Mixed scheduling disciplines for network flows

Hanhua Feng Vishal Misra
Columbia University, New York, NY 10027

ABSTRACT

We introduce a novel method to prove that the FBPS discipline has optimal mean sojourn time and mean slowdown ratio for DHR service time distributions in an $M/G/1$ queue. We then discuss the problems related to FBPS, and propose a new scheduling discipline to overcome these problems.

1. INTRODUCTION

Recently researchers have shown a great interest in scheduling disciplines such as shortest remaining processing time (SRPT) for network flow scheduling [1], since it provides the optimal mean sojourn time (response time). However, operating systems, software applications and devices like routers usually do not know the remaining job size, which makes the SRPT scheduling discipline difficult to implement. Various alternatives have been sought. Among them, the foreground-background processor-sharing (FBPS) discipline[1] can be considered as a temporal reverse of SRPT; it prefers the job that has the least attained service time. Researchers have found for heavy-tailed job size distributions, FBPS performs comparably to SRPT.

There is a problem in both SRPT and FBPS. If some small jobs and a large job are in the system, the large job won't wait for a long time because of the small jobs. However if a moderately large job and a large job are in the system, the service for the large job has to be interrupted for a long time. Generally speaking, both these disciplines do not give smooth service particularly expected for network clients, and user impatience or some protocol defined timeout could result in wasted resources.

In a recent abstract, Guo and Matta[2] studied by simulation a special multilevel processor sharing (ML-PS) scheduling discipline, in which those jobs that have so far received less than a certain amount of service time (the threshold) get more processor time than each of the other jobs. One extreme is ML-PRIO, letting the former jobs occupy the whole processor if there are such jobs. They showed that in this case, if the service time distribution is bounded-Pareto, the response time of short jobs will get significantly improved while the response time of large jobs increased by only a little. ML-PRIO can be considered as a degraded FBPS scheduler that can only identify two discretized service times: large and small. If the threshold is not very large, it will not have the problem in SRPT and FBPS mentioned above.

In this abstract, we first prove that, if the service time distribution has a decreasing hazard rate (DHR), the FBPS is optimal for both mean sojourn time and mean slowdown ratio (which is defined as the ratio of sojourn times to service time), among all scheduling disciplines that do not make use of the remaining service time. Then, we present another mixed scheduling discipline called FLIPS, that also does not suffer from the previous problem in SRPT and FBPS. This discipline, as well as the ML-PRIO, belongs to a family of multilevel scheduling disciplines studied by Kleinrock[3] in the early 70's.

[1] It has other names in the literatures: feedback(FB), least attained service time(LAST), or least attained service(LAS).

2. THE OPTIMALITY OF FBPS

Let $F(x)$ be the service time distribution of a job and $f(x)$ be its density function. A distribution is said to have a decreasing hazard rate (DHR) if $\frac{f(x)}{1-F(x)}$ is monotonically decreasing, and to have a increasing hazard rate (IHR) if $\frac{f(x)}{1-F(x)}$ is monotonically increasing. Then we have the following theorems.

Theorem 2.1. *For a DHR service time distribution, the expected sojourn time of an $M/G/1$ queue under the FBPS discipline is optimal among all scheduling disciplines in which the remaining service time is unknown.*

Theorem 2.2. *For a DHR service time distribution, the expected slowdown ratio of an $M/G/1$ queue under the FBPS discipline is optimal among all scheduling disciplines in which the remaining service time is unknown.*

In a recent report, for DHR distributions, it was showed that the mean service time of FBPS is better than processor-sharing(PS) [7]. We will now present a proof of these two theorems, which are more general. Righter et al. proved Theorem 2.1 with a discrete model[5, 6].

Lemma 2.3. *Let $[a,b]$ be a closed interval on $\mathbf{R}^+$. Let $d(x)$ be a non-negative, monotonically increasing integrable function bounded in $[a,b]$, $g(x)$ be an integrable bounded function such that $g(a) = g(b) = 0$, $h(x)$ is a function either non-positive, non-negative, or zero. For any $\xi \in [a,b]$, suppose $g(\xi^-) \triangleq \lim_{x \uparrow \xi} g(x)$ and $g(\xi^+) \triangleq \lim_{x \downarrow \xi} g(x)$ are well-defined, (so for $h(x)$), and*

$$h(\xi^-)g(\xi^-) + \int_a^\xi d(x)g(x)dx \geq 0, \tag{1}$$

then

$$\int_a^b g(x)dx \geq 0,$$

PROOF OF LEMMA 2.3. For any $\epsilon > 0$ we have

$$d((\xi+\epsilon)^-)g((\xi+\epsilon)^-) + \int_a^{\xi+\epsilon} d(x)g(x)dx \geq 0.$$

Letting $\epsilon \to 0$,

$$h(\xi^+)g(\xi^+) + \int_a^{\xi} d(x)g(x)dx \geq 0,$$

Without lost of generality, we can let $g(x)$ be right-continuous, i.e., $g(x) = g(x^+)$.

Since $g(x)$ is integrable, it must also be measurable. Therefore sets $\{x : g(x) > 0\}$ and $\{x : g(x) < 0\}$ should be Borel sets, which contains countable open or closed intervals.

We only prove for the case that the number of these intervals are finite. It is easy to extend it to the case where the intervals are countably infinite, by taking approximations of $g(x)$.

For the finite case of n intervals, this theorem is proved by induction. Let (a_i, b_i) be maximum intervals such that either $g(x) > 0$ or $g(x) < 0$ for all $x \in (a_i, b_i)$, where $i = 0, 1, 2, \ldots$.

Clearly, for any $\xi \in \{a_i\}_{i=0}^{n-1} \cup \{b_i\}_{i=0}^{n-1}$, since (a_i, b_i) is maximum, we have $g(x^-)g(x) \leq 0$. Therefore for these ξ's,

$$\int_a^{\xi} d(x)g(x)dx \geq 0. \tag{2}$$

In the first interval (a_0, b_0), $g(x)$ must be positive, otherwise (2) will not be satisfied at $\xi = b_0$. Thus $\int_0^{b_0} g(x) \geq 0$. Since $d(x)$ is increasing, we also have

$$0 \leq \int_a^{b_0} d(x)g(x)dx \leq d(b_0)\int_a^{b_0} g(x)dx.$$

Suppose we have

$$\int_a^{b_i} d(x)g(x)dx \leq d(b_i)\int_a^{b_i} g(x)dx \tag{3}$$

for $i = 0, 1, \ldots, k-1$, $k \geq 0$.

Then for interval (a_n, b_n), there are two cases:

(1) In the case that $g(x) > 0$ for all $x \in (a_n, b_n)$, we have

$$\begin{aligned}
&\int_a^{b_k} d(x)g(x)dx \\
&\quad\leq\quad d(b_{k-1})\int_a^{b_{k-1}} g(x)dx + d(b_k)\int_{a_k}^{b_k} g(x)dx \\
&\quad\leq\quad d(b_k)\int_a^{b_k} g(x)dx.
\end{aligned}$$

(2) In the case that $g(x) < 0$ for all $x \in (a_n, b_n)$, we have

$$\begin{aligned}
&\int_a^{b_k} d(x)g(x)dx \\
&\quad=\quad \int_a^{b_{k-1}} d(x)g(x)dx - \int_{a_k}^{b_k} d(x)|g(x)|dx \\
&\quad\leq\quad d(b_{k-1})\int_a^{b_{k-1}} g(x)dx - d(a_k)\int_{a_k}^{b_k} |g(x)|dx \\
&\quad\leq\quad d(b_{k-1})\int_a^{b_k} g(x)dx \\
&\quad\leq\quad d(b_k)\int_a^{b_k} g(x)dx
\end{aligned}$$

Hence for either case, (3) is true for $i = k$. Let $k = n-1$, we have

$$0 \leq \int_a^{b_{n-1}} d(x)g(x)dx \leq d(b_{n-1})\int_a^{b_{n-1}} g(x)dx.$$

Since either $b = b_{n-1}$ or $g(x) \equiv 0$ for $x \in (b_{n-1}, b]$, we have

$$\int_a^{b} g(x)dx \geq 0.$$

□

To prove 2.1, we first introduce some notations. We denote the sojourn time of a job of size x by $T(x)$, and by $ET(x)$ its expectation. The sojourn time of a job with any size is denoted by random variable T.

We denote by random variable $N(x)$ the number of jobs which have so far received less than x seconds of services, and by random variable $U(x)$ their remaining service time before x seconds, and by random variable $R(t, x)$ the remaining service time before x seconds for a single job with attained service time of t, where $t \leq x$.

Lemma 2.4. *Among all disciplines, the FBPS discipline minimizes* $E[U(x)]$.

PROOF OF LEMMA 2.4. At any time if there is no arrival, $U(x)$ decreases by a fixed rate, which is independent of the scheduling descipline, as long as the processor is serving a job that has an attained service time smaller than x. For any $x \geq 0$, FBPS discipline devotes all the processor time to serve jobs with attained service smaller than or equal to x, if these jobs are present. Thus the average of total remaining service time of these jobs, $E[U(x)]$, should be minimum. □

PROOF OF THEOREM 2.1. According to [4], we have

$$\frac{d}{dx}EN(x) = \lambda[1 - F(x)]\frac{d}{dx}ET(x). \tag{4}$$

By definition of $U(x)$ and $R(x, t)$ we have

$$E[R(x,t)] = (x-t)\frac{1-F(x^-)}{1-F(t)} + \int_t^{x^-} (\tau - t)\frac{dF(\tau)}{1-F(t)}$$

and

$$\begin{aligned}
U(x) &= \int_{0^-}^{x^-} R(x,t)dN(t) \\
EU(x) &= \int_{0^-}^{x^-} E[R(x,t)dN(t)] \\
&= \int_{0^-}^{x^-} E[\frac{dN(t)}{dt}]E[R(x,t)]dt \qquad (5) \\
&= \int_{0^-}^{x^-} \frac{dEN(t)}{dt}\left[(x-t)\frac{1-F(x^-)}{1-F(t)}\right. \\
&\qquad \left. + \int_t^{x^-} \frac{\tau - t}{1-F(t)}dF(\tau)\right]dt \\
&= \lambda\int_{0^-}^{x^-} [1-F(t)]\frac{dET(t)}{dt}\left[(x-t)\frac{1-F(x^-)}{1-F(t)}\right. \\
&\qquad \left. + \int_t^{x^-} \frac{\tau - t}{1-F(t)}dF(\tau)\right]dt \\
&= \lambda\int_{0^-}^{x^-} \left[(x-t)[1-F(x^-)]\right. \\
&\qquad \left. + \int_t^{x^-} (\tau - t)dF(\tau)\right] dET(t) \qquad (6)
\end{aligned}$$

Since

$$\begin{aligned}
&\int_t^{x^-} (\tau - t) dF(\tau) \\
&= -\int_t^{x^-} (\tau - t) d[1 - F(\tau)] \\
&= -[(\tau - t)[1 - F(\tau)]]_{\tau=t}^{x^-} + \int_t^{x^-} [1 - F(\tau)] d\tau \\
&= -(x - t)[1 - F(x^-)] + \int_t^{x^-} [1 - F(\tau)] d\tau
\end{aligned}$$

we have

$$\begin{aligned}
EU(x) &= \lambda \int_{0-}^{x^-} \left[\int_t^{x^-} [1 - F(\tau)] d\tau \right] dET(t) \\
&= \lambda \left[ET(t) \int_t^{x^-} [1 - F(\tau)] d\tau \right]_{t=0-}^{x^-} \\
&\quad -\lambda \int_{0-}^{x^-} ET(t) d \left\{ \int_t^{x^-} [1 - F(\tau)] d\tau \right\} \\
&= \lambda \int_{0-}^{x^-} ET(t)[1 - F(t)] dt \\
&= \lambda \int_{0-}^{x^-} [ET(t) f(t)] \frac{1 - F(t)}{f(t)} dt,
\end{aligned}$$

By Lemma 2.4 we have $U(x) - U(x)_{\text{FBPS}} \geq 0$, therefore,

$$\begin{aligned}
&0 \leq U(x) - U(x)_{\text{FBPS}} \\
&= \lambda \int_{0-}^{x^-} [ET(t) - ET(t)_{\text{FBPS}}] f(x) \frac{1 - F(t)}{f(t)} dt \quad (7)
\end{aligned}$$

Define

$$\begin{aligned}
d(x) &\triangleq \lambda \frac{1 - F(x)}{f(x)}, \\
g(x) &\triangleq [ET(x) - ET(x)_{\text{FBPS}}] f(x), \text{ and} \\
h(x) &\triangleq 0,
\end{aligned}$$

by Lemma 2.3, we have

$$\begin{aligned}
0 &\leq \int_{0-}^{\infty} g(x) dx \\
&= \int_{0-}^{\infty} [ET(x) - ET(x)_{\text{FBPS}}] f(x) dx \\
&= E[T] - E[T]_{\text{FBPS}},
\end{aligned}$$

which means $E[T] \geq E[T]_{\text{FBPS}}$. □

Remark 2.5. *The condition that a scheduling discipline should not use any information about remaining job size is necessary, because in (5) we assume $dN(t)/dt$ and $R(t)$ are independent. Therefore, the SRPT scheduling discipline, using the job size information, will not satisfy this assumption. Similarly, the SRPT discipline does not satisfy Kleinrock's conservative law [4] that states $\int_{0-}^{\infty} T(x)[1 - F(x)]dx$ is constant.*

Remark 2.6. *The bounded-Pareto distribution is not DHR because of the effects of truncation on both sides, therefore the FBPS is not strictly optimal for this distribution. However, the property of DHR can be approximately used in so-called heavy-tailed property distributions, so one can say that the FBPS has almost the best mean sojourn time.*

PROOF OF THEOREM 2.2. From (7), we can also get

$$\begin{aligned}
&0 \leq U(x) - U(x)_{\text{FBPS}} \\
&= \lambda \int_{0-}^{x^-} \left[\frac{ET(t)}{t} - \frac{ET(t)_{\text{FBPS}}}{t} \right] f(t) \frac{t[1 - F(t)]}{f(t)} dt
\end{aligned}$$

Let

$$\begin{aligned}
g(x) &\triangleq \left[\frac{ET(x)}{x} - \frac{ET(x)_{\text{FBPS}}}{x} \right] f(x), \\
d(x) &\triangleq \lambda \frac{x[1 - F(x)]}{f(x)}, \text{ and} \\
h(x) &\triangleq 0,
\end{aligned}$$

where $d(t)$ is an increasing function because x is an increasing function and $\frac{1-F(x)}{f(x)}$ is also increasing for DHR distributions. By applying Lemma 2.3 again we have

$$\int_{0-}^{\infty} \frac{ET(x)}{x} f(x) dx \geq \int_{0-}^{\infty} \frac{ET(x)_{\text{FBPS}}}{x} f(x) dx.$$

Hence we have

$$\begin{aligned}
E\left[\frac{T(x)}{x}\right] &\triangleq \int_{0-}^{\infty} \frac{T(x)}{x} dF(x) \\
&= E\left[\frac{ET(x)}{x}\right] \\
&\geq E\left[\frac{ET(x)_{\text{FBPS}}}{x}\right] \\
&= E\left[\frac{T(x)_{\text{FBPS}}}{x}\right].
\end{aligned}$$

□

Remark 2.7. *For job size distributions such that $\frac{f(x)}{x[1-F(x)]}$ is decreasing, FBPS will be optimal for mean slowdown ratio. This relaxed condition has many other distributions included, besides DHR distributions. One example is the 2nd-order Erlang distribution*

$$f_{E_2}(x) = \lambda^2 x e^{-\lambda x}.$$

Its distribution function is

$$F_{E_2}(x) = 1 - (1 + \lambda x) e^{-\lambda x},$$

therefore

$$\frac{f_{E_2}(x)}{x[1 - F_{E_2}(x)]} = \frac{\lambda^2 x e^{-\lambda x}}{x(1 + \lambda x) e^{-\lambda x}} = \frac{\lambda^2}{1 + \lambda x}$$

is monotonically decreasing, so FBPS has the optimal mean slowdown ratio for the 2nd-order Erlang distribution.

3. THE FLIPS SCHEDULER

We can observe two problems in the FBPS scheduling discipline. The first problem is that FBPS must obey the Kleinrock's conservative law, therefore when it gives quick responses for small jobs, it might starve large jobs, although the situation is not very bad for heavy-tailed service time distributions. Figure 1 shows that the slowdown ratio of large jobs under FBPS when the load is high. The second problem is that the FBPS does not provide smooth service, as mentioned in the introduction. To avoid these two problems, we present the FLIPS scheduler.

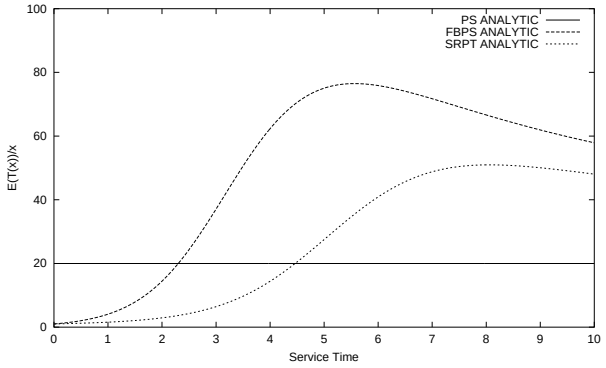

Figure 1: Analytic results of mean slowdown ratios of an $M/M/1$ queue ($\mu = 1.0$) at load $= 0.95$, as functions of service time, under various scheduling disciplines.

The FLIPS scheduler basically involves two levels with a threshold of service times. Unlike ML-PS being both processor sharing, this scheduling discipline is a mixture of FBPS and processor sharing (PS): given a threshold, namely p, if some jobs in the system have not yet received an amount service time that is more than p, the job that have so far received the least service time takes the whole processor. If all jobs in the system have received more than p amount of service time, each of these jobs would get an equal share of the processor. Intuitively we see that, if the threshold is very large, the FLIPS scheduler behaves just like an FBPS scheduler whereas, if the threshold is very small, it would be basically processor sharing. The sojourn time for jobs whose size is smaller than p are the same as under FBPS discipline, since they both ignore jobs having more attained service time – how they get scheduled is irrelevant. For jobs whose size is greater than p, the sojourn times will go beyond those under the processor sharing discipline, in which FLIPS and ML-PS behave the same, because jobs that have received no more than p time would always get scheduled first: how they get scheduled is irrelevant. Figure 2 shows the slowdown ratio as a function of service time, for various scheduling disciplines including PS, FBPS, ML-PRIO and FLIPS with same thresholds, in an $M/G/1$ queue.

Simulation results showed that at the threshold point the slowdown ratio function is not continuous for ML-PRIO, which means, if the attained service time is near the threshold, users would experience a sudden slowdown. With FLIPS, although the slowdown ratio increases sharply after the threshold point, but it is basically continuous at the threshold point.

Acknowledgements: We thank Urtzi Ayesta, Sem Borst and Rudesindo Núñez-Queija for discussions of the problems in the previous version of this paper.

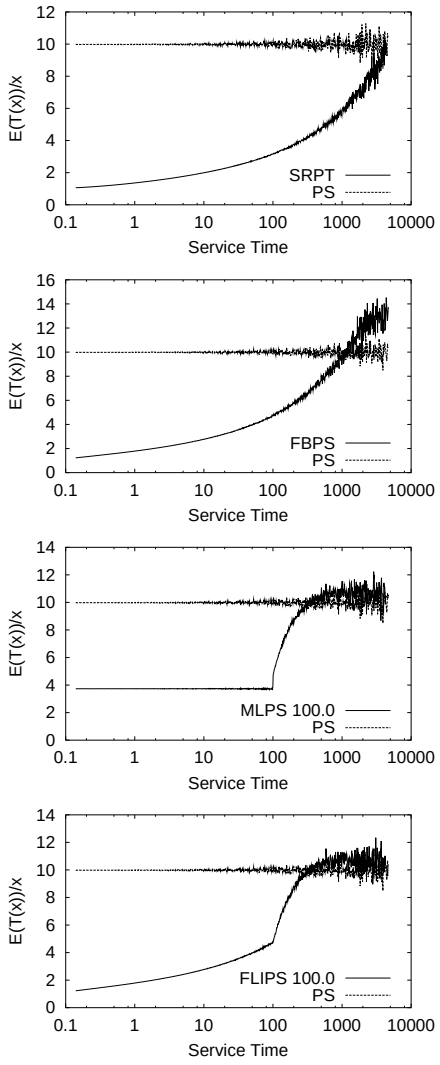

Figure 2: Simulation result of mean slowdown ratios of an $M/G/1$ queue at load $= 0.9$, with bounded-Pareto distributed service time (index $= 1.1$), as functions of service time, under various scheduling disciplines.

4. REFERENCES

[1] N. Bansal and M. Harchol-Balter, Analysis of SRPT scheduling: Investigating unfairness, ACM SIGMETRICS 2001, Cambridge, MA

[2] L. Guo and I. Matta, Scheduling flows with unknown sizes: approximate analysis, ACM SIGMETRICS 2002.

[3] L. Kleinrock and R.R. Muntz, Processor-sharing queueing models of mixed scheduling disciplines for time-shared systems, Journal of the ACM, 19, 464-482, 1972

[4] L. Kleinrock, R.R. Muntz and J. Hsu, Tight bounds on average response time for processor-sharing models of time-shared computer systems, Information Processing 71, TA-2, 50-58, August 1971.

[5] R. Righter and J.G. Shanthikumar, Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. Prob. Eng. Inf. Science, 3, 323-334.

[6] R. Righter, J.G. Shanthikumar, and G. Yamazaki, On extremal service disciplines in single-stage queueing systems, Journal of Applied Probability, 27, 409-416.

[7] A. Wierman, N. Bansal and M. Harchol-Balter, A note on comparing response times in the $M/GI/1/FB$ and $M/GI/1/PS$ queues, technique report CMU-CS-02-177, School of Computer Science, Carnegie Mellon University (Pittsburgh, PA 15213), September 2002.