Queueing Syst (2006) 53: 19–30
DOI 10.1007/s11134-006-7584-x

Fluid and diffusion limits for transient sojourn times of processor sharing queues with time varying rates

Robert C. Hampshire · Mor Harchol-Balter · William A. Massey

Abstract We provide an approximate analysis of the transient sojourn time for a processor sharing queue with time varying arrival and service rates, where the load can vary over time, including periods of overload. Using the same asymptotic technique as uniform acceleration as demonstrated in [12] and [13], we obtain fluid and diffusion limits for the sojourn time of the $M_t/M_t/1$ processor-sharing queue. Our analysis is enabled by the introduction of a "virtual customer" which differs from the notion of a "tagged customer" in that the former has no effect on the processing time of the other customers in the system. Our analysis generalizes to non-exponential service and interarrival times, when the fluid and diffusion limits for the queueing process are known.

Keywords Processor sharing · Fluid limits · Diffusion limits · Transient behavior · Time-varying queues · Uniform acceleration · Sojourn times · Virtual customers.

1. Introduction

The processor sharing discipline has been used to model many aspects of computer systems, including the quantum-based time sharing of the CPU by computer operating systems (see Kleinrock [10]) and (elastic) traffic modeling in communication networks (see Nunez-Queija [16], Roberts [18] and Bonald and Proutière [2]) and scheduling in Web servers [6].

Under processor sharing (PS), the service capacity is at all times equally shared among all the jobs present. If there are n jobs present, each one receives a fraction $1/n$ of the total service capacity. This scheduling policy induces simple formulas in the case of an underloaded (stable) M/G/1/PS queue. For example, in Kleinrock [10] it is shown that

$$\lim_{t\to\infty} \mathsf{P}(Q(t) = n) = \begin{cases} (1-\rho)\rho^n & \text{if} \rho < 1, \\ 0 & \text{if} \rho \geq 1, \end{cases} \tag{1.1}$$

where $Q(t)$ denotes the number of jobs in the system at time t. Moreover, $\rho = \lambda \cdot E[S]$ where S denotes the random size (service requirement) of a job and λ is the mean Poisson customer arrival rate. Furthermore, the expected sojourn time for a job with size (service requirement) x is known to be

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

The popularity of the PS queue is due in large part to Kleinrock [10] who uses processor-sharing as an approximation of round-robin quantum-based scheduling. This is work that primarily deals with a *stationary queue* having load $\rho < 1$. The survey papers of Yashkov [23, 24] provide a detailed overview of the results on stationary PS queues, including results by Coffman, Muntz and Trotter [4], Morrison [15], Guillemin and Boyer [7]. All these papers deal with various aspects of the sojourn time distribution for the $M/M/1/PS$. These sojourn time analyses are often based on the notion of tracking a "tagged customer" who arrives into the system and interacts with the customers there. Masuyama and Takine [14] obtain similar results for a $MAP/M/1/PS$ queue. In the early 1980's Ott [17], Schassberger [19] and Yashkov [22]

R. C. Hampshire (✉) · W. A. Massey
Department of Operations Research and Financial Engineering, Princeton University
e-mail: {rhampshi, wmassey}@princeton.edu

M. Harchol-Balter
School of Computer Science, Carnegie Mellon University
e-mail: harchol@cs.cmu.edu

all derive independently the Laplace-Stieltjes transform (LST) for the sojourn time in the $M/G/1/PS$ queue. Zwart and Boxma [25] show how to eliminate the contour integrals in the above results and obtain a more explicit formula, which they use to efficiently compute the moments of sojourn time. They also prove that for heavy-tailed service demand distributions, the sojourn time distribution has the same tail index as the service distribution.

While stationary behavior of the M/G/1/PS queue is well-studied, it is important to understand the *transient* behavior of the M/G/1/PS queue as well. When a job arrives into the system, the job finds a specific number of existing jobs with existing remaining sizes (remaining service requirements), not a steady-state distribution. Initial work on the M/M/l/PS transient queue was done by Sengupta and Jagerman [21] who produce Laplace transforms for the transient behavior. Similar work on the M/G/l/PS transient queue was done later by Kitaev [9]. This was followed by a key paper in the area of transient analysis of the M/G/1/PS queue by Jean-Marie and Robert [8] who provide a fluid approximation for the case of fixed load $\rho > 1$. For this overloaded regime, they derive the asymptotic growth rate for the number of customers after t time units of overload (for large t), where this rate is the solution to a simple integral Equation. They also derive the asymptotic behavior of residual service times. These results were generalized by Chen, Kella, and Weiss [3] who also develop a fluid approximation, based on a time and space scaling. They too consider fixed load ρ and examine three regimes: $\rho < 1$, $\rho = 1$, and $\rho > 1$. They also allow for general conditions on what the tagged arrival sees in terms of the number of jobs found in the system with their residual job sizes.

There are two areas that the prior work on transient analysis does not address, and these form the primary contributions of this paper. First, the prior work all deals with a *fixed load* ρ. In practice there are short-term *fluctuations* in load, which have a dramatic impact on the sojourn time, and are not captured by the existing constant-rate models. To motivate the importance of capturing load fluctuations, consider the performance graph of an Apache web server shown in Figure 1, taken from [20]. In this figure, instantaneous load fluctuates between 1.2 and 0.2, where the time-average load is 0.7. As the authors in [20] point out, steady-state queueing formulas for load 0.7 result in a very poor prediction of mean sojourn time (or equivalently, the mean response time). The analysis techniques that we introduce in this paper result in simple formulas that capture the effect of fluctuating loads and rates on transient sojourn time. Second, the prior work provides only a *first-order* (mean) approximation of sojourn time. Since the fluid approximations correspond to a strong law of large numbers, it is natural to extend these results to diffusion limits, which correspond to central limit theorems. We thus strengthen the prior work by using diffusion limits,

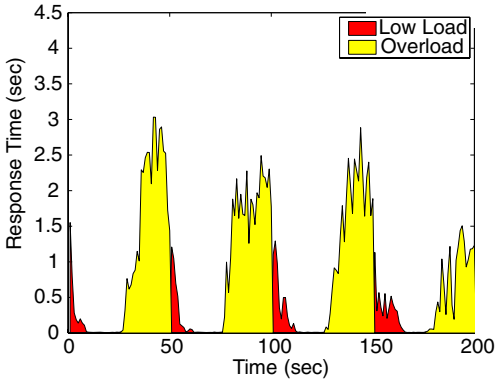


Fig. 1 Response time (averaged over 1 second intervals) for an Apache web server servicing HTTP requests, where load fluctuates between $\rho = 1.2$ and $\rho = 0.2$ (for 25 seconds each).

which we rigorously establish for exponential service distributions. This provides us with estimations of the standard deviations about the mean and enables us to estimate the *distribution* of sojourn time, rather than just its mean.

In this paper, we provide fluid and diffusion limits for the transient sojourn time of an $M_t/M_t/1/PS$ queue. Using this extended Kendall notation, the M_t for an arrival process denotes a non-homogeneous Poisson process. Similarly, the M_t for a service time distribution denotes the times between jumps for a non-homogeneous Poisson process. Our primary technique is uniform acceleration [13, 12], however we use it differently from how it has been used in the past. First and foremost, we are for the first time applying uniform acceleration to the sojourn time of a processor-sharing queue. Second, practical considerations have motivated us to extend the traditional uniform acceleration analysis and allow for general starting conditions in terms of the number of jobs seen by an arrival. Third, we introduce the notion of a "virtual customer," which differs from the traditional "tagged customer" in that the virtual customer has no effect on the experience of the other customers in the system.

The sojourn time $T(x)$ of the virtual customer with a job of size x can be determined from the following formula

$$x = \int_0^{T(x)} \frac{dt}{1 + Q(t)}. \tag{1.2}$$

This would be the exact sojourn time of a "real customer" if the effect of the virtual customer on $Q(t)$ were taken into account. Such a job arriving at time 0 increases the number of jobs in the system by 1, and is served at the instantaneous rate $1/(1 + Q(t))$ at time t.

Uniform acceleration is an asymptotic analysis method where we scale the arrival and service rates by a factor η. In the context of formula (1.2), this means that $\lambda(t) \Longrightarrow \eta\lambda(t)$ and $\mu(t) \Longrightarrow \eta\mu(t)$ and the corresponding queueing

process is referred to as $Q^\eta(t)$, where we now take the limit as $\eta \to \infty$.

It is not clear why uniform acceleration should tell us anything about the original unscaled system. For example, in the case of a stationary queue, scaling the arrival and service rates each by a factor of η should lead to a drop in the mean sojourn time by a factor of η. By contrast, for the case of transient (nonstationary) queues, we prove in Theorem 2.4, that we can induce an asymptotic analysis on $T(x)$ in terms of our asymptotic analysis of $Q^\eta(t)$. We also provide some underlying motivation for why the sojourn time obtained in the accelerated regime is indicative of what happens in the original (non-accelerated) system for the transient queue.

Our derivation of the sojourn time behavior of the $M_t/M_t/1/PS$ queue leads to some surprising results. First, we find that under systems with time-fluctuating load, the expected slowdown experienced by a job (its sojourn time divided by its size) is no longer constant. This is in sharp contrast to the classical M/G/1/PS queue with fixed load $\rho < 1$, which is characterized by constant slowdown (which is also referred to as "fairness" [1]). Second, in studying the sojourn time distribution, we observe a point mass. This indicates that for a given fixed job size, we can explicitly determine a point at which the distribution has positive mass.

We end the paper with an exploration of numerical examples that show how well the asymptotic results for the $M_t/M_t/1/PS$ queue approximate the mean, variance, and distribution of these sojourn times.

2. Uniform acceleration for the $M_t/M_t/1/PS$ queue

Since the behavior of the $M_t/M_t/1$ queueing process is independent of any work conserving queueing discipline, the sample path behavior of the $M_t/M_t/1/FIFO$ and $M_t/M_t/1/PS$ queues are identical. We can apply the asymptotic results to this system as found in Massey [13]. We use the asymptotic analysis of *uniform acceleration*, whereby we scale the time dependent arrival and service rates, $\lambda(t)$ and $\mu(t)$ respectively, by the same parameter $\eta > 0$. This results in an $M_t/M_t/1$ queueing process with arrival and service rates $\eta\lambda(t)$ and $\eta\mu(t)$, respectively, and we denote this queueing process as $\{Q^\eta(t)|t \geq 0\}$. We then analyze the behavior of this process asymptotically by letting our scale factor η become very large.

We start by recalling the following asymptotic analysis for the transition probabilities of the $M_t/M_t/1$ queueing process:

Theorem 2.1 (Massey, 1985). *If λ and μ are continuously differentiable functions of t, we have $\rho^*(t) < 1$ where*

$$\rho^*(t) \equiv \sup_{0 \leq s < t} \frac{\int_s^t \lambda(r)dr}{\int_s^t \mu(r)dr}, \tag{2.1}$$

and $\rho(t) = \lambda(t)/\mu(t)$, then

$$\mathrm{P}(Q^\eta(t) = n) = (1 - \rho(t))\rho(t)^n + \frac{\rho'(t)}{\eta\mu(t)}\left(\frac{\rho(t)}{(1-\rho(t))^2} - \frac{n(n+1)}{2}\right)\rho(t)^{n-1} + O\left(\frac{1}{\eta^2}\right), \tag{2.2}$$

as $\eta \to \infty$. Moreover, if $\rho^*(t) \geq 1$, then

$$\lim_{\eta\to\infty} \mathrm{P}(Q^\eta(t) = n) = 0 \tag{2.3}$$

for all $t > 0$.

From this analysis there are three different regimes of asymptotic behavior that are labeled as follows: underloaded ($\rho^*(t) < 1$), critically loaded ($\rho^*(t) = 1$) and overloaded ($\rho^*(t) > 1$). Observe that the parameter that determines these asymptotic regimes is *not* given by $\rho(t)$. Also, observe that the results of this theorem are non-trivial (non-zero) only for the underloaded case.

Mandelbaum and Massey [12] apply the analysis of uniform acceleration directly to the random sample path behavior of the $M_t/M_t/1$ queue.

Theorem 2.2 (Mandelbaum and Massey, 1995). *If λ and μ are locally integrable functions and $Q^\eta(0) = Q(0)$ for all $\eta > 0$, then $\lim_{\eta\to\infty} Q^\eta(t)/\eta = Q^{(0)}(t)$* a.s. *uniformly on compact sets, where*

$$Q^{(0)}(t) = \int_0^t (\lambda(s) - \mu(s))ds - \inf_{0\leq s\leq t} \int_0^s (\lambda(r) - \mu(r))dr. \tag{2.4}$$

Moreover, $\lim_{\eta\to\infty} \sqrt{\eta}\left(Q^\eta(t)/\eta - Q^{(0)}(t)\right) \stackrel{d}{=} Q^{(1)}(t)$, *where*

$$Q^{(1)}(t) = \mathrm{W}\left(\int_0^t (\lambda(s) + \mu(s))ds\right) - \inf_{s\in\Phi(t)} \mathrm{W}\left(\int_0^s (\lambda(r) + \mu(r))dr\right), \tag{2.5}$$

$\{W(t)|t \geq 0\}$ is standard (mean 0, variance t) Brownian motion, and finally

$$\Phi(t) = \left\{s \middle| \int_s^t (\lambda(r) - \mu(r))dr = Q^{(0)}(t) \text{ and } 0 \leq s \leq t\right\}. \tag{2.6}$$

The results of Theorem 2.1 are non-zero precisely when the results of Theorem 2.2 are zero. Conversely, the results of Theorem 2.2 are non-zero when the results of Theorem 2.1

are zero. Thus these two types of uniform acceleration analysis literally complement each other.

The deterministic process $\{Q^{(0)}(t)|t \geq 0\}$ is called the $M_t/M_t/1$ queueing *fluid limit*. Observe that $Q^{(0)}(t) > 0$ if and only if $\rho^*(t) > 1$. The fluid limit is then an estimate of the backlog for the original (non accelerated, $\eta = 1$) queueing process. Notice that since we always have $\rho(t) \leq \rho^*(t)$, it is possible to have $\rho(t) < 1$ (or $\lambda(t) < \mu(t)$), but still have $\rho^*(t) \geq 1$. This is due to a backlog of jobs, acquired during a period of overloading in the past, that have not been flushed out of the queue by time t.

The random process $\{Q^{(1)}(t)|t \geq 0\}$ is the first order "correction term" to the fluid limit and gives a sense of how the original queueing process deviates from the fluid model. For simplicity, we refer to it as the $M_t/M_t/1$ queueing *diffusion limit*. Technically, it may *not* be a diffusion since the sample paths at fixed time points may have a non-zero probability of a discontinuity. However, these discontinuities only occur during transitions *from* the overloaded regime to the underloaded one.

Now we generalize these limit theorems to the *extended* case of a uniformly accelerated initial load where we set $Q^\eta(0) = \eta \cdot Q(0)$. Below we have our extended fluid and diffusion limits.

Theorem 2.3. *If $Q^\eta(0) = \eta \cdot Q(0)$, then*

$$Q^{(0)}(t) = Q(0) + \int_0^t (\lambda(s) - \mu(s))ds - \inf_{0\leq s\leq t}\left(Q(0) + \int_0^s (\lambda(r) - \mu(r))dr\right) \wedge 0, \quad (2.7)$$

and

$$Q^{(1)}(t) = \begin{cases} Q_0^{(1)}(t) \\ \text{if } Q(0) < \sup_{0\leq s\leq t} \int_0^s (\mu(r) - \lambda(r))\, dr, \\ Q_0^{(1)}(t) \vee W\left(\int_0^t (\lambda(s) + \mu(s))ds\right) \\ \text{if } Q(0) = \sup_{0\leq s\leq t} \int_0^s (\mu(r) - \lambda(r))\, dr, \\ W\left(\int_0^t (\lambda(s) + \mu(s))ds\right) \\ \text{if } Q(0) > \sup_{0\leq s\leq t} \int_0^s (\mu(r) - \lambda(r))\, dr, \end{cases} \quad (2.8)$$

where $\left\{ Q_0^{(1)}(t) \mid t \geq 0 \right\}$ is the $M_t/M_t/1$ queueing diffusion limit given by (2.5) when $Q(0) = 0$.

Proof: Both limits follow from the fact that the process $\{Q^\eta(t) \mid t \geq 0\}$ can be written as

$$Q^\eta(t) = Z^\eta(t) - \inf_{0\leq s\leq t} Z^\eta(s) \wedge 0 = \max\left(Q_0^\eta(t), Z^\eta(t)\right), \quad (2.9)$$

where $\left\{Q_0^\eta(t) \mid t \geq 0\right\}$ is another $M_t/M_t/1$ queueing process with the same arrival and service rates but with $Q(0) = 0$, and

$$Z^\eta(t) = \eta \cdot Q(0) + \Pi_1\left(\int_0^t \eta\lambda(s)\, ds\right) - \Pi_2\left(\int_0^t \eta\mu(s)\, ds\right), \quad (2.10)$$

where $\{\Pi_i(t) \mid t \geq 0\}$ for $i = 1, 2$ are two independent, standard (rate 1) Poisson processes that are used to construct both queueing processes Q^η and Q_0^η. Using the theory of strong approximations for Poisson processes, we have

$$\lim_{\eta\to\infty} \sup_{0\leq s\leq t} \left|\frac{1}{\eta} Z^\eta(s) - Q(0) - \int_0^s (\lambda(r) - \mu(r))\, dr\right| = 0 \text{ a.s.} \quad (2.11)$$

and we can construct a Brownian motion, $\{W(t)|t \geq 0\}$, such that

$$\lim_{\eta\to\infty} \sup_{0\leq s\leq t} \left|\sqrt{\eta}\left(\frac{1}{\eta} Z^\eta(s) - Q(0) - \int_0^s (\lambda(r) - \mu(r))\, dr\right) - W\left(\int_0^s (\lambda(r) + \mu(r))\, dr\right)\right| \stackrel{d}{=} 0. \quad (2.12)$$

□

Applying the asymptotics of (2.11), (2.12) and Theorem 2.2 to (2.9) gives us the desired result. ∎

Suppose that we have a single virtual customer with a job of size x sharing a unit processing rate with $Q(t)$ customers. Combining our notion of a virtual customer with uniform acceleration, we transform this system into η virtual customers with jobs of size x/η sharing a unit processing rate with $Q^\eta(t)$ customers. This gives us

$$x/\eta = \int_0^{T^\eta(x)} \frac{dt}{\eta + Q^\eta(t)} \quad \Longrightarrow x = \int_0^{T^\eta(x)} \frac{dt}{1 + Q^\eta(t)/\eta}. \quad (2.13)$$

Now we can state our main result and its proof.

Theorem 2.4. *Given an $M_t/M_t/1/PS$ queue and for all $x \geq 0$, we have the strong law of large numbers limit*

$$\lim_{\eta\to\infty} T^\eta(x) = T^{(0)}(x) \text{ a.s.} \qquad \text{where}$$

$$x = \int_0^{T^{(0)}(x)} \frac{dt}{1 + Q^{(0)}(t)}, \quad (2.14)$$

and the central limit theorem

$$\lim_{\eta\to\infty}\sqrt{\eta}\,\left(T^{\eta}(x)-T^{(0)}(x)\right)$$
$$\stackrel{d}{=} T^{(1)}(x)\equiv T^{(0)'}(x)\cdot\int_0^{T^{(0)}(x)}\frac{Q^{(1)}(t)\,dt}{\left(1+Q^{(0)}(t)\right)^2}. \qquad (2.15)$$

Proof: To prove the strong law limit result (2.14), it suffices to show that every convergent subsequence $\left\{T^{\eta(k)}(x)|k=1,2,\ldots\right\}$, where $\eta(k)\to\infty$, converges to $T^{(0)}(x)$.

Let $U=\lim_{k\to\infty}T^{\eta(k)}(x)$. Since it follows that

$$\left|\int_0^{T^{\eta(k)}(x)}\frac{dt}{1+Q^{\eta(k)}(t)/\eta(k)}-\int_0^{U}\frac{dt}{1+Q^{(0)}(t)}\right| \le\left|T^{\eta(k)}(x)-U\right|, \qquad (2.16)$$

we then have

$$x=\lim_{k\to\infty}\int_0^{T^{\eta(k)}(x)}\frac{dt}{1+Q^{\eta(k)}(t)/\eta(k)}=\int_0^{U}\frac{dt}{1+Q^{(0)}(t)}. \qquad (2.17)$$

Since $1/(1+Q^{(0)}(t))$ is never zero, we must then have $U=T^{(0)}(x)$.

To prove the central limit theorem result (2.15), we first observe that

$$\int_{T^{(0)}(x)}^{T^{\eta}(x)}\frac{dt}{1+Q^{(0)}(t)}$$
$$=\int_0^{T^{\eta}(x)}\frac{dt}{1+Q^{(0)}(t)}-\int_0^{T^{(0)}(x)}\frac{dt}{1+Q^{(0)}(t)} \qquad (2.18)$$
$$=\int_0^{T^{\eta}(x)}\frac{dt}{1+Q^{(0)}(t)}-\int_0^{T^{\eta}(x)}\frac{dt}{1+Q^{\eta}(t)/\eta} \qquad (2.19)$$
$$=\int_0^{T^{\eta}(x)}\frac{\left(Q^{\eta}(t)/\eta-Q^{(0)}(t)\right)dt}{\left(1+Q^{(0)}(t)\right)\left(1+Q^{\eta}(t)/\eta\right)}. \qquad (2.20)$$

Now we obtain the identity

$$\sqrt{\eta}\,\left(T^{\eta}(x)-T^{(0)}(x)\right)$$
$$=\left(\frac{1}{T^{\eta}(x)-T^{(0)}(x)}\int_{T^{(0)}(x)}^{T^{\eta}(x)}\frac{dt}{1+Q^{(0)}(t)}\right)^{-1}$$
$$\times\int_0^{T^{\eta}(x)}\frac{\sqrt{\eta}\,\left(Q^{\eta}(t)/\eta-Q^{(0)}(t)\right)dt}{\left(1+Q^{(0)}(t)\right)\left(1+Q^{\eta}(t)/\eta\right)}. \qquad (2.21)$$

□

The rest follows since $\sqrt{\eta}\,\left(Q^{\eta}(t)/\eta-Q^{(0)}(t)\right)$ converges in distribution to a random process and all the other limits converge to constants. ■

It should be pointed out that the proof given here holds for any processor sharing queueing system where the queueing process for the number of customers in the system has fluid and diffusion limits.

We note that Theorem 2.4 applies to the *transient* sojourn times of the $M_t/M_t/1/PS$ queue. We emphasize the word transient to stress the fact that we are *not* analyzing sojourn times where the initial queue has a steady state distribution. The results that we discuss here do not necessarily have any implications for constant rate queues in steady state.

Figure 2 illustrates the heuristic idea behind Theorem 2.4. The original queue is represented in the top part of the figure. To illustrate our scale transformation, the jobs are now "quantized" into a series of shaded squares that represent unit processing times that can be viewed as "CPU time units." The bottom part of Figure 2 represents an *accelerated* queue with scale factor 2, where both the number of new jobs and the processing rates of these jobs are doubled. Observe that this is equivalent to leaving the processing rate untouched but breaking each job in the original queue into two identical "subjobs", that are each half the size of the original total job.

Figure 2 illustrates a specific example where the job sizes in the original queue are all multiples of the scale factor 2 and all jobs are initially present. In this specific case shown, the completion times of the final subjob for the accelerated queue are identical to the corresponding completion times of the jobs for the original queue. We refer to this property as "scale invariance" for the processor sharing service discipline. While it is clear that scale invariance holds for the specific example illustrated in Figure 2, and probably generalizes to other cases of deterministic job sizes with appropriately sized jobs, this argument cannot be generalized indefinitely. Yet we hope that this provides some intuition behind Theorem 2.4, specifically why the accelerated system allows us to understand the behavior of the nonaccelerated system.

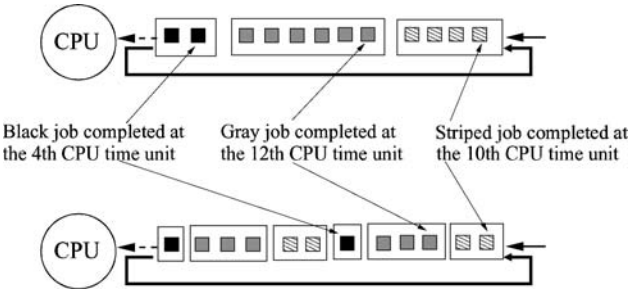


Fig. 2 Scale invariance for processor sharing and limited round robin

For deterministic job sizes and arrival processes, we can then define a precise notion of "scale invariance" for the processor sharing service discipline and a limited version of round robin. For the round robin case, we are limited to the case of the job sizes in the original queue initially being integer multiples of the scale factor 2 times the CPU time unit and all jobs are present initially. In either case, the completion time of the final subjobs for the accelerated queue is identical to the completion times of the corresponding total jobs for the original queue. All these results generalize to any scale factor of integer size η.

For a random processor sharing queue, we define a stochastic generalization of this type of acceleration. For the example of a scale factor equal to 2, we still assume that the initial load of jobs in the system is deterministic and we double this number. For the job arrival process however, we replace this notion of doubling (scaling by 2) by the superposition of two i.i.d. replicas of the original arrival process. Note that for deterministic processes, this is the same as doubling. The sizes of these arriving subjobs are i.i.d. random variables but each one has the same distribution as one half the size of a random job (half as many CPU time units) in the original queue. For the case of the $M_t/M_t/1/PS$ queue, this notion of "stochastic acceleration" is identical to uniform acceleration with scale factor $\eta = 2$. Now we extend the definition of uniform acceleration to the case of the scale factor η being any integer. Instead of scale invariance, we have the fluid and diffusion limits of Theorem 2.4 as $\eta \to \infty$.

In the next section, we reduce these formulas to the constant rate case. We can obtain more explicit formulas that give us insight into how the sojourn time responds to periods of underloading and overloading.

3. Exact formulas for the constant rate case

We have the following formula for the fluid limit of the $M/M/1/PS$ queue, which is a special case of Chen, Kella and Weiss [3].

Theorem 3.1. *Given an $M/M/1/PS$ queue with arrival rate λ and service rate μ, we have*

$$Q^{(0)}(t) = (Q(0) + (\lambda - \mu)t)^+, \tag{3.1}$$

for all $t \geq 0$. Moreover, for all job sizes $x \geq 0$, we have

$$T^{(0)}(x) = x + \int_0^x \left((Q(0)+1)\cdot e^{(\lambda-\mu)y} - 1\right)^+ dy. \tag{3.2}$$

Finally, for all job sizes $x \geq 0$, we have

$$Q^{(0)}\left(T^{(0)}(x)\right) = \left((Q(0)+1)\cdot e^{(\lambda-\mu)x} - 1\right)^+. \tag{3.3}$$

For the case of $\lambda < \mu$, let t_* equal the first time that the extended fluid limit process empties. Similarly, let x_* equal the size of the smallest *virtual* job starting at time 0 that finishes just as the fluid limit first empties. We can write them explicitly as

$$t_* \equiv \frac{Q(0)}{\mu - \lambda} \quad \text{and} \quad x_* \equiv \frac{\log(Q(0)+1)}{\mu - \lambda}. \tag{3.4}$$

Using t_*, the extended diffusion limit for the queueing process can be written as follows:

Theorem 3.2. *If $\lambda > \mu$ or $\lambda = \mu$ and $Q(0) > 0$, then*

$$\left\{ Q^{(1)}(t) \mid t \geq 0 \right\} \stackrel{d}{=} \left\{ W((\lambda+\mu)t) \mid t \geq 0 \right\}. \tag{3.5}$$

If $\lambda = \mu$ and $Q(0) = 0$, then

$$\left\{ Q^{(1)}(t) \mid t \geq 0 \right\} \stackrel{d}{=} \left\{ W(2\lambda t) - \inf_{0\leq s\leq t} W(2\lambda s) \;\middle|\; t \geq 0 \right\}. \tag{3.6}$$

Finally, if $\lambda < \mu$ and $Q(0) \geq 0$, then

$$Q^{(1)}(t) \stackrel{d}{=} \begin{cases} W((\lambda+\mu)t) & \text{if } t < t_*, \\ W((\lambda+\mu)t)^+ & \text{if } t = t_*, \\ 0 & \text{if } t > t_*. \end{cases} \tag{3.7}$$

Note that for the case of $\lambda < \mu$ and $Q(0) > 0$, almost all the sample paths of $\{Q^{(1)}(t) | t \geq 0\}$ are discontinuous at t_*. Probabilistically speaking, half of them are only left continuous (i.e. $W((\lambda+\mu)t_*) > 0$) and the other half are only right continuous at t_*.

Excluding the critical loading case of $\lambda = \mu$ and $Q(0) = 0$, we can totally characterize the distribution of $T^{(1)}(x)$.

Theorem 3.3. *If $Q(0) > 0$ or $\lambda \neq \mu$, then $T^{(1)}(x)$ is a Gaussian random variable with $E\left[T^{(1)}(x)\right] = 0$. Moreover, we have*

$$\mathsf{Var}\left[T^{(1)}(x)\right] = \frac{\left(1+Q(0)\right)(\lambda+\mu)}{(\lambda-\mu)^3}\left[e^{2(\lambda-\mu)x} - 2(\lambda-\mu)xe^{(\lambda-\mu)x} - 1\right], \tag{3.8}$$

provided that either $\lambda > \mu$ or the conjunction of $\lambda < \mu$ and $x < x_*$ holds. When $\lambda = \mu$, this formula reduces by L'Hopital's rule to

$$\mathsf{Var}\left[T^{(1)}(x)\right] = \left(1+Q(0)\right)\frac{2\lambda}{3}x^3. \tag{3.9}$$

Otherwise, when $\lambda < \mu$ and $x \geq x_*$, we have

$$\mathsf{Var}\left[T^{(1)}(x)\right] = \frac{\lambda+\mu}{(\mu-\lambda)^3}\left[Q(0)+1 - 2\cdot\log\left(Q(0)+1\right) - \frac{1}{Q(0)+1}\right]. \tag{3.10}$$

Proof of Theorem 3.1: The extended constant rate fluid limit (3.1) follows immediately from Equation (2.7). The constant rate formula given by (3.3) follows from differentiating $T^{(0)}(x)$ and subtracting one from it. It remains to derive a closed form solution to the extended fluid limit of the sojourn time.

If $\lambda = \mu$, then $Q^{(0)}(t) = Q(0)$ for all $t \geq 0$, and so $T^{(0)}(x) = (Q(0)+1)\cdot x$.

If $\lambda > \mu$, then

$$x = \int_0^{T^{(0)}(x)} \frac{dt}{1+Q(0)+(\lambda-\mu)t} = \frac{1}{\lambda-\mu}\log\left(1+\frac{\lambda-\mu}{Q(0)+1}T^{(0)}(x)\right). \tag{3.11}$$

Solving for $T^{(0)}(x)$ gives us

$$T^{(0)}(x) = (Q(0)+1)\cdot\frac{e^{(\lambda-\mu)x}-1}{\lambda-\mu} = x + \int_0^x \left((Q(0)+1)\cdot e^{(\lambda-\mu)y}-1\right)^+ dy. \tag{3.12}$$

The last step follows from $e^{(\lambda-\mu)x}$ being an *increasing* function of x and *greater* than one when $\lambda > \mu$ and $x > 0$.

Now let $\lambda < \mu$. Since $t_* = \inf\{t | Q^{(0)}(t) = 0\}$, we have the two cases of $T^{(0)}(x) < t_*$ and $T^{(0)}(x) \geq t_*$. If $T^{(0)}(x) < t_*$, then we have the same equation for $T^{(0)}(x)$ as (3.11) for the case of $\lambda > \mu$. By continuity, we then have $T^{(0)}(x_*) = t_*$.

If $T^{(0)}(x) \geq t_*$, then $x \geq x_*$, and so we have $x - x_* = T^{(0)}(x) - t_*$. This equation simply states the fact that x_* is the amount of the fluid model job of size x that was processed with the initial load of fluid model jobs. Since $\lambda < \mu$, the fluid level is zero for all time after t_*. This means that the remaining job amount of $x - x_*$ has the server all to itself. Consequently,

$$T^{(0)}(x) = x - x_* + t_*, \tag{3.13}$$

which gives us

$$T^{(0)}(x) = x + \int_0^{x_*}\left((Q(0)+1)\,e^{(\lambda-\mu)y}-1\right)dy = x + \int_0^x \left((Q(0)+1)\,e^{(\lambda-\mu)y}-1\right)^+ dy. \tag{3.14}$$

The last step follows from $e^{(\lambda-\mu)x}$ being a *decreasing* function of x when $\lambda < \mu$. This argument also proves that this last formula is true for the previous case of $T^{(0)}(x) < t_*$. ∎

Proof of Theorem 3.3: When $\lambda > \mu$, let $\alpha \equiv (\lambda-\mu)/(1+Q(0))$. Using the identities

$$\int_0^t \frac{s\,ds}{(1+\alpha s)^2} = \frac{1}{\alpha^2}\left[\log(1+\alpha t) + \frac{1}{1+\alpha t} - 1\right] \tag{3.15}$$

and

$$\int_0^t \frac{\log(1+\alpha s)\,ds}{(1+\alpha s)^2} = \frac{1}{\alpha}\left[1 - \frac{\log(1+\alpha t)}{1+\alpha t} - \frac{1}{1+\alpha t}\right], \tag{3.16}$$

gives us

$$\begin{aligned}
&\mathsf{Var}\left[T^{(1)}(x)\right] \\
&= (1+Q(0))^2\cdot e^{2(\lambda-\mu)x}\times\int_0^{T^{(0)}(x)} \int_0^{T^{(0)}(x)} \frac{\mathsf{Cov}\left[W((\lambda+\mu)s),\,W((\lambda+\mu)t)\right]ds\,dt}{(1+Q(0)+(\lambda-\mu)s)^2(1+Q(0)+(\lambda-\mu)t)^2} \\
&= \frac{2(\lambda+\mu)e^{2(\lambda-\mu)x}}{(1+Q(0))^2}\cdot\int_0^{T^{(0)}(x)}\left(\int_0^t \frac{s\,ds}{\left(1+\alpha s\right)^2}\right)\frac{dt}{\left(1+\alpha t\right)^2} \\
&= \frac{2(\lambda+\mu)e^{2(\lambda-\mu)x}}{(1+Q(0))^2\alpha^2}\cdot \int_0^{T^{(0)}(x)} \frac{\log(1+\alpha t)+1/(1+\alpha t)-1}{(1+\alpha t)^2}dt \\
&= \frac{(\lambda+\mu)(1+Q(0))}{(\lambda-\mu)^3}\cdot\left[e^{2(\lambda-\mu)x} - 2(\lambda-\mu)xe^{(\lambda-\mu)x} - 1\right].
\end{aligned}$$

For the case of $\lambda < \mu$, observe that when $x < x_*$, we have

$$T^{(1)}(x) = T^{(0)\prime}(x)\cdot\int_0^{T^{(0)}(x)} \frac{W((\lambda+\mu)t)\,dt}{(1+Q(0)+(\lambda-\mu)t)^2}. \tag{3.17}$$

It follows that the variance formula is the same here as for the case of $\lambda > \mu$.

For the case of $\lambda < \mu$ and $x \geq x_*$, we have $Q^{(0)}(T^{(0)}(x)) = Q^{(0)}(T^{(0)}(x_*)) = 0$. Therefore,

$$T^{(1)}(x) = T^{(0)\prime}(x_*)\cdot\int_0^{T^{(0)}(x_*)} \frac{W((\lambda+\mu)t)\,dt}{(1+Q(0)+(\lambda-\mu)t)^2}. \tag{3.18}$$

If follows that the variance formula for $T^{(1)}(x)$ is computed here by applying the previous variance formula to $T^{(1)}(x_*)$. ∎

4. Numerics for the mean and variance

The aim of this section is to compare our analytical results for these $M_t/M_t/1/PS$ sojourn times with results from simulation. The limit theorems of Section 2 suggest the following fluid and diffusion limit approximations for the mean and variance of the sojourn times,

$$\mathsf{E}\left[T(x)\right] \approx T^{(0)}(x) + \mathsf{E}\left[T^{(1)}(x)\right] \quad \text{and}$$
$$\mathsf{Var}\left[T(x)\right] \approx \mathsf{Var}\left[T^{(1)}(x)\right], \tag{4.1}$$

where in general we have $\mathsf{E}[T^{(1)}(x)] = 0$. Now we give some general conditions for computing $\mathsf{Var}\left[T^{(1)}(x)\right]$.

Suppose that the queueing process alternates between periods of underloading and overloading, where the times of critical loading are discrete, isolated points. If we set σ_n and τ_n to be respectively, the starting time for the n-th period of overloading and the ending time for the n-th period of overloading, then we have

$$\int_{\sigma_n}^{\tau_n} (\lambda(s) - \mu(s))ds = 0 \tag{4.2}$$

which implies

$$Q^{(0)}(t) = \begin{cases} \int_{\sigma_n}^{t} (\lambda(s) - \mu(s))\, ds & \text{if } \sigma_n \le t < \tau_n, \\ 0 & \text{otherwise,} \end{cases} \tag{4.3}$$

and

$$Q^{(1)}(t) = \begin{cases} W\left(\int_0^t (\lambda(s) + \mu(s))\, ds\right) & \\ -W\left(\int_0^{\sigma_n} (\lambda(s) + \mu(s))\, ds\right) & \text{if } \sigma_n \le t < \tau_n, \\ \left(W\left(\int_0^t (\lambda(s) + \mu(s))\, ds\right)\right. & \\ \left. -W\left(\int_0^{\sigma_n} (\lambda(s) + \mu(s))\, ds\right)\right)^+ & \text{if } t = \tau_n, \\ 0 & \text{otherwise.} \end{cases} \tag{4.4}$$

Theorem 4.1. *If we are given an $M_t/M_t/1$ queue where critical loading only occurs at isolated time points, then $T^{(1)}(x)$ is a Gaussian random variable, with $\mathsf{E}\left[T^{(1)}(x)\right] = 0$ and*

$$\mathsf{Var}\left[T^{(1)}(x)\right] = T^{(0)'}(x)^2 \cdot \sum_{n=0}^{\infty} \int_{\sigma_n(x)}^{\tau_n(x)} \int_{\sigma_n(x)}^{\tau_n(x)} \frac{\left(\int_{\sigma_n(x)}^{s\wedge t} (\lambda(r) + \mu(r))\, dr\right) ds\, dt}{\left(1 + Q^{(0)}(s)\right)^2 \left(1 + Q^{(0)}(t)\right)^2}, \tag{4.5}$$

with

$$\sigma_n(x) \equiv \sigma_n \wedge T^{(0)}(x) \quad \text{and} \quad \tau_n(x) \equiv \tau_n \wedge T^{(0)}(x). \tag{4.6}$$

Proof: We have

$$T^{(1)}(x) = T^{(0)'}(x)^2 \cdot \int_0^{\infty} \frac{Q^{(1)}(t)\, dt}{\left(1 + Q^{(0)}(t)\right)^2} = T^{(0)'}(x)^2 \cdot \sum_{n=0}^{\infty} \int_{\sigma_n(x)}^{\tau_n(x)} \frac{Q^{(1)}(t)\, dt}{\left(1 + Q^{(0)}(t)\right)^2} \tag{4.7}$$

Since Brownian motion has the independent increment property, then by (4.4) $Q^{(1)}(s)$ and $Q^{(1)}(t)$ are independent random variables whenever s and t belong to disjoint periods of overloading. This means that

$$\begin{aligned} \mathsf{Var}\left[T^{(1)}(x)\right] &= T^{(0)'}(x)^2 \cdot \mathsf{Var}\left[\sum_{n=0}^{\infty} \int_{\sigma_n(x)}^{\tau_n(x)} \frac{Q^{(1)}(t)\, dt}{\left(1 + Q^{(0)}(t)\right)^2}\right] \\ &= T^{(0)'}(x)^2 \cdot \sum_{n=0}^{\infty} \mathsf{Var}\left[\int_{\sigma_n(x)}^{\tau_n(x)} \frac{Q^{(1)}(t)\, dt}{\left(1 + Q^{(0)}(t)\right)^2}\right] \\ &= T^{(0)'}(x)^2 \cdot \sum_{n=0}^{\infty} \int_{\sigma_n(x)}^{\tau_n(x)} \int_{\sigma_n(x)}^{\tau_n(x)} \frac{\mathsf{Cov}\left[Q^{(1)}(s), Q^{(1)}(t)\right] ds\, dt}{\left(1 + Q^{(0)}(s)\right)^2 \left(1 + Q^{(0)}(t)\right)^2}. \end{aligned}$$

The last step follows from setting

$$\mathsf{Cov}\left[Q^{(1)}(s), Q^{(1)}(t)\right] = \int_{\sigma_n(x)}^{s\wedge t} (\lambda(r) - \mu(r))\, dr, \tag{4.8}$$

which completes the proof. □

Figure 3 shows the approximations for mean and variance given by Equation (4.1) as compared with simulation, as well as the relative error. The left column of Figure 3 assumes a non-homogeneous Poisson arrival process with mean rate $\lambda(t) = 1.2 + 0.2 * \sin(2\pi * 0.2 * t)$, while the right column assumes a non-homogeneous Poisson arrival process with

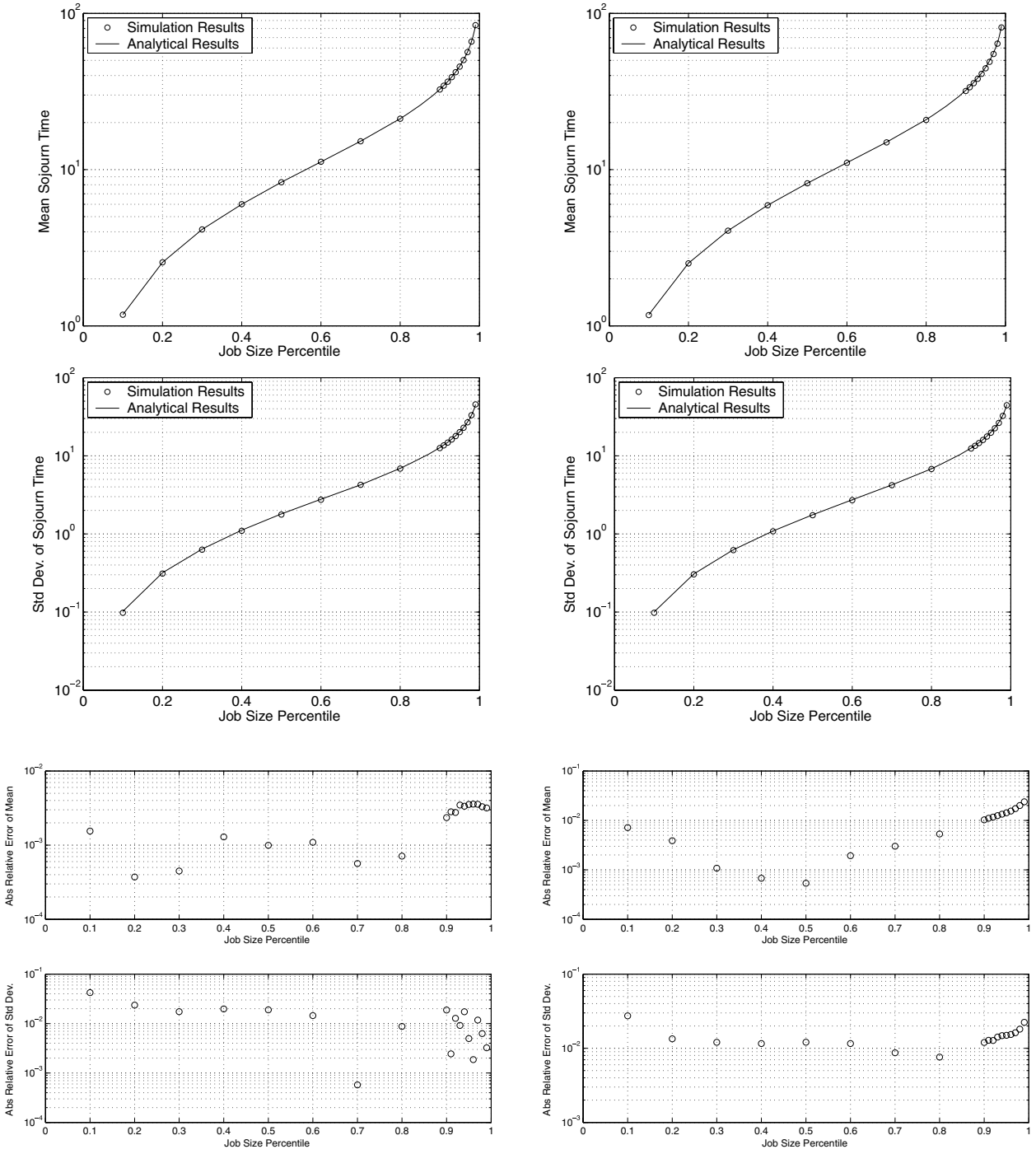
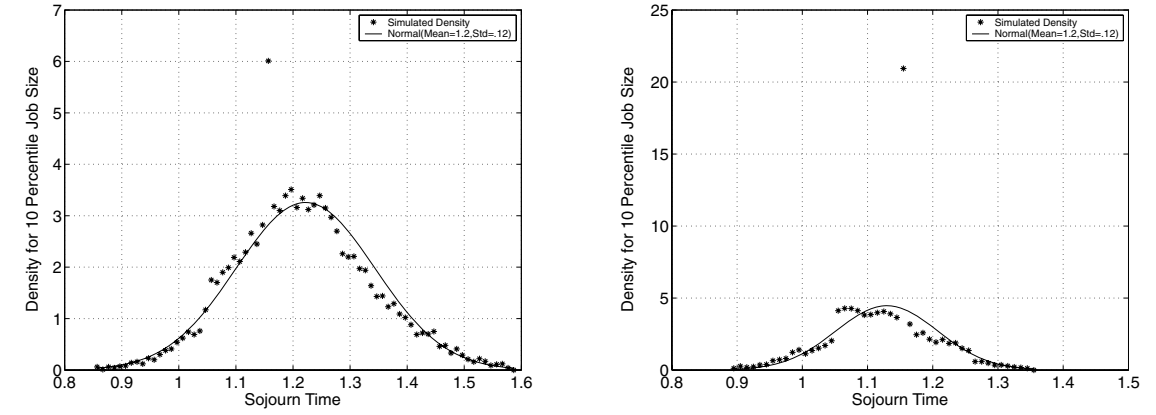


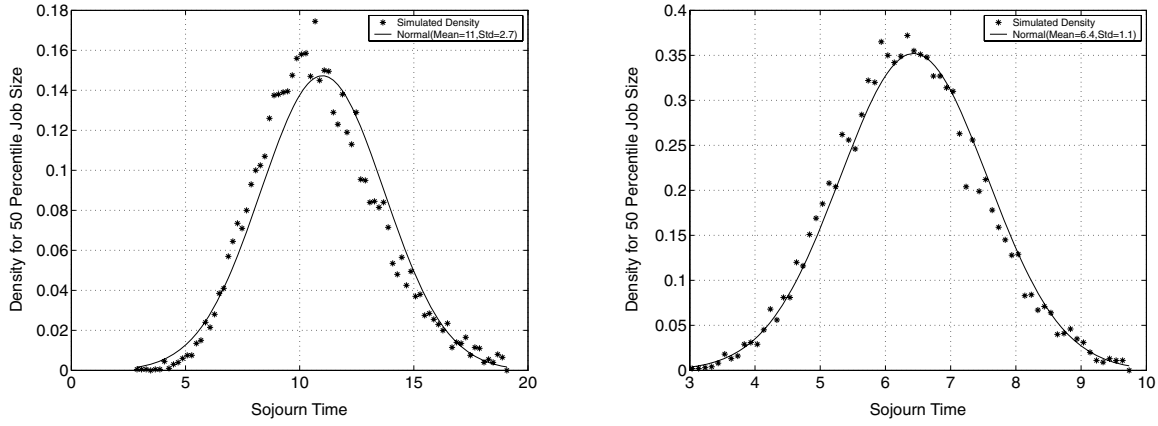
Fig. 3 Mean, standard deviation and relative errors for the sojourn times of $M_t/M/1/PS$ when $\lambda(t) = 1.2 + .2 * \sin 0.4\pi t$ (left column) and $\lambda(t) = 1.2 + .2 * \sin 20\pi t$ (right column), where $\mu = 1$. Here, we assume throughout that $Q(0) = 10$.

mean rate $\lambda(t) = 1.2 + 0.2 * \sin(2\pi * 10.0 * t)$. Throughout, all jobs have sizes that are exponentially distributed with mean 1, and the initial number of jobs in the system, $Q(0)$ is fixed at 10. All the simulation results presented are derived from 10,000 realizations.

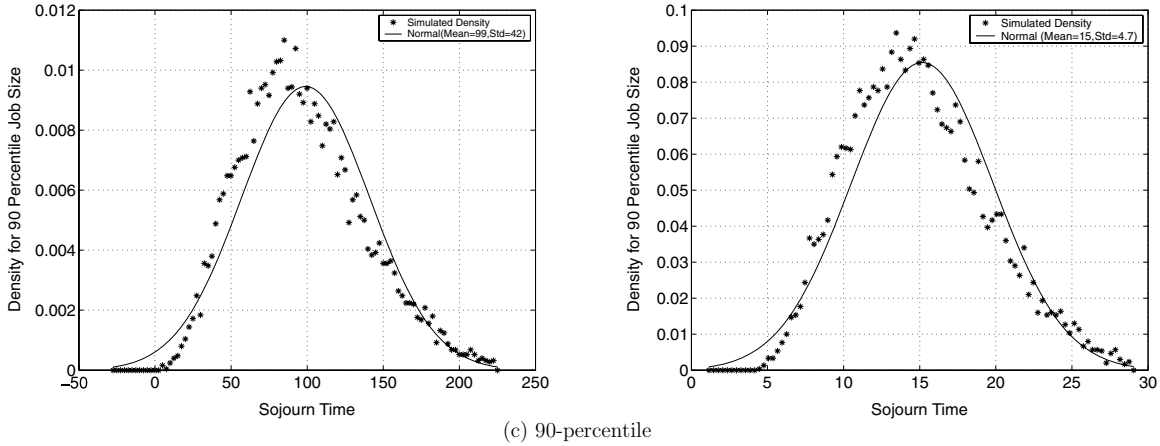
The maximum relative error of the fluid limit for the smaller frequency case (i.e. 0.2 versus 10.0) is less than the higher frequency case. This is consistent with results from the theory of uniform acceleration as seen in Theorem 3.1. The leading order terms of Equation (3.2, 3.3 and 3.4) do not depend on the rate of change in the offered load, $\rho'(t)$. However their correction terms grow in magnitude as $\rho'(t)$ becomes larger. Therefore it is reasonable to observe a smaller relative error in the slowly varying rate case.



(a) 10-percentile



(b) 50-percentile



(c) 90-percentile

Fig. 4 Empirical results. Density function for M/M/1/PS, restricted to jobs in (a) 10-percentile, (b) 50-percentile, (c) 90-percentile. Left column shows overloaded case where $\lambda = 2.0$ and $\mu = 1.0$. Right column shows underloaded case where $\lambda = 0.5$ and $\mu = 1$. Here, we assume throughout that $Q(0) = 10$.

5. Bimodality and numerics for the distribution

In this section, we first show that the sojourn time distribution always has a point mass distribution at the constant $(1 + Q(0))x$.

Theorem 5.1. *Conditioned on $Q(0)$, we can define the following two independent events:*

1. *The first $Q(0)$ i.i.d. exponential service times are all larger than x.*
2. *The number of non-homogeneous Poisson arrivals for the next $(1 + Q(0))x$ time units is zero.*

It then follows that the intersection of these two events 1 and 2 implies the event $\{T(x) = (1 + Q(0))\, x\}$.

Moreover, if $T(x)$ is the sojourn time for the $M_t/M_t/1/PS$ queue with some initial load $Q(0)$, then we have

$$\mathsf{P}\left(T(x) = (1 + Q(0))\, x\right) = \exp\left(-\int_0^{(1+Q(0))x} \lambda_t\, dt - Q(0) \cdot \int_0^x \mu_t\, dt\right), \tag{5.1}$$

which equals the probability for the intersection of these two events 1 and 2.

Proof: When one of these two events does not happen, then there is some exponentially distributed random variable that is less than $(1 + Q(0))x$. This random variable, conditioned on being less than the constant $(1 + Q(0))x$, has a density. Since our underlying queueing process is Markovian, the remaining time until $T(x)$ occurs is independent of this conditioned random variable.

Now we use the fact that if X and Y are two independent random variables and if X has a probability density, then so does $X + Y$. In fact the new density for this sum is the convolution of the density for X with the probability distribution for Y. □

If we use the notion of generalized functions, we can let $f_{T(x)}(t)$ denote the "density" of $T(x)$, where we may use delta functions to allow for the possibility of point mass distributions. Given our limit theorems, we know that this sojourn time distribution is asymptotically normal. This suggests that the actual sojourn time distribution is approximately bimodal with peaks about the values $(1 + Q(0))x$ and $T^{(0)}(x)$. Our density approximation formula is then

$$f_{T(x)}(t) \approx e^{-\beta(x)} \cdot \delta(t - (1 + Q(0))x) + (1 - e^{-\beta(x)}) \cdot \frac{1}{\sqrt{2\pi\, v(x)}} e^{-(t-m(x))^2/(2v(x))} \tag{5.2}$$

Table 1 Values for $\beta(x)$ used in Figure 4.

$\beta(x)$	$\lambda = 2.0$	$\lambda = 0.5$
$x = 0.10536$ (10th percentile job size)	3.37152	1.63308
$x = 0.69315$ (50th percentile job size)	22.1808	10.743825
$x = 2.30259$ (90th percentile job size)	73.68288	35.690145

where $\delta(\cdot)$ is the delta function, $m(x) \equiv T^{(0)}(x)$, $v(x) \equiv \mathsf{Var}\left[T^{(1)}(x)\right]$ and

$$\beta(x) \equiv \int_0^{(1+Q(0))x} \lambda(s)\, ds + Q(0) \int_0^x \mu(s)\, ds. \tag{5.3}$$

For the constant rate case, we have $\beta(x) = (\lambda \cdot (Q(0) + 1) + \mu \cdot Q(0))x$.

For the graphs of our numerical examples in Figure 4, we set $\mu = 1$, $Q(0) = 10.0$ and either $\lambda = 2.0$ or $\lambda = 0.5$. Given an exponentially distributed service time with mean one, we have $x = 0.10536$ for the 10 percentile job size, $x = 0.69315$ for the 50th percentile job size and $x = 2.30259$ for the 90th percentile job size. The values for $\beta(x)$ are given by Table 1.

In Figure 4 the simulated densities for the 10, 50 and 90-th percentile job sizes are compared with their respective normal approximations. The left column of Figure 4 displays this comparison for an overloaded $M/M/1/PS$ with arrival rate $\lambda = 2$, while the right column considers an underloaded $M/M/1/PS$ with arrival rate $\lambda = 0.5$.

The bimodal phenomenon discussed earlier in this section is readily apparent in the two graphs for the 10 percentile job size. Here, the point masses occur with probabilities 0.034 and 0.20. The bimodality is observable here because $\beta(x)$ is sufficiently small. Since β increases linearly in x, the point mass probability decreases exponentially. Thus the bimodal behavior of the sojourn times of the 50-th and 90-th job size percentiles are *not* observable, since their point mass probabilities equal 2.33×10^{-10} (50%-tile,left), 2.17×10^{-5} (50%-tile,right), 9.83×10^{-33} (90%-tile,left), and 3.13×10^{-16} (90%-tile,right).

6. Conclusion

We introduce the notion of a virtual customer for an approximate analysis of the sojourn time for a processor sharing queue. This creates a virtual job of a known size that is affected by the other jobs in the queue, but does *not* affect the response times of those other jobs.

We extend previous asymptotic results for the $M_t/M_t/1/PS$ queueing process to accommodate a non-zero, scaled initial load. These results can then be transformed into fluid and diffusion limits for the sojourn times. These sojourn time formulas hold in general for any queueing process that has fluid and diffusion limits.

Our uniform acceleration scaling gives a simpler analysis of the mean and variance of the sojourn time, yet yields results that are a good approximation of the original stochastic model. We then obtain a time-varying analysis of the response time for systems that may experience alternating periods of underloading and overloading. Our numerical examples of overload behavior show that our approximations work well for a wide range of virtual job sizes. This type of behavior cannot be captured by steady-state models.

The density of the virtual response time is found to be well approximated by the convolution of a normal density and a point mass. Guided by simulation of the sojourn times of the $M_t/M_t/1/PS$ queue, we show that the sojourn time density may sometimes have a bimodal property. Even though we have convergence to the normal distribution, in practice the point mass contribution may not decay to zero quickly enough. In our numerical examples, we see when such a point mass term can contribute to our approximation of the density.

In future work we will extend this fluid and diffusion analysis to sojourn times for heterogeneous classes of customers with general job size (service) distributions, customer abandonment (having jobs "time out" after a given amount of time), and weighted processor sharing.

References

1. N. Bansal and M. Harchol-Balter, Analysis of SRPT scheduling: Investigating unfairness, in: *Proceedings of Association of Computing Machinery Sigmetrics 2001 Conference on Measurement and Modeling of Computer Systems*, Cambridge, MA, (2001).
2. T. Bonald and A. Proutière, On performance bounds for the integration of elastic and adaptive streaming flows, in: *Proceedings of Association of Computing Machinery Sigmetrics/Performance Joint International Conference on Measurement and Modeling of Computer Systems*, (2004) 235–245.
3. H. Chen, O. Kella, and G. Weiss, Fluid approximations for a processor sharing queue, Queueing Systems: Theory and Applications, 27 (1997) 99–125.
4. E.G. Coffman, R.R. Muntz, and H. Trotter, Waiting time distribution for processor-sharing systems, Journal of the Association of Computing Machinery 17 (1970) 123–130.
5. F. Delcoigne, A. Proutière, and G. Règniè, Modeling integration of streaming and data traffic, Performance Evaluation 55 (2004) 185–209.
6. M. Harchol-Balter, B. Schroeder, N. Bansal, and M. Agrawal, Size-based scheduling to improve web performance, Association of Computing Machinery Transactions on Computer Systems 21(2) (2003) 207–233.
7. F. Guillemin and J. Boyer, Analysis of the $M/M/1$ queue with processor sharing via spectral theory, Queueing Systems 39(4) (2001) 377–397.
8. A. Jean-Marie and P. Robert, On the transient behavior of the processor sharing queue, Queueing Systems 17 (1994) 129–136.
9. M. Yu. Kitaev, The M/G/1 processor-sharing model: Transient behavior, Queueing Systems 14 (1993) 239–273.
10. L. Kleinrock. *Queueing Systems, Volume II: Computer Applications* (John Wiley & Sons, 1976).
11. J.D.C. Little, A proof of the queueing formula $L = \lambda W$, Operations Research 9 (1961) 383–387.
12. A. Mandelbaum and W.A. Massey, Strong approximations for time dependent queues, Mathematics of Operations Research 20(1) (1995) 33–64.
13. W.A. Massey, Asymptotic analysis of the time dependent M/M/1 queue, Mathematics of Operations Research 10 (1985) 305–327.
14. H. Masuyama and T. Takine, Sojourn time distribution in a MAP/M/1 processor-sharing queue, Operations Research Letters 31(5) (2003) 406–412.
15. J. Morrison, Response time for a processor-sharing system, SIAM J. Appl. Math. 45(1) (1985) 152–167.
16. R. Nunez-Queija, Processor sharing models for integrated services networks, Ph.D. Thesis Eindhoven University of Technology (2000).
17. T. Ott, The sojourn time distribution in the M/G/1 queue with processor sharing, J. Appl. Prob. 21 (1984) 360–378.
18. J.W. Roberts, Engineering for quality of service, in: K. Park and W. Willinger (eds.), *Self-Similar Network Traffic and Performance Evaluation* (Wiley, New York, 2000) pp. 401–420.
19. R. Schassberger, A new approach to the M/G/1 processor sharing queue, Adv. Appl. Prob. 16 (1984) 202–213.
20. B. Schroeder and M. Harchol-Balter, Web servers under overload: How scheduling can help, *18th International Teletraffic Congress* (Berlin, Germany, 2003).
21. B. Sengupta and D.L. Jagerman, A conditional response time of the M/M/1 processor-sharing queue, AT& T Technical Journal 64 (1985) 409–421.
22. S.F. Yashkov, A derivation of response time distribution for a M/G/1 processor-sharing queue, Problems of Control and Information Theory 12(2) (1983) 133–148.
23. S.F. Yashkov, Processor-sharing queues: Some progress in analysis, Queueing Systems 2 (1987) 1–17.
24. S.F. Yashkov, Mathematical problems in the theory of shared-processor systems, Journal of Soviet Mathematics 58 (1992) 101–147.
25. B. Zwart and O.J. Boxma, Sojourn time asymptotics in the M/G/1 processor sharing queue, Queueing Systems 35 (2000) 141–166.