On Saturation Effects in Coupled Speed Scaling

Maryam Elahi Carey Williamson

Department of Computer Science, University of Calgary

Abstract. In coupled speed scaling systems, the speed of the CPU is adjusted dynamically based on the number of jobs present in the system. In this paper, we use Markov chain analysis to study the autoscaling properties of an M/GI/1/PS system. In particular, we study the saturation behaviour of the system under heavy load. Our analytical results show that the mean and variance of system occupancy are not only finite, but tightly bounded by polynomial functions of the system load and the speed scaling exponent. We build upon these results to study the speed, utilization, and mean busy period of the M/GI/1/PS. Discrete-event simulation results confirm the accuracy of our analytical models.

1 Introduction

Coupled speed scaling systems adjust the CPU speed dynamically based on the number of jobs in the system. These dynamic speed scaling systems provide tradeoffs between response time and energy consumption [1, 2]. Specifically, running the CPU faster improves the response time, but consumes more energy.

Within a speed scaling system, the two most important considerations are the scheduler and the speed scaling function. The scheduler determines which job is executed next, and the speed scaling function determines the speed at which that job is executed. A popular approach for the latter is *job-count-based* speed scaling, in which the service rate is a function of the current system occupancy [3, 8, 22]. We refer to this as *coupled* speed scaling, since the service rate is coupled to the system occupancy.

In this paper, we focus on the autoscaling properties of coupled speed scaling systems under heavy load. In particular, we consider sustained offered loads that drive the system toward *saturation*, in which the utilization becomes arbitrarily close to unity (i.e., $U \rightarrow 1$). Note that if there is no limit to the maximum service rate, then the system will automatically adjust (i.e., autoscale) its service rate to accommodate whatever load is presented to it.

Our current paper is motivated by some of our own prior work on the autoscaling properties of coupled speed scaling systems [10]. In particular, our prior work used discrete-event simulation to show that the mean system occupancy remained finite in coupled speed scaling systems under heavy load (see Figure 1). Furthermore, the mean occupancy was estimated as $E[N] \approx \rho^{\alpha}$ [10].

In our current paper, we present analytical results that bound the mean and variance of occupancy under heavy load. Specifically, we use Markov chain analysis to study the dynamics of a Processor Sharing (PS) speed scaling system,

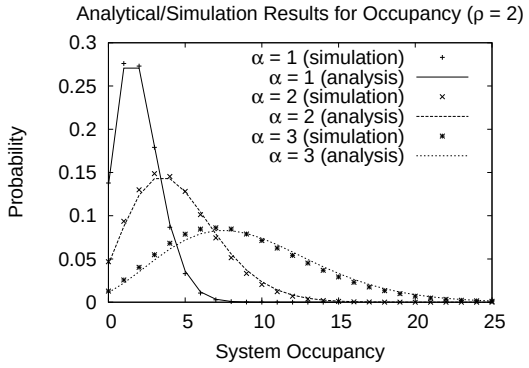


Fig. 1. Distribution of System Occupancy (based on Figure 2b in [10])

and derive tight bounds on system performance. We then extend our model to analyze the mean busy period for PS under coupled speed scaling. Finally, we use discrete-event simulation to verify the accuracy of our analytic model, and to extend our observations to Shortest Remaining Processing Time (SRPT) systems.

The main insights from our work are the following. First, we show that the mean and variance of the system speed are bounded, even under heavy (but finite) offered load. Second, we show that the mean and variance of system occupancy are tightly bounded, and are polynomial functions of ρ and α. Third, we show that the mean busy period in a PS-based coupled speed scaling system grows at least exponentially with offered load. Finally, we show that the mean busy period for an SRPT-based system grows much faster than that for the corresponding PS-based system.

The rest of this paper is organized as follows. Section 2 reviews prior literature on speed scaling systems. Section 3 presents our system model. Section 4 presents our analytical and numerical results. Section 5 presents simulation results. Finally, Section 6 concludes the paper.

2 Background and Related Work

Prior research on speed scaling systems appears in two different research communities: theory and systems. Theoretical work typically focuses on the optimality of speed scaling systems under some simplifying assumptions (e.g., unbounded service rates, known job sizes). Systems work typically focuses on "good" solutions, rather than optimal ones [7, 8], and especially those that are robust to unknown job sizes, scheduling overheads, as well as finite and discrete system

speeds. In this literature review, we focus primarily on the theoretical work as relevant background context for our paper.

In speed scaling systems, there are many tradeoffs between service rate, response time, and energy consumption. Yao *et al.* [24] analyzed dynamic speed scaling systems in which jobs have explicit deadlines, and the service rate is unbounded. Bansal *et al.* [5] considered an alternative approach that minimizes system response time, within a fixed energy budget. Others have focused on finding the optimal fixed rate at which to serve jobs in a system with dynamically-settable speeds [11, 22, 23].

Several studies indicate that energy-proportional speed scaling is nearly optimal [3, 6]. In this model, the power consumption $P(s)$ of the system depends only on the speed s, which itself depends on the number of jobs in the system. Bansal, Chan, and Pruhs [6] showed that SRPT with the speed scaling function $P^{-1}(n+1)$ is 3-competitive for an arbitrary power function P. Andrew *et al.* [3] showed that the optimal policy is SRPT with a job-count-based speed scaling function of the form $s = P^{-1}(n\beta)$.

Fairness in dynamic speed scaling systems is also an important consideration. In particular, speed scaling systems induce tradeoffs between fairness, robustness, and optimality [3]. PS is always fair, providing the same expected slowdown for all jobs, even under speed scaling. However, the unfairness of SRPT is magnified under speed scaling, since large jobs tend to run only when the system is nearly empty, and hence at lower speeds. While PS is good for fairness, it is suboptimal for both response time and energy [3].

3 System Model

3.1 Model Overview and Assumptions

We consider a single-server system with dynamically adjustable service rates. Service rates are changed only when the system occupancy changes (i.e., at job arrival and departure points). There is no cost incurred for changing the service rate, and no limit on the maximum possible service rate. (Prior work by others has considered bounded service rates [11].)

The workload presented to the server is a sequence of jobs with random arrival times and sizes. We assume that the arrival process is Poisson, with mean arrival rate λ. The size (work) of a job represents the time it takes to complete the job when the service rate is $\mu = 1$. We assume that job sizes are exponentially distributed and independent. Unless stated otherwise, we assume that the mean job size is $E[X] = 1$. Table 1 summarizes our model notation.

In this paper, we consider two specific work-conserving scheduling policies, namely PS and SRPT. PS shares the CPU service rate equally amongst all jobs present in the system, while SRPT works exclusively on the job with the least remaining work. We assume that the schedulers know all job sizes upon arrival, or can at least estimate them dynamically [8]. A job in execution may be preempted and later resumed without any context-switching overhead.

Table 1. Model Notation

Symbol	Description
λ	Mean job arrival rate
μ	Service rate
μ_n	Service rate in state n
$E[X]$	Average size (work) for each job
ρ	Offered load $\rho = \lambda/\mu = \lambda E[X]$
p_n	Steady-state probability of n jobs in the system (a.k.a. $\pi(n)$)
U	System utilization $U = 1 - p_0$
n	Number of jobs
$\phi(n)$	CPU speed as a function of number of jobs
t	Time in seconds
$n(t)$	Number of jobs in system at time t
$s(t)$	CPU speed at time t
$P(s)$	Power consumption when running at speed s
α	Exponent in power consumption function $P(s) = s^\alpha$

A speed scaling function, $s(t)$, specifies the speed of the system at time t. For coupled speed scaling, the speed at time t depends on the number of jobs in the system, denoted by $n(t)$, and thus is influenced by the scheduling policy. The best known policy uses the speed function $s(t) = P^{-1}(n(t)\beta)$ [3]. In this paper, we assume $\beta = 1$. We also consider $P(s) = s^\alpha$, which is commonly used in the literature to model the power consumption of the CPU. Therefore, in the coupled speed scaling model, we use $s(t) = \sqrt[\alpha]{n(t)} = n(t)^{1/\alpha}$, where $\alpha \geq 1$. When time t is not relevant, we use $\phi(n)$ to denote the CPU speed for n jobs.

In our work, we focus on the PS scheduling policy, which is an example of a symmetric scheduling policy [12]. Such policies do not prioritize based on job size, or any other job trait, but merely treat all arrivals equivalently. Symmetric policies have the important property that their departure process is stochastically identical to their arrival process when time is reversed. Therefore, in the M/GI/1 model, where arrivals are Poisson, the queue occupancy states form a birth-death process regardless of the form of the job size distribution. This result is formalized in the following theorem, the proof of which is given in [12]. A proof for the special case of PS scheduling appears in [16].

Theorem 1 [12]. In an M/GI/1 queue with a symmetric scheduling policy, the limiting probability that the queue contains n jobs is:

$$\pi(n) = \frac{\rho^n}{\prod_{i=0}^{\infty} \phi(n)} \pi(0), \quad \text{for } n > 0,$$

where the probability $\pi(0)$ of the system being empty is given by:

$$\pi(0) = \frac{1}{1 + \sum_{n=1}^{\infty} \frac{\rho^n}{\prod_{i=1}^{n} \phi(i)}}.$$

Theorem 1 indicates that all symmetric polices have the same occupancy distribution for the same $\phi(n)$ function. Furthermore, this occupancy distribution is insensitive to the job size distribution, and depends only on the mean job size.

Although FCFS is not a symmetric policy, it is interesting to note that the occupancy distribution for M/M/1 FCFS is equivalent to the occupancy distribution for M/GI/1 symmetric policies with $\phi(n) = 1$. In fact, the occupancy distribution under all non-size-based policies is equivalent for a general $\phi(n)$ [12, 23]. Therefore, in a single-server with some $\phi(n)$ speed-scaling discipline, in order to study the occupancy distribution under M/GI/1 PS, it suffices to study the occupancy distribution under M/M/1 FCFS. In our work, we consider the special case of $\phi(n) = n^{1/\alpha}$, and derive results for the average speed, occupancy, and expected busy period length.

3.2 Markov Chain Model

We consider the dynamics of a system with sub-linear speed scaling. Specifically, we consider running the system at speed $s = n^{1/\alpha}$ when the system occupancy is n jobs. We consider $1 \leq \alpha \leq 3$, which is the relevant range of interest for Dynamic Voltage and Frequency Scaling (DVFS) on modern processors [20, 23]. Note that α need not be an integer, but is treated as such in the discussion.

The parameter α determines the set of distinct speeds available in our speed scaling system. For the special case $\alpha = 1$, the speeds scale linearly with occupancy, much like the M/M/∞ queue, which provides a natural validation point for our model. For $\alpha = 2$, speeds scale less than linearly with system occupancy, following the "square root speed scaling" approach recommended in the literature (i.e., the system speed when there are n jobs in the system is $\sqrt{n} = n^{1/2}$). For $\alpha = 3$, speeds scale even more slowly with growing system occupancy: the system speed when there are n jobs in the system is $\sqrt[3]{n} = n^{1/3}$). In the limiting case of $\alpha = \infty$, the speeds scale so slowly that they are effectively constant (i.e., single-speed system). This provides another validation point for our model.

Figure 2 shows the Markov chain for our speed scaling system. The key difference from Kleinrock's classic M/M/∞ model is the change in the service rates $\mu_n = n^{1/\alpha}\mu$. Analysis of this chain produces steady-state probabilities p_n that are analogous to those for the M/M/∞ chain, except for the effect of the $1/\alpha$ exponent on all of the service rates.

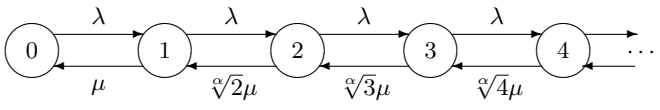


Fig. 2. Markov Chain for Coupled Speed Scaling System Model

4 Analytical and Numerical Results

In this section, we consider the M/M/1 queue with FCFS scheduling and $\phi(n)$-coupled speed scaling, where $\phi(n) = n^{1/\alpha}$ for $\alpha \geq 1$.

In the context of the "dynamic service rate" control problem, the M/M/1 FCFS queue with adjustable service rates has been studied in the literature [4, 11, 14, 23], and elegant results for state-dependent speeds that optimize the linear combination of average occupancy and average energy consumption are presented in [11, 23]. However, the proof for the formulation of the occupancy distribution is not provided explicitly. For the sake of completeness, we briefly discuss here the special case of $n^{1/\alpha}$-coupled speed-scaling systems.

Consider an $n^{1/\alpha}$-coupled speed-scaling system for some $\alpha > 0$, and with non-preemptive, non-size-based scheduling. Assume inter-arrival times are exponentially distributed with rate λ, and job sizes are exponentially distributed with rate μ. Let $\rho = \lambda/\mu$.

In this system, the queue occupancy evolves as a birth-death process since the time between transitions is exponentially distributed. The CTMC for this model is similar to the single-speed M/M/∞ in that transitions between states occur upon state-independent arrivals with rate λ, and state-dependent departures with rates μ_n. Unlike the M/M/∞, however, this is a single server model, with at most one job in service at any point in time. When in state $n > 0$, provided that no arrival occurs, the time to the next departure is the remaining work of the job in service divided by the service rate $n^{1/\alpha}$. Since the service requirements are exponentially distributed with rate μ, and the exponential distribution is closed under scaling by a positive factor, the time until the next departure is also exponentially distributed with rate $\mu_n = \mu n^{1/\alpha}$. Therefore, the queue occupancy forms a birth-death process, and the limiting probabilities (if they exist) are:

$$\pi(n) = \pi(0) \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}} = \pi(0) \prod_{i=0}^{n-1} \frac{\lambda/\mu}{(i+1)^{1/\alpha}} = \pi(0) \frac{\rho^n}{(n!)^{1/\alpha}}, \quad \text{for } n > 0,$$

where:

$$\pi(0) = \frac{1}{\sum_{i=0}^{\infty} \frac{\lambda_0 \lambda_1 \ldots \lambda_{i-1}}{\mu_1 \mu_2 \ldots \mu_i}} = \frac{1}{\sum_{i=0}^{\infty} \frac{\rho^i}{(i!)^{1/\alpha}}}.$$

To show that the limiting probabilities exist, and that the chain is ergodic, it suffices to show that the infinite sums converge. Based on the ratio test for convergence of an infinite series, the series $\sum_{i=0}^{\infty} a_n$ converges if $\lim_{n\to\infty} |\frac{a_{n+1}}{a_n}| < 1$. This condition holds in our case, since $\alpha > 0$ and $\mu > 0$. Specifically,

$$\lim_{n\to\infty} \frac{\rho^{n+1}/(n+1)!^{1/\alpha}}{\rho^n/(n!)^{1/\alpha}} = \lim_{n\to\infty} \frac{\rho}{(n+1)^{1/\alpha}} < 1.$$

Note that our speed scaling system is just a special case of Theorem 1 with $\phi(n) = n^{1/\alpha}$. Therefore, M/GI/1 queues with $n^{1/\alpha}$-coupled speed-scaling and with symmetric scheduling policies, including PS, have the same occupancy distribution as M/M/1 FCFS with $n^{1/\alpha}$-coupled speed-scaling.

Unfortunately, we do not have closed form expressions for the foregoing steady-state probabilities. However, it is possible to numerically evaluate the mean and higher moments of the occupancy distribution. In fact, we can derive bounds for the mean and variance of the occupancy distribution (see Section 4.2). In the remainder of this section, we make a few observations about the shape of the occupancy distribution.

The steady-state probability distribution in our system is a function of the average load ρ and the speed-scaling parameter α. Recall that ρ is a function of the arrival rate λ, and the job size based on rate μ. Note that increasing or decreasing the arrival rate, while adjusting the average job size to keep the average load constant, results in the same occupancy distribution. The parameter α determines the set of distinct speeds available in the coupled speed-scaling system. For the special case $\alpha = 1$, the speeds scale linearly with occupancy, similar to the M/M/∞ queue. For $\alpha > 1$, speeds scale sub-linearly with system occupancy. For very large α, the speeds scale so slowly that the system effectively behaves like a single-speed system.

The parameter α has three main impacts on the occupancy distribution, as illustrated in prior work [10]. The first effect of increasing α is to shift the occupancy distribution to the right (see Figure 1). This is intuitively expected, since the slower service rates lead to a larger queue of jobs in the system. However, as the backlog of jobs grows, the service rate is also increased, which eventually stabilizes the system. This pendulum effect keeps the mode of the occupancy distribution close to ρ^α, which determines the average speed (ρ) required to serve the load arriving to the system. The second effect of $\alpha > 1$ is the distortion of the Poisson distribution observed for system occupancy when $\alpha = 1$. While the structure of the distribution is similar to Poisson, the state probabilities degenerate, and the Coefficient of Variation (CoV) is greater than that for a Poisson distribution. The particular relationship observed is $Var[N] \approx \alpha E[N]$ (see Figure 4(b) for graphical evidence of this observation). In the limiting case of $\alpha \to \infty$, this distribution degenerates to an equal but negligible probability for all states, indicating an unstable (infinite) queue. The third effect that we observe when increasing α is the decline in $\pi(0)$, which is the probability of having an idle system. We call this the *saturation effect*, which is our main focus in this paper. We explore the effect of α on the utilization, and the expected busy period length, in Sections 4.3 and 4.4, respectively.

4.1 Mean and Variance of Speed

We first establish some fundamental results regarding the mean and variance of the system speed for coupled speed scaling systems. Let random variables S and N denote the speed of the server and the system occupancy, respectively. In the $n^{1/\alpha}$-coupled speed-scaling system, $S = N^{1/\alpha}$.

Theorem 2. In an M/M/1 queue with $n^{1/\alpha}$-coupled speed-scaling, $E[S] = \rho$. Furthermore, $Var[S] < 1$ for any $\alpha \geq 2$.

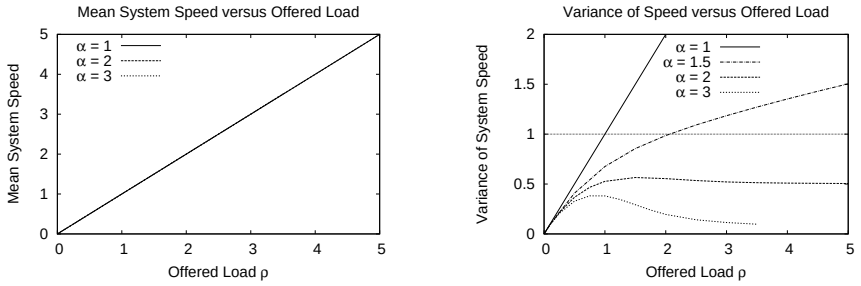
The proof of the first half of Theorem 2 is fairly straightforward, based on the definition of $E[S]$. Intuitively, we expect the steady-state average speed to

be equal to the incoming load for any stable system (assuming the speed is 0 when there are no jobs in the system). In [23], the general lower bound for the time average speed in all stable speed-scaling systems is argued to be $\bar{S} \geq \rho$. Furthermore, $\bar{S} = \rho$ for systems that run at speed 0 when the system is empty. Our result involves a stochastic proof for the special case of M/M/1 with $n^{1/\alpha}$-coupled speed-scaling [9].

The proof for the second part of Theorem 2 is a bit more involved, but we sketch it here. The essence is to use the fact that $Var[S] = E[S^2] - E[S]^2$, and to derive a bound on $E[S^2]$, since $E[S]^2 = \rho^2$. The algebraic derivation culminates in $Var[S] \leq 1 - \pi(0) < 1$ (see Chapter 4 of [9] for details).

Figure 3 shows the effects of α and ρ on the system speed. Figure 3(a) shows the mean speed, which scales linearly with ρ, regardless of the value of α. Figure 3(b) shows the variance of speed. When $\alpha = 1$, $Var[S] = \rho$, since the speeds follow the Poisson distribution with rate ρ (analagous to occupancy). For $\alpha \geq 2$, the variance is always less than unity. The theoretical bound (indicated by the horizontal line) is not especially tight, but it is a provable bound [9].

For $\alpha \geq 2$, the variance of speed initially increases with ρ up to a point, before decreasing and seemingly converging to a value well below 1. The intuition behind this result is that for larger α, the speed changes are quite gradual, especially when the occupancy is high. Thus the variance of the speed remains low even when the occupancy fluctuates a lot.



(a) Mean System Speed

(b) Variance of System Speed

Fig. 3. Analytical Results for System Speed in Coupled Speed Scaling Systems

4.2 Mean and Variance of Occupancy

We next establish results concerning the mean and the variance of system occupancy. The following Theorem 3 shows that the upper bound for occupancy exceeds the lower bound by at most a polynomial function of α, which we specify

in Definition 1 below. Furthermore, Theorem 3 provides an upper bound on the variance of system occupancy.

Definition 1. Let $f : \mathbb{N} \to \mathbb{N}$ be $f(\alpha) = \alpha - 1$ for $\alpha \in \{1, 2, 3\}$ and $f(\alpha) = \alpha(\alpha - 1)$ for $\alpha \geq 4$.

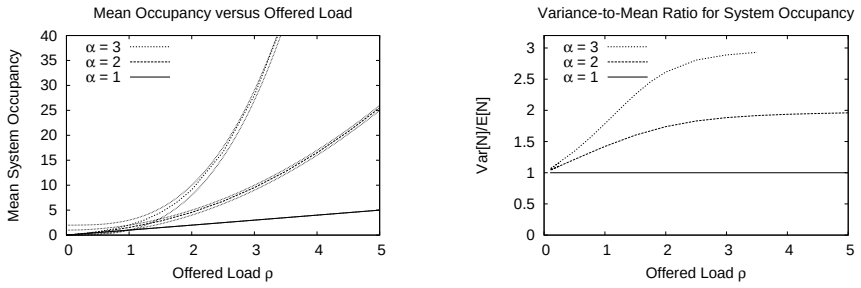
Theorem 3. Consider an M/M/1 system with $n^{1/\alpha}$-coupled speed-scaling, where $\alpha \in \mathbb{N}$. For $f(\alpha)$ as defined above, the mean system occupancy $E[N]$ satisfies:

$$\rho^\alpha \leq E[N] \leq \rho^\alpha + f(\alpha).$$

Furthermore, for $\alpha \geq 2$, $Var[N] \leq E[N](f(\alpha) + 2\alpha - 1)$.

Proof. See Chapter 4 of [9] for details.

Figure 4 shows the effects of α and ρ on the mean and variance of occupancy. For mean occupancy, the bounds are quite tight, as shown in Figure 4(a). Furthermore, these $E[N]$ values can be used in conjunction with Little's Law to determine the mean time T in the system. For $\alpha = 1$, the occupancy distribution is Poisson, so the mean and variance are both equal to ρ. Unlike the variance of speed, which is less than 1 for $\alpha \geq 2$, the variance of occupancy is an increasing function of the average load (since $E[N]$ increases with load). Numerical and simulation results show that, under heavy load, $Var[N] \approx \alpha E[N]$, which is even less than the given bound. Figure 4(b) illustrates this phenomenon, by plotting the variance-to-mean ratio for system occupancy as load is increased.



(a) Mean System Occupancy (b) Variance-to-Mean Ratio

Fig. 4. Analytical Results for System Occupancy in Coupled Speed Scaling Systems

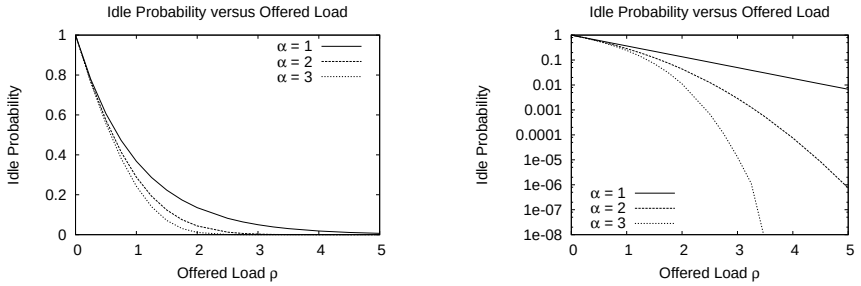
4.3 System Saturation

We next explore the saturation effect, in which system utilization U approaches unity. Conversely, the probability $\pi(0)$ of an empty system approaches zero.

Figure 5 shows $\pi(0)$ as a function of load, both on a linear scale (Figure 5(a)) and a logarithmic scale (Figure 5(b)). We see that with an increase in α, the

probability of the idle state decreases quickly. Recall that $U = 1 - \pi(0)$ is the utilization of the system. For $\rho > 1.6$, systems with $\alpha \geq 2$ are utilized more than 90% of the time, and for $\rho \geq 4$, the utilization exceeds 99.99% in these systems.

For an arbitrary small threshold $\epsilon > 0$, one can define a "saturation load" ρ at which $\pi(0) \leq \epsilon$. For example, when $\epsilon = 10^{-4}$, the saturation loads would be near 9.2 for $\alpha = 1$, 3.9 for $\alpha = 2$, and 2.75 for $\alpha = 3$. For $\epsilon = 10^{-6}$, the saturation loads would be 13.5 ($\alpha = 1$), 4.9 ($\alpha = 2$), and 3.2 ($\alpha = 3$).



(a) Idle Probability (linear scale) (b) Idle Probability (log scale)

Fig. 5. Analytical Results for Saturation in Coupled Speed Scaling Systems

Recall that in single-speed systems, the mean occupancy under M/M/1 (FCFS or PS) is $\frac{\rho}{1-\rho}$. Furthermore, the average load ρ is equal to the utilization U. Therefore, when utilization approaches 1, the mean occupancy under FCFS (equivalently under PS) grows very quickly. In coupled speed-scaling systems, however, M/M/1 FCFS (equivalently PS) with $n^{1/\alpha}$-coupled speed-scaling maintains robust performance even when the utilization is close to 1. That is, the mean occupancy is always polynomial in ρ with degree α (see Theorem 3).

4.4 Mean Busy Period

In this section, we analyze the expected busy period length under M/M/1 with $n^{1/\alpha}$-coupled speed-scaling. Recall that the length of a busy period, denoted by B, is defined to be the time from when the system becomes busy until the next time that all jobs have left the system, and the system becomes idle. The length of an idle period is denoted by I.

Our main result is that the mean busy period grows at least exponentially with ρ. We achieve this result by first establishing the following Theorem 4, and then focusing on Corollary 1.

Theorem 4. Consider M/M/1 with $n^{1/\alpha}$-coupled speed-scaling with load $\rho = \lambda/\mu$, where λ is the arrival rate and μ is the rate of the (exponential) job

size distribution. Then, the expected busy period length exists, and it satisfies:

$$E[B] = \frac{1}{\lambda} \sum_{i=1}^{\infty} \frac{\rho^i}{(i!)^{1/\alpha}}$$

Proof. In a birth-death process, it is known that B and I form an alternating renewal process, for which the following equality holds [16]:

$$U = \frac{E[B]}{E[B] + E[I]}$$

where U is the limiting probability of the system being busy (i.e., utilization). Therefore, the expected busy period length can be derived as a function of the expected idle period length and the utilization as follows:

$$E[B] = \frac{UE[I]}{1 - U}$$

Note that the length of the idle period is the time until the next arrival. Since the arrival process is Poisson with rate λ, $E[I] = 1/\lambda$. By definition, $U = 1-\pi(0)$. Based on Theorem 1, the system is ergodic, and $\pi(0) = \frac{1}{\sum_{i=0}^{\infty} \frac{\rho^i}{(i!)^{1/\alpha}}} > 0$. Therefore,

$$E[B] = \frac{1 - \pi(0)}{\pi(0)\lambda} = \frac{1}{\lambda}\left(\frac{1}{\pi(0)} - 1\right) = \frac{1}{\lambda}\left(\sum_{i=0}^{\infty} \frac{\rho^i}{(i!)^{1/\alpha}} - 1\right) = \frac{1}{\lambda}\sum_{i=1}^{\infty} \frac{\rho^i}{(i!)^{1/\alpha}}. \quad \square$$

Corollary 1. In an M/M/1 with $n^{1/\alpha}$-coupled speed-scaling, for any $\alpha \geq 1$, $E[B]$ satisfies:

$$E[B] \geq \frac{1}{\lambda}(e^{\rho} - 1)$$

Proof. It is known that $\sum_{i=0}^{\infty} \frac{\rho^i}{i!} = e^{\rho}$. The result then follows directly. $\square$

This corollary shows that the mean busy period grows at least exponentially with load ρ, as stated earlier. Note, however, that the busy period duration is sensitive to *both* the arrival rate and the average load, while U is only a function of the average load.

Figure 6 illustrates the effects of α and ρ on the expected busy period length. There are three pairs of lines in this graph, corresponding to $\alpha = 3$ (highest pair), $\alpha = 2$ (middle pair), and $\alpha = 1$ (lowest pair), respectively. Within each pair of lines, the flatter one shows the expected busy period length when the load is changed via the arrival rate (i.e., $\rho = \lambda$, since $\mu = 1$), while the steeper line shows the expected busy period length when the load is changed via the mean of the job size distribution (i.e. $\rho = E[X] = 1/\mu$, since $\lambda = 1$). As expected, the trend is similar in both cases, with the lines differing by a factor of λ (note the logarithmic vertical scale on the graph). The lines also differ at the leftmost edge of the graph, since $E[B] \approx E[X]$ when the load is very light (i.e., $\rho << 1$).

The lower bound given by Corollary 1 is for the general case of $\alpha \geq 1$. It is tight for $\alpha = 1$ (see LB points in Figure 6), but very loose for $\alpha \geq 2$, for which $E[B]$ grows much faster than e^{ρ}, since the system saturates sooner, and much more dramatically. (It is more like $e^{\rho^{\alpha}/\alpha}$, but we have no proof for this yet).

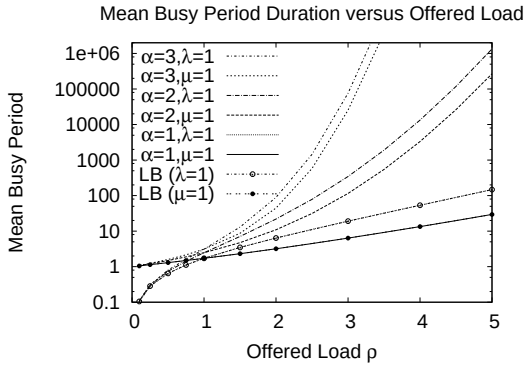


Fig. 6. Analytical Results for Busy Period in Coupled Speed Scaling Systems

5 Simulation Results

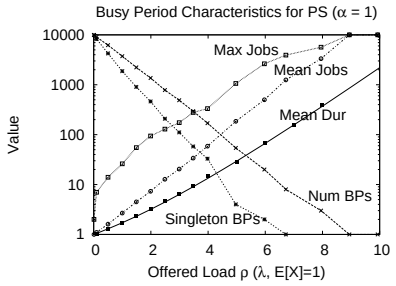
In this section, we use discrete-event simulation to explore saturation effects in coupled speed scaling systems. Our simulator supports different schedulers (e.g., FCFS, PS, SRPT) and speed scaling functions (e.g., coupled, decoupled), and reports results for speeds, response times, energy, and busy period structure [19]. We use this simulator to study the autoscaling dynamics of PS and SRPT.

In our first experiment, we use our simulator to explore the busy period structure of PS-based speed scaling systems. As the load offered to a speed scaling system is increased, the number of busy periods diminishes until there is a single massive busy period that includes all jobs. We refer to this phenomenon as *saturation*, since $U \to 1$.

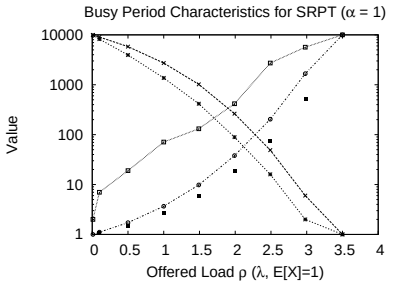
Figure 7(a) illustrates the saturation effect, based on simulation of a PS-based system with linear speed scaling (i.e., $\alpha = 1$). The horizontal axis shows the offered load based on the arrival rate λ, assuming that the mean job size $E[X] = 1$, while the vertical axis shows the value of different busy period metrics, on a logarithmic scale.

The downward-sloping diagonal line on the graph shows the number of busy periods observed, in a simulation run with a total of 10,000 jobs. Furthermore, the dashed line just beneath it shows the number of busy periods that have only a single job. At light load, there are thousands of busy periods, and most have just a single job. As the load increases, the number of busy periods decreases, as does the number of singleton busy periods. The straight-line behaviour on this log-linear plot indicates exponential decline, consistent with the mathematical model.

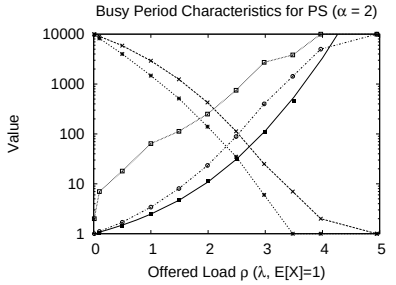
In Figure 7(a), the upward-sloping dotted line shows the average number of jobs per busy period, while the line above it shows the maximum number of



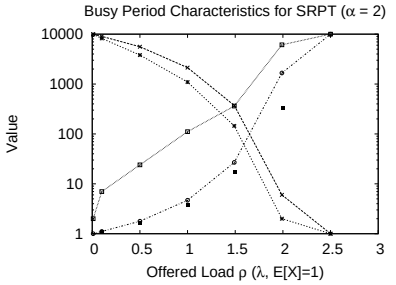
(a) PS ($\alpha = 1$)



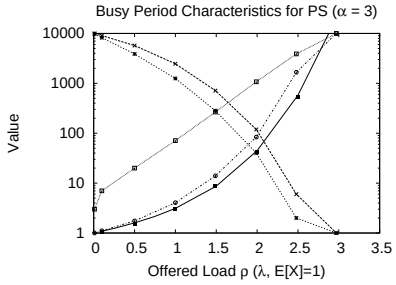
(d) SRPT ($\alpha = 1$)



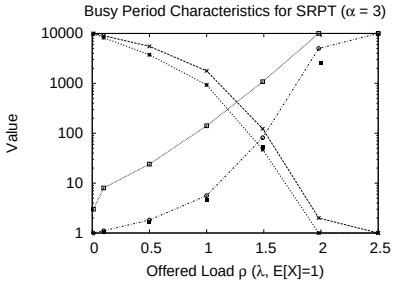
(b) PS ($\alpha = 2$)



(e) SRPT ($\alpha = 2$)



(c) PS ($\alpha = 3$)



(f) SRPT ($\alpha = 3$)

Fig. 7. Busy Period Characteristics for PS and SRPT Scheduling (simulation)

jobs observed in any of the busy periods seen. Both of these lines increase with load, and asymptotically approach a limit that reflects a single massive busy period containing all of the jobs. For $\alpha = 1$, this limit is near $\lambda = 9$, though the simulation results are somewhat noisy near this point.

Analytically, from the Poisson distribution, we know that $p_0 = e^{-\lambda}$ when $\alpha = 1$. For some suitably chosen small $\epsilon > 0$, this formula can be used to determine the load λ at which $p_0 \leq \epsilon$. For example, for $\epsilon = 0.0001$, solving $\lambda = -ln(\epsilon)$ yields $\lambda = 9.2$, which closely matches the simulation results.

The solid line in Figure 7(a) shows the mean busy period duration calculated using our analytical result from Theorem 4, while the black squares show the simulation results. The close agreement provides validation for our model.

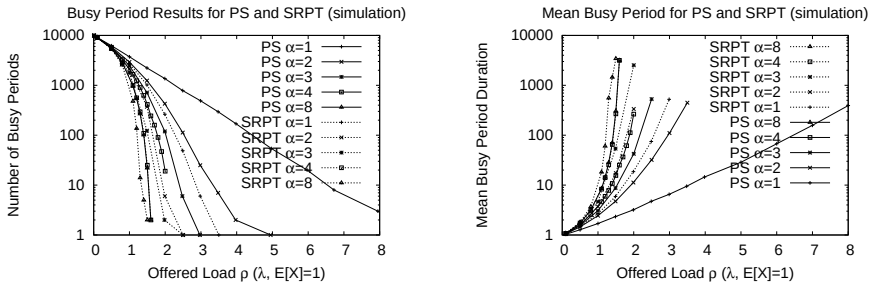
Figures 7(b) and (c) show the results for PS when $\alpha = 2$ and $\alpha = 3$, respectively. Note that the horizontal scales of these graphs differ from those in Figure 7(a). The busy period dynamics in these graphs are structurally similar to Figure 7(a), wherein light load has many very small busy periods, while heavier loads have fewer and larger busy periods. The primary differences from Figure 7(a) are the distinct downward curvature for the lines showing the number of busy periods, implying a decrease that is faster than exponential as load is increased. Furthermore, the point at which saturation occurs, as load is increased, arises sooner when α is larger. For example, the load levels at which saturation occurs in the simulation are $\lambda = 5$ for $\alpha = 2$, and $\lambda = 3$ for $\alpha = 3$. These closely match the predictions from our saturation analysis.

Despite the saturation of the utilization U, the speed scaling system still remains stable, even if the load is further increased. The probability of the system returning to the empty state becomes very small, but the system is still recurrent.

The right-hand side of Figure 7 shows the busy period results from our second set of simulation experiments, for an SRPT-based speed scaling system. Note that on each row of graphs, the horizontal scales for PS and SRPT plots differ.

The main observation here is that the saturation load for SRPT is different than under PS scheduling. In particular, the SRPT system saturates sooner. One implication of this observation is that there exist load levels at which SRPT is beyond saturation, while PS is not. In such scenarios, there will be significant unfairness for large jobs under SRPT (i.e., starvation). That is, while the average number of jobs in the system is the same for both PS and SRPT, the SRPT system tends to retain the largest jobs, causing anomalously high response times [10].

Another anecdotal observation from the simulation results is that the busy period structure for an SRPT system with speed scaling exponent α is qualitatively similar to that for a PS system with speed scaling exponent 2α (at least over the range of parameters considered here). Figure 8 illustrates this result, both for the number of busy periods in Figure 8(a), and the busy period duration in Figure 8(b). Furthermore, the saturation point in Figure 8(a) asymptotically approaches 1 (as expected) when α is increased from 1 to 8.



(a) PS ($\alpha = 1$) (d) SRPT ($\alpha = 1$)

Fig. 8. Busy Period Comparison for PS and SRPT Scheduling (simulation)

6 Conclusions

In this paper, we have used mathematical analysis and simulation to explore the autoscaling properties of dynamic speed scaling systems. We have assumed coupled (i.e., job-count-based) speed scaling, with PS as a representative symmetric scheduler. We focus particularly on heavy loads that cause the system to approach saturation (i.e., $U \to 1$).

The main conclusions from our work are the following. First, the mean and variance of the system speed are bounded, as long as the offered load is finite. Second, the mean and variance of system occupancy are tightly bounded by polynomial functions of ρ and α. Third, the mean busy period in a PS-based coupled speed scaling system grows at least exponentially with offered load when $\alpha = 1$, and even faster than this when $\alpha > 1$. Finally, we show that SRPT-based systems saturate sooner than the corresponding PS-based system. While such a system remains stable (in terms of job occupancy), it can manifest extreme unfairness due to starvation of the largest jobs.

Our ongoing work is exploring tighter bounds for the mean busy period in both PS and SRPT systems.

Acknowledgements

The authors thank the QEST 2018 reviewers for their contructive and insightful comments on an earlier version of this paper, and Philipp Woelfel for many hours discussing and analyzing proofs. Financial support for this work was provided by Canada's Natural Sciences and Engineering Research Council (NSERC).

References

1. S. Albers, F. Mueller, and S. Schmelzer, "Speed Scaling on Parallel Processors", *Proceedings of ACM SPAA*, pp. 289-298, 2007.

2. S. Albers, "Energy-Efficient Algorithms", *Communications of the ACM*, Vol. 53, No. 5, pp. 86-96, May 2010.
3. L. Andrew, M. Lin, and A. Wierman, "Optimality, Fairness, and Robustness in Speed Scaling Designs", *Proceedings of ACM SIGMETRICS*, pp. 37-48, June 2010.
4. B. Ata and S. Shneorson, "Dynamic Control of an M/M/1 Service System with Adjustable Arrival and Service Rates", *Management Science*, Vol. 52, No. 11, pp. 1778-1791, 2006.
5. N. Bansal, T. Kimbrel, and K. Pruhs, "Speed Scaling to Manage Energy and Temperature", *Journal of the ACM*, Vol. 54, 2007.
6. N. Bansal, H. Chan, and K. Pruhs, "Speed Scaling with an Arbitrary Power Function", *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, 2009.
7. M. Dell'Amico, D. Carra, M. Pastorelli, and P. Michiardi, "Revisiting Size-based Scheduling with Estimated Job Sizes", *Proceedings of IEEE MASCOTS*, Paris, France, pp. 411-420, September 2014.
8. M. Dell'Amico, D. Carra, M. Pastorelli, and P. Michiardi, "PSBS: Practical Size-Based Scheduling", *IEEE Trans. Computers*, Vol. 65, No. 7, pp. 2199-2212, 2016.
9. M. Elahi, *Optimality and Fairness in Speed-Scaling Systems*, PhD Dissertation, Department of Computer Science, University of Calgary, September 2017.
10. M. Elahi and C. Williamson, "Autoscaling Effects in Speed Scaling Systems", *Proceedings of IEEE MASCOTS*, London, UK, pp. 307-312, September 2016.
11. J. George and J. Harrison, "Dynamic Control of a Queue with Adjustable Service Rate", *Operations Research*, Vol. 49, No. 5, pp. 720-731, September-October 2001.
12. F. Kelly, *Reversibility and Stochastic Networks*, Wiley, 1979.
13. L. Kleinrock, *Queueing Systems, Volume 1: Theory*, Wiley, 1975.
14. D. Low, "Optimal Dynamic Pricing Policies for an M/M/s Queue", *Operations Research*, Vol. 22, No. 3, pp. 545-561, 1974.
15. D. Lu, H. Shen, and P. Dinda, "Size-based Scheduling Policies with Inaccurate Scheduling Information", *Proceedings of IEEE/ACM MASCOTS*, Volendam, Netherlands, pp. 31-38, October 2004.
16. S. Ross, *Stochastic Processes*, Wiley, 1983.
17. L. Schrage, "A Proof of the Optimality of the Shortest Remaining Processing Time Discipline", *Operations Research*, Vol. 16, pp. 678-690, 1968.
18. B. Schroeder and M. Harchol-Balter, "Web Servers Under Overload: How Scheduling Can Help", *ACM Transactions on Internet Technology*, Vol. 6, No. 1, pp. 20-52, February 2006.
19. A. Skrenes and C. Williamson, "Experimental Calibration and Validation of a Speed Scaling Simulator", *Proceedings of IEEE MASCOTS*, London, UK, pp. 105-114, September 2016.
20. D. Snowdon, E. Le Sueur, S. Petters, and G. Heiser, "Koala: A Platform for OS-level Power Management", *Proceedings of ACM EuroSys*, pp. 289-302, 2009.
21. M. Weiser, B. Welch, A. Demers, and S. Shenker, "Scheduling for Reduced CPU Energy", *Proceedings of USENIX OSDI*, 1994.
22. A. Wierman, L. Andrew, and A. Tang, "Power-Aware Speed Scaling in Processor Sharing Systems", *Proceedings of IEEE INFOCOM*, April 2009.
23. A. Wierman, L. Andrew, and A. Tang, "Power-Aware Speed Scaling in Processor Sharing Systems: Optimality and Robustness", *Performance Evaluation*, Vol. 69, pp. 601-622, 2012.
24. F. Yao, A. Demers, and S. Shenker, "A Scheduling Model for Reduced CPU Energy", *Proceedings of ACM FOCS*, pp. 374-382, 1995.