On the inapproximability of $M/G/K$: Why two moments of job size distribution are not enough *

Varun Gupta, Mor Harchol-Balter
Computer Science Department
Carnegie Mellon University
{varun,harchol}@cs.cmu.edu

J.G. Dai
School of Industrial and Systems Engineering
Georgia Institute of Technology
dai@gatech.edu

Bert Zwart†
CWI
Amsterdam, The Netherlands
bert.zwart@cwi.nl

Abstract

The $M/G/K$ queueing system is one of the oldest model for multi-server systems, and has been the topic of performance papers for almost half a century. However, even now, only coarse approximations exist for its mean waiting time. All the closed-form (non-numerical) approximations in the literature are based on (at most) the first two moments of the job size distribution. In this paper we prove that no approximation based on only the first two moments can be accurate for all job size distributions, and we provide a lower bound on the inapproximability ratio, which we refer to as "the gap." This is the first such result in the literature to address "the gap." The proof technique behind this result is novel as well and combines mean value analysis, sample path techniques, scheduling, regenerative arguments, and asymptotic estimates. Finally, our work provides insight into the effect of higher moments of the job size distribution on the mean waiting time.

1 Introduction

The $M/G/K$ queueing system is one of the oldest and most classical example of multi-server systems. Such multi-server systems are commonplace in a wide range of applications, ranging from call centers to manufacturing systems to computer systems, because they are cost-effective and their serving capacity can be easily scaled up or down.

An $M/G/K$ system consists of K identical servers and a First-Come-First-Serve (FCFS) queue. The jobs (or customers) arrive according to a Poisson process (the symbol M) with rate λ and their service requirements (job sizes) are assumed to be independent and identically distributed random variables having a general distribution (the symbol G); we use X to denote such a generic random variable. If an arriving job finds a free server, it immediately enters service, otherwise it waits in the FCFS queue. When a server becomes free, it chooses the next job to process from the head of

*Preprint of the article appearing in *Queueing Systems: Theory and Applications*. The original publication is available at www.springerlink.com.

†This work was done when the author was an Associate Professor in the School of Industrial and Systems Engineering, Georgia Institute of Technology.

the FCFS queue. We denote the load of this $M/G/K$ system as $\rho = \frac{\lambda \mathbf{E}[X]}{K}$, and assume $\rho < 1$ so that a steady-state distribution exists. We will focus on the metric of mean waiting time in this work, denoted as $\mathbf{E}\left[W^{M/G/K}\right]$, and defined to be the expected time from the arrival of a customer to the time it enters service. Throughout the paper, we assume $\mathbf{E}[X] = 1$. This is without loss of generality since the arrival rate, the mean job size and the mean waiting time can be scaled appropriately for general values of $\mathbf{E}[X]$.

Even though the $M/G/K$ queue has received a lot of attention in the queueing literature, an exact analysis for even simple metrics like mean waiting time for the case $K \geq 2$ still eludes researchers. To the best of our knowledge, the first approximation for the mean waiting time for an $M/G/K$ queue was given by Lee and Longton [26] nearly half a century ago:

$$\mathbf{E}\left[W^{M/G/K}\right] \approx \left(\frac{C^2+1}{2}\right) \mathbf{E}\left[W^{M/M/K}\right] \tag{1}$$

where $\mathbf{E}\left[W^{M/M/K}\right]$ is the mean waiting time with exponentially distributed job sizes with the same mean, $\mathbf{E}[X]$, as in the $M/G/K$ system, and C^2 is the squared coefficient of variation[1] (SCV) of X. Many other authors have also proposed simple approximations for the mean waiting time, [19, 20, 25, 31, 32, 48], but all these closed-form approximations involve only the first two moments of the job size distribution.

Whitt [47], while referring to (1) as "usually an excellent approximation, even given extra information about the service-time distribution," hints that approximations based on two moments of the job size distribution may be inaccurate when C^2 is large. Similar suggestions have been made by many authors, but there are very limited numerical experiments to support this. While a high C^2 may not be of major concern in many applications like manufacturing or customer contact centers, the invalidity of the approximation (1) *is* a major problem in computer and communication systems. In Table 1, we consider two values of C^2, $C^2 = 19$ and $C^2 = 99$, and parameterize the distrbutions so that they have these C^2 values. Such high values of C^2 are typical for workloads encountered in computer systems, such as the sizes of files transferred over the internet [2], and the CPU requests of UNIX jobs [12] and supercomputing jobs [17]. We consider a range of distributions (Weibull, lognormal, truncated Pareto[2]) used in the literature to model computer systems workloads and compare the mean waiting time obtained via simulations to the mean waiting time predicted by the approximation in (1). As can be seen, there is a huge disagreement between the simulated mean waiting time and the 2-moment approximation (1). Further, the simulated mean waiting times are consistently smaller than the analytical approximation. Also observe that different distributions with the same mean and C^2 result in very different mean waiting times.

In this paper, we investigate the above experimental findings, illuminating how other characteristics of the job size distribution may affect the mean waiting time, $\mathbf{E}\left[W^{M/G/K}\right]$. We do so by choosing a specific class of distributions, the hyper-exponential distributions, which are mixtures of exponential distributions. Use of hyper-exponential distributions allows us the freedom to evaluate the effect of different characteristics of the distribution while preserving the first two (and even higher) moments.

Our foremost goal is to study the range of possible values of $\mathbf{E}\left[W^{M/G/K}\right]$ for general job size distributions with some given first two moments. We refer to this range as "the gap". To define

[1] The squared coefficient of variation of a positive random variable X is defined as $C^2 = var(X)/\left(\mathbf{E}[X]\right)^2$

[2] The cumulative distribution function of a truncated Pareto distribution with support $[x_{min}, x_{max}]$ and parameter α is given by:

$$F(x) = \frac{x_{min}^{-\alpha} - x^{-\alpha}}{x_{min}^{-\alpha} - x_{max}^{-\alpha}} \qquad\qquad x_{min} \leq x \leq x_{max}$$

Therefore, specifying the first two moments and the α parameter uniquely defines a truncated Pareto distribution.

	$C^2 = 19$	$C^2 = 99$
	$\mathbf{E}[W]$	$\mathbf{E}[W]$
2-moment approximation (Eqn. 1)	6.6873	33.4366
Weibull	6.0691±0.0138	25.9896±0.1773
Truncated Pareto ($\alpha = 1.1$)	5.5277±0.0216	24.6049±0.2837
Lognormal	4.9937±0.0249	19.5430±0.4203
Truncated Pareto ($\alpha = 1.3$)	4.8788±0.0249	18.7738±0.3612
Truncated Pareto ($\alpha = 1.5$)	3.9466±0.0321	10.6487±0.5373

Table 1: Simulation results for the 95% confidence intervals of the mean waiting time for an $M/G/K$ with $K = 10$ and $\rho = 0.9$. The first line shows the mean waiting time given by the analytical 2-moment approximation in Equation (1). All job size distributions throughout the paper have $\mathbf{E}[X] = 1$.

the gap, set

$$W_h^{C^2} = \sup\left\{ \mathbf{E}\Big[W^{M/G/K}\Big] \;\Big|\; \mathbf{E}[X] = 1, \mathbf{E}\big[X^2\big] = C^2 + 1 \right\}, \tag{2}$$

and

$$W_l^{C^2} = \inf\left\{ \mathbf{E}\Big[W^{M/G/K}\Big] \;\Big|\; \mathbf{E}[X] = 1, \mathbf{E}\big[X^2\big] = C^2 + 1 \right\}. \tag{3}$$

The gap spans $(W_l^{C^2}, W_h^{C^2})$. As one of the major contributions of this paper, we prove a lower bound on the gap for the case $\rho < \frac{K-1}{K}$ (at least one spare server) in Theorem 1, and for the case $\rho > \frac{K-1}{K}$ (no spare servers) in Theorem 2. We believe that the bounds presented in Theorem 1 for the case $\rho < \frac{K-1}{K}$ are tight, and conjecture tight bounds for the case $\rho > \frac{K-1}{K}$ in Section 7, Conjecture 1.

Theorem 1 *For any finite C^2 and $\rho < \frac{K-1}{K}$,*

$$W_h^{C^2} \geq (C^2 + 1)\mathbf{E}\Big[W^{M/D/K}\Big]$$
$$W_l^{C^2} \leq \mathbf{E}\Big[W^{M/D/K}\Big]$$

and thus,

$$\frac{W_h^{C^2}}{W_l^{C^2}} \geq C^2 + 1$$

where $\mathbf{E}\big[W^{M/D/K}\big]$ is the mean waiting time when the job size distribution is deterministic 1*.*

Theorem 2 *For any finite C^2 and $\rho \geq \frac{K-1}{K}$,*

$$W_h^{C^2} \geq \left(\frac{C^2+1}{2}\right)\mathbf{E}\Big[W^{M/M/K}\Big]$$
$$W_l^{C^2} \leq \mathbf{E}\Big[W^{M/M/K}\Big] + \frac{1}{1-\rho}\left[\rho - \frac{K-1}{K}\right]\frac{C^2-1}{2}$$

and thus,

$$\frac{W_h^{C^2}}{W_l^{C^2}} \geq \frac{\left(\frac{C^2+1}{2}\right)\mathbf{E}\big[W^{M/M/K}\big]}{\mathbf{E}\big[W^{M/M/K}\big] + \frac{1}{1-\rho}\left[\rho - \frac{K-1}{K}\right]\frac{C^2-1}{2}}$$

where $\mathbf{E}\left[W^{M/M/K}\right]$ *is the mean waiting time when the job size distribution is exponential with mean* 1.

Theorem 1 will be proved in Section 4 and follows by combining a result of Daley [7] with some new observations. Theorem 2 is far more intricate to prove, and forms the bulk of the paper (Section 5).

We now make a few important observations on the gap:

- Since we prove a lower bound for $W_h^{C^2}$ and an upper bound for $W_l^{C^2}$, Theorems 1 and 2 give a lower bound on the span of the gap for general distributions.
- The gap can be quite large if the C^2 of the job size distribution is high. In particular, when $\rho < \frac{K-1}{K}$, Theorem 1 proves that the maximum possible mean waiting time is at least $\left(C^2+1\right)$ times the minimum possible mean waiting time.
- The lower bound on $W_h^{C^2}$ in Theorem 2 is the same as the 2-moment approximation in (1). (The lower bound on $W_h^{C^2}$ in Theorem 1 is very close but slightly higher than the 2-moment approximation.)
- Theorems 1 and 2 prove that *any* approximation based only on the first two moments will be inaccurate for some distribution because the span of possible values of mean waiting time is large.

Another interesting point is that the lower bound on the gap depends on the load, ρ. The case $\rho \geq \frac{K-1}{K}$ is commonly known in the queueing literature as *0-spare servers* and the case $\rho < \frac{K-1}{K}$ is known as *at least 1 spare server*. The presence of spare servers is known to play a crucial role in determining whether the mean waiting time is infinite given that the second moment of the job size distribution is infinite (see [35] and references therein) and on the tail of the waiting time distribution (see [13]). Observe that in our results too, the number of spare servers (zero or at least one) affects whether C^2 shows up in the lower bound of the gap. When there is even just one spare server, the lower bound is independent of C^2, which suggests that having even one spare server might potentially reduce most of the effect of C^2 on the mean waiting time.

Proving Theorem 1 $\left(\rho < \frac{K-1}{K}\right)$ essentially involves looking at two extreme two-point job size distributions and finding the mean waiting time under those extremal job size distributions. To prove Theorem 2 $\left(\rho \geq \frac{K-1}{K}\right)$, we look at two extreme distributions in the class of 2-phase hyperexponential distributions and obtain the mean waiting time under those job size distributions. We believe that it is not hard to tighten the bound in Theorem 2 by extending our proof technique to work with two-point distributions, and proving a wider "gap" than we do in this paper. However, presently, we focus on 2-phase hyperexponential distributions for ease of exposition and to elucidate the basic steps in obtaining the bound. Clearly the span described by Theorem 1 is non-empty for all $C^2 > 0$. The span described by Theorem 2 is non-empty only when $C^2 > 1$ even though the theorem is true for all values of C^2. In fact, Proposition 1 shows that our lower bound for the span of the gap is strictly positive when $K \geq 2$ and $C^2 > 1$:

Proposition 1 *Let* $\mathbf{E}\left[W^{M/M/K}\right]$ *be the mean waiting time in an* $M/M/K$ *with mean job size* 1. *For all values of* $K \geq 2$, $\rho \in \left[\frac{K-1}{K}, 1\right)$ *and* $C^2 > 1$,

$$\left(\frac{C^2+1}{2}\right)\mathbf{E}\left[W^{M/M/K}\right] > \mathbf{E}\left[W^{M/M/K}\right] + \frac{1}{1-\rho}\left[\rho - \frac{K-1}{K}\right]\frac{C^2-1}{2}.$$

We provide a proof of the proposition in Appendix B.

The bounds on $W_h^{C^2}$ and $W_l^{C^2}$ in Theorem 2 are identical for $K = 1$, and in fact in this case agree with the well-known Pollaczek Khintchine formula

$$\mathbf{E}\Big[W^{M/G/1}\Big] = \left(\frac{C^2+1}{2}\right)\mathbf{E}\Big[W^{M/M/1}\Big], \tag{4}$$

which shows that the mean waiting time is completely determined by C^2 and $\mathbf{E}[X]$.

Similar results on the gap for the mean queue length of a $GI/M/1$ queue were derived by Whitt [46] by considering extremal interarrival time distributions. For the $GI/M/1$ queue, proving such theorems is simplified due to the availability of the exact expression for the mean queue length.

Outline

Section 2 reviews existing work on obtaining closed-form, numerical and heavy-traffic approximations for $\mathbf{E}\big[W^{M/G/K}\big]$. In Section 3 we seek insights into why the first two moments of the job size distribution are insufficient for approximating the mean delay. We also seek answer to the question: "Which characteristics of the job size distribution, outside of the first two moments, are important in determining the mean waiting time?" Our insights stem from numerical experiments based on the 2-phase hyperexponential class of job size distributions. These insights help us later in proving Theorem 2. Sections 4 and 5 are devoted to proving Theorems 1 and 2, respectively. In Section 6, we address the question of the effect of higher moments of job size distribution on the mean waiting time. We present some results and conjecture on the exact span of $\mathbf{E}\big[W^{M/G/K}\big]$ given the first two moments of the job size distribution via tight two-moment bounds in Section 7. We conclude in Section 8.

2 Prior Work

While there is a large body of work on approximating the mean waiting time of an $M/G/K$ system, all the closed-form approximations only involve at most the first two moments of the job size distribution. As mentioned earlier, to the best of our knowledge, the first approximation for the mean waiting time for an $M/G/K$ queue was given in (1) by Lee and Longton [26]. This approximation is very simple, is exact for $K = 1$ and was shown to be asymptotically exact in heavy traffic by Köllerström [25]. The same expression is obtained by Nozaki and Ross [31] by making approximating assumptions about the $M/G/K$ system and solving for exact state probabilities of the approximating system, and by Hokstad [19] by starting with the exact equations and making approximations in the solution phase. Boxma et al. [32] obtain a closed-form approximation for the mean waiting time in an $M/D/K$ system, extending the heavy traffic approximation of Cosmetatos [6]. Takahashi [40] obtains expressions for mean waiting time by assuming a parametric formula. Kimura [22] uses the method of system interpolation to derive a closed-form approximation for the mean waiting time that combines analytical solutions of simpler systems.

There is also a large literature on numerical methods for approximating the mean waiting time by making much weaker assumptions and solving for state probabilities. For example, Tijms et al. [18] assume that if a departure from the system leaves behind k jobs where $1 \leq k < K$, then the time until the next departure is distributed as the minimum of k independent random variables, each of which is distributed according to the equilibrium distribution of X. If, however, the departure leaves behind $k \geq K$ jobs, then the time until the next departure is distributed as X/K. Similar approaches are followed in [19, 20, 28, 29, 36]. Miyazawa [29] uses "basic equations" to provide a unified view of approximating assumptions made in [31], [19] and [18], and to derive new approximation formulas. Boxma et al. [32] also provide a numerical approximation for $M/G/K$ which is reasonably accurate for job size distributions with low variability ($C^2 \leq 1$) by assuming a para-

metric form and matching the heavy traffic and light traffic behaviors. Other numerical algorithms include [9, 10, 11]. While these numerical methods are accurate and usually give an approximation for the entire waiting time distribution, the final expressions do not give any structural insight into the behavior of the queueing system and the effect of $M/G/K$ parameters on waiting time.

Heavy traffic, light traffic and diffusion approximations for the $M/G/K$ system have been studied in [5, 21, 25, 43, 47, 48]. The diffusion approximations used in [43] are based on many-server diffusion limits. Motivated by call center applications, there is now a huge body of literature for multiserver systems with a large number of exponential servers; see the survey paper [14] and references therein.

Bounds on the mean waiting time for $M/G/K$ queues (and more generally for $GI/GI/K$ queues) have mainly been obtained via two approaches. The first approach is by assuming various orderings (stochastic ordering, increasing convex ordering) on the distribution of job sizes (see [8, 30, 38, 44, 45]), but these tend to be very loose as approximations. Moreover, one does not always have the required strong orderings on the job size distribution. The second, and more practical, approach that started with the work of Kingman [23] is obtaining bounds on mean delay in terms of the first two moments of the interarrival and job size distributions. The best known bounds of this type for $GI/GI/K$ mean waiting time are presented by Daley [7]. Scheller-Wolf and Sigman [34] derive bounds for the case $K\rho < \lfloor \frac{K}{2} \rfloor$ which are in many cases superior to the bounds in [7]. Daley [7] also conjectures tight upper and lower bounds on $GI/GI/K$ mean waiting time in terms of the first two moments of interarrival and job size distributions, and proves a tight lower bound

$$\inf \mathbf{E}\left[W^{GI/GI/K}\right] = 0, \quad \text{when } \rho < 1 - \frac{1}{K}.$$

While bounds for $GI/GI/K$ mean waiting time are more general, they can also be loose when applied to $M/G/K$. Recently, Bertsimas and Natarajan [3] have proposed a computational approach based on semidefinite optimization to obtain bounds on the moments of waiting time in $GI/GI/K$ queues given the information of moments of the job size and the interarrival time distributions.

We differ from the prior work in that we prove $\mathbf{E}\left[W^{M/G/K}\right]$ is inapproximable within a certain factor based on just the knowledge of the first two moments of the job size distribution.

3 Insights on why two-moment approximations are not enough

Our goal in this section is to illustrate the inadequacy of the first two moments of the job size distribution for approximating $\mathbf{E}\left[W^{M/G/K}\right]$. To do this, we restrict our attention to the class of two-phase hyperexponential distributions, denoted by H_2 (see Definition 1 below). Distributions in the H_2 class are mixtures of two exponential distributions and thus have three degrees of freedom. Having three degrees of freedom provides us a method to create a set of distributions with any given first two moments ($C^2 > 1$ in the case of H_2) and analyze the effect of some other characteristic. A natural choice for this third characteristic is the *third moment* of the distribution[3]. The H_2 distribution is also convenient because it allows us to capture the effect of *small vs. large jobs* (the two phases of the hyperexponential) – an insight which will be very useful to us.

Definition 1 *Let $\mu_1 > \mu_2 \ldots > \mu_n > 0$. Let $p_i > 0$, $i = 1, \ldots, n$, be such that $\sum_{i=1}^{n} p_i = 1$. We*

[3]In [9, 47], the authors use the quantity r, which denotes the fraction of load contributed by the branch with the smaller mean, as the third parameter to specify the H_2 distribution. We choose the third moment because it is more universal and better understood than r. Further, r is an increasing function of the third moment.

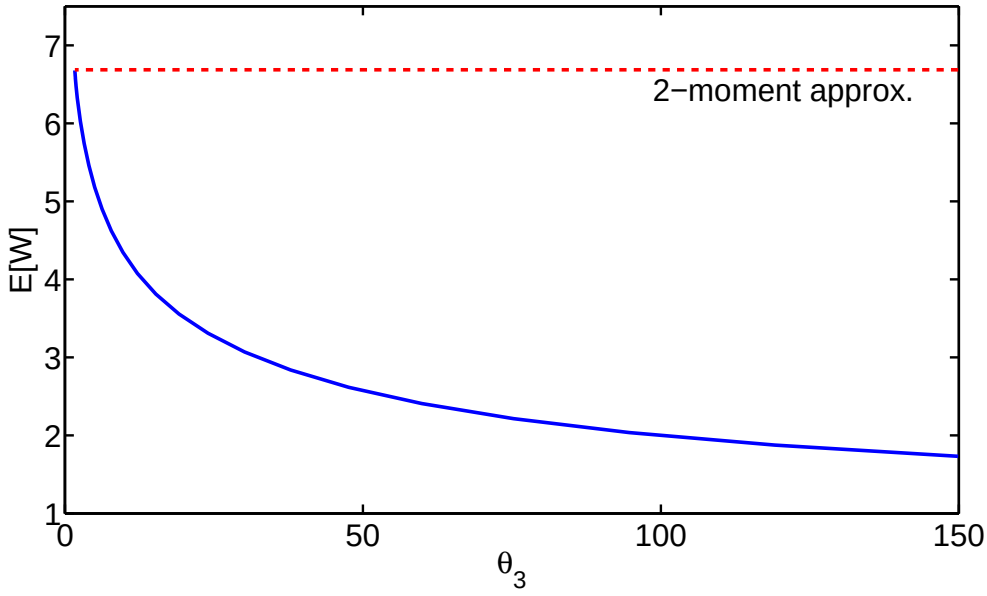


Figure 1: Illustration of the inadequacy of two-moment approximations for mean delay $\mathbf{E}\left[W^{M/G/K}\right]$. As shown, the normalized 3rd moment, θ_3, of the job size distribution has a big effect on mean waiting time of an $M/H_2/10$ system (solid line). The parameters of the job size distribution were held constant at $\mathbf{E}[X] = 1$ and $C^2 = 19$ with load $\rho = 0.9$. The dashed line shows the standard two-moment approximation of (1). The values on the $x-$axis are the normalized third moment (5).

define the $n-$phase hyperexponential distribution, H_n, with parameters μ_i, p_i, $i = 1, \ldots, n$, as:

$$H_n \sim \begin{cases} \text{Exp}(\mu_1) & \textit{with probability } p_1 \\ \text{Exp}(\mu_2) & \textit{with probability } p_2 \\ \vdots & \\ \text{Exp}\,(\mu_n) & \textit{with probability } p_n \end{cases}$$

where $\text{Exp}(\mu_i)$, $i = 1, \ldots, n$, *are n independent exponential random variables with mean* $\frac{1}{\mu_i}$, $i = 1, \ldots, n$.

Definition 2 *Let $\mu_1 > \mu_2 \ldots > \mu_{n-1} > 0$. Let $p_i > 0$, $i = 0, \ldots, n-1$, be such that $\sum_{i=0}^{n-1} p_i = 1$. We define the $n-$phase degenerate hyperexponential distribution, H_n^*, with parameters p_0, μ_i, p_i, $i = 1, \ldots, n-1$, as:*

$$H_n^* \sim \begin{cases} 0 & \textit{with probability } p_0 \\ \text{Exp}(\mu_1) & \textit{with probability } p_1 \\ \vdots & \\ \text{Exp}\,(\mu_{n-1}) & \textit{with probability } p_{n-1} \end{cases}$$

where $\text{Exp}(\mu_i)$, $i = 1, \ldots, n-1$, *are $n-1$ independent exponential random variables with mean* $\frac{1}{\mu_i}$, $i = 1, \ldots, n-1$.

Figure 1 shows the mean waiting time for an $M/H_2/K$ system evaluated numerically using matrix analytic methods. The dashed line shows the standard two moment approximation of (1). Note

that the $x-$axis is actually not showing $\mathbf{E}\left[X^3\right]$ but rather a normalized version of the third moment, θ_3, which we define as:

$$\theta_3 = \frac{\mathbf{E}\left[X^3\right]\mathbf{E}[X]}{\mathbf{E}\left[X^2\right]^2}. \tag{5}$$

The above normalization for the third moment with respect to the first two moments is analogous to the definition of the squared coefficient of variation, $C^2 = \frac{\mathbf{E}\ X^2}{\mathbf{E}[X]^2} - 1$, which is the scale-invariant normalization of the second moment with respect to the first moment. For positive distributions, θ_3 takes values in the range $[1, \infty)$, and our ongoing work on approximations for $\mathbf{E}\left[W^{M/G/K}\right]$ based on higher moments of job size distribution suggests that θ_3 is the right variable to look at. We will use the normalized third moment, θ_3, throughout the paper.

Our first interesting observation is that the $M/H_2/K$ mean waiting time actually *drops with an increase in the normalized third moment* of X. We also observe that the existing two moment approximation is insufficient as it sits at one end of the spectrum of possible values for $\mathbf{E}\left[W^{M/H_2/K}\right]$. For lower values of the third moment the approximation is good, but it is very inaccurate for high values. Moreover, *any* approximation based only on the first two moments will be inaccurate for some distribution because the span of possible values of mean waiting time for the same first two moments of the job size distribution is large.

While the drop in mean waiting time with increasing θ_3 seems very counterintuitive, this phenomenon can partially be explained by looking at how increasing θ_3 alters the distribution of load among the small and large jobs. Let $\rho(x)$ represent the fraction of load made up by jobs of size smaller than x. If $f(x)$ represents the probability density function of the job size distribution, then,

$$\rho(x) = \frac{1}{\mathbf{E}[X]} \int_0^x uf(u)du.$$

In Figure 2, we show the $\rho(x)$ curves for distributions in the H_2 class with mean 1, $C^2 = 19$ and different values of θ_3. As reference, we also show the $\rho(x)$ curve for the exponential distribution with mean 1. As can be seen from Figure 2, increasing θ_3 while holding fixed the first two moments of the H_2 distribution, causes the load to (almost monotonically) shift towards smaller jobs. While the large jobs also become larger, they become rarer at an even faster rate so that in the limit as $\theta_3 \to \infty$, the $\rho(x)$ curve for the H_2 distribution converges to the $\rho(x)$ curve for the exponential distribution with the same mean. Thus as θ_3 increases, the $M/H_2/K$ system sees smaller jobs more often, thereby causing a smaller mean waiting time. In fact, this behavior would hold for any $M/G/K$ system where the job size distribution is a mixture of two scaled versions of an arbitrary distribution.

Based on the numerical evidence of the huge variation in $\mathbf{E}\left[W^{M/H_2/K}\right]$, a natural question that arises is: Can this span of possible values of $\mathbf{E}\left[W^{M/H_2/K}\right]$ be quantified? Lemmas 3 and 4 in Section 5 answer this question. Lemma 3 is obtained by considering the case of a distribution in the H_2 class with a low θ_3. In particular, we consider the case of an H_2^* distribution (see Definition 2) which we can prove has the lowest possible third moment of all distributions in the H_2 family (with any given first two moments), and we derive the exact mean waiting time under the H_2^* jobs size distribution. Likewise, Lemma 4 is derived by considering the case of an H_2 distribution where θ_3 goes to ∞ and we derive the asymptotic mean waiting time for that situation. Since we restrict our attention to a subset of the entire space of distributions with given first two moments, our results provide a lower bound on the exact span of $\mathbf{E}\left[W^{M/G/K}\right]$. We conjecture the exact span of $\mathbf{E}\left[W^{M/G/K}\right]$ in Section 7, Conjecture 1.

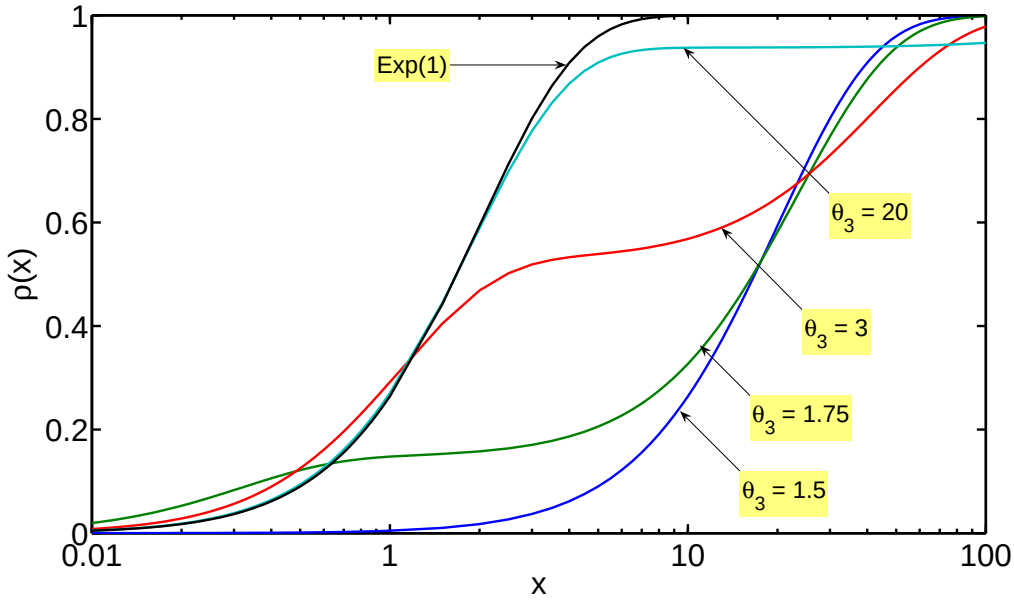


Figure 2: Illustration of the effect of the normalized 3rd moment, θ_3, on the distribution of load as a function of job size for the H_2 class of distributions. The first two moments were held constant at $\mathbf{E}[X] = 1$ and $C^2 = 19$. The distribution of the load for exponential distribution with mean 1, labeled Exp(1), is shown for reference.

4 Proof of Theorem 1

To obtain the bounds on $W_h^{C^2}$ and $W_l^{C^2}$ in Theorem 1, it suffices to show the existence of job size distributions with SCV C^2 which give the desired expressions for mean waiting times. To obtain an upper bound on $W_l^{C^2}$, we use a corollary of [7], Proposition 3.15:

Lemma 1 (Daley [7, Proposition 3.15]) *For any $C^2 > 0$ and $0 < \epsilon < \sqrt{\frac{1}{C^2}}$, define the following two-point job size distribution:*

$$D^{(\epsilon)} \sim \begin{cases} 1 - \epsilon\sqrt{C^2} & \text{with probability } \frac{1}{1+\epsilon^2} \\ 1 + \frac{\sqrt{C^2}}{\epsilon} & \text{with probability } \frac{\epsilon^2}{1+\epsilon^2}. \end{cases}$$

For $\rho < \frac{K-1}{K}$ and any given GI arrival process,

$$\lim_{\epsilon \to 0} \mathbf{E}\left[W^{GI/D^{(\epsilon)}/K}\right] = \mathbf{E}\left[W^{GI/D/K}\right]$$

where $\mathbf{E}\left[W^{GI/D/K}\right]$ is the mean waiting time when the job size distribution is deterministic 1.

By definition, each distribution in the $D^{(\epsilon)}$ family has mean 1 and SCV C^2. The bound on $W_l^{C^2}$ follows by setting $GI \equiv M$.

To obtain a lower bound on $W_h^{C^2}$, we consider the following two-point distribution:

$$D_2^* \sim \begin{cases} 0 & \text{with probability } \frac{C^2}{C^2+1} \\ C^2 + 1 & \text{with probability } \frac{1}{C^2+1}. \end{cases}$$

It is easy to verify that the above distribution has mean 1, squared coefficient of variation C^2 and $\theta_3 = 1$. We denote the $M/G/K$ system with D_2^* job size distribution as $M/D_2^*/K$.

The bound on $W_h^{C^2}$ follows from the following lemma:

Lemma 2 *For any $\rho < 1$ and $C^2 > 0$,*

$$\mathbf{E}\Big[W^{M/D_2^*/K}\Big] = (C^2+1)\mathbf{E}\Big[W^{M/D/K}\Big].$$

Proof: Since the scheduling discipline is size independent, the distribution of waiting time experienced by zero-sized jobs and non-zero jobs is identical. Further, to find the waiting time distribution experienced by non-zero sized jobs, we can ignore the presence of zero-sized jobs. The waiting time distribution of the non-zero sized jobs is thus equivalent to the waiting time distribution in an $M/D/K$ system with arrival rate $\frac{\lambda}{C^2+1}$ and mean job size (C^2+1). The latter system, however, is just an $M/D/K$ system with arrival rate λ and mean job size 1 seen on a slower time scale, slowed by a factor (C^2+1). Hence, the mean waiting time of the original system is also (C^2+1) times the mean waiting time of an $M/D/K$ system with arrival rate λ and mean job size 1. ■

5 Proof of Theorem 2

As in the proof of Theorem 1, to obtain the bounds on $W_h^{C^2}$ and $W_l^{C^2}$ in Theorem 2, it suffices to show the existence of job size distributions with SCV C^2 which give the desired mean waiting times. To handle the case $\rho > \frac{K-1}{K}$, we resort to job size distributions in the class of 2-phase hyperexponentials.[4]

To obtain a lower bound on $W_h^{C^2}$, we consider the following degenerate hyperexponential distribution:

$$H_2^* \sim \begin{cases} 0 & \text{with probability } \frac{C^2-1}{C^2+1} \\ \text{Exp}\left(\frac{2}{C^2+1}\right) & \text{with probability } \frac{2}{C^2+1}. \end{cases}$$

It is easy to verify that the above distribution has mean 1, squared coefficient of variation C^2 and $\theta_3 = \frac{3}{2}$. The H_2^* distribution as defined above has the lowest third moment among all the H_n distributions with mean 1 and SCV C^2:

Claim 1 *Let $\cup_{n>1}\{H_n|C^2\}$ be the set of all hyperexponential distributions with finite number of phases, mean 1 and squared coefficient of variation C^2 ($C^2 > 1$). The H_2^* distribution lying in this set has the smallest third moment among all the distributions in $\cup_{n>1}\{H_n|C^2\}$.*

Proof: See Appendix A. ■

The bound on $W_h^{C^2}$ in Theorem 2 follows from the following lemma which can be proved along the lines of Lemma 2:

Lemma 3 *For any $\rho < 1$ and $C^2 > 1$,*

$$\mathbf{E}\Big[W^{M/H_2^*/K}\Big] = \left(\frac{C^2+1}{2}\right)\mathbf{E}\Big[W^{M/M/K}\Big].$$

Note that the bound obtained from Lemma 3 is weaker than the bound from Lemma 2 since $\mathbf{E}\big[W^{M/M/K}\big] < 2 \cdot \mathbf{E}\big[W^{M/D/K}\big]$. We present Lemma 3 here for comparison with the corresponding upper bound on $W_l^{C^2}$ in Lemma 4 and the 2-moment approximation (1), which involve $\mathbf{E}\big[W^{M/M/K}\big]$.

[4]The reader may wonder why we don't use the two-point job size distributions from Section 4. The use of two-point distributions may lead to stronger bounds, as we conjecture in Section 7, and we believe that our proof technique can be extended to the class of two-point job size distributions. But the additional complexity of doing so is beyond the scope of this paper.

To obtain a bound on $W_l^{C^2}$, we consider a sequence of systems parameterized by a parameter ϵ in which we *fix the first two moments of the job size distribution* analogous to Lemma 1. The parameter ϵ allows for increasing the third moment as ϵ goes to 0. More precisely, we consider the sequence of queues $M/H_2^{(\epsilon)}/K$ (see Section 5.2, Definition 3) as $\epsilon \to 0$ and prove the following limit theorem:

Lemma 4 *For any finite C^2,*

$$\lim_{\epsilon \to 0} \mathbf{E}\Big[W^{M/H_2^{(\epsilon)}/K}\Big] = \begin{cases} \mathbf{E}\big[W^{M/M/K}\big] & \text{if } \rho < \frac{K-1}{K} \\ \mathbf{E}\big[W^{M/M/K}\big] + \frac{1}{1-\rho}\left[\rho - \frac{K-1}{K}\right]\frac{C^2-1}{2} & \text{if } \rho \geq \frac{K-1}{K} \end{cases}$$

where $\mathbf{E}\big[W^{M/M/K}\big]$ is the mean waiting time when the job size distribution is exponential with mean 1.

The rest of this section is devoted to proving Lemma 4. Since the proof of Lemma 4 involves a new technique, we begin in Section 5.1 with a high level proof idea. Subsequent subsections will provide the rigorous lemmas.

5.1 Proof idea

The key steps involved in the analysis are as follows:

1. We first observe that the $H_2^{(\epsilon)}$ job size distribution is made up of two classes of jobs – small jobs and large jobs. We use N_s and N_ℓ to denote the number of small and large jobs in system, respectively.

2. We show that the expected number of large jobs, $\mathbf{E}\Big[N_\ell^{M/H_2^{(\epsilon)}/K}\Big]$, vanishes as ϵ goes to zero; therefore it suffices to consider only small jobs (see Section 5.3).

3. For each $M/H_2^{(\epsilon)}/K$ system, we construct another system, $U^{(\epsilon)}$, which stochastically upper bounds the number of small jobs in the corresponding $M/H_2^{(\epsilon)}/K$ system. That is,

   $$N_s^{M/H_2^{(\epsilon)}/K} \leq_{st} N_s^{U^{(\epsilon)}}$$

   (see Section 5.4).

4. To analyze $N_s^{U^{(\epsilon)}}$, we consider two kinds of periods: **good** periods – when there are no large jobs in the system, and **bad** periods – when there is at least one large job in the system. Our approach is to obtain upper bounds on the mean number of small jobs during the good and bad periods, $\mathbf{E}\Big[N_s^{U^{(\epsilon)}} | \text{ good period}\Big]$ and $\mathbf{E}\Big[N_s^{U^{(\epsilon)}} | \text{ bad period}\Big]$, respectively, and obtain an upper bound on $\mathbf{E}\Big[N_s^{U^{(\epsilon)}}\Big]$ using the law of total probability:

   $$\mathbf{E}\Big[N_s^{U^{(\epsilon)}}\Big] = \mathbf{E}\Big[N_s^{U^{(\epsilon)}} | \text{ good period}\Big]\mathbf{Pr}[\text{good period}] + \mathbf{E}\Big[N_s^{U^{(\epsilon)}} | \text{ bad period}\Big]\mathbf{Pr}[\text{bad period}]$$

   We obtain upper bounds on the mean number of small jobs during the good and bad periods using the following steps (see Section 5.5):

   (a) We first look at the number of small jobs only at *switching points*. That is, we consider the number of small jobs only at the instants when the system switches from a good period to a bad period and vice versa.

(b) To obtain bounds on the number of small jobs at the switching points, we define a random variable Δ, which upper bounds the *increment* in the number of small jobs during a bad period. Further, by our definition, the upper bound Δ is independent of the number of small jobs at the beginning of the bad period. To keep the analysis simple, this independence turns out to be crucial.

(c) Next we obtain a stochastic upper bound on the number of small jobs at the end of a good period by solving a fixed point equation of the form

$$A \stackrel{d}{=} \Phi(A + \Delta)$$

where A is the random variable for (the stochastic upper bound on) the number of small jobs at the end of a good period, and Φ is a function that maps the number of small jobs at the beginning of a good period to the number of small jobs at the end of the good period.

(d) Finally, we obtain the mean number of small jobs *during* the good and bad periods from the mean number of small jobs at the switching points.

5. Similar to $U^{(\epsilon)}$, for each $M/H_2^{(\epsilon)}/K$ system, we also construct a system, $L^{(\epsilon)}$, which stochastically lower bounds the number of small jobs in the corresponding $M/H_2^{(\epsilon)}/K$ system. That is,

$$N_s^{M/H_2^{(\epsilon)}/K} \geq_{st} N_s^{L^{(\epsilon)}}$$

(see Section 5.6). We omit the analysis of $L^{(\epsilon)}$ since it is similar to analysis of $U^{(\epsilon)}$. Note, that we indeed obtain

$$\mathbf{E}\Big[N_s^{U^{(\epsilon)}}\Big] = \mathbf{E}\Big[N_s^{L^{(\epsilon)}}\Big] + o(1)$$

Convergence of $\mathbf{E}\Big[N^{M/H_2^{(\epsilon)}/K}\Big]$ follows from convergence of its upper and lower bounds.

6. Finally, we use Little's law to obtain mean waiting time, $\mathbf{E}\Big[W^{M/H_2^{(\epsilon)}/K}\Big]$, from the mean number of waiting jobs, $\mathbf{E}\Big[N^{M/H_2^{(\epsilon)}/K}\Big] - K\rho$.

5.2 Preliminaries

Below we give a formal definition of the $H_2^{(\epsilon)}$ class of job size distributions.

Definition 3 *We define a family of distributions parameterized by ϵ as follows:*

$$H_2^{(\epsilon)} = \begin{cases} \text{Exp}\left(\mu_s^{(\epsilon)}\right) & \textit{with probability } p^{(\epsilon)} \\ \text{Exp}\left(\mu_\ell^{(\epsilon)}\right) & \textit{with probability } 1 - p^{(\epsilon)} \end{cases}$$
$$\mu_s^{(\epsilon)} > \mu_\ell^{(\epsilon)}$$

where $\mu_s^{(\epsilon)}$, $\mu_\ell^{(\epsilon)}$ *and* $p^{(\epsilon)}$ *satisfy,*

$$\frac{p^{(\epsilon)}}{\mu_s^{(\epsilon)}} + \frac{1-p^{(\epsilon)}}{\mu_\ell^{(\epsilon)}} = \mathbf{E}\Big[X^{(\epsilon)}\Big] = 1$$
$$2\frac{p^{(\epsilon)}}{\left(\mu_s^{(\epsilon)}\right)^2} + 2\frac{1-p^{(\epsilon)}}{\left(\mu_\ell^{(\epsilon)}\right)^2} = \mathbf{E}\left[\left(X^{(\epsilon)}\right)^2\right] = C^2+1$$
$$6\frac{p^{(\epsilon)}}{\left(\mu_s^{(\epsilon)}\right)^3} + 6\frac{1-p^{(\epsilon)}}{\left(\mu_\ell^{(\epsilon)}\right)^3} = \mathbf{E}\left[\left(X^{(\epsilon)}\right)^3\right] = \frac{1}{\epsilon}$$

For proving the upper bound on the lower bound $W_l^{C^2}$ of $\mathbf{E}[W]$, we look at $\mathbf{E}\Big[W^{M/H_2^{(\epsilon)}/K}\Big]$ as $\epsilon \to 0$. That is, the third moment of service time goes to ∞. Below we present some elementary results on the asymptotic behavior[5] of the parameters of the $H_2^{(\epsilon)}$ distribution, which will be used in the analysis in Section 5.5.

Lemma 5 *The* $\mu_s^{(\epsilon)}$, $\mu_\ell^{(\epsilon)}$ *and* $p^{(\epsilon)}$ *can be expressed in terms of* ϵ *as :*

$$\mu_s^{(\epsilon)} = 1 + \frac{3}{2}(C^2-1)^2\epsilon + \Theta(\epsilon^2)$$
$$\mu_\ell^{(\epsilon)} = 3(C^2-1)\epsilon + 18C^2(C^2-1)\epsilon^2 + \Theta(\epsilon^3)$$
$$p^{(\epsilon)} = 1 - \frac{9}{2}(C^2-1)^3\epsilon^2 + \Theta(\epsilon^3)$$

Proof in Appendix A.

Corollary 1 *As* $\epsilon \to 0$,

$$\begin{array}{rclcrcl} p^{(\epsilon)} & \to & 1 & , & \mu_s^{(\epsilon)} & \to & 1 \\ \frac{1-p^{(\epsilon)}}{\mu_\ell^{(\epsilon)}} & \to & 0 & , & \frac{1-p^{(\epsilon)}}{{\mu_\ell^{(\epsilon)}}^2} & \to & \frac{C^2-1}{2} \end{array}$$

[5]We will use the following asymptotic notation frequently in this paper: We say a function $h(\epsilon)$ is:

1. $\Theta(g(\epsilon))$ if

$$0 < \liminf_{\epsilon\to 0} \frac{h(\epsilon)}{g(\epsilon)} \leq \limsup_{\epsilon\to 0} \frac{h(\epsilon)}{g(\epsilon)} < \infty$$

   Intuitively, this means that the functions h and g grow at the same rate, asymptotically, as $\epsilon \to 0$.

2. $o(g(\epsilon))$ if

$$\lim_{\epsilon\to 0} \frac{h(\epsilon)}{g(\epsilon)} = 0$$

   Intuitively, h becomes insignificant when compared with g, asymptotically, as $\epsilon \to 0$.

3. $O(g(\epsilon))$ if

$$\limsup_{\epsilon\to 0} \frac{h(\epsilon)}{g(\epsilon)} < \infty$$

   That is, h is either $\Theta(g(\epsilon))$ or $o(g(\epsilon))$.

Corollary 1 formalizes the observation we made from Figure 2: As the third moment grows, asymptotically, all the load is made up *only* by the small jobs, whose mean approaches 1. While the mean size of the large jobs also grows linearly in the third moment (asymptotically), the probability that a large job arrives vanishes at a faster rate. Thus, intuitively, our $M/H_2^{(\epsilon)}/K$ system rarely encounters a large job in the limit as $\epsilon \to 0$.

It is important to point out that, as $\epsilon \to 0$, the $H_2^{(\epsilon)}$ distribution converges in distribution to the Exp(1) distribution. Thus, the stationary queue length and waiting time distributions of the sequence of $M/H_2^{(\epsilon)}/K$ systems also converge in distribution to the queue length and waiting time distributions of the corresponding $M/M/K$ system [4, 37]. However, convergence in distribution of the waiting time *does not* imply convergence of the mean waiting time; namely, it is possible that

$$\lim_{\epsilon \to 0} \mathbf{E}\Big[W^{M/H_2^{(\epsilon)}/K}\Big] \neq \mathbf{E}\Big[W^{M/M/K}\Big]. \tag{6}$$

Indeed, (6) can be verified for $K = 1$ where the mean waiting time is given by the Pollaczek-Khintchine formula (4). Lemma 4 proves that the non-convergence (6) also holds for the $M/H_2^{(\epsilon)}/K$ system when $\rho > \frac{K-1}{K}$.

Daley [7] proved an analogous non-convergence result by considering a class of job size distributions, $S^{(\epsilon)}$, which includes $H_2^{(\epsilon)}$ job size distributions. He further conjectured [7, Conjecture 3.19] an expression for the difference,

$$\lim_{\epsilon \to 0} \mathbf{E}\Big[W^{GI/S^{(\epsilon)}/K}\Big] - \mathbf{E}\Big[W^{GI/S/K}\Big],$$

where S denotes the limiting job size distribution. The proof of Lemma 4 verifies Daley's conjecture for the case of Poisson arrival process and H_2 job size distribution.

5.3 Bounding the number of large jobs

The following lemma proves that to bound the mean number of jobs in an $M/H_2^{(\epsilon)}/K$ system within $o(1)$, it suffices to consider only the small jobs.

Lemma 6 $\mathbf{E}\Big[N_\ell^{M/H_2^{(\epsilon)}/K}\Big] = o(1)$

Proof: We will upper bound the expected number of large customers in the system by (a) giving high priority to the small customers and letting the large jobs receive service only when there are no small jobs in the system, and (b) by allowing the large customers to be served by at most one server at any time. Further, we increase the arrival rate of small customers to λ and increase the mean size of the small customers to 1. By not being work conserving, increasing the arrival rate, and making small jobs stochastically larger, the modified system can become overloaded. However, since we are only interested in the asymptotic behavior as $\epsilon \to 0$, it suffices to find an ϵ' such that the above system is stable for all $\epsilon < \epsilon'$. This is indeed true for $\epsilon' = \frac{1}{6}\left[\frac{K\rho(C^2+1)^2}{4(1-\rho)} + 1\right]^{-1}$ (See proof of Lemma 9).

For brevity, we use $M(a)/M(b)/k$ to denote an $M/M/k$ queue with arrival rate a and service rate b. Let $\overline{N_\ell}^{(\epsilon)}$ be the steady-state number of customers in an $M\left(\lambda(1-p^{(\epsilon)})\right)/M\left(\mu_\ell^{(\epsilon)}\right)/1$ queue with service interruptions, where the server is interrupted for the duration of the busy period of an $M(\lambda)/M(1)/K$ queue. It is easy to see that

$$\mathbf{E}\Big[N_\ell^{M/H_2^{(\epsilon)}/K}\Big] \leq \mathbf{E}\Big[\overline{N_\ell}^{(\epsilon)}\Big].$$

The proof is completed by the following lemma:

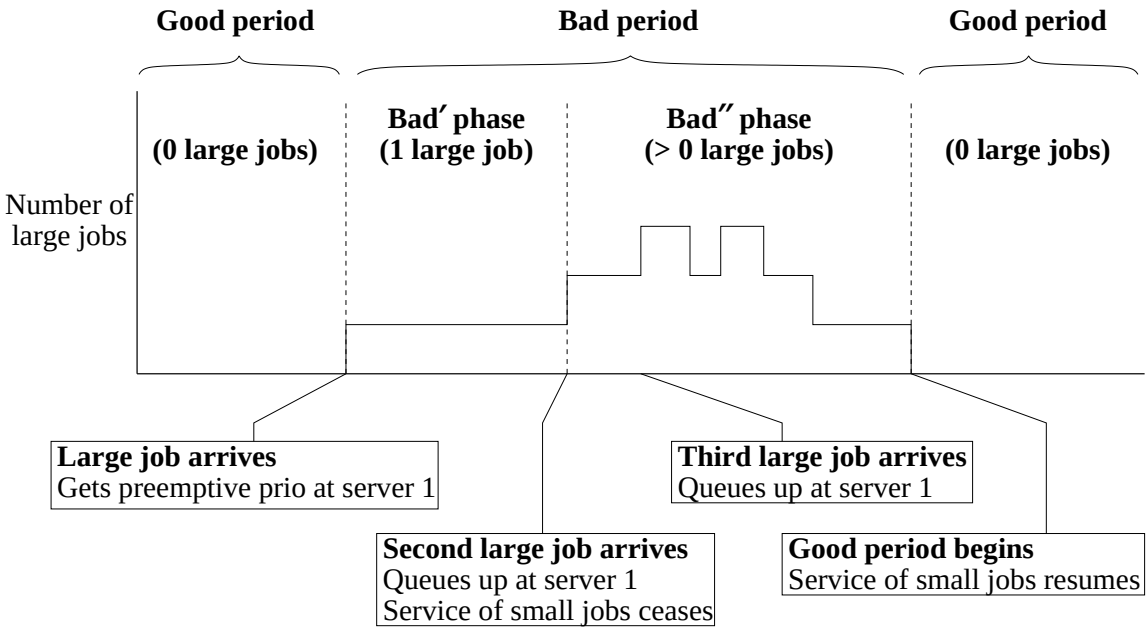


Figure 3: Construction of system $U^{(\epsilon)}$ which upper bounds the number of jobs in an $M/H_2^{(\epsilon)}/K$

Lemma 7 $\mathbf{E}\left[\overline{N_\ell}^{(\epsilon)}\right] = o(1)$

Proof in Appendix A. ∎

5.4 Construction of $U^{(\epsilon)}$: the upper bounding system for $N_s^{M/H_2^{(\epsilon)}/K}$

Figure 3 illustrates the behavior of system $U^{(\epsilon)}$, which upper bounds the number of small jobs in an $M/H_2^{(\epsilon)}/K$. Denote periods where there are no large jobs (including when the system is idle) as *good* periods, and periods when there is at least 1 large job as a *bad* period. During a good period, the small jobs receive service according to a normal K server FIFO system. As soon as a large job arrives, we say that a bad period begins. The bad period consists of up to 2 phases, called *bad′* and *bad″*. A bad′ phase spans the time from when a large job first arrives until either it leaves or a second large job arrives (whichever happens earlier). A bad″ phase occurs if a second large job arrives while the first large job is still in the system, and covers the period from when this 2nd large job arrives (if it does) until there are no more large jobs in the system.

The large job starting a bad period preempts the small job at server 1 (if any) and starts receiving service. The small jobs are served by the remaining $(K-1)$ servers. If a second large job arrives during a bad period while the first large job is still in system, starting a bad″ phase, we cease serving the small jobs and continue serving the large jobs by *only* server 1 until this busy period of large jobs ends (there are no more large jobs). When the last large job leaves, we resume the service of small jobs according to a normal K server FIFO system.

Analyzing system $U^{(\epsilon)}$ is simpler than analyzing the corresponding $M/H_2^{(\epsilon)}/K$ system because in $U^{(\epsilon)}$, the large jobs form an $M/M/1$ system independent of the small jobs, due to preemptive priority and service by only one server. The small jobs operate in a random environment where they have either K, $(K-1)$ or 0 servers.

Lemma 8 *The number of small jobs in an $M/H_2^{(\epsilon)}/K$ system, $N_s^{M/H_2^{(\epsilon)}/K}$, is stochastically upper bounded by the number of small jobs in the corresponding system $U^{(\epsilon)}$, $N_s^{U^{(\epsilon)}}$.*

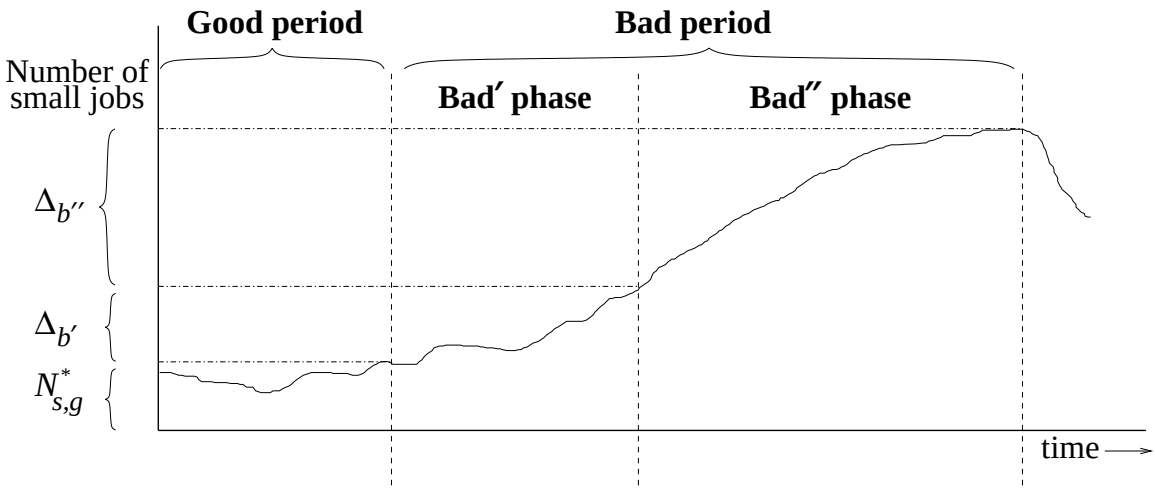


Figure 4: Notation used for analysis of system $U^{(\epsilon)}$

Proof: Straightforward using stochastic coupling. ■

Stability of system $U^{(\epsilon)}$: Since system $U^{(\epsilon)}$ is not work conserving, there are values of ϵ for which it is unstable, even when $\rho < 1$. Therefore we restrict our attention to the following range of ϵ:

Lemma 9 *The upper bounding system, $U^{(\epsilon)}$, is stable for $\epsilon < \epsilon'$ where*

$$\epsilon' = \frac{1}{6}\left[\frac{K\rho(C^2+1)^2}{4(1-\rho)} + 1\right]^{-1}.$$

Proof in Appendix A.

5.5 Analysis of system $U^{(\epsilon)}$

Figure 4 introduces the notation we will use in this section. Since in this section we focus only on the analysis of system $U^{(\epsilon)}$, we will omit superscripting the random variables used in analysis by $U^{(\epsilon)}$ for readability. Unless explicitly superscripted, random variables correspond to the $U^{(\epsilon)}$ system. We define the following random variables:

- $N^*_{s,g} \equiv$ the number of small jobs *at the end* of a good period, that is, when the system switches from a good to a bad period
- $N^*_{s,b} \equiv$ the number of small jobs *at the end* of a bad period, that is, when the system switches from a bad to a good period
- $N_{s,g} \equiv$ the time stationary number of small jobs *during* a good period
- $N_{s,b} \equiv$ the time stationary number of small jobs *during* a bad period
- $\Delta_{b'} \equiv$ the *increment* in the number of small jobs during a bad$'$ period (when small jobs have $(K-1)$ servers available)
- $\Delta_{b'}(n) \equiv$ the *increment* in the number of small jobs during a bad$'$ period given that the bad$'$ period begins with n small jobs
- $\Delta_{b''} \equiv$ the *increment* in the number of small jobs during a bad$''$ period (where the service of small jobs has been blocked)

- $\Delta_b = \Delta_{b'}(0) + \Delta_{b''}$

We denote the fraction of time spent in a good, bad, bad$'$ and bad$''$ phase by $\mathbf{Pr}[g]$, $\mathbf{Pr}[b]$, $\mathbf{Pr}[b']$ and $\mathbf{Pr}[b'']$ respectively.

By the law of total probability,

$$\mathbf{E}[N_s] = \mathbf{E}[N_{s,g}]\mathbf{Pr}[g] + \mathbf{E}[N_{s,b}]\mathbf{Pr}[b] \tag{7}$$

In Section 5.5.1, we derive stochastic upper bounds on $N_{s,g}$ and $N_{s,b}$, which give us an upper bound, (9), on $\mathbf{E}[N_s]$. In Sections 5.5.2 and 5.5.3, we derive expressions for the quantities appearing in (9). These are used to obtain the final upper bound on $\mathbf{E}[N_s]$ in Section 5.5.4.

5.5.1 Stochastic Bounds

Obtaining a stochastic upper bound on $N_{s,g}$: Let $\Phi(A)$ be a mapping between non-negative random variables where $\Phi(A)$ gives the random variable for the number of small jobs at the end of a good period, given that the number at the beginning of the good period is given by A. Let $\bar{N}^*_{s,g}$ be the solution to the following fixed point equation:

$$\bar{N}^*_{s,g} \stackrel{d}{=} \Phi(\bar{N}^*_{s,g} + \Delta_b) \tag{8}$$

Lemma 10

$$N_{s,g} \stackrel{d}{=} N^*_{s,g} \leq_{st} \bar{N}^*_{s,g}$$

Proof sketch: The first relation follows since the length of a good period is exponential and its termination is independent of the number of small jobs. Hence, by *conditional* PASTA [42] (see also [16] for a similar use of conditional PASTA),

$$N_{s,g} \stackrel{d}{=} N^*_{s,g}$$

Intuitively, Δ_b stochastically upper bounds the increment in the number of small jobs during a bad period since it assumes there were zero small jobs at the beginning of the bad period and hence ignores the departures of those small jobs. Therefore, solving the fixed point equation (8) gives a stochastic upper bound on $N^*_{s,g}$. A formal proof of the stochastic inequality is in Appendix A. ■

Obtaining a stochastic upper bound on $N_{s,b}$: The required upper bound is given by the following lemma.

Lemma 11

$$N_{s,b} \leq_{st} \bar{N}^*_{s,g} + \Delta_{b'}(0) + \mathbf{I}_{b''|b} A_\lambda\left(T_{b''e}\right)$$

where $A_\lambda\left(T_{b''e}\right)$ *is the number of arrivals of a Poisson process (with rate* λ*) during a random time interval* $T_{b''e}$ *denoting the* excess *of the length of a bad*$''$ *period, and where* $\mathbf{I}_{b''|b}$ *denotes an indicator random variable which is* 1 *with probability* $\mathbf{Pr}[b'']/\mathbf{Pr}[b]$.

Proof sketch: Observe that the first term in the upper bound is a stochastic upper bound on the number of small jobs at the beginning of a bad period. The second term denotes a stochastic upper bound on the increment in the number of small jobs during a bad$'$ phase. Finally, the third term denotes the "average increment" in the number of small jobs during a bad$''$ phase. See Appendix A for the complete proof. ■

Combining the bounds on $N_{s,g}$ and $N_{s,b}$, we get an upper bound on $\mathbf{E}[N_s]$:

$$\mathbf{E}[N_s] \leq \mathbf{E}\left[\bar{N}^*_{s,g}\right]\mathbf{Pr}[g] + \mathbf{E}\left[\bar{N}^*_{s,g} + \Delta_{b'}(0) + \mathbf{I}_{b''|b} A_\lambda\left(T_{b''e}\right)\right]\mathbf{Pr}[b] \tag{9}$$

To complete the proof, we need expressions for each of the quantities in equation (9). In Section 5.5.2 we will obtain expressions for $\mathbf{E}[\Delta_{b'}(0)]$ for the cases $\rho < \frac{K-1}{K}$ and $\rho \geq \frac{K-1}{K}$. In Section 5.5.3 we will obtain $\mathbf{E}\left[\bar{N}^*_{s,g}\right]$. However, to do this, we will need the first two moments of Δ_b, $\mathbf{E}[\Delta_b]$ and $\mathbf{E}\left[\Delta_b^2\right]$, which are also derived in Section 5.5.2.

To obtain $\mathbf{Pr}[b]$, recall that the large jobs form an $M/M/1$ system. Hence (see Lemma 5 for expressions for p and μ_ℓ),

$$\begin{aligned}\mathbf{Pr}[b] = \mathbf{Pr}[\geq 1 \text{ large job}] &= \frac{\lambda(1-p^{(\epsilon)})}{\mu_\ell^{(\epsilon)}} \\ &= \frac{3K\rho(C^2-1)^2\epsilon}{2} + \Theta(\epsilon^2)\end{aligned} \tag{10}$$

The following asymptotic behavior of $\frac{\mathbf{Pr}[b'']}{\mathbf{Pr}[b]}\mathbf{E}[A_\lambda\left(T_{b''e}\right)]$ is proved in the proof of Lemma 14:

$$\frac{\mathbf{Pr}[b'']}{\mathbf{Pr}[b]}\mathbf{E}[A_\lambda\left(T_{b''e}\right)] = \Theta(1) \tag{11}$$

In Section 5.5.4, we perform the final calculations by substituting the above quantities into (9).

5.5.2 Obtaining $\mathbf{E}[\Delta_b]$ and $\mathbf{E}\left[\Delta_b^2\right]$

Recall that we defined,

$$\Delta_b = \Delta_{b'}(0) + \Delta_{b''}$$

where $\Delta_{b'}(0)$ is the random variable for the number small jobs at the end of a bad$'$ phase given that it starts with 0 small jobs and $\Delta_{b''}$ is the number of small of jobs that arrive during a bad$''$ phase.

Lemma 12 gives the expressions for $\mathbf{E}[\Delta_{b'}(0)]$ and $\mathbf{E}\left[\Delta^2_{b'}(0)\right]$. Lemma 14 gives the asymptotic expressions for $\mathbf{E}[\Delta_{b''}]$ and $\mathbf{E}\left[\Delta^2_{b''}\right]$ which will be sufficient for our purposes of obtaining $\mathbf{E}[N_s]$ within $o(1)$.

Lemma 12
Case: $\rho < \frac{K-1}{K}$

$$\begin{aligned}\mathbf{E}[\Delta_{b'}(0)] &= O(1) \\ \mathbf{E}\left[\Delta^2_{b'}(0)\right] &= O(1)\end{aligned}$$

Case: $\rho > \frac{K-1}{K}$

$$\begin{aligned}\mathbf{E}[\Delta_{b'}(0)] &= \frac{K\left(\rho - \frac{K-1}{K}\right)}{3(C^2-1)\epsilon} + \Theta(1) \\ \mathbf{E}\left[\Delta^2_{b'}(0)\right] &= \frac{2}{9}\frac{K^2\left(\rho - \frac{K-1}{K}\right)^2}{(C^2-1)^2\epsilon^2} + \Theta\left(\frac{1}{\epsilon}\right)\end{aligned}$$

Proof: We can think of $\Delta_{b'}(0)$ as the number of jobs in an $M/M/K-1$ with arrival rate $\lambda_s = \lambda p$ and service rate μ_s at time $T \sim \text{Exp}\,(\beta)$ ($\beta = \lambda(1-p) + \mu_\ell$) given that it starts empty. Let us call this $N^{M(\lambda_s)/M(\mu_s)/K-1}(T)$. Let $N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)$ be the number of jobs in an $M/M/1$ with arrival rate λ_s and service rate $(K-1)\mu_s$ at time T given that it starts empty. Then,

$$N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T) \leq_{st} N^{M(\lambda_s)/M(\mu_s)/K-1}(T) \leq_{st} \quad N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T) + (K-1) \quad (12)$$

To see why (12) is true, first note that using coupling, $N^{M(\lambda_s)/M(\mu_s)/K-1}(T)$ can be (stochastically) sandwiched between $N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)$ and the number of jobs in an $M/M/K-1$ where the service is stopped when the number of jobs goes below $K-1$. Finally, again using coupling, the number of jobs in this latter system can be stochastically upper bounded by $N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T) + (K-1)$.

Therefore, using (12), we only need to evaluate the first and second moments of $N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)$ to obtain $\mathbf{E}[\Delta_{b'}(0)]$ and $\mathbf{E}\left[\Delta^2_{b'}(0)\right]$ within an error of $\Theta(1)$ and $\Theta(\mathbf{E}[\Delta_{b'}(0)])$, respectively. We do this next.

Case: $\rho < \frac{K-1}{K}$

For this case the $M/M/K-1$ system is stable during bad$'$ phases, and hence

$$\mathbf{E}[\Delta_{b'}(0)] = O(1)$$
$$\mathbf{E}\left[\Delta^2_{b'}(0)\right] = O(1).$$

Case: $\rho > \frac{K-1}{K}$

The following lemma gives the expressions for the first and second moments of $N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)$ for the case $\rho > \frac{K-1}{K}$.

Lemma 13 *Let $T \sim \text{Exp}(\beta)$ and $\lambda_s > (K-1)\mu_s$. Then,*

$$\mathbf{E}\left[N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)\right] = \frac{\lambda_s - (K-1)\mu_s}{\beta} + \Theta(1)$$
$$\mathbf{E}\left[(N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T))^2\right] = 2\left(\frac{\lambda_s - (K-1)\mu_s}{\beta}\right)^2 + \Theta\left(\frac{1}{\beta}\right).$$

Proof of Lemma 13: See Appendix A.

Now, using the inequality (12) and Lemma 13, and substituting in the expressions for μ_s, λ_s and μ_ℓ from Lemma 5 :

$$\begin{aligned}
\mathbf{E}[\Delta_{b'}(0)] &= \mathbf{E}\left[N^{M(\lambda_s)/M(\mu_s)/K-1}(T)\right] \\
&\leq \mathbf{E}\left[N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)\right] + O(1) \\
&= \frac{\lambda_s - (K-1)\mu_s}{\beta} + \Theta(1) \\
&= \frac{\lambda p - (K-1)\mu_s}{\lambda(1-p) + \mu_\ell} + \Theta(1) \\
&= \frac{\lambda(1-\Theta(\epsilon^2)) - (K-1)(1+\Theta(\epsilon))}{\lambda\Theta(\epsilon^2) + (3(C^2-1)\epsilon + \Theta(\epsilon^2))} + \Theta(1) \\
&= \frac{K\left(\rho - \frac{K-1}{K}\right)}{3(C^2-1)\epsilon} + \Theta(1)
\end{aligned}$$

and,

$$\begin{aligned}
\mathbf{E}\big[\Delta_{b'}^2(0)\big] &= \mathbf{E}\left[\left(N^{M(\lambda_s)/M(\mu_s)/K-1}(T)\right)^2\right] \\
&\leq \mathbf{E}\left[\left(N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)\right)^2\right] + O\left(\frac{1}{\epsilon}\right) \\
&= 2\left(\frac{\lambda_s - (K-1)\mu_s}{\beta}\right)^2 + \Theta\left(\frac{1}{\beta}\right) + O\left(\frac{1}{\epsilon}\right) \\
&= 2\left(\frac{K\left(\rho - \frac{K-1}{K}\right)}{3(C^2-1)\epsilon}\right)^2 + \Theta\left(\frac{1}{\epsilon}\right)
\end{aligned}$$

■

Lemma 14 *The asymptotics for the first and second moments of $\Delta_{b''}$ are given by:*

$$\begin{aligned}
\mathbf{E}[\Delta_{b''}] &= O(1) \\
\mathbf{E}\big[\Delta_{b''}^2\big] &= \Theta\left(\frac{1}{\epsilon}\right)
\end{aligned}$$

Proof: See Appendix A. ■

5.5.3 Obtaining $\mathbf{E}\big[\bar{N}_{s,g}^*\big]$

We will use the following lemma to obtain $\mathbf{E}\big[\bar{N}_{s,g}^*\big]$.

Lemma 15 *Consider an $M/M/K$ system with arrival rate λ and mean job size μ^{-1}. We interrupt this $M/M/K$ system according to a Poisson process with rate α, and at every interruption, a random number of jobs are added to the system. The number of jobs injected are i.i.d. random variables which are equal in distribution to some non-negative random variable Δ. Let $N^{(Int)}$ denote the number of jobs in this $M/M/K$ system. If $\mathbf{E}[\Delta] = o\left(\frac{1}{\alpha}\right)$, we have,*

$$\mathbf{E}\Big[N^{(Int)}\Big] = \mathbf{E}\Big[N^{M/M/K}\Big] + \frac{\frac{\alpha}{2}\mathbf{E}\big[\Delta^2\big]}{K\mu - \lambda} + o(1).$$

Proof in Appendix A.

To use the above lemma, we will consider an $M/M/K$ with arrival rate $\lambda p^{(\epsilon)}$, mean job size $\frac{1}{\mu_1^{(\epsilon)}}$, $\alpha = \lambda(1-p^{(\epsilon)})$ and $\Delta \stackrel{d}{=} \Delta_b$. Using the expression for $\mathbf{E}[\Delta_b]$ derived in Section 5.5.2, one can check that the condition of Lemma 15 is met. Therefore,

$$\mathbf{E}\big[\bar{N}_{s,g}^*\big] = \mathbf{E}\Big[N^{M/M/K}\Big] + \frac{1}{2}\frac{\lambda(1-p)\mathbf{E}\big[\Delta_b^2\big]}{K\mu - \lambda} + o(1) \tag{13}$$

Substituting $\mathbf{E}\big[\Delta_b^2\big]$ from Section 5.5.2 and using Lemma 5,
Case: $\rho < \frac{K-1}{K}$

$$\begin{aligned}
\mathbf{E}\big[\bar{N}_{s,g}^*\big] &= \mathbf{E}\Big[N^{M/M/K}\Big] + \frac{1}{2}\frac{\lambda\left(\frac{9}{2}(C^2-1)^3\epsilon^2\right)\Theta\left(\frac{1}{\epsilon}\right)}{K\mu - \lambda} + o(1) \\
&= \mathbf{E}\Big[N^{M/M/K}\Big] + o(1)
\end{aligned}$$

Case: $\rho > \frac{K-1}{K}$

$$\begin{aligned}
&\mathbf{E}\left[\bar{N}^*_{s,g}\right] \\
=&\mathbf{E}\left[N^{M/M/K}\right] + \frac{1}{2}\frac{\lambda\left(\frac{9}{2}(C^2-1)^3\epsilon^2\right)\left(\frac{2}{9}\frac{K^2\left(\rho-\frac{K-1}{K}\right)^2}{(C^2-1)^2\epsilon^2}\right)}{K\mu-\lambda} + o(1) \\
=&\mathbf{E}\left[N^{M/M/K}\right] + \frac{K^2\rho}{1-\rho}\left[\rho-\frac{K-1}{K}\right]^2\frac{C^2-1}{2} + o(1)
\end{aligned}$$

5.5.4 Putting it together: Upper bound on $\mathbf{E}[N_s]$

Recall the expression for upper bound on $\mathbf{E}[N_s]$ from equation (9):

$$\mathbf{E}[N_s] \leq \mathbf{E}\left[\bar{N}^*_{s,g}\right](1-\mathbf{Pr}[b]) + \mathbf{E}\left[\bar{N}^*_{s,g} + \Delta_{b'}(0) + \mathbf{I}_{b''|b}A_\lambda\left(T_{b''e}\right)\right]\mathbf{Pr}[b]$$

Substituting the expressions for $\mathbf{E}\left[\bar{N}^*_{s,g}\right]$ from Section 5.5.3, $\mathbf{E}[\Delta_{b'}(0)]$ from Lemma 12, $\mathbf{Pr}[b]$ from Equation (10) and $\frac{\mathbf{Pr}[b'']}{\mathbf{Pr}[b]}\mathbf{E}[A_\lambda\left(T_{b''e}\right)]$ from Equation (11) into the above equation, we get:

Case: $\rho < \frac{K-1}{K}$

$$\mathbf{E}[N_s] \leq \mathbf{E}\left[N^{M/M/K}\right] + o(1)$$

Case: $\rho > \frac{K-1}{K}$

$$\begin{aligned}
\mathbf{E}[N_s] \leq& \left(\mathbf{E}\left[N^{M/M/K}\right] + \frac{K^2\rho}{1-\rho}\left[\rho-\frac{K-1}{K}\right]^2\frac{C^2-1}{2}\right) \\
&+ \left(\frac{K\left(\rho-\frac{K-1}{K}\right)}{3(C^2-1)\epsilon} + \Theta(1)\right)\left(\frac{3K\rho(C^2-1)^2\epsilon}{2}\right) + o(1) \\
=& \mathbf{E}\left[N^{M/M/K}\right] + \frac{K^2\rho}{1-\rho}\left[\rho-\frac{K-1}{K}\right]^2\frac{C^2-1}{2} + K^2\rho\left[\rho-\frac{K-1}{K}\right]\frac{C^2-1}{2} + o(1) \\
=& \mathbf{E}\left[N^{M/M/K}\right] + \frac{K\rho}{1-\rho}\left[\rho-\frac{K-1}{K}\right]\frac{C^2-1}{2} + o(1)
\end{aligned}$$

Case: $\rho = \frac{K-1}{K}$

The critical case $\rho = \frac{K-1}{K}$ is difficult to handle directly. However, we can infer the limit

$$\lim_{\epsilon\to 0}\mathbf{E}\left[N^{M/H_2^{(\epsilon)}/K}\right] = \mathbf{E}\left[N^{M/M/K}\right]$$

from the preceding analysis to obtain upper bounds for the cases $\rho < \frac{K-1}{K}$ and $\rho > \frac{K-1}{K}$, and the matching lower bounds obtained via analysis of the system described in Section 5.6 as follows. For each ϵ, let $f^{(\epsilon)} : [0,1) \to \Re_0^+$ denote the function mapping the load ρ to the mean number of jobs in an $M/H_2^{(\epsilon)}/K$ system, $\mathbf{E}\left[N^{M/H^{(\epsilon)}/K}\right]$. Let $f(\cdot)$ be the point-wise limit of $f^{(\epsilon)}(\cdot)$ as $\epsilon \to 0$. Since each $f^{(\epsilon)}$ is a monotonic function, f is also monotonic. Further,

$$\lim_{\rho\uparrow\frac{K-1}{K}} f(\rho) = \lim_{\rho\downarrow\frac{K-1}{K}} f(\rho) = \mathbf{E}\left[N^{M/M/K}\right].$$

Thus we conclude,

$$f\left(\frac{K-1}{K}\right) = \lim_{\epsilon\to 0}\mathbf{E}\left[N^{M/H_2^{(\epsilon)}/K}\right]\Big|_{\rho=\frac{K-1}{K}} = \mathbf{E}\left[N^{M/M/K}\right].$$

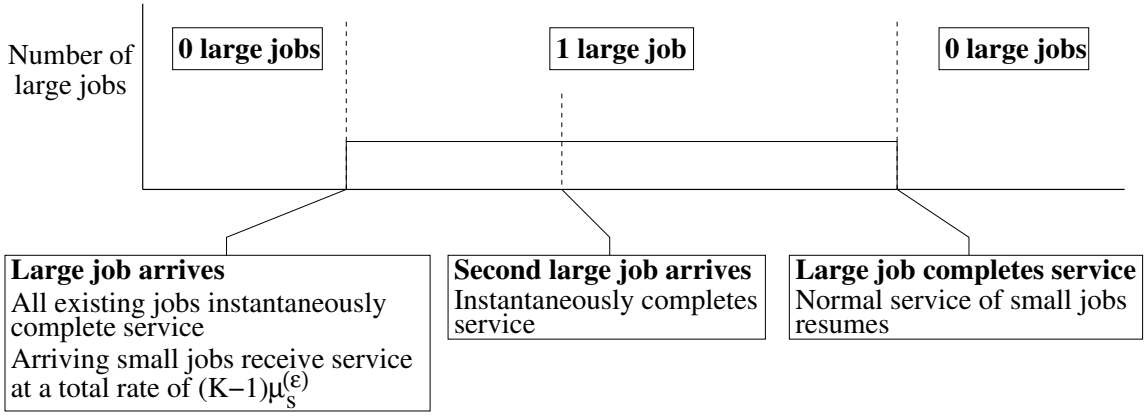


Figure 5: Construction of system $L^{(\epsilon)}$ which lower bounds the number of jobs in an $M/H_2^{(\epsilon)}/K$

5.6 Construction of $L^{(\epsilon)}$: the lower bounding system

Case: $\rho > \frac{K-1}{K}$

Figure 5 shows the behavior of system $L^{(\epsilon)}$ for this case. As before, denote the periods where there are no large jobs in the system as *good* periods, and periods when there is at least 1 large job as *bad* periods. During a good period, the small jobs receive service according to a normal K server FIFO system. As soon as a large job arrives to begin the bad period, all the small jobs currently in the system instantaneously complete service. That is, the system restarts with 1 large job. Any large jobs that arrive during this bad period complete service instantaneously. Further, whenever there are fewer than $(K-1)$ small jobs in the system during a bad period, they are collectively served at a total rate of $(K-1)\mu_s^{(\epsilon)}$.

Case: $\rho \leq \frac{K-1}{K}$

For this case we can consider an alternate lower bounding system which simplifies the analysis. In the lower bounding system, $L^{(\epsilon)}$, all large jobs instantaneously complete service on arrival. Thus the number of large jobs is always 0 and the number of small jobs behaves as in an $M/M/K$ with arrival rate $\lambda p^{(\epsilon)}$ and mean job size $\frac{1}{\mu_s^{(\epsilon)}}$.

Lemma 16 *The number of small jobs in an $M/H_2^{(\epsilon)}/K$ system, $N_s^{M/H_2^{(\epsilon)}/K}$, is stochastically lower bounded by the number of small jobs in the corresponding system $L^{(\epsilon)}$, $N_s^{L^{(\epsilon)}}$.*

Proof: Straightforward using stochastic coupling. ■

Sketch of Analysis of $L^{(\epsilon)}$

Case: $\rho > \frac{K-1}{K}$

The analysis of system $L^{(\epsilon)}$ is simplified because the large jobs form an $M/M/1/1$ system independent of the small jobs. The length of a bad period is distributed as $\text{Exp}\left(\mu_\ell^{(\epsilon)}\right)$ and the length of a good period is distributed as $\text{Exp}\left(\lambda(1-p^{(\epsilon)})\right)$. Further, during a bad period, the number of small jobs behaves as in an $M/M/1$ queue with arrival rate $\lambda p^{(\epsilon)}$ and service rate $(K-1)\mu_s^{(\epsilon)}$ starting with an empty system. Therefore, the distribution of the number of small jobs at the end of bad periods (and hence, by conditional PASTA, the distribution of the time average number of small jobs during the bad periods) in system $L^{(\epsilon)}$ can be derived along the lines of proof of Lemma 12. To complete the proof we need to find the stationary mean number of small jobs at the end of

good periods (and hence, by conditional PASTA, the stationary mean number of small jobs during the good periods). This is equivalent to finding the mean number of jobs in an $M/M/K$ at time $T \sim \text{Exp}\,(\lambda(1-p))$, starting at $t = 0$ with number of jobs sampled from the distribution of the number of small jobs at the end of bad periods. To do this, we start with Eqn. (43), proceed as in the proof of Lemma 13 by finding the root of the denominator in the interval $[0, 1)$ and equating the numerator to zero at this root. We then follow the proof of Lemma 15 to obtain the mean number of jobs at time T.

Case: $\rho \leq \frac{K-1}{K}$

As stated earlier, in constructing the lower bound system $L^{(\epsilon)}$, we assume that the large jobs complete service instantaneously on arrival. Therefore, the number of large jobs in the system is 0 with probability 1. The distribution of the time average number of small jobs in the system is given by the stationary distribution in an $M/M/K$ FCFS system.

6 Effect of higher moments

In Theorems 1 and 2, we proved that the first two moments of the job size distribution alone are insufficient to approximate the mean waiting time accurately. In Section 3, by means of numerical experiments, we observed that within the H_2 class of distributions, the normalized third moment of the job size distribution has a significant impact on the mean waiting time. Further, we observed that for H_2 job size distributions, increasing the normalized third moment causes the mean waiting time to drop. It is, therefore, only natural to ask the following questions: Are three moments of the job size distribution sufficient to accurately approximate the mean waiting time, or do even higher moments have an equally significant impact? Is the qualitative effect of 4th and higher moments similar to the effect of the 3rd moment or is it the opposite? In this section, we touch upon these interesting and largely open questions.

	$C^2 = 19$		$C^2 = 99$	
	$\mathbf{E}[W]$	θ_3	$\mathbf{E}[W]$	θ_3
2-moment approx. (Eqn. 1)	6.6873	-	33.4366	-
Weibull	6.0691	4.2	25.9896	8.18
Truncated Pareto ($\alpha = 1.1$)	5.5241	4.24	24.5788	6.30
Lognormal	4.9937	20	19.5548	100
Truncated Pareto ($\alpha = 1.3$)	4.8770	7.59	18.8933	16.85
Truncated Pareto ($\alpha = 1.5$)	3.9504	20	10.5404	100

Table 2: Results from simulating an $M/G/K$ with $K = 10$ and $\rho = 0.9$ (confidence intervals omitted). All job size distributions have $\mathbf{E}[X] = 1$.

We first revisit the simulation results of Table 1. Table 2 shows the simulation results of Table 1 again, but with an additional column – the normalized third moment of the job size distribution. We have omitted the confidence intervals in Table 2. Observe that the lognormal distribution and the Pareto distribution with $\alpha = 1.5$ have *identical first three moments*, yet exhibit very different mean waiting times. This behavior is compounded when the system load is reduced to $\rho = 0.6$ (Table 3). As we saw in Section 3, the disagreement in the mean waiting time for the lognormal and the truncated Pareto distribution can be partly explained by the very different looking $\rho(x)$ curves for these distributions, shown in Figure 6. The bulk of the load in the lognormal distribution is comprised of larger jobs as compared to the truncated Pareto distribution.

The example of lognormal and Pareto ($\alpha = 1.5$) distributions suggests that even knowledge of

	$C^2 = 19$		$C^2 = 99$	
	$\mathbf{E}[W]$	θ_3	$\mathbf{E}[W]$	θ_3
2-moment approx. (Eqn. 1)	0.2532	-	1.2662	-
Weibull	0.1374	4.2	0.4638	8.18
Truncated Pareto ($\alpha = 1.1$)	0.0815	4.24	0.2057	6.30
Lognormal	0.0854	20	0.2154	100
Truncated Pareto ($\alpha = 1.3$)	0.0538	7.59	0.0816	16.85
Truncated Pareto ($\alpha = 1.5$)	0.0355	20	0.0377	100

Table 3: Results from simulating an $M/G/K$ with $K = 10$ and $\rho = 0.6$ (confidence intervals omitted). All job size distributions have $\mathbf{E}[X] = 1$.

three moments of the job size distribution may not be sufficient for accurately approximating the mean waiting time. *So what is the effect of higher moments on the mean waiting time?* To begin answering this question, we will follow a similar approach as in Section 3 where we looked at the H_2 job size distribution. However, we first need to expand the class of job size distributions to allow us control over the 4th moment. For this purpose, we choose the *3-phase degenerate hyperexponential* class of distribution, denoted by H_3^*. Analogous to the H_2^* distribution, H_3^* is the class of mixtures of three exponential distributions where the mean of one of the phases is 0 (see Definition 2). Compared to the H_2 class, the H_3^* class has one more parameter and thus four degrees of freedom, which allows us control over the 4th moment while holding the first three moments fixed.

We now extend the numerical results of Figure 1 by considering job size distributions in the H_3^* class with the same mean and SCV as the example illustrated in Figure 1. However, to demonstrate the effect of the 4th moment, we choose two values of θ_3 and plot the $\mathbf{E}[W]$ curves as a function of the 4th moment in Figure 7. As a frame of reference, we also show the mean waiting time under the H_2 job size distribution (with the same first three moments as H_3^*) and that under H_2^* distribution (with the same first two moments as H_3^*).

As is evident from Figure 7, the fourth moment can have as significant an impact on the mean waiting time as the third moment. As the 4th moment is increased, the mean waiting time increases from $\mathbf{E}\left[W^{M/H_2/K}\right]$ to $\mathbf{E}\left[W^{M/H_2^*/K}\right]$. Therefore, the qualitative effect of the 4th moment is *opposite* to that of the third moment.

The effect of the fourth moment also helps explain the disagreement between the mean waiting time for the lognormal, the truncated Pareto ($\alpha = 1.5$) and the H_2 distributions. For the case $C^2 = 19$, the lognormal distribution has a much higher 4th moment ($\mathbf{E}\left[X^4\right] = 64 \times 10^6$) than the Pareto ($\mathbf{E}\left[X^4\right] = 5.66 \times 10^6$) and the H_2 ($\mathbf{E}\left[X^4\right] = 4.67 \times 10^6$) distribution with $\theta_3 = 20$. While this is a possible cause for a higher mean waiting time under the lognormal distribution, there is still disagreement between the mean waiting time under the lognormal distribution and the H_3^* distribution (see Figure 7) with the same first 4 moments, indicating that even higher moments are playing an important role as well!

In conclusion, by looking at a range of distributions including hyperexponential, Pareto and lognormal distributions, we see that the moments of the job size distribution may not be sufficient to accurately predict the mean waiting time. Further, for distributions such as the lognormal distribution which are not uniquely determined by their moments, no finite number of moments may suffice. Other characteristics, such as the distribution of load among the small and large job sizes, may lead to more accurate approximations. (We make stronger conjectures on the exact effect of the higher moments in [15].)

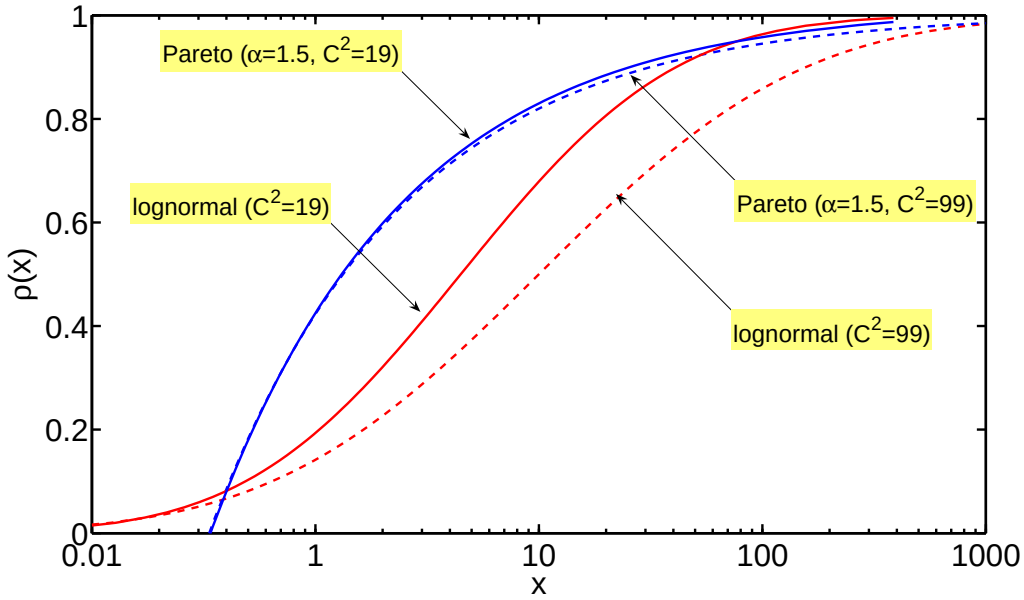


Figure 6: The distribution of load as a function of job size for the lognormal and bounded Pareto ($\alpha = 1.5$) distributions for two values of squared coefficient of variation. Although the lognormal and Pareto distributions have identical first three moments, the distribution of load among different job sizes is drastically different.

7 On tight bounds for $\mathbf{E}\left[W^{M/G/K}\right]$

In Theorem 1 we proved a lower bound on $W_h^{C^2}$ and an upper bound on $W_l^{C^2}$, respectively, by considering two-point job size distributions. In Theorem 2 we proved bounds on $W_h^{C^2}$ and $W_l^{C^2}$ by considering job size distributions which are mixtures of two exponential random variables. However, all known tight bounds for $GI/GI/1$ involving the first two moments of the job size distribution are obtained by considering two-point distributions. Thus, we conjecture that the bounds in Theorem 1 are tight, whereas the bounds in Theorem 2 can be tightened as described in the conjecture below:

Conjecture 1 *For any finite C^2,*

$$W_h^{C^2} = \left(C^2 + 1\right) \mathbf{E}\left[W^{M/D/K}\right] \quad \textit{for all } \rho < 1$$

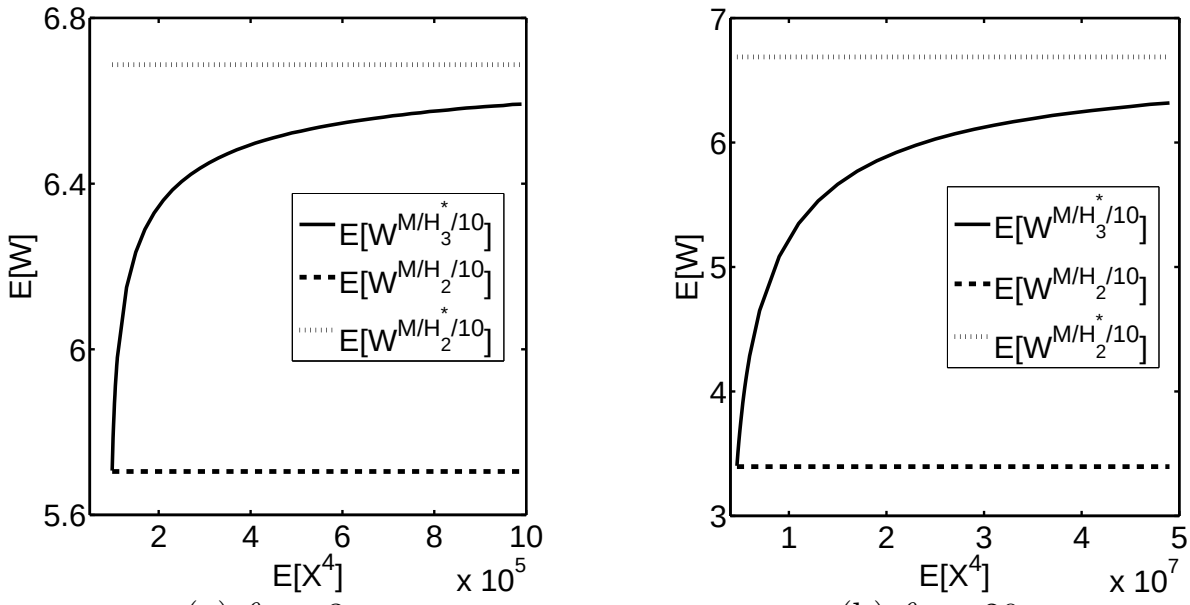
and,

$$W_l^{C^2} = \begin{cases} \mathbf{E}\left[W^{M/D/K}\right] & \textit{if } \rho < \frac{K-1}{K} \\ \mathbf{E}\left[W^{M/D/K}\right] + \frac{1}{1-\rho}\left[\rho - \frac{K-1}{K}\right]\frac{C^2}{2} & \textit{if } \frac{K-1}{K} \leq \rho < 1 \end{cases}$$

where $\mathbf{E}\left[W^{M/D/K}\right]$ is the mean waiting time when all the jobs have a constant size 1.

8 Conclusions

In this paper, we addressed the classical problem of approximating the mean waiting time of an $M/G/K$ queueing system. While there is a huge body of work on developing closed-form approximations for the mean waiting time, all such approximations are based only on the first two



(a) $\theta_3 = 3$ (b) $\theta_3 = 20$

Figure 7: Illustration of the effect of 4th moment of the service distribution on mean waiting time of an $M/H_3^*/10$ system for two values of the normalized third moment. Dashed line shows the mean waiting time under an H_2 service distribution with the same first three moments and the light dotted line shows the mean waiting time under an H_2^* service distribution with the same first two moments as the H_3^* distribution. The mean and squared coefficient of variation of the job size distribution were held constant at $\mathbf{E}[X] = 1$ and $C^2 = 19$ with load $\rho = 0.9$ (same as Figure 1).

moments of the job size distribution. In this work, we proved that it is impossible to develop any approximation, based on only the first two moments, that is accurate for all job size distributions. Specifically, we proved that specifying the first two moments of the job size distribution insufficiently limits the range of possible values of mean waiting time: The maximum value of this range can be as much as $(C^2 + 1)$ times the minimum value.

Further, we suggest that *moments* are not the ideal job size characteristic on which to base approximations for mean waiting time[5]. The moment sequence *can* be useful if one of the moments (appropriately normalized) is small. As an example, if the job size distribution has a small normalized third moment, then an approximation based on only the first two moments is likely to be accurate. However, there are also many job size distributions like the lognormal distribution (whose moments are all high), for which moments are not useful in accurately predicting mean waiting time. Other characteristics, such as the distribution of load among different job sizes, may be more representative for the purpose of approximating mean waiting time.

9 Acknowledgements

We would like to thank the referees for numerous helpful comments on improving the exposition of the paper.

Varun Gupta and Mor Harchol-Balter were supported by NSF grant CNS-0719106 (SGER: Collaborative Research: CSR-SMA) and by the Microsoft Breakthrough Research Grant 2007. Jim Dai

[5]We refer the reader to [15] for stronger conjectures on inapproximability gap and bounds given information about higher moments.

and Bert Zwart were supported in part by NSF grants CMMI-0727400 and CNS-0718701.

References

[1] I. J. B. F. Adan and J. Resing. *Queueing theory.* Eindhoven University of Technology, 2002.

[2] Paul Barford and Mark Crovella. Generating representative web workloads for network and server performance evaluation. *Proceeding of ACM SIGMETRICS/Performance'98*, pages 151–160, 1998.

[3] Dimitris Bertsimas and Karthik Natarajan. A semidefinite optimization approach to the steady-state analysis of queueing systems. *Queueing Syst.*, 56(1):27–39, 2007.

[4] A.A. Borovkov. *Stochastic Processes in Queueing Theory.* Nauka, Moscow, 1972.

[5] D.Y. Burman and D.R. Smith. A light-traffic theorem for multi-server queues. *Math. Oper. Res.*, 8:15–25, 1983.

[6] G.P. Cosmetatos. Some approximate equilibrium results for the multiserver queue $(M/G/r)$. *Operational Research Quarterly*, 27:615–620, 1976.

[7] Daryl J. Daley. Some results for the mean waiting-time and workload in $GI/GI/k$ queues. In Jewgeni H. Dshalalow, editor, *Frontiers in queueing: models and applications in science and engineering*, pages 35–59. CRC Press, Inc., Boca Raton, FL, USA, 1997.

[8] D.J. Daley and T. Rolski. Some comparibility results for waiting times in single- and many-server queues. *J. Appl. Prob.*, 21:887–900, 1984.

[9] Jos H. A. de Smit. A numerical solution for the multiserver queue with hyper-exponential service times. *Oper. Res. Lett.*, 2(5):217–224, 1983.

[10] Jos H. A. de Smit. The queue $GI/M/s$ with customers of different types or the queue $GI/H_m/s$. *Adv. in Appl. Probab.*, 15(2):392–419, 1983.

[11] Jos H. A. de Smit. The queue $GI/H_m/s$ in continuous time. *J. Appl. Probab.*, 22(1):214–222, 1985.

[12] Allen Downy and Mor Harchol-Balter. Exploiting process lifetime distributions for dynamic load balancing. *ACM Transactions on Computer Systems*, 15(3):253–285, August 1997.

[13] Serguei Foss and Dmitry Korshunov. Heavy tails in multi-server queue. *Queueing Syst.*, 52(1):31–48, 2006.

[14] Noah Gans, Ger Koole, and Avi Mandelbaum. Telephone call centers: tutorial, review, and research prospects. *Manufacturing and Service operations Management*, 5:79–141, 2003.

[15] Varun Gupta, J.G. Dai, Mor Harchol-Balter, and Bert Zwart. The effect of higher moments of job size distribution on the performance of an $M/G/K$ queueing system. Technical Report CMU-CS-08-106, School of Computer Science, Carnegie Mellon University, 2008.

[16] Varun Gupta, Mor Harchol-Balter, Alan Scheller-Wolf, and Uri Yechiali. Fundamental characteristics of queues with fluctuating load. In *Proceedings of ACM SIGMETRICS*, pages 203–215, 2006.

[17] Mor Harchol-Balter and Bianca Schroeder. Evaluation of task assignment policies for supercomputing servers. In *Proceedings of 9th IEEE Symposium on High Performance Distributed Computing (HPDC '00)*, 2001.

[18] M.H. van Hoorn H.C. Tijms and A. Federgruen. Approximations for the steady-state probabilities in the $M/G/c$ queue. *Adv. Appl. Prob.*, 13:186–206, 1981.

[19] Per Hokstad. Approximations for the $M/G/m$ queue. *Operations Research*, 26(3):510–523, 1978.

[20] Per Hokstad. The steady state solution of the $M/K_2/m$ queue. *Adv. Appl. Prob.*, 12(3):799–823, 1980.

[21] T. Kimura. Diffusion approximation for an $M/G/m$ queue. *Operations Research*, 31:304–321, 1983.

[22] T. Kimura. Approximations for multi-server queues: system interpolations. *Queueing Systems*, 17(3-4):347–382, 1994.

[23] J.F.C. Kingman. Inequalities in the theory of queues. *J. R. Statist. Soc.*, 32(1):102–110, 1970.

[24] Leonard Kleinrock. *Queueing Systems, Volume I: Theory.* Wiley-Interscience, 1975.

[25] Julian Köllerström. Heavy traffic theory for queues with several servers. I. *J. Appl. Prob.*, 11:544–552, 1974.

[26] A.M. Lee and P.A. Longton. Queueing process associated with airline passenger check-in. *Operations Research Quarterly*, 10:56–71, 1959.

[27] Hau Leung Lee and Morris A. Cohen. A note on the convexity of performance measures of $M/M/c$ queueing systems. *J. Appl. Probab.*, 20(4):920–923, 1983.

[28] B.N.W. Ma and J.W. Mark. Approximation of the mean queue length of an $M/G/c$ queueing system. *Operations Research*, 43(1):158–165, 1995.

[29] Masakiyo Miyazawa. Approximation of the queue-length distribution of an $M/GI/s$ queue by the basic equations. *J. Appl. Prob.*, 23:443–458, 1986.

[30] Alfred Müller and Dietrich Stoyan. *Comparison methods for stochastic models and risks.* Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2002.

[31] S.A. Nozaki and S.M. Ross. Approximations in finite-capacity multi-server queues with Poisson arrivals. *J. Appl. Prob.*, 15(4):826–834, 1978.

[32] J. Cohen O. Boxma and N. Huffels. Approximations in the mean waiting time in an $M/G/s$ queueing system. *Operations Research*, 27:1115–1127, 1979.

[33] Sheldon M. Ross. *Stochastic Processes, 2nd Edition.* Wiley, 1996.

[34] Alan Scheller-Wolf and Karl Sigman. New bounds for expected delay in FIFO $GI/GI/c$ queues. *Queueing Systems*, 26(1-2):169–186, 1997.

[35] Alan Scheller-Wolf and Rein Vesilo. Structural interpretation and derivation of necessary and sufficient conditions for delay moments in FIFO multiserver queues. *Queueing Syst.*, 54(3):221–232, 2006.

[36] D. Stoyan. Approximations for $M/G/s$ queues. *Math. Operationsforsch. Statist. Ser. Optimization*, 7:587–594, 1976.

[37] Dietrich Stoyan. A continuity theorem for queue size. *Bull. Acad. Sci. Pollon.*, 21:1143–1146, 1973.

[38] Dietrich Stoyan. *Comparison methods for queues and other stochastic models.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1983. Translation from the German edited by Daryl J. Daley.

[39] Hideaki Takagi. *Queueing Analysis, Vol. 1: Vacation and Priority Systems.* North-Holland, 1991.

[40] Y. Takahashi. An approximation formula for the mean waiting time of an $M/G/c$ queue. *J. Opns. Res. Soc. Japan*, 20:147–157, 1977.

[41] Hermann Thorisson. The queue $GI/GI/k$: finite moments of the cycle variables and uniform rates of convergence. *Comm. Statist. Stochastic Models*, 1(2):221–238, 1985.

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

[43] W. Whitt. A diffusion approximation for the G/GI/n/m queue. *Operations Research*, 52:922–941, 2004.

[44] Ward Whitt. The effect of variability in the $GI/G/s$ queue. *J. Appl. Prob.*, 17:1062–1071, 1980.

[45] Ward Whitt. Comparison conjectures about the $M/G/s$ queue. *OR Letters*, 2(5):203–209, 1983.

[46] Ward Whitt. On approximations for queues, I: Extremal distributions. *AT&T Bell Laboratories Technical Journal*, 63:115–138, 1984.

[47] Ward Whitt. Approximations for the $GI/G/m$ queue. *Production and Operations Management*, 2(2):114–161, 1993.

[48] D.D. Yao. Refining the diffusion approximation for the $M/G/m$ queue. *Operations Research*, 33:1266–1277, 1985.

A Proofs

Proof of Claim 1: The proof will proceed in two steps. We first show that the H_2^* distribution lying in $\{H_2|C^2\}$ has the smallest third moment in $\{H_2|C^2\}$ for all $C^2 > 1$. Then we will give a method, which given any n-phase hyperexponential distribution for $n > 2$, allows one to create an $(n-1)$-phase hyperexponential distribution with the same first two moments but a smaller third moment. Using this method one can, in the end, obtain an H_2 distribution with a smaller third moment and combine it with first step of the proof to prove the claim.

Step 1: Let X be a random variable distributed according to the following H_2 distribution:

$$X \sim \begin{cases} \text{Exp}\,(\mu_1) & \text{w.p. } p \\ \text{Exp}\,(\mu_2) & \text{w.p. } 1-p \end{cases}$$

The expressions for first 3 moments of X are given by:

$$\mathbf{E}[X] = \frac{p}{\mu_1} + \frac{1-p}{\mu_2} \tag{14a}$$

$$\mathbf{E}\left[X^2\right] = 2\frac{p}{\mu_1^2} + 2\frac{1-p}{\mu_2^2} \tag{14b}$$

$$\mathbf{E}\left[X^3\right] = 6\frac{p}{\mu_1^3} + 6\frac{1-p}{\mu_2^3} \tag{14c}$$

Performing $\frac{(14c)}{6} \times \frac{(14a)}{1} - \frac{(14b)}{2} \times \frac{(14b)}{2}$, we obtain the following relation between the moments of X and the parameters of the distribution:

$$\frac{\mathbf{E}\left[X^3\right]\mathbf{E}[X]}{6} - \frac{\mathbf{E}\left[X^2\right]^2}{4} = \frac{p(1-p)}{\mu_1\mu_2}\left[\frac{1}{\mu_1} - \frac{1}{\mu_2}\right]^2$$

It is easy to see that since the right hand side is non-negative, $\frac{3\mathbf{E}\left[X^2\right]^2}{2\mathbf{E}[X]}$ is a lower bound on the smallest possible value of $\mathbf{E}\left[X^3\right]$ given the first two moments, and this lower bound is realized by letting $\mu_1 \to \infty$ (or $\mu_2 \to \infty$), that is, by the degenerate hyperexponential distribution.

Step 2: If the H_n distribution has a phase with mean 0, then pick any two phases with non-zero mean. Replace these two phases with the H_2^* distribution with the same first two moments as those of the conditional distribution, conditioned on being in these two phases. Merge the phases with 0 mean. Using step 1 above, this replacement necessarily creates an $(n-1)$-phase hyperexponential distribution with smaller third moment while preserving the first two. If the H_n distribution has no phase with mean 0, perform the above step twice to reduce the number of phases by 1. ■

Proof of Lemma 5: Suppressing the superscript, we have the following equations from Definition 3:

$$\frac{p}{\mu_s} + \frac{1-p}{\mu_\ell} = 1 \tag{15a}$$

$$\frac{p}{\mu_s^2} + \frac{1-p}{\mu_\ell^2} = \frac{C^2+1}{2} \tag{15b}$$

$$\frac{p}{\mu_s^3} + \frac{1-p}{\mu_\ell^3} = \frac{1}{6\epsilon} \tag{15c}$$

Performing (15b) − (15a) × (15a):

$$p(1-p)\left(\frac{1}{\mu_s} - \frac{1}{\mu_\ell}\right)^2 = \frac{C^2-1}{2} \tag{16}$$

Performing (15c) × (15a) − (15b) × (15b):

$$\frac{p(1-p)}{\mu_s\mu_\ell}\left(\frac{1}{\mu_s} - \frac{1}{\mu_\ell}\right)^2 = \frac{1}{6\epsilon} - \frac{(C^2+1)^2}{4} \tag{17}$$

The above two equations give:

$$\mu_s\mu_l = \frac{\frac{C^2-1}{2}}{\frac{1}{6\epsilon} - \frac{(C^2+1)^2}{4}} \tag{18}$$

From equations (15a) and (15b),

$$p(\mu_\ell - \mu_s) = \mu_s\mu_\ell - \mu_s$$
$$p(\mu_\ell^2 - \mu_s^2) + \mu_s^2 = \frac{C^2+1}{2}(\mu_s\mu_\ell)^2$$

Substituting $p(\mu_\ell - \mu_s)$ as $\mu_s\mu_\ell - \mu_s$ in the second equation gives:

$$\begin{aligned}\mu_s + \mu_\ell &= 1 + \frac{C^2+1}{2}\mu_s\mu_\ell \\ &= 1 + \frac{\frac{C^2+1}{2}\cdot\frac{C^2-1}{2}}{\frac{1}{6\epsilon} - \frac{(C^2+1)^2}{4}}\end{aligned}$$

Finally,

$$\begin{aligned}\mu_s\mu_\ell &= \frac{\frac{C^2-1}{2}}{\frac{1}{6\epsilon} - \frac{(C^2+1)^2}{4}} = 3(C^2-1)\epsilon\left(1 - \frac{3(C^2+1)^2}{2}\epsilon\right)^{-1} \\ &= 3(C^2-1)\epsilon\left[1 + \frac{3}{2}(C^2+1)^2\epsilon + \frac{9}{4}(C^2+1)^4\epsilon^2 + \Theta(\epsilon^3)\right] \\ \mu_s + \mu_\ell &= 1 + \frac{\frac{C^2+1}{2}\cdot\frac{C^2-1}{2}}{\frac{1}{6\epsilon} - \frac{(C^2+1)^2}{4}} = 1 + \frac{3}{2}(C^2+1)(C^2-1)\epsilon\left(1 - \frac{3(C^2+1)^2}{2}\epsilon\right)^{-1} \\ &= 1 + \frac{3}{2}(C^2+1)(C^2-1)\epsilon\left[1 + \frac{3}{2}(C^2+1)^2\epsilon + \frac{9}{4}(C^2+1)^4\epsilon^2 + \Theta(\epsilon^3)\right]\end{aligned}$$

It is straightforward to verify that the expressions for μ_s and μ_ℓ in Lemma 5 satisfy the above equations. The expression for p then follows from $p = 1 - \mu_\ell\frac{\mu_s - 1}{\mu_s - \mu_l}$. ■

Proof of Lemma 7: Recall that $\overline{N_\ell}^{(\epsilon)}$ is defined to be the steady-state number of customers in an $M\left(\lambda(1-p^{(\epsilon)})\right)/M\left(\mu_\ell^{(\epsilon)}\right)/1$ queue with service interruptions where the server is interrupted for the duration of the busy period of an $M(\lambda)/M(1)/K$ queue. The busy period of an $M(\lambda)/M(1)/K$ queue has finite second moment [41], and hence the second moment of the service interruptions is also finite. Let $B_{\lambda,1,K}$ be the busy period of this queue. Define $\rho_\ell^{(\epsilon)} = \lambda(1-p^{(\epsilon)})/\mu_\ell^{(\epsilon)}$.

Our aim is to prove:

$$\mathbf{E}\left[\overline{N_\ell}^{(\epsilon)}\right] = o(1)$$

The lemma follows by specializing results for the $M/G/1$ queue with server breakdowns to the special case considered here, see e.g. Adan & Resing [1, page 101]. For completeness, we provide a new proof of the $M/G/1$ queue with breakdowns by viewing it as a special case of an $M/G/1$ with setup times [39, page 130]. Let G be a so-called *generalized* service time, which is the service time of a large customer plus the total duration of service interruptions while that customer was in service. Let $\alpha = \lambda(1-p^{(\epsilon)})$ denote the arrival rate of the customers. The breakdowns (busy periods of the $M(\lambda)/M(1)/K$ queue) arrive at a rate λ when the system is "up", and let $\widetilde{B}_{\lambda,1,K}(s)$ denote the Laplace transform of the duration of these breakdowns. We can now view the $M(\alpha)/M(\mu_\ell^{(\epsilon)})/1$ queue with breakdowns as an $M/G/1$ queue with service distribution given by the generalized service time, G, and a setup time I at the beginning of each busy period, where the Laplace transform of I, $\widetilde{I}(s)$, satisfies:

$$\widetilde{I}(s) = \frac{\alpha}{\alpha+\lambda} + \frac{\lambda}{\alpha+\lambda}\cdot\widetilde{B}_{\lambda,1,K}(\alpha)\cdot\widetilde{I}(s) + \frac{\lambda}{\alpha+\lambda}\cdot\frac{\alpha}{\alpha-s}\left(\widetilde{B}_{\lambda,1,K}(s) - \widetilde{B}_{\lambda,1,K}(\alpha)\right) \qquad (19)$$

In the above equation, the first term denotes the event that the customer arrives before the breakdown, the second term denotes the event that the breakdown arrives before the customer, but no customers arrive during this breakdown, and the third term denotes the event that the breakdown arrives before the customer and a customer arrives during this breakdown. By differentiating (19) with respect to s once and twice, and evaluating at $s = 0$, the first two moments of I are obtained, respectively, as:

$$\mathbf{E}[I] = \left(\frac{\lambda}{\alpha+\lambda}\right) \cdot \frac{\mathbf{E}[B_{1,\lambda,K}] - \frac{1-\widetilde{B}_{1,\lambda,K}(\alpha)}{\alpha}}{1 - \frac{\lambda}{\alpha+\lambda} \cdot \widetilde{B}_{1,\lambda,K}(\alpha)} \tag{20}$$

$$\mathbf{E}\big[I^2\big] = \left(\frac{\lambda}{\alpha+\lambda}\right) \frac{\mathbf{E}\Big[B^2_{1,\lambda,K}\Big] - 2\frac{\mathbf{E}\ B_{1,\lambda,K}}{\alpha} + 2\frac{1-\widetilde{B}_{1,\lambda,K}(\alpha)}{\alpha^2}}{1 - \frac{\lambda}{\alpha+\lambda} \cdot \widetilde{B}_{1,\lambda,K}(\alpha)} \tag{21}$$

Define $\overline{V_\ell}^{(\epsilon)}$ to be the system time (response time) of large customers in the modified queue. From [39, page 130], we get

$$\begin{aligned}\mathbf{E}\Big[\overline{V_\ell}^{(\epsilon)}\Big] &= \mathbf{E}[G] + \left(\frac{\rho_G}{1-\rho_G}\right)\frac{\mathbf{E}\big[G^2\big]}{2\mathbf{E}[G]} + \frac{2\mathbf{E}[I] + \alpha\mathbf{E}\big[I^2\big]}{2(1+\alpha\mathbf{E}[I])} \\ &= \mathbf{E}[G] + \left(\frac{\rho_G}{1-\rho_G}\right)\frac{\mathbf{E}\big[G^2\big]}{2\mathbf{E}[G]} + \left(\frac{\lambda\mathbf{E}[B_{\lambda,1,K}]}{1+\lambda\mathbf{E}[B_{\lambda,1,K}]}\right)\frac{\mathbf{E}\Big[B^2_{\lambda,1,K}\Big]}{2\mathbf{E}[B_{\lambda,1,K}]}. \end{aligned} \tag{22}$$

Here $\rho_G = \rho_\ell^{(\epsilon)}(1 + \mathbf{E}[B_{\lambda,1,K}]/\lambda)$. The first two moments of G are given by

$$\mathbf{E}[G] = \frac{1}{\mu_\ell^{(\epsilon)}}\left(1 + \frac{\mathbf{E}[B_{\lambda,1,K}]}{\lambda}\right) \tag{23}$$

and that

$$\mathbf{E}\big[G^2\big] = \frac{2}{\left(\mu_\ell^{(\epsilon)}\right)^2}\left(1 + \frac{\mathbf{E}[B_{\lambda,1,K}]}{\lambda}\right)^2 + \frac{1}{\mu_\ell^{(\epsilon)}}\lambda\mathbf{E}\big[B^2_{\lambda,1,K}\big]. \tag{24}$$

From these equations, it follows that $\mathbf{E}[G] = \Theta(1/\epsilon)$ and $\mathbf{E}\big[G^2\big] = \Theta(1/\epsilon^2)$. This implies $\mathbf{E}\Big[\overline{V_\ell}^{(\epsilon)}\Big] = \Theta(1/\epsilon)$. By Little's law, $\mathbf{E}\Big[\overline{N_\ell}^{(\epsilon)}\Big] = \lambda(1-p^{(\epsilon)})\mathbf{E}\Big[\overline{V_\ell}^{(\epsilon)}\Big]$, which implies $\mathbf{E}\Big[\overline{N_\ell}^{(\epsilon)}\Big] = \Theta(\epsilon)$. ∎

Proof of Lemma 9: Consider a further modification of system $U^{(\epsilon)}$ where the small jobs are not served during the entire bad period. That is, even when there is only a single large job in the system, we stop serving small jobs. The fraction of time this modified system $U^{(\epsilon)}$ is busy with large jobs is given by $\lambda\frac{1-p^{(\epsilon)}}{\mu_\ell^{(\epsilon)}} = K\rho\frac{1-p^{(\epsilon)}}{\mu_\ell^{(\epsilon)}}$. The load of the small jobs is less than ρ. Thus, system $U^{(\epsilon)}$ will be stable if $\rho < 1 - K\rho\frac{1-p^{(\epsilon)}}{\mu_\ell^{(\epsilon)}}$.

Since $p^{(\epsilon)} \le 1$ and $\mu_s^{(\epsilon)} \ge 1$, we have

$$\frac{1-p^{(\epsilon)}}{\left(\mu_\ell^{(\epsilon)}\right)^2} \le \frac{C^2+1}{2}$$

$$\frac{1-p^{(\epsilon)}}{\left(\mu_\ell^{(\epsilon)}\right)^3} \ge \frac{1}{6\epsilon} - 1$$

Now,

$$\frac{1-p^{(\epsilon)}}{\mu_\ell^{(\epsilon)}} = \frac{\left(\frac{1-p^{(\epsilon)}}{\mu_\ell^{(\epsilon)\ 2}}\right)^2}{\frac{1-p^{(\epsilon)}}{\mu_\ell^{(\epsilon)\ 3}}} \leq \frac{\left(\frac{C^2+1}{2}\right)^2}{\frac{1}{6\epsilon}-1}$$

Thus,

$$\begin{aligned} \epsilon &< \frac{1}{6}\left[\frac{K\rho(C^2+1)^2}{4(1-\rho)}+1\right]^{-1} \\ \implies K\rho\frac{\left(\frac{C^2+1}{2}\right)^2}{\frac{1}{6\epsilon}-1} &< 1-\rho \\ \implies K\rho\frac{1-p^{(\epsilon)}}{\mu_\ell^{(\epsilon)}} &< 1-\rho \end{aligned}$$

■

Proof of Lemma 10: Recall that $\Phi(A)$ was defined as the mapping between non-negative random variables where $\Phi(A)$ gives the random variable for the number of jobs at the end of a good period given that the number at the beginning of the good period is A. Let $\Psi(A)$ be another mapping between random variables defined by:

$$\Psi(A) = \Delta_{b''} + \sum_{i=0}^{\infty}(i+\Delta_{b'}(i))\mathbf{I}_{\{A=i\}}$$

That is, $\Psi(A)$ gives the number of small jobs at the end of a bad period given that the number at the start is A. Further, the following facts can be easily verified via coupling:

1. $A_1 \leq_{st} A_2 \implies \Phi(A_1) \leq_{st} \Phi(A_2)$
2. $\Delta_{b'}(0) \geq_{st} \Delta_{b'}(1) \geq_{st} \ldots \Delta_{b'}(i) \geq \Delta_{b'}(i+1) \geq \ldots$

The last fact implies $\Psi(A) \leq_{st} A + \Delta_{b'}(0) + \Delta_{b''} \stackrel{def}{=} A + \Delta_b$. This gives us a way to stochastically upper bound $N^*_{s,g}$. We defined $\bar{N}^*_{s,g}$ to be the solution to the following fixed point equation:

$$\bar{N}^*_{s,g} \stackrel{d}{=} \Phi(\bar{N}^*_{s,g} + \Delta_b)$$

Also,

$$N^*_{s,g} \stackrel{d}{=} \Phi(\Psi(N^*_{s,g}))$$

Let $Y(0) = \bar{Y}(0) = 0$. Further, let $Y(n+1) = \Phi(\Psi(Y(n)))$ and $\bar{Y}(n+1) = \Phi(\bar{Y}(n) + \Delta_b)$. Since the Markov chains defined by the transition functions $\Phi(\Psi(\cdot))$ and $\Phi(\cdot + \Delta_b)$ are positive recurrent (we proved system $U^{(\epsilon)}$ stable for $\epsilon < \epsilon'$ but the proof implies the stability of this system as well) and irreducible,

$$\begin{aligned} N^*_{s,g} &= \lim_{n\to\infty} Y(n) \\ \bar{N}^*_{s,g} &= \lim_{n\to\infty} \bar{Y}(n) \end{aligned}$$

Since $Y(n) \leq_{st} \bar{Y}(n)$ for all n by induction, $N^*_{s,g} \leq_{st} \bar{N}^*_{s,g}$. ∎

Proof of Lemma 11: Let $N_{s,b'}$ denote the number of small jobs during the bad$'$ phase and $N_{s,b''}$ denote the number of jobs during the bad$''$ phase. We will stochastically bound $N_{s,b'}$ and $N_{s,b''}$ separately using stochastic coupling.

Bound for $N_{s,b'}$: We know that the lengths of bad$'$ phases of system $U^{(\epsilon)}$ are i.i.d. random variables. Let $T_{b'}$ denote a random variable which is equal in distribution to these. It is easy to see that $N_{s,b'}$ is equal in distribution to the number of small jobs in the following regenerative process. The system regenerates after i.i.d. periods whose lengths are equal in distribution to $T_{b'}$. At each regeneration the system starts with a random number of small jobs sampled from the distribution of $N^*_{s,g}$ and then the system evolves as an $M/M/K-1$ with arrival rate λp and service rate μ_s until the next renewal.

Now, $N_{s,b'}$ can be stochastically upper bounded by the number in system in another regenerative process where the renewals happen in the same manner but at every renewal the system starts with a random number of jobs sampled from the distribution of $\bar{N}^*_{s,g}$. These jobs never receive service. However, we also start another $M/M/K-1$ from origin (initially empty) with arrival rate λp and service rate μ_s and look at the total number of small jobs.

Finally, since $T_{b'}$ is an exponential random variable, by PASTA, the distribution of number of jobs at a randomly chosen time (or as $t \to \infty$) is the same as the number of jobs at a random chosen renewal. Therefore,

$$N_{s,b'} \leq_{st} \bar{N}^*_{s,g} + \Delta_{b'}(0) \tag{25}$$

Bound for $N_{s,b''}$: To obtain a stochastic upper bound on $N_{s,b''}$, we follow the same procedure as above. It is easy to see that $N_{s,b''}$ is stochastically upper bounded by the number of jobs in the following regenerative system. The renewals happen after i.i.d. intervals which are equal in distribution to $T_{b''}$, the random variable for the length of a bad$''$ phase in system $U^{(\epsilon)}$. At every renewal, the system starts with a random number of jobs sampled from the distribution of $\bar{N}^*_{s,g} + \Delta_{b'}(0)$ and external arrivals happen at a rate λ (there are no departures) until the next renewal. Let $T_{b''e}$ denote the age (and equal in distribution to the excess) of $T_{b''}$ and $A_\lambda(T)$ denote the number of arrivals in time T of a Poisson process with rate λ. This gives us the following stochastic bound on $N_{s,b''}$,

$$N_{s,b''} \leq_{st} \bar{N}^*_{s,g} + \Delta_{b'}(0) + A_\lambda\left(T_{b''e}\right) \tag{26}$$

The excess of $T_{b''}$ comes into the picture because we need the number of jobs at a randomly chosen instant of time during the bad$''$ phase. The time elapsed since the starting of a bad$''$ phase until this randomly chosen instant of time is distributed as $T_{b''e}$, the excess of $T_{b''}$. Finally, combining (25) and (26),

$$N_{s,b} \leq_{st} \bar{N}^*_{s,g} + \Delta_{b'}(0) + \mathbf{I}_{b''|b} A_\lambda\left(T_{b''e}\right) \tag{27}$$

∎

Proof of Lemma 13: The z-transform of $N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)$ is given by [16, Theorem 4]:

$$\widehat{N}^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)(z) = \frac{\beta z - (K-1)\mu_s(1-z)p_0}{\beta z - ((K-1)\mu_s - \lambda_s z)(1-z)} \tag{28}$$

where,

$$p_0 = \frac{\beta\xi}{(K-1)\mu_s(1-\xi)}$$

and ξ is the root of the polynomial in the denominator of $\widehat{N}^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)(z)$ in the interval $(0,1)$. Let η be the other root (lying in $(1,\infty)$). Therefore, we can write (28) as,

$$\begin{aligned}\widehat{N}^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)(z) &= \frac{\beta z - (K-1)\mu_s(1-z)\frac{\beta\xi}{(K-1)\mu_s(1-\xi)}}{-\lambda_s(z-\xi)(z-\eta)} \\ &= \frac{\beta}{-\lambda_s(1-\xi)(z-\eta)} \\ &= \frac{1-\eta}{z-\eta} \end{aligned} \tag{29}$$

The last step follows since $\widehat{N}^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)(z)|_{z=1} = 1$. By differentiating the transform in (29) and evaluating the derivatives at $z=1$, we have

$$\begin{aligned}\mathbf{E}\Big[N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)\Big] &= \frac{1}{\eta-1} \\ \mathbf{E}\Big[(N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T))^2\Big] &= \frac{2}{(\eta-1)^2} + \frac{1}{\eta-1}\end{aligned}$$

Factoring the denominator of (28), we can write η as the larger root of the quadratic equation:

$$z^2\lambda_s - z(\lambda_s + \beta + (K-1)\mu_s) + (K-1)\mu_s$$

That is,

$$\begin{aligned}\eta &= \frac{\lambda_s + \beta + (K-1)\mu_s + \sqrt{(\lambda_s+\beta+(K-1)\mu_s)^2 - 4\lambda_s(K-1)\mu_s}}{2\lambda_s} \\ &= \frac{\lambda_s + \beta + (K-1)\mu_s + \sqrt{(\lambda_s+\beta-(K-1)\mu_s)^2 + 4\beta(K-1)\mu_s}}{2\lambda_s} \\ &= \frac{\lambda_s + \beta + (K-1)\mu_s + (\lambda_s+\beta-(K-1)\mu_s)\sqrt{1+4\frac{\beta(K-1)\mu_s}{(\lambda_s+\beta-(K-1)\mu_s)^2}}}{2\lambda_s} \\ &= \frac{\lambda_s + \beta + (K-1)\mu_s + (\lambda_s+\beta-(K-1)\mu_s)\left(1+2\frac{\beta(K-1)\mu_s}{(\lambda_s+\beta-(K-1)\mu_s)^2} + \Theta(\beta^2)\right)}{2\lambda_s} \\ &= 1 + \frac{\beta}{\lambda_s}\cdot\left(1+\frac{(K-1)\mu_s}{(\lambda_s+\beta-(K-1)\mu_s)}\right) + \Theta(\beta^2) \\ &= 1 + \frac{\beta}{\lambda_s-(K-1)\mu_s} + \Theta(\beta^2)\end{aligned}$$

which results in the expressions in the lemma. ∎

Proof of Lemma 14: Recall that $\Delta_{b''}$ is the random variable denoting the number of small jobs that arrive during time $T_{b''}$, where $T_{b''}$ is the random variable for the length of the bad$''$ phase of a bad period. Using $A_\lambda(T)$ to denote the number of Poisson (with rate λ) arrivals in a random time interval T, we have $\Delta_{b''}$ is equal in distribution to $A_{\lambda p}(T_{b''})$. The following equalities are easy to prove:

$$\mathbf{E}[A_\lambda(T)] = \lambda\mathbf{E}[T] \tag{30}$$

$$\mathbf{E}\Big[(A_\lambda(T))^2\Big] = \lambda^2\mathbf{E}\big[T^2\big] + \lambda\mathbf{E}[T] \tag{31}$$

Thus we need the first two moments of $T_{b''}$ to obtain the first two moments of $\Delta_{b''}$. The Laplace transform of $T_{b''}$, $\widetilde{T_{b''}}(s)$, is given by:

$$\widetilde{T_{b''}}(s) = \frac{\mu_\ell}{\mu_\ell + \lambda(1-p)} + \frac{\lambda(1-p)}{\mu_\ell + \lambda(1-p)}\widetilde{B}^2(s) \tag{32}$$

where $\widetilde{B}(s)$ is the Laplace transform for the length of busy periods of an $M/M/1$ with arrival rate $\lambda(1-p)$ and service rate μ_ℓ. To see this, note that with probability $\frac{\mu_\ell}{\mu_\ell+\lambda(1-p)}$, the large job starting the bad phase leaves before another large job arrives and thus bad$''$ phase has length 0. With probability $\frac{\lambda(1-p)}{\mu_\ell+\lambda(1-p)}$, a large job arrives and starts the bad$''$ phase. In this case, the length of the bad$''$ phase is the time for an $M/M/1$ with arrival rate $\lambda(1-p)$ and service rate μ_ℓ to become empty starting with 2 jobs in the system. This is just the sum of two independent $M/M/1$ busy periods.

By differentiating the transform in (32) and evaluating at $s = 0$, we obtain:

$$\mathbf{E}[T_{b''}] = \frac{\lambda(1-p)}{\mu_\ell + \lambda(1-p)}\left(\frac{2}{\mu_\ell - \lambda(1-p)}\right) \;=\; \Theta(1) \tag{33}$$

$$\mathbf{E}\left[T_{b''}^2\right] = \frac{\lambda(1-p)}{\mu_\ell + \lambda(1-p)}\left(\frac{4\mu_\ell}{(\mu_\ell - \lambda(1-p))^3}\right) \;=\; \Theta\left(\frac{1}{\epsilon}\right) \tag{34}$$

Obtaining $\mathbf{E}[\Delta_{b''}]$ and $\mathbf{E}\left[\Delta_{b''}^2\right]$: Substituting $\lambda \equiv \lambda p$ and $T \equiv T_{b''}$ in (30)-(31) and using (33)-(34), we get the following asymptotics which will be sufficient for our purposes:

$$\mathbf{E}[\Delta_{b''}] = \lambda p \mathbf{E}[T_{b''}] = \Theta(1) \tag{35}$$

$$\mathbf{E}\left[\Delta_{b''}^2\right] = \lambda^2 p^2 \mathbf{E}\left[T_{b''}^2\right] + \lambda p \mathbf{E}[T_{b''}] = \Theta\left(\frac{1}{\epsilon}\right) \tag{36}$$

Obtaining $\mathbf{E}[A_\lambda(T_{b''e})]$: $A_\lambda(T_{b''e})$ denotes the number of Poisson (with rate λ) arrivals in a random time interval given by $T_{b''e}$ – the stationary age (equivalently excess) of a renewal process where renewals intervals are i.i.d. according to $T_{b''}$. Note that $A(T_{b''e})$ is not equal in distribution to $\Delta_{b''}$ since $T_{b''}$ is not an exponential random variable. From (30),

$$\mathbf{E}[A_\lambda(T_{b''e})] = \lambda \mathbf{E}[T_{b''e}]$$

From the formula for stationary age (equivalently excess) of a renewal process [33],

$$\mathbf{E}[T_{b''e}] = \frac{\mathbf{E}\left[T_{b''}^2\right]}{2\mathbf{E}[T_{b''}]} = \Theta\left(\frac{1}{\epsilon}\right)$$

Combining, we get the following asymptotics for $\mathbf{E}[A_\lambda(T_{b''e})]$ which will be sufficient for our purposes:

$$\mathbf{E}[A_\lambda(T_{b''e})] = \Theta\left(\frac{1}{\epsilon}\right) \tag{37}$$

$$\frac{\mathbf{Pr}[b'']}{\mathbf{Pr}[b]}\mathbf{E}[A_\lambda(T_{b''e})] = \frac{\mathbf{E}[T_{b''}]}{\left(\frac{1}{\mu_\ell - \lambda(1-p)}\right)}\mathbf{E}[A_\lambda(T_{b''e})] \;=\; \Theta(1) \tag{38}$$

■

Proof of Lemma 15: Recall that $N^{(Int)}$ denotes the number of jobs in the interrupted $M/M/K$ system. Let $\widehat{N^{(Int)}}(z)$ be the z-transform of $N^{(Int)}$ and let $\widehat{\Delta}(z)$ be the z-transform of Δ. Since

the interruptions happen according to a Poisson process, $N^{(Int)}$ also denotes the random variable for the number of jobs *just before* the interruptions. Let f map the z-transform of the distribution of number of jobs in an $M/M/K$ at time $t = 0$ to the z-transform of the distribution of number of jobs after the $M/M/K$ system has run (uninterrupted) for $T \sim \text{Exp}(\alpha)$ time. The solution for $\widehat{N^{(Int)}}(z)$ is given by the following fixed point equation:

$$\widehat{N^{(Int)}}(z) = f\left(\widehat{N^{(Int)}}(z)\widehat{\Delta}(z)\right)$$

Our next goal is to derive the function $f(\cdot)$. Let $p_i(t)$ denote the probability that there are i jobs in the $M/M/K$ system at time t. We can write the following differential equations for $p_i(t)$:

$$\frac{d}{dt}p_0(t) = -\lambda p_0(t) + \mu p_1(t) \tag{39}$$

$$\frac{d}{dt}p_i(t) = \lambda p_{i-1}(t) - (\lambda + i\mu)p_i(t) + (i+1)\mu p_{i+1}(t) \qquad \ldots 1 \le i \le K-1 \tag{40}$$

$$\frac{d}{dt}p_i(t) = \lambda p_{i-1}(t) - (\lambda + K\mu)p_i(t) + K\mu p_{i+1}(t) \qquad \ldots i \ge K \tag{41}$$

Let $\widehat{\Pi}(z,t) = \sum_{i=0}^{\infty} p_i(t)z^i$. Multiplying (39) by z^0 and the set of equations (40) and (41) by z^i and summing, we have:

$$\begin{aligned}\frac{\partial}{\partial t}\widehat{\Pi}(z,t) = \widehat{\Pi}(z,t)\left[K\mu\left(\frac{1}{z}-1\right) + \lambda(z-1)\right] \\ + \mu\left(1-\frac{1}{z}\right)\left[Kp_0(t) + (K-1)zp_1(t) + \ldots + z^{K-1}p_{K-1}(t)\right]\end{aligned} \tag{42}$$

Let $\widehat{\Pi}_\alpha(z) = \int_0^\infty \widehat{\Pi}(z,t)\alpha e^{-\alpha t}dt$ and $p_{i,\alpha} = \int_0^\infty p_i(t)\alpha e^{-\alpha t}dt$. Integrating by parts, we get:

$$\begin{aligned}\widehat{\Pi}_\alpha(z) &= \int_0^\infty \widehat{\Pi}(z,t)\alpha e^{-\alpha t}dt \;=\; \int_0^\infty \left(\widehat{\Pi}(z,t)\right)\left(d\left(-e^{-\alpha t}\right)\right) \\ &= \left[-\widehat{\Pi}(z,t)e^{-\alpha t}\right]_{t=0}^{\infty} - \int_{t=0}^{\infty}\left(-e^{-\alpha t}\right)\left(d\widehat{\Pi}(z,t)\right) \\ &= \widehat{\Pi}(z,0) + \frac{1}{\alpha}\int_{t=0}^{\infty}\alpha e^{-\alpha t}\left(\widehat{\Pi}(z,t)\left[K\mu\left(\frac{1}{z}-1\right) + \lambda(z-1)\right]\right. \\ &\qquad \left. + \mu\left(1-\frac{1}{z}\right)\left[Kp_0(t) + (K-1)zp_1(t) + \ldots + z^{K-1}p_{K-1}(t)\right]\right) \\ &= \widehat{\Pi}(z,0) + \frac{\widehat{\Pi}_\alpha(z)}{\alpha}\left[K\mu\left(\frac{1}{z}-1\right) + \lambda(z-1)\right] + \frac{\mu}{\alpha}\left(1-\frac{1}{z}\right)\left[Kp_{0,\alpha} + \ldots + z^{K-1}p_{K-1,\alpha}\right]\end{aligned} \tag{43}$$

To obtain $\widehat{N^{(Int)}}(z)$, we substitute $\widehat{\Pi}_\alpha(z) = \widehat{N^{(Int)}}(z)$, $\widehat{\Pi}(z,0) = \widehat{N^{(Int)}}(z)\widehat{\Delta}(z)$ and $p_{i,\alpha} = p_i = \mathbf{Pr}\left[N^{(Int)} = i\right]$. This gives:

$$\widehat{N^{(Int)}}(z) = \frac{\mu\left[Kp_0 + (K-1)zp_1 + \ldots + z^{K-1}p_{K-1}\right]}{(K\mu - \lambda z) - \alpha z\left(\frac{1-\widehat{\Delta}(z)}{1-z}\right)} \tag{44}$$

Since $\widehat{N^{(Int)}}(1) = 1$, and $\lim_{z\to 1}\frac{1-\widehat{\Delta}(z)}{1-z} = \mathbf{E}[\Delta]$, we get

$$Kp_0 + (K-1)p_1 + \ldots + p_{K-1} = K - \frac{\lambda + \alpha\mathbf{E}[\Delta]}{\mu} \tag{45}$$

The sum on the left is precisely the expected number of idle servers at $T \sim \text{Exp}(\alpha)$. Let

$$C = 0 \cdot K \cdot p_0 + (K-1) \cdot 1 \cdot p_1 + (K-2) \cdot 2 \cdot p_2 + \ldots + 1 \cdot (K-1) \cdot p_{K-1}$$

Then,

$$\begin{aligned}
\mathbf{E}\left[N^{(Int)}\right] &= \frac{d}{dz}\widehat{N^{(Int)}}(z)\Big|_{z=1} \\
&= \frac{\mu \frac{d}{dz}\left[Kp_0 + (K-1)zp_1 + \ldots + z^{K-1}p_{K-1}\right]}{(K\mu - \lambda z) - \alpha z\left(\frac{1-\widehat{\Delta}(z)}{1-z}\right)}\Bigg|_{z=1} \\
&\quad - \frac{\mu\left[Kp_0 + (K-1)zp_1 + \ldots + z^{K-1}p_{K-1}\right]}{\left((K\mu - \lambda z) - \alpha z\left(\frac{1-\widehat{\Delta}(z)}{1-z}\right)\right)^2}\frac{d}{dz}\left((K\mu - \lambda z) - \alpha z\left(\frac{1-\widehat{\Delta}(z)}{1-z}\right)\right)\Bigg|_{z=1} \\
&= \frac{\mu C}{K\mu - \lambda - \alpha\mathbf{E}[\Delta]} \\
&\quad - \frac{\widehat{N^{(Int)}}(1)}{K\mu - \lambda - \alpha\mathbf{E}[\Delta]}\left(-\lambda - \alpha\frac{1-\widehat{\Delta}(z)}{1-z} - \alpha z\left(\frac{1-\widehat{\Delta}(z) - (1-z)\frac{d\widehat{\Delta}(z)}{dz}}{(1-z)^2}\right)\right)\Bigg|_{z=1}
\end{aligned}$$

and applying L'Hospital's rule to the last term,

$$\begin{aligned}
&= \frac{\mu C}{K\mu - \lambda - \alpha\mathbf{E}[\Delta]} \\
&\quad - \frac{1}{K\mu - \lambda - \alpha\mathbf{E}[\Delta]}\left(-\lambda - \alpha\mathbf{E}[\Delta] - \alpha\left(\frac{-\frac{d}{dz}\widehat{\Delta}(z) - (1-z)\frac{d^2\widehat{\Delta}(z)}{dz^2} + \frac{d}{dz}\widehat{\Delta}(z)}{-2(1-z)}\right)\Bigg|_{z=1}\right) \\
&= \frac{\mu C}{K\mu - \lambda - \alpha\mathbf{E}[\Delta]} - \frac{1}{K\mu - \lambda - \alpha\mathbf{E}[\Delta]}\left(-\lambda - \alpha\mathbf{E}[\Delta] - \alpha\frac{\mathbf{E}\left[\Delta^2\right] - \mathbf{E}[\Delta]}{2}\right) \\
&= \frac{\mu C}{K\mu - \lambda - \alpha\mathbf{E}[\Delta]} + \frac{\lambda + \frac{\alpha}{2}\left(\mathbf{E}\left[\Delta^2\right] + \mathbf{E}[\Delta]\right)}{K\mu - \lambda - \alpha\mathbf{E}[\Delta]} \qquad (46)
\end{aligned}$$

To calculate C we need the following relations obtained by matching the coefficients of z^i, $i = 0, \ldots, K-1$, from (43):

$$\begin{aligned}
-\lambda p_{0,\alpha} + \mu p_{1,\alpha} &= \alpha\left[p_{0,\alpha} - p_0(0)\right] \\
\lambda p_{i-1,\alpha} - (\lambda + i\mu)p_{i,\alpha} + (i+1)\mu p_{i+1,\alpha} &= \alpha\left[p_{i,\alpha} - p_i(0)\right] \qquad \ldots 1 \le i \le K-1
\end{aligned}$$

which yields $p_{i,\alpha} = p_{0,\alpha}\frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i + \Theta(\alpha)$. Let π_i be the stationary probabilities of an $M/M/K$ system with arrival rate λ and mean job size $\frac{1}{\mu}$. We can use (45) to write:

$$K\pi_0 + (K-1)\pi_1 + \ldots + \pi_{K-1} = K - \frac{\lambda}{\mu}$$

or equivalently,

$$\pi_0\left(K \cdot 1 + (K-1) \cdot \frac{\lambda}{\mu} + \ldots + 1 \cdot \frac{1}{(K-1)!}\left(\frac{\lambda}{\mu}\right)^{K-1}\right) = K - \frac{\lambda}{\mu}$$

Rewriting (45) and using the facts $p_i = p_0 \frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i + \Theta(\alpha)$ and $\alpha\mathbf{E}[\Delta] = o(1)$:

$$p_0\left(K\cdot 1 + (K-1)\cdot\frac{\lambda}{\mu} + \ldots + 1\cdot\frac{1}{(K-1)!}\left(\frac{\lambda}{\mu}\right)^{K-1}\right) + \Theta(\alpha) = K - \frac{\lambda + o(1)}{\mu}$$

which gives $p_0 = \pi_0 + o(1)$, and hence $p_i = \pi_i + o(1)$ for $i \leq K-1$. Using this, we have:

$$\begin{aligned}
\frac{\mu C + \lambda}{K\mu - \lambda - \alpha\mathbf{E}[\Delta]} &= \frac{\mu\left(\sum_{i=0}^{K-1} i(K-i)p_i\right) + \lambda}{K\mu - \lambda + \alpha\mathbf{E}[\Delta]} \\
&= \frac{\mu\left(\sum_{i=0}^{K-1} i(K-i)\pi_i\right) + \lambda + o(1)}{K\mu - \lambda + o(1)} \\
&= \frac{\mu\left(\sum_{i=0}^{K-1} i(K-i)\pi_i\right) + \lambda}{K\mu - \lambda} + o(1) \\
&= \mathbf{E}\left[N^{M/M/K}\right] + o(1)
\end{aligned}$$

where $\mathbf{E}\left[N^{M/M/K}\right]$ is the mean number of jobs in a stationary $M/M/K$ queue with arrival rate λ and service rate μ. To see that $\mathbf{E}\left[N^{M/M/K}\right]$ can be written in the above form, note that when $\Delta \equiv 0$, $\mathbf{E}\left[N^{(Int)}\right] = \frac{\mu C+\lambda}{K\mu-\lambda}$ but $\mathbf{E}\left[N^{(Int)}\right] = \mathbf{E}\left[N^{M/M/K}\right]$. Finally,

$$\mathbf{E}\left[N^{(Int)}\right] = \mathbf{E}\left[N^{M/M/K}\right] + \frac{\frac{\alpha}{2}\mathbf{E}\left[\Delta^2\right]}{K\mu - \lambda} + o(1)$$

since $\alpha\mathbf{E}[\Delta] = o(1)$. ■

B Proof of Proposition 1

Our aim is to prove that for $K \geq 2$, $\rho \geq (K-1)/K$ and $C^2 > 1$

$$\frac{C^2-1}{2}\mathbf{E}\left[W^{M/M/K}\right] > \frac{1}{1-\rho}\left[\rho - \frac{K-1}{K}\right]\frac{C^2-1}{2}. \tag{47}$$

Recall that we take $\mathbf{E}[X] = 1$ without loss of generality so that $\rho \geq (K-1)/K$ is equivalent to $\lambda \geq K-1$. Let $C(K,\lambda)$ be the probability of wait in an $M/M/K$. It is easily shown that

$$\mathbf{E}\left[W^{M/M/K}\right] = \frac{C(K,\lambda)}{K-\lambda}. \tag{48}$$

Therefore, using $\rho = \frac{\lambda}{K}$, (47) holds if (we have assumed $\frac{C^2-1}{2} > 0$)

$$C(K,\lambda) > [\lambda - (K-1)]. \tag{49}$$

It is known that $C(K,\lambda)$ is a strictly convex function in λ on $[0,K]$ (see [27]). Since (49) trivially holds for $\lambda = K-1$, and since the right hand side of (49) has derivative (w.r.t. λ) 1, it suffices to show that

$$\frac{d}{d\lambda}C(K,\lambda)\bigg|_{\lambda\to K} < 1. \tag{50}$$

Let A_λ be a random variable that is Poisson with mean λ. It is well known ([24], page 103) that

$$C(K,\lambda) = \frac{1}{\rho + (1-\rho)\frac{P(A_\lambda \leq K)}{P(A_\lambda = K)}}. \tag{51}$$

Using this expression, we find that

$$\begin{aligned}\frac{d}{d\lambda}C(K,\lambda)\bigg|_{\lambda\to K} &= \frac{d}{d\lambda}\frac{1}{\frac{\lambda}{K}+\left(1-\frac{\lambda}{K}\right)\frac{P(A_\lambda\leq K)}{P(A_\lambda=K)}}\bigg|_{\lambda\to K}\\ &= -\frac{\frac{1}{K}-\frac{1}{K}\frac{P(A_\lambda\leq K)}{P(A_\lambda=K)}+\left(1-\frac{\lambda}{K}\right)\frac{d}{d\lambda}\frac{P(A_\lambda\leq K)}{P(A_\lambda=K)}}{\left(\frac{\lambda}{K}+\left(1-\frac{\lambda}{K}\right)\frac{P(A_\lambda\leq K)}{P(A_\lambda=K)}\right)^2}\bigg|_{\lambda\to K}\\ &= \frac{1}{K}\frac{P(A_K\leq K-1)}{P(A_K=K)}\\ &= \frac{1}{K}\sum_{k=0}^{K-1}\frac{P(A_K=k)}{P(A_K=K)}.\end{aligned}$$

Now, note that at $\lambda = K$

$$\frac{P(A_K=K-1)}{P(A_K=K)} = \frac{K^{K-1}/(K-1)!}{K^K/K!} = 1.$$

If $k < K-1$ we find that

$$\frac{P(A_K=k)}{P(A_K=k+1)} = \frac{k+1}{K} < 1,$$

which implies that

$$\frac{P(A_K=k)}{P(A_K=k+1)} < 1, \quad k < K-1.$$

Consequently, for $K \geq 2$, we see that

$$\frac{d}{d\lambda}C(K,\lambda)\bigg|_{\lambda\to K} = \frac{1}{K}\sum_{k=0}^{K-1}\frac{K^k/k!}{K^K/K!} < 1, \tag{52}$$

which completes the proof of the proposition.