The Foreground-Background queue: a survey

Misja Nuyens* Adam Wierman†

September 12, 2007

Abstract

Computer systems researchers have begun to apply the Foreground-Background (FB) scheduling discipline to a variety of applications, and as a result, there has been a resurgence in theoretical research studying FB. In this paper, we bring together results from both of these research streams to provide a survey of state-of-the-art theoretical results characterizing the performance of FB. Our emphasis throughout is on the impact of these results on computer systems.

Keywords: scheduling policies, FB, FBPS, LAS, LAST, SET, SEPT, M/G/1 queue

1 Introduction

Scheduling is a common mechanism for improving computer-system performance without purchasing additional resources. Simple policies such as First-Come-First-Served (FCFS) and Processor-Sharing (PS), which shares the service capacity equally among all jobs in the system, are most commonly used in computer systems. However, many recent system designs use policies that give priority to jobs with small service demands in order to reduce the mean response time (sojourn time, waiting time) and mean queue length, see, e.g., [27, 47].

The emergence of policies that prioritize small jobs is motivated by the Shortest-Remaining-Processing-Time (SRPT) policy, which always serves the job in the system that needs the least amount of service in order to complete: SRPT is known to be optimal with respect to mean response time and mean queue length [58, 59]. The improvement of SRPT over FCFS and PS with respect to mean response time is quite dramatic as soon as there is even moderate variability in the service distribution. This is particularly true for heavy-tailed service distributions, which appear frequently as models for service-demand distributions in computer systems, see, e.g., [17, 62].

*Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands, mnuyens@few.vu.nl

†Computer Science Department, California Institute of Technology, 1200 E. California Boulevard, MC 256-80, Pasadena, CA 91125, USA, adamw@caltech.edu.

Though SRPT is optimal with respect to mean response time, it is often not possible to use SRPT in computer systems, because in many cases, the scheduler is *blind* to the service demands of jobs. For instance, an operating system does not know how long a process will need to run, and a router does not know the length of a flow in the network. However, even when the scheduling policy is blind, i.e., it cannot use job-size information to prioritize, the scheduler can use other statistics in order to prioritize jobs with small service demands. One such statistic is the *age* or *attained service* of a job, which is defined as the amount of service already given to the job.

The Foreground-Background (FB) discipline uses the age of a job as an indication of the remaining size of the job. In particular, FB works according to the following priority rule: priority is given to the job that has received the least amount of service. If there are n such jobs, then they are served simultaneously, i.e., each of them is served at rate $1/n$. Equivalently, a queue using the FB discipline always shares the server evenly among the *youngest* jobs in the system. Nowadays, the motivation for using the age of a job as an indication of its remaining size is that under heavy-tailed distributions, jobs that have received a large amount of service are likely to be very large, and thus have remaining sizes that are still large. So, under heavy-tailed distributions, FB is acting as a "poor man's" SRPT: without knowledge of remaining sizes, it does its best to give some priority to jobs with small remaining sizes. In fact, it can be shown that, among blind policies, FB minimizes both the mean response time and queue-length distribution under a certain class of service distributions that tend to have high variance. Further, the improvement FB provides over PS and FCFS under these distributions is significant, though it is not as dramatic as the improvement provided by SRPT.

The combination of the growing acceptance of heavy-tailed distributions as models for the service-demand distributions in computer systems and the need for blind scheduling in many computer applications has led to the investigation of FB as an alternative for scheduling flows in routers [45, 46, 47], scheduling processes in operating systems [24, 61], and many other settings. However, the broad acceptance of designs based on FB has been hindered by a number of practical worries. In particular, system designers worry about the performance of jobs with large service demands (large jobs) under FB. It is clear that these jobs are biased against, and thus worries about the "unfairness" and "starvation" experienced by large jobs are pervasive. Another roadblock to the acceptance of FB in practical applications is that, though FB performs very well under some classes of heavy-tailed distributions, there are classes of light-tailed distributions where FB performs quite badly. Thus, it is important for system designers to understand how to determine if using FB is appropriate in their situation.

Over the last five years, parallel streams of research studying FB have emerged. While many researchers have focused on traditional queueing analysis of FB, other researchers focused on addressing the practical roadblocks to the acceptance of FB in computer systems. Consequently, results about FB are scattered across the literature. In many cases, even the name of the policy is

not consistent across domains: instead of FB, different acronyms are sometimes used, e.g., Least-Attained-Service (LAS) and Shortest Elapsed Time (SET). In this paper, we survey recent results from both theoretical and practical work with the goal of providing an up-to-date reference point for researchers interested in applying or analyzing FB. By bringing together the results from both of these research streams, we can provide a complete picture of the behavior of FB and how it compares with other policies. Where possible, we will contrast the behavior of FB with the behavior of the two most common blind policies used in computer systems, FCFS and PS. We also compare the behavior of FB with that of SRPT in order to illustrate the penalty that FB pays for not using job-size information.

The most complete reference of results on FB to this point is Yashkov's survey of PS queues from fifteen years ago [72], which includes some detail about the FB queue, but as the reader will see, there has been significant progress in our understanding of FB since that point.

This survey is organized as follows. We begin in Section 2 with a more detailed description of the workings of FB. Then, we discuss the historic evolution of our understanding of FB queues in Section 3. Following these introductory sections, we move into the body of the survey, the results about the FB queue. We start in Section 4 by describing the optimality results that make FB such an attractive discipline to consider. Then, we move to qualitative results on FB. In Section 5 we describe the behavior of the mean response time and mean queue length of FB queues. In this section we focus on (i) the impact of variability in the service distribution on the mean response time and (ii) the growth rate of the mean response time as a function of the load in heavy-traffic. Following this, we move beyond the mean behavior of the FB queue and discuss the distributional behavior of the response time and queue length in Section 6. In this section we deal with a variety of distributional measures including the moments of response time and queue length, the tail behavior of the response time, and the tail behavior of the maximal queue length. Finally, in Section 7 we discuss work describing the experience of large job sizes under FB. We conclude by highlighting a number of future research topics in Section 8.

In this paper, we use the following notation. The generic service time is denoted by X, its cumulative distribution function by F, and we let $\bar{F} = 1 - F$. If the service distribution has a density, it is denoted by f. The upper endpoint of the service distribution, x_U, is defined as $x_U = \sup\{x : F(x) < 1\}$. In case of M/GI/1 queues, the arrival rate is denoted by λ. Unless stated otherwise, we consider queues for which the stability condition $\rho = \lambda EX < 1$ holds. The response time in the stationary queue is denoted by V, and the stationary queue length by Q. Service disciplines (policies) are written in sans-serif font, e.g., FCFS.

A function f satisfies $f(x) = \Omega(g(x))$ if $\liminf f(x)/g(x) > 0$, satisfies $f(x) = O(g(x))$ if $\limsup f(x)/g(x) < \infty$, and $f(x) = \Theta(g(x))$ if $f(x) = O(g(x))$ and $f(x) = \Omega(g(x))$. Further, $f(x) = o(g(x))$ if $\limsup f(x)/g(x) = 0$, and $f(x) \sim g(x)$ means that $\lim f(x)/g(x) = 1$. Finally, $\stackrel{d}{=}$ means "equal in distribution", and $a \wedge b$ stands for $\min\{a, b\}$.

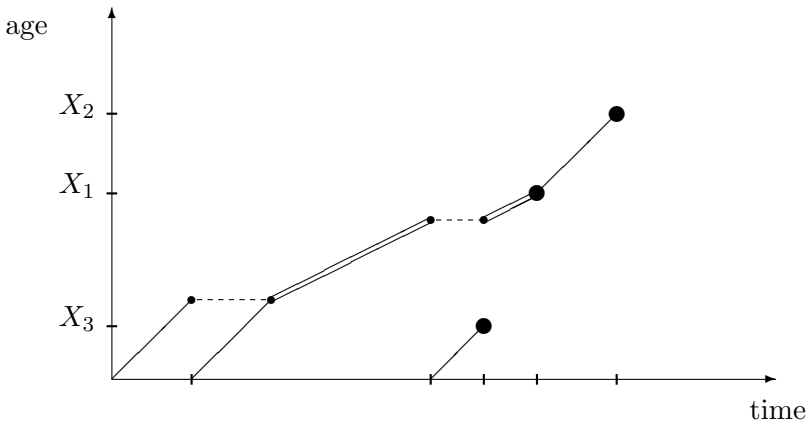


Figure 1: The age process of three jobs in the FB queue, with service times X_1, X_2 and X_3. Small circles indicate that the server switches, large circles denote departures.

2 An introduction to FB

FB works according to the following simple priority rule: priority is given to the job that has received the least amount of service. If there are n such jobs, for some $n \in \mathbb{N}$, then they are served simultaneously, i.e., each of them is served at rate $1/n$. See Figure 1 for an example of this operation.

As we briefly discussed in the introduction, the motivation for using the age of a job as an indication of the remaining size of a job is that, under many heavy-tailed distributions, jobs that have received a large amount of service are likely to be *much* larger, and thus have large remaining sizes. Specifically, consider the class of distributions that have a decreasing failure rate (DFR), i.e., $\mu(x) = f(x)/\bar{F}(x)$ is non-increasing for all $x \geq 0$. The failure rate can intuitively be seen as an "instantaneous departure probability." For DFR service distributions, larger jobs have a smaller failure rate, and thus are less likely to complete when they are served. So, the FB priority rule corresponds to *greedy* scheduling in this setting. This intuition is the reason that FB is appealing: in many computer applications, DFR distributions such as the Pareto distribution have been suggested to model the service distribution. However, the same logic implies that FB is likely to behave quite badly if the service distribution has an increasing failure rate (IFR), for example, if the service distribution is uniform.

To get a feeling for the evolution of the queue under the FB discipline, let us consider what happens when a new job arrives to the FB queue. Since that job is strictly the youngest in the queue, it is served immediately. Then, as the queue evolves, there are three possible scenarios:

1. The new job needs at least as much service as the age of the job(s) that was (were) preempted at its arrival. In this case, after some time the job *joins* a *cohort*, a group of jobs with the same age. This happens to the second customer in Figure 1.

2. The new job needs less service than the age of the jobs in the cohort that was preempted. In this case, the new job leaves the queue before joining the older cohort and the server returns to the cohort that was preempted. This is illustrated by the third customer in Figure 1.

3. Before joining another cohort or leaving the queue, the new job is preempted itself by the arrival of another new job. This happens to the first customer in Figure 1.

Since a job with service time x is younger than x throughout its stay in the queue, it has *priority* over all jobs older than x. As a consequence, the time such a job spends in the system is the same as if all service times would be truncated at x, i.e., if all service times y would have value $y \wedge x$ instead. Hence, in the FB queue, small jobs do not suffer from the presence of large jobs in their midst. The response times of small jobs are therefore *insensitive* to the shape of the tail of the service distribution. This property turns out to be very useful when studying response times in the FB queue. One consequence of the isolation of small jobs is felt by the jobs with large service demands: these jobs receive service primarily when no other jobs are present in the queue, see Theorem 7.1 and the related discussion. As a result, one of the main issues addressed by recent studies of FB queues is determining the price that large jobs have to pay as a result of the priority FB provides for small jobs, i.e., the question of how "unfairly" large jobs are treated.

3 The history of FB

FB first emerged in the literature in the second half of the 1960s. The term FB, or rather FB_n, was used as an abbreviation for both *Foreground-Background* and *Feedback* queueing systems. These different names referred to the same model, see Schrage [57], Coffman and Kleinrock [16], and the survey article on time-sharing models by McKinney [36]. The FB_n queue with so-called *quantum size* q is a one-server queue with n states, or priority classes. This queue operates as follows. Upon arrival in the queue, a job enters the first (or highest priority) state. Within each priority state, the priority of jobs depends on their arrival time to that state, in a FCFS manner. Jobs are served one at a time and uninterruptedly for a time period of length q. After the server has completed a job's service request in a certain state, a job from the highest (non-empty) priority state is selected for service. If a job does not leave the queue during its time in the kth state, it moves to state $k+1$, which has lower priority, and waits until it is served in that state. In the nth and final state, jobs are served only if there are no jobs in other states. In that final state, they are served until they leave the system. So, for example, FB_1=FCFS. Interest in the FB_n model with n states and positive quantum size q has faded. However, in many applications it is more practical to implement FB_n

than FB and, thus, it is still a relevant discipline. In particular, Aalto et al. [1, 3, 4] have recently published a number of interesting papers on this policy.

To a large extent, people have moved to studying the limiting case where first $n \to \infty$ and then $q \to 0$. After Kleinrock devoted a section of [31] to this limiting case of the FB_n model, the term Foreground-Background (FB) became the generally accepted name for this model in the queueing community, and so it is in this paper. But, in the literature many other names for FB appear. To distinguish FB from the FB_n model, some authors prefer to use the term Foreground-Background *Processor Sharing* (FBPS), FB_∞, or *Generalized* Foreground-Background (GFB). Others have come across the policy independently and named it based on its priority rule. For example, in the computer science community the acronyms LAS (Least Attained Service first) and LAST (Least Attained Service Time first) are common. Further, in the worst-case scheduling community the acronyms SET (Shortest Elapsed Time) or SEPT (Shortest Elapsed Processing Time) tend to be used. These should not be confused with the Shortest Expected Time and Shortest Expected Processing Time disciplines. Furthermore, in Pechinkin [44], FB is disguised as 'advantageous sharing of a processor'. Due to this cacophony of names, some results on FB queues are difficult to find in the literature. One of the goals of the present survey is to unite the FB world again, and to provide a clear overview of all available results.

Though FB was first discussed as early as the 1960s, it was given very little attention until the 1980s, when more results started to appear. During the 1960s and 1970s much of the research on FB queues was done by Schrage [57] and Kleinrock [31], who derived the mean and Laplace transform of the response time for a job of size x under FB. However, little else about FB was studied. There seem to be two reasons for this lack of attention: (i) in those days, there was less interest in queues with heavy-tailed characteristics, and (ii) the analysis of the FB queue tends to be more difficult than queues with other common policies. However, around 1980 interest in FB began to build. Pechinkin [44], Schassberger [56] and Yashkov [70] obtained expressions for the generating functional of the steady-state queue length, and using these expressions along with the earlier analyses of Kleinrock and Schrage, people began to study behavioral properties of FB. Yashkov [72] provided a survey of most results known at the time. Soon after, during the early 1990s, Righter and Shanthikumar [52, 60] proved a set of optimality results for the queue length of FB.

Since 2000, there have been a host of new results about FB. Much of this recent work on FB has been motivated by proposals of computer system designers suggesting the use of FB in a variety of applications such as scheduling flows at routers [45, 46, 47] and scheduling processes in operating systems [24, 61]. These proposals have created the need for a more detailed study of the behavior of FB with respect to traditional queueing measures, in addition to introducing a number of new, non-traditional queueing measures. Areas of recent progress in understanding the behavior of FB queues include the distribution of the response time, the performance of large jobs under FB, and

the maximal queue length.

4 The optimality of FB

The fundamental motivation for the use of FB in practical settings is that, under a large class of practical distributions, FB minimizes the queue-length distribution and mean response time among *all* blind scheduling policies. In this section, we will summarize these optimality properties of FB.

The first result about the optimality of FB was mentioned by Yashkov in [71]: FB minimizes the mean queue length, and thus the mean response time, across all blind disciplines when the service distribution has a decreasing failure rate (DFR). Soon after, Righter and Shantikumar [52] proved that FB minimizes not only the mean queue length, but even the marginal distribution of the queue length. To state their theorem, let $Q(t)^{\mathsf{P}}$ denote the queue length at time t in the queue with discipline P. Furthermore, a random variable X is called *stochastically smaller* than Y, denoted $X \leq_{st} Y$, if $P(X \leq x) \geq P(Y \leq x)$ for all x. Then regardless of the number of jobs present at $t = 0$ and their ages, we have the following result.

Theorem 4.1 *Consider a GI/GI/1 queue. Let* P *be a blind policy and let* NP *be a policy that is both non-preemptive and blind. If the service distribution belongs to the class DFR, then for every* $t \geq 0$*,*

$$Q(t)^{\mathsf{FB}} \leq_{st} Q(t)^{\mathsf{P}} \leq_{st} Q(t)^{\mathsf{NP}}.$$

Further, for IFR service distributions the inequalities are reversed.

The condition in Theorem 4.1 that the service distribution should have a decreasing failure rate is not very surprising in the light of the following. The failure rate of a job, $f(x)/\bar{F}(x)$, can be seen as its "instantaneous departure probability". Therefore, FB is taking a greedy approach as it minimizes the queue length in the short run by serving the job that is most likely to complete soon, i.e., the job with the highest failure rate.

The proof of Theorem 4.1 uses a coupling argument based on exactly the above intuition, i.e., any policy other than FB must pay a price for not serving the job that is most likely to complete. Although Theorem 4.1 was originally proven in the GI/GI/1 setting, the proof also goes through when the arrival process is allowed to be any sequence, e.g., a deterministic sequence.

It is worthwhile to mention two other approaches that result in a similar, but weaker version of Theorem 4.1. Specifically, these approaches can be used to prove that the mean queue length and mean response time of FB is optimal among blind policies. The first approach is to make use of the index approach developed by Gittins [25] and Klimov [32, 33]. When the service distribution is DFR, FB can be viewed as the policy that always schedules the job with the highest so-called Gittins index, From this, the optimality then follows immediately. This approach has recently been

applied to develop optimal blind policies for non-DFR service distributions, see Aalto and Ayesta [5].

A second approach makes use of the following extremal property of FB found in Righter et al. [53] and Aalto et al. [3]: for any service distribution, each sample path, and any x and t, FB minimizes the so-called unfinished truncated work, $U_x(t)$, i.e., the amount of work needed to serve all jobs until their age is at least x. This property implies a stochastic ordering of the corresponding processes $\{U_x^{\mathsf{FB}}(t); t \geq 0\}$ and $\{U_x^{\mathsf{P}}(t); t \geq 0\}$, where P is an arbitrary blind policy. This ordering can then be used to prove that FB minimizes the mean response time among blind policies for DFR service distributions (see Theorem 1 in [3]). This approach can also be applied for proving the optimality of FB among blind policies with respect to weighted response-time measures. In particular, Feng and Misra [22] have used this approach to show that if $f(x)/(x\bar{F}(x))$ is decreasing, then FB optimizes the mean *slowdown* (or stretch), where the slowdown of a job of size x is defined as $V(x)/x$, and $V(x)$ is the response time of a job of size x.

Although FB optimizes the marginal distributions of the queue length for DFR service distributions, a stronger condition on the density of the service distribution is needed to obtain an optimality result for the law of the whole queue-length process, $\{Q(t), t \geq 0\}$. This stronger condition is that the density f be *log-convex*, i.e., $\log f$ is convex. By integration, one can show that the class of log-convex densities is a subclass of DFR and the class of log-concave densities is a subclass of IFR. Note that the class of distributions with a log-convex density includes many well-known distributions, e.g., Pareto distributions with density $f(x) = c\alpha/(cx+1)^{(\alpha+1)}, \alpha, c > 0$ and gamma distributions with density $f(x) = \lambda^r x^{r-1} \exp(-\lambda x)/\Gamma(r), \lambda, x \geq 0$, and $0 < r <\leq 1$. For more results on these classes of distributions we refer to Shaked and Shanthikumar [60].

Using the stronger condition of log-convex service densities, Righter proved in [60] the following result by means of another coupling argument:

Theorem 4.2 *Let* P *be a blind policy and let* NP *be a policy that is both non-preemptive and blind. If the service distribution has a log-convex density, then*[1]

$$\{Q(t)^{\mathsf{FB}}, t \geq 0\} \leq_{st} \{Q(t)^{\mathsf{P}}, t \geq 0\} \leq_{st} \{Q(t)^{\mathsf{NP}}, t \geq 0\}.$$

For service times with a log-concave density, the inequalities are reversed.

So far we have seen that FB optimizes the queue-length distribution under DFR distributions and the queue length process under log-convex distributions. However, one expects that FB may

[1]The stochastic ordering of processes is a generalization of stochastic ordering for random variables, and can be defined similarly, see also Section 4.B.7 of Shaked and Shanthikumar [60]. We say that two processes $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$ are stochastically ordered, notation $\{X(t), t \geq 0\} \leq_{st} \{Y(t), t \geq 0\}$, if there exist processes $\{\bar{X}(t), t \geq 0\}$ and $\{\bar{Y}(t), t \geq 0\}$, defined on a common probability space, such that $P(\bar{X}(t) \leq \bar{Y}(t)\ \forall t) = 1$ and $\{X(t), t \geq 0\} \stackrel{d}{=} \{\bar{X}(t), t \geq 0\}$, $\{Y(t), t \geq 0\} \stackrel{d}{=} \{\bar{Y}(t), t \geq 0\}$.

perform well even when the service distribution is outside of these classes. In particular, one expects that FB will perform well whenever the age of a job is positively correlated with the remaining size of a job. Thus, it is natural to wonder whether FB will also have optimality properties under service distributions with an increasing mean residual life (IMRL), since under these distributions a job with a larger age has a larger expected remaining size. However, studying the behavior of FB under IMRL distributions has proven to be tricky. In [53], Righter et al. state that FB minimizes the mean queue length under IMRL distributions, but the proof of the result contains an error that cannot be immediately fixed, as was noted by Aalto et al. [3]. The proof considers the unfinished work of jobs with age less than x, but they do not take into account that this quantity makes a vertical downward jump whenever a job reaches age x. A similar result of Feng and Misra [22] contains the same mistake.

Further, Aalto and Ayesta [2] have recently found a counter-example to the idea that FB optimizes mean queue length under IMRL distributions. They showed that a hybrid policy that combines FB and FCFS can have smaller mean queue length than FB under IMRL service distributions having the form

$$f(x) = \begin{cases} c^{-x} \log c, & 0 \leq x \leq c, \\ cx^{-c-1}, & x > c, \end{cases}$$

with $1 < c < e$. Notice that this distribution has a failure rate that is first increasing and then decreasing, so in the light of Theorem 4.1, it is not too surprising that a combination of FCFS and FB can provide a smaller mean queue length than FB under service distributions. Of course this does not mean that FB does not perform well under IMRL service distributions. As we will see later, the mean queue length and mean response time of FB under IMRL distributions is smaller than that of other common blind policies, e.g., PS and FCFS.

5 The mean performance of FB

To this point, we have seen that FB minimizes/maximizes the queue-length distribution under DFR/IFR distributions, and thus also the mean queue length EQ and the mean response time EV. However, these results do not say to what extent the mean response time of FB is better/worse than that of other blind policies. Further, the results provide no indication of how much worse the performance of FB is than the performance of policies that can use job-size information, e.g., SRPT. Also, it is important to understand how FB performs under distributions that are not DFR or IFR. These and related questions are of fundamental importance if one is considering to use FB in practice. To answer them, we need explicit expressions for the mean response time and mean queue length under FB. In this section we will focus entirely on mean response time. Using Little's Law, $EQ = \lambda EV$, the results can easily be translated to the mean queue length as well.

As is common under priority-based policies, the approach for deriving the mean response time of FB is to first study the conditional response time of FB, $V(x)^{\mathsf{FB}}$, which is defined as the response time experienced by a job of size x. The first derivation of $EV(x)^{\mathsf{FB}}$ was by Schrage [57] and holds for the M/GI/1 queue:

$$EV(x)^{\mathsf{FB}} = \frac{x}{1-\rho(x)} + \frac{\lambda m_2(x)}{2(1-\rho(x))^2} = \frac{x}{1-\lambda\int_0^x \bar{F}(t)dt} + \frac{\lambda\int_0^x t\bar{F}(t)dt}{(1-\lambda\int_0^x \bar{F}(t)dt)^2}, \tag{1}$$

where $m_i(x) = i\int_0^x t^{i-1}\bar{F}(t)dt$ are the moments of $X \wedge x$, and $\rho(x) = \lambda m_1(x)$.

Though this expression may look complicated at first, it is actually quite natural. Consider the experience of a job of size x, j_x, under FB. Since no job older than x will ever receive service while there is a job younger than x in the system, we can transform the service distribution from X to $X \wedge x$ without affecting the response time of j_x. Further, notice that the transformed system is still work conserving. Finally, notice that j_x finishes exactly when this transformed system goes idle. Define $L_x(y)$ as the length of a busy period, started with a job of size y, where arrivals occur at rate λ and with service times distributed as $X \wedge x$. Denoting the steady-state workload of the transformed system by W_x^{FB}, we then have

$$V(x)^{\mathsf{FB}} \stackrel{\mathrm{d}}{=} L_x(x + W_x^{\mathsf{FB}}) \stackrel{\mathrm{d}}{=} L_x(x) + L_x(W_x^{\mathsf{FB}}). \tag{2}$$

Since $L_x(y)$ is the same for all work-conserving disciplines, equation (1) follows from $E[L_x(Y)|Y] = Y/(1-\rho(x))$, $E[E[L_x(Y)|Y]] = EY/(1-\rho(x))$, and $EW_x^{\mathsf{FB}} = \lambda m_2(x)/(2(1-\rho(x)))$.

Before moving to a discussion of the overall mean response time of FB, it is worthwhile to make a few observations about the behavior of $V(x)^{\mathsf{FB}}$ described by (1). An important observation we can make about (1) is that it clearly indicates that FB results in a strong bias towards small job sizes, regardless of the service distribution. In fact, small job sizes are insensitive to the tail behavior of the service distribution. To illustrate this, notice that as $x \to 0$, $EV(x) \sim x$, which indicates that the smallest jobs have response times that are as small as if they were served in isolation, regardless of the system load. In fact, small job sizes are isolated from large job sizes even when $\rho > 1$. In particular, FB will remain stable for all job sizes x such that $\rho(x) < 1$.

Large job sizes experience a very different performance of the queue than small job sizes, regardless of the service distribution: Harchol-Balter and Wierman [28] and [40] derived from (1) that $EV(x) \sim x/(1-\rho)$ for $x \to \infty$. This asymptotic relation indicates that the largest job sizes are receiving service primarily when there are no other jobs in the system. In particular, the average service rate given to large jobs converges to the total service rate, namely 1, reduced by the load of jobs that pass through the system in the meantime, which in the limit is ρ, since the system remains stable.

Moving to the overall mean response time, we can now use (1) to calculate EV^{FB} as follows:

$$EV^{\mathsf{FB}} = \int_0^\infty EV(x)^{\mathsf{FB}} dF(x) = \int_0^\infty \left(\frac{x}{1-\lambda\int_0^x \bar{F}(t)dt} + \frac{\lambda\int_0^x t\bar{F}(t)dt}{(1-\lambda\int_0^x \bar{F}(t)dt)^2} \right) dF(x). \tag{3}$$

Using (3) we can easily obtain the stability conditions for an FB queue. In particular, by combining equation (3) with the relation $EV(x)^{\mathsf{FB}} \sim x/(1-\rho)$ as $x \to \infty$, we can prove that $EV^{\mathsf{FB}} < \infty$ and $EQ^{\mathsf{FB}} < \infty$ whenever $EX < \infty$ and $\rho < 1$.

Beyond showing stability, (3) can also be used to bound the mean response time under FB. In particular, combining the results of Yashkov [72] and Wierman et al. [67], we have the following bounds:

Theorem 5.1 *In an M/GI/1 queue,*

$$\frac{EX}{\rho}\log\left(\frac{1}{1-\rho}\right) \leq EV^{\mathsf{FB}} \leq \frac{EX}{2}\frac{(2-\rho)}{(1-\rho)^2}. \tag{4}$$

These bounds give an indication of how well FB can do when it is at its best and how poorly it behaves when it is at its worst. The upper bound follows from (3) by a series of inequalities. For deterministic service distributions, all these inequalities are actually equalities, and thus the upper bound is attained. The lower bound follows from partial integration of (3) and ignoring a number of positive terms. As a consequence, the lower bound itself is never attained, but it has been shown to be asymptotically tight in heavy traffic. Specifically, it has been shown that as $\rho \to 1$, $EV^{\mathsf{FB}} = \Theta(\log(1/(1-\rho))$ under certain Pareto service distributions [9, 10], see also Theorem 5.3 below.

Though (3) can be used to bound the mean response time of FB, its complicated form means that it is difficult to gain an understanding of the *behavior* of EV^{FB}. For example, the impact of (i) the variability in the service distribution and (ii) the load on EV^{FB} are not obvious from (3). In the next two subsections we will describe some recent work that has begun to characterize the impact of these parameters.

5.1 The impact of variability

We have already seen a number of indications of how large an effect the variability of the service distribution can have on the response times of FB. We have seen that FB minimizes mean response time across blind policies under DFR distributions, which tend to be highly variable, and maximizes mean response time across blind policies under IFR distributions, which tend to have low variability. From these results, the expectation is that FB will perform well for all highly variable distributions and poorly for all distributions with low variability. There exist many statements in the literature to this effect. For instance, Yashkov [72] writes that, in the stationary FB queue, "EV decreases with an increase in the *dispersion* of $F(x)$, and conversely increases as the dispersion of $F(x)$ decreases." This seems to be supported by Figure 2. However, the story is not so simple.

Throughout this section, we will use the squared coefficient of variation, defined by $C^2[X] = Var[X]/E[X]^2$, in order to characterize the variability or dispersion of the service distribution. Distributions with $C^2[X] > 1$ are said to have high variability and distributions with $C^2[X] < 1$

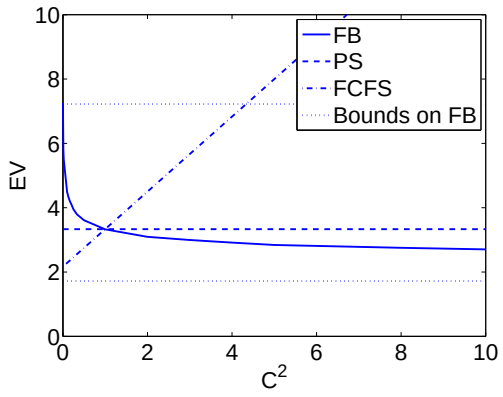


Figure 2: An illustration of the impact of service time variability on the mean response time of FB. The service distributions in this figure are Weibull with mean 1 and the load is 0.7. The bounds on FB are from Theorem 5.1

are said to have low variability. It can be seen that DFR and IMRL distributions all have $C^2[X] \geq 1$, while IFR and DMRL distributions all have $C^2[X] \leq 1$. For the exponential distribution, $C^2[X] = 1$.

A good starting point for discussing the effect of variability on $E[V^{\mathsf{FB}}]$ is the M/M/1 queue. In this setting, the service distribution has a constant failure rate, and is thus both IFR and DFR. Thus, all blind policies have the same queue-length distribution and mean response time, and in particular

$$E[V^{\mathsf{FB}}] = E[V^{\mathsf{FCFS}}] = E[V^{\mathsf{PS}}] = \frac{E[X]}{1-\rho}.$$

Interestingly, the M/M/1 queue serves as the crossover point for the mean response times of FCFS and PS, i.e.,

$$E[V^{FCFS}] \geq E[V^{\mathsf{PS}}] \Leftrightarrow C^2[X] \geq 1.$$

Based on this observation, Coffman and Denning [15] made the natural parallel conjecture for FB:

$$E[V^{\mathsf{FB}}] \leq E[V^{\mathsf{PS}}] \Leftrightarrow C^2[X] \geq 1. \tag{5}$$

The conjecture is supported, for example, by Figure 2, where $E[V^{\mathsf{FB}}]$ is decreasing in $C^2[X]$, while $E[V^{\mathsf{PS}}]$ is constant and $E[V^{\mathsf{FCFS}}]$ is increasing. Conjecture (5) was thought to be true until recently when Wierman et al. [64] provided a counterexample. They showed that when the service distribution is such that $P(X = 1) = 4/5 + \epsilon$ and $P(X = 6) = 1/5 - \epsilon$ for some $0 < \epsilon < 1/10$, then $C^2[X] > 1$, but $E[V^{\mathsf{FB}}] > E[V^{\mathsf{PS}}]$. Using a similar distribution, Feng and Misra [23] go on to show that not only does $C^2[X] > 1$ not imply that $E[V^{\mathsf{FB}}] < E[V^{\mathsf{PS}}]$, but there exist distributions with

$C^2[X] > 1$ for which FB has mean response time arbitrarily close to the upper bound in (4), which is the maximal mean response time across all work conserving policies.

Surprisingly, these results show that variability, as measured by $C^2[X]$, is not a strong enough criterion to guarantee that FB performs well, i.e., better than PS. However, we still expect that FB should be guaranteed to perform well for some subclass of highly variable distributions. Very recently, Aalto and Ayesta [2] show that this is indeed the case: though we discussed earlier that FB does not optimize mean response time under IMRL distributions, FB does have smaller mean response time than PS under all IMRL distributions:

Theorem 5.2 *In an M/GI/1 queue with an IMRL service distribution,*

$$EV^{\mathsf{FB}} \leq EV^{\mathsf{PS}} = \frac{EX}{1-\rho}.$$

Further, the inequality is reversed under DMRL service distributions.

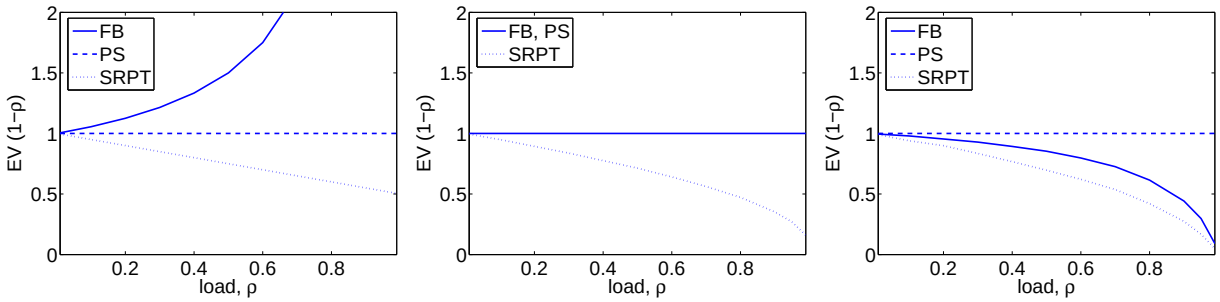
5.2 The impact of load

We will now discuss how the mean response time of FB grows with the system load. There are two types of results along these lines. The first type characterizes the growth rate of EV^{FB} as $\rho \to 1$, i.e., the *heavy-traffic growth rate* of FB. The heavy-traffic growth rate is a key metric to many computer applications, since systems are often run at very high loads. The second type characterizes the growth rate of the queue length when the system is unstable, i.e., the behavior of $Q(t)/t$ when $\rho > 1$. This growth rate is important in many practical settings since computer applications tend to experience periods of overload, and minimizing the number of jobs that build up during these periods limits the impact of these overload periods. For instance, resource provisioning must be done with the overload periods in mind. We will start by presenting results characterizing the heavy-traffic growth rate of EV^{FB} as $\rho \to 1$ and then discuss the behavior of FB when $\rho > 1$.

For many common policies, determining the heavy-traffic growth rate is quite easy. For instance, $EV^{\mathsf{PS}} = EX/(1-\rho)$, thus the mean response time is $\Theta(1/(1-\rho))$ as $\rho \to 1$ regardless of the service distribution. Similarly, the growth rate of EV^{FCFS} is $\Theta(1/(1-\rho))$ whenever $EX^2 < \infty$. In contrast, determining the heavy-traffic growth rate of FB is a difficult task due to the complex form of EV^{FB} in (3).

As an indication of this difficulty, notice that the bounds in (4) show that the behavior of EV^{FB} as a function of load strongly depends on the service distribution: the lower bound grows as $\Theta(\log(1/(1-\rho)))$, while the upper bound grows as $\Theta(1/(1-\rho)^2)$. Therefore, many of the results characterizing the heavy-traffic growth rate of FB do so for only a small class of service distributions. Summarizing the results of Nuyens [40], Bansal and Gamarnik [9], and Wierman et al. [67], we have the following

Theorem 5.3 *Consider an M/GI/1 queue.*



(a) Deterministic job sizes (b) Exponential job sizes (c) Pareto job sizes, $\alpha = 1.5$.

Figure 3: An illustration of the impact of the service distribution on the growth rate of EV^{FB} with load.

(i) *If the service distribution is deterministic, then* $EV^{\mathsf{FB}} = \Theta\left(\frac{1}{(1-\rho)^2}\right)$ *as* $\rho \to 1$.

(ii) *If* $x_U < \infty$, *and the service distribution is of the form* $\bar{F}(x) \sim \alpha(x_U - x)^\beta$ *as* $x \to x_U$ *for some* $\alpha, \beta > 0$, *then* $EV^{\mathsf{FB}} = \Theta\left(\frac{1}{(1-\rho)^{1+1/(\beta+1)}}\right)$ *as* $\rho \to 1$.

(iii) *If the service distribution is exponential, then* $EV^{\mathsf{FB}} = \Theta\left(\frac{1}{1-\rho}\right)$ *as* $\rho \to 1$.

(iv) *If the service distribution is Pareto*(α) *with* $\bar{F}(x) = (k/x)^{-\alpha}$ *for* $x \geq k$, *then as* $\rho \to 1$,

$$EV^{\mathsf{FB}} = \begin{cases} \Theta\left(\log\left(\frac{1}{1-\rho}\right)\right), & \text{if } 1 < \alpha < 2 \\ \Theta\left(\log^2\left(\frac{1}{1-\rho}\right)\right), & \text{if } \alpha = 2 \\ \Theta\left((1-\rho)^{-\frac{\alpha-2}{\alpha-1}}\right), & \text{if } \alpha > 2. \end{cases}$$

The results in Theorem 5.3 illustrate the contrasting impact of load on FB under different service distributions. These are also illustrated in Figure 3. We can see that in cases (i) and (ii) the heavy-traffic growth rate of FB is strictly worse than that of PS and FCFS. This is not surprising since under bounded distributions, the failure rate increases unboundedly as $x \to x_U$. In contrast, the heavy traffic growth rate in case (iv) is much smaller than that of PS and FCFS. In fact, under Pareto distributions with $1 < \alpha < 2$, the heavy-traffic growth rate matches that in the lower bound on EV^{FB} in (4), which indicates that FB performs best when the service distribution is very heavy-tailed. Further, under Pareto service distributions, FB nearly matches the behavior of SRPT, the policy with the smallest mean response time. In fact, Bansal and Wierman have shown in [10] that FB nearly matches SRPT for all regularly varying distributions - a generalization of Pareto distributions, see Definition 6.2 below.

Theorem 5.3 illustrates a trend that we are seeing throughout this survey: the heavy-traffic growth rate of FB is better than that of PS and FCFS when the service distribution is "highly variable", and worse when the service distribution is "lightly variable." However, apart from the four classes of distributions studied in Theorem 5.3, it has not been determined what properties

of the service distribution lead to good or bad heavy-traffic growth rates under FB. As a step towards answering this question, we prove the following new theorem. The theorem shows that a key determining factor as to whether the heavy traffic growth rate of FB is better or worse than PS and FCFS is whether or not there is an upper bound on the service distribution. We include the proof of the theorem in the appendix in order to provide an example of how results about the heavy-traffic growth rate tend to be proven while not interrupting the flow of the survey.

Theorem 5.4 *Consider an M/GI/1 queue with a continuous service distribution with failure rate* $\mu(x)$.

(i) If the service distribution is bounded, then $EV^{\mathsf{FB}} = \Omega(1/(1-\rho))$ *as* $\rho \to 1$.

(ii) If the service distribution is unbounded and $m_2(x)\mu(x) = O(1)$*, then* $EV^{\mathsf{FB}} = O(1/(1-\rho))$ *as* $\rho \to 1$.

Note that the condition (ii) holds for most well-behaved unbounded distributions. For instance, if $EX^2 < \infty$, then it simply requires that $\mu(x)$ is bounded, which occurs under all common unbounded distributions – though it is possible to construct examples where this is not the case, e.g., $f(x) = \sum_{n=1}^{\infty} I_{[n,n+2^{-n}]}(x)$, where I_A is the indicator function of the set A. On the other hand, if $EX^2 = \infty$, then $\mu(x)m_2(x) = O(1)$ requires a tradeoff between the growth of the second moment and the rate of decrease of the hazard rate. However, this tradeoff is met under most common distributions. For example, under regularly varying distributions, which include all Pareto distributions, $\mu(x) = \Theta(1/x)$ and $m_2(x) = O(x)$.

We will now move from discussing the behavior of EV^{FB} in heavy traffic (as $\rho \to 1$) to discussing the behavior of FB under overload (when $\rho > 1$). In this setting, the queue is unstable, so the goal is to characterize the growth rate of the queue length over time. Policies that have smaller growth rates are much more practical in computer systems since they limit the amount of over-provisioning necessary to handle periods of overload.

Balkema and Verwijmeren [8] have characterized the growth rate of Q^{FB} as follows:

$$\frac{Q(t)^{\mathsf{FB}}}{t} \to \lambda \bar{F}(x^*) \quad \text{a.s. as } t \to \infty,$$

where the critical service time x^* is the unique solution of $\rho(x) = 1$. Notice that the growth rate of the queue length in overload depends only on the behavior of the largest job sizes. This is in stark contrast with the behavior of PS and FCFS. For example, under PS, Jean-Marie and Robert [29] have proven that

$$\frac{Q(t)^{\mathsf{PS}}}{t} \to \eta^* \quad \text{a.s. as } t \to \infty,$$

where η^* is the unique, positive solution to

$$\lambda \int_0^{\infty} e^{-\eta x} \bar{F}(x) dx = 1.$$

Further, under FCFS it is straightforward to see that

$$\frac{Q(t)^{\mathsf{FCFS}}}{t} \to \lambda - \frac{1}{EX} \quad \text{a.s. as } t \to \infty.$$

From the above, we see that there is an interesting contrast between the growth rate of the queue length under FCFS and those of FB and PS. If $EX = \infty$, then $Q(t)^{\mathsf{FCFS}}/t \to \lambda$ a.s. and the fraction of customers that manage to leave the system is negligible. However, under PS and FB, a positive fraction of all customers leaves the queue.

Beyond this observation, we can also compare the growth rates of the queue lengths under these policies numerically. In particular, Nuyens [40] has observed that under Pareto service times, the asymptotic growth rate of the queue length of FB is smaller than those of PS and FCFS. Further, for exponential service distributions, the queue lengths under FB and PS are stochastically equal, and for deterministic distributions the growth rate of FCFS is smaller than those of PS and FB.

6 Beyond the mean performance of FB

In the previous section we focused on the mean performance of FB. Though the mean performance is clearly an important measure for practical settings, the distributional performance is often just as important. In fact, users even seem to prefer response times that are larger on average if the response times are less variable, and thus more predictable [20, 74]. Further, understanding the distributional behavior of response times and queue lengths is fundamental when considering QoS, admission control, and capacity planning applications, where guarantees of the form "95% of the time the response time (buffer size) is smaller than C" are desired.

In this section, we will summarize the results characterizing the behavior of the response-time distribution (Section 6.1) and the queue-length distribution (Section 6.2). Direct analysis of the distributional behavior of response times and queue lengths is typically only possible in very specialized settings, such as the M/M/1 queue, and under only very simple policies, such as FCFS. Thus, under FB, studies of the distributional behavior of response time and queue length typically use asymptotic scalings of the distributions.

6.1 The response-time distribution

In order to study the distribution of V^{FB}, we will use the same approach that we used earlier when studying EV^{FB}. In particular, we will start by studying the distribution of the conditional response time $V(x)^{\mathsf{FB}}$, and then apply the results in order to characterize the distribution of V^{FB}.

We saw during our derivation of $EV(x)^{\mathsf{FB}}$ that it is not too difficult to characterize the entire distribution of $V(x)^{\mathsf{FB}}$. In fact, the Laplace transform of $V(x)^{\mathsf{FB}}$, defined as $\mathcal{L}_{V(x)^{\mathsf{FB}}}(s) = Ee^{sV(x)^{\mathsf{FB}}}$

and derived by Kleinrock [31], follows from (2):

$$\mathcal{L}_{V(x)^{\mathsf{FB}}}(s) = \frac{(1-\rho(x))(s+\lambda-\lambda\mathcal{L}_{L_x}(s))}{s} e^{-x(s+\lambda-\lambda\mathcal{L}_{L_x}(s))}. \tag{6}$$

From (6) it is possible to calculate the moments of $V(x)^{\mathsf{FB}}$. For example, the following expression for the variance of $V(x)$, which also appears in [31]:

$$Var[V(x)]^{\mathsf{FB}} = \frac{\lambda x m_2(x)}{(1-\rho(x))^3} + \frac{\lambda m_3(x)}{3(1-\rho(x))^3} + \frac{3}{4}\left(\frac{\lambda m_2(x)}{(1-\rho(x))^2}\right)^2. \tag{7}$$

Although expressions for the moments often appear complex, they can typically be decomposed into pieces that have a natural interpretation. For example, the first term in (7) is $Var[L_x(x)]$, and the other two terms combine to form $Var[L_x(W_x^{\mathsf{FB}})]$. Many of the comments we made about $EV(x)^{\mathsf{FB}}$ also apply to $Var[V(x)]^{\mathsf{FB}}$ and other higher moments. For instance, it follows from (7) that $\lim_{x\to 0} Var[V(x)]^{\mathsf{FB}} = O(x^3)$. Hence, the response time of very small jobs nearly matches their service time, i.e., it is almost as if they are served in isolation. In addition, $Var[V(x)]^{\mathsf{FB}}$ only depends on the service distribution up to x, since $m_i(x)$ is independent of the shape distribution beyond x.

Using the Laplace transform of $V(x)^{\mathsf{FB}}$, it is easy to write the Laplace transform of the overall response time of FB by integration. However, as in the case of EV^{FB}, understanding the distributional behavior of V^{FB} in this way is not straightforward. Only very recently have distributional results for V^{FB} begun to appear.

One of the fundamental questions to answer about the distribution of V^{FB} is what requirements on the service distribution guarantee finiteness of the moments of V^{FB}. Nuyens [40] recently used differentiation of the transform in (6) to prove that

$$E[V(x)^n]^{\mathsf{FB}} = \frac{x^n}{(1-\rho(x))^n} + \begin{cases} O(x^{n-1}) & \text{if } \exists\, \alpha \geq 2 : EX^\alpha < \infty, \\ o(x^{n+1-\alpha}) & \text{if } \exists\, 1 < \alpha < 2 : EX^\alpha < \infty \end{cases} \tag{8}$$

Since $V(x) \geq x$, using (8) yields the following result.

Theorem 6.1 *In the M/GI/1* FB *queue, for any $n \in \mathbb{N}$, we have $E[V^n]^{\mathsf{FB}} < \infty$ if and only if $EX^n < \infty$.*

This result is quite appealing since these moment conditions are the weakest possible. We believe that the result holds for all $n \in [1, \infty)$, but the current proof of Theorem 6.1 depends on the differentiation of transforms, and can therefore not be generalized to include all $n \in [1, \infty)$.

Furthermore, Theorem 6.1 illustrates the benefits of using FB over non-preemptive policies, for which $E[V^n]^{\mathsf{FB}} < \infty$ if and only if $EX^{n+1} < \infty$. Note though that FB is far from unique in this behavior. As noted by Yashkov in [71], PS also requires only $EX^n < \infty$ in order to have $E[V^n]^{\mathsf{PS}} < \infty$, and Nuyens et al. [42] have proven that $V(x)^{\mathsf{SRPT}} \leq_{st} V(x)^{\mathsf{FB}}$ and thus SRPT also requires only $EX^n < \infty$ in order to have $E[V^n]^{\mathsf{SRPT}} < \infty$.

Theorem 6.1 provides a first indication of the similarities between the tail behavior of the service distribution and the tail behavior of the response time under FB. In fact, we will see that under heavy-tailed service distributions this parallel is even stronger. To illustrate this, let us consider the class of *regularly varying* service distributions. This class is a generalization of Pareto distributions and thus includes a number of practical distributions.

Definition 6.2 *We say that L is a slowly varying function if $\lim_{x\to\infty} L(yx)/L(x) = 1$ for all $y > 1$. We say that $\bar{F}$ is regularly varying with index α when $\bar{F}(x) = L(x)x^{-\alpha}$, where $L(x)$ is a slowly varying function.*

When the service distribution is regularly varying, Núñez-Queija and others showed in [12, 38, 39] that the tail of the service distribution and the tail of the response time are asymptotically equivalent in the M/GI/1 FB queue. Very recently, this result was generalized to the GI/GI/1 FB queue by Nuyens et al. [42]:

Theorem 6.3 *In a GI/GI/1 queue where $\bar{F}$ is regularly varying*[2]*,*

$$P(V^{\mathsf{FB}} > x) \sim P(X > (1-\rho)x), \text{ as } x \to \infty. \tag{9}$$

Theorem 6.3 can be interpreted as stating that whenever a job experiences a long response time, the most likely cause of the long response time is that the job itself is very large. Similar results have been proven for a number of other policies as well. For example, (9) has also been shown to hold for SRPT [39, 42] and PS [26, 75].

The proof of Theorem 6.3 uses at its core a technique developed by Guillemin et al. [26] for the analysis of the tail behavior of PS. In particular, [26] introduces two conditions on $V(x)$ that are sufficient for showing that (9) holds:

(i) $V(x)/x \to 1/(1-\rho)$ in probability as $x \to \infty$[3], and

(ii) there exists some k such that $P(V(x) > kx) = o(\bar{F}(x))$.

Thus, the two conditions together show that the conditional response time is tightly concentrated around $x/(1-\rho)$ as $x \to \infty$, which in the heavy-tailed setting is enough to guarantee that (9) holds. The main difficulty in applying these conditions to FB is in showing that the second condition holds. To accomplish this, Nuyens et al. [42] apply a probabilistic approach using an explicit random walk representation of the workload made up of jobs having service distribution $X \wedge y$, seen at the arrival of a tagged job with size y.

[2]This result has actually been shown to hold for the class of distributions that are of *intermediate regular variation* at infinity, which is slightly more general. See [38] for a discussion.

[3]Guillemin et al. [26] actually required almost sure convergence, but this was weakened by Borst et al. [13] to require only convergence in probability.

Relation (9) is a very appealing property for computer systems, because it means large job sizes cannot cause the system to provide poor overall response times: the effect of a large job is only felt by other large jobs. This is in contrast to the behavior of FCFS, under which the arrival of a large job causes subsequent arrivals of all sizes to experience large response times. In fact, under FCFS and all non-preemptive policies, the most likely reason for a job to experience a large response time is to find at the server, upon arrival, a job with a large remaining size. The result is that FCFS and all other non-preemptive policies have a response-time tail that is 'one degree (or order) heavier' than the tail of the service distribution, i.e., it is of the order $xP(X > x)$, see Borst et al. [12] for a survey.

Not only is the response-time tail of FB lighter than that of FCFS, the tail of FB seems to be "asymptotically optimal": there seem to be no policies with a smaller tail of the response-time distribution than the one described in (9). In any case, this behavior is asymptotically near optimal, in the sense that no policy can have a response time tail more than a multiplicative constant smaller. This is quite encouraging since the service distributions in many computer applications are modeled using Pareto distributions.

However, the story is not completely rosy. As we have seen throughout this survey, while FB performs well under heavy-tailed service distributions, it can perform quite badly under light-tailed service distributions. In particular, Mandjes and Nuyens [34] have shown that in an M/GI/1 queue with a light-tailed service distribution, the response-time tail of FB is asymptotically equivalent, on a logarithmic scale, to that of the busy period, which can be far from optimal. This result was later generalized to the GI/GI/1 queue by Nuyens et al. [42]:

Theorem 6.4 *In a GI/GI/1 queue with a light tailed service distribution, i.e., $E[e^{sX}] < \infty$ for some $s > 0$, we have*

$$\log P(V^{\mathsf{FB}} > x) \sim \log P(L > x), \text{ as } x \to \infty, \tag{10}$$

where L is an independent random variable with the same distribution as a busy period. Further, the logarithmic decay rate is

$$\lim_{x\to\infty} \frac{1}{x} \log P(V^{\mathsf{FB}} > x) = -\sup_{s\geq 0} \left\{ s + \Phi_A^{-1}\left(\frac{1}{\Phi_X(s)}\right)\right\}, \tag{11}$$

where Φ_X and Φ_A are the generating functions of X and the generic interarrival time A, and Φ_A^{-1} is the inverse of Φ_A.

The proof of Theorem 6.4 has two parts. First, it is shown that in the light-tailed setting, the response time of a job cannot have a heavier tail than that of the busy period, L. Then, in order to prove a lower bound, the proof identifies an event A_y with $P(A_y) > 0$ that causes the response time of a job to be larger than L_y, the busy period in a queue with service times distributed like $X \wedge y$.

After taking logarithms, dividing by x, and taking the limit $x \to \infty$, the effect of $P(A_y)$ disappears and the proof is then completed by showing that the distribution of L_y converges, roughly speaking, to that of L as $y \to x_U$.

Though the logarithmic decay rate in Theorem 6.4 appears complicated, it can be illustrative in special cases. In particular, for the M/GI/1 queue, the right-hand side of (11) becomes $\sup_{s \geq 0}\{s + \lambda - \lambda E[e^{sX}]\}$, and for the M/M/1 queue with mean service time $1/\mu$, it further simplifies to $(\sqrt{\mu} - \sqrt{\lambda})^2$.

Theorem 6.4 can be interpreted as saying that if a job experiences a large response time, it is likely the result of arriving during a very long busy period. Again this behavior is similar to that of a number of other policies: under additional conditions, (10) holds for both SRPT [43] and PS [35], though in both cases there exist some light-tailed distributions under which the response-time distribution has a lighter tail than that of FB. For (10) to hold under SRPT there must not be any probability mass in x_U, the right endpoint of the service distribution [43]: if there is mass in x_U, the response-time tail of SRPT is lighter than that of FB. For (10) to hold under PS, the service distribution must satisfy $\frac{1}{x} \log P(X > c \log x) \to 0$ for all $c > 0$ as $x \to \infty$, which eliminates distributions with bounded support or very light tails, e.g., the deterministic distribution. For the M/D/1 queue, it has recently been shown that the response-time tail of PS is lighter than that of FB [21].

As in the case of heavy-tailed service distributions, (10) is in contrast with the behavior of FCFS, under which a large response time is likely due to seeing a large workload in the queue upon arrival. Interestingly, in the light-tailed setting, under work-conserving disciplines, the tail of the workload is much lighter than the tail of the busy period. It should be noted though that the poor behavior of FB with respect to the response-time tail is not necessarily indicative of other performance measures. For example, some gamma distributions are both light-tailed and DFR. So, for those service distributions, the tail of the response time is as heavy as possible (Theorem 6.4), but the queue-length distribution and mean response time are optimal among blind policies (Theorem 4.1). Additionally, it is important to point out that the poor behavior of the response-time tail under FB in the case of light-tailed service distributions is merely caused by the behavior of the largest jobs. In particular, if we look at the distribution of the response time experienced by jobs of size x, Nuyens et al. [42] have proven the following:

Theorem 6.5 *In a GI/GI/1 queue with a light tailed service distribution, i.e., $E[e^{sX}] < \infty$ for some $s > 0$, for all x,*

$$\log P(V(x)^{\mathsf{FB}} > y) \sim \log P(L_x > y) \quad \text{as } y \to \infty,$$

where L_x is the length of a busy period in a queue with generic service time $X \wedge x$.

Theorem 6.5 implies that for many job sizes, the tail of the conditional response time is lighter

than under FCFS. In fact, calculations similar to those performed in Nuyens and Zwart [43] show that for the M/M/1 queue, regardless of the load, at least 63% of the jobs prefer FB over FCFS as far as the response time tail is concerned. In addition, this percentage increases to 100% as $\rho \to 1$.

Let us conclude this section by pointing out that Theorems 6.3 and 6.4 illustrate a trade-off that seems to be a general tendency: policies that have near optimal response-time tail behavior under heavy-tailed service distributions, behave poorly under light-tailed service distributions. In particular, it seems unlikely that any policy can obtain the "best of both worlds." The reason for this tradeoff is simple. Policies that behave well in the heavy-tailed setting, have a response-time tail that is asymptotically equivalent to the tail of the service distribution, which in this setting is in turn asymptotically equivalent to the length of a busy period. In particular, De Meyer and Teugels [19] have shown that (9) holds for the length of a busy period as well. Thus, if a policy behaves well in the heavy-tailed setting, it has a response-time tail asymptotically equivalent to the length of a busy period. However, if the response time is asymptotically equivalent to a busy period in the light-tailed setting as well, it is far from optimal there.

6.2 The queue-length distribution

Understanding the queue-length distribution of FB is key when considering applying the policy in applications such as routers, where capacity planning is fundamental.

Though the mean response time and mean queue length are related by Little's Law, distributional forms of Little's Law do not apply because FB allows later arrivals to overtake earlier arrivals in the queue. And, in fact, the analysis of the queue-length distribution of FB has proceeded much more slowly than the analysis of the response-time distribution. The first derivation of the generating function of the queue-length distribution did not occur until nearly 15 years after the derivation of the Laplace transform of response time. Pechinkin [44] was the first to derive the generating function of Q^{FB}:

Theorem 6.6 *Let Q^{FB} be the number of jobs in the stationary M/GI/1 queue. Then for $z < 1$,*

$$Ez^{Q^{\mathsf{FB}}} = (1-\rho)\exp\Big(-\int_0^\infty z\frac{\partial v(t,z)}{\partial z}dt\Big), \tag{12}$$

where $v(t,z)$ is the unique non-negative root of the equation

$$v(t,z) = \lambda\big(1-\int_0^t e^{-v(t,z)x}dF(x) - z(1-F(t))e^{-v(t,z)t}\big). \tag{13}$$

Yashkov [70] obtained the counterpart of (12) for the case of batch arrivals. From the proof of Theorem 6.6 it follows that $v(t,1) = 0$. This allows for computing the moments of Q by differentiating (12).

Using Theorem 6.6, the primary question of interest is to determine the requirements for the finiteness of moments of Q^{FB}. Through a detailed analysis of the derivative of $v(t,z)$, Nuyens [40] proved the following:

Theorem 6.7 *In the M/GI/1* FB *queue, if $EX^{\alpha} < \infty$ for some $\alpha > 1$, then all moments of Q^{FB} are finite.*

There is a strong contrast between these moment conditions and those of FCFS: FCFS requires $EX^{n+1} < \infty$ in order for EQ^n to be finite. Further, FB almost matches the behavior of SRPT and PS: under SRPTand PS all moments of Q are finite if $EX < \infty$. In fact, an interesting question that is left open at this point is to determine whether $EX < \infty$ is also a strong enough condition to guarantee that all moments of Q^{FB} are finite.

Beyond the transform and moment conditions of Q^{FB}, there are only a few scattered results in the literature. These results focus on two practical questions: (i) what happens to Q^{FB} in heavy traffic and (ii) what is the distribution of the maximal queue length in a busy period? These two questions are fundamental issues when addressing the task of capacity planning in computer networks.

In answer to the first question, Nagorenko and Pechinkin [37] prove the following asymptotic characterization of the distribution of Q^{FB} under heavy traffic, which can provide a useful rule-of-thumb for buffer management.

Theorem 6.8 *For service distributions with tail $\bar{F}(x) \sim ax^b e^{-cx}$, for some $a > 0, b \geq 0$ and $c > 0$, the stationary queue length Q^{FB} in the M/GI/1* FB *queue satisfies*

$$\lim_{\rho \uparrow 1} P(Q^{\mathsf{FB}}/EQ^{\mathsf{FB}} < x) = 1 - e^{-x}, \qquad x \geq 0.$$

In addition, for the M/D/1 FB queue, an expression for the Laplace transform of $\lim_{\rho\uparrow 1} Q^{\mathsf{FB}}/EQ^{\mathsf{FB}}$ has been derived by Yashkov and Yashkova in [73].

A partial answer to the second question can be found in Borel [11]. *Avant la lettre,* he studied the M/D/1 FB queue, and derived the distribution of the maximal queue length during a busy period, denoted by Q^{FB}_{max}:

Theorem 6.9 *In the M/D/1* FB *queue with arrival rate λ and service times equal to 1,*

$$P(Q^{\mathsf{FB}}_{max} = n) = \lambda^{n-1} e^{-\lambda n} \frac{n^{n-1}}{n!} \sim e^{n(\log\lambda + 1 - \lambda)}/(n\sqrt{n}\lambda\sqrt{2\pi}), \quad \text{as } n \to \infty.$$

Beyond the M/D/1 setting, such asymptotics have not been derived. However, Nuyens [41] proved the following bound on Q^{FB}_{max} under log-convex service distributions.

Theorem 6.10 *In an M/GI/1 queue where the service distribution has a log-convex density,*

$$P(Q_{max}^{\mathsf{FB}} > n) \leq \rho^n, \qquad n = 0, 1, \ldots. \tag{14}$$

Amazingly, this result indicates that for log-convex service distributions, the bound on the distribution of the maximal queue length under FB is not much larger than the stationary queue length distribution of PS. Recall that $P(Q^{\mathsf{PS}} = n) = (1-\rho)\rho^n$. Further, the bound on the distribution of Q_{max}^{FB} is insensitive to the form of the service distribution beyond its mean.

Theorem 6.10 is proved by first calculating the maximal queue length of FB*, an artificial discipline very similar to FB: under FB*, the first customer in a busy period has lowest priority, and all other customers are treated according to FB. The busy period of FB* can be decomposed in a random number of sub-busy periods that behave like FB busy periods, and so the maximal queue length of FB* can be expressed in terms of the FB maximum. This similarity, in combination with the optimality of FB (Theorem 4.2), is enough to find the bound in (14).

Using the regenerative structure of the queue length process, it is possible to relate the maximal queue length during a busy period to the maximal queue length over the time interval $[0, t]$ for $t \to \infty$, see the survey article of Asmussen [6]. In particular, define $Q_{max}^{\mathsf{FB}}(t)$ as the maximal queue length over the time interval $[0, t]$ under FB. Then Nuyens [41] used Theorem 6.10 to prove the following:

Theorem 6.11 *Consider an M/GI/1* FB *where service times have a log-convex density. Then for any $x > 0$, the inequality*

$$P(Q_{max}^{\mathsf{FB}}(t) > a \log t + b + x) \leq \rho^x$$

holds for t large enough and $a = -1/(\log \rho)$, $b = -(\log \lambda + \log(1-\rho))/(\log \rho) + 1$.

The key observation about Theorem 6.11 is that it indicates that the maximal queue length under FB grows logarithmically in time, with rate at most $a = -1/\log \rho$. If we consider the case of heavy traffic ($\rho \to 1$), this growth rate satisfies $-1/\log \rho \sim 1/(1-\rho)$. Comparing this growth rate with that of other policies is difficult because little is known about the maximal queue length under other disciplines, but we can provide some comparison with the behavior of FCFS.

Cohen [18] has described the asymptotic behavior of the maximal queue length in the first n busy periods in the FCFS queue for light-tailed distributions, i.e., under the condition that $E[e^{sX}] < \infty$ for some $s > 0$. Combining the analysis in [18] with the technique from Asmussen [6] we described above, it can be seen that the maximal queue length under FCFS in the time interval $[0, t]$ grows logarithmically in t with rate $1/\log(1+\theta^*)$, where θ^* is the positive solution of the equation $\theta + 1 = E[e^{\theta X}]$. Note that θ^* can be interpreted as the decay rate of the busy-period distribution, see, e.g., Mandjes and Zwart [35]. In heavy traffic, $\theta^* \to 0$, so the logarithmic growth rate behaves like $1/\theta^*$, and $\theta^*(\rho) \sim 2(1-\rho)/(1+C^2[X])$ for $\rho \to 1$, see [35]. Unfortunately, due to

the conditions on the service distributions, we can only compare the asymptotic growth rates under FB and FCFS for light-tailed log-convex densities, for example the gamma distributions mentioned before Theorem 4.2. For those distributions, we can conclude that in heavy traffic the growth rate of the maximal queue length under FB is smaller than under FCFS, since the coefficient of variation satisfies $C^2[X] > 1$ for all log-convex densities.

7 The performance of large jobs under FB

To this point, we have concerned ourselves primarily with traditional queueing metrics such as measures of the queue-length distribution and response-time distribution. With respect to these measures, we have repeatedly seen that FB performs well when the service distribution is heavy-tailed, but that it can behave very poorly if the service distribution is light-tailed. Since many computer applications have service distributions that are typically modeled as heavy-tailed distributions, these results suggest that FB is quite applicable in practical applications. However, there is a key worry that has traditionally kept FB from being used in practice: how bad is the performance of large jobs? Phrased differently, the worry is that large job sizes are treated "unfairly."

Addressing this issue is a difficult task because of the amorphous nature of "fairness," and this difficulty is likely the reason that the fairness of FB went unstudied for so long. Only very recently did Harchol-Balter et al. [28] provide a first analysis of the fairness of FB. This was followed shortly by a number of more detailed analyses, e.g., Wierman and Harchol-Balter [65] and Rai et al. [47], which were motivated by the need to address fairness concerns in the context of web servers and routers.

The notion of fairness that emerged in these papers is derived intuitively from a Aristotle's notion of fairness, which is based on the idea that: *like cases should be treated alike, different cases should be treated differently, and different cases should be treated differently in proportion to the difference at stake* [51]. In the context of scheduling queues, this matches the common intuition that small jobs should have small response times, large jobs should have large response times, and the differences in response times of small and large jobs should be proportional to the differences the job sizes. Specifically, the response time for a job of size x, $V(x)$, should be proportional to x.

The first analysis of the fairness of FB by Harchol-Balter et al. [28] focused only on the experience of the largest job sizes, motivated by the intuition that the largest jobs will be treated the most unfairly under FB. Surprisingly, this analysis showed that in the M/GI/1 setting these largest jobs are treated no worse under FB than under PS, which is intuitively the most fair policy, since all jobs in the system share the server evenly at all times. The results in [28] were later generalized by Nuyens et al. [42], who proved that:

Theorem 7.1 *In a GI/GI/1 queue,*

$$\lim_{x\to\infty} \frac{V(x)^{\mathsf{FB}}}{x} = \lim_{x\to\infty} \frac{V(x)^{\mathsf{PS}}}{x} = \frac{1}{1-\rho} \quad a.s.$$

This result may be interpreted as follows: during the response time of an exceptionally large job, the service rate it gets is the total service rate (namely 1) reduced by the load of jobs that pass through the system in the meantime, which is ρ in the limit, since the system remains stable. Wierman and Harchol-Balter [66] have also shown that this limit is achieved by almost all common preemptive policies. Thus, it seems one need not worry about the largest jobs being treated unfairly since most common preemptive policies treat them asymptotically equivalently.

Surprisingly though, it is not always the largest job sizes that receive the most unfair treatment under FB. The following theorem summarizes the results of Wierman and Harchol-Balter [65], Rai et al. [47], and Brown [14].

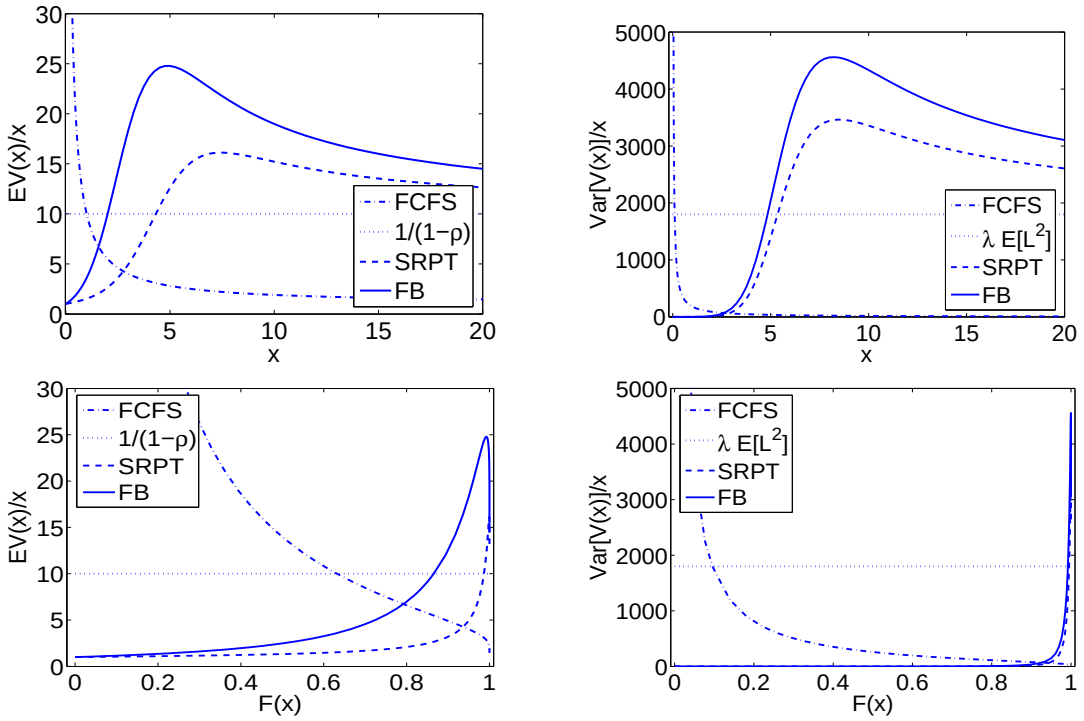
Theorem 7.2 *For all $0 < \rho < 1$ and all continuous service distributions with $EX^2 < \infty$, $EV(x)^{\mathsf{FB}}/x$ is not monotonically increasing. Further, $EV(x)^{\mathsf{FB}}/x$ converges from above to $1/(1-\rho)$ as $x \to \infty$. However, if F is regularly varying with index $\alpha \in (1, 3/2)$, then $EV(x)^{\mathsf{FB}}/x$ is monotonically increasing.*

So, the largest job sizes are not treated worse under FB than under PS, but some range of large, but not the largest, job sizes *may* receive unfairly long response times. Figure 4 illustrates this "hump" behavior, which is similar to the behavior of SRPT, but in stark contrast to that of PS. Wierman and Harchol-Balter [65] termed this behavior of $V(x)$ "Sometimes Unfair", since under some service distributions with $EX^2 < \infty$, there exists some x for which $EV(x)^{\mathsf{FB}}/x > 1/(1-\rho)$, but under other service distributions, $EV(x)^{\mathsf{FB}}/x \leq 1/(1-\rho)$ for all x.[4] The value $1/(1-\rho)$ is a natural criterion for fairness because (i) $\min_{\mathsf{P}} \max_x EV(x)^{\mathsf{P}}/x = 1/(1-\rho)$ [65] and (ii) the intuitively-fair PS has $EV(x)^{\mathsf{PS}}/x = 1/(1-\rho)$. Notice that this criterion for fairness is very useful from a practical perspective as well since PS is typically the default scheduling policy used in computer systems where FB is being suggested as an alternative. For more details on this metric and criterion, see the recent survey of Wierman [68].

It is important to note that FB is far from alone in being classified as Sometimes Unfair. In fact, among blind scheduling policies, only policies for which $EV(x) = x/(1-\rho)$ for all x, avoid being at least Sometimes Unfair. To see why this is true, recall Kleinrock's conservation law for M/G/1 queues:

$$\lambda \int_0^\infty EV(x)\bar{F}(x)dx = \frac{\lambda E[X^2]}{2(1-\rho)}. \tag{15}$$

[4]Note that in [65], Wierman and Harchol-Balter say that FB is "Always Unfair", but that is because they were considering only service distributions with $EX^2 < \infty$. Later, results of Brown [14] for the case when $EX^2 = \infty$ changed the classification of FB.



(a) $EV(x)/x$ (b) $Var[V(x)]/x$

Figure 4: An illustration of the behavior of $EV(x)^{\mathsf{FB}}$ and $Var[V(x)]^{\mathsf{FB}}$. In all cases the service distribution is exponential and the load is 0.9. The top row shows the behavior as a function of x while the bottom row shows the behavior as a function of $F(x)$, i.e., the percentile of x.

A policy P only avoids being at least Sometimes Unfair if $EV(x)^{\mathsf{P}} \leq x/(1-\rho)$ holds for each service distribution and all x. Using this bound in the conservation law, we see that

$$\lambda \int_0^\infty EV(x)^{\mathsf{P}} \bar{F}(x)dx \leq \frac{\lambda}{1-\rho} \int_0^\infty x\bar{F}(x)dx = \frac{\lambda E[X^2]}{2(1-\rho)}. \tag{16}$$

Therefore, the inequality must be an equality everywhere, so that $EV(x)^{\mathsf{P}} = x/(1-\rho)$ for all x. The only policies known to satisfy this condition are the so-called symmetric policies of Kelly [30], such as PS and PLCFS.

Given that FB always treats some range of job sizes unfairly, it becomes important to characterize (i) what percentage of jobs are treated unfairly, (ii) which job sizes are treated unfairly, and (iii) how unfairly are these jobs treated. These questions are addressed Wierman and Harchol-Balter [65], Rai et al. [47], and Brown [14] . We summarize their results in the following theorem:

Theorem 7.3 *In an M/GI/1 queue, for all x,*

$$EV(x)^{\mathsf{FB}} \leq \left(\frac{1-\rho/2}{1-\rho}\right) EV(x)^{\mathsf{PS}}.$$

Further, $EV(x)^{\mathsf{FB}} \leq EV(x)^{\mathsf{PS}}$ for all x such that $\rho(x) \leq \max(2\rho/3, \rho/(1+\sqrt{1-\rho}))$.

Theorem 7.3 shows that even when a job size is treated unfairly, $EV(x)^{\mathsf{FB}}$ is not too much larger than $EV(x)^{\mathsf{PS}}$, unless ρ is close to 1. However, $\rho/(1+\sqrt{1-\rho}) \to 1$ as $\rho \to 1$, so in heavy traffic, very few jobs are treated unfairly under FB. Furthermore, we see that not too large a fraction of jobs is treated worse under FB than under PS. This fraction is especially small under heavy-tailed distributions, where a larger percentage of the load is made up by a small percentage of large jobs. For example, if the load is 0.9, then any x such that $\rho(x) < 0.68$ will have $EV(x)^{\mathsf{FB}} \leq EV(x)^{\mathsf{PS}}$. For the exponential distribution, this condition is satisfied by 86% of arrivals. Under heavy-tailed distributions, this condition may be satisfied by more than 95% of arrivals.

Finally, if the degree of unfairness of FB is judged to be too large, one might turn to hybrid policies, investigated by Rai et al. [48], that combine aspects of FCFS and FB in order to shrink the worst case of $V(x)/x$.

So far we have focused on fairness with respect to mean response times, but worries about unfairness are not limited to the mean. In fact, worries about large job sizes being starved of service also lead to the suggestion that response times of large jobs are unfairly variable, in addition to being unfairly long. To address such worries, Wierman and Harchol-Balter studied the behavior of higher moments of $V(x)^{\mathsf{FB}}$ in [66]. Interestingly, the non-monotonic behavior of $EV(x)^{\mathsf{FB}}/x$ seems to extend to higher moments as well. In particular, Wierman and Harchol-Balter [66] have proven that the behavior of $Var[V(x)]^{\mathsf{FB}}/x$ parallels that of $EV(x)^{\mathsf{FB}}/x$ when $\lambda E[L^2]$ is used in place of $1/(1-\rho)$, see Figure 4. Recall that L is the length of a busy period.

Further, experimental evidence suggests that this non-monotonic behavior extends beyond $EV(x)^{\mathsf{FB}}/x$ and $Var[V(x)]^{\mathsf{FB}}/x$. Wierman and Harchol-Balter [66] conjecture that for any $n \in \mathbb{N}$, the so-called nth normalized *cumulant moment* of $V(x)^{\mathsf{FB}}$ will also show non-monotonic behavior when $I_{[n=1]} + \lambda E[L^n]$ is used in place of $1/(1-\rho)$. For more information on this conjecture, we refer to [66, 68].

8 Future research topics

From this survey, it is clear that we have come a long way towards characterizing the performance of FB in single server queues. However, it should also be clear that many interesting questions remain.

For example, much progress has been made towards characterizing under which service distributions FB is appropriate. We have seen that for DFR service distributions, FB optimizes the queue-length distribution among all blind policies, and that for IMRL distributions FB still has a smaller mean response time than PS. But, we have also seen that IMRL is not a strong enough condition for FB to minimize the mean response time. It would therefore be interesting to see if there is a class larger than DFR, but not containing all IMRL distributions, for which FB does minimize the mean response time.

Another open problem is for which service distributions the mean response time achieves the theoretical upper and lower bounds. So far, it has only been shown that the lower bound is achieved for regularly varying service distributions (in the limit as the system load approaches 1), and that the upper bound is tight for deterministic distributions.

Further, we have seen that in many cases FB can perform nearly as well as SRPT, which optimizes the queue-length distribution among all scheduling policies. For instance, for regularly varying service distributions, the mean response time under FB is within a constant of SRPT for all loads, and the tail behavior of the response time is asymptotically equivalent to that of SRPT. However, is it necessary for the service distribution to be regularly varying for FB to match the performance of SRPT?

On the other hand, we have seen that FB can also perform far from optimal. For instance, for deterministic service distributions, FB has mean response times that are as large as possible under any work conserving policy, and for light-tailed service distributions, FB has a tail behavior that matches the heaviest tail possible under work conserving policies. However, we have also seen that FB can perform badly for distributions with high variability; thus, an important task that remains is to better characterize under which classes of service distributions FB performs badly.

In addition to characterizing under which service distributions FB is appropriate, another interesting set of open questions is related to the behavior of FB in heavy-traffic. As we saw in Section 5.2, the behavior of FB as $\rho \to 1$ is much more complicated than that of PS and FCFS. Consequently, results to this point have focused on the behavior of EV^{FB} and EQ^{FB} as $\rho \to 1$, rather than on the distributional behavior of V^{FB} and Q^{FB}. Obtaining such distributional results is an important task, both practically and theoretically.

Another set of open problems about FB is related to the topic of fairness. We have seen that under FB there is a range of large jobs, but not the largest jobs, that has unfairly large values of $EV(x)^{\mathsf{FB}}$ and $Var[V(x)]^{\mathsf{FB}}$. However, there are a number of interesting questions that remain. For instance, do higher cumulant moments of $V(x)^{\mathsf{FB}}$ mimic the behavior of $EV(x)^{\mathsf{FB}}/x$ and $Var[V(x)]^{\mathsf{FB}}$? And more broadly, the characterization of fairness in terms of $V(x)$ is only one of many notions of fairness, and the question of how FB performs under other fairness definitions, e.g., those in [7, 49, 50, 54, 55], has only begun to be addressed. A very practical open problem is the task of adjusting FB so as to improve its fairness while not increasing response times and queue lengths too much. Recent papers by Feng et al. [24] and Rai et al. [48] have begun to address this issue, but many issues remain.

It is also interesting to consider the behavior of FB beyond the single server setting. With the growing trend towards server-farm architectures in computer systems, this is another important direction for future work. However, to this point there have been only a few proposals, e.g., Wu and Down [69]. Finally, an important future research direction is understanding the behavior of FB and other priority-based disciplines in networks of queues. Such an understanding is key to the

adoption of FB in applications such as routers. However, this will not be straightforward: recent work of Verloop et al. [63] shows that global use of FB in a resource sharing network can render the system unnecessarily unstable. But, limiting the use of FB to the edges of networks may avoid such issues.

Acknowledgments We would like to thank all four referees for their suggestions and comments. These have clearly improved the presentation and readability of the paper. The core of this paper was composed during a visit of Wierman to the EURANDOM institute in the Netherlands, supported by an STW grant.

References

[1] S. Aalto and U. Ayesta. Mean delay analysis of multi level processor sharing disciplines. In *Proceedings of Infocom, Barcelona, Spain,* pages 23–29, 2006.

[2] S. Aalto and U. Ayesta. On the non-optimality of the foreground-background discipline for imrl service times. *Journal of applied probability*, 43(2):523–534, 2006.

[3] S. Aalto, U. Ayesta, and E. Nyberg-Oksanen. Two-level processor-sharing scheduling disciplines: mean delay analysis. In *Proceedings ACM Sigmetrics, New York, USA,* pages 97–105, 2004.

[4] S. Aalto, U. Ayesta, and E. Nyberg-Oksanen. M/G/1 MLPS compared to M/G/1 PS. *Operations research letters*, 33:519–524, 2004.

[5] S. Aalto and U. Ayesta. Mean delay optimization for the M/G/1 queue with Pareto type service times. *Extended abstract in ACM Sigmetrics, San Diego, USA,* pages 383–384, 2007.

[6] S. Asmussen. Extreme values for queues via cycle maxima. *Extremes*, 1(2):137–168, 1998.

[7] B. Avi-Itzhak and H. Levy. On measuring fairness in queues. *Advances in applied probability*, 36(3):919–936, 2004.

[8] A. Balkema and S. Verwijmeren. Stability of an M/G/1 queue with thick tails and excess capacity. *Proceedings of the 19th Seminar on Stability Problems for Stochastic Models, Part II (Vologda, 1998), Journal of Mathematical Science*, 99(4):1386–1392, 2000.

[9] N. Bansal and D. Gamarnik. Handling load with less stress. Queueing systems, 54(1):45-54, 2006.

[10] N. Bansal and A. Wierman. Competitive analysis of M/GI/1 queueing policies. Technical Report CMU-CS-02-201, Carnegie Mellon University, 2002.

[11] E. Borel. Sur l'emploi du théorème de Bernoulli pour faciliter le calcul d'une infinité de coefficients. Application au problème de l'attente à un guichet. *Comptes Rendus de l'Académie des Sciences, Paris*, 214:452–456, 1942.

[12] S. Borst, O. Boxma, R. Núñez Queija, and A. Zwart. The impact of the service discipline on delay asymptotics. *Performance Evaluation*, 54:175–206, 2003.

[13] S. Borst, R. Núñez Queija, and A. Zwart. Sojourn time asymptotics in Processor Sharing queues. *Queueing Systems*, 53:31–51, 2006.

[14] P. Brown. Comparing FB and PS scheduling policies, In *ACM Sigmetrics Performance Evaluation Review*, 34(3):18–22, 2006.

[15] E. Coffman and P. Denning. *Operating systems theory.* Prentice-Hall, 1973.

[16] E. Coffman and L. Kleinrock. Feedback queueing models for time-shared systems. *Journal of the Association for Computing Machinery*, 15:549–576, 1968.

[17] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. In *Proceedings ACM Sigmetrics, Philadelphia, USA*, pages 160–169, 1996.

[18] J. Cohen. *The single server queue.* North-Holland, 1982.

[19] A. De Meyer and J. Teugels. On the asymptotic behaviour of the distributions of the busy period and service time in M/G/1. *Journal of Applied Probability*, 17(3):802–813, 1980.

[20] B. Dellart. How tolerable is delay? consumers evaluations of internet web sites after waiting. *Journal of Interactive Marketing*, 13:41–54, 1999.

[21] R. Egorova, B. Zwart, and O. Boxma. Sojourn time tails in the M/D/1 processor sharing queue. *Probability in the Engineering and Informational Sciences,* 20(3):429-446, 2006.

[22] H. Feng and V. Misra. Mixed scheduling disciplines for network flows. *ACM Sigmetrics Performance Evaluation Review*, 31(2):36–39, 2003.

[23] H. Feng and V. Misra. On the relationship between coefficient of variation and the performance of M/G/1-FB queues. *ACM Sigmetrics Performance Evaluation Review*, 32(2):17–19, 2004.

[24] H. Feng, V. Misra, and D. Rubenstein. PBS: A Unified Priority-Based CPU Scheduler. In *Proceedings of ACM Sigmetrics, San Diego, USA*, pages 203–214, 2007.

[25] J. Gittins. *Multi-armed bandit allocation indices.* Wiley, Chichester, 1989.

[26] F. Guillemin, P. Robert, and A. Zwart. Tail asymptotics for processor sharing queues. *Advances in Applied Probability*, 36(2):525–543, 2004.

[27] M. Harchol-Balter, B. Schroeder, N. Bansal, and M. Agrawal. Size-based scheduling to improve web performance. *ACM Transactions on Computer Systems (TOCS)*, pages 207–233, 2003.

[28] M. Harchol-Balter, K. Sigman, and A. Wierman. Asymptotic convergence of scheduling policies with respect to slowdown. *Performance Evaluation*, 49:241–256, 2002.

[29] A. Jean-Marie and P. Robert. On the transient behaviour of the processor sharing queue. *Queueing Systems*, 17:129–136, 1994.

[30] F. Kelly. *Reversibility and Stochastic Networks*. John Wiley & Sons, 1979.

[31] L. Kleinrock. *Queueing systems, Volume II: Computer Applications*. Wiley, 1976.

[32] G. Klimov. Time-sharing service systems, I. Theory of Probability and its Applications, 19:532-551, 1974.

[33] G. Klimov. Time-sharing service systems, II. Theory of Probability and its Applications, 23:314-321, 1978.

[34] M. Mandjes and M. Nuyens. Sojourn times in the M/G/1 FB queue with light-tailed service times. *Probability in the Engineering and Informational Sciences*, 19(3):351–361, 2005.

[35] M. Mandjes and A. Zwart. Large deviations for waiting times in processor sharing queues. *Queueing Systems*, 52(4):237-250, 2006.

[36] J. McKinney. A survey of analytical time-sharing models. *Computing Surveys*, 1(2):105–116, 1969.

[37] V. Nagonenko and A. Pechinkin. Heavy loading in FBPS shared-processor systems. *Izv. Akad. Nauk. SSSR, Tekh. Kibernet*, 6:62–67, 1980.

[38] R. Núñez Queija. *Processor-Sharing Models for Integrated-Services Networks*. PhD thesis, Eindhoven University, 2000.

[39] R. Núñez Queija. Queues with equally heavy sojourn time and service requirement distributions. *Annals of Operations Research*, 113:101-117, 2002.

[40] M. Nuyens. *The Foreground-Background Queue*. PhD thesis, University of Amsterdam, available from `http://www.few.vu.nl/~mnuyens/publications`, 2004.

[41] M. Nuyens. The maximum queue length for heavy-tailed service times. *Queueing Systems*, 47:107–116, 2004.

[42] M. Nuyens, A. Wierman, and A. Zwart. Preventing large sojourn times using SMART scheduling. *To appear in Operations Research.*

[43] M. Nuyens and A. Zwart. A large-deviations analysis of the GI/GI/1 SRPT queue. *Queueing Systems*, 54(2):85-97, 2006.

[44] A. Pechinkin. Stationary probabilities in a system with discipline of advantageous distribution of a process. *Engineering Cybernetics*, 18(5):52–56, 1981.

[45] I. Rai, E. Biersack, and G. Urvoy-Keller. LAS scheduling approach to avoid bandwidth hogging in heterogeneous TCP networks. In *7th IEEE International Conference on High Speed Networks and Multimedia Communications HSNMC'04, Toulouse, France*, pages 179–190, 2004.

[46] I. Rai, E. Biersack, and G. Urvoy-Keller. Size-based scheduling to improve the performance of short TCP flows. In *IEEE Network Magazine*, 19(1):12–17 2005.

[47] I. Rai, G. Urvoy-Keller, and E. Biersack. Analysis of LAS scheduling for job size distributions with high variance. In *Proceedings of ACM Sigmetrics, San Diego, USA*, pages 218–228, 2003.

[48] I. Rai, G. Urvoy-Keller, and E. Biersack. Performance models for LAS-based scheduling disciplines in a packet switched network. In *Proceedings of ACM Sigmetrics, New York, USA*, pages 106–117, 2004.

[49] D. Raz, H. Levy, and B. Avi-Itzhak. A resource-allocation queueing fairness measure. In *Proceedings of ACM Sigmetrics, New York, USA*, pages 130–141, 2004.

[50] D. Raz, H. Levy and B. Avi-Itzhak. On the twin measure and system predictability and fairness. *ACM Sigmetrics Performance Evaluation Review*, 34(3):15–17, 2006.

[51] R. Rhodes, M. P. Battin, and A. Silvers. *Medicine and social justice.* Oxford University Press, 2002.

[52] R. Righter and J. Shanthikumar. Scheduling multiclass single server queueing systems to stochastically maximize the number of succesful departures. *Probability in the Engineering and Informational Sciences*, 3:323–333, 1989.

[53] R. Righter, J. Shanthikumar, and G. Yamazaki. On extremal service disciplines in single-stage queueing systems. *Journal of Applied Probability*, 27:409–416, 1990.

[54] W. Sandmann. A discrimination frequency based queueing fairness measure with regard to job seniority and service requirement. In *Proceedings of Euro NGI Conference on Next Generation Internet Networks, Rome, Italy*, pages 106-113, 2005.

[55] W. Sandmann. Analysis of a queueing fairness measure. In *Proceedings of the 13th GI/ITG Conference on Measurement, Modelling and Evaluation of Computer and Communication Systems, Nurnberg, Germany*, pages 219–231, 2006.

[56] R. Schassberger. The steady state distribution of spent service times present in the M/G/1 foreground-background processor-sharing queue. *Journal of Applied Probability*, 25(7):194–203, 1988.

[57] L. Schrage. The queue M/G/1 with feedback to lower priority queues. *Management Science*, 18(7):466–474, 1967.

[58] L. Schrage. A proof of the optimality of the shortest remaining service time discipline. *Queueing Systems*, 16:670–690, 1968.

[59] L. Schrage and L. Miller. The M/G/1 queue with the shortest remaining processing time first discipline. *Operations Research*, 14(4):670–684, 1966.

[60] M. Shaked and J. Shanthikumar. *Stochastic orders and their applications.* Academic Press, 1994.

[61] A. Silberschatz, P.B. Galvin, and G. Gagne. *Applied Operating Systems Concepts.* John Wiley & Sons, 2000.

[62] M. Taqqu, W. Willinger, and A. Erramilli. A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks. In *Stochastic Networks: Theory and Applications*, pages 339–366. Oxford University Press, 1996.

[63] I. Verloop, S. Borst, R. Núñez Queija.Stability of size-based scheduling disciplines in resource-sharing networks, In *Performance Evaluation* 62, Special issue - Proceedings Performance 2005, 247–262, 2005.

[64] A. Wierman, N. Bansal, and M. Harchol-Balter. A note comparing response times in the M/GI/1/FB and M/GI/1/PS queues. *Operations Research Letters*, 32(1):73–76, 2004.

[65] A. Wierman and M. Harchol-Balter. Classifying scheduling policies with respect to unfairness in an M/GI/1. In *Proceedings of ACM Sigmetrics, San Diego, USA*, pages 238–249, 2003.

[66] A. Wierman and M. Harchol-Balter. Classifying scheduling policies with respect to higher moments of response time in an M/GI/1. In *Proceedings of ACM Sigmetrics, Banff, CA*, pages 229–240, 2005.

[67] A. Wierman and M. Harchol-Balter. Nearly insensitive bounds for SMART scheduling. In *Proceedings of ACM Sigmetrics, Banff, CA*, pages 205–216, 2005.

[68] A. Wierman. Fairness and classifications. In *ACM Sigmetrics Performance Evaluation Review*, 34(4): 4-12, 2007.

[69] R. Wu and D. Down. Scheduling multi-server systems using foreground-background processing. In *Proceedings of the 42nd Annual Allerton Conference on Communications, Control, and Computing, Urbana, Illinois*, 2004. **NO PAGE NUMBER**

[70] S. Yashkov. Analysis of a system with priority-based processor-sharing. *Automatic Control and Computer Sciences*, 18(3):27–36, 1984.

[71] S. Yashkov. Processor-sharing queues: some progress in analysis. *Queueing Systems*, 2:1–17, 1987.

[72] S. Yashkov. Mathematical problems in the theory of shared-processor systems. *Journal of Soviet Mathematics*, 58(2):101–147, 1992.

[73] S. Yashkov and A. Yashkova. A note on a heavy taffic limit theorem for the M/D/1-FBPS queue. *Information Processes*, 4(3):251–255, 2004.

[74] M. Zhou and L. Zhou. How does waiting duration information influence customers' reactions to waiting for services. *Journal of Applied Social Psychology*, 26:1702–1717, 1996.

[75] A. Zwart and O. Boxma. Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Systems*, 35:141–166, 2000.

A Proof of Theorem 5.4

We will start by proving the result in the case of a bounded service distribution. Let x_U be the upper bound of the service distribution and define $\bar{\rho}(x) = \lambda \int_0^x tf(t)dt$. Note that $\rho(x) \geq \bar{\rho}(x)$ and $\bar{\rho}'(x) = \lambda x f(x)$. Then after removing the first term in (1), we have for all $y \geq 0$,

$$\begin{aligned} EV^{\mathsf{FB}} &\geq \int_0^{x_U} \frac{\lambda m_2(x) f(x)}{2(1-\bar{\rho}(x))^2} dx \\ &\geq \frac{m_2(y)}{2} \int_y^{x_U} \frac{\lambda x f(x)}{(1-\bar{\rho}(x))^2} \frac{1}{x} dx \\ &\geq \frac{m_2(y)}{2x_U} \int_y^{x_U} \frac{\bar{\rho}'(x)}{(1-\bar{\rho}(x))^2} dx = \Omega\left(\frac{1}{1-\rho}\right) \text{ as } \rho \to 1. \end{aligned}$$

To prove the result in the case of an unbounded service distribution, note that $\rho'(x) = \lambda \bar{F}(x)$. Then for all $z \geq 0$,

$$\begin{aligned} EV^{\mathsf{FB}} &= \int_0^{\infty} \frac{x}{1-\rho(x)} f(x) dx + \int_0^{\infty} \frac{\lambda m_2(x) f(x)}{2(1-\rho(x))^2} dx \\ &\leq \frac{EX}{1-\rho} + \int_0^z \frac{\lambda}{2} \frac{m_2(x)}{(1-\rho(x))^2} f(x) dx + \int_z^{\infty} m_2(x)\mu(x) \frac{\rho'(x)}{2(1-\rho(x))^2} dx. \end{aligned} \tag{17}$$

Since $m_2(x)\mu(x) = O(1)$, there exists an x_0 and an N such that $m_2(x)\mu(x) \leq N$ for $x \geq x_0$. Taking $z = x_0$ in (17) yields that

$$\begin{aligned} EV^{\mathsf{FB}} &\leq \frac{EX}{1-\rho} + \frac{\lambda x_0^2 F(x_0)}{2(1-\rho(x_0))^2} + \int_{x_0}^{\infty} m_2(x)\mu(x)\frac{\rho'(x)}{2(1-\rho(x))^2}dx \\ &\leq \frac{EX}{1-\rho} + O(1) + \frac{N}{1-\rho} = O\left(\frac{1}{1-\rho}\right) \text{ as } \rho \to 1. \end{aligned}$$

This concludes the proof.