{m-1} , \dots) \;=\; \mathcal{P}(\mathbf{C}_{m+1} \in \cdot \mid \tilde{\mathbf{B}}_m). \tag{2.1}$$

To gain some insight, we now indicate that for a very large class of GI/GI/c queues the c-cycle chain has a natural regeneration structure that is directly related to the conventional cycles. Define

$$p(\tilde{\mathbf{b}}) \;=\; \mathcal{P}\Big(\begin{matrix}\text{all servers become idle at} \\ \text{some instant during } \mathbf{C}_{m+1}\end{matrix} \;\Big|\; \tilde{\mathbf{B}}_m = \tilde{\mathbf{b}} \Big)$$

and

$$\nu(\cdot) \;=\; \mathcal{P}\Big(\tilde{\mathbf{B}}_{m+1} \in \cdot \;\Big|\; \begin{matrix}\text{all servers become idle at} \\ \text{some instant durring } \mathbf{C}_{m+1}\end{matrix} \Big).$$

Then

$$\mathcal{P}(\tilde{\mathbf{B}}_{m+1} \in \cdot \mid \tilde{\mathbf{B}}_m = \tilde{\mathbf{b}}) \;\geq\; p(\tilde{\mathbf{b}})\, \nu(\cdot).$$

The above minorization determines a regeneration that can be interpreted in terms of a weighted coin toss that is performed for each c-cycle. (See [10, ch. 4].) Given $\{\tilde{\mathbf{B}}_{m+1} = \tilde{\mathbf{b}}\}$ the probability of head for c-cycle $\mathbf{C}_{m+1}$ is $p(\tilde{\mathbf{b}})$. If the result of this coin toss is a head then $\tilde{\mathbf{B}}_{m+1}$ can be interpreted as an independent sample from the distribution $\nu(\cdot)$ and this outcome causes a regeneration: given a head for c-cycle $\mathbf{C}_{m+1}$ the future c-cycles $\mathbf{C}_{m+2},\mathbf{C}_{m+3}, \dots$ are conditionally independent of the past c-cycles $\mathbf{C}_m,\mathbf{C}_{m-1}, \dots$. Notice that it is this regeneration structure that relates our c-cycle definition to the conventional definition.

Not all queues regenerate in the above fashion. (It may happen that $\mathcal{P}(p(\tilde{\mathbf{B}}_m) \geq 0) = 0$. There is a well known D/D/2 example in which this does happen. See [10, p. 57].) However, all that we actually require in terms of stability of the c-cycle chain is listed in the definition below and this stability condition does not necessarily rely on the above regeneration.

Definition: Let L_m denote the number of jobs that arrive during the m'th cycle $\mathbf{C}_m$. We shall say that the GI/GI/c queueing system is *stable* if the c-cycle chain $\{\mathbf{C}_m\}$ is ergodic and if under the stationary distribution we have $E[L_m] < \infty$.

We do not investigate conditions for stability here. We simply adopt the above definition as a hypothesis. For the purpose of reference below, it is appropriate mention that one criteria for stability depends on the *utilization* parameter

$$\rho \;=\; \frac{\text{arrival rate}}{\text{total service rate}} \;=\; \frac{E[A_k]^{-1}}{\sum_{i=1}^{c} E[B_j^{(i)}]^{-1}}. \tag{2.2}$$

It is known, at least in the case of homogeneous service distributions [11], that the queueing process is stable if and only if $\rho < 1$. We shall see in the next section that ρ plays a key role in our analysis.

Now, let $L_m^{(n)}$ denote the number of jobs that arrive during the m'th cycle and find at least n jobs in queue (that is, $Q_k \geq n$). Let $\mathcal{F}_M$ denote the history of the first M cycles, and let $Q^{(M)}$ denote the queue length seen upon arrival by a job that is randomly selected from these M cycles. If the queue is stable, then by the ergodicity of the c-cycle chain we have

$$\mathcal{P}(\, Q^{(M)} \geq n \mid \mathcal{F}_M \,) \;=\; \frac{\text{total \# of jobs that find > n queue length}}{\text{total \# of jobs}}$$

$$=\; \frac{\sum_{m=1}^{M} L_m^{(n)}}{\sum_{m=1}^{M} L_m} \;\rightarrow\; \frac{E_\pi[L_1^{(n)}]}{E_\pi[L_1]}$$

as $M \rightarrow \infty$ almost surely. In the last line, $E_\pi[\cdot]$ denotes the expectation with respect to the stationary distribution of the c-cycle chain. It is apparent that all we really require is the stationary distribution of $\tilde{\mathbf{B}}_0$. We now have that the stationary queue length distribution can be expressed as

$$\mathcal{P}(\, Q_k \geq n \,) \;=\; \frac{E_\pi[L_1^{(n)}]}{E_\pi[L_1]}. \tag{2.3}$$

Likewise, let $L_m^{(w)}$ denote the number of jobs in that arrived during the m'th cycle and experience a waiting time of at least w (that is, $W_k \geq w$). Then

$$\mathcal{P}(\, W_k > w \,) \;=\; \frac{E_\pi[L_1^{(w)}]}{E_\pi[L_1]}. \tag{2.4}$$

Before presenting our theorems, we must first indicate how to construct the characteristic equation $\Lambda(\theta) = 0$ for the case of multiple servers with possibly different service time distributions. This equation is to be expressed in terms of the inter-arrival time and service time Laplace transforms

$$A^*(\eta) \;=\; E[\,\exp(-\eta A_k)\,] \quad \text{and} \quad B_i^*(\alpha_i) \;=\; E[\,\exp(-\alpha_i B_j^{(i)})\,]. \tag{2.5}$$

The problem is that with c possibly different service time transforms $B_i^*(\alpha_i)$, we actually have a c-dimensional vector to work with; $\alpha = (\alpha_1,..,\alpha_c)$. The following "basic lemma" provides the appropriate reduction to a scalar parameter η. The idea is that we restrict attention to a curve $\alpha(\eta)$ in $[0,\infty)^c$ that is defined such that at each point on the curve all of the service time transforms ($B_i^*(\alpha_i)$, i = 1,..,c) have the same value.

Basic Lemma: Define

$$\overline{\eta} \;=\; \sup\left\{ \eta : \begin{array}{c} \text{there exists an } \alpha \in [0,\infty)^c \text{ with } \eta = \sum_{i=1}^{c} \alpha_i \\ \text{and } B_i^*(-\alpha_i) = B_1^*(-\alpha_1) < \infty \text{ for } i = 2,..,c \end{array} \right\}. \tag{2.6}$$

Then for each $\eta \in [0,\overline{\eta})$ there is a unique $\alpha(\eta) \in [0,\infty)^c$ such that $\sum_{i=1}^{c} \alpha_i(\eta) = \eta$ and $B_i^*(-\alpha_i(\eta)) = B_1^*(-\alpha_1(\eta)) < \infty$ for i = 2,..,c. Moreover, $\alpha(\eta)$ is an analytic curve on $(0,\overline{\eta})$ with $\alpha(0) = \mathbf{0}$. In the case $\overline{\eta} < \infty$, define $\alpha(\overline{\eta}) = \lim_{\eta \uparrow \overline{\eta}} \alpha(\eta)$ allowing $+\infty$ as a limit. Next define

$$\Lambda(\eta) \;=\; \log(A^*(\eta)) \;+\; \log(\,B_1^*(-\alpha_1(\eta))\,) \tag{2.7}$$

for $\eta \in [0,\overline{\eta}]$ and $\Lambda(\eta) = +\infty$ for $\eta > \overline{\eta}$. Then $\Lambda(\eta)$ is a lower semicontinuous proper convex function, and moreover, $\Lambda(\eta)$ is analytic on $(0,\overline{\eta})$. ❑

In the case of homogeneous service distributions, we have $\eta = (\eta/c,..,\eta/c)$. In this case $\Lambda(\eta)$ of formula (2.6) agrees with Takahashi's formula [4]. (See equation (1.6).) With a little manipulation, the results of [5] can also be expressed in terms of definition (2.7).

There is the possibility certain pathological behaviors of the function $\Lambda(\eta)$ that do not occur in the case of phase type service distributions. We may have $\Lambda(\eta) \to -\infty$ as $\eta \to \infty$, or we may have $0 < \overline{\eta} < \infty$ and $\Lambda(\overline{\eta}) < 0$. In both cases there is no positive solution of $\Lambda(\theta) = 0$. In general, we define

$$\theta \;=\; \sup\{\,\eta : \Lambda(\eta) \le 0\,\}. \tag{2.8}$$

This allows the possibility $\theta = +\infty$ which occurs in the first case mentioned above. Also, (2.8) yields $\theta = \overline{\eta}$ in the case that $0 < \overline{\eta} < \infty$ and $\Lambda(\overline{\eta}) < 0$. Define

$$\omega \;=\; \inf_{0 \le \eta \le \theta} A^*(\eta). \tag{2.9}$$

For $\eta \ge 0$, $0 < A^*(\eta) < \infty$, $A^*(\eta)$ is continuous and $A^*(\eta)$ is strictly decreasing (since $F_A(0) < 1$). It follows that $\omega = A^*(\theta)$ when $\theta < \infty$ and $\omega = 0$ when $\theta = \infty$.

Define $\overline{Q}$ = maximum Q_k over the first c-cycle with $\bar{\mathbf{B}}_0$ being sampled from the stationary distribution $\pi(\cdot)$, where $\pi(\cdot)$ is the c-cycle stationary distribution. It turns out that $\overline{Q}$ is a key variable in our analysis of the stationary queue length, but we note that the distribution of $\overline{Q}$ is of interest in its own right as it is related to the problem of buffer overflow in systems with a finite queue buffer. See [6], [8] and [12].

Theorem 1: Assume that the queueing system is stable and operating under the stationary distribution. Then

$$\lim_{n\to\infty} \frac{1}{n} \log(\mathcal{P}(Q_k \geq n)) \;=\; \lim_{n\to\infty} \frac{1}{n} \log(\mathcal{P}(\overline{Q} \geq n)) \;=\; \log(\omega) \qquad (2.10)$$

If in addition we have $0 < \theta < \infty$, $\Lambda(\theta) = 0$ and $\Lambda'(\theta) < \infty$, then there exists constants K_Q and $K_{\overline{Q}}$ in $(0,\infty)$ such that

$$\lim_{n\to\infty} \omega^{-n} \mathcal{P}(\overline{Q} \geq n) \;=\; K_{\overline{Q}} \qquad (2.11)$$

and

$$\lim_{n\to\infty} \omega^{-n} \mathcal{P}(Q_k \geq n) \;=\; K_Q. \qquad (2.12)$$

❑

Our analysis of the stationary waiting time distribution is a direct parallel of the queue length analysis. Define $\overline{W}$ = maximum W_k over the first (stationary) busy period.

Theorem 2: Assume that the queueing system is stable and operating under the stationary distribution. Then

$$\lim_{w\to\infty} \frac{1}{w} \log(\mathcal{P}(W_k \geq w)) \;=\; \lim_{w\to\infty} \frac{1}{w} \log(\mathcal{P}(\overline{W} \geq w)) \;=\; \log(\theta) \qquad (2.13)$$

If in addition we have $0 < \theta < \infty$, $\Lambda(\theta) = 0$ and $\Lambda'(\theta) < \infty$, then there exists constants K_W and $K_{\overline{W}}$ in $(0,\infty)$ such that

$$\lim_{n\to\infty} e^{\theta w} \mathcal{P}(\overline{W} \geq w) \;=\; K_{\overline{W}} \qquad (2.14)$$

and

$$\lim_{n\to\infty} e^{\theta w} \mathcal{P}(W_k > n) \;=\; K_W. \qquad (2.15)$$

❑

Finally, define Q_k^{max} = max $\{Q_{k'}$: $k' \leq k\}$ and W_k^{max} = max $\{W_{k'}$: $k' \leq k\}$. Then as a consequence of Theorems 1 and 2 we have. Following the work of Iglehart [12] (see also Anderson [12]) we can apply Theorems 1 and 2 to prove the following theorem.

Theorem 3: Assume that the queue is stable and that is regenerates with $E[L] < \infty$ where L is distributed as the length of a regeneration cycle. Then

$$\frac{Q_k^{max}}{-\log_\omega(k)} \;\to\; 1 \quad \text{and} \quad \frac{W_k^{max}}{\log(k)} \;\to\; 1 \qquad (2.9)$$

where the convergence is in probability. Define $\beta = E[L]^{-1}$. If we have $0 < \theta < \infty$, $\Lambda(\theta)$ $= 0$ and $\Lambda'(\theta) < \infty$, then there exists constants $K^*_{\overline{Q}}$ and $K^*_{\overline{W}}$ such that

$$\exp(-\beta\omega^{n-1}) \;\le\; \liminf_{k\to\infty} \mathcal{P}(\, Q_k^{\max} \;<\; n - \log_\omega(K_Q^* k)\,)$$
$$=\; \limsup_{k\to\infty} \mathcal{P}(\, Q_k^{\max} \;<\; n - \log_\omega(K_Q^* k)\,) \;\le\; \exp(-\beta\omega^{n}) \qquad (2.10)$$

and

$$\lim_{k\to\infty} \mathcal{P}(\, W_k^{\max} \;<\; w + \log(K_W^* k)\,) \;=\; \exp(-\beta e^{-w}). \qquad (2.11)$$

□

3. ANALYSIS OF QUEUE LENGTH

Appealing to formula (2.3), we now concentrate on the behavior of a single c-cycle busy period that begins with the arrival of job $k = 0$ at time $t = 0$ in order to determine the large n behavior of the expectation $E_\pi[L_1^{(n)}]$. $\bar{\mathbf{B}}_0 = (\bar{B}_0^{(1)}, .. ,\bar{B}_0^{(c)})$ is the vector of residual service times for the jobs being serviced at time $t = 0$. As discussed above, the distribution of this vector is assumed to be the stationary distribution.

Define

$$J_k^{(i)} \;=\; \inf\left\{ j : \sum_{k'=1}^{k} A_{k'} \;<\; \bar{B}_0^{(i)} \;+\; \sum_{j'=1}^{j} B_{j'}^{(i)} \right\} \qquad (3.1)$$

(with $\inf \varnothing = \infty$). Prior to the end of the first busy period, $J_k^{(i)}$ is the service index of the job being processed by the i'th server at the instant that the k'th job arrives. At the instant that job k arrives, a total of k+c jobs have entered the system. (This includes the job that arrives at time $t = 0$ and the $c - 1$ jobs being processed at time $t = 0$.) The total number of jobs to have exited the system is $\sum_{i=1}^{c} J_k^{(i)}$. Hence, the first job to arrive and find an idle server is the job with arrival index

$$K_0 \;=\; \inf\left\{ k : \sum_{i=1}^{c} J_k^{(i)} \;\ge\; k+1 \right\}. \qquad (3.2)$$

Next define

$$\tilde{K}_n \;=\; \inf\left\{ k : \sum_{i=1}^{c} J_k^{(i)} \;= k-n \right\}. \qquad (3.3)$$

To see the meaning of (3.3), note that Q_k = # of jobs that have arrived − # of jobs served and being served = $k + c \;-\; (\, \sum_{i=1}^{c} J_k^{(i)} \;+\; c\,) \;=\; k \;-\; \sum_{i=1}^{c} J_k^{(i)}$. Hence, on the event $\{\tilde{K}_n < K_0\}$, we have $\tilde{K}_n = \inf\{k\colon Q_k = n\}$. In words, given that the queue length exceeds n during the cycle, $\tilde{K}_n$ is the arrival index of the first job to arrive to find n jobs in queue. Finally, define

$$K_n = \min\,\{K_0,\tilde{K}_n\}, \qquad (3.4)$$

$$\overline{Q} \;=\; \max\,\{\, Q_k : k \le K_0\,\}, \qquad (3.5)$$

and

$$\mathcal{F}_k \;=\; \sigma(A_1,..,A_k;\ \tilde{\mathbf{B}}_0; B_j^{(i)}, j = 1,..J_k^{(i)}, i = 1,..,c). \tag{3.6}$$

for $k \geq 0$. The following lemma simply summarizes some immediate facts.

Lemma 3.1: $\mathcal{F}_k$ is an increasing sequence of σ-fields, K_0 is an $\mathcal{F}_k$-stopping time, $\{\tilde{K}_n\}$ and $\{K_n\}$ are increasing sequences of $\mathcal{F}_k$-stopping times, and $\mathcal{P}(\ K_n < \infty\ ) = 1$. Define $\mathcal{J}_n^{(i)} = J_{K_n}^{(i)}$. Then on the event $\{\overline{Q} \geq n\} = \{K_n = \tilde{K}_n\} = \{\tilde{K}_n < K_0\}$ we have

$$\sum_{i=1}^{c} \mathcal{J}_n^{(i)} \;=\; \tilde{K}_n - n. \tag{3.7}$$

Next we present our exponential change in distribution for the queueing process busy period which we call the *(α,n)-conjugate distribution*. We write $\mathcal{P}^{(\alpha,n)}(\cdot)$. Define $\mathcal{D} = \{\alpha : \alpha_i \geq 0 \text{ and } \Lambda_i(\alpha_i) < \infty\}$ = a closed rectangle in $[0,\infty)^c$ containing $\mathbf{0}$. For any $\alpha \in \mathcal{D}$ define

$$dF_i^{(\alpha)}(b) \;=\; \frac{e^{\alpha_i b}\, dF_i(b)}{B_i^*(-\alpha_i)} \tag{3.8}$$

and

$$dF_A^{(\alpha)}(a) \;=\; \frac{e^{-\eta a}\, dF_A(a)}{A^*(\eta)} \tag{3.9}$$

where $\eta = \eta(\alpha) = \sum_{i=1}^{c} \alpha_i > 0$. The residual service time vector $\tilde{\mathbf{B}}_0$ is not twisted, that is, $\tilde{\mathbf{B}}_0$ is sampled from the stationary distribution of the c-cycle Markov chain $\{\mathbf{C}_m\}$. This specifies restriction of $\mathcal{P}^{(\alpha,n)}(\cdot)$ to $\mathcal{F}_0 = \sigma(\tilde{\mathbf{B}}_0)$. We extend to $\mathcal{F}_k$ recursively. Given $\mathcal{F}_{k-1}$, we determine if $K_n < k$ or $K_n \geq k$. If $K_n \geq k$, then A_k is sampled from $F_A^{(\alpha)}(\cdot)$ and each collection $\{B_j^{(i)}: j = J_{k-1}^{(i)}+1,..,J_k^{(i)}\}$ is an i.i.d. sequence sampled from $F_i^{(\alpha)}(\cdot)$ stopped at time $J_k^{(i)}$. Otherwise, if $K_n < k$, then A_k is sampled from $F_A(\cdot)$ and the random variables $\{B_j^{(i)}, j = J_{k+n-1}^{(i)}+1,..,J_{k+n}^{(i)}\}$ is an i.i.d. sequence sampled from $F_i(\cdot)$ stopped at time $J_k^{(i)}$.

In words, the (α,n)-conjugate distribution twists the i.i.d. inter-arrival time and service sequence distributions according to formulas (3.8) and (3.9), but only up to the arrival of job K_n. Below we shall select the twisting parameter vector α in such a way that the queueing process is unstable up to arrival index K_n. After this instant the process reverts to its nominal stable evolution.

We shall also refer to the *α-conjugate distribution* which is simply the i.i.d. sequence distribution generated by $F_A^{(\alpha)}(\cdot)$ and $F_i^{(\alpha)}(\cdot)$, $i = 1,..,c$. One can think of the α-conjugate distribution as the (α,∞)-conjugate distribution.

For each $i = 1, .., c$, we define

$$S_k^{(i)} = \begin{array}{c}\text{the service time remaining for the job being}\\ \text{served by server i at the instant that job k arrives}\end{array}$$

$$= \tilde{B}_0^{(i)} + \sum_{j=1}^{J_k^{(i)}} B_j^{(i)} - \sum_{k'=1}^{k} A_{k'}. \qquad (3.10)$$

In vector notation we write $\mathbf{S}_k = (S_k^{(1)}, .. ,S_k^{(c)})$.

We have the following change of measure formula. We note that proof is a standard and straightforward computation. In the notation, $\alpha\cdot\beta$ shall denote the Euclidean inner product.

Lemma 3.2: For any $\alpha \in \mathcal{D}$, the process distributions $\mathcal{P}(\cdot)$ and $\mathcal{P}^{(\alpha,n)}(\cdot)$ are mutually absolutely continuous and

$$\frac{d\mathcal{P}^{(\alpha,n)}}{d\mathcal{P}} = \exp\Big(\alpha\cdot\mathbf{S}_{K_n} - \sum_{i=1}^{c} \mathcal{J}_n^{(i)} \log(B_i^*(-\alpha_i)) - K_n \log(A^*(\eta)) \Big). \qquad (3.11)$$

where $\eta = \sum_{i=1}^{c} \alpha_i$.

Having defined an exponential change of measure, for a general vector of twisting parameters α, recall our key Lemma which defines and characterizes the curve $\alpha(\eta)$ and the function $\Lambda(\eta)$ (formula (2.7)). In particular, recall that the curve $\alpha(\eta)$ is defined by the relationships $B_i^*(-\alpha_i(\eta)) = B_1^*(-\alpha_i(\eta))$ for $i = 2,..,c$. In (3.11) we now see the reason for this definition; it equates the coefficients of the $\mathcal{J}_n^{(i)}$'s. This in turn allows us to apply formula (3.7) which significantly reduces the complexity of (3.11). The following lemma presents this reduction.

Lemma 3.3: For any $\eta \in (0,\overline{\eta})$, or for $\eta = \overline{\eta}$ if $\Lambda(\overline{\eta}) < \infty$, formula (3.11) reduces to

$$\frac{d\mathcal{P}^{(\eta,n)}}{d\mathcal{P}} = \exp\big(\alpha(\eta)\cdot\mathbf{S}_{K_n} - (K_n - n)\,\Lambda(\eta) \big)\, A^*(\eta)^{-n} \qquad (3.12)$$

on the event $\{\overline{Q} \geq n\}$.

Hereafter we shall refer to the $(\alpha(\eta),n)$-conjugate distribution (or the $\alpha(\eta)$-conjugate distribution) as simply the *(η,n)-conjugate distribution* (or the η-conjugate distribution). All notation is modified accordingly by replacing the vector α with the scalar η and we shall simply write α_i instead of $\alpha_i(\eta)$.

The proof of the Basic Lemma is for the most part a straightforward application of convex function theory. In the process, one obtains the derivative

$$\Lambda'(\eta) \; = \; (\rho(\eta) \, - \, 1\,)\, E^{(\eta)}[A_1] \tag{3.13}$$

where

$$\rho(\eta) \; = \; \frac{E^{(\alpha)}[A_1]^{-1}}{\sum_{i=1}^{c} E^{(\alpha)}[B_1^{(i)}]^{-1}}\,. \tag{3.14}$$

This is an interesting observation because recalling formula (2.3) we see that the parameter $\rho(\eta)$ determines the stability, or more importantly the instability of the η-conjugate distribution. By our stability assumption, we have $\rho = \rho(0) < 1$ which is equivalent to $\Lambda'(0) < 0$. We shall shortly set $\eta = \theta$ where θ is the unique positive solution of $\Lambda(\theta) = 0$ (which we assume exists for the sake of discussion). By convexity, we must have $\Lambda'(\theta) > 0$ which implies $\rho(\theta) > 1$ which in turn implies that the θ-conjugate distribution is unstable! So, we now see how the (θ,n)-conjugate distribution behaves; it is unstable up to time K_n and then reverts back to the original stable behavior. Notice that this instability will greatly increase the likelihood of the event $\{\overline{Q} \ge n\} = \{K_n = \tilde{K}_n\}$.

We mention that the above characterization of the (θ,n)-conjugate distribution as typical behavior of busy periods that cause large backlogs is also apparent in Anatharam's paper [14].

We conclude this section by summarizing the proof of Theorem 1. Appealing to formula (3.2), it is sufficient to consider $E_\pi[L_1^{(n)}]$. From Lemma 3.3, we have

$$E_\pi[L_1^{(n)}] \; = \; E_\pi^{(\eta,n)}[\, L_1^{(n)} \; \exp(\, - \, \alpha(\eta) \cdot \mathbf{S}_{K_n}) + (K_n - n)\, \Lambda(\eta) \,) \, ; \, \overline{Q} \ge n \,]\, A^*(\eta)^n$$

In particular, if $\Lambda(\theta) = 0$, then

$$E_\pi[L_1^{(n)}] \; = \; E_\pi^{(\theta,n)}[\, L_1^{(n)} \; \exp(\, - \, \alpha(\theta) \cdot \mathbf{S}_{K_n}) \,) \, ; \, \overline{Q} \ge n \,]\, A^*(\theta)^n . \tag{3.15}$$

Theorem 1 follows by demonstrating that the expectation above tends to a limit, in particular, we have

$$\lim_{n \to \infty} E_\pi^{(\theta,n)}[\, L_1^{(n)} \; \exp(\, - \, \alpha(\theta) \cdot \mathbf{S}_{K_n}) \,) \, ; \, \overline{Q} \ge n \,] \; = \; K_Q \, E_\pi[L_1]$$

The fact that this expectation does not vanish follows intuitively from the instability of the θ-conjugate distribution. The technical step is to demonstrate that the residual service time vectors $\mathbf{S}_{K_n}$ have a limiting distribution as $n \to \infty$. To do this we observe that $\{\mathbf{S}_k\}$ is an irreducible and aperiodic Markov chain. Using renewal theory, we can prove that $\{\mathbf{S}_k\}$ positive recurrent. This ergodicity of $\{\mathbf{S}_k\}$ is not sufficient to obtain the desired convergence in distribution $\mathbf{S}_{K_n}$, however, it is a key step.

4. ANALYSIS OF WAITING TIME

Unlike the queue length analysis of Section 3, waiting time analysis depends on the service priority to be employed. For example, the classical analysis based on (1.4) and (1.5) are valid only for FIFO service priority. Here we shall only consider FIFO priority, however, it will be evident that the analysis could be modified for other kinds of priorities as well.

Even though there is actually only one queue, for the purpose of waiting time analysis it is convenient to think of c separate queues; one queue for each server. The service priority can be thought of as a function that assigns the jobs to the various queues. In the real queueing process this assignment is actually performed as the jobs leave the queue and enter service. However, by allowing the queue assignment function to know the service times of the jobs in queue, there is an equivalent assignment function that assigns jobs to these conceptual server queues as they arrive at the system. Define $\hat{J}_k^{(i)}$ to be the service index for the last job in the i'th queue at the instant that job k arrives. (This is in contrast to $J_k^{(i)}$ = the service index of the job being processed by the i'th server at the instant that the k'th job arrives.) Next define

$$W_k^{(i)} = \bar{B}_0^{(i)} + \sum_{j'=1}^{\hat{J}_k^{(i)}} B_{j'}^{(i)} - \sum_{k'=1}^{k} A_{k'}. \qquad (4.1)$$

Notice that (4.1) is similar to (3.10). In fact, in this section $W_k^{(i)}$ plays a role that is analogous to the role of the residual service time $S_k^{(i)}$ in Section 3. For FIFO priority, the waiting time for job k is $W_k = \min_{i=1,..,c} W_k^{(i)}$. K_0 is as in (3.2), but instead of $\tilde{K}_n$ we now use $\tilde{K}_w = \inf\{ k\colon W_k \geq w \}$ and $K_w = \min\{K_0, \tilde{K}_w\}$.

Next, we redefine the conjugate distribution. The *(α,w)-conjugate distribution* is defined just like the (α,n)-conjugate distribution, except that we replace $J_k^{(i)}$ by $\hat{J}_k^{(i)}$ and we replace K_n by K_w. The key difference between the two definitions is that in the new definition at the instant that job K_w arrives we have twisted the distributions of the service times for jobs that are in queue (but not already serviced). In contrast, in the Section 3 definition the service time distribution twisting is applied only for jobs that have been or are being serviced, but not to jobs in queue when job K_n arrives.

Using the above definitions we can prove Theorem 2 in a fashion completely analogous to the way we prove Theorem 1. For brevity we shall not investigate the now obvious parallels further.

5. CONCLUSION AND NUMERICAL EXAMPLES

We conclude with some numerical data obtained using the Monte Carlo simulation. We actually first estimate the expected c-cycle length $E_\pi[L_1]$ and the quantities $E_\pi[L_1^{(n)}]$ and $E_\pi[L_1^{(w)}]$, and then apply these estimates to formulas (2.3) and (2.4) to obtain estimates of $\mathcal{P}(Q_k \geq n)$ and $\mathcal{P}(W_k > w)$.

Our algorithm is as follows. First, i.i.d. samples of the residual service time vector and the c-cycle length $(\tilde{\mathbf{B}}_0,L_1)$ were obtained using the regeneration scheme described in Section 2. The queueing system was simulated in a continuous fashion keeping track of the instances when new c-cycles begin. At the beginning of each new c-cycle, the algorithm stores the residual service time vector and the length of the last c-cycle. When a regeneration occurs, that is, when the service system completely empties out during some c-cycle, the simulation stops and one of the $(\tilde{\mathbf{B}}_0,L_1)$ pairs is randomly selected (with uniform likelihood) from the stored regeneration block. Thus, successive $(\tilde{\mathbf{B}}_0,L_1)$ are i.i.d. Moreover, it turns out that this is equivalent to sampling from the stationary distribution. The i.i.d. samples of L_1 were used to estimate the expected c-cycle length $E_\pi[L_1]$. To estimate $E_\pi[L_1^{(n)}]$ (or $E_\pi[L_1^{(w)}]$) we used the importance sampling technique. Busy periods were simulated using the i.i.d. samples of $\tilde{\mathbf{B}}_0$ and (θ,n)-conjugate distribution. Unbiased Monte Carlo estimates are then obtained using formulas (3.11) as the importance sampling weighting function. We direct the reader to references [6], [7] and [8] for more discussion of importance sampling.

For particular example presented here, inter-arrival time distribution is uniform on the interval [0,1] and service time distributions are uniform on the interval [0,.8c]. In this way, the utilization factor (see (2.2)) is $\rho = .8$ for all values of c. Moreover, the characteristic equation is exactly the same for different c. By numerical solution of the characteristic equation, we determine that $\theta = 1.486$ and $\omega = .5206$ for all $c \geq 1$. Table 1 and Figures 1 and 2 present the numerical results of the simulations.

Table 1: Summary of c-cycle simulation data

c	Expected c-cycle Length	Expected # of c-cycle per regeneration	K_Q	K_W
1	3.13	1.00	.469	.077
3	4.04	9.27	.371	.072
5	4.93	60.19	.306	.064
7	5.62	481.0	.286	.062

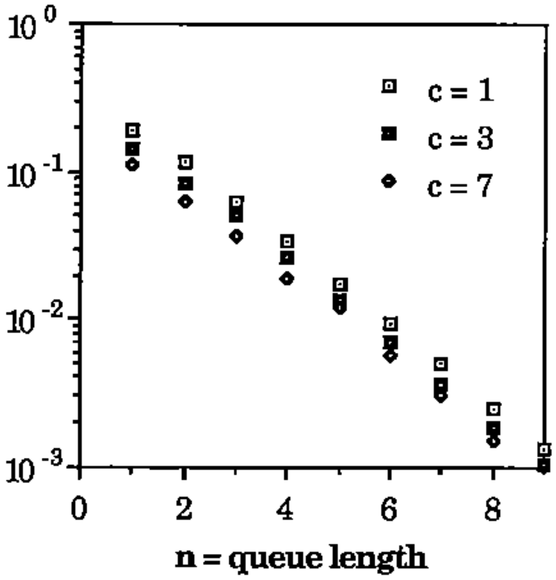

Figure 1: Estimate of $\mathcal{P}(Q_k \geq n)$

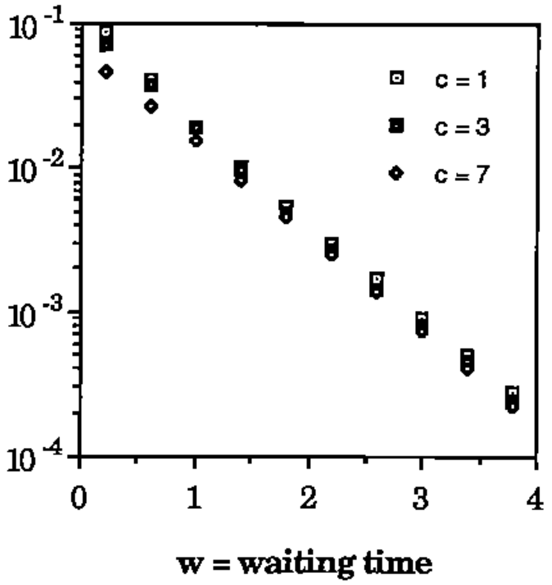

Figure 2: Estimate of $\mathcal{P}(W_k > w)$

REFERENCES

[1] Feller, W. *An Introduction to Probability and its Applications, Vol. II*. (John Wiley & Sons, New York, 1971)

[2] Amussen, S., Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the GI/G/1 queue, *Adv. Appl. Prob.* **14** (1982) 143-170.

[3] Kingman, J. F. C., On the algebra of queues, *J. Appl. Prob.* **3** (1966) 285-326.

[4] Takahashi, Y., Asymptotic exponentially of the tail of the waiting-time distribution in a PH/PH/c queue, *Adv. Appl. Probab.* **13** (1981) 619-630.

[5] Neuts, M. and Takahashi, Y., Asymptotic behavior of the stationary distribution in the GI/PH/c queue with heterogeneous servers, *Z. Wahr.* **57** (1981) 441-452.

[6] Parekh, S. and Walrand, J. A quick simulation method for excessive backlogs in networks of queues, *IEEE Trans. Auto. Control* **AC-34** (1989) 54-66.

[7] Glynn, P. W. and Iglehart, D. L., Importance sampling for stochastic simulation, *Management Sci.* **35** (1989) 1367-1392.

[8] Sadowsky, J. S., Large deviations theory and efficient simulation of excessive backlogs in a GI/GI/m queue, *IEEE Trans. Auto. Control.* (1990).

[9] Ney, P. and Nummelin, E., Markov additive processes I: eigenvalue properties and limit theorems, II: large deviations, *Ann. Probab.*, **15** (1987) 561-609.

[10] Nummelin, E., *General Irreducible Markov Chains and Non-negative Operators*, (Cambridge University Press, Cambridge, 1984).

[11] Kiefer, J. and Wolfowitz, J., On the theory of queues with many servers, *Trans. Amer. Math. Soc.* **78** (1955) 1-18.

[12] Iglehart, D., Extreme values in the GI/GI/1 queue, *Ann. Math. Statist.*, **43** (1972) 627-635.

[13] Anderson, C. W., Extreme value theory for a class of discrete distributions with applications to some stochastic processes, *J. Appl. Prob.*, **7** (1970) 99-113.

[14] Anatharam, V., How large delays build up in a GI/G/1 queue, *Queueing Systems* **5** (1988) 345-368.