

IBM Research Report

Handling Load with Less Stress

Nikhil Bansal, David Gamarnik

IBM Research Division

Thomas J. Watson Research Center

P.O. Box 218

Yorktown Heights, NY 10598



Research Division

Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Handling load with less stress

Nikhil Bansal*

David Gamarnik[†]

July 20, 2004

Abstract

In this work we study how the average performance of a system degrades as the load nears its peaks capacity. We restrict our attention to the performance measures of average sojourn time and the large deviation rates of buffer overflow probabilities. While it is common knowledge that the average sojourn time in an M/M/1 system under FIFO varies as $1/(1 - \rho)$ with the load ρ , we show that it could be substantially better for other queueing systems. For example, we show that for an M/G/1 system where the job sizes have the Pareto distribution with $\alpha < 2$, the average sojourn time under the policy Shortest job first (SJF) varies only as $\log(1/(1 - \rho))$ with load, which is an exponential improvement over the $1/(1 - \rho)$ dependence. We investigate this phenomenon further for other distributions and other policies.

We then study how much performance improvement can be obtained by policies which are more restricted. We first consider “blind” scheduling policies in M/G/1 type queueing systems, where a “blind” scheduling policy is one that has no knowledge of the sizes of jobs in hand and is thus restricted (for example, it cannot schedule jobs in the SJF fashion). We show that a logarithmic dependence on $1/(1 - \rho)$ is possible even for blind policies. Next, we consider even more restricted policies (that we call threshold based policies). Threshold based policies allow the jobs to be partitioned into two job classes, but cannot reorder jobs within a class and are essentially like FIFO. We show that these policies can also outperform the naive FIFO significantly. Finally we study these policies in the regime of large deviations rates asymptotics.

1 Introduction

It is common experience that as a system becomes more loaded the performance of the system degrades. The measure of such performance could be for example the response times observed by the users, the average queue lengths, the buffer overflow probability and so on.

It is natural to expect that when the processing times are known in advance of scheduling decisions, better performance bounds can be achieved. In this paper we quantify the advantage of knowing the processing times in advance in various simple queueing systems. We will be concerned with average sojourn times and large deviations rates in M/M/1 and variants of M/G/1 system.

Let us first consider average sojourn times. It is well known that for simple queueing policies such as first in first out (FIFO) or processor sharing (PS) the dependence of average sojourn time on load is directly proportional to $1/(1 - \rho)$, where ρ is the load of the system [4]. We call such dependence linear in $1/(1 - \rho)$. We first investigate whether this dependence on load is inherent for any scheduling policy. Note that such a question is naturally dependent on the choice of the job size distribution. For example, if all job sizes are

*IBM T.J. Watson Research Center, Yorktown Heights, 10598, Email: nikhil@us.ibm.com.

[†]IBM T.J. Watson Research Center, Yorktown Heights, 10598, Email: daveg@us.ibm.com.

identical (i.e. we have an $M/D/1$ system, then any queueing policy is essentially similar to FIFO and hence the linear dependence in $1/(1 - \rho)$ of average sojourn time cannot be avoided.

We thus ask the question if there are job size distribution and scheduling policies for which this dependence can be improved. Somewhat surprisingly we show that this it is possible, and moreover this improvement can be quite dramatic. Our first result is that the average sojourn time depends on load as $\log 1/(1 - \rho)$ for the scheduling policies Shortest Job First (SJF) and Shortest Remaining Processing Time (SRPT) when the job sizes have a Pareto distribution with $\alpha < 2$ (i.e. the job size distribution is $Pr[\text{Size} \geq x] = (k/x)^\alpha$). Note that the dependence on load is logarithmic on $1/(1 - \rho)$, which is substantially better than a linear dependence. Moreover it also turns out that we do not lose a lot in the constant term, the constant term is essentially the average job size (and hence the best possible). We then investigate the performance for SJF and SRPT for other distributions, specifically for Pareto distributions with $\alpha \geq 2$. In this case the improvement in average sojourn time is not as substantial as for the case $\alpha < 2$, but we show that it is still sublinear in $1/(1 - \rho)$ (i.e. it varies as $1/(1 - \rho)^\beta$, for some $\beta < 1$).

Notice that the policies SJF and SRPT use the knowledge of job sizes while scheduling. This is not ideal as far as a real system (such as an operating system) is concerned, because the information about the size of a job may not be available apriori to the scheduler. Formally, we call a scheduling policy *blind* if no information (except for distribution function) is available for the processing times of jobs in the queueing system. We next investigate the average sojourn time for blind policies. It is a folklore result that for an $M/M/1$ system any blind policy has average sojourn time of $E[S]/(1 - \rho)$, where $E[S]$ is the average job size [3]. We ask whether it is an inherent limitation of blind policies that for any $M/G/1$ system the average sojourn time will depend linearly on $1/(1 - \rho)$. Again somewhat surprisingly we answer this question in the negative. We show that the blind policy Feedback (FB) (this policy at every time works on the job that has received the least amount of processing at that time), is essentially similar to *SJF* and *SRPT*. In particular, the average sojourn time under FB depends on load as $\log 1/(1 - \rho)$ when the job sizes have a Pareto distribution with $\alpha < 2$. Moreover, for $\alpha \geq 2$, the average sojourn depends sublinearly on $1/(1 - \rho)$. These results argue that the policy *FB* is substantially better than other blind policies such as *FIFO* or *PS*, especially at higher loads and variable workloads.

Next, we consider different types of policies that we call threshold based policies. A threshold based policy with threshold level a is a policy which classifies jobs into those with processing times $\leq a$ and $> a$, but otherwise is oblivious to the actual processing times. Within each threshold class the policy as a very simple structure in that it processes the jobs in the FIFO order. Thus threshold policies can thus be viewed as partially blind policies with a simple scheduling rule. Our motivation for considering these policies comes from applications such as routers, where it is usually impossible to have a complicated scheduling rule. Surprisingly, we discover that even threshold type policy can achieve a significant improvement in performance. We show that even for an $M/M/1$ system, there is a threshold based policy that can achieve an average sojourn time of $O(E[S]/((1 - \rho) \log(e/(1 - \rho))))$, [1]. This is interesting for two reasons. First, a threshold policy breaks (though not substantially) the barrier of $E[S]/(1 - \rho)$ for blind $M/M/1$ policies. Second, this average sojourn time is infact the best achievable under any policy. In particular, it is known that SRPT (the optimum policy for minimizing the average sojourn time) has an average sojourn time of $\theta(E[S]/((1 - \rho) \log(1/(1 - \rho))))$. Thus even though the threshold policy has a much simpler structure it achieves a performance almost similar to that of SRPT.

Finally, we analyze the performance of threshold based policies with respect to buffer overflow probabilities. We consider an $M/M/1$ system and show a simple threshold rule can increase the large deviation rate of the buffer overflow probability by a constant factor. Since the overflow probability depends expo-

nentially on the large deviation rate, this implies a significant reduction in the buffer overflow probability. 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). We show that by simply partitioning the processing capacity of the router into two dedicated servers and using the threshold based rule gives significant reductions in the overflow probability. Note that if one of the servers is idle, the full system capacity may not be used. Hence, it is somewhat surprising that the overflow probability is reduced. However, the main reason for the improvement is that at high loads the servers are idle very rarely, and the variance reduction in the job size distribution at each of the dedicated servers is the dominating factor that causes the improvement.

The paper is structured as follows. In the following section we analyze performance of SJF in M/G/1 queueing system when the processing times have a Pareto distribution. We also analyze SRPT and the blind policy FB. In Section 3 we analyze the performance of static priority rules based on job lengths in the same M/G/1 queueing system. Specifically, we will consider the case of two priority classes corresponding to jobs with lengths $\leq a$ and $> a$. In the same section we consider an M/M/1 system with threshold type scheduling policy. In Section 4 we consider an M/M/1 system and analyze large deviations rates for queue lengths corresponding to the dedicated processor sharing policy.

2 SJF policy in M/G/1 queueing system

Recall (Pollaczek-Khintchine formula) that the expected sojourn time in M/G/1 queueing system under FIFO scheduling policy is $\Theta(1/(1-\rho))$. As we will show below this performance measure can be significantly improved if SJF policy is used, assuming the job processing times are available to the scheduler.

Let $E[T(x)]_{SJF}$ denote the average sojourn time of a job of size x under SJF, and let $E[T]_{SJF}$ denote the average sojourn time under SJF. As usual, we let λ denote the mean arrive rate and μ denote the average service rate. The load ρ is thus equal to λ/μ . The standard Pareto(α) distribution is given as $\bar{F}(x) = (\frac{k}{x})^\alpha$ for $x \geq k$ and 1 for $x \leq k$. The parameter α is assumed to be bigger than 1 so that the mean is finite. Thus the density function is given as $f(x) = \alpha k^\alpha x^{-\alpha-1}$ for $x \geq k$ and 0 otherwise.

The main result of this section is as follows.

Theorem 1 *For the Shortest Job First policy with Pareto(α), $\alpha > 1$ job size distributions, the average sojourn time $E[T]_{SJF}$ is*

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

The *big-O* notation in the expressions for the sojourn time above only suppresses constant factors depending on α .

As we see the performance improvement over $\Theta(1/(1-\rho))$ is quite dramatic when the processing times have a Pareto distribution.

Proof: Consider an M/G/1 queue with the service time distributed according to $f(x)$ and mean arrive rate λ . It is well known that, [6], the average sojourn time for a job of size x is $E[T(x)]_{SJF} = E[W(x)] + E[R(x)]$ where

$$E[W(x)] = \frac{\lambda \int_0^x t^2 f(t) dt}{2(1-\rho(x))^2}$$

$$E[R(x)] = \frac{x}{1 - \rho(x)}$$

Here $\rho(x) = \lambda \int_0^x tf(t)dt$ denotes the load made up by jobs of size less than or equal to x .

Without loss of generality we will assume that $\rho \geq \frac{1}{2}$. For the case when $\rho \leq 1/2$, it follows trivially that $E[T]_{SJF} = O(k)$.

We begin by bounding the contribution of the residence time.

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

Notice that the analysis above holds for any job size distribution $f(x)$. Moreover, the average residence time has only a logarithmic dependence on $(1 - \rho)$.

By a simple calculation, it follows that the average job size is $\int_0^\infty xf(x)dx = \frac{\alpha}{\alpha-1}k$. Similarly, $\rho(x) = \lambda \int_0^x tf(t)dt = \rho(1 - (k/x)^{\alpha-1})$, for $x \geq k$. Finally, we have that $\int_0^\infty x^2 f(x)dx = \frac{\alpha}{\alpha-2}k^2$.

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

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

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

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

By plugging in the value of $\rho(x)$ it follows that,

$$1 - \rho(x) = 1 - \rho + \rho \left(\frac{k}{x} \right)^{\alpha-1} \tag{5}$$

Before we begin our analysis for the waiting time, we define a threshold $x_0 = k(\frac{\rho}{1-\rho})^{\frac{1}{\alpha-1}}$. Observe that since $\rho \geq 1/2$, x_0 is greater than k . The threshold x_0 has the following nice property, which will allow us to bound the quantity $(1 - \rho(x))$ by a much simpler expression without losing more than a factor of 2.

Lemma 1 For $x \leq x_0$,

$$[\rho(k/x)^{\alpha-1}]^2 \leq (1 - \rho(x))^2 \leq 4[\rho(k/x)^{\alpha-1}]^2 \tag{6}$$

For $x \geq x_0$,

$$(1 - \rho)^2 \leq (1 - \rho(x))^2 \leq 4(1 - \rho)^2 \tag{7}$$

Proof: By the choice of x_0 , it is easy to see that $\rho(k/x)^{\alpha-1} \geq 1 - \rho$ for $x \leq x_0$. By Equation 5, it follows that $1 - \rho(x) \leq 2\rho(k/x)^{\alpha-1}$, which implies (6). Similarly, (7) follows as $\rho(k/x)^{\alpha-1} \leq 1 - \rho$ for $x \geq x_0$. ■

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 Equation 3 and Lemma 1,

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

where $g(x) = 2[\rho(k/x)^{\alpha-1}]^2$.

We now bound the first term,

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

Finally, we bound the second term,

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

The second last step above follows from our choice of $x_0 = k(\frac{\rho}{1-\rho})^{\frac{1}{\alpha-1}}$. The last step follows since $\rho = \lambda/\mu$ and $\mu = \frac{\alpha-1}{\alpha k}$.

2. When $\alpha = 2$: By Equation 3 we have,

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

The last step follows by substituting the values of λ and x_0 .

As $\alpha = 2$ and $\rho \geq 1/2$, the average sojourn time is easily seen to be $O(k \ln^2(\frac{\rho}{1-\rho}))$.

3. When $\alpha > 2$: By Equation 3 and Lemma 1,

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

where $g(x) = 2[\rho(k/x)^{\alpha-1}]^2$.

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 $O(k \ln(\frac{1}{1-\rho}))$, the proof of the theorem follows by considering the obtained expression for the average waiting time in all the three cases. ■

2.1 The results for SRPT

Since SRPT is the optimum policy for minimizing the average sojourn time under any G/G/1 queueing system [5], it follows that the average sojourn times under SRPT are no worse than those under SJF. Thus we trivially get that,

Corollary 2 *For the SRPT policy with Pareto(α), $\alpha > 1$ job size distributions the average sojourn time $E[T]_{SRPT}$ is*

$$E[T]_{SRPT} = \begin{cases} O(k \ln(\frac{1}{1-\rho})) & \text{if } \alpha < 2 \\ O(k \ln^2(\frac{1}{1-\rho})) & \text{if } \alpha = 2 \\ O(k(1-\rho)^{-\frac{(\alpha-2)}{(\alpha-1)}}) & \text{if } \alpha > 2 \end{cases} \quad (8)$$

2.2 The blind policy FB

Next, we consider the blind policy FeedBack (FB) which is a close approximation to that used by many practical systems such as the Unix operating system. Even though FB is blind it turns out that for the

Pareto distribution, the average sojourn time under FB is exactly similar (up to a constant factor dependent on α) to that SJF . This shows that FB is a significantly better policy as compared to other blind policies such as Processor Sharing or FCFS where the dependence on load is $1/(1 - \rho)$.

Corollary 3 *For the FB policy with Pareto(α), $\alpha > 1$ job size distributions the average sojourn time $E[T]_{FB}$ is bounded by*

$$E[T]_{FB} = \begin{cases} O(k \ln \left(\frac{1}{1-\rho} \right)) & \text{if } \alpha < 2 \\ O(k \ln^2 \left(\frac{1}{1-\rho} \right)) & \text{if } \alpha = 2 \\ O(k(1 - \rho)^{-\frac{(\alpha-2)}{(\alpha-1)}}) & \text{if } \alpha > 2 \end{cases} \quad (9)$$

Proof: Let $\rho_1(x) = \lambda \int_0^x t f(t) dt + \lambda x(1 - F(x))$. It is well known, [4], that the average sojourn time of a job of size x under FB is given by

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

Note that this expression is very similar to that of SJF except that in the denominators we have $1 - \rho_1(x)$ instead of $1 - \rho(x)$ and in the numerator of FB there is the additional $x^2(1 - F(x))$ term.

We first show that $1 - \rho_1(x)$ is within a constant factor of $1 - \rho(x)$

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

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

which implies that the contribution due to the denominators of the waiting time under SJF and FB do not differ by a factor more than α .

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 SJF for this case, we only use the fact that for SJF, $\int_0^x t^2 f(t) dt = O(k^\alpha x^{2-\alpha})$ and hence the expression for FB changes by a constant factor only.

Similarly for $\alpha = 2$, $x^2(1 - F(x)) = k^2$. Again under SJF for this case, the corresponding term $\int_0^x t^2 f(t) dt$ is $O(k^2 \log(x/k))$ which is at least k^2 for $x = \Omega(k)$. Thus the expression for the waiting time for FB changes by only a constant factor as compared with that for SJF.

Finally for $\alpha \geq 2$, $x^2(1 - F(x)) = k^\alpha x^{-\alpha+2}$. In bounding the waiting time under SJF 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 SJF also holds for FB (upto constant factors).

3 Threshold Based Policies

We consider the classic $M/G/1$ preemptive priority queueing system. There are n job classes. Class 1 jobs receive the most favorable treatment; class k jobs receive the least favorable treatment. Jobs within each class are served according to the FCFS policy. At any time the job from the least numbered class is executed. In particular, if a job of class j is receiving service when a job of class $i < j$ arrives then the arriving job preempts the class j job.

Let λ_i denote the arrival rate of class i jobs, and let f_i denote the pdf of the job size distribution of class i jobs. Let $E[S_i]$ and $E[S_i^2]$ denote the first and the second moment respectively of the job size distribution of class i jobs. Finally, let $\rho_i = \lambda E[S_i]$ denote the load of class i jobs, and a_i denotes $\sum_{j=1}^i \rho_j$.

It is a classic result that the mean sojourn time of jobs of class i is given by

$$E[T_i] = \frac{E[S_i]}{1 - a_{i-1}} + \frac{\sum_{j=1}^i \lambda_j E[S_j^2]}{2(1 - a_{i-1})(1 - a_i)} \quad (10)$$

Let $p_i = \lambda_i/\lambda$ denote the fraction of jobs that belong to class i . The average sojourn time is clearly

$$E[T] = \sum_{i=1}^k p_i E[T_i] \quad (11)$$

For our purposes we will consider the following special case of the setting above. We will only have two job classes. We will assume that the jobs arrive from a distribution with pdf $f(x)$ and there is a threshold x_0 that determines which class the job falls in. We will say that a job belongs to class 1 if its size is below x_0 , and belongs to class 2 otherwise. Thus we consider a threshold type policy with threshold level x_0 . Notice that in our setting, $p_1 = F(x_0)$ and $p_2 = \overline{F}(x_0)$ where $F(x) = \int_0^x f(t)dt$. In our setting, the expression for the average sojourn time given by the Equations 10 and 11 simplifies to the following:

Lemma 2

$$\begin{aligned} E[T] &= E[S] + \frac{\overline{F}(x_0)\rho_1 E[S_2]}{1 - \rho_1} + \frac{p_1 \lambda_1 E[S_1^2]}{2(1 - \rho_1)} \\ &\quad + \frac{\lambda \overline{F}(x_0) E[S^2]}{2(1 - \rho_1)(1 - \rho)} \end{aligned} \quad (12)$$

Proof: By Equation 10,

$$\begin{aligned} E[T_1] &= E[S_1] + \frac{\lambda_1 E[S_1^2]}{2(1 - \rho_1)} \\ E[T_2] &= \frac{E[S_2]}{1 - \rho_1} + \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2(1 - \rho_1)(1 - \rho)} \\ &= \frac{E[S_2]}{1 - \rho_1} + \frac{\lambda E[S^2]}{2(1 - \rho_1)(1 - \rho)} \end{aligned}$$

The desired result follows by applying Equation 11 and observing that $E[S] = p_1 E[S_1] + p_2 E[S_2]$. ■

Finally, observing that $\overline{F}(x_0)E[S_2] \leq E[S]$, $E[T]$ in Equation 12 can be bounded above by

$$E[T] \leq \frac{E[S]}{1 - \rho_1} + \frac{p_1 \lambda_1 E[S_1^2]}{2(1 - \rho_1)} + \frac{\lambda \overline{F}(x_0) E[S^2]}{2(1 - \rho_1)(1 - \rho)} \quad (13)$$

3.1 Pareto Distributions

We construct a threshold policy for the Pareto distribution which improves the dependence on $1/(1 - \rho)$. First, we note that the second moment of processing times is infinite for $\alpha \leq 2$ (even if we only restrict the distribution to jobs of size $> x_0$ for any finite x_0). Since, the average sojourn time for the second class (see Lemma 2) depends on the second moment of the processing times, it follows that the average sojourn time under a threshold policy is finite only when $\alpha > 2$. Thus we only restrict our attention to the case when $\alpha > 2$.

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

$$E[T] \leq \frac{4k(\alpha - 1)}{(\alpha - 2)(1 - \rho)^{\frac{\alpha-1}{\alpha}}}$$

As we see even a simple threshold type policy which only distinguishes jobs with processing times $\leq x_0$ from $> x_0$ achieves a significant performance improvement over FIFO.

Proof: To avoid trivialities, let us assume that $\rho \geq 1/2$. Choose x_0 such that $\bar{F}(x_0) = 1 - \rho$. Thus, $(k/x_0)^\alpha = (1 - \rho)$ and hence that $x_0 = k(1 - \rho)^{-1/\alpha}$.

As $E[S_1^2] \leq E[S^2]$, $p_1 \lambda_1 \leq \lambda$ and $\bar{F}(x_0) = 1 - \rho$, Equation 13 reduces to,

$$E[T] \leq \frac{E[S]}{1 - \rho_1} + \frac{\lambda E[S^2]}{(1 - \rho_1)}$$

Now, $E[S] = \alpha k / (\alpha - 1)$ and $E[S^2] = \alpha k / (\alpha - 2)$, thus using $\rho < 1$ and $\alpha / (\alpha - 1) < (\alpha - 1) / (\alpha - 2)$, we get

$$\begin{aligned} E[T] &\leq \frac{\alpha k}{(\alpha - 1)(1 - \rho_1)} + \frac{\rho(\alpha - 1)k}{(\alpha - 2)(1 - \rho_1)} \\ &\leq \frac{2(\alpha - 1)k}{(\alpha - 2)(1 - \rho_1)} \end{aligned} \quad (14)$$

Using Equation 5 it follows that $1 - \rho_1 = 1 - \rho + \rho(k/x_0)^{\alpha-1}$. Since, $(k/x_0) = (1 - \rho)^{1/\alpha}$, it follows that

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

Ignoring the term $1 - \rho$, and using that fact that $\rho \geq 1/2$, we get that

$$1 - \rho_1 \geq \frac{1}{2} \cdot (1 - \rho)^{\frac{\alpha-1}{\alpha}} \quad (15)$$

Together with Equation 14 this implies the desired result.

3.2 Exponential distributed job sizes

We now consider an $M/M/1$ system, and show that there is a simple threshold policy such that average sojourn time is similar (up to constant factors) to that of SRPT and hence achievable by any policy [1, 5].

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

$$E[T] \leq \frac{6}{\mu(1 - \rho) \log \left(\frac{e^2}{1 - \rho} \right)}$$

Proof: If $\rho \leq 1/2$, choosing $x_0 = \infty$, it is easy to see that $E[T] \leq 2/\mu$. So, we assume that $\rho \geq 1/2$.

Choose x_0 such that $\bar{F}(x_0) = 1 - \rho$. Thus, $e^{-\mu x_0} = (1 - \rho)$ or $x_0 = \frac{1}{\mu} \log \frac{1}{1-\rho}$. Using Equation 13 we obtain

$$E[T] \leq \frac{E[S]}{1 - \rho_1} + \frac{p_1 \lambda_1 E[S_1^2]}{2(1 - \rho_1)} + \frac{\lambda E[S^2]}{2(1 - \rho_1)}$$

As $E[S_1^2] \leq E[S^2]$ and $p_1 \lambda_1 \leq \lambda$ we get,

$$E[T] \leq \frac{E[S]}{1 - \rho_1} + \frac{\lambda E[S^2]}{1 - \rho_1}$$

Since $\lambda E[S] \leq 1$ and $E[S^2] = 2E[S]^2$ for the exponential distribution, we get that,

$$E[T] \leq 3 \frac{E[S]}{1 - \rho_1} \quad (16)$$

By a direct calculation it can be seen that

$$\rho_1 = \rho \mu \int_0^{x_0} t f(t) dt = \rho^2 - \rho(1 - \rho) \log \left(\frac{1}{1 - \rho} \right)$$

Thus,

$$\begin{aligned} 1 - \rho_1 &= 1 - \rho^2 + \rho(1 - \rho) \log \left(\frac{1}{1 - \rho} \right) \\ &\geq \frac{1}{2} \cdot (1 - \rho) \log \left(\frac{e^2}{1 - \rho} \right) \end{aligned}$$

Thus by equation 16,

$$E[T] \leq \frac{6}{\mu(1 - \rho) \log \left(\frac{e^2}{1 - \rho} \right)}$$

4 Large deviation of queue lengths

We consider an M/M/1 queue with the arrival rate $\lambda < 1$ and the processing rate $\mu = 1$. Thus $\rho = \lambda$. Recall that in steady state $\Pr(Q \geq m) = \rho^m = e^{\log(\rho)m}$. In other words the large deviation rate for the queue length is $\log(\rho)$, which is approximately $1 - \rho$ when ρ is closed to the unity. We consider now a certain threshold based dedicated scheduling policy which works as follows. The jobs are partitioned into two classes: those with processing time $\leq a$ and those with $> a$, for some threshold a to be defined. For some parameter $\psi \in (0, 1)$ we allocated exactly ψ portion of the server to serving class $\leq a$ jobs and remaining portion to class $> a$ jobs. The value of ψ will be specified later. The partition of the server is dedicated in the sense that even if there are no class $\leq a$ jobs in the system, the server gives only $1 - \psi$ portion of its capacity to class $> a$ jobs and vice vs.

We compute the large deviations rates θ_1, θ_2 corresponding to steady state queue lengths of classes $\leq a$ and $> a$ and we will show that for certain parameters a, ψ , $\max(\theta_1, \theta_2) < \log(\rho)$. Since the overall large deviation for steady state queue length is given as $\min(\theta_1, \theta_2)$ we obtain that our policy achieves an improvement over FIFO policy in M/M/1 in terms of the large deviations rates performance.

Recall (Theorem 4.3 [2]) that in G/G/1 policy the large deviations rate

$$\lim_{x \rightarrow \infty} \frac{\log(\Pr(Q(t) > x))}{x}$$

for the steady state queue length is given as $\Lambda_A(\theta)$ where θ is the smallest negative root of the equation

$$\Lambda_A(\theta) + \Lambda_B(-\theta) = 0 \quad (17)$$

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

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

implying $\theta = \lambda - \mu$ and the large deviation rate becomes $\log E[e^{(\lambda-\mu)A}] = \log(\lambda/\mu) = \log(\rho)$ as expected. We use formula (17) to compute the large deviations rates for individual queues when we use our scheduling policy.

4.1 The system with size $> a$

Let us consider first the system with $> a$. In this case, the distribution of the sizes of the jobs that arrive in this system 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 λe^{-a} . We will set $\psi = e^{-a}(a + 1)$.

However, note that since we are only giving a portion ψ to this server, this will stretch the processing times in this system by a factor of $1/\psi$. In particular this implies that a job which has a size x in the original system will have size x/ψ in this system. To remove this effect of scaling we do the following trick: Observe that if we scale the speed and the arrival of any queueing system simultaneously by a factor c , then the distribution of lengths 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 processing 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 now compute the relevant quantities. Clearly, $\Lambda_B(-\theta) = \log \int_a^\infty e^a e^{-x} e^{-\theta x} dx$ which is $\log(e^{-\theta a}/(\theta + 1))$.

Since the interarrival distribution is still Poisson with rate λ' 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

$$\frac{\lambda'}{\lambda' - \theta} \cdot e^{-\theta a}/(\theta + 1) = 1 \quad (18)$$

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

To see this, we rewrite the above equation as

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

Next we observe that since $\theta^* < 0$, $1 - e^{-\theta^* a} \leq a\theta^* - a^2\theta^{*2}/2$. Thus,

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

Since $\theta^* < 0$ this implies that

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

Hence

$$\theta^* \geq \frac{(\lambda - 1)(a + 1)}{(a + 1) + \lambda a^2/2}$$

Since $\theta^* < 0$, the above equation implies that as $\lambda \rightarrow 1$, then $\theta^* \rightarrow 0$. Our goal now is to obtain

$$\lim_{\lambda \rightarrow 1} \frac{\theta^*}{\lambda - 1}$$

We know that,

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

and hence,

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

Applying the L'Hopital rule twice,

$$\lim_{\theta^* \rightarrow 0} \frac{(a+1)\theta^{*2} + (a+1)\theta^* - 1 + e^{-\theta^*a} - \theta^*}{\theta^*(1 - e^{-\theta^*a} + \theta^*)} = \frac{2(a+1) + a^2}{2(a+1)}$$

Now, as $\lambda \rightarrow 1$, or equivalently as $\theta^* \rightarrow 0$, the large deviation rate $\log \frac{\lambda'}{\lambda' - \theta^*}$ approaches $\theta^*/\lambda' = (a+1)\theta^*$. This gives that the large deviation rate is $(\lambda - 1)2(a+1)^2/(a^2 + 2a + 2)$. For $a = 1$ this is $1.6(\lambda - 1)$.

4.2 When size $\leq a$

In this system the arrival rate of the jobs is $(1 - e^{-a})\lambda$, and the speed of the server is $1 - \psi = 1 - (a+1)e^{-a}$. We again apply the scaling idea from above. We scale the job sizes by the factor $1/(1 - \psi)$, so that the jobs sizes at our system still have the distribution $(1 - e^{-a})^{-1}e^{-x}$ for $x \leq a$ and 0 otherwise. Let m denote $(1 - (a+1)e^{-a})/(1 - e^{-a})$. Note that after the scaling the arrival rate $\lambda' = \lambda/m$.

We again compute the relevant quantities. By a simple calculation it follows that

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

So, we have to solve for the smallest root of

$$\frac{\lambda/m}{\lambda/m - \theta} \cdot \frac{1}{(\theta + 1)(1 - e^{-a})} (1 - e^{-a(1+\theta)}) = 1$$

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

$$\lambda(1 - e^{-a-a\theta}) = (\lambda - m\theta)(\theta + 1)(1 - e^{-a})$$

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

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

Since $1 - e^{-a\theta} \leq a\theta$ and $\theta < 0$, then

$$\lambda e^{-a}a \geq \lambda(1 - e^{-a}) - (\theta + 1)(1 - (a+1)e^{-a}),$$

or $\theta \geq \lambda - 1$, implying $\theta \rightarrow 0$ as $\lambda \rightarrow 1$ from below.

Again as above one can show that as $\lambda \rightarrow 1$, then $\theta^* \rightarrow 0$. So,

We use the approximation

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

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

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

If $\lambda \rightarrow 1$, the large deviation rate $\log \frac{\lambda}{\lambda - m\theta^*}$ evaluates to $m\theta^*/\lambda$. Plugging the value of θ^* this evaluates to

$$\frac{1 - (a+1)e^{-a}}{1 - e^{-a}} \cdot \frac{(\lambda - 1)(1 - (a+1)e^{-a})}{1 - (a+1)e^{-a} - \lambda e^{-a} \frac{a^2}{2}}$$

If we choose $\alpha = 1$, and choose $\lambda \rightarrow 1$, this evaluates to $1.3755(\lambda - 1)$.

Thus in both cases we obtain that large deviations rates improve by a factor at least $\min(1.6, 1.3755) = 1.3755$. We summarize our results.

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 θ_1, θ_2 for the two queue lengths satisfy*

$$\lim_{\rho \rightarrow 1} \frac{\min(|\theta_1|, |\theta_2|)}{|\log \rho|} \geq 1.3755.$$

5 Conclusions

We believe that the results in this paper are a first step towards understanding how the performance of queueing policies is affected based on knowledge of job sizes. We show that some simple policies result in qualitative and sometimes exponential improvements for sojourn times when compared to naive FIFO or processor sharing policies. It would be very interesting to obtain a more complete understanding of the tradeoffs involved and how other measures such as variance and higher moments of sojourn times, the sojourn times of the few largest jobs etc are dependent on the policies and the job size distributions.

References

- [1] N. Bansal, *M/M/1/SRPT has average sojourn time $\theta(1/\log(1/(1-\rho)) \cdot 1/(\mu(1-\rho)))$* , To appear in Operations Research Letters.
- [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] L. Kleinrock, *Queueing systems*, John Wiley and Sons, 1975.
- [5] L.E. Schrage, *A proof of the optimality of the shortest processing remaining time discipline*, Operations Research **16** (1968), 678–690.
- [6] Ronald W. Wolff, *Stochastic modeling and the theory of queues*, Prentice Hall, 1989.