Analysis of Scheduling Policies under Correlated Job Sizes

Varun Gupta
Michelle Burroughs
Mor Harchol-Balter

March 2010
CMU-CS-10-107

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Keywords: Scheduling, Correlation, Auto-correlation, $M/G/1$, Asymptotic analysis, Fluid analysis

Abstract

Correlations in traffic patterns are an important facet of the workloads faced by real systems, and one that has far-reaching consequences on the performance and optimization of the systems involved. While there has been considerable amount of work on understanding the effect of correlations between successive interarrival times, there is very little analytical work in understanding the effect of correlations between successive service requirements (job sizes). All the prior work on analyzing the effects of correlated job sizes is limited to First-Come-First-Served scheduling. This leaves open fundamental questions such as: How do various scheduling policies interact with correlated job sizes? Can scheduling be used to mitigate the harmful effects of correlations?

In this paper we take the first step towards answering these questions. We assume a simple and intuitive model for job size correlations and present the first asymptotic analysis of various common size-independent scheduling policies when the job size sequence exhibits high correlation. Our analysis reveals that the characteristics of various scheduling policies, as well as their performance relative to each other, are markedly different under the assumption of i.i.d. job sizes versus correlated job sizes. Further, among the class of size-independent scheduling policies, there is no single scheduling policy that is optimal for all degrees of correlations and thus any optimal policy must learn the correlations. We support the asymptotic analysis with numerical algorithms for exact performance analysis under an arbitrary degree of correlation, with simulations, and finally verify the lessons from our correlation model on real world traces.

1 Introduction

Motivation

The $M/G/1$ single-server queue has been used as a guiding model for performance analysis of widely varying systems, such as buffers for network switches, web server downlinks, and the CPU scheduler. There is a considerable body of analytical work surrounding the $M/G/1$ queue, including analysis of different scheduling policies and their effects on response times of jobs (defined to be the time from the arrival to the completion of a job) [5]. However, almost all of the exact analysis has been performed under the assumptions of (*i*) Poisson arrival process and (*ii*) independent and identically distributed (*i.i.d.*) job sizes.

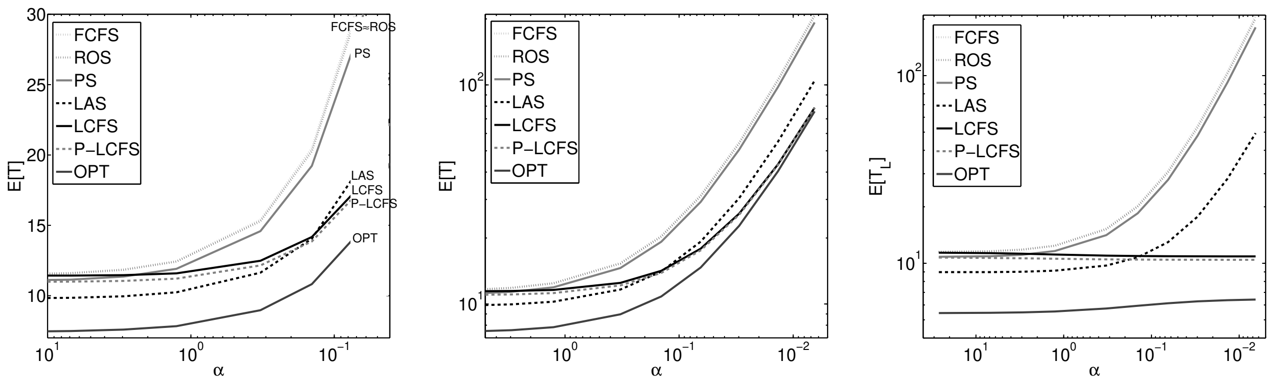
Long ago, the need was recognized to relax these assumptions, as real systems workloads exhibit significant correlation patterns and these patterns tend to greatly affect the accuracy of the traditional results [8, 19]. Primarily, there are three kinds of correlations that exist in real workloads [10]: (*i*) correlations between consecutive interarrival times, (*ii*) correlations between interarrival times and the subsequent service requirements, and (*iii*) correlations between consecutive service requirements (job sizes). Examples of correlation in interarrival processes include [10] (network traffic), and [7, 12, 24] (web servers). Examples of correlation in interarrival time and job sizes include [4, 10, 12, 13]. Examples of correlation in sequential job sizes include [10], [15, 9, 23] (supercomputing), and [18] (disk request sizes). In this paper we focus on studying the effects of correlations of type (*iii*).

While there has been a lot of *analytical* work studying the effect of all three types of correlation on mean response time in single server queues, all of this work has assumed First-Come-First-Served (FCFS) queues only. Fendick et al.[10] examine all three types of correlation via a Brownian approximation and propose a stationary workload approximation based on heavy traffic limits. Adan and Kulkarni [2] also use analysis to study autocorrelation and cross-correlation of interarrival and service times in a MAP/G/1/FCFS queue. Riska et al. [21] use matrix-analytic methods to model correlated arrival streams in a MAP/PH/1 queue, and to numerically calculate mean response time under FCFS. Ghosh and Squillante [12] propose a refinement to the Fendick et al. [10] approximation for FCFS queues, and propose approximations for a multi-class priority system with FCFS scheduling within each class. Cidon et al. [4] derive the Laplace transform of the workload using the theory of linear functional equations in a queue with an Interrupted Poisson arrival process and where the size of a job is positively correlated with the interarrival time preceding the job.

The effect of correlation has also been carefully studied via *simulation*, see for example [16, 17, 22, 26]. In all except [17], FCFS scheduling was assumed. In [17] the authors examine an approximation of Shortest-Job-First (SJF) scheduling, which the authors call SWAP, and compare it against FCFS scheduling via simulation. In [18], the authors propose and evaluate a scheme that predicts the future service requirements based on the auto-correlation function, and drops large jobs to improve performance of an FCFS server.

In summary, while there has been a lot of prior work dealing with correlations in successive job sizes, it has almost exclusively dealt with FCFS scheduling. Important questions have remained unanswered: How do different scheduling policies react to correlations in job sizes? Can scheduling be used to allay the detrimental effect of correlated job sizes on the system performance?

In this paper, we take an important first step by analyzing the mean response time under various scheduling policies in the presence of correlated job sizes. We restrict ourselves to the class of *size-independent* policies. That is, we will look at policies which know the generative correlation model, but not the actual realizations of the sizes (or the size-class) of jobs. In most applications, including CPU scheduling, IP flow scheduling, scheduling of database queries etc., the job sizes are often not known a priori, and hence size-independent policies are more realistic. Further, if the job sizes are known to the scheduler, then SRPT (Shortest Remaining Processing Time first) is already known to minimize mean response time irrespective of the arrival pattern (hence also under correlated arrivals). We will consider the question of how does the optimality of size-independent policies is affected by the presence or absence of correlation in the job sizes.



(a) (b) (c)

Figure 1: An example of the effect of job-size correlation on scheduling policies: (a) mean response time versus α for low to medium correlation; (b) mean response time versus α for medium to high correlation; (c) mean response time of the "little" (L) jobs versus α. Here $\rho = 0.97$ and $C^2 \approx 1.08$. Note that the $\mathbf{E}[T]$ ordering changes from FCFS=ROS=LCFS>PS=P-LCFS>LAS>OPT at $\alpha = \infty$ (*i.i.d.* job sizes) to FCFS≈ROS>PS>LAS>LCFS=P-LCFS=OPT as $\alpha \to 0$ (high correlation).

The MMAP Correlation Model

We assume the following simple *Markov Modulated Arrival Process (MMAP)* model for job-size correlations: jobs belong to one of two classes called little (L) and huge (H), where jobs of class

L (respectively H) are Exponentially distributed with mean $\frac{1}{\mu_L}$ (respectively $\frac{1}{\mu_H} > \frac{1}{\mu_L}$) [1]. Therefore, our jobs belong to a 2-phase hyperexponential (H_2) distribution. Further, the system operates under a 2-state Markovian environment process with states L and H: while the environment process is in state L all arrivals are of class L, and while in state H all arrivals are of class H. The time spent in state L during each visit are *i.i.d.* Exponentially distributed with mean $\frac{1}{\alpha_L}$, and those in state H are *i.i.d.* Exponentially distributed with mean $\frac{1}{\alpha_H}$. Denote $\alpha = \alpha_L + \alpha_H$, and $p = \frac{\alpha_H}{\alpha}$. The arrivals occur according to a Poisson process with rate λ independent of the environment process. Thus if we look at a random arrival, it is of class L with probability p, and of class H with probability $1-p$. We will use $\rho = \lambda \cdot \left(\frac{p}{\mu_L} + \frac{1-p}{\mu_H}\right)$ to denote the long run fraction of time the system is busy. If we fix the job size distribution and arrival rate (i.e. μ_L, μ_H, p, λ) and set $\alpha = \infty$, then the job sizes form an *i.i.d.* stream. However, as we decrease α and thereby increase the mean residence time per sojourn of L and H states, we increase the correlation among successive job sizes, since the probability that a class L job is followed by another class L job ($p_{L,L} = p + \frac{\lambda(1-p)}{\lambda+\alpha}$) increases. An alternate approximate way to view the MMAP correlation is the following: The arrivals occur according to a Poisson process independent of the job size sequence. With probability q, the class of a job is the same as the class of the immediately preceding job, and with probability $1-q$ it is an independent sample dictated by the H_2 job size distribution, where $q = \frac{\lambda}{\lambda+\alpha}$. While more intuitive, this alternate way is only an approximation of the MMAP model in that under the MMAP model an interarrival time is not Exp(λ) if we condition on the classes of the jobs that arrive immediately before and immediately after it.

Let $\cdots, X_{-2}, X_{-1}, X_0, X_1, X_2 \cdots$ represent the sequence of job sizes. An appealing property of the above correlation model is the simple closed-form autocorrelation function (acf). In particular, the lag n correlation is given by:

$$cor(X_m, X_{m+n}) = \frac{\mathbf{E}[X_m X_{m+n}] - \mathbf{E}[X_m]\mathbf{E}[X_{m+n}]}{\sqrt{var(X_m)}\sqrt{var(X_{m+n})}} = \frac{1}{2}\left(\frac{C^2-1}{C^2}\right)\left(\frac{\lambda}{\lambda+\alpha}\right)^n \quad \ldots n \geq 1$$

where $C^2 = \frac{var(X_0)}{\mathbf{E}[X_0]^2} > 1$ denotes the squared coefficient of variation (SCV) of the H_2 job size distribution.

Scope of the MMAP correlation model: The MMAP correlation model that we analyze in this paper is similar to the model used in [2].In particular, richer MMAP models (with more than 2 phases), see [18] for example, are more useful for modeling more general auto-correlation functions. However, the goal of this paper is to explore qualitative behavior of different scheduling policies in the presence of correlated job sizes, and to gain insights for these behaviors and the effect of various system parameters on the performance via an analytically tractable correlation model. We believe

[1]Note that the mean sizes of the two classes can in fact be close. We have chosen the names of the classes to map to low (L) and high (H) load, respectively, in Section 2.

Scheduling Policy	Description
First-Come-First-Served (FCFS)	Jobs are served in the order of arrival.
Last-Come-First-Served (LCFS)	Whenever a job completes service, the next job to be served is the one that arrived last.
Preemptive LCFS with Resume (P-LCFS)	New arrivals immediately begin service by preempting the job at the server. On a service completion, the next job to resume service is the one that arrived last.
Least-Attained-Service (LAS)	The job with the least amount of received service (age) gets to serve.
Threshold-Age-Based-Priority (LAS-a)	Jobs with age $> a$ serve FCFS in low priority queue. Jobs with age $< a$ serve FCFS in high priority queue with preemptive priority over low priority (older than a) jobs.
Processor Sharing (PS)	If there are n jobs in the system, each job gets $\frac{1}{n}$th of the server's capacity.
Random-Order-of-Service (ROS)	Whenever a job completes service, the next job to be served is picked uniformly at random from amongst the jobs currently in the queue.
Optimal Omniscient (OPT)	A hypothetical optimal scheduling scheme that knows the class of all jobs, and gives preemptive priority to class L jobs.

Table 1: A glossary of scheduling policies analyzed in this paper.

our analysis is a critical first step towards understanding the effect of correlations. For example, while a lot of prior work on modeling correlations has focused on matching the autocorrelation function, we will later see that our MMAP model dispels the common wisdom that the autocorrelation function is the only important factor determining the performance of a system. That is, we will see that the performance of various scheduling policies will depend critically on all the system parameters, μ_L, μ_H, p and α, and not just α, ρ and C^2. We believe that the behavior of scheduling policies under correlated job sizes discovered in this paper would extend to more general correlation structures, and we partially test this via real-world traces in Section 4, but we leave it as a topic for future research.

Summary of Contributions

Given the above correlation model, we proceed to analyze a wide range of size-independent scheduling policies (see Table 1 for a list of scheduling policies analyzed in this paper). We now summarize some of our **findings**:

Most of our results look at the effect of the parameter α on mean response time, $\mathbf{E}[T]$. We prove that, although all scheduling policies we consider are hurt by increasing the correlation, the degree to which correlation affects different policies varies widely. We consider two regimes: *(i)* $\mu_L > \mu_H > \lambda$, where the server is never in overload, and *(ii)* $\mu_L > \lambda > \mu_H$, where the system is in overload, during bursts of H jobs, although it is still stable on average. For the no-overload regime, we prove that, as α decreases (correlation increases), all size-independent scheduling policies become the same with respect to mean response time. For the transient-overload regime, we prove that as correlation decreases, there can be a large (up to a factor of $\frac{\mu_L}{\mu_H}$) difference in $\mathbf{E}[T]$ between the policies. Also, the ordering of policies from "best" to "worst" mean response time changes a lot under correlation. An example is shown in Figure 1(a). Some particularly interesting findings include:

- While LAS has provably the best mean response time (among size-independent policies) when $\alpha \to \infty$ for an H_2 job size distribution due to its decreasing failure rate [20], it is provably sub-optimal when $\alpha \to 0$.
- LCFS, which is worst (along with FCFS and ROS) when $\alpha = \infty$ is provably best when $\alpha \to 0$.
- P-LCFS is also provably best when $\alpha \to 0$, which is interesting because at $\alpha \to \infty$ (*i.i.d.* case) LCFS and P-LCFS can be far apart for high variability job size distributions.
- P-LCFS and PS are provably equal at $\alpha \to \infty$, but PS can be arbitrarily worse than P-LCFS as $\alpha \to 0$.

The effect of correlation on the mean response time of the L jobs, $\mathbf{E}[T_L]$, is even more pronounced. In particular, we prove that:

- While $\mathbf{E}[T_L]$ increases for most policies, as α decreases (correlation increases), $\mathbf{E}[T_L]$ always *decreases* for PLCFS and for LCFS. An example is shown in Figure 1(b).
- LAS performs poorly for $\mathbf{E}[T_L]$ and even worse for $\mathbf{E}[T_L^2]$.

The above findings are important because they reverse our intuition for which policies are best under no correlation ($\alpha \to \infty$). In particular, while LAS is designed to help the little jobs, by biasing towards jobs with least attained service, it fails to do this under correlation, and policies like LCFS which are entirely oblivious to job size distribution can actually help the little jobs.

The above results are primarily obtained by using fluid analysis and looking at asymptotic behavior of response time as $\alpha \to 0$, see Section 2. However, the effect of correlation under moderate α is also interesting. To study the moderate α regime, we derive numerical algorithms to analyze LCFS, OPT, PLCFS, and FCFS, see Section 3. For the other policies, we resort to simulations, see Section 4. These numerical and simulation results are useful for understanding the behavior of scheduling policies for intermediate α values and for getting a feel for how quickly scheduling policies converge to their asymptotically-limiting ($\alpha \to 0$) behavior. Figure 3 shows this well, illustrating that LAS is particularly slow to converge to its asymptotic behavior, compared with the other scheduling policies. All the above findings assume a Poisson arrival process and correlations between successive job sizes, as specified by our model. To see how our messages carry through to real-world scenarios, we end Section 4 with some trace-driven simulation studies, the first involving packets at a router, and the second involving supercomputing jobs.

Outline

The paper is structured as follows: We have already described our correlation model in Section 1. Section 2 presents the asymptotic analysis of all the scheduling policies described in Table 1 in the limit $\alpha \to 0$. Section 3 presents our algorithms for exact numerical analysis of some of the scheduling policies for general values of the correlation parameter q. Section 4 provides simulation results, where the theoretical asymptotic results ($\alpha \to 0$) are juxtaposed with simulations and numerical results. Finally, we conclude in Section 5.

Notation	Meaning	Notation	Meaning
$\mathbf{E}[T_L^\pi], \mathbf{E}[T_H^\pi], \mathbf{E}[T^\pi]$	mean response time of a class {L, H, avg} job under policy π	$\mathbf{E}[D_L^\pi], \mathbf{E}[D_H^\pi], \mathbf{E}[D^\pi]$	mean delay of a class {L, H, avg} job under scheduling policy π
$\mathbf{E}[T_L^\pi(x)], \mathbf{E}[T_H^\pi(x)]$	mean response time of a class L, H job of size x under scheduling policy π	W_L, W_H	stationary workload conditioned on being in state L, H
r_L r_H ρ	$= 1 - \frac{\lambda}{\mu_L}$ $= 1 - \frac{\lambda}{\mu_H}$ $= \lambda\,(p/\mu_L + (1-p)/\mu_H)$	$r_L(x)$ $r_H(x)$ $\rho(x)$	$= 1 - \lambda s_L(x)$ $= 1 - \lambda s_H(x)$ $= \lambda(p s_L(x) + (1-p) s_H(x))$
$s_L(x)$ $s_H(x)$	$= \mathbf{E}[\min\{\mathrm{Exp}(\mu_L), x\}] = \frac{1-e^{-\mu_L x}}{\mu_L}$ $= \mathbf{E}[\min\{\mathrm{Exp}(\mu_H), x\}] = \frac{1-e^{-\mu_H x}}{\mu_H}$	W_L^*, W_H^*	stationary fluid workload in a system with flow rates r_L and r_H, conditioned on being in state L, H
$g(x) = \Theta(h(x))$ as $x \to x_0$	$0 < \liminf\limits_{x\to x_0} \frac{g(x)}{h(x)} \le \limsup\limits_{x\to x_0} \frac{g(x)}{h(x)} < \infty$	$W_L^*(x), W_H^*(x)$	stationary fluid workloads in a system with flow rates $r_L(x)$, $r_H(x)$
$g(x) = o(h(x))$ as $x \to x_0$	$\lim_{x\to x_0} \frac{g(x)}{h(x)} = 0$	$\widetilde{X}(s) = \mathbf{E}\left[e^{-sX}\right]$	Laplace transform of r.v. X

Table 2: Notation used in Section 2.

2 Asymptotic Analysis Of Scheduling Policies as $\alpha \to 0$

Our goal in this section is to obtain an understanding of the "first-order effect" of correlations in the job sizes by considering the limiting case where the correlation approaches its maximum value under our model, that is, $\alpha \to 0$.[2] While this extremal case implies arbitrarily long consecutive streaks of only L and only H arrivals, an understanding of the behavior of the various scheduling policies under

[2]The analysis of the asymptote $\alpha \to 0$ should be seen analogously to heavy traffic analysis where the traffic intensity ρ is allowed to approach 1 to observe "first order" effect of system parameters (variance, cross-correlations) on the system performance.

this asymptote gives us insights into why different scheduling policies react differently to correlation in job sizes, and will guide us in design of policies which are robust to the correlations.

As pointed out above, the asymptote $\alpha \to 0$ implies that we have arbitrarily long streaks of L only and H only jobs. Depending on whether $\mu_H > \lambda$ or not (note we are guaranteed $\mu_L > \lambda$ by definition), we will have a system that is either stable or under transient overload during the H states. We consider the two cases separately.

In Section 2.1, we present the asymptotic results for the simpler case $\mu_H > \lambda$. The remainder of the section is devoted to the non-trivial case of $\mu_H < \lambda$. We will start the analysis of this case in Section 2.2 by analyzing the stationary workload in the system and providing intuition for the goal of asymptotic analysis. The remaining sections 2.3-2.8 will present the results for asymptotic mean response time for various scheduling policies for the case $\mu_H < \lambda$. A large number of scheduling policies that we will analyze will involve asymptotic analysis of busy periods. We have chosen to present the main results on busy period analysis in Appendix A and focus on the messages in the main body. For ready reference, we have summarized the notation used in this section in Table 2.

Note on scaling and asymptotic notation: We perform the asymptotic analysis of the scheduling policies by considering a sequence of systems, indexed by the parameter α. The system with index α is obtained by setting the switching rates of the environment process as $\alpha_H = p \cdot \alpha$ and $\alpha_L = (1-p)\alpha$, where p, μ_L, μ_H and λ are held constant. We are interested in seeing the behavior of the scheduling policies in the asymptote $\alpha \to 0$, and hence the expressions for mean response times presented in this section will be written in the *asymptotic notation*: We say that a function $g(\alpha)$ is of a 'smaller order' than $h(\alpha)$ (and make the limit $\alpha \to 0$ implicit), denoted $g(\alpha) = o(h(\alpha))$, when $\frac{g(\alpha)}{h(\alpha)} \to 0$ when $\alpha \to 0$ (see Table 2). When we write the expressions for the mean response time under the αth system, we only identify the dominant term in the expression, expressing the remaining terms which become negligible in comparison as $\alpha \to 0$ as of being smaller order than the dominant term. Similarly, we say $g(\alpha)$ is of 'the same order' as $h(\alpha)$ (again with the limit $\alpha \to 0$ implicit), denoted $g(\alpha) = \Theta(h(\alpha))$ when intuitively $\frac{g(\alpha)}{h(\alpha)}$ is eventually bounded between two strictly positive constants. Thus, for example, a $\Theta(1)$ function is eventually bounded between two strictly positive constants as $\alpha \to 0$. In proving theorems about response time, it will often suffice to just argue about the asymptotic order of busy period durations, probabilities and related quantities.

2.1 Analysis for case $\mu_H > \lambda$

Let T_L^π and T_H^π denote the random variables for response time of class L and class H jobs, respectively, under scheduling policy π (see Table 2). When $\mu_H > \lambda$, the system is stable during both L and H states, and we have the following intuitive result.

Theorem 1 *Let π be any work-conserving, size-independent policy. When $\mu_H > \lambda$,*

$$\lim_{\alpha\to 0} \mathbf{E}[T_L^\pi] = \frac{1}{\mu_L - \lambda} \quad ; \quad \lim_{\alpha\to 0} \mathbf{E}[T_H^\pi] = \frac{1}{\mu_H - \lambda}.$$

Proof: The basic intuition behind the theorem is that since the system is stable under both L and H states, the workload that is present when the environment switches state is uniformly bounded in q (this is not true when $\mu_H < \lambda$ and the H state is in overload). Thus, asymptotically as $q \to 1$ and the residence time in an environment state during each visit increases as $\frac{1}{1-q}$, during each sojourn in state L (or H), the system converges to the stationary distribution of an $M/M/1$ queue with service rate μ_L (or μ_H) as $q \to 1$.

More formally, whenever the environment switches state from H to L or L to H, there is some time until the system first empties out. Because we have assumed a work conserving policy, this time can be stochastically upper bounded by the equilibrium distribution of the busy period duration in an $M/M/1$ with only class H jobs (we omit the proof due to space). Until the system switches again, it goes through *i.i.d.* busy periods of an $M/M/1$ with only L or only H jobs depending on the current environment state. As $q \to 1$, the contribution of these *i.i.d.* busy periods washes away the initial transient effect, and, asymptotically, the system behaves as a probabilistic mixture of two separate $M/M/1$ queues. ■

Remark 1: Theorem 1 says that as job sizes become more and more correlated, the behavior of all work-conserving, size-independent scheduling policies will tend to become the same, provided $\mu_H > \lambda$. Since LAS is optimal (among size-independent policies) at each extreme, we intuitively expect LAS to be near-optimal through the entire range of α, and thus for all levels of correlation. We verify that this is indeed true in Section 4, Figure 2.

2.2 Preliminaries: Workload analysis via Fluid model for the case $\mu_H < \lambda$

We begin our study of the case $\mu_H < \lambda$ by finding the distribution of stationary workload during the L and H states, respectively. To do this, we first introduce the *fluid model* of our MMAP correlation model.

Definition 1 *Under the fluid model, we assume that the workload increases at a constant rate of $-r_H$ during the H states (see Table 2), and decreases at a constant rate of r_L during the L states as long as the workload is positive.*

We now present the expression for the stationary workload under the fluid model for our system, deferring the proof to the end of the section.

Lemma 1 *Let W_L^* and W_H^* denote the random variables for the stationary workload during L and H states under the fluid model, respectively (we will superscript random variables by $*$ when referring to the fluid model). Let $\widetilde{W_L^*}(s) = \mathbf{E}\left[e^{-sW_L^*}\right]$ and $\widetilde{W_H^*}(s) = \mathbf{E}\left[e^{-sW_H^*}\right]$ denote their Laplace transforms.*

Then,

$$\widetilde{W_H^*}(s) = \frac{\gamma_H - \gamma_L}{s + (\gamma_H - \gamma_L)} \tag{1}$$

$$\widetilde{W_L^*}(s) = \left(1 - \frac{\gamma_L}{\gamma_H}\right) + \frac{\gamma_L}{\gamma_H} \cdot \frac{\gamma_H - \gamma_L}{s + (\gamma_H - \gamma_L)} \tag{2}$$

where $\gamma_L = \frac{\alpha_L}{r_L}$ *and* $\gamma_H = -\frac{\alpha_H}{r_H}$.
Thus the workload during the H states, W_H^**, is distributed as an* $\mathrm{Exp}(\gamma_H - \gamma_L)$ *random variable, and the workload during the L states,* W_L^**, is a mixture of an* $\mathrm{Exp}\,(\gamma_H - \gamma_L)$ *random variable and an atom at* 0*. Further, the mean of* W_L^* *and* W_H^* *are of the order* $\Theta\left(\frac{1}{\alpha}\right)$*. Thus, as* $\alpha \to 0$ *, the fluid workload diverges at a rate of* $\frac{1}{\alpha}$*.*

Lemma 2

$$W_L \stackrel{d}{=} W_L^* + o(\alpha^{-1}) \;\; ; \qquad W_H \stackrel{d}{=} W_H^* + o(\alpha^{-1})$$

Remark 2: Lemma 2 says that, asymptotically as $\alpha \to 0$, the workload of the stochastic system converges in distribution to the workload under the fluid model.

Proof of Lemma 1: We first note that by conditional PASTA [25], W_L^* and W_H^* are equal in distribution to the stationary workload at the end of L and H states respectively. Let T_L and T_H be Exponentially distributed random variables with mean $\frac{1}{\alpha_L}$ and $\frac{1}{\alpha_H}$, respectively. We have the following stochastic fixed point equations:

$$W_H^* \stackrel{d}{=} W_L^* - r_H T_H \; ; \quad W_L^* \stackrel{d}{=} \max\left\{W_H^* - r_L T_L, 0\right\}$$

Taking Laplace transforms of the above equations, we get the following fixed point equations:

$$\widetilde{W_H^*}(s) = \widetilde{W_L^*}(s) \cdot \frac{\alpha_H/r_H}{\alpha_H/r_H - s}; \quad \widetilde{W_L^*}(s) = \frac{s\widetilde{W_H^*}(\alpha_L/r_L) - (\alpha_L/r_L)\widetilde{W_H^*}(s)}{s - \alpha_L/r_L}$$

which yield the expressions in Lemma 1. ∎

Proof of Lemma 2: The lemma is proven by starting with Theorem 9 which gives the exact expressions for the Laplace transforms of W_L and W_H. The transforms of W_L and W_H can be recognized as mixtures of a point mass at zero, and two Exponential distributions. We consider the case of W_L here. According to Theorem 9:

$$\widetilde{W_L}(s) = \frac{(1-\rho)\alpha m_L m_H - s m_L g_H \pi_L(0)}{\alpha_L g_H m_L + \alpha_H g_L m_H - s g_L g_H} \tag{3}$$

where,

$$m_L = \mu_L + s \quad ; \quad m_H = \mu_H + s$$
$$g_L = \mu_L - \lambda + s \quad ; \quad g_H = \mu_H - \lambda + s$$
$$\pi_L(0) = \frac{(1-\rho)\alpha(\mu_H + \xi)}{\xi(\mu_H - \lambda + \xi)}$$

and ξ denotes the unique root of the denominator of (3) (viewed as a cubic in s) in the interval $(0, +\infty)$. The quantity $\pi_L(0)$ denotes the long run fraction of time that the system is empty conditioned on being in state L. Taking the limit $\alpha \to 0$, we get

$$\xi \sim (\lambda - \mu_H) + \frac{p\alpha\lambda}{\lambda - \mu_H} + \Theta(\alpha^2)$$

and thus,

$$\begin{aligned}\pi_L(0) &= \frac{(1-\rho)\alpha(\mu_H + \xi)}{\xi(\mu_H - \lambda + \xi)} \\ &\sim \frac{(1-\rho)\alpha(\lambda + \Theta(\alpha))}{(\lambda - \mu_H + \Theta(\alpha))\left(\frac{p\alpha\lambda}{\lambda - \mu_H} + \Theta(\alpha^2)\right)} \\ &\sim \frac{1-\rho}{p} + \Theta(\alpha)\end{aligned}$$

Note that the above is not in disagreement with the result $\mathbf{Pr}[W_L^* = 0] = \left(1 - \frac{\gamma_L}{\gamma_H}\right)$ as the latter is only equivalent to $\mathbf{Pr}\Big[W_L = o\left(\frac{1}{\alpha}\right)\Big]$. The other roots of the denominator of (3) in the limit $\alpha \to 0$ are given by:

$$\chi \sim (\lambda - \mu_L) - \frac{p\alpha\lambda}{\mu_L - \lambda} + \Theta(\alpha^2)$$
$$\eta \sim -\frac{\alpha\mu_L\mu_H(1-\rho)}{(\mu_L - \lambda)(\lambda - \mu_H)} + \Theta(\alpha^2)$$

Canceling the common factor $(s - \xi)$, and noting that $\frac{\alpha\mu_L\mu_H(1-\rho)}{(\mu_L-\lambda)(\lambda-\mu_H)} = (\gamma_H - \gamma_L)$, we can rewrite:

$$\begin{aligned}\widetilde{W_L}(s) &= \pi_L(0) + K_1\frac{-\chi}{s - \chi} + K_2\frac{-\eta}{s - \eta} \\ &= \frac{1-\rho}{p} + K_1\frac{\mu_L - \lambda + \Theta(\alpha)}{s + (\mu_L - \lambda + \Theta(\alpha))} + K_2\frac{\gamma_H - \gamma_L}{s + (\gamma_H - \gamma_L)}\end{aligned}$$

Matching the coefficients of s, we get $K_1 = \frac{r_L}{1-r_L}\left(\frac{1-\rho}{p}\right) + \Theta(\alpha)$ and $K_2 = 1 - \frac{1-\rho}{p(1-r_L)} + \Theta(\alpha) = \frac{\gamma_L}{\gamma_H} + \Theta(\alpha)$.

Thus we have proved that, as $\alpha \to 0$, W_L is a mixture of an Exponential distribution with mean $\frac{1}{\gamma_H - \gamma_L}$ with probability $\sim \frac{\gamma_L}{\gamma_H}$, and with the remaining probability the stationary distribution of an $M/M/1$ with arrival rate λ and service rate μ_L. ■

Goals of asymptotic analysis Since we are interested in analyzing work-conserving policies, the stationary workload, W, is the same across policies. What differs from one policy to another is what types of jobs make up that workload. Since we restrict ourselves to size-independent policies, we can bound the mean remaining size of any job under our H_2 job size distribution between $\frac{1}{\mu_L}$ and $\frac{1}{\mu_H}$. This gives bounds on $\mathbf{E}[N^\pi]$ – the mean number of jobs in the system for any work-conserving policy π – as $\mu_H \mathbf{E}[W] \le \mathbf{E}[N^\pi] \le \mu_L \mathbf{E}[W]$. Finally, by applying Little's law, we get $\frac{\mu_H}{\lambda}\mathbf{E}[W] \le \mathbf{E}[T^\pi] \le \frac{\mu_L}{\lambda}\mathbf{E}[W]$. Since $\mathbf{E}[W]$ diverges as $\frac{1}{\alpha}$ as $\alpha \to 0$, we have the following.

Lemma 3 *When $\mu_H < \lambda$ in the MMAP model, the mean response time of any work-conserving size-independent scheduling policy π grows as $\mathbf{E}[T^\pi] = \frac{K^\pi}{\alpha} + o(\frac{1}{\alpha})$, for some constant K^π which depends only on the scheduling policy and the parameters μ_H, μ_L, p and λ.*

Our goal is to identify the K^π for different policies. Note again the analog to heavy traffic analysis, where space (response time, number of jobs in system, etc.) is scaled by $(1-\rho)$ and analyzed in the limit $\rho \to 1$.

2.3 FCFS

Theorem 2 *In the regime $\mu_H < \lambda$,*

$$\mathbf{E}\Big[D_L^{FCFS}\Big] = \frac{(1-p)}{p(1-\rho)}\left(\frac{\lambda}{\mu_H} - 1\right)^2 \frac{1}{\alpha} + o\left(\frac{1}{\alpha}\right)$$
$$\mathbf{E}\Big[D_H^{FCFS}\Big] = \frac{1}{(1-\rho)}\left(1 - \frac{\lambda}{\mu_L}\right)\left(\frac{\lambda}{\mu_H} - 1\right)\frac{1}{\alpha} + o\left(\frac{1}{\alpha}\right)$$

Proof: By conditional PASTA, the delay of class L jobs is distributed as W_L, and that of class H as W_H. Applying Lemmas 2 and 1, the result is immediate. ■

Remark 3: We already see a divergence in the behavior of scheduling policies when job sizes become correlated. When $\alpha = \infty$ (*i.i.d.* case), and under a Poisson arrival process, the mean delay under FCFS depends only on the first two moments of the job size distribution. However, as $\alpha \to 0$, it depends on all the parameters of the H_2 job size distribution.

2.4 P-LCFS, LCFS and OPT

While it is hard to characterize the optimal size-independent policy when job sizes are correlated since the optimal policy might (and will) exploit the correlation structure to predict classes of future jobs based on observed history of job sizes, we can obtain a trivial lower bound by considering an omniscient scheduler – that is, a scheduler that knows the *class* of each job in the system, but not the exact size, and gives preemptive priority to class L jobs. We call this policy OPT.

Theorem 3 *When $\mu_H < \lambda$, we have for each policy $\pi \in \{LCFS, P\text{-}LCFS, OPT\}$:*

$$\mathbf{E}[D_L^\pi] = \Theta(1)$$
$$\mathbf{E}[D_H^\pi] = \left[\frac{\mu_H}{\lambda(1-p)}\right] \frac{(1-p)\lambda}{(1-\rho)} \left(\frac{1}{\mu_H} - \frac{1}{\mu_L}\right) \left(\frac{\lambda}{\mu_H} - 1\right) \frac{1}{\alpha} + o\left(\frac{1}{\alpha}\right)$$

Corollary 1 *For $\pi \in \{LCFS, P\text{-}LCFS, OPT\}$, when $\mu_H < \lambda$: $\lim_{\alpha \to 0} \frac{\mathbf{E}\left[T^{FCFS}\right]}{\mathbf{E}[T^\pi]} = \frac{\lambda}{\mu_H}$.*

Proof of Theorem 3: We first consider class L jobs. Under OPT, class L jobs get priority, and hence their response time is stochastically upper bounded by that of an $M/M/1$ with arrival rate λ and service rate μ_L, and is $\Theta(1)$. Under P-LCFS, the response time of class L jobs is the busy period started by $\text{Exp}(\mu_L)$ work in state L. By Theorem 11, Case 2 (see Appendix A), this is $\Theta(1)$. Under LCFS, the delay of class L jobs is a busy period started either by $\text{Exp}(\mu_L)$, $\text{Exp}(\mu_H)$ or 0 work. Again, by Theorem 11, Case 2, this is $\Theta(1)$.
To understand the delay of class H jobs, note that the above implies that the mean number of class L jobs in the system, and hence their contribution to the total workload is $\Theta(1)$. However, the stationary average workload is $\Theta(\alpha^{-1})$, and hence this must be composed (aside from a $\Theta(1)$ term) of class H jobs alone. Since, all scheduling policies are size-independent, the mean residual size of these class H jobs is $\frac{1}{\mu_H}$, yielding the mean number of class H jobs of $\frac{p\mathbf{E}[W_L]+(1-p)\mathbf{E}[W_H]}{1/\mu_H}$. By Little's law, we obtain the mean delay of class H jobs as $\frac{p\mathbf{E}[W_L]+(1-p)\mathbf{E}[W_H]}{\lambda(1-p)/\mu_H}$. ■

Remark 4: The above proof does not extend to other policies in Table 1 as their $\mathbf{E}[T_L]$ is not $\Theta(1)$.
Remark 5: For the metric of mean response time, all three policies – LCFS, P-LCFS and OPT – are asymptotically optimal. However, the mean response times of class L jobs under the three policies are different, although always $\Theta(1)$, and given by the following lemma, whose proof we omit.

Lemma 4 *When $\mu_H < \lambda$, the class L mean response time under OPT, LCFS and P-LCFS are given by:*

$$\mathbf{E}\left[T_L^{OPT}\right] = \frac{1}{\mu_L - \lambda} + o(1)$$
$$\mathbf{E}\left[T_L^{P-LCFS}\right] = \mathbf{E}\left[B_L^L\right] + o(1) = \frac{1-\rho_H}{\mu_L(1-\rho)} + o(1)$$
$$\mathbf{E}\left[T_L^{LCFS}\right] = \theta_H(1 - \frac{\lambda}{\mu_L})\mathbf{E}\left[B_L^H\right] + \frac{\lambda}{\mu_L}\mathbf{E}\left[B_L^L\right] + \frac{1}{\mu_L} + o(1)$$

where $\theta_H = \frac{(1-p)(\lambda-\mu_H)}{(1-p)\lambda+(p-\rho)\mu_H}$, and expressions for $\mathbf{E}\left[B_L^L\right]$ and $\mathbf{E}\left[B_L^H\right]$ are given in Corollary 3 (see Appendix A).

Remark 6: Comparing with $\mathbf{E}\Big[T_L^{P-LCFS}|\alpha=\infty\Big] = \frac{1}{\mu_L}\cdot\frac{1}{1-\rho}$, we see that the extreme correlated $\mathbf{E}[T_L]$ for P-LCFS is always lower than the uncorrelated $\mathbf{E}[T_L]$. We can also prove a similar result for LCFS.

Remark 7: A further difference between the three policies emerges if one looks at higher order metrics, such as $\mathbf{E}[(T_L^{\pi})^2]$. As a byproduct of the proof of Theorem 11 (Case 2), we can see that $\mathbf{E}\Big[(T_L^{P-LCFS})^2\Big] = \Omega\left(\frac{1}{\alpha}\right)$, while it is $\Theta(1)$ for OPT. Thus, while simple policies such as P-LCFS and LCFS are asymptotically optimal, there still are benefits in investing in a learning-based scheduling policy when one cares about more fine-grained metrics than just the mean response time.

2.5 LAS

The asymptotic analysis of LAS presented below builds on the analysis under *i.i.d.* arrivals given in [6]. In short, to analyze the response time of a tagged arrival of size x, we consider a modified system where jobs of original size s are truncated to size $\min\{s,x\}$ when they enter the system. Under LAS, the response time of the tagged arrival is given by the busy period generated by the work it sees on arrival in this modified system. The expressions for the mean response time of class L and H jobs of size x are given by:

Theorem 4 *When $\mu_H < \lambda$, the mean response time of a job of size x under the LAS scheduling policy is given by:* ***Case*** $\lambda s_H(x) > 1$***:***

$$\mathbf{E}\Big[T_L^{LAS}(x)\Big] = \frac{\mathbf{E}[W_L^*(x)]}{1-\rho(x)} + o\left(\frac{1}{\alpha}\right)$$

$$\mathbf{E}\Big[T_H^{LAS}(x)\Big] = \frac{1}{\alpha_H} + \frac{\mathbf{E}[W_H^*(x)] + \frac{\lambda s_H(x)-1}{\alpha_H}}{1-\rho(x)} + o\left(\frac{1}{\alpha}\right)$$

Case $\lambda s_H(x) < 1$***:***

$$\mathbf{E}\Big[T_L^{LAS}(x)\Big] = \mathbf{E}\Big[T_L^{M/M/1/LAS}(x)\Big] + o(1)$$

$$\mathbf{E}\Big[T_H^{LAS}(x)\Big] = \mathbf{E}\Big[T_H^{M/M/1/LAS}(x)\Big] + o(1)$$

where $\mathbf{E}\Big[T_L^{M/M/1/LAS}(x)\Big]$ *and* $\mathbf{E}\Big[T_H^{M/M/1/LAS}(x)\Big]$ *denote the mean response time of a job of size x under LAS scheduling in $M/M/1$ queues with arrival rate λ, and job size distribution* $\exp(\mu_L)$ *and* $\exp(\mu_H)$*, respectively.*

Proof: Case $\lambda s_H(x) > 1$: In this case, the modified system with truncated job sizes is in transient overload during the H states. Theorem 11, Case 1 (see Appendix A), gives us the expression for the required mean busy period.

Case $\lambda s_H(x) < 1$: In this case, the modified system with truncated job sizes is stable during the H states. As $\alpha \to 0$, the system looks like a mixture of two independent stable $M/G/1$ queues with the modified job size distributions (similar to Theorem 1). The mean response time of a type L job of size x in this modified system thus converges to the mean response time of a job of size x under an $M/M/1/LAS$ system with arrival rate λ and job sizes *i.i.d.* $\text{Exp}(\mu_L)$. A similar argument applies to type H jobs of size x. ∎

Remark 8: Under *i.i.d.* H_2 job sizes, LAS is the optimal size-independent scheduling policy and in particular better than LCFS and P-LCFS for minimizing the mean response time because it isolates the class L jobs from class H jobs. Intuitively we expect this behavior to carry over when correlations are introduced, *but this is not the case.* Not only does LAS perform suboptimally, but the mean response time of L jobs under LAS grows as $\Theta\left(\frac{1}{\alpha}\right)$, while it is $\Theta(1)$ under LCFS and P-LCFS. The reason for this counter-intuitive behavior lies in the fraction of L jobs that do not get isolation and hence experience $\Theta\left(\frac{1}{\alpha}\right)$ mean response time. Under LCFS and P-LCFS, this fraction is $\Theta(\alpha)$ with a net effect of $\Theta(1)$. Under LAS, however, all L jobs with a size bigger than $\frac{1}{\mu_H}\log\left(\frac{\mu_H}{\lambda-\mu_H}\right)$, which is a $\Theta(1)$ fraction, experience $\Theta\left(\frac{1}{\alpha}\right)$ mean response time.

2.6 LAS-a

Recall that under the LAS-a scheduling policy with parameter a, jobs of age less than a get preemptive priority over jobs of age larger than a. Within the same priority class, the jobs are served in FCFS order.

Theorem 5 *When $\mu_H < \lambda$, the mean response time of a job of size x under the LAS-a scheduling policy with threshold a is given by:*

Case $\lambda s_H(a) > 1$***:***

Subcase $x < a$ ***:***

$$\mathbf{E}\left[T_L^{LAS-a}(x)\right] = \mathbf{E}[W_L^*(a)] + o(1/(1-q))$$
$$\mathbf{E}\left[T_H^{LAS-a}(x)\right] = \mathbf{E}[W_H^*(a)] + o(1/(1-q))$$

Subcase $x > a$ ***:***

$$\mathbf{E}\left[T_L^{LAS-a}(x)\right] = \frac{\mathbf{E}[W_L]}{1-\rho(a)} + o\left(\frac{1}{1-q}\right)$$
$$\mathbf{E}\left[T_H^{LAS-a}(x)\right] = \frac{1}{\alpha_H} + \frac{\mathbf{E}[W_H] - \frac{r_H(a)}{\alpha_H}}{1-\rho(a)} + o\left(\frac{1}{1-q}\right)$$

Case $\lambda s_H(a) < 1$***:***

Subcase $x < a$ ***:***

$$\mathbf{E}\left[T_L^{LAS-a}(x)\right] = \mathbf{E}\left[T_L^{M/M/1/LAS-a}(x)\right] + o(1)$$
$$\mathbf{E}\left[T_H^{LAS-a}(x)\right] = \mathbf{E}\left[T_H^{M/M/1/LAS-a}(x)\right] + o(1)$$

where $\mathbf{E}\Big[T_L^{M/M/1/LAS-a}(x)\Big]$ *and* $\mathbf{E}\Big[T_H^{M/M/1/LAS-a}(x)\Big]$ *are the mean response time of a job of size* x *under LAS-a scheduling in M/M/1 queues with arrival rate* λ*, and job size distribution* $\mathrm{Exp}(\mu_L)$ *and* $\mathrm{Exp}(\mu_H)$*, respectively. These, in turn are given by the response time of a job of size* $x < a$ *in* $M/G/1/FCFS$ *queues with arrival rate* λ *and job size distributions* $\min\{\mathrm{Exp}\,\mu_L, a\}$ *and* $\min\{\mathrm{Exp}\,\mu_H, a\}$*, respectively.*

Subcase $x > a$ ***:***

$$\mathbf{E}\Big[T_L^{LAS-a}(x)\Big] = \frac{\mathbf{E}[W_L]}{1-\rho(a)}(1-u_\alpha^L) + \frac{\mathbf{E}[W_L]}{r_L(a)}u_\alpha^L$$
$$\mathbf{E}\Big[T_H^{LAS-a}(x)\Big] = \frac{\mathbf{E}[W_H]}{1-\rho(a)}(1-u_\alpha^H) + \frac{\mathbf{E}[W_H]}{r_H(a)}u_\alpha^H$$

$$\textit{where } u_\alpha^L = \left[\frac{1-\widetilde{W}_L\left(\frac{\alpha_L}{r_L(a)}+\frac{\alpha_H}{r_H(a)}\right)}{\mathbf{E}[W_L]\left(\frac{\alpha_L}{r_L(a)}+\frac{\alpha_H}{r_H(a)}\right)}\right]$$

$$\textit{and } u_\alpha^H = \left[\frac{1-\widetilde{W}_H\left(\frac{\alpha_L}{r_L(a)}+\frac{\alpha_H}{r_H(a)}\right)}{\mathbf{E}[W_H]\left(\frac{\alpha_L}{r_L(a)}+\frac{\alpha_H}{r_H(a)}\right)}\right]$$

Proof: **Case:** $\lambda s_H(a) > 1$

Subcase $x < a$ **:** When a job of size $x < a$ arrives, its delay is precisely the amount of workload in the high priority queue it sees on arrival. Since the high priority queue is in transient overload during H states, these workloads are asymptotically given by the stationary fluid level in the high priority queue $W_L^*(a)$ and $W_H^*(a)$.

Subcase $x > a$ **:** When a job of size $x > a$ arrives, its delay is given by the busy period generated by all the workload it sees on arrival, and the new arrivals into the high priority queue. Since the workload it sees is $\Theta(\frac{1}{1-q}) = \Theta(\alpha^{-1})$, and the high priority queue is in transient overload during H states, the expression for the mean busy period is obtained by application of Theorem 11, Case 1.

Case: $\lambda s_H(a) < 1$

Subcase $x < a$ **:** This case is similar to $\lambda s_H(x) < 1$ case under LAS scheduling.

Subcase $x > a$ **:** When a job of size $x > a$ arrives into the system, all the work in the high priority queue, as well as the low priority queue needs to be processed before this tagged job can depart. This workload is W_L or W_H depending on the class of the arriving job. Further, while the work ahead of the tagged user and the work associated with the tagged user is being processed, new arrivals into the high priority queue also delay the tagged user. However, the workload ahead of the tagged user is decreasing during both the L and the H states. Thus the mean busy period is given by an application of Theorem 12, Case 2. ∎

2.7 PS

Analysis of PS requires a totally different approach than what we have used thus far. To analyze PS, we will approximate the evolution of number of class L and class H jobs in the system via mean field ordinary differential equations (ODEs). We then solve these ODEs to obtain closed-form dynamics of the number of L and H jobs conditioned on the state. While in this section we will only work with the solutions of these ODEs, we would like to point out that the number of jobs in the corresponding stochastic system are within $o\left(\frac{1}{\alpha}\right)$ (in fact, $\Theta\left(\frac{1}{\alpha^{0.5}}\right)$ using the framework of Kurtz[14]) of the solution of the mean field ODEs. We have so far been unable to obtain the stationary distributions for the number of class L and H jobs from our analysis, but are still able to draw useful conclusions about the behavior of the system. Further, in Section 4, we use our analysis to efficiently obtain the mean response time under the asymptotic regime $\alpha \to 0$ via numerical simulation. (See remarks at the end of this section.)

To analyze PS, let $x_L(t)$ and $x_H(t)$ denote the number of class L and class H jobs in the system at time t under the mean field approximation, and $f(t) = \frac{x_L(t)}{x_H(t)}$. Let $w(t) = \frac{x_L(t)}{\mu_L} + \frac{x_H(t)}{\mu_H}$ denote the workload in the system at time t. We can approximate the (stochastic) dynamics of the system under PS during L states (deterministically) as:

$$\frac{dx_L}{dt} = \lambda - \frac{x_L}{x_L + x_H}\mu_L \ ; \quad \frac{dx_H}{dt} = -\frac{x_H}{x_L + x_H}\mu_H$$

and during H states as:

$$\frac{dx_L}{dt} = -\frac{x_L}{x_L + x_H}\mu_L \ ; \quad \frac{dx_H}{dt} = \lambda - \frac{x_H}{x_L + x_H}\mu_H$$

The above ODEs are justified because as $\alpha \to 0$, x_L and x_H are of the order $\frac{1}{\alpha}$. Thus the ratio $\frac{x_L}{x_L+x_H}$ changes on a much slower time scale than the mean interarrival or interdeparture times.

Theorem 6 *The dynamics of the number of class L and H jobs during the L states under the mean field approximation and PS scheduling satisfies:*

$$\frac{f(t) + \frac{\mu_L}{\mu_H}}{f(0) + \frac{\mu_L}{\mu_H}} = \frac{(w(0) - r_L t)^+}{w(0)} \left(\frac{f(t) + \frac{\lambda}{\lambda+\mu_H-\mu_L}}{f(0) + \frac{\lambda}{\lambda+\mu_H-\mu_L}} \right)^{\frac{\mu_H}{\lambda+\mu_H-\mu_L}} \tag{4}$$

and during H states satisfies:

$$\frac{\frac{1}{f(t)} + \frac{\mu_H}{\mu_L}}{\frac{1}{f(0)} + \frac{\mu_H}{\mu_L}} = \frac{w(0) - r_H t}{w(0)} \left(\frac{\frac{1}{f(t)} + \frac{\lambda}{\lambda+\mu_L-\mu_H}}{\frac{1}{f(0)} + \frac{\lambda}{\lambda+\mu_L-\mu_H}} \right)^{\frac{\mu_L}{\lambda+\mu_L-\mu_H}} . \tag{5}$$

Proof: We will prove the dynamics for L states. The dynamics for H states is obtained by flipping the L's and the H's. Recall $f = \frac{x_L}{x_H}$ and $w = \frac{x+L}{\mu_L} + \frac{x_H}{\mu_H}$. We first write the ode for f:

$$
\begin{aligned}
\frac{df}{dt} = \frac{d\left(\frac{x_L}{x_H}\right)}{dt} &= \frac{1}{x_H}\frac{dx_L}{dt} - \frac{x_L}{x_H^2}\frac{dx_H}{dt} \\
&= \frac{x_H\left(\lambda - \frac{x_L}{x_L+x_H}\mu_L\right) - x_L\left(-\frac{x_H}{x_L+x_H}\mu_H\right)}{x_H^2} \\
&= \frac{x_H\lambda - (\mu_L-\mu_H)\frac{x_L x_H}{x_L+x_H}}{x_H^2}
\end{aligned}
$$

Since $x_L = f x_H$, we have $x_H = w\frac{1}{\frac{f}{\mu_L}+\frac{1}{\mu_H}}$ and $x_L = w\frac{f}{\frac{f}{\mu_L}+\frac{1}{\mu_H}}$, which gives:

$$
\begin{aligned}
\frac{df}{dt} &= \frac{\lambda}{w}\left(\frac{f}{\mu_L}+\frac{1}{\mu_H}\right) - (\mu_L-\mu_H)\frac{f}{w(1+f)}\left(\frac{f}{\mu_L}+\frac{1}{\mu_H}\right) \\
&= \frac{1}{w}\left(\frac{f}{\mu_L}+\frac{1}{\mu_H}\right)\left(\frac{\lambda(1+f)-(\mu_L-\mu_H)f}{1+f}\right)
\end{aligned}
$$

or,

$$
\frac{f+1}{\left(f+\frac{\mu_L}{\mu_H}\right)\left(f+\frac{\lambda}{\lambda+\mu_H-\mu_L}\right)}df = \frac{\lambda+\mu_H-\mu_L}{\mu_L}\frac{dt}{w}
$$

or,

$$
\frac{f+1}{\left(f+\frac{\mu_L}{\mu_H}\right)\left(f+\frac{\lambda}{\lambda+\mu_H-\mu_L}\right)}df = \frac{\lambda+\mu_H-\mu_L}{-\mu_L\left(1-\frac{\lambda}{\mu_L}\right)}\frac{dw}{w}
$$

Now noting that $\frac{\frac{a-1}{a-b}}{f+a} + \frac{\frac{1-b}{a-b}}{f+b} = \frac{f+1}{(f+a)(f+b)}$, where $a = \frac{\mu_L}{\mu_H}$ and $b = \frac{\lambda}{\lambda+\mu_H-\mu_L}$,

$$
(a-1)d\log(f+a) + (1-b)d\log(f+b) = (a-b)\frac{\lambda+\mu_H-\mu_L}{-\mu_L\left(1-\frac{\lambda}{\mu_L}\right)}d\log w
$$

Substituting back a and b,

$$
\left(\frac{\mu_L-\mu_H}{\mu_H}\right)d\log\left(f+\frac{\mu_L}{\mu_H}\right) - \left(\frac{\mu_L-\mu_H}{\lambda+\mu_H-\mu_L}\right)d\log\left(f+\frac{\lambda}{\lambda+\mu_H-\mu_L}\right) = \frac{(\mu_L-\mu_H)}{\mu_H}d\log w
$$

or,

$$
d\log\left(f+\frac{\mu_L}{\mu_H}\right) = d\log w + \left(\frac{\mu_H}{\lambda+\mu_H-\mu_L}\right)d\log\left(f+\frac{\lambda}{\lambda+\mu_H-\mu_L}\right)
$$

finally giving,

$$\frac{f(t)+\frac{\mu_L}{\mu_H}}{f(0)+\frac{\mu_L}{\mu_H}} = \frac{(w(0)-r_L t)^+}{w(0)} \left(\frac{f(t)+\frac{\lambda}{\lambda+\mu_H-\mu_L}}{f(0)+\frac{\lambda}{\lambda+\mu_H-\mu_L}}\right)^{\frac{\mu_H}{\lambda+\mu_H-\mu_L}}. \tag{6}$$

■

Remark 9: We use Theorem 6 to simulate the limit $\alpha \to 0$ in Section 4 as follows: Given the number of L and H jobs at the moment system switches environment, we obtain the number of L and H jobs at the next environment switch by first sampling the duration until the next switch. We then use Theorem 6 to obtain the number of L and H jobs at this next switch, and repeat. By conditional PASTA, we note that the distribution of L and H jobs at the switching epochs suffices to obtain the stationary distribution of L and H jobs in the system, and hence the mean response time.

Remark 10: During the H states, the system workload diverges. Theorem 6 shows that under PS, this divergence happens along $f(t) \to 0$ (or $x_L = 0$), and further $f(t) \to 0$ as approximately $f(t) \sim \left(\frac{w(t)}{w(0)}\right)^{-\frac{\lambda+\mu_L-\mu_H}{\lambda-\mu_H}}$. To see why this is true, note that $\frac{w(t)}{w(0)}$ increases, and the exponent of $f(t)$ on LHS of (5) is -1 and on the RHS is $\frac{-\mu_L}{\lambda+\mu_L-\mu_H}$, for an effective exponent of $\frac{-(\lambda-\mu_H)}{\lambda+\mu_L-\mu_H}$ on the LHS. Thus $f(t)$ increases as approximately $\left(\frac{w(t)}{w(0)}\right)^{-\frac{\lambda+\mu_L-\mu_H}{\lambda-\mu_H}}$ when t is sufficiently large.

Remark 11: During the L states, the system workload approaches 0. However, under PS there is an interesting dichotomy. When $\mu_L < \lambda + \mu_H$, the workload goes to 0 along the line $f(t) \to \infty$ (or $x_H = 0$), and further $f(t) \to \infty$ as approximately $f(t) \sim \left(\frac{w(t)}{w(0)}\right)^{-\frac{\lambda+\mu_H-\mu_L}{\mu_L-\lambda}}$. The explanation for this case is similar to that for H states: $\frac{w(t)}{w(0)}$ goes to 0, and since $\frac{f(t)+\frac{\lambda}{\lambda+\mu_H-\mu_L}}{f(0)+\frac{\lambda}{\lambda+\mu_H-\mu_L}}$ is bounded below by a positive constant ($f(t)$ is positive and increasing during L states), we must have $f(t) \to \infty$. However, when $\mu_L > \lambda + \mu_H$, the workload goes to 0 along the line $f(t) \to \frac{\lambda}{\mu_L-\lambda-\mu_H}$, and further $f(t) \sim \frac{\lambda}{\mu_L-\lambda-\mu_H} - C \cdot \left(\frac{w(t)}{w(0)}\right)^{\frac{\mu_L-\lambda-\mu_H}{\mu_H}}$. Therefore, when $\mu_L > \lambda + \mu_H$, we have $\frac{x_L(t)}{x_H(t)} < \frac{\lambda}{\mu_L-\lambda-\mu_H}$ for all t under the mean field approximation.

The last remark entails the following:

Corollary 2 *When $\mu_L > \lambda + \mu_H$, we have the following bound on the mean response time under PS (as $\alpha \to 0$):*

$$\mathbf{E}\Big[T^{PS}\Big] \le \frac{\mu_L \mu_H}{\lambda(\mu_L-\lambda)} \mathbf{E}[W].$$

Recall that the mean response time of OPT is given by $\frac{\mu_H}{\lambda}\mathbf{E}[W]$. Thus, PS mean response time is within a factor $\frac{\mu_L}{\mu_L-\lambda}$ of OPT. Further, when $\frac{\lambda}{\mu_L} + \frac{\mu_H}{\lambda} < 1$, it follows $\mathbf{E}\Big[T^{PS}\Big] < \mathbf{E}\Big[T^{FCFS}\Big]$.

Proof: When $\mu_L > \lambda + \mu_H$, then we necessarily have $f(t) = \frac{x_L(t)}{x_H(t)} \le \frac{\lambda}{\mu_L-\lambda-\mu_H}$ for all t. This is because during L states, $f(t)$ increases asymptoting to $\frac{\lambda}{\mu_L-\lambda-\mu_H}$, while during H states $f(t)$ decreases asymptoting to 0 (when $\mu_L < \lambda + \mu_H$, $f(t)$ can increase arbitrarily during L states). We can thus obtain an upper bound by assuming $\frac{\mathbf{E}\left[N_L^{PS}\right]}{\mathbf{E}\left[N_H^{PS}\right]} = \frac{\lambda}{\mu_L-\lambda-\mu_H}$. Combining this with the relations $\frac{\mathbf{E}\left[N_L^{PS}\right]}{\mu_L} + \frac{\mathbf{E}\left[N_H^{PS}\right]}{\mu_H} = \mathbf{E}[W]$, and $\mathbf{E}\left[T^{PS}\right] = \frac{1}{\lambda}(\mathbf{E}\left[N_L^{PS}\right] + \mathbf{E}\left[N_H^{PS}\right])$ gives the bound in the corollary. Noting that $\mathbf{E}\left[T^{FCFS}\right] = \mathbf{E}[W]$, the final observation of the corollary follows. ■

2.8 ROS

Analysis of ROS parallels that of PS. Let $x_L(t)$ and $x_H(t)$ denote the number of class L and class H jobs in the system at time t under the mean field approximation. Let $w(t) = \frac{x_L(t)}{\mu_L} + \frac{x_H(t)}{\mu_H}$ denote the workload in the system at time t. We can approximate the dynamics of the system under ROS during L states as:

$$\frac{dx_L}{dt} = \lambda - \frac{x_L}{\frac{x_L}{\mu_L} + \frac{x_H}{\mu_H}} \; ; \quad \frac{dx_H}{dt} = -\frac{x_H}{\frac{x_L}{\mu_L} + \frac{x_H}{\mu_H}}$$

and during H states as:

$$\frac{dx_L}{dt} = -\frac{x_L}{\frac{x_L}{\mu_L} + \frac{x_H}{\mu_H}} \; ; \quad \frac{dx_H}{dt} = \lambda - \frac{x_H}{\frac{x_L}{\mu_L} + \frac{x_H}{\mu_H}}$$

We now give intuition for the above equations. Consider a period of time where there are n total departures, but during which the ratio $\frac{x_L}{x_H}$ remains constant. Under ROS, of these n departures, an expected $\frac{x_L}{x_L+x_H}n$ are of class L, and $\frac{x_H}{x_L+x_H}n$ are of class H. The duration of this period is in expectation $\frac{\frac{x_L}{\mu_L}+\frac{x_H}{\mu_H}}{x_L+x_H} \cdot n$. Therefore, the departure rate of type L jobs is $\frac{\text{num departures}}{\text{total duration}} = \frac{x_L}{\frac{x_L}{\mu_L}+\frac{x_H}{\mu_H}}$. Solving the above set of ODEs, we obtain the following:

Theorem 7 *The dynamics of the number of class L and H jobs during the L states under the mean field approximation and ROS scheduling satisfies:*

$$x_H(t) = x_H(0)\left(\frac{(w(0)-r_Lt)^+}{w(0)}\right)^{\frac{1}{1-\frac{\lambda}{\mu_L}}}$$

$$x_L(t) = \mu_L(w(0)-r_Lt)^+ - x_H(0)\frac{\mu_L}{\mu_H}\left(\frac{(w(0)-r_Lt)^+}{w(0)}\right)^{\frac{1}{1-\frac{\lambda}{\mu_L}}}$$

and during H states satisfies:

$$x_L(t) = x_L(0)\left(\frac{w(0)-r_H t}{w(0)}\right)^{\frac{-1}{\frac{\lambda}{\mu_H}-1}}$$

$$x_H(t) = \mu_H(w(0)-r_H t) - x_L(0)\frac{\mu_H}{\mu_L}\left(\frac{w(0)-r_H t}{w(0)}\right)^{\frac{-1}{\frac{\lambda}{\mu_H}-1}}.$$

The following theorem compares ROS and PS under the mean field regime.

Theorem 8 *Under the mean field approximation, the number of jobs in the system under ROS is stochastically larger than the number of jobs in the system under PS.*

Proof: We first couple the environment processes (and hence also the workload processes) of the ROS and PS systems. We will now show that

$$n^{PS}(t) = x_L^{PS}(t) + x_H^{PS}(t) \le x_L^{ROS}(t) + x_H^{ROS}(t) = n^{ROS}(t), \quad \forall t.$$

To prove the above, we will use the fact that if for some function $h(t)$, $h(0) = 0$ and $\frac{dh(t)}{dt} > 0$ whenever $h(t) < 0$, then $h(t) \ge 0$, $\forall t \ge 0$. We will consider $h(t) = n^{ROS}(t) - n^{PS}(t)$. Let $w(t) = \frac{x_L^{PS}}{\mu_L} + \frac{x_H^{PS}}{\mu_H} = \frac{x_L^{ROS}}{\mu_L} + \frac{x_H^{ROS}}{\mu_H}$.

$$\begin{aligned}
\frac{d}{dt}(n^{ROS} - n^{PS}) &= \frac{x_L^{PS}\mu_L + x_H^{PS}\mu_H}{x_L^{PS} + x_H^{PS}} - \frac{x_L^{ROS} + x_H^{ROS}}{w} \\
&= \frac{x_L^{PS}\mu_L + \mu_H\left(\mu_H w - \frac{\mu_H}{\mu_L}x_L^{PS}\right)}{x_L^{PS} + \mu_H w - \frac{\mu_H}{\mu_L}x_L^{PS}} - \frac{x_L^{ROS} + \left(\mu_H w - \frac{\mu_H}{\mu_L}x_L^{ROS}\right)}{w} \\
&= \frac{w\frac{\mu_L-\mu_H}{\mu_L}(x_L^{PS}\mu_L - x_L^{ROS}\mu_H) - x_L^{PS}x_L^{ROS}\frac{(\mu_L-\mu_H)^2}{\mu_L^2}}{w(x_L^{PS} + x_H^{PS})} \\
&= \frac{(\mu_L-\mu_H)}{\mu_L^2}\cdot\frac{w\mu_L(x_L^{PS}\mu_L - x_L^{ROS}\mu_H) - x_L^{PS}x_L^{ROS}(\mu_L-\mu_H)}{w(x_L^{PS} + x_H^{PS})}
\end{aligned}$$

For a fixed w, x_L^{ROS}, the above is an increasing function of x_L^{PS} in the interval $[0, w\mu_L)$ (in fact in $[0,\infty)$, if we consider an extended definition), and the value at $x_L^{PS} = x_L^{ROS}$ (and hence at $n^{PS} = n^{ROS}$) is given by $\frac{(\mu_L-\mu_H)^2}{\mu_L^2}\cdot\frac{(w\mu_L - x_L^{PS})x_L^{PS}}{w(x_{PS}^L + x_H^{PS})} > 0$. Therefore, $\frac{d}{dt}(n^{ROS} - n^{PS}) > 0$, whenever $n^{ROS} \le n^{PS}$, proving the theorem. ■

We in fact conjecture that FCFS has a larger mean response time than ROS, and thus under correlated job sizes FCFS>ROS>LCFS while all three are equal when job sizes are *i.i.d.*. We do see this ordering in our simulations, even for the case when $\mu_H > \lambda$ and hence system is always stable, but are so far unable to prove a comparison result for FCFS and ROS. Additionally, this would imply FCFS>PS, which currently we only prove to be true when $\frac{\lambda}{\mu_L} + \frac{\mu_H}{\lambda} < 1$.

3 Exact Numerical Analysis of FCFS, OPT, P-LCFS and LCFS for $0 < q < 1$

In this section we present numerical algorithms for exact analysis of a few scheduling policies for general values of the parameter α, and for general settings of the parameters μ_L, μ_H, λ and p.

3.1 FCFS

We begin by noting that the delay of a class L job, by conditional PASTA [25], is given by the stationary workload conditioned on the system being in state L (similarly for class H job). The next theorem provides the expressions for the Laplace transforms of these random variables. The mean stationary workloads during L and H states (and hence the mean delay of class L and H jobs, respectively) can be obtained by differentiating the transforms at $s = 0$.

Theorem 9 *Let $\widetilde{W_L}(s)$ and $\widetilde{W_H}(s)$ denote the transform for the stationary workloads during the L and H states respectively. The expressions for $\widetilde{W_L}(s)$ is given by:*

$$\widetilde{W_L}(s) = \frac{(1-\rho)\alpha m_L m_H - s m_L g_H \pi_L(0)}{\alpha_L g_H m_L + \alpha_H g_L m_H - s g_L g_H} \tag{7}$$

where,

$$m_L = \mu_L + s \quad ; \quad m_H = \mu_H + s$$
$$g_L = \mu_L - \lambda + s \quad ; \quad g_H = \mu_H - \lambda + s$$
$$\pi_L(0) = \frac{(1-\rho)\alpha(\mu_H + \xi)}{\xi(\mu_H - \lambda + \xi)}$$

and ξ denotes the unique root of the denominator of (7) in the interval $(0, +\infty)$. The quantity $\pi_L(0)$ denotes the long run fraction of time that the system is empty conditioned on being in state L. The expression for $\widetilde{W_H}(s)$ is obtained by flipping μ_H and μ_L, and α_L and α_H.

Proof: As the first step, we need to consider the transient workload in an $M/G/1$. Consider an $M/G/1$ with arrival rate λ, *i.i.d.* job sizes $X_1, X_2, \ldots$ with Laplace transform of the job size distribution given by $\mathbf{E}\left[e^{-sX_1}\right] = \widetilde{X}(s)$. We can write the following equation for the evolution of the workload $W(t)$ in this $M/G/1$:

$$W(t+\delta t) = W(t) - \delta t \mathbf{1}_{W(t)>0} + \sum_n X_n \mathbf{1}_{n\text{th arrival in }(t,t+\delta t)}$$

Let $\widetilde{W_t}(s) = \mathbf{E}\left[e^{-sW(t)}\right]$. Taking Laplace transforms in the above equation, and then letting $\delta t \to 0$,

$$\frac{d}{dt}\widetilde{W_t}(s) = \widetilde{W_t}(s)\left(s - \lambda(1 - \widetilde{X}(s))\right) - s\mathbf{Pr}[W_t = 0]$$

Let T be an Exp(ν) random variable and $\widetilde{W_T}(s) = \mathbf{E}\left[e^{-sW(T)}\right]$. Using integration by parts, we get:

$$\begin{aligned}
\widetilde{W_T}(s) &\equiv \int_{u=0}^{\infty} \widetilde{W_u}(s)\nu e^{-\nu u} du \\
&= \left[\frac{\widetilde{W_u}(s)\nu e^{-\nu u}}{-\nu}\right]_{u=0}^{\infty} + \int_{u=0}^{\infty} \frac{d\widetilde{W_u}(s)}{du} e^{-\nu u} du \\
&= \widetilde{W_0}(s) + \frac{1}{\nu}\int_{u=0}^{\infty}\left(\widetilde{W_u}(s)\left[s - \lambda(1 - \widetilde{X}(s))\right] - s\mathbf{Pr}[W_u = 0]\right)\nu e^{-\nu u} du \\
&= \widetilde{W_0}(s) + \frac{1}{\nu}\left(\widetilde{W_T}(s)[s - \lambda(1 - \widetilde{X}(s))] - s\mathbf{Pr}[W(T) = 0]\right)
\end{aligned}$$

Specializing to our problem, we obtain the following two equations by applying the above equation during L and H states, and noting that by PASTA $\widetilde{W_L}(s)$ and $\widetilde{W_H}(s)$ also denote the stationary workloads at the *ends* of L and H states, respectively:

$$\begin{aligned}
\widetilde{W_L}(s) &= \widetilde{W_H}(s) + \frac{s}{\alpha_L}\left[\widetilde{W_L}(s)\left(1 - \frac{\lambda}{\mu_L + s}\right) - \pi_L(0)\right] \\
\widetilde{W_H}(s) &= \widetilde{W_L}(s) + \frac{s}{\alpha_H}\left[\widetilde{W_H}(s)\left(1 - \frac{\lambda}{\mu_H + s}\right) - \pi_H(0)\right]
\end{aligned}$$

Eliminating $\widetilde{W_H}(s)$, and $\pi_H(0)$ by using the fact $\frac{\pi_L(0)}{\alpha_L} + \frac{\pi_H(0)}{\alpha_H} = (1-\rho)\left(\frac{1}{\alpha_L} + \frac{1}{\alpha_H}\right)$, we obtain the expression for $\widetilde{W_L}(s)$ shown in the Lemma. It now remains to determine the unknown $\pi_L(0)$. To obtain this, we note that the polynomial in the denominator of $\widetilde{W_L}(s)$ is a cubic in s which approaches $-\infty$ as $s \to \infty$. Further, the denominator is positive at $s = 0$ but negative at $s = \lambda - \mu_L < 0$. Therefore there is exactly one root of the denominator in the interval $(0, +\infty)$, which we denote by ξ, at which there is a degeneracy in the denominator. Since the transform must converge in $Re(s) > 0$, the numerator must share this root, and this gives us the unknown $\pi_L(0)$. ■

Finally, we obtain the response time of class L and class H jobs, respectively, as:

$$\begin{aligned}
\mathbf{E}\left[T_L^{FCFS}\right] &= \mathbf{E}[W_L] + \frac{1}{\mu_L} \\
\mathbf{E}\left[T_H^{FCFS}\right] &= \mathbf{E}[W_H] + \frac{1}{\mu_H}
\end{aligned}$$

Remark 12: As noted in the proof, the complexity of solving the mean response time (and indeed higher moments) conditioned on the job class via the above method is the same as that of finding the roots of a cubic polynomial.

3.2 OPT

Recall that since we are looking at policies which are size-independent but might still exploit the correlation structure, we can lower bound the response time of any policy in this class of scheduling policies by a hypothetical policy OPT that perfectly knows the class of each job. Under OPT, class L mean delay is obtained as the mean workload during the L states but with $\mu_H = \infty$. Let this quantity be denoted by $\mathbf{E}\Big[D_L^{OPT}\Big]$, which can be obtained by the results in Section 3.1. However, since OPT is a work conserving policy, if we denote the mean delay of class H jobs under OPT by $\mathbf{E}\Big[D_H^{OPT}\Big]$, then via Little's law we get:

$$\frac{\lambda p}{\mu_L}\left(\frac{1}{\mu_L}+\mathbf{E}\Big[D_L^{OPT}\Big]\right)+\frac{\lambda(1-p)}{\mu_H}\left(\frac{1}{\mu_H}+\mathbf{E}\Big[D_H^{OPT}\Big]\right)=p\mathbf{E}[W_L]+(1-p)\mathbf{E}[W_H] \tag{8}$$

where $\mathbf{E}[W_L]$ and $\mathbf{E}[W_H]$ denote the stationary mean workload during L and H states, respectively, and are independent of the scheduling policy. Finally,

$$\mathbf{E}\Big[D^{OPT}\Big]=p\mathbf{E}\Big[D_L^{OPT}\Big]+(1-p)\mathbf{E}\Big[D_H^{OPT}\Big].$$

3.3 P-LCFS

The analysis of Preemptive Last-Come-First-Served requires a different approach than FCFS – that of busy period recurrences.

Consider a tagged class L arrival. The response time of the tagged job is the duration of the busy period started by the tagged job on its arrival, only including all subsequent arriving work. We call this *a class L busy period started in state* L, and denote it by B_L^L. The superscript signifies that a job of *class* L starts the busy period, and the subscript signifies that the busy period starts in *state* L, because the tagged job's arrival necessarily occurs in state L. In general, B_s^c denotes the random variable for the busy period started by a class c job in state s. While the response time of a class L job under P-LCFS is given by B_L^L (and likewise for a class H job, B_H^H), we will later see that B_H^L and B_L^H will be needed in the analysis. The next theorem presents expressions for the expected values of these busy periods.

Theorem 10 *Let B_s^c $(c, s \in \{L, H\})$ denote the random variable for the busy period started by a class c job in state s. We use $\overline{c}/\overline{s}$ to denote the complementary class/state of c/s. The following set of recurrences solve for the expected values of the four kinds of busy periods:*

$$\mathbf{E}[B_s^c] \quad = \quad \frac{1}{\alpha_s+\lambda+\mu_c}+\frac{\alpha_s}{\alpha_s+\lambda+\mu_s}\mathbf{E}[B_{\overline{s}}^c]+\frac{\lambda}{\alpha_s+\lambda+\mu_c}\left[\mathbf{E}[B_s^s]+X_s^c\right] \tag{9}$$

$$\textit{where } X_s^c \;=\; p_{H|_s^s}\mathbf{E}[B_H^c] + p_{L|_s^s}\mathbf{E}[B_L^c]$$

Here p_{s_1,c_2} $(s_1, c_2 \in \{L, H\})$ *denote the probabilities that given the system is currently in state* s_1*, the next arrival is of class* c_2 *and are given by:*

$$p_{L,H} = 1 - p_{L,L} = \frac{\alpha}{\alpha+\lambda}(1-p)$$
$$p_{H,L} = 1 - p_{H,H} = \frac{\alpha}{\alpha+\lambda}p$$

Also, $p_{s_2|_{s_1}^{c_1}}$ *denote the probabilities that a busy period started by a job of class* c_1 *in state* s_1 *ends in state* s_2*, and are obtained by solving the following set of recurrences:*

$$p_{s|_s^c} = \frac{\mu_c}{\alpha_s+\lambda+\mu_c} + \frac{\alpha_s}{\alpha_s+\lambda+\mu_c}p_{s|_{\bar{s}}^c} + \frac{\lambda}{\alpha_s+\lambda+\mu_c}\left[p_{L|_s^s}p_{s|_L^c} + p_{H|_s^s}p_{s|_H^c}\right] \tag{10}$$

Proof: We will first show how recurrences denoted by (9) are obtained by considering the example of B_H^L, the busy period started by a class L job in state H. For this setting, (9) becomes:

$$\mathbf{E}\Big[B_H^L\Big] \;=\; \frac{1}{\alpha_H+\lambda+\mu_L} + \left(\frac{\alpha_H}{\alpha_H+\lambda+\mu_L}\mathbf{E}\Big[B_L^L\Big] + \frac{\lambda}{\alpha_H+\lambda+\mu_L}\Big[\mathbf{E}\Big[B_H^H\Big] + X_L^H\Big]\right) \tag{11}$$

$$\text{where } X \;=\; p_{H|_H^H}\mathbf{E}\Big[B_H^L\Big] + p_{L|_H^H}\mathbf{E}\Big[B_L^L\Big]$$

The first term in (11) denotes the mean time until one of the following events happens: the class L job completes service, an external arrival occurs, or the environment switches state. With probability $\frac{\mu_L}{\alpha_H+\lambda+\mu_L}$, the class L job completes service and the busy period ends, otherwise the busy period increases by the term in brackets. With probability $\frac{\alpha_H}{\alpha_H+\lambda+\mu_L}$, the environment switches state to L and the remaining busy period will be given by B_L^L. With the remaining probability, an external arrival occurs which increases the busy period in the last parentheses.

The external arrival must necessarily be of class H since it occurred during an H state. We will refer to this external arrival as the *new H* job. The new H job creates its own busy period, of type B_H^H. This busy period created by the new H job completes in some state s, where $s = H$ with probability $p_{H|_H^H}$ and $s = L$ with probability $p_{L|_H^H}$. Only after that busy period completes, can we resume the original L job, which (by memorylessness) has remaining size L and thus generates a busy period of type B_s^L, since the system state is now s. This explains the X_H^L term above.

The proof of (10) is completely analogous to the proof of (9), by conditioning on the first event following the start of the busy period, and also conditioning on the state in which the busy period of the new arrival (if it is the first event to occur) ends. ∎

Finally:

$$\mathbf{E}\Big[T_L^{P-LCFS}\Big] = \mathbf{E}\Big[B_L^L\Big] \quad ; \quad \mathbf{E}\Big[T_H^{P-LCFS}\Big] = \mathbf{E}\Big[B_H^H\Big].$$

Remark 13: The system of equations (9) is a linear system once we have all the coefficients. The system (10) to obtain those coefficients is not a linear system, and the complexity of solving (10) boils down to finding the roots of a cubic polynomial. Note that similar to (9), we could also have written recurrence relations for the Laplace transforms for the busy periods to obtain higher moments of the response time. We omit them here for lack of space.

3.4 LCFS

Under the LCFS scheduling discipline, suppose that a tagged arrival sees a job, j, in service with remaining service requirement, j_e. Then the tagged arrival doesn't get to serve until the completion of a busy period started by j_e. Translating to our model, if a class L (respectively, H) job on arrival finds the server empty, then its time in queue is zero. However if it finds a class c job at the server, then its time in queue is distributed as B_L^c (respectively B_H^c), whose mean is given by Theorem 10. To complete the analysis of LCFS, we therefore only need to find the fraction of jobs of either class that (i) find the server idle, (ii) find the server busy with a class L job, or (iii) find the server busy with a class H job. Let f_L^{idle}, f_L^L and f_L^H denote the above fractions, for class L arrivals, where a subscript of H would be used for class H arrivals. Below we give a step-by-step method for determining these fractions.

Step 1: Let $N_s^c(L)$ denote the mean number of class L arrivals during the (non-preemptive) service of a class c job started in state s, and $N_s^c(H)$ similarly denote the mean number of class H arrivals. The first step is to determine the $N_s^c(L)$ and $N_s^c(H)$ for $s, c \in \{L, H\}$. We illustrate the recurrences that solve these by showing the case $c = L$.

$$\begin{aligned} N_L^L(L) &= \frac{\lambda}{\alpha_L + \lambda + \mu_L}\Big(1 + N_L^L(L)\Big) + \frac{\alpha_L}{\alpha_L + \lambda + \mu_L} N_H^L(L) \\ N_H^L(L) &= \frac{\lambda}{\alpha_H + \lambda + \mu_L} N_H^L(L) + \frac{\alpha_H}{\alpha_H + \lambda + \mu_L} N_L^L(L) \end{aligned}$$

Step 2: Let b_L denote the fraction of idle periods that end with a class L arrival (equivalently the fraction of busy periods for the server that begin in state L). By writing a 2-state Markov chain (with states L and H) to track the class of the arrival at the beginning of busy periods, we have:

$$\Big(p_{H|_L^L} p_{H,H} + p_{L|_L^L} p_{L,H}\Big)\, b_L = \Big(p_{L|_H^H} p_{L,L} + p_{H|_H^H} p_{H,L}\Big)\,(1 - b_L)$$

where we we transition from state L to state H if the busy period started by the class L job in state L ends in state H and the next arrival is of class H, and analogously for transitions from state H to state L.

Step 3: For any non-preemptive scheduling policy, we consider 3 kinds of events: idle periods, service of a class L job, and service of a class H job. Let e^{idle}, e^L and e^H denote the fraction of each of the above event types, respectively. By observing that the mean duration of each occurrence of the above event types is $\frac{1}{\lambda}$, $\frac{1}{\mu_L}$ and $\frac{1}{\mu_H}$ respectively, and the total fraction of time spent in each of the event types is $(1-\rho)$, $\frac{\lambda}{\mu_L}$ and $\frac{\lambda}{\mu_H}$, respectively, we can obtain e^{idle}, e^L and e^H by solving:

$$\frac{\lambda p}{\mu_L} = \frac{\frac{e^L}{\mu_L}}{\frac{e^{idle}}{\lambda} + \frac{e^L}{\mu_L} + \frac{e^H}{\mu_H}} \;; \quad \frac{\lambda(1-p)}{\mu_H} = \frac{\frac{e^H}{\mu_H}}{\frac{e^{idle}}{\lambda} + \frac{e^L}{\mu_L} + \frac{e^H}{\mu_H}}; \quad 1-\rho = \frac{\frac{e^{idle}}{\lambda}}{\frac{e^{idle}}{\lambda} + \frac{e^L}{\mu_L} + \frac{e^H}{\mu_H}}$$

Step 4: Finally we obtain f_L^{idle}, f_L^L and f_L^H as follows. For succinctness, define the auxiliary variables:

$$g_L^L = f_L^{idle} + f_L^L p_{L|_L^L} + f_L^H p_{L|_L^H}$$
$$g_H^H = f_H^{idle} + f_H^L p_{H|_H^L} + f_H^H p_{H|_H^H}$$

and $g_H^L = 1 - g_L^L$, $g_L^H = 1 - g_H^H$. Here g_s^c denotes the probability that under LCFS, a class c arrival begins service in state s, and is obtained by conditioning on the state on arrival. Now we can write the following set of equations to solve for the desired f_L^{idle}, f_L^L and f_L^H:

$$f_L^{idle} = \frac{e^{idle} b_L}{N^{total}(L)} \;; \quad f_L^L = \frac{e^L[g_L^L N_L^L(L) + g_H^L N_H^L(L)]}{N^{total}(L)} \;; \quad f_L^H = \frac{e^H[g_L^H N_L^H(L) + g_H^H N_H^H(L)]}{N^{total}(L)}$$

where

$$N^{total}(L) = e^{idle} b_L + e^L[g_L^L N_L^L(L) + g_H^L N_H^L(L)] + e^H[g_L^H N_L^H(L) + g_H^H N_H^H(L)]$$

Similarly, there are three more equations with $f_H^{(\cdot)}$ on the LHS. In the above equations, $N^{total}(L)$ denotes the expected number of class L arrivals during a random (average) event. The numerator of f_L^H, for example, denotes the contribution to $N^{total}(L)$ only due to the arrivals during events corresponding to service of class H jobs. Thus, f_L^H denotes the fraction of all class L arrivals that occur while the server is busy with a class H job.

Finally,

$$\mathbf{E}\Big[T_L^{LCFS}\Big] = \frac{1}{\mu_L} + f_L^L \mathbf{E}\Big[B_L^L\Big] + f_L^H \mathbf{E}\Big[B_L^H\Big]$$
$$\mathbf{E}\Big[T_H^{LCFS}\Big] = \frac{1}{\mu_H} + f_H^L \mathbf{E}\Big[B_H^L\Big] + f_H^H \mathbf{E}\Big[B_H^H\Big].$$

4 Evaluation via Simulations

While Section 2 provided fluid asymptotics as $\alpha \to 0$ for a wide range of size-independent scheduling policies, we are only able to perform exact analysis of the case $0 < \alpha < \infty$ for a smaller subset (FCFS, OPT, LCFS, P-LCFS), via algorithms proposed in Section 3. This section studies the full range of policies via simulation for all α. We start with simulation and numerical results for our MMAP model and then present results for trace-based simulations.

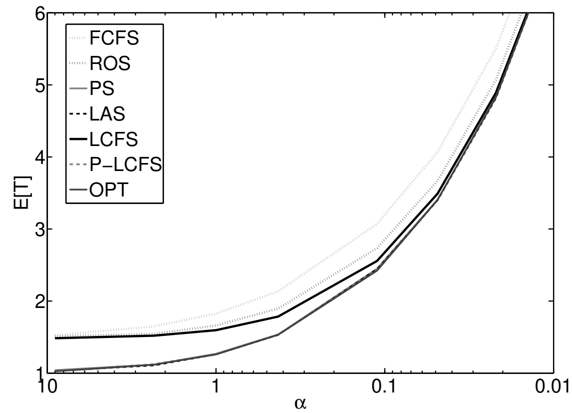
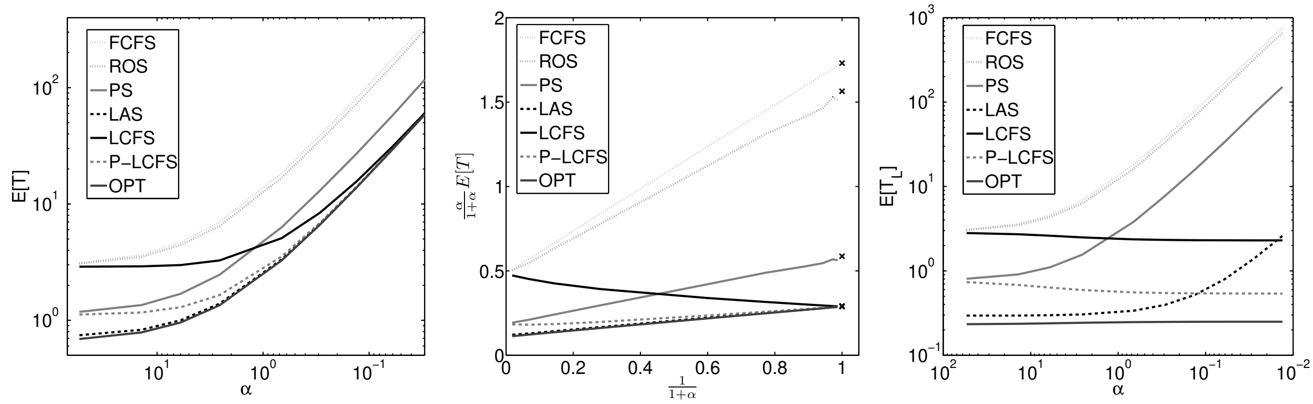


Figure 2: Effect of job size correlation when $\mu_H > \lambda$. The parameters chosen were $\mu_L = 50.73, \mu_H = 1.0055, p = 0.5073, \lambda = 1$ ($\rho = 0.5, C^2 \approx 2.9$).

MMAP under No transient overload: In Figure 2, we see the effect of correlation on scheduling policies when $\mu_H > \lambda$, so that there is no transient overload in H states. We see that for moderate α, the mean response times of the different scheduling policies range from $\mathbf{E}[T] = 1$ to about $\mathbf{E}[T] = 1.5$, with FCFS being the worst and LAS being the best. As α decreases, we see that the relative performance difference between scheduling policies begin to vanish ($\mathbf{E}[T]$ ranges from 6.9 to 7.5 for $\alpha \approx 0.01$). This behavior as $\alpha \to 0$ is consistent with Theorem 1. Observe also that while FCFS, ROS and LCFS are equal at the two extremes ($\alpha = \infty$ and $\alpha \to 0$), for $0 < \alpha < \infty$ they are ordered as FCFS>ROS>LCFS with respect to $\mathbf{E}[T]$.

MMAP under Transient overload: Figure 3 shows the effect of correlation in the more interesting case of $\mu_H < \lambda$, implying that there is transient overload during the H states. Figure 3(a) shows the $\mathbf{E}[T]$ vs. α curves for the different scheduling policies. The x-axis shows α on a log-scale so as to clearly illuminate both low and high α values. We see that FCFS is the worst policy and LAS is optimal or close to optimal throughout the range of α shown. On the other hand, P-LCFS starts out equal to PS when $\alpha = \infty$ and is clearly suboptimal; yet for low α (high correlation), P-LCFS approaches and even overtakes LAS, and becomes optimal. This is consistent with Theorem 3. Similarly, LCFS starts out equal to FCFS when $\alpha = \infty$ and is worst in performance, but becomes optimal as $\alpha \to 0$, again confirming Theorem 3.



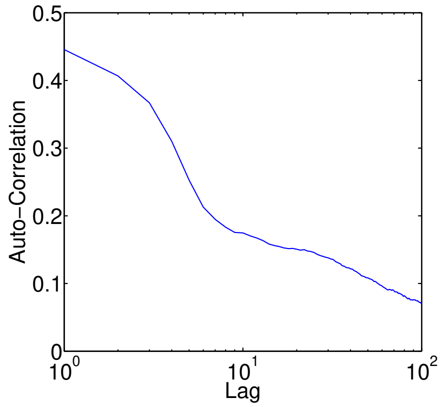
(a) $\mathbf{E}[T]$ vs. α (b) The fluid asymptote (c) $\mathbf{E}[T_L]$ vs. α

Figure 3: Effect of job size correlation when $\mu_H < \lambda$. The parameters chosen were $\mu_L = 10, \mu_H = 1, p = 0.95, \lambda = 6$ ($\rho = 0.87, C^2 \approx 4.66$).

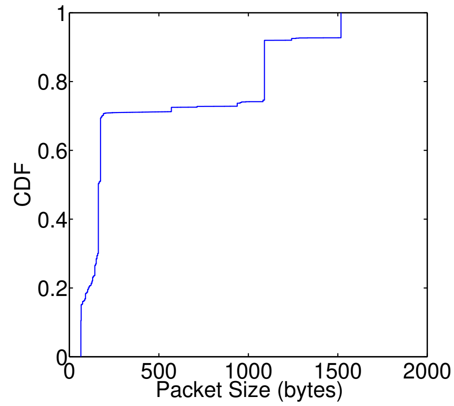
A major difference between Figure 3(a) (transient overload) and Figure 2 (no overload) is that the policies clearly do not converge to each other in Figure 3 as $\alpha \to 0$, whereas they do in Figure 2. Furthermore, for each policy π in Figure 3(a), the $\mathbf{E}[T]$ curve asymptotes to a line on the plotted scale, which corresponds to $\mathbf{E}[T^\pi] \sim \frac{K^\pi}{\alpha}$ as in Lemma 3. Thus the mean response times grow unboundedly as $\alpha \to 0$, unlike in Figure 2.

Figure 3(b) verifies the expressions for K^π obtained from our asymptotic analysis by showing $\left(\frac{\alpha}{1+\alpha}\right)\mathbf{E}[T]$ as a function of $\frac{1}{1+\alpha}$. We choose to scale $\mathbf{E}[T]$ by $\frac{\alpha}{1+\alpha}$ (instead of α) to show the results for $\alpha = \infty$ asymptote and the $\alpha \to 0$ asymptote in the same plot. In the former case, $\lim_{\alpha\to\infty} \frac{\alpha}{1+\alpha}\mathbf{E}[T^\pi] = \mathbf{E}[T^\pi]$ and in the latter case $\lim_{\alpha\to 0} \frac{\alpha}{1+\alpha}\mathbf{E}[T^\pi] = \lim_{\alpha\to 0} \alpha\mathbf{E}[T^\pi] = K^\pi$. The x-axis shows $\frac{1}{1+\alpha}$ which is bounded between 0 and 1 (unlike α). The $\alpha\mathbf{E}[T^\pi]$ curves clearly converge to the analytically obtained values of K^π which are marked with a small $\mathbf{x}$. In the limit $\alpha \to 0$, $\mathbf{E}[T]$ for the different policies follows the order LCFS = P-LCFS < LAS < PS < ROS < FCFS. Note the difference between LAS and LCFS = P-LCFS as $\alpha \to 0$ is very slight; this contrasts with the results in Figure 1 where the difference between LAS and P-LCFS was significant.

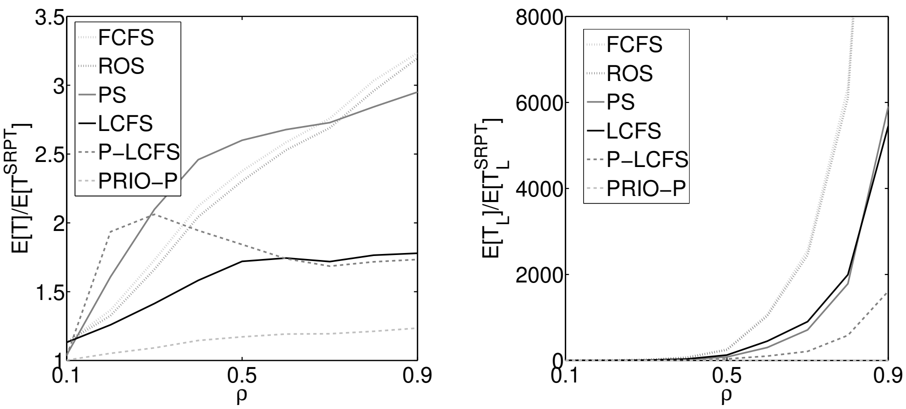
Figure 3(c) shows mean response time for "little" (class L) jobs, denoted $\mathbf{E}[T_L]$, versus α. For the L jobs, there is a wide range (several orders of magnitude difference) in performance across policies. Several policies (FCFS, ROS, PS, LAS) show increases in $\mathbf{E}[T_L]$ proportional to $\frac{1}{\alpha}$ (though this is less obvious in the case of LAS, since convergence is slower for this policy); however, other policies (LCFS, P-LCFS) show a decrease in $\mathbf{E}[T_L]$ as α decreases, as pointed out in Remark 6. Under the first group of polices, $\mathbf{E}[T_L]$ suffers from increased correlation, because L jobs are affected by H jobs. For LCFS and P-LCFS, this is not the case, since an L job is only affected by H jobs if the H job arrives during the L job's busy period. This happens with probability proportional to α, which becomes zero as $\alpha \to 0$.



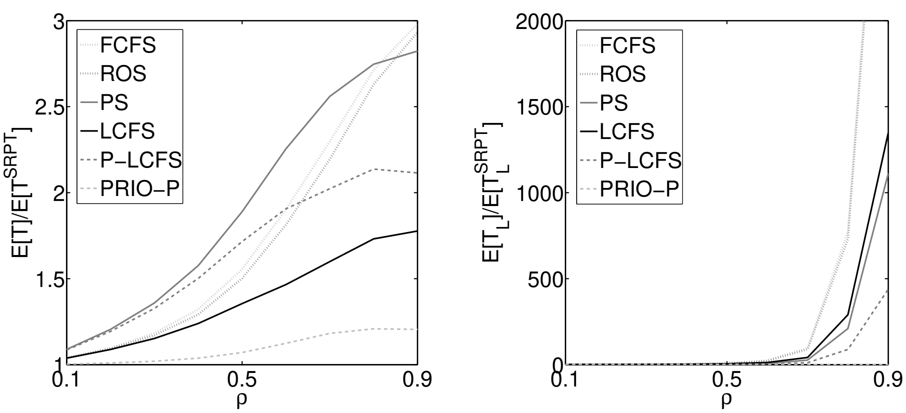
(a) Bellcore trace acf



(b) Bellcore Job size cdf

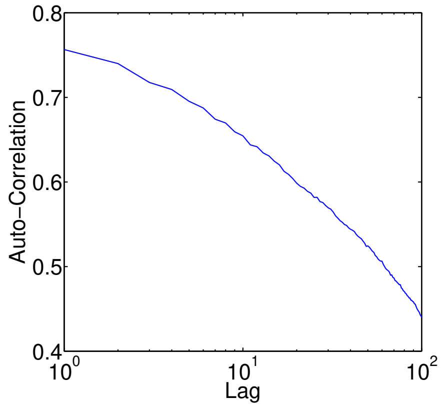


(c) Results for arrivals from trace

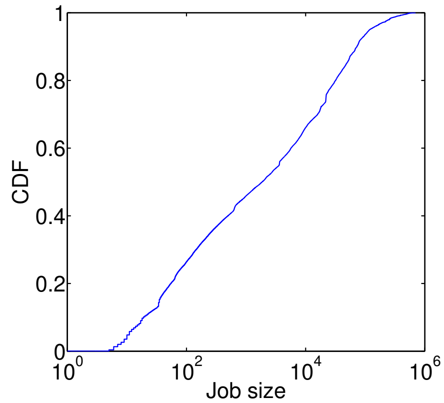


(d) Results for Poisson arrivals

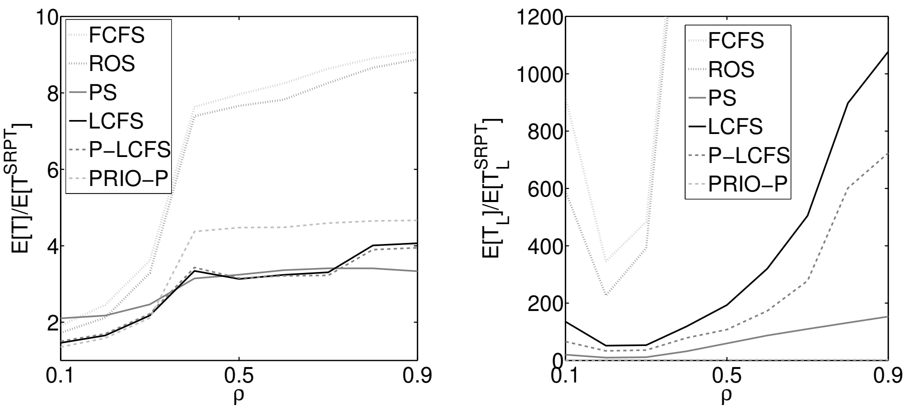
Figure 4: Trace-based experiments: Simulation results for the Bellcore trace. The top-left plot shows the autocorrelation function for the sequence of job sizes; the top-right plot shows the cdf of the job size distribution; the center row plots show the performance (as the ratio of $\mathbf{E}[T]$ to $\mathbf{E}\left[T^{SRPT}\right]$, and of $\mathbf{E}[T_L]$ to $\mathbf{E}\left[T_L^{SRPT}\right]$, respectively) when the interarrival times are taken from the trace; the two bottom row plots show the performance obtained by creating a synthetic Poisson arrival process.



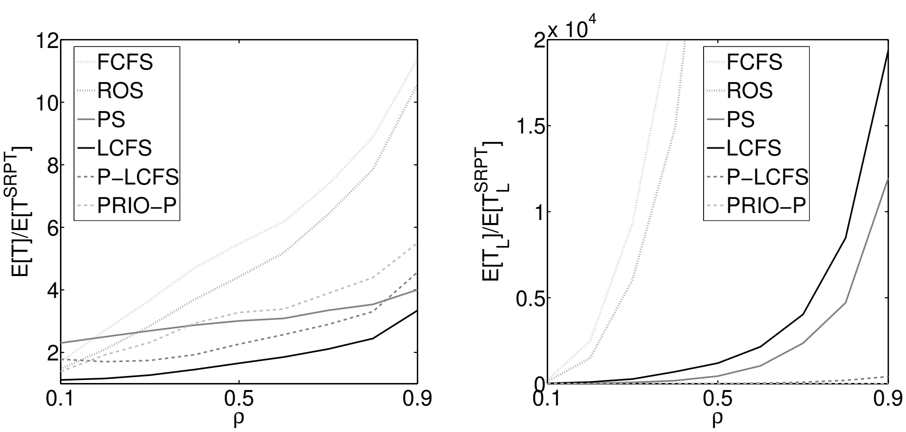
(a) SHARCNET trace acf



(b) SHARCNET Job size cdf



(c) Results for arrivals from trace



(d) Results for Poisson arrivals

Figure 5: Trace-based experiments: Simulation results for the SHARCNET trace. The top-left plot shows the autocorrelation function for the sequence of job sizes; the top-right plot shows the cdf of the job size distribution; the center row plots show the performance (as the ratio of $\mathbf{E}[T]$ to $\mathbf{E}\left[T^{SRPT}\right]$, and of $\mathbf{E}[T_L]$ to $\mathbf{E}\left[T_L^{SRPT}\right]$, respectively) when the interarrival times are taken from the trace; the two bottom row plots show the performance obtained by creating a synthetic Poisson arrival process.

Trace-based experiments: While we garnered useful intuition by analyzing the MMAP correlation model, it is not obvious to what extent our results would extend to real-world applications. To investigate this, we consider two very different traces, one involving packets sizes (Bellcore) and a second involving supercomputing job sizes (SHARCNET). We have simulated FCFS, ROS, PS, LCFS and P-LCFS. In addition, we simulate PRIO-P, which gives preemptive priority to class L jobs and hence is similar to the OPT policy, but is not necessarily the optimal size-independent policy because class L and H jobs are no longer Exponentially distributed. We also simulate SRPT (Shortest Remaining Processing Time) policy, and our plots show the mean response time under the simulated policies normalized by the mean response time under SRPT scheduling.

Figure 4 shows the results of our experiments with a trace of packet sizes seen on the Bellcore Ethernet [11]. The autocorrelation function of packet sizes (Figure 4(a)) shows significant sequential job size correlation – the lag-1 correlation is approximately 0.45 with correlation persisting even at lags of up to 100 (unlike MMAP model where the correlation decreases exponentially in lag). Figure 4(b) shows the job size distribution which is almost a trimodal distribution. To perform the simulations, we modify the base trace as follows: In the first set of experiments (Figure 4(c)), we scale the interarrival times from the trace to vary the 'load'. In the second set of experiments (Figure 4(d)), we keep the same sequence of job sizes as the original trace, but create a new Poisson arrival process to eliminate the effect of correlations in the arrival process (the arrival process is bursty) and eliminate correlations between interarrival times and job sizes (the correlations between a job size and immediately following interarrival time is -0.15). We see that with respect to $\mathbf{E}[T]$, the ordering of the policies largely obeys FCFS≈ROS ≈PS>LCFS≈P-LCFS>PRIO-P>SRPT. This is consistent with the ordering we obtained via analysis using the MMAP correlation model. We also see that $\mathbf{E}\left[T^{FCFS}\right]$ is up to 1.8 times worse than $\mathbf{E}\left[T^{LCFS}\right]$ which contrasts with the uncorrelated case where they are equal. We also investigate the effect of scheduling on the little jobs by defining "little" to be packets of size less than 400 bytes. Under our criterion, the L jobs make up 70% of the packets, and 25% of the total bytes. We find that $\mathbf{E}\left[T_L^{FCFS}\right]$ is up to 3 to 4 times worse than $\mathbf{E}\left[T_L^{LCFS}\right]$ and almost 10 times worse than $\mathbf{E}\left[T_L^{P-LCFS}\right]$. We also see that PS outperforms LCFS but not P-LCFS in terms of $\mathbf{E}[T_L]$. This can be explained by the fact that under the uncorrelated case PS and P-LCFS have identical performance and outperform LCFS which suffers due to job size variability. Under moderate correlation, we see a behavior that is the mixture of uncorrelated and high-correlation cases: job size variability is still hurting class L jobs under LCFS and thus gives them worse performance than PS, however due to correlation P-LCFS is able to perform better than PS. The same observations hold under a Poisson arrival process, but the gains are more moderate. This suggests that in the presence of cross-correlations and bursty arrivals, the effect of scheduling will be even more pronounced.

Figure 5 shows the results for the SHARCNET trace [1], which is a supercomputing workload. Here job size is defined as the run time of jobs submitted to the server, and there is very high autocorrelation in the sequence of job sizes (lag-1 autocorrelation is over .7, and even lag-100 correlation is over .4). The ordering of policies with respect to $\mathbf{E}[T]$ largely obeys FCFS>ROS>PRIO-P>PS≈P-LCFS≈LCFS>SRPT. The gains of utilizing LCFS instead of FCFS for the SHARCNET trace are even more significant, as the ratio of $\mathbf{E}\left[T^{FCFS}\right]$ to $\mathbf{E}\left[T^{LCFS}\right]$ can be over 2. For the SHARCNET trace, we defined "little" jobs as those smaller than 54000 seconds (86% jobs, 25% of total load). There is again a significant difference between $\mathbf{E}\left[T_L^{FCFS}\right]$ and $\mathbf{E}\left[T_L^{LCFS}\right]$, up to 4X when scaling the original interarrival times, and 15X to 20X when the arrival process has been converted to a Poisson process. Comparing $\mathbf{E}[T_L]$ for LCFS, PS and P-LCFS, we see that PS does better than LCFS which can be explained by the presence of job size variability (our MMAP simulation results also suggest that for moderate correlations, PS still outperforms LCFS). However the ordering of PS and P-LCFS under arrival times from the SHARCNET trace switches when a Poisson arrival process is considered. While under a Poisson arrival process, PS performs worse than P-LCFS as predicted by our analysis of the MMAP correlation model, under the arrival sequence from the SHARCNET trace, PS outperforms P-LCFS. This suggests that the correlation between the arrival times (the SHARCNET arrival sequence has extremely bursty and variable interarrival times compared to the Bellcore trace) is also an important aspect to consider to fully understand the the effect of scheduling under correlated traffic pattern.

5 Conclusions

To the best of our knowledge, this is the first paper to study analytically how common scheduling policies, like PS, LAS, ROS, PLCFS, LCFS, etc. are affected by correlation among consecutive job sizes. We find the ranking of scheduling policies, from highest to lowest mean response time ($\mathbf{E}[T]$), changes dramatically under correlation: LCFS which performs poorly under no correlation becomes optimal among size-independent policies under high correlation; the optimal size-independent policy for i.i.d. job sizes, LAS, becomes sub-optimal under high correlation; the mean response times of policies which are insensitive to job-size variability when job sizes are i.i.d., like PS and P-LCFS, now depend on the entire job-size distribution; to cite a few examples. When examining the mean response time of "little" jobs only ($\mathbf{E}[T_L]$), the change in ranking is even more dramatic, with correlation actually making some policies like LCFS and P-LCFS perform better, and making other policies like LAS perform far worse.

We have only scratched the surface of how correlation in job sizes affects performance. First, our correlation model is very simple, chosen specifically for analytical tractability and to gain insights; extending the results presented here to richer models is left for future work. Second, while this paper shows that LCFS performs optimally among size-independent policies under very high correlation, the paper does not answer the question of which scheduling policy is best under moderate correlation. Furthermore, we have not even explored policies which might exploit the correlation structure to improve performance. Third, our model only captures correlations in consecutive job sizes, but we believe that the techniques introduced herein can be applied to understanding the effect of all three types of correlation on the performance of scheduling policies.

References

[1] `http://www.cs.huji.ac.il/labs/parallel/workload/`.

[2] I. J. B. F. Adan and V. G. Kulkarni. Single-server queue with Markov-dependent inter-arrival and service times. *QUESTA*, 45:113–134, 2003.

[3] G. L. Choudhury, A. Mandelbaum, M. I. Reiman, and W. Whitt. Fluid and diffusion limits for queues in slowly changing environments. *Stoch. Mod.*, 13:121–146, 1997.

[4] I. Cidon, R. Guérin, A. Khamisy, and M. Sidi. Analysis of a correlated queue in a communication system. In *INFOCOM'93*, pages 209–216, 1993.

[5] J. Cohen. *The single server queue*. North Holland, 1969.

[6] R. Conway, W. Maxwell, and M. Miller. *Theory of Scheduling*. Addision-Wesley, 1967.

[7] M. E. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. In *ACM SIGMETRICS'96*, pages 160–169, May 1996.

[8] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Trans. on Networking*, 4:209–223, 1996.

[9] D. G. Feitelson. Packing schemes for gang scheduling. In *IPPS '96*, pages 89–110, London, UK, 1996. Springer-Verlag.

[10] K. Fendick, V. Saksena, and W. Whitt. Dependence in packet queues. *IEEE Trans. Commun.*, 37:1173–1183, 1989.

[11] H. J. Fowler, W. E. Leland, and B. Bellcore. Local area network traffic characteristics, with implications for broadband network congestion management. *IEEE Journal on Selected Areas in Communications*, 9:1139–1149, 1991.

[12] S. Ghosh and M. Squillante. Analysis and control of correlated web server queues. *Computer Communications*, 27(18):1771–1785, 2004.

[13] L. Kleinrock. *Communication nets; stochastic message flow and delay*. Dover Publications, Incorporated, 1972.

[14] T. Kurtz. *Approximation of Population Processes*. SIAM Press, 1981.

[15] H. Li, D. Groep, and L. Wolters. Workload characteristics of a multi-cluster supercomputer. pages 176–193. Springer Verlag, 2004.

[16] M. Livny, B. Melamed, and A. K. Tsiolis. The impact of autocorrelation on queuing systems. *Manage. Sci.*, 39(3):322–339, 1993.

[17] N. Mi, G. Casale, and E. Smirni. Scheduling for performance and availability in systems with temporal dependent workloads. In *DSN'08*, pages 336–345, 2008.

[18] N. Mi, G. Casale, Q. Zhang, A. Riska, and E. Smirni. Autocorrelation-driven load control in distributed systems. In *MASCOTS'09*, 2009.

[19] V. Paxson and S. Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, 1995.

[20] R. Righter, J. G. Shanthikumar, and G. Yamazaki. On extremal service disciplines in single-stage queueing systems. *J. Appl. Probab.*, 27(2):409–416, 1990.

[21] A. Riska, M. Squillante, S.-Z. Yu, Z. Liu, and L. Zhang. Matrix-analytic analysis of a $MAP/PH/1$ queue fitted to web server data. *Matrix-Analytic Methods: Theory and Applications*, pages 335–356, 2002.

[22] E. Smirni, Q. Zhang, N. Mi, A. Riska, and G. Casale. New results on the performance effects of autocorrelated flows in systems. In *IEEE IPDPS'07*, pages 1–6, 2007.

[23] B. Song, C. Ernemann, and R. Yahyapour. Parallel computer workload modeling with markov chains. In *Proc. of the 10th Job Scheduling Strategies for Parallel Processing (JSSPP)*, pages 47–62. Springer, 2004.

[24] M. S. Squillante, D. D. Yao, and L. Zhang. Internet traffic: periodicity, tail behavior, and performance implications. *System performance evaluation: methodologies and applications*, pages 23–37, 2000.

[25] E. van Doorn and J. Regterschot. Conditional PASTA. *Oper. Res. Lett.*, 7:229–232, 1988.

[26] Q. Zhang, N. Mi, A. Riska, and E. Smirni. Load unbalancing to improve performance under autocorrelated traffic. In *ICDCS'06*, Lisboa, Portugal, 2006.

A Asymptotic Expressions for Mean Busy Periods

Busy periods form the core of the analysis for scheduling policies, and therefore we deal with the problem of finding busy periods in as much generality as possible.

We consider a system with an environment controlled by a 2 state Markov chain with states L and H. The time spent in state L during each visit is $\text{Exp}(\alpha_L)$ and time spent in state H is $\text{Exp}(\alpha_H)$. Let $\alpha = \alpha_L + \alpha_H$, $p = \frac{\alpha_H}{\alpha}$. The arrivals occur at a rate λ in each state. The arrivals during an L state have *i.i.d.* general job sizes and we use S_L to denote such a generic random variable. Similarly, the arrivals during an H state have *i.i.d.* general job sizes distributed with the same distribution as a random variable S_H. We will assume $\mathbf{E}[S_L] < \mathbf{E}[S_H]$. We index this system by α.

The scaling: We consider a sequence of systems, indexed by α, obtained by setting the switching rates as $\alpha_L + \alpha_H = \alpha$, while fixing $p = \frac{\alpha_H}{\alpha}$. We start the αth system in a prescribed state with initial workload (a random variable) denoted by W_α. We will say that the workload sequence W_α is $\Theta(g(\alpha))$ if the sequence $\left\{\frac{W_\alpha}{g(\alpha)}\right\}$ is uniformly integrable and $\lim_{\alpha\to 0} \frac{W_\alpha}{g(\alpha)} \xrightarrow{d} \overline{W}$, where $\overline{W}$ is some non-degenerate random variable. Similarly, we say $W_\alpha = o(h(\alpha))$ or $\omega(h(\alpha))$ if $W_\alpha = \Theta(g(\alpha))$, and $\lim_{\alpha\to 0} \frac{g(\alpha)}{h(\alpha)} = 0$, or $\lim_{\alpha\to 0} \frac{h(\alpha)}{g(\alpha)} = 0$, respectively.

Goal: Let $B_L(W_\alpha)$ and $B_H(W_\alpha)$ denote the random variables for the busy periods started by work W_α in states L and H, respectively, in the αth system. We will be interested in obtaining the mean busy period in the asymptotic regime $\alpha \to 0$. That is, we are interested in obtaining the dominant term in $\mathbf{E}[B_L(W_\alpha)]$ or $\mathbf{E}[B_H(W_\alpha)]$, as the switching rate $\alpha \to 0$.

Notation: $\widetilde{S_L}(s) = \mathbf{E}\left[e^{-sS_L}\right]$; $\widetilde{S_H}(s) = \mathbf{E}\left[e^{-sS_H}\right]$

$$r_L = 1 - \lambda\mathbf{E}[S_L] \ ; \ r_H = 1 - \lambda\mathbf{E}[S_H]$$
$$\rho = \lambda(p\mathbf{E}[S_L]+(1-p)\mathbf{E}[S_H])$$

We first present the theorems on asymptotic expressions for the mean busy periods. After presenting the theorems, we present a brief proof sketch to elucidate how the theorems were derived, and then present detailed proofs. Theorem 11 considers the case $\lambda\mathbf{E}[S_H] > 1$, and Theorem 12 considers the case $\lambda\mathbf{E}[S_H] < 1$.

Theorem 11 *Let $r_H < 0$. That is, the system is under temporary overload during H states.*
Case 1: $W_\alpha = \omega(1)$, $\mathbf{Pr}\left[\overline{W} = 0\right] = 0$*:*

$$\mathbf{E}[B_L(W_\alpha)] = \frac{\mathbf{E}[W_\alpha]}{1-\rho} + o(W_\alpha)$$
$$\mathbf{E}[B_H(W_\alpha)] = \frac{\mathbf{E}[W_\alpha] + \frac{1-\rho-r_H}{\alpha_H}}{1-\rho} + o(\max\left\{W_\alpha, \alpha^{-1}\right\})$$

Case 2: *$W_\alpha = \Theta(1)$:*

$$\mathbf{E}[B_L(W_\alpha)] = \frac{\mathbf{E}\left[\overline{W}\right]}{r_L} + p_{switch} \cdot (1 - Q_f)\frac{1-\rho-r_H}{\alpha_H(1-\rho)} + o(1)$$
$$\mathbf{E}[B_H(W_\alpha)] = (1 - P_f) \cdot \frac{\mathbf{E}\left[\overline{W}\right] + \frac{1-\rho-r_H}{\alpha_H}}{1-\rho} + O(1)$$

where, p_{switch} denotes the probability that the environment state switches to H before the busy period started by $\overline{W}$ in state L ends. We call this event a 'switch'. The expression for p_{switch} is given by $p_{switch} = \frac{\mathbf{E}\left[\overline{W}\right]\alpha_L}{r_L} + o(\alpha)$. The quantity Q_f denotes the probability that, given a 'switch' occurs, the residual busy period is finite if the H state were to last indefinitely from then on:

$$Q_f = \widetilde{V}(\lambda(1-p_f)) + o(1)$$

where $\widetilde{V}(\cdot)$ is given by [3]: $\widetilde{V}(s) = \frac{r_L \cdot \frac{1-\widetilde{\overline{W}}(s)}{\mathbf{E}[\overline{W}]}}{s-\lambda(1-\widetilde{S_L}(s))}$, and $p_f \in (0,1)$ solves the fixed point equation[4]: $p_f = \widetilde{S_H}(\lambda(1-p_f))$.

The quantity P_f denotes the probability that the busy period started by $\overline{W}$ during an H state is finite if the H state were to last indefinitely and is given by $P_f = \widetilde{\overline{W}}(\lambda(1-p_f))$.

Corollary 3 *Consider the case $S_L \sim \mathrm{Exp}\,(\mu_L)$ and $S_H \sim \mathrm{Exp}\,(\mu_H)$, $\mu_L > \lambda > \mu_H$. Let B_s^c ($c, s \in \{L, H\}$) denote the busy period duration started by a class c job in environment state s. Then,*

$$\mathbf{E}\left[B_L^L\right] = \frac{1}{\mu_L - \lambda}\left(1 + \frac{1-p}{p} \cdot \frac{\lambda - \mu_H}{\mu_L - \mu_H} \cdot \frac{\frac{\lambda}{\mu_H} - \rho}{1-\rho}\right) + o(1)$$
$$\mathbf{E}\left[B_L^H\right] = \frac{\mu_L}{\mu_H(\mu_L - \lambda)}\left(1 + \frac{1-p}{p}(1 - Q_{f_H})\frac{\frac{\lambda}{\mu_H} - \rho}{1-\rho}\right) + o(1)$$

and:

$$\mathbf{E}\left[B_H^H\right] = \left(1 - \frac{\mu_H}{\lambda}\right) \cdot \frac{\frac{\lambda}{\mu_H} - \rho}{\alpha_H(1-\rho)} + o(\alpha^{-1})$$
$$\mathbf{E}\left[B_H^L\right] = \left(1 - \frac{\mu_L}{\mu_L + \lambda - \mu_H}\right) \cdot \frac{\frac{\lambda}{\mu_H} - \rho}{\alpha_H(1-\rho)} + o(\alpha^{-1}).$$

[3] The function $\widetilde{V}(s)$ denotes the Laplace transform of the workload in the system just before the 'switch' event occurs. $\widetilde{V}(s)$ is obtained as the Laplace transform of the stationary workload conditioned on server being busy in an $M/G/1$ with repeated vacations, with service distribution S_L and *i.i.d.* vacations distributed as $\overline{W}$

[4]The quantity p_f denotes the probability that a busy period started by an H job in an H state is finite if the H state were to last indefinitely.

$$\textit{In the above,}\quad 1 - Q_{f_H} = 1 - \widetilde{V_H}(\lambda(1-\phi_f))$$

$$\textit{where}\quad \phi_f = \frac{\mu_H}{\lambda}\ ;\ \widetilde{V_H}(s) = \frac{\left(1-\frac{\lambda}{\mu_L}\right)\left(\frac{\mu_H}{\mu_H+s}\right)}{1-\frac{\lambda}{\mu_L}\left(\frac{\mu_L}{\mu_L+s}\right)}.$$

Theorem 12 *Let $r_H > 0$. That is, the system is stable during H states.*
Case 1: $W_\alpha = \omega(\alpha^{-1})$

$$\mathbf{E}[B_L(W_\alpha)] = \frac{\mathbf{E}[W_\alpha]}{1-\rho} + o(W_\alpha)$$
$$\mathbf{E}[B_H(W_\alpha)] = \frac{\mathbf{E}[W_\alpha]}{1-\rho} + o(W_\alpha)$$

Case 2: $W_\alpha = \Theta(\alpha^{-1})$

$$\mathbf{E}[B_L(W_\alpha)] = \frac{\mathbf{E}[W_\alpha]}{1-\rho}(1-u_\alpha) + \frac{\mathbf{E}[W_\alpha]}{r_L}u_\alpha + o(\alpha^{-1})$$
$$\mathbf{E}[B_H(W_\alpha)] = \frac{\mathbf{E}[W_\alpha]}{1-\rho}(1-u_\alpha) + \frac{\mathbf{E}[W_\alpha]}{r_H}u_\alpha + o(\alpha^{-1})$$
$$\textit{where}\quad u_\alpha \equiv \left[\frac{1-\widetilde{W_\alpha}\left(\frac{\alpha_L}{r_L}+\frac{\alpha_H}{r_H}\right)}{\mathbf{E}[W_\alpha]\left(\frac{\alpha_L}{r_L}+\frac{\alpha_H}{r_H}\right)}\right],\ 0 < u_\alpha < 1$$
$$\textit{and}\quad \lim_{\alpha\to 0} u_\alpha = u = \left[\frac{1-\widetilde{\overline{W}}\left(\frac{1-p}{r_L}+\frac{p}{r_H}\right)}{\mathbf{E}\left[\overline{W}\right]\left(\frac{1-p}{r_L}+\frac{p}{r_H}\right)}\right]$$

and recall $\overline{W} = \lim_{\alpha\to 0}\alpha W_\alpha$.
Case 3: $W_\alpha = o(\alpha^{-1})$

$$\mathbf{E}[B_L(W_\alpha)] = \frac{\mathbf{E}[W_\alpha]}{r_L} + o(W_\alpha)$$
$$\mathbf{E}[B_H(W_\alpha)] = \frac{\mathbf{E}[W_\alpha]}{r_H} + o(W_\alpha)$$

Proof Sketch: Recall our fluid model, in which the workload decreases at deterministic rates r_L during the L states, and increases at rate $-r_H$ during the H states. We would like to believe that given an initial workload W_α, asymptotically the busy period started by it is the same as the duration of the busy period started by W_α under the fluid model. However, this is only partially true. When $W_\alpha = \Theta(\alpha^{-1})$, this asymptotic equivalence is justified by [3, Theorem 1(b)] which proves the convergence of sample paths of the stochastic and fluid systems (although one needs to do a bit more work to convert it to convergence of busy periods). For the remaining cases, we must consider the tree of events that may occur until each leaf corresponds to an empty system, or one with workload that is $\Theta(\alpha^{-1})$ so that we can apply [3, Theorem 1(b)]. We describe this below.

Case: $W_\alpha = \omega(\alpha^{-1})$: In this case, the initial workload is of a higher order than the scale at which the system switches. Thus, asymptotically, the number of times the system switches states before W_α drains goes to ∞ as $\alpha \to 0$, and the workload sees the "average system" during its sojourn. Thus the mean busy period is $\frac{\mathbf{E}[W_\alpha]}{1-\rho} + o(W_\alpha)$.

Case: $W_\alpha = \Theta(\alpha^{-1})$: As noted above, in this case from [3, Theorem 1(b)] asymptotically the mean busy period is given by the busy period under the fluid model. The expressions in the theorems are obtained by setting up and solving recurrences for the mean busy period under the fluid model.

Remark 14: When $r_H > 0$, the mean busy period started in state s is a convex combination of the busy period if the state s were to last indefinitely, and the busy period of the "average system", with the coefficient being a function of the Laplace transform of the workload. This contrasts sharply with the case $r_H < 0$.

Case: $W_\alpha = o(\alpha^{-1})$, $r_H > 0$: In this case, the system is stable in both states. Consider a busy period starting in state L. If the L state were to last forever, the busy period would exactly be $\frac{\mathbf{E}[W_\alpha]}{r_L}$. However, since we may switch at rate $\Theta(\alpha)$, there is a $o(1)$ probability that the system switches to state H before the busy period finishes. If this switch were to happen, the remaining busy period would be stochastically bounded by a $\Theta(W_\alpha)$ random variable, as the system is always stable, thus giving a $o(W_\alpha)$ contribution to the overall busy period after multiplying by the probability of switching. Thus asymptotically, the mean busy period started by W_α workload in state L would be $\frac{\mathbf{E}[W_\alpha]}{r_L} + o(W_\alpha)$.

Case: $W_\alpha = \Theta(1)$, $r_H < 0$: This case is the most non-trivial of all, and clearly explains the failure of fluid modeling of busy periods. First, consider a busy period started in state H by $W_\alpha = \Theta(1)$ work. The fluid model would imply that the workload keeps increasing at rate $-r_H$ until the system switches to L. At this point we have $\Theta(\alpha^{-1})$ workload built up, and we could apply [3, Theorem 1(b)]. *However, given that we start with $\Theta(1)$ workload in state H (which is in transient overload), there is still a constant probability that the stochastic busy period started by the $\Theta(1)$ workload is finite*! This probability is given by P_f in the statement of Theorem 11, and given that this event does not happen, we can use the fluid busy period expressions, which is how we arrive at Theorem 11. Next, consider a busy period started in state L by $W_\alpha = \Theta(1)$ work. In this case, with $\Theta(\alpha)$ probability (given by p_{switch}), there is a class H arrival before the busy period ends. We are now in state H with $\Theta(1)$ workload (whose transform is given by $\widetilde{L}(s) \cdot \widetilde{S_H}(s)$). Given that a class H arrival happens, the residual busy period (from our argument above) is $\Theta(\alpha^{-1})$. After multiplying it with p_{switch}, we see that the contribution of this term to the overall busy period is $\Theta(1)$, and hence is of the same asymptotic order as the duration of the busy period started in state L conditioned on it ending in state L ($= \frac{\mathbf{E}[W_\alpha]}{r_L} + o(1)$). Therefore, we need to be precise with each of the terms involved, and applying the fluid method does not yield the correct expressions.

Proof of Theorem 11:

Case 1: $W_\alpha = \omega(1)$, $\mathbf{Pr}\left[\overline{W} = 0\right] = 0$: We first show that under the fluid regime, the expressions for the busy periods are as given in the theorem. Then we will argue that when $W_\alpha = \omega(1)$, the fluid approximation for the busy period is within a small additive term of the stochastic busy period.

Let W_α be deterministic x. Then we can write the following recurrence relation for the fluid busy period started in L or H state by workload x.

$$\begin{aligned}
\mathbf{E}[B_H(x)] &= \frac{1}{\alpha_H} + \mathbf{E}\Big[B_L\left(x - \frac{r_H}{\alpha_H}\right)\Big] \\
\mathbf{E}[B_L(x)] &= \mathbf{E}\Big[\min\left\{\frac{x}{r_L}, T_L\right\}\Big] + \mathbf{E}\Big[B_H\left(x - r_L \min\left\{\frac{x}{r_L}, T_L\right\}\right) \cdot \mathbf{1}_{\{x > r_L \cdot T_L\}}\Big]
\end{aligned}$$

where $T_L \sim \text{Exp}\,(\alpha_L)$.

Now we assume $\mathbf{E}[B_L(x)] = b_L x$ and $\mathbf{E}[B_H(x)] = a_H + b_H x$ for some constants b_L, a_H, b_H, and then verify that these forms are indeed correct by identifying the unknown constants. Under the assumed forms for fluid busy periods, the recurrences reduce to:

$$\begin{aligned}
a_H + b_H x &= \frac{1}{\alpha_H} + b_L x - b_L \frac{r_H}{\alpha_H} \\
b_L x &= \frac{1 - e^{-\frac{\alpha_L}{r_L}x}}{\alpha_L} + a_H(1 - e^{-\frac{\alpha_L}{r_L}x}) + b_H x - b_H r_L \frac{1 - e^{-\frac{\alpha_L}{r_L}x}}{\alpha_L}
\end{aligned}$$

Since the above equations should be satisfied for all x, we get:

$$\begin{aligned}
b_L &= b_H \\
a_H &= \frac{1}{\alpha_H}\left(1 - b_L r_H\right) \\
a_H &= \frac{1}{\alpha_L}\left(b_H r_L - 1\right)
\end{aligned}$$

which gives:

$$b_L = b_H = \frac{1}{1-\rho} \quad ; \quad a_H = \frac{1 - \rho - r_H}{\alpha_H(1-\rho)}$$

yielding the expressions in the theorem statement.

Now we verify that when $W_\alpha = \omega(1)$, the fluid busy period expressions are asymptotically correct. In the simple case $W_\alpha = \omega(\alpha^{-1})$, the system switches on a faster time-scale ($\Theta(\alpha^{-1})$) than the initial amount of work ($\omega(\alpha^{-1})$). Thus this workload sees the "average" system (rather than the transient system) and its busy period is simply $\frac{\mathbf{E}[W_\alpha]}{1-\rho} + o(W_\alpha)$.

When the workload is $\Theta(\alpha^{-1})$, then using [3], the sample paths of the stochastic system (scaled by α) converge as $\alpha \to 0$ to the fluid sample path in the space $D[0, \infty)$. Thus, the mean busy period of the stochastic system is within $o(\alpha^{-1})$ of the mean busy period of the fluid system, asymptotically.

Now consider the case $W_\alpha = \Theta(g(\alpha))$ where $g(\alpha) = \omega(1)$, but $g(\alpha) = o(\alpha^{-1})$ (for example $g(\alpha) = \frac{1}{\sqrt{\alpha}}$).
Subcase 1: Busy period beginning in state H: We will show that even though the initial workload is $o(\alpha^{-1})$, since it is $\omega(1)$, with overwhelming probability, the sample paths will follow the fluid trajectory. Let the Laplace transform of W_α be given by $\widetilde{W_\alpha}(s) = \mathbf{E}\left[e^{-sW_\alpha}\right]$. Since reordering the jobs served in a busy period does not change the busy period duration, consider the case where the initial workload W_α is served first. If the H state were to last forever, the $z-$transform for the number of arrivals of class H jobs while workload W_α is served is given by $\widetilde{W_\alpha}(\lambda(1-z))$. The main idea is to show that since the H state is in overload, with probability tending to 1, at least one of the class H job will start a busy period that lasts until the end of the H state, whereby by the Strong Law of Large Numbers the accumulated workload will be $\Theta(\alpha^{-1})$. Consider the busy period that one class H job starts, provided the H state continues forever. The Laplace transform of the busy period in an $M/G/1$ with only class H jobs, denoted by $\widetilde{B_H}(s)$, satisfies:

$$\widetilde{B_H}(s) = \widetilde{S_H}(s + \lambda(1 - \widetilde{B_H}(s))).$$

Since the $M/G/1$ is in overload, there is a constant probability that the busy period is infinite. The probability that the busy period is finite is obtained as

$$p_f = \lim_{s \to 0} \widetilde{B_H}(s).$$

Taking limit in the expression for $\widetilde{B_H}(s)$, we obtain:

$$p_f = \widetilde{S_H}(\lambda(1 - p_f))$$

The busy period started by W_α, given the H phase lasts forever, is finite if and only if the busy period started by each H arrival while W_α was served is finite. This probability, then is given by

$$\begin{aligned}\mathbf{Pr}[\text{busy period started by } W_\alpha \text{ during H is finite}] &= \sum_{i=0}^{\infty} \mathbf{Pr}[i \text{ arrivals during } W_\alpha] \cdot p_f^i \\ &= \widetilde{W_\alpha}(\lambda(1 - p_f)) \to 0\end{aligned}$$

The last fact is true since $\frac{W_\alpha}{g(\alpha)} \to \overline{W}$, $\widetilde{W_\alpha}(s) \to \widetilde{\overline{W}}(s \cdot g(\alpha)) \to 0$ as $\alpha \to 0$ ($\widetilde{\overline{W}}(s)$ is a decreasing function from 1 to 0, and $g(\alpha) = \omega(1)$). The fact that $\lim_{s\to\infty} \widetilde{\overline{W}}(s) = 0$ follows from the assumption $\mathbf{Pr}\left[\overline{W} = 0\right] = 0$.

Therefore, with probability approaching 1, the busy period started by W_α in phase H (under the assumption that the H phase lasts forever) is not finite. In other words, during the H phase, the workload increases asymptotically along the fluid trajectory, and then the system switches to the L phase. Since the work built up during the H state is $\Theta(\alpha^{-1})$, the workload follows the fluid trajectory after switching to the L state. Therefore, the expression for the mean busy period started in H phase by $\omega(1)$ work is indeed given by the mean busy period under the fluid regime within a $o(\max\{W_\alpha, \alpha^{-1}\})$ term.

Subcase 2: Busy period beginning in state L: Now we consider the case where the busy period starts in the L phase. If the L phase were to last forever, the workload in the system, scaled by $g(\alpha)$, would follow the fluid trajectory, and hence the mean busy period would be the mean busy period under the fluid regime within a $o(W_\alpha)$ term. However, with probability $\Theta(\alpha \cdot g(\alpha))$ the system switches to H state before the fluid workload reaches 0. Conditioned on switching to the H state before the period ends, the workload at the beginning of the H state is again $\Theta(g(\alpha))$. We have already argued above that subsequently the workload follows the fluid trajectory – and the residual busy period will be $\Theta(\alpha^{-1})$ within an $o(\alpha^{-1})$ term. Therefore, the mean busy period started in L phase will be the mean busy period under the fluid regime, within a $o(W_\alpha)$ term.

Case 2: $W_\alpha = \Theta(1)$: We first consider the case where the busy period begins in the H state by workload W with Laplace transform $\widetilde{W}(s)$. As we have argued above, since the H state is in overload, there is a constant probability that the busy period does not end before the system switches to the L state. This probability is given by $1 - P_f$ where,

$$P_f = \widetilde{W}(\lambda(1 - p_f))$$

and p_f is the solution to the fixed point equation

$$p_f = \widetilde{S_H}(\lambda(1 - p_f)).$$

P_f denotes the probability that a busy period started by work W in the $M/G/1$ under overload is finite, and p_f is the probability that a busy period started by a single class H job is finite.

Given that the busy period does not end before the system switches, the amount of workload that builds up in the system is given by $T_H(\frac{\lambda}{\mu_H} - 1) + o(\alpha^{-1})$ where T_H denotes the duration of the H state and is $\Theta(\alpha^{-1})$. We can thus apply the previous case and conclude that the mean busy period in this case, that is with probability $1 - P_f$, is given by $\frac{1}{\alpha_H} - \frac{r_H}{\alpha_H(1-\rho)}$. In simpler terms, we are starting the busy period with $\Theta(1)$ work in the H state. With $\Theta(1)$ probability, the busy period does not end in the H phase, in which case, we start the subsequent L state with $\Theta(\alpha^{-1})$ work, with an overall contribution to the mean busy period of $\Theta(\alpha)$. If however, the original busy period ends in the H state itself, then this event contributes a $\Theta(1)$ term and hence is asymptotically negligible compared to the contribution of the event where the busy period does not end in the H state.

Now we consider the case where the busy period begins in an L state. Again, we have two cases – either the busy period ends in the L state itself, or the system switches to an H state before the busy period ends. If the busy period ends in the L state, an event which happens with probability

$1-\Theta(\alpha)$, then the mean busy period conditioned on this event is given by $\frac{\mathbf{E}[W]}{r_L}$. However, the system can switch with probability $\Theta(\alpha)$, and the contribution of the residual busy period conditioned on this event can be $\Theta(\alpha^{-1})$ (from the previous subcase). Therefore, this event also contributes a $\Theta(1)$ term to the mean busy period, and we handle this event next.

Consider an $M/G/1$ busy period started by work W. We let this $M/G/1$ evolve in the L state, and consider an independent Poisson(α_L) marking process. Our aim is to find the workload in the $M/G/1$ when the first mark arrives during the busy period. The probability that no mark arrives is given by $1-\frac{\mathbf{E}[W]\alpha_L}{r_L}$, which we denote by $1-p_{switch}$ in the theorem statement. Thus, with probability p_{switch}, at least one mark arrives, or equivalently, the environment processes switches before the busy period ends and hence the busy period now evolves in the H state.

The subsequent busy period (that which evolves after the system switches to H) is given by the the busy period that starts in H state with workload $\widetilde{V}(s)$, where $\widetilde{V}(s)$ denotes the transform of the workload that is seen by the Poisson(α_L) marking process *conditioned on being the first mark of a busy period.* We will now argue that this is asymptotically given by the stationary workload in an $M/G/1$ conditioned on the server being busy, with exceptional service distribution for the job that starts the busy period given by W, and service distribution S_L. We first note that if we have such an $M/G/1$ where we consider the distribution of workload seen by all marks, then this is indeed the stationary workload conditioned on server being busy, and hence given by stationary delay in the $M/G/1$ system with special first service seen by arrivals finding the server busy (this expression, $\widetilde{V}(s) = \frac{r_L \cdot \frac{1-\widetilde{\overline{W}}(s)}{\mathbf{E}\left[\overline{W}\right]}}{s-\lambda(1-\widetilde{S_L}(s))}$, is given in the theorem statement and derived in Appendix B). However, we are interested in the workload that the first mark sees in a busy period, call this W_1. We will argue that as the probability of marking goes to 0, the workload seen by the first mark converges in distribution to the stationary workload conditioned on the server being busy (and that as the probability of marking goes to 0, this sequence of random variables remains uniformly integrable so that the Laplace transforms converge). We first note that the workload seen by the first mark is stochastically bounded above by the supremum of the workload in a busy period started by work W, denote this by W_1^*. Further, conditioned on a second marked arrival, we can upper bound the work that this mark sees by the supremum of the workload in the busy period started by W_1^* (which is an upper bound on the workload after the arrival of the first mark), denote this by W_2^*. Similarly, we can obtain an upper bound on the workload seen by the nth marked arrival in a busy period. We also have the trivial lower bound of 0 on the workload seen by the nth marked arrival in a busy period. Note that both these upper and lower bounds are independent of the marking probability. Let p_i denote the probability that there are i marked arrivals in a busy period. We can thus sandwich the stationary workload of the $M/G/1$ conditioned on it being busy by $\frac{p_1 \cdot W_1}{\sum_{i=1}^{\infty} p_i}$ and $\frac{p_1 \cdot W_1 + \sum_{i=2}^{\infty} p_2 \cdot W_i^*}{\sum_{i=1}^{\infty} p_i}$. However, as the marking probability ($\Theta(\alpha)$) goes to 0, $p_i \sim \Theta(\alpha^i)$. Therefore, W_1 converges to the stationary workload in the $M/G/1$ with special service, conditioned on server being busy. ■

Proof of Theorem 12:

Recall that in the workload is decreasing during both the L and H states. There is a negative drift of $r_L = 1 - \frac{\lambda}{\mu_L}$ during the L phase and a negative drift of $r_H = 1 - \frac{\lambda}{\mu_H}$ during the H phase.

Case 1: $W_\alpha = \omega(\alpha^{-1})$: As in the proof of Theorem 11, since the system switches at a faster time scale ($\Theta(\alpha^{-1})$) than the initial workload ($\omega(\alpha^{-1})$), the workload during its sojourn sees an average system, and hence the busy period is $\frac{\mathbf{E}[W_\alpha]}{1-\rho} + o(W_\alpha)$.

Case 2: $W = \Theta(\alpha^{-1})$: We begin by noting that since the initial workload is $\Theta(\alpha^{-1})$, the workload trajectory of the stochastic system, scaled by α, converges to the fluid trajectory. Hence the busy period of the stochastic system is given by the fluid busy period and an additional $o(\alpha^{-1})$ term.

We now set up the recurrences for busy periods started by deterministic work x during the H and L phases under the fluid regime:

$$\mathbf{E}[B_H(x)] = \mathbf{E}\Big[\min\Big\{\frac{x}{r_H}, T_H\Big\}\Big] + \mathbf{E}\Big[B_L\left(x - r_H \min\Big\{\frac{x}{r_H}, T_H\Big\}\right) \cdot \mathbf{1}_{\{x > r_H \cdot T_H\}}\Big]$$
$$\mathbf{E}[B_L(x)] = \mathbf{E}\Big[\min\Big\{\frac{x}{r_L}, T_L\Big\}\Big] + \mathbf{E}\Big[B_H\left(x - r_L \min\Big\{\frac{x}{r_L}, T_L\Big\}\right) \cdot \mathbf{1}_{\{x > r_L \cdot T_L\}}\Big]$$

where T_H is an $\mathrm{Exp}\,(\alpha_H)$ random variable and T_L is an $\mathrm{Exp}\,(\alpha_L)$ random variable.

We now guess and verify that $\mathbf{E}[B_H(x)]$ and $\mathbf{E}[B_L(x)]$ have the following function form:

$$B_i(x) = a_i + b_i x + c_i e^{-\left(\frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}\right)x}$$

where a_i, b_i and c_i, $i \in \{L, H\}$, are constants to be determined.[5]

[5] The above 'guess' is in fact an educated attempt arrived at by looking at an alternate discrete system but which is similar on fluid scale to the system we want to analyze. This is achieved by looking at a system where arrival rate is 0, but server capacity (rate) switches between r_L and r_H. If we then denote the expected length of the busy period started by n jobs in state L as $B_L(n)$ (similarly $B_H(n)$ denotes the expected length of the busy period started by n jobs in state H), and define

$$\widehat{B_L}(z) = \sum_{z=0}^{\infty} z^n B_L(n); \quad \widehat{B_H}(z) = \sum_{z=0}^{\infty} z^n B_H(n),$$

we can set up the following recurrences,

$$\widehat{B_L}(z)((\alpha_L + r_L) - r_L z) = \frac{z}{1-z} + \alpha_L \widehat{B_H}(z)$$
$$\widehat{B_H}(z)((\alpha_H + r_H) - r_H z) = \frac{z}{1-z} + \alpha_H \widehat{B_L}(z)$$

which solve for,

$$\widehat{B_L}(z) r_L r_H (1-z)\left[\left(1 + \frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}\right) - z\right] = \frac{z}{1-z}(\alpha + r_H(1-z))$$

The above transform gives the following form for $B_L(n)$: $B_L(n) = a_L + b_L n + c_L r^n$ for some constants a_L, b_L, c_L and

Since $B_i(0) = 0$, we have $a_i = -c_i$. Since the Laplace transform for $x - r_i \min\left\{\frac{x}{r_i}, T_i\right\}$ is

$$\mathbf{E}\left[e^{-s\left(x - \min\left\{\frac{x}{r_i}, T_i\right\}\right)}\right] = \frac{se^{-\frac{\alpha_i}{r_i}x} - \frac{\alpha_i}{r_i}e^{-sx}}{s - \frac{\alpha_i}{r_i}}$$

and $\mathbf{E}\left[\min\left\{\frac{x}{r_i}, T_i\right\}\right] = \frac{1-e^{-\frac{\alpha_i}{r_i}x}}{\alpha_i}$, our recurrences become:

$$\begin{aligned}
a_L + b_L x + c_L e^{-\left(\frac{\alpha_L}{r_L}+\frac{\alpha_H}{r_H}\right)x} &= \frac{1 - e^{-\frac{\alpha_L}{r_L}x}}{\alpha_L} + a_H + b_H\left(x - \frac{1 - e^{-\frac{\alpha_L}{r_L}x}}{\frac{\alpha_L}{r_L}}\right) \\
&\quad + c_H\left(\frac{\left(\frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}\right)e^{-\frac{\alpha_L}{r_L}x} - \frac{\alpha_L}{r_L}e^{-\left(\frac{\alpha_L}{r_L}+\frac{\alpha_H}{r_H}\right)x}}{\frac{\alpha_H}{r_H}}\right) \\
a_H + b_H x + c_H e^{-\left(\frac{\alpha_L}{r_L}+\frac{\alpha_H}{r_H}\right)x} &= \frac{1 - e^{-\frac{\alpha_H}{r_H}x}}{\alpha_H} + a_L + b_L\left(x - \frac{1 - e^{-\frac{\alpha_H}{r_H}x}}{\frac{\alpha_H}{r_H}}\right) \\
&\quad + c_L\left(\frac{\left(\frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}\right)e^{-\frac{\alpha_H}{r_H}x} - \frac{\alpha_H}{r_H}e^{-\left(\frac{\alpha_L}{r_L}+\frac{\alpha_H}{r_H}\right)x}}{\frac{\alpha_L}{r_L}}\right)
\end{aligned}$$

Since the above equalities hold for all x, we have the following relations:

$$\begin{aligned}
b_L &= b_H \\
a_L &= \frac{1}{\alpha_L} + a_H - \frac{b_H}{\frac{\alpha_L}{r_L}} \\
a_H &= \frac{1}{\alpha_H} + a_L - \frac{b_L}{\frac{\alpha_H}{r_H}} \\
0 &= -\frac{1}{\alpha_L} + \frac{b_H}{\frac{\alpha_L}{r_L}} + c_H\frac{\frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}}{\frac{\alpha_H}{r_H}} \\
0 &= -\frac{1}{\alpha_H} + \frac{b_L}{\frac{\alpha_H}{r_H}} + c_L\frac{\frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}}{\frac{\alpha_L}{r_L}} \\
0 &= c_L\frac{\alpha_H}{r_H} + c_H\frac{\alpha_L}{r_L}
\end{aligned}$$

$r = \left(1 + \frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}\right)^{-1}$.

which together with $a_i = -c_i$ finally yield:

$$
\begin{aligned}
b_L = b_H &= \left(\frac{\frac{r_L}{\alpha_L} + \frac{r_H}{\alpha_H}}{\frac{1}{\alpha_L} + \frac{1}{\alpha_H}} \right)^{-1} = \frac{1}{1-\rho} \\
c_L \frac{\alpha_H}{r_H} &= \frac{\frac{1}{r_H} - \frac{1}{r_L}}{\left(\frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}\right)\left(\frac{r_L}{\alpha_L} + \frac{r_H}{\alpha_H}\right)} \\
c_H \frac{\alpha_L}{r_L} &= \frac{\frac{1}{r_L} - \frac{1}{r_H}}{\left(\frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}\right)\left(\frac{r_L}{\alpha_L} + \frac{r_H}{\alpha_H}\right)}
\end{aligned}
$$

equivalently

$$
\begin{aligned}
c_L &= \frac{r_L - r_H}{\alpha_L \alpha_H \left(\frac{r_L}{\alpha_L} + \frac{r_H}{\alpha_H}\right)^2} \cdot \frac{r_H}{\alpha_H} \\
c_H &= -\frac{r_L - r_H}{\alpha_L \alpha_H \left(\frac{r_L}{\alpha_L} + \frac{r_H}{\alpha_H}\right)^2} \cdot \frac{r_L}{\alpha_L}
\end{aligned}
$$

and,

$$
\begin{aligned}
a_L &= -c_L \\
a_H &= -c_H
\end{aligned}
$$

Therefore the expected busy period started by a workload of size x during L and H phases, respectively, are given by

$$
\mathbf{E}[B_L(x)] = \frac{x}{1-\rho} - \frac{r_L - r_H}{\alpha_L \alpha_H \left(\frac{r_L}{\alpha_L} + \frac{r_H}{\alpha_H}\right)^2} \cdot \frac{r_H}{\alpha_H} \cdot \left(1 - e^{-\left(\frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}\right)x}\right) \tag{12}
$$

$$
\mathbf{E}[B_H(x)] = \frac{x}{1-\rho} + \frac{r_L - r_H}{\alpha_L \alpha_H \left(\frac{r_L}{\alpha_L} + \frac{r_H}{\alpha_H}\right)^2} \cdot \frac{r_L}{\alpha_L} \cdot \left(1 - e^{-\left(\frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}\right)x}\right) \tag{13}
$$

We can express $\mathbf{E}[B_L(x)]$ and $\mathbf{E}[B_H(x)]$ in the following more convenient/intuitive form:

$$\mathbf{E}[B_L(x)] = \frac{x}{1-\rho} - \left(\frac{x}{1-\rho} - \frac{x}{r_L}\right) \cdot \left[\frac{1 - e^{-\left(\frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}\right)x}}{\left(\frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}\right)x}\right] \tag{14}$$

$$\mathbf{E}[B_H(x)] = \frac{x}{1-\rho} - \left(\frac{x}{1-\rho} - \frac{x}{r_H}\right) \cdot \left[\frac{1 - e^{-\left(\frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}\right)x}}{\left(\frac{\alpha_L}{r_L} + \frac{\alpha_H}{r_H}\right)x}\right] \tag{15}$$

which show that $\mathbf{E}[B_L(x)]$ and $\mathbf{E}[B_H(x)]$ are a weighted averages of the busy periods of the $\alpha = 0$ and $\alpha \to \infty$ cases.

Now taking the expectation over x (which is distributed as W_α), we obtain the expressions given in the theorem.

Case 3: $W_\alpha = o(\alpha^{-1})$: Since the system is stable during both the L and H states, the busy period is $\Theta(W_\alpha)$ (being upper bounded by the busy period started by W_α in an $M/G/1$ with service distribution S_H). Suppose the busy period starts in the L state. If the L state were to last forever, the busy period would indeed be $\frac{\mathbf{E}[W_\alpha]}{r_L}$. Now either the system switches to the H state before this busy period ends, and this event happens with probability $1 - o(1)$. In this case, the length of the busy period conditioned on it being smaller than $\text{Exp}(\alpha_L)$ will be $\frac{\mathbf{E}[W_\alpha]}{r_L} + o(W_\alpha)$ since $W_\alpha = o(\alpha^{-1})$. However, if the system switches before the busy period ends, which happens with probability $o(1)$, the residual busy period is still $\Theta(W_\alpha)$. The overall contribution of the second event to the mean busy period started by W_α is $o(W_\alpha)$. By law of total probability, the mean busy started in L phase is $\frac{\mathbf{E}[W_\alpha]}{r_L} + o(W_\alpha)$.

The proof for busy periods started during H phases is identical. ■

B Analysis of $M/G/1$ busy period with special first service

We consider a busy period started by work W in an $M/G/1$ with arrival rate λ, and a general service distribution. The Laplace transform of W is given by $\widetilde{W}(s)$, and of the service distribution is given by $\widetilde{S}(s)$. We want to find the transform of the workload seen by an arbitrary arrival that arrives during the busy period. This workload is given by the stationary workload in an $M/G/1$ with special first service (note that this is not just a set-up time, but the distribution of the job that starts the busy period is W), given the system is busy. The stationary workload is in turn given by the stationary delay in this special $M/G/1$ conditioned on the job finding the server busy.

Replicating the analysis for $M/G/1$, let $\hat{N}(z)$ denote the z-transform of the number of jobs left behind by a departure. Let $\widehat{Q_W}(z) = \tilde{W}(\lambda(1-z))$ be the transform of the number of Poisson(λ) arrivals during W, and $\widehat{Q_S}(z) = \tilde{S}(\lambda(1-z))$. Then,

$$\hat{N}(z) = p_0 \frac{z\widehat{Q_W}(z) - \widehat{Q_S}(z)}{z - \widehat{Q_S}(z)} \tag{16}$$

where,

$$p_0 = \frac{1-\rho}{1+\lambda\mathbf{E}[W]-\rho} \tag{17}$$

is the idle probability.

The Laplace transform of the customer average response time is given by

$$\begin{aligned}
\widetilde{R}(s) &= \hat{N}\left(1 - \frac{s}{\lambda}\right) &(18)\\
&= \frac{1-\rho}{1+\lambda\mathbf{E}[W]-\rho} \cdot \frac{(\lambda - s)\widetilde{W}(s) - \lambda\widetilde{S}(s)}{(\lambda - s) - \lambda\widetilde{S}(s)} &(19)\\
&= p_0\widetilde{W}(s) + p_0\widetilde{S}(s)\frac{1-\widetilde{W}(s)}{s - \lambda(1-\widetilde{S}(s))} &(20)\\
&= p_0\widetilde{W}(s) + (1-p_0)\widetilde{S}(s)\frac{(1-\rho)\frac{1-\widetilde{W}(s)}{\mathbf{E}[W]}}{s-\lambda(1-\widetilde{S}(s))} &(21)
\end{aligned}$$

Thus, the transform of the workload, given that the server is busy is given by

$$\widetilde{C}(s) = \frac{(1-\rho)\frac{1-\widetilde{W}(s)}{\mathbf{E}[W]}}{s-\lambda(1-\widetilde{S}(s))}. \tag{22}$$