Queueing Syst (2006) 54:45–54
DOI 10.1007/s11134-006-8218-z

# Handling load with less stress

**Nikhil Bansal · David Gamarnik**

Received: 19 July 2004 / Revised: 9 March 2006


**Abstract** We study how the average performance of a system degrades as the load nears its peak capacity. We restrict our attention to the performance measures of average sojourn time and the large deviation rates of buffer overflow probabilities. We first show that for certain queueing systems, the average sojourn time of requests depends much more weakly on the load $\rho$ than the commonly observed $1/(1-\rho)$ dependence for most queueing policies. For example, we show that for an M/G/1 system under the preemptive Shortest Job First (pSJF) policy, the average sojourn time varies as $\log(1/(1-\rho))$ with load for a certain class of distributions.

We observe that such results hold even for more restricted policies. We give some examples of non-preemptive policies and policies that do not use the knowledge of job sizes while scheduling, where the dependence of average sojourn time on load is significantly better than $1/(1-\rho)$. Similar results hold even for very simple non-preemptive threshold based policies that partition all the jobs into two job classes based on a fixed threshold and do FIFO within each class. Finally we study the large deviations rate of the queue length under a simple dedicated partition-based policy.

**Keywords** M/G/1 queues · Average solourn time · Heavy traffic · Heavy tailed distributions · Large deviations.

## 1 Introduction

We consider two commonly used performance measures in queueing systems: The average sojourn time and the large deviations rate for the steady state queue length. The sojourn time of a job is defined as the difference between its completion time and its arrival time, and the average sojourn time refers to the sojourn time averaged over all jobs. The large deviations rate for the steady state queue length $Q$ is defined as

$$\lim_{x\to\infty} \frac{\log(\Pr(Q > x))}{x}.$$

We will restrict our attention to M/G/1 queueing systems, that is, where the job arrival process is Markovian and the job sizes are independent and identically distributed. For an M/G/1 system, note that the performance measures considered in this paper are completely determined by the service policy used, the load of the system ($\rho$) and the job size distribution. We will be interested in how these performance measures degrade as the load approaches 1.

Recall that for any M/G/1 system the Processor Sharing (PS) policy has an average sojourn time of $E[S]/(1-\rho)$, and the classic Pollaczek-Khintchine formula which shows that the average sojourn time under the First-In-First out (FIFO) policy for any M/G/1 system is $\rho E[S^2]/(2E[S](1-\rho)) + E[S]$. Here, the random variable $S$ denotes the size of a job. Note that for both of these policies, the average sojourn time varies as $1/(1-\rho)$ with load. The $1/(1-\rho)$ dependence of the average sojourn time also holds more generally for certain classes of policies. It is known [3] that for a general M/G/1 system, any policy that is both non-preemptive and non-size based, has an average sojourn time of $\rho E[S]^2/2E[S](1-\rho) + E[S]$. Non-sized based policies are policies that do not use the knowledge of job size in their scheduling decisions. Some common examples are FIFO, PS, Last-Come-First-Served (LCFS) and Foreground-Background (FB). Moreover, for an M/M/1 system any non-size based policy (possibly preemptive) has an average

N. Bansal (✉) · D. Gamarnik
IBM T.J. Watson Research Center, Yorktown Heights, 10598
e-mail: {nikhil,daveg}@us.ibm.com.

sojourn time of $E[S]/(1-\rho)$, [3]. In addition, various conservation laws are known for fairly general classes of scheduling policies, [7][Page 197], [13][Page 440] and [8]. These laws essentially state that for any policy in the class, a certain weighted sum of sojourn times of jobs is proportional to $1/(1-\rho)$.

We are interested in the question whether there exist queueing systems where the average sojourn time can have a "better" dependence on load than $1/(1-\rho)$. The notion of dependence on load is formally defined in Section 1.1, but intuitively we say that the dependence is better than $1/(1-\rho)$, if the average sojourn time at load $\rho$ divided by $1/(1-\rho)$ approaches 0 as $\rho$ approaches 1. While there has been extensive work on studying the optimality properties and obtaining exact/approximate results for different performance measures for several queueing policies (see for example [11, 13, 6, 7] and the references there in), the question addressed above does not seem to have received much attention. To the best of our knowledge, the only result in this spirit is due to Bansal [1] who showed that for an M/M/1 system, the policy Shortest Remaining Processing Time (SRPT) has an average sojourn time of $(1+o(1))E[S]/((1-\rho)\log(e/(1-\rho)))$. Recently, Wierman et al. (Theorem 5.8 in [12]) showed a very general lower bound on the average sojourn time achievable in any M/G/1 system. They show that for any job size distribution and any scheduling policy, the average sojourn time must be at least $(E[S]\cdot\log(1/(1-\rho))/\rho)$.

In this paper we show that the lower bound due to Wierman et al. is tight up to a constant factor independent of $\rho$. In particular, the average sojourn time depends on load as $\log(1/(1-\rho))$ for the preemptive Shortest Job First policy under certain heavy-tailed distributions. This closes the (large) gap between the lower bound due to [12] and previously known upper bound due to [1].

We next explore if such an improvement holds for a more restricted class of policies. As mentioned previously, for any policy that is both non-preemptive and non-size based, the average sojourn time varies as $1/(1-\rho)$ for all job size distributions. Thus, any policy with load dependence better than $1/(1-\rho)$ must either be preemptive, or size-based. We show that preemption suffices to improve the dependence on load. In particular, there exists a non-size based (but preemptive) policy for which the load varies as $\log(1/(1-\rho))$ for certain job size distributions. Similarly, we show that being non-size based suffices. In particular, there exists a non-preemptive (but size-based) policy which has a load dependence better than $1/(1-\rho)$. In Section 3 we consider a class of policies that we call threshold based policies (defined formally in Section 1.1) which are non-preemptive and make very limited use of the knowledge of job sizes. We show the improved load dependence holds even for these threshold based policies.

In the second part of this paper we consider the performance of scheduling policies with respect to buffer overflow probabilities. Our motivation for this problem comes from the application to routers. In routers there is a fixed size buffer and there are packets that require a fixed amount of size to store. The amount of time required to transmit the packet from the head of buffer is distributed exponentially (based on the complexity of decoding the address).

Recall that in an M/M/1 system at steady state at load $\rho$, the queue length $Q$ satisfies, $\Pr(Q \geq m) = \rho^m = e^{\log(\rho)m}$. In other words the large deviations rate for the queue length is $\log(\rho)$, which is approximately $-(1-\rho)$ when $\rho$ is close to 1. We show that for an M/M/1 system, applying a very simple dedicated partition-based policy (described formally in Section 1.1) can increase the large deviations rate by a constant factor. While the properties of threshold and class-based policies have been extensively studied previously, and it is known that they can lead to substantial performance improvements, particularly for highly variable job size distributions [13, 5], we do not know of any related previous work in the regime of large deviations rate.

### 1.1 Preliminaries

Throughout this paper we will consider M/M/1 and M/G/1 settings only. We use $S$ to denote the random variable that corresponds to the service time or the size of a job. We use $f(x)$ and $F(x)$ to denote the density function and the cdf of the job size distribution. The first and second moments of job sizes are denoted by $E[S]$ and $E[S^2]$ respectively. We also use $\mu$ to denote the service rate which is equal to $1/E[S]$ and $\rho$ to denote the load, defined as $\lambda/\mu$.

We will only consider two types of job size distributions. The first is the exponential distribution with rate parameter $\mu$, defined by $f(x) = \mu e^{-\mu x}$ for $x \geq 0$. The standard Pareto($\alpha$) distribution is given as $F(x) = 1 - (\frac{k}{x})^{\alpha}$ for $x \geq k$ and 0 for $x \leq k$. Thus the density function is $f(x) = \alpha k^{\alpha} x^{-\alpha-1}$ for $x \geq k$ and 0 otherwise. The parameter $\alpha$ is assumed to be greater than 1 so that the mean job size is finite. For a scheduling policy $P$, we use $E[T(x)]_P$ to denote the average sojourn time of a job of size $x$ under $P$, and $E[T]_P$ to denote the average sojourn time under $P$. Clearly, $E[T]_P = \int_0^{\infty} f(x)E[T(x)]_P dx$.

We will be interested in the performance measure of a policy as the load approaches 1. In particular, for a fixed job size distribution and a scheduling policy, we increase the arrival rate $\lambda$ and study the performance as the load approaches 1. For a function $f$, we say that the average sojourn time varies as $O(f(\rho))$ if there is a constant $c$ independent of $\rho$, and a constant $\rho_0 < 1$, such that for all $\rho_0 \leq \rho < 1$, the average sojourn time at load $\rho$ does not exceed $c \cdot f(\rho)$. For two functions $f(\rho)$ and $g(\rho)$, we say that $f$ has a better dependence on $\rho$ than $g$, or that $f(\rho) = o(g(\rho))$, if for every $\epsilon > 0$

there exists $\rho_\epsilon < 1$ such that $f(\rho) \leq \epsilon \cdot g(\rho)$ for all $\rho_\epsilon < \rho < 1$.

We use pSJF to denote the Preemptive Shortest Job First policy, which at any time works on the job with the smallest size. Clearly, pSJF is both preemptive and is size-based. We use nSJF to denote the non-preemptive version of Shortest Job First policy. FB will denote the Foreground-Background policy, which at any time works on the job that has received the least amount of service thus far. Note that FB is a preemptive, non-size based policy.

A threshold based policy with threshold level $a$ classifies jobs into two classes, small and big, depending on whether the service time of a job is less than $a$ or not. The policy is otherwise oblivious to the actual service times. Within each class the jobs are processed in the FIFO order. The policy is non-preemptive and gives preference to small jobs over big jobs. In particular, whenever a job finishes, the next job to be executed is big if and only if there is no small job in the system. As the average sojourn time varies as $O(1/(1-\rho))$ for any non-preemptive, non-size based policy, the class of threshold based policies is one of the most restricted classes of policies for which one can hope to have a load dependence better than $1/(1-\rho)$.

Our dedicated partition-based policy in Section 4 works as follows. The jobs are divided into two classes, small and big, for some threshold $a$, depending on whether the service time is less than $a$ or not. For some parameter $\psi \in (0, 1)$ we allocate exactly $\psi$ portion of the server to serving small jobs and remaining portion to big jobs. Within each class the jobs are served in FIFO order. The partition of the server is dedicated in the sense that even if there are no small jobs in the system, the server gives only $1-\psi$ portion of its capacity to the big jobs and vice vs. In particular, if one of the classes is empty, the full system capacity may not be used.

### 1.2 Our results

We show the following results about average sojourn time:

- For an M/G/1 system with the pSJF policy, the average sojourn time is $O(\log(1/(1-\rho)))$ if the job sizes have a Pareto distribution with parameter $\alpha \in (1, 2)$. This shows that the lower bound due to Wierman et al [12] on the sojourn time of any scheduling policy in an M/G/1 system is tight.
  For other values of $\alpha$, the average sojourn time is $O(\log^2(1/(1-\rho)))$ for $\alpha = 2$, and $O((1-\rho)^{-(\alpha-2)/(\alpha-1)})$ for $\alpha > 2$.
- For Pareto job size distributions, FB has identical performance as pSJF (up to constant factors that only depend on $\alpha$). Thus being size-based is not necessary to obtain load dependence better than $1/(1-\rho)$.
  For Pareto distributions with $\alpha > 2$, we show that the average sojourn time under the nSJF policy is $O((1-\rho)^{-(\alpha-2)/(\alpha-1)})$, which implies that preemption is not necessary either.
- For the simple M/M/1 system, there is a threshold based policy with an average sojourn time of $(2 + o(1))E[S]/((1-\rho)\log(e/(1-\rho)))$, which is twice worse than that achievable by any arbitrary scheduling policy in an M/M/1 system [1]. For Pareto distribution with $\alpha > 2$, there exist threshold policies for which the average sojourn performance varies as $O((1-\rho)^{-(\alpha-1)/\alpha})$.

For the large deviations rate of the buffer overflow probabilities we consider an M/M/1 system and show that a dedicated partition-based policy can increase the large deviations rate by a constant factor. In particular, for a certain choice of $a$ and $\psi$, the large deviations rate is about $1.37(\rho - 1)$ as the load approaches 1. Since the overflow probability depends exponentially on the large deviations rate, this implies a significant reduction in the buffer overflow probability. Since the partitions of the server are dedicated and the full system capacity may not be used if one of the classes is empty, we find it somewhat surprising that the overflow probability is reduced.

## 2 Average sojourn time

We analyze the load dependence for three well studied policies: pSJF, FB and nSJF. Recall that FB is non-size based, nSJF is non-preemptive where as pSJF is completely general. FB is the natural candidate for a non-size based policy, as FB achieves the optimum average sojourn time among all non-sized based policies for job size distributions with a decreasing failure rate [9, 4]. While SRPT is the optimum policy for minimizing the average sojourn time for any job size distribution [10], we consider pSJF instead of SRPT as it has a relatively simpler analytic expression for the average sojourn time.

### 2.1 Preemptive SJF policy in the M/G/1 queueing system

We study the load dependence of sojourn times under the Preemptive Shortest Job First (pSJF) policy for Pareto distributions. Let $E[T(x)]_{pSJF}$ denote the average sojourn time of a job of size $x$ under pSJF, and let $E[T]_{pSJF}$ denote the average sojourn time under pSJF. Recall that the Pareto distribution has density $f(x) = \alpha k^\alpha x^{-\alpha-1}$ for $x \geq k$ and 0 otherwise. The mean job size $E[S]$ is $\int_0^\infty x f(x)dx = k\alpha/(\alpha-1)$. Let $\rho(x) = \lambda \int_0^x t f(t)dt$ denote the load made up by jobs of size less than or equal to $x$. The main result of this section is as follows.

**Theorem 1.** *For the Preemptive SJF policy with Pareto job size distribution, the average sojourn time $E[T]_{pSJF}$ satisfies*

$$E[T]_{pSJF} = \begin{cases} O(\ln\left(\frac{1}{1-\rho}\right)) & \text{if } \alpha < 2 \\ O(\ln^2\left(\frac{1}{1-\rho}\right)) & \text{if } \alpha = 2 \\ O((1-\rho)^{-\frac{(\alpha-2)}{(\alpha-1)}}) & \text{if } \alpha > 2 \end{cases} \quad (1)$$

**Proof:** Consider an M/G/1 queue with the service time distributed according to $f(x)$ and mean arrival rate $\lambda$. It is well known that, [13], the average sojourn time for a job of size $x$ is $E[T(x)]_{pSJF} = E[W(x)] + E[R(x)]$ where

$$E[W(x)] = \frac{\lambda \int_0^x t^2 f(t)\,dt}{2(1-\rho(x))^2} \quad \text{and} \quad E[R(x)] = \frac{x}{1-\rho(x)}.$$

We begin by bounding the contribution of the residence time.

$$\begin{aligned} E[R] &= \int_0^\infty E[R(x)] f(x)\,dx \\ &= \int_0^\infty \frac{x f(x)\,dx}{1-\rho(x)} = \frac{1}{\lambda}\int_0^\rho \frac{d\rho(x)}{1-\rho(x)} \\ &\qquad (\text{Since } d\rho(x) = \lambda x f(x)\,dx) \\ &= \frac{1}{\lambda}\ln(1/(1-\rho)) \end{aligned} \quad (2)$$

Notice that the analysis above in fact holds for any job size density $f(x)$ and that the average residence time has a logarithmic dependence on $(1-\rho)$. Henceforth, we restrict our attention to bounding the contribution due to the waiting time.

By a simple calculation, $\rho(x) = \lambda \int_0^x t f(t)\,dt = \rho(1 - (k/x)^{\alpha-1})$, for $x \geq k$ and $\int_0^\infty x^2 f(x)\,dx = \frac{\alpha}{\alpha-2}k^2$.

As $f(x) = 0$, for $x \leq k$ for the Pareto distribution, we can write the average waiting time as

$$E[W] = \int_k^\infty \lambda f(x) \frac{\int_k^x t^2 f(t)\,dt}{2(1-\rho(x))^2}\,dx. \quad (3)$$

We first focus on the term $\int_k^x t^2 f(t)\,dt$. Again, by a straightforward calculation it can be verified that,

$$\int_k^x t^2 f(t)\,dt = \begin{cases} \frac{\alpha}{2-\alpha}k^\alpha(x^{2-\alpha} - k^{2-\alpha}) & \text{if } \alpha < 2 \\ 2k^2\ln(x/k) & \text{if } \alpha = 2 \\ \frac{\alpha}{\alpha-2}k^\alpha(k^{2-\alpha} - x^{2-\alpha}) & \text{if } \alpha > 2 \end{cases} \quad (4)$$

Consider the threshold $x_0 = k(\frac{\rho}{1-\rho})^{\frac{1}{\alpha-1}}$. Note that if $\rho \geq 1/2$, then $x_0$ is at least $k$. If $x \leq x_0$, then $(k/x)^{\alpha-1}\rho \geq 1-\rho$ and by the definition of $\rho(x)$ it follows that

$$[(k/x)^{\alpha-1}\rho]^2 \leq (1-\rho(x))^2 \leq 4[(k/x)^{\alpha-1}\rho]^2. \quad (5)$$

Similarly, for $x \geq x_0$, we have that $(k/x)^{\alpha-1}\rho \leq 1-\rho$ and hence

$$(1-\rho)^2 \leq (1-\rho(x))^2 \leq 4(1-\rho)^2. \quad (6)$$

We now upper bound $E[W]$ in equation (3). We will consider three cases, depending on whether $\alpha < 2$, $\alpha = 2$ or $\alpha > 2$.

1. When $\alpha < 2$: By equations (3), (4), (5) and (6),

$$\begin{aligned} E[W] &\leq \int_k^{x_0} \frac{\lambda\alpha k^\alpha}{x^{\alpha+1}} \cdot \frac{\alpha}{2-\alpha} \cdot \frac{k^\alpha(x^{2-\alpha} - k^{2-\alpha})}{g(x)}\,dx \\ &\quad + \int_{x_0}^\infty \frac{\lambda\alpha k^\alpha}{x^{\alpha+1}} \cdot \frac{\alpha}{2-\alpha} \cdot \frac{k^\alpha(x^{2-\alpha} - k^{2-\alpha})}{2(1-\rho)^2}\,dx \\ &\leq \int_k^{x_0} \frac{\lambda\alpha k^{2\alpha}}{x^{\alpha+1}} \cdot \frac{\alpha}{2-\alpha} \cdot \frac{x^{2-\alpha}}{g(x)}\,dx \\ &\quad + \int_{x_0}^\infty \frac{\lambda\alpha k^{2\alpha}}{x^{\alpha+1}} \cdot \frac{\alpha}{2-\alpha} \cdot \frac{x^{2-\alpha}}{2(1-\rho)^2}\,dx \end{aligned}$$

where $g(x) = 2[(k/x)^{\alpha-1}\rho]^2$.
The first term in the equation above can be simplified as

$$\begin{aligned} &\int_k^{x_0} \frac{\lambda\alpha k^{2\alpha}}{x^{\alpha+1}} \cdot \frac{\alpha}{2-\alpha} \cdot \frac{x^{2-\alpha}}{g(x)}dx \\ &= \int_k^{x_0} \lambda\alpha k^2 \cdot \frac{\alpha}{2-\alpha} \cdot \frac{1}{2\rho^2 x}dx \\ &= k(\alpha-1) \cdot \frac{\alpha}{2-\alpha} \cdot \frac{1}{2\rho}\frac{1}{\alpha-1}\ln\left(\frac{\rho}{1-\rho}\right) \end{aligned}$$

The second term can be simplified as

$$\begin{aligned} &\int_{x_0}^\infty \frac{\lambda\alpha k^{2\alpha}}{x^{\alpha+1}} \cdot \frac{\alpha}{2-\alpha} \cdot \frac{x^{2-\alpha}}{2(1-\rho)^2}\,dx \\ &= \int_{x_0}^\infty \lambda\alpha k^{2\alpha} \cdot \frac{\alpha}{2-\alpha} \cdot \frac{x^{1-2\alpha}}{2(1-\rho)^2}\,dx \\ &= \lambda\alpha k^{2\alpha} \cdot \frac{\alpha}{2-\alpha} \cdot \frac{1}{2\alpha-2}\frac{x_0^{2-2\alpha}}{2(1-\rho)^2} \\ &= k(\alpha-1) \cdot \frac{\alpha}{2-\alpha} \cdot \frac{1}{2\alpha-2}\frac{1}{2\rho} \end{aligned}$$

The last step follows by substituting the values of $x_0$ and $\rho$.

2. When $\alpha = 2$: By equations (3), (5) and (6) we have,

$$\begin{aligned} E[W] &\le \int_k^{x_0} \frac{2\lambda k^2}{x^3} \frac{2k^2x^2\ln(x/k)}{2\rho^2k^2}dx \\ &\quad + \int_{x_0}^{\infty} \frac{2\lambda k^2}{x^3} \frac{2k^2\ln(x/k)}{2(1-\rho)^2}dx \\ &= \int_k^{x_0} \frac{2\lambda k^2\ln(x/k)}{\rho^2 x}dx + \int_{x_0}^{\infty} \frac{2\lambda k^4\ln(x/k)}{x^3(1-\rho)^2}dx \\ &= \frac{\lambda k^2\ln^2(x_0/k)}{\rho^2} + \frac{\lambda k^4(1+2\ln(x_0/k))}{4x_0^2(1-\rho)^2} \\ &= \frac{k\ln^2\left(\frac{\rho}{1-\rho}\right)}{2\rho} + \frac{k(2\ln\left(\frac{\rho}{1-\rho}\right)+1)}{8\rho} \end{aligned}$$

The last step follows by substituting the values of $\lambda = \rho(\alpha-1)/(\alpha k) = \rho/(2k)$ and $x_0$.
As $\alpha = 2$, the average sojourn time is easily seen to be $O(\ln^2\left(\frac{\rho}{1-\rho}\right))$.

3. When $\alpha > 2$: By equations (3), (5) and (6)

$$\begin{aligned} E[W] &\le \int_k^{x_0} \frac{\lambda\alpha k^{2\alpha}}{x^{\alpha+1}} \cdot \frac{\alpha}{\alpha-2} \cdot \frac{k^{2-\alpha}}{g(x)}dx \\ &\quad + \int_{x_0}^{\infty} \frac{\lambda\alpha k^{2\alpha}}{x^{\alpha+1}} \cdot \frac{\alpha}{\alpha-2} \cdot \frac{k^{2-\alpha}}{2(1-\rho)^2}dx \end{aligned}$$

where $g(x) = 2[(k/x)^{\alpha-1}\rho]^2$.
We now calculate the first term,

$$\begin{aligned} &\int_k^{x_0} \lambda k\alpha \cdot \frac{\alpha}{\alpha-2} \cdot \frac{k^{3-\alpha}x^{\alpha-3}}{2\rho^2}dx \\ &= \frac{(\alpha-1)}{2\rho} \cdot \frac{\alpha k}{(\alpha-2)^2} \cdot \left[\left(\frac{\rho}{1-\rho}\right)^{\frac{\alpha-2}{\alpha-1}} - 1\right] \end{aligned}$$

The second term is simply

$$\begin{aligned} &\int_{x_0}^{\infty} \frac{\lambda\alpha k^{2\alpha}}{x^{\alpha+1}} \cdot \frac{\alpha}{\alpha-2} \cdot \frac{k^{2-\alpha}}{2(1-\rho)^2}dx \\ &= \lambda\alpha k^{\alpha+2} \cdot \frac{1}{\alpha-2} \cdot \frac{x_0^{-\alpha}}{2(1-\rho)^2} \\ &= \frac{k(\alpha-1)}{2(\alpha-2)} \cdot \rho^{-\frac{1}{\alpha-1}} \cdot (1-\rho)^{-\frac{\alpha-2}{\alpha-1}} \end{aligned}$$

To finish the proof, by equation (2), we know that the average residence time is always $O(\ln(\frac{1}{1-\rho}))$. The proof of the theorem follows by considering the obtained expression for the average waiting time in all the three cases. □

### 2.2 Foreground-background policy in the M/G/1 system

We now show that a similar result to Theorem 1 holds for the non size-based policy Foreground Background. In particular, for the Pareto distribution the average sojourn time under FB is similar to that under pSJF up to constant factors more than $\alpha^2$.

**Theorem 2.** *For the FB policy with Pareto job size distribution, the average sojourn time $E[T]_{FB}$ satisfies*

$$E[T]_{FB} = \begin{cases} O(\ln\left(\frac{1}{1-\rho}\right)) & \text{if } \alpha < 2 \\ O(\ln^2\left(\frac{1}{1-\rho}\right)) & \text{if } \alpha = 2 \\ O((1-\rho)^{-\frac{(\alpha-2)}{(\alpha-1)}}) & \text{if } \alpha > 2 \end{cases} \tag{7}$$

**Proof:** Let $\rho_1(x) = \lambda\int_0^x tf(t)\,dt + \lambda x(1-F(x))$. It is well known, [6], that the expected sojourn time of a job of size $x$ under FB is

$$E[T(x)]_{FB} = \frac{\lambda}{2}\frac{\int_0^x t^2 f(t) + x^2(1-F(x))}{(1-\rho_1(x))^2} + \frac{x}{1-\rho_1(x)}.$$

This expression is similar to that of pSJF except that in both the denominators we have $1-\rho_1(x)$ instead of $1-\rho(x)$ and in the numerator of the waiting time term under FB there is the additional $x^2(1-F(x))$ term.

We first show that $1-\rho_1(x)$ is always within a constant factor of $1-\rho(x)$ for the Pareto distribution.

$$\begin{aligned} \rho_1(x) &= \lambda\int_0^x tf(t)\,dt + \lambda x(1-F(x)) \\ &= \rho\cdot(1-(k/x)^{\alpha-1}) + \lambda k(k/x)^{\alpha-1} \\ &= \rho\cdot(1-(k/x)^{\alpha-1}) + (\alpha-1)(\rho/\alpha)(k/x)^{\alpha-1} \\ &= \rho\cdot(1-\frac{1}{\alpha}(k/x)^{\alpha-1}) \end{aligned}$$

Thus,

$$1 \le \frac{1-\rho(x)}{1-\rho_1(x)} \le \frac{\rho-\rho(x)}{\rho-\rho_1(x)} = \alpha$$

which implies that the contribution due to the denominators of the waiting time under pSJF and FB do not differ by factors that depend on $\alpha$.

We now account for the term $x^2(1-F(x))$. Notice that for the Pareto distribution for $\alpha < 2$, $x^2(1-F(x)) = k^\alpha x^{2-\alpha}$. In bounding the average waiting time under pSJF for this case, we only use the fact that for pSJF, $\int_0^x t^2 f(t)dt = O(k^\alpha x^{2-\alpha})$ and hence the expression for FB changes by a constant factor only.

Similarly for $\alpha = 2$, $x^2(1 - F(x)) = k^2$. Under pSJF, the corresponding term $\int_0^x t^2 f(t)dt$ is $O(k^2 \log(x/k))$. Thus the expression for the waiting time for FB changes by only a constant factor as compared with that for pSJF.

Finally for $\alpha \geq 2$, $x^2(1 - F(x)) = k^\alpha x^{-\alpha+2}$. In bounding the waiting time under pSJF in this case, we only use the fact that $\int_0^x t^2 f(t)dt = O(k^\alpha x^{-\alpha+2})$. Thus the result for the waiting time under pSJF also holds for FB (up to constant factors). □

### 2.3 Non-preemptive SJF policy in the M/G/1 system

Recall that if the job size distribution has infinite variance, then the average sojourn time for any non-preemptive policy is infinite. Hence throughout this subsection we assume that $\alpha > 2$.

The expression for the expected response time for a job of size $x$ under the nSJF policy is well known [13].

In particular, for a continuous job size distribution the expected sojourn time for a job of size $x$ is

$$E[T(x)]_{nSJF} = \lambda \frac{\int_0^\infty t^2 f(t)\,dt}{2(1-\rho(x))^2} + x$$
$$= \rho \frac{E[S^2]}{2E[S](1-\rho(x))^2} + x.$$

**Theorem 3.** *For the non-preemptive SJF policy with Pareto($\alpha$), $\alpha > 2$ job size distributions the average sojourn time $E[T]_{nSJF}$ satisfies*

$$E[T]_{nSJF} = O((1-\rho)^{-(\frac{\alpha-2}{\alpha-1})}) \tag{8}$$

**Proof:** As in the proof of Theorem 1, we choose $x_0 = k(\frac{\rho}{1-\rho})^{\frac{1}{\alpha-1}}$ and rewrite the above integral as

$$E[T] = \int_x E[T(x)] f(x)\,dx$$
$$= \rho \frac{E[S^2]}{E[S]} \left( \int_k^{x_0} \frac{f(x)}{2(1-\rho(x))^2}\,dx + \int_{x_0}^\infty \frac{f(x)}{2(1-\rho(x))^2} dx \right) + E[S].$$

By equations (5) and (6), this sum can be upper bounded by

$$\rho \frac{E[S^2]}{E[S]} \left( \int_k^{x_0} \frac{f(x)}{2((k/x)^{\alpha-1}\rho)^2}\,dx + \int_{x_0}^\infty \frac{f(x)}{2(1-\rho)^2}\,dx \right).$$

Since $f(x) = \alpha k^\alpha x^{-\alpha-1}$ for $x \geq k$, $E[S^2] = \alpha k^2/(\alpha-2)$ and $E[S] = \alpha k/(\alpha-1)$, the above quantity is equal to

$$\rho \frac{(\alpha-1)k}{\alpha-2} \left( \int_k^{x_0} \frac{\alpha k^\alpha x^{-\alpha-1}}{2(k/x)^{2\alpha-2}\rho^2} dx + \frac{k^\alpha x_0^{-\alpha}}{2(1-\rho)^2} \right)$$

which by a straightforward calculation can be seen to be $O((1-\rho)^{-(\frac{\alpha-2}{\alpha-1})})$. □

## 3 Non-preemptive threshold based policies

We consider the non-preemptive threshold based policy defined in Section 1.1, with the threshold level $x_0$.

We view the class of these policies as an extreme simplification of nSJF. Note that nSJF can be viewed as threshold based policy with infinitely many thresholds (one for each $x$ for every $x > 0$). Also note that if $x_0$ is set to infinity then the policy corresponds to FIFO. Since the average sojourn time depends as $1/(1-\rho)$ on load for non-preemptive, non-size based policies, the threshold policy defined above is one of the most restricted classes of policies for which one can hope to have a load dependence better than $1/(1-\rho)$.

Let $\rho_1$ denote the load comprised of class 1 jobs. Using the well-known results for average sojourn time under non-preemptive class based priority systems [13][Page 441], we have that

$$E[T] = E[S] + \rho \frac{E[S^2]}{2E[S]} \left( \frac{F(x_0)}{1-\rho_1} + \frac{1-F(x_0)}{(1-\rho_1)(1-\rho)} \right). \tag{9}$$

### 3.1 Pareto distributions

We construct a threshold policy for the Pareto distribution which has a better dependence on load than $1/(1-\rho)$. Since Pareto distributions have infinite variance for $\alpha \leq 2$ and our threshold policy is non-preemptive, we restrict our attention to the case when $\alpha > 2$.

**Theorem 4.** *Suppose the service times in an M/G/1 system have a Pareto distribution with parameter $\alpha > 2$ and a threshold based priority scheduling policy is used with threshold level $x_0 = k(1-\rho)^{-1/\alpha}$. Then*

$$E[T] = O((1-\rho)^{-(\frac{\alpha-1}{\alpha})}).$$

**Proof:** Choose $x_0$ such that $F(x_0) = \rho$. Thus, $(k/x_0)^\alpha = (1-\rho)$ and hence $x_0 = k(1-\rho)^{-1/\alpha}$. Thus by equation (9)

we get,

$$E[T] = E[S] + \rho\frac{E[S^2]}{2E[S]}\left(\frac{F(x_0)}{1-\rho_1} + \frac{1}{1-\rho_1}\right)$$
$$\leq E[S] + \rho\frac{E[S^2]}{2E[S]}\left(\frac{2}{1-\rho_1}\right).$$

By a straightforward calculation, $1-\rho_1 = 1-\rho+(\frac{k}{x_0})^{\alpha-1}\rho$. As $(k/x_0) = (1-\rho)^{1/\alpha}$, it follows that

$$1-\rho_1 = 1-\rho+\rho(1-\rho)^{\frac{\alpha-1}{\alpha}}$$

which implies the desired result. □

### 3.2 Exponentially distributed job sizes

We now show that even for the M/M/1 system, there is a simple threshold policy such that the average sojourn time is within a factor 2 of the smallest average sojourn time achievable by any arbitrary policy.

**Theorem 5.** *Consider an M/M/1 system with processing rate $\mu$. Suppose a threshold based priority scheduling policy is used with threshold level $x_0 = \frac{1}{\mu}\log\frac{\rho}{1-\rho}$. Then*

$$E[T] \leq \frac{2+o(1)}{\mu(1-\rho)\log\left(\frac{e}{1-\rho}\right)}.$$

**Proof:** Choose $x_0$ such that $F(x_0) = \rho$. Thus, $e^{-\mu x_0} = 1-\rho$ or $x_0 = \frac{1}{\mu}\log\frac{1}{1-\rho}$. Using Equation (9) we obtain

$$E[T] \leq E[S] + \rho\frac{E[S^2]}{2E[S]}\left(\frac{\rho+1}{1-\rho_1}\right).$$

For the exponential distribution, we have that $E[S] = 1/\mu$ and $E[S^2] = 2/\mu^2$. Thus

$$E[T] = \frac{1}{\mu}\cdot\left(\frac{\rho+1}{1-\rho_1}\right) \leq \frac{2}{\mu(1-\rho_1)}.$$

A direct calculation gives that $\rho_1 = \rho\mu\int_0^{x_0} tf(t)dt = \rho^2 - \rho(1-\rho)\log\left(\frac{1}{1-\rho}\right)$. Thus,

$$1-\rho_1 = 1-\rho^2+\rho(1-\rho)\log\left(\frac{1}{1-\rho}\right)$$
$$\geq \rho(1-\rho)\log\left(\frac{e}{1-\rho}\right).$$

As $1/\rho = 1+o(1)$ as $\rho\to 1$, this implies the desired result. □

## 4 Large deviation of queue lengths

We consider an M/M/1 queue with arrival rate $\lambda < 1$ and processing rate $\mu = 1$. Thus $\rho = \lambda$. Recall that in steady state $\Pr(Q \geq m) = \rho^m = e^{\log(\rho)m}$, and hence the large deviations rate for the queue length is $\log(\rho)$, which is approximately $\rho - 1$ when $\rho$ is close to 1. We now consider the class of dedicated partition-based policies described in Section 1.1. A policy in this class is completely characterized by the parameters $a$ and $\psi$. We compute the large deviations rates $\theta_1, \theta_2$ corresponding to steady state queue lengths of classes with small jobs and big jobs respectively. We will show that for certain parameters $a, \psi$, $\max(\theta_1, \theta_2) < \log(\rho)$. Since the overall large deviations rate for steady state queue length is given as $\min(\theta_1, \theta_2)$, our policy achieves an improvement over FIFO policy in M/M/1 in terms of the large deviations rates performance.

By Theorem 4.3 [2], the large deviations rate for the steady state queue length in a G/G/1 system with the FIFO policy (provided it exists) is given as $\Lambda_A(\theta)$ where $\theta$ is the smallest negative root of the equation

$$\Lambda_A(\theta) + \Lambda_B(-\theta) = 0 \tag{10}$$

and $\Lambda_A(s) = \log E[e^{sA}]$, $\Lambda_B(s) = \log E[e^{sB}]$ and $A, B$ are random inter-arrival and service times. In particular for the M/M/1 system the equation becomes

$$\frac{\lambda}{\lambda-\theta}\cdot\frac{\mu}{\mu+\theta} = 1$$

implying $\theta = \lambda - \mu$ and the large deviations rate becomes $\log E[e^{(\lambda-\mu)A}] = \log(\lambda/\mu) = \log(\rho)$ as expected. [2] also give a sufficient condition (see Assumption B in [2]) for the large deviations rate to exist. We use formula (10) to compute the large deviations rates for individual queues in our scheduling policy.

### 4.1 The system with large jobs

Let us consider first the system with large jobs. In this case, the distribution of job sizes is given by $f(x) = e^a e^{-x}$, if $x > a$ and 0 otherwise. Thus the average job size is $a+1$. The arrival rate of jobs to this system is $\lambda e^{-a}$. We will set $\psi = e^{-a}(a+1)$.

However, note that since we are only giving a portion $\psi$ to this server, this will stretch the service times in this system by a factor of $1/\psi$. In particular this implies that a

job of size $x$ in the original system will have size $x/\psi$ in this system. To remove this effect of scaling we do the following trick: Observe that if we scale the processing and the arrival rate of any queueing system simultaneously by a factor $c$, then the distribution of the queue length remains unchanged. So, in this case we will choose $c = 1/\psi = e^a/(a+1)$. This transformation has the effect that a job of size $x$ in the original system also requires $x$ units of service in our current system. After this scaling however, the arrival rate becomes $\lambda' = \lambda e^{-a}/\psi = \lambda/(a+1)$. Thus, from now on we will consider the system with the job size distribution $f(x) = e^a e^{-x}$, if $x > a$ and 0 otherwise, and arrival rate $\lambda' = \lambda/(a+1)$.

We note that Assumption B in [2] is trivially satisfied for our system as both the inter-arrival times and the service times are i.i.d. and have finite moments of all orders, which implies the existence of the large deviations rate. We now compute the relevant quantities. Clearly,

$$\Lambda_B(-\theta) = \log \int_a^\infty e^a e^{-x} e^{-\theta x} dx = \log\left(\frac{e^{-\theta a}}{\theta+1}\right).$$

Since the inter-arrival distribution is still Poisson with rate $\lambda'$ we have that $\Lambda_A(\theta) = \log(\frac{\lambda'}{\lambda'-\theta})$. Thus to obtain the large deviations rate, we solve for the smallest root of

$$\frac{\lambda'}{\lambda'-\theta} \cdot \frac{e^{-\theta a}}{\theta+1} = 1 \tag{11}$$

Clearly, the smallest root $\theta^*$ of this equation is negative (otherwise the large deviations rate is undefined). We now show that as $\lambda \to 1$, then $\theta^* \to 0$. Of course this is fully expected since as the system approaches heavy-traffic the large deviations rate must approach zero.

To see this, we rewrite equation (11) as

$$\lambda(1 - e^{-\theta a}) = (a+1)\theta^2 - \lambda\theta + (a+1)\theta \tag{12}$$

As $\theta^* < 0$, we have that $1 - e^{-\theta^* a} \le a\theta^*$. Thus, by equation (12)

$$\lambda a\theta^* \ge (a+1)\theta^{*2} - \lambda\theta^* + (a+1)\theta^*$$

As $\theta^* < 0$, this implies that

$$\lambda a \le (a+1)(\theta^* + 1) - \lambda$$

and hence,

$$\theta^* \ge (\lambda - 1)$$

Since $\theta^* < 0$, the above equation implies that as $\lambda \to 1$, then $\theta^* \to 0$. Our goal now is to obtain $\lim_{\lambda\to 1} \frac{\theta^*}{\lambda-1}$. As $\theta^*$ satisfies equation (12), we have that

$$\frac{\lambda - 1}{\theta^*} = \frac{(a+1)\theta^{*2} + (a+1)\theta^* - 1 + e^{-\theta^* a} - \theta^*}{\theta^*(1 - e^{-\theta^* a} + \theta^*)} \tag{13}$$

Applying the L'Hopital rule twice,

$$\lim_{\theta^*\to 0} \frac{(a+1)\theta^{*2} + (a+1)\theta^* - 1 + e^{-\theta^* a} - \theta^*}{\theta^*(1 - e^{-\theta^* a} + \theta^*)} = \frac{2(a+1) + a^2}{2(a+1)}. \tag{14}$$

As $\lambda' = 1/(a+1)$, the large deviations rate

$$\Lambda_A(\theta^*) = \log \frac{\lambda'}{\lambda' - \theta^*} = -\log(1 - (a+1)\theta^*).$$

As $\lim_{x\to 0}(\log(1+x))/x = 1$, and by equations (13) and (14) it follows that

$$\lim_{\lambda\to 1} \frac{\Lambda_A(\theta^*)}{\lambda - 1} = \lim_{\lambda\to 1} \frac{\theta^*}{\lambda - 1} \cdot \frac{\Lambda_A(\theta^*)}{\theta^*} = \frac{2(a+1)}{2(a+1) + a^2} \cdot (a+1) = \frac{2(a+1)^2}{2(a+1) + a^2}.$$

For $a = 1$, this implies that

$$\lim_{\lambda\to 1} \frac{\Lambda_A(\theta^*)}{\lambda - 1} = 1.6 \tag{15}$$

### 4.2 The system with small jobs

In this system the arrival rate of the jobs is $(1 - e^{-a})\lambda$, and the speed of the server is $1 - \psi = 1 - (a+1)e^{-a}$. As previously, for computational convenience, we scale both jobs sizes and the arrival rate by the factor $1/(1-\psi)$. The jobs sizes now have the distribution $(1 - e^{-a})^{-1} e^{-x}$ for $x \le a$ and 0 otherwise. The arrival rate $\lambda' = \lambda(1 - e^{-a})/(1 - (a+1)e^{-a})$.

Again, the existence of the large deviations rate is guaranteed as Assumption B in [2] is easily satisfied by our system. We now compute the relevant quantities. By a simple calculation it follows that

$$\Lambda_B(-\theta) = \log \frac{1 - e^{-a(1+\theta)}}{(\theta+1)(1 - e^{-a})}$$

As $\Lambda_A(\theta) = \log(\lambda'/(\lambda' - \theta))$, we need to solve for

$$\frac{\lambda'}{\lambda' - \theta} \cdot \frac{1 - e^{-a(1+\theta)}}{(\theta+1)(1 - e^{-a})} = 1 \tag{16}$$

As usual let $\theta^* < 0$ denote the smallest root of this equation. We first note that $\theta^* \geq -1$, because $(1 - e^{-a(1+\theta)})/((\theta+1)(1-e^{-a})) < 1$ for $\theta < -1$ and clearly $\lambda'/(\lambda' - \theta) < 1$ for any $\theta < 0$, and hence equation (16) cannot be satisfied for $\theta < -1$.

We will be interested in the behavior of $\theta^*$ as $\lambda$ approaches 1. In particular, we will be interested in how $\theta^*$ approaches 0 as $\lambda$ approaches 1. We also fix $a = 1$. Then equation (16) can be written as

$$\frac{e - e^{-\theta^*}}{(\theta^*+1)(e-1)} = 1 - \frac{\theta^*}{\lambda'} \tag{17}$$

We first show the weaker statement that $\theta^* > -0.8$ as $\lambda$ approaches 1. For notational convenience, we will use $f(\theta^*)$ to denote the function $(1 - e^{-(1+\theta^*)})/(1+\theta^*)$. By definition of $\lambda'$, it follows that $\lim_{\lambda\to 1} \lambda' = (e-1)/(e-2) > 2.39$. As $\theta^* \geq -1$, we have that

$$\lim_{\lambda\to 1} 1 - \frac{\theta^*}{\lambda'} \leq \lim_{\lambda\to 1}\left(1 + \frac{1}{\lambda'}\right) \leq \frac{3.39}{2.39}. \tag{18}$$

By (18) and by equation (17), it follows that

$$\lim_{\lambda\to 1} f(\theta^*) \leq \frac{3.39(1-e^{-1})}{2.29} < 0.9.$$

As $(1 - e^{-x})/x$ is a decreasing function of $x$ for $x \geq 0$, and as $\theta^* \geq -1$, it follows that $f(\theta^*)$ is a decreasing function of $\theta^*$. Now, $f(-0.8) > 0.906$, which implies that $\lim_{\lambda\to 1} \theta^* > -0.8$.

We are now ready to show that $\theta^*$ actually approaches 0 as $\lambda$ approaches 1. Using the Taylor expansion of $e^x$, and that $-0.8 \leq \theta^* < 0$, we get that

$$\begin{aligned} e^{-\theta^*} &= 1 - \theta^* + \theta^{*2}/2! - \theta^{*3}/3! + \ldots \\ &\leq 1 - \theta^* + \theta^{*2}/2! + 0.8\theta^{*2}(1/3! + 1/4! + \ldots) \\ &= 1 - \theta^* + \theta^{*2}(0.8e - 1.5) \\ &\leq 1 - \theta^* + 0.7\theta^{*2}. \end{aligned} \tag{19}$$

Plugging inequality (19) into equation (17), we find

$$\begin{aligned} 1 - \frac{\theta^*}{\lambda'} &\geq \frac{e - 1 + \theta^* - 0.7\theta^{*2}}{(\theta^*+1)(e-1)} \\ &= \frac{(e-1)(\theta^*+1) + \theta^*(2-e) - 0.7\theta^{*2}}{(\theta^*+1)(e-1)}. \end{aligned}$$

Subtracting 1 from both sides, and multiplying by $-\lambda'(\theta^* + 1)/\theta^* > 0$, we get

$$\theta^* + 1 \geq \lambda'\left(\frac{e - 2 + 0.7\theta^*}{e-1}\right).$$

After substituting $\lambda' = \lambda(e-1)/(e-2)$, and rearranging terms, we get

$$\theta^*\left(1 - \frac{0.7\lambda}{e-2}\right) \geq \lambda - 1.$$

This implies that $\theta^* \to 0$ when $\lambda \to 1$.

We now evaluate $\lim_{\lambda\to 1} \frac{\theta^*}{\lambda - 1}$. Using equation (16), we write $\lambda$ as a function of $\theta^*$. Applying the L'Hopital rule twice and replacing $\lim_{\lambda\to 1}$ by $\lim_{\theta^*\to 0}$, we get

$$\lim_{\lambda\to 1} \frac{\theta^*}{\lambda - 1} = \frac{2e-4}{2e-5}. \tag{20}$$

The limit of the ratio large deviations rate divided by $\lambda - 1$ as $\lambda \to 1$ can be evaluated as

$$\begin{aligned} \lim_{\lambda\to 1} \frac{\Lambda_A(\theta^*)}{\lambda - 1} &= \lim_{\lambda\to 1} \frac{\theta^*}{\lambda - 1} \cdot \lim_{\theta^*\to 0} \frac{\Lambda_A(\theta^*)}{\theta^*} \\ &= \frac{2e-4}{2e-5} \cdot \lim_{\theta^*\to 0}\left(\frac{1}{\theta^*} \cdot \log\frac{\lambda'}{\lambda' - \theta^*}\right) \\ &= \frac{2e-4}{2e-5} \cdot \frac{e-2}{e-1} \approx 1.3755 \end{aligned} \tag{21}$$

The second step follows from equation (20) and the final step follows as $\lambda' = \lambda(e-1)/(e-2)$.

By equations (15) and (21) we have that

**Theorem 6.** *Given an M/M/1 queueing system operating under a dedicated processor sharing policy, there exists a threshold value a such that the large deviations rates $\theta_1, \theta_2$ for the two queue lengths satisfy*

$$\lim_{\rho\to 1} \frac{\min(|\theta_1|, |\theta_2|)}{|\log \rho|} \geq 1.37.$$

## References


1. N. Bansal, *On the average sojourn time under M/M/1/SRPT*, Operations Research Letters 33(2005) 195–200.
2. D. Bertsimas, I. Paschalidis, and J. Tsitsiklis, *Large deviations behavior of acyclic networks of G/G/1 queues*, The Annals of Applied Probability 8(1998) 1027–1069.
3. R. W. Conway, W. L. Maxwell, and L. W. Miller, *Theory of scheduling*, Addison-Wesley Publishing Company, 1967.
4. H. Feng and V. Misra, *Mixed scheduling disciplines for network flows (the optimality of FBPS)*, Proceedings of The Fifth Workshop on MAthematical performance Modeling and Analysis (MAMA), 2003.
5. M. Harchol-Balter, *Task assignment with unknown duration*, Journal of the ACM (JACM) 49(2002) 260–288.
6. L. Kleinrock, *Queueing systems*, John Wiley and Sons, 1975.
7. L. Kleinrock, *Queueing systems vol. 2: Computer applications*, John Wiley and Sons, 1976.

8. T.M. O'Donovan, *Distribution of attained service and residual service in general queueing systems*, Operations Research 22(1974) 570–575.
9. R. Righter and J.G. Shanthikumar, *Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures*, Probability in the Engineering and Informational Sciences 3(1989) 323–333.
10. L.E. Schrage, *A proof of the optimality of the shortest remaining processing time discipline*, Operations Research 16(1968) 678–690.
11. D. Stoyan, *Comparison methods for queues and other stochastic models*, Wiley, 1983.
12. A. Wierman, M. Harchol-Balter, and T. Osogami, *Nearly insensitive bounds on SMART scheduling*, ACM Sigmetrics, (2005) 205–216.
13. R. W. Wolff, *Stochastic modeling and the theory of queues*, Prentice Hall, 1989.