The effect of higher moments of job size distribution on the performance of an $M/G/K$ queueing system

Varun Gupta[*] **Jim Dai**[†] **Mor Harchol-Balter**[*]
Bert Zwart[†]

February 2008
CMU-CS-08-106

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[*] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[†] School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Keywords: Multi-server systems; $M/G/K$; two-moment approximation; inapproximability; higher moments;

Abstract

The $M/G/K$ queueing system is 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 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. This is the first such result in the literature. 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 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, 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 the FCFS queue. We denote the load of this $M/G/K$ system as $\rho = \frac{\lambda \mathbf{E}[X]}{K} < 1$. 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 [22] 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, [16, 17, 21, 27, 28, 40], but all these closed-form approximations involve only the first two moments of the job size distribution.

Whitt [39], 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 too 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 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 $C^2 = 99$. 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 [10] and the supercomputing jobs [14]. 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

[1] The squared coefficient of variation of a 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:

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

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 approx. (Eqn. 1)	6.6873	33.4366
Weibull	6.0691±0.0138	25.9896±0.1773
Trunc. Pareto ($\alpha = 1.1$)	5.5277±0.0216	24.6049±0.2837
Lognormal	4.9937±0.0249	19.5430±0.4203
Trunc. Pareto ($\alpha = 1.3$)	4.8788±0.0249	18.7738±0.3612
Trunc. 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 $\rho = 0.9$. All job size distributions have $\mathbf{E}[X] = 1$.

a huge disagreement between the actual mean waiting time and the 2-moment approximation (1). Further, even among the distributions considered in Table 1, there is a substantial variation in the mean waiting times.

In this paper, we support the above experimental findings with an investigation of 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. Hyper-exponential distributions allow 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 $\mathbf{E}\left[W^{M/G/K}\right]$ when the first two moments of X are fixed. We refer to this range as "the gap". To define the gap, set

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

and

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

The gap spans $(W_l^{C^2}, W_h^{C^2})$. As one of the major contributions of this paper, we prove the following theorems:

Theorem 1.1 *For any load $\rho < 1$ and finite C^2,*

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

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 1.2 *For any finite C^2,*

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

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

That is, we derive a lower bound for $W_h^{C^2}$ and an upper bound for $W_l^{C^2}$. Therefore, Theorems 1.1 and 1.2 only give a lower bound on the span of the gap for general distributions. Observe that the gap can be quite large if the C^2 of the job size distribution is high. In particular, when $\rho < \frac{K-1}{K}$, the maximum possible mean waiting time is at least $\left(\frac{C^2+1}{2}\right)$ times the minimum possible mean waiting time. We thus prove a lower bound on the error of approximation (1). Note that the lower bound on $W_h^{C^2}$ in Theorem 1.1 is the same as the 2-moment approximation in (1). Further, Theorems 1.1 and 1.2 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, in Theorem 1.2, the lower bound 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 criterion 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 not (see [29] and references therein). Theorem 1.2 indicates that even when the C^2 of the job size distribution is finite, having a spare server can potentially reduce the effect of C^2 on the mean waiting time.

To prove Theorems 1.1 and 1.2, 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. Since the hyperexponential distributions only allow $C^2 > 1$, the span described by the theorems is non-empty only when $C^2 > 1$ even though the theorems are true for all values of C^2. In fact, we are able to show that our lower bound for the span of the gap is strictly positive when $K \geq 2$ and $C^2 > 1$:

Proposition 1.3 *Let* $\mathbf{E}\left[W^{M/M/K}\right]$ *be the mean waiting time in an* $M/M/K$ *with mean job size* 1. *Define:*

$$\underline{W_h^{C^2}} \triangleq \left(\frac{C^2+1}{2}\right)\mathbf{E}\left[W^{M/M/K}\right]$$

and,

$$\overline{W_l^{C^2}} \triangleq \begin{cases} \mathbf{E}\left[W^{M/M/K}\right] & \text{if } \rho < \frac{K-1}{K} \\ \mathbf{E}\left[W^{M/M/K}\right] + \frac{1}{1-\rho}\left[\rho - \frac{K-1}{K}\right]\frac{C^2-1}{2} & \text{if } \rho \geq \frac{K-1}{K} \end{cases}$$

For all values of $\rho \in (0,1)$, $K \geq 2$ *and* $C^2 > 1$,

$$\underline{W_h^{C^2}} > \overline{W_l^{C^2}}.$$

We provide a proof of the proposition in Appendix B.

The two bounds can actually be shown to be identical for $K = 1$, and in fact agree with the well-known Pollaczek Khintchine formula

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

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

Results similar to Theorems 1.1 and 1.2, were derived for the mean queue length of a $GI/M/1$ queue by Eckberg [11] and extended by Whitt [38] 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. In fact, for $GI/M/1$ queues, tight bounds on the mean queue length given the first n moments of the interarrival time distribution can be obtained by employing the theory of complete Tchebycheff systems [18].

Outline

Section 2 reviews existing work on obtaining closed-form, numerical and heavy-traffic approximations for $\mathbf{E}\left[W^{M/G/K}\right]$. As mentioned earlier, we prove Theorems 1.1 and 1.2 by looking at two extreme distributions in the class of 2-phase hyperexponential distributions. Therefore, in Section 3, we begin with some numerical experiments based on the 2-phase hyperexponential distributions. These experiments help us answer the question: "Which characteristics of the job size distribution, outside of the first two moments, are important in determining the mean waiting time?" Sections 4 and 5 are devoted to proving Theorems 1.1 and 1.2, respectively. In Section 6, we address the question of effect of higher moments of job size distribution on the mean waiting time. In Section 7, we state conjectures on the exact values of $W_h^{C^2}$ and $W_l^{C^2}$ and the effect of higher moments of job size distribution on $\mathbf{E}\left[W^{M/G/K}\right]$. 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 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 by Lee and Longton [22]:

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

This approximation is very simple, is exact for $K = 1$ and was shown to be asymptotically exact in heavy traffic by Köllerström [21]. The same expression is obtained by Nozaki and Ross [27] by making approximating assumptions about the $M/G/K$ system and solving for exact state probabilities of the approximating system, and by Hokstad [16] by starting with the exact equations and making approximations in the solution phase. Boxma et al. [28] obtain a closed-form approximation for the mean waiting time in an $M/D/K$ system, extending the heavy traffic approximation of Cosmetatos [5]. Takahashi [33] obtains expressions for mean waiting time by assuming a parametric formula. Kimura [20] 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. [15] 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 [16, 17, 24, 25, 30]. Boxma et al. [28] 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 parametric form and matching the heavy traffic and light traffic behaviors. Other numerical algorithms include [7, 8, 9]. 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 [4, 19, 21, 35, 39, 40]. The diffusion approximations used in [35] 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 [12] and references therein.

Bounds on the mean waiting time for $M/G/K$ queues (and more generally for $G/G/K$ queues) have been obtained by assuming various orderings (stochastic ordering, increasing convex ordering) on the distribution of job sizes (see [6, 26, 32, 36, 37]), but these tend to be very loose as approximations. Moreover, one does not always have the required strong orderings on the job size distribution.

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 job size distribution.

3 Experiments with the H_2 distribution

Our goal in this section is to study the effect of characteristics other than the first two moments of the job size distribution on $\mathbf{E}\left[W^{M/G/K}\right]$. To do this, we restrict our attention to the class of two-phase hyperexponential distributions, denoted by H_2 (see Definition 3.1 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 gives us a method to create a set of distributions with any given first two moments 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 3.1 *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 define the $n-$phase hyperexponential distribution, H_n, with parameters μ_i, p_i, $i = 1, \ldots, n$, as:*

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

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

Definition 3.2 *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, $i = 1, \ldots, n$, as:*

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

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

[3] In [7, 39], the authors use

$$r = \frac{p_1/\mu_1}{p_1/\mu_1 + p_2/\mu_2}$$

as the third parameter to specify the H_2 distribution. We choose the third moment because it is more universal and well understood than r. Further, r is an increasing function of the third moment.

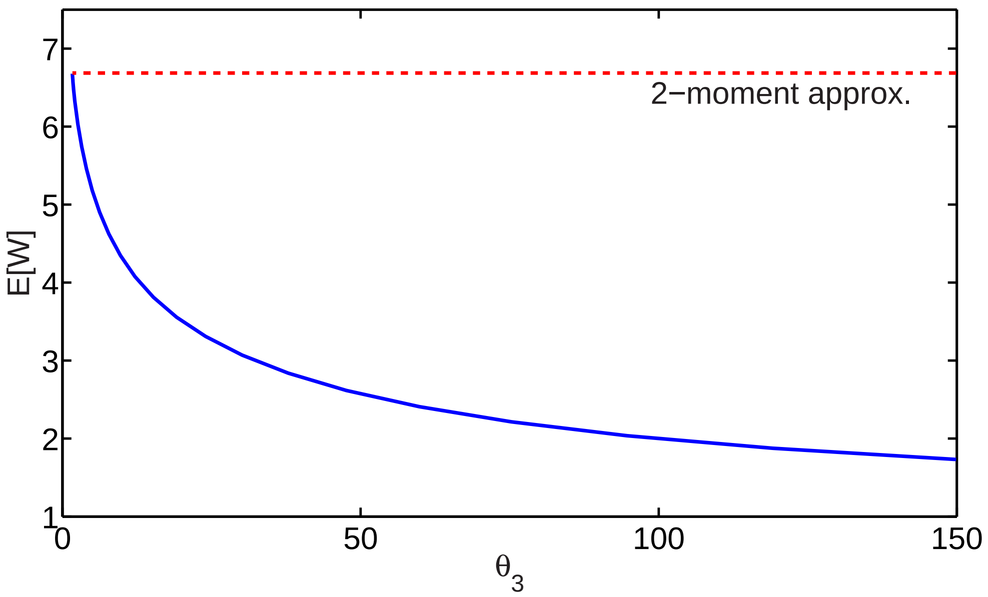


Figure 1: Illustration of the effect of the normalized 3rd moment, θ_3, of the job size distribution on mean waiting time of an $M/H_2/10$ system (solid line). The parameters of the job size distribution were held constant at $\mathbf{E}[X] = 1$ and $C^2 = 19$ with load $\rho = 0.9$. The dashed line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

Figure 1 shows the $M/H_2/K$ 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:

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

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 third moment* of X. We also observe that the existing two moment approximation is grossly 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 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

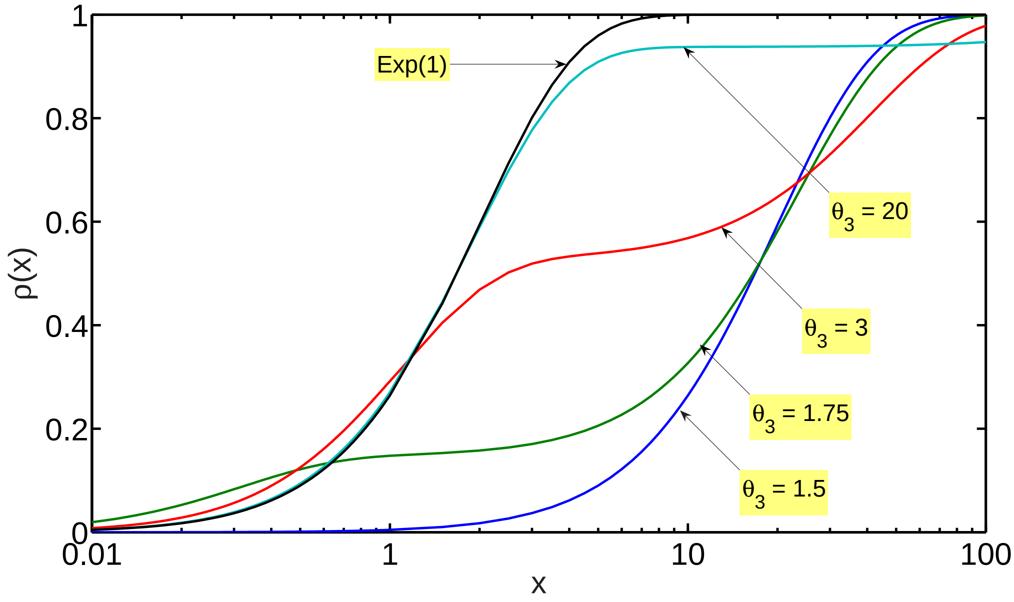


Figure 2: Illustration of the effect of the normalized 3rd moment, θ_3, on the distribution of load as a function of job size for the H_2 class of distributions. The first two moments were held constant at $\mathbf{E}[X] = 1$ and $C^2 = 19$. The distribution of the load for exponential distribution with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

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. 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.

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? Theorems 1.1 and 1.2 answer this question. Theorem 1.1 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 3.2) 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 H_2^* jobs size distribution. Likewise, Theorem 1.2 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, Theorems 1.1 and 1.2 provide bounds on the exact span of $\mathbf{E}\left[W^{M/G/K}\right]$. (We state conjectures about the exact span in Section 7.)

4 Proof of Theorem 1

To prove Theorem 1.1, it suffices to show the existence of a job size distribution with SCV C^2 which gives the desired expression for mean waiting time. For this purpose, we consider the following degenerate hyperexponential distribution [4]:

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

It is easy to verify that the above distribution has mean 1, squared coefficient of variation C^2 and $\theta_3 = \frac{3}{2}$. We denote the $M/G/K$ system with H_2^* job size distribution as $M/H_2^*/K$. It is interesting to note that 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 4.1 *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. ■

We now analyze the mean waiting time in an $M/H_2^*/K$ system. 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 $M/M/K$ system with arrival rate $\frac{2\lambda}{C^2+1}$ and mean job size $\frac{C^2+1}{2}$. The latter system, however, is just an $M/M/K$ system with arrival rate λ and mean job size 1 seen on a slower time scale, slowed by a factor $\frac{C^2+1}{2}$. Hence, the mean waiting time of the original system is also $\frac{C^2+1}{2}$ times the mean waiting time of an $M/M/K$ system with arrival rate λ and mean job size 1. That is,

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

5 Proof of Theorem 2

As in the proof of Theorem 1.1, to prove Theorem 1.2, it suffices to show the existence of a job size distribution with SCV C^2 which gives the desired mean waiting time. However, in this case, the distribution we need is the limit of a sequence of distributions and hence the analysis will involve finding the limit of mean waiting times of a sequence of $M/G/K$ systems. Since the proof 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

We consider a sequence of systems where we *fix the first two moments of the job size distribution* and look at increasing values of θ_3 as parameterized by a parameter ϵ which allows for increasing the third moment as ϵ goes to 0. More precisely, the system parameterized by ϵ is the $M/H_2^{(\epsilon)}/K$ (see Section 5.2, Definition 5.1) and we will consider the behavior of this sequence of systems as $\epsilon \to 0$.

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, respectively.

[4] We use the notation $\text{Exp}(\mu)$ to denote an exponential random variable with mean $\frac{1}{\mu}$.

2. We show that the expected number of large jobs, $\mathbf{E}\left[N_\ell^{M/H_2^{(\epsilon)}/K}\right]$, 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,

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

(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, $\mathbf{E}\left[N_s^{U^{(\epsilon)}} \mid \text{good period}\right]$ and $\mathbf{E}\left[N_s^{U^{(\epsilon)}} \mid \text{bad period}\right]$, respectively, and obtain an upper bound on $\mathbf{E}\left[N_s^{U^{(\epsilon)}}\right]$ using the law of total probability:

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

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

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

where A is the random variable for (the stochastic upper bound on) the number of small jobs at the end of a good period.

(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,

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

(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

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

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

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

5.2 Preliminaries

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

Definition 5.1 *We define a family of distributions parameterized by ϵ as follows:*

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

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

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

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

Lemma 5.2 *The $\mu_s^{(\epsilon)}$, $\mu_\ell^{(\epsilon)}$ and $p^{(\epsilon)}$ can be expressed in terms of ϵ as* [5]*:*

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

Corollary 5.3 *As $\epsilon \to 0$,*

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

[5] We say a function $h(\epsilon)$ is $\Theta(g(\epsilon))$ if

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

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

Corollary 5.3 is saying that 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 $\epsilon \to 0$. Note that, as $\epsilon \to 0$, the $H_2^{(\epsilon)}$ distribution converges in distribution to the $\mathrm{Exp}(1)$ 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 [3, 31]. However, convergence in distribution *does not* imply convergence of the means. Indeed, in the case where $\rho > \frac{K-1}{K}$, we find that the mean of the limiting system is not $\mathbf{E}\left[W^{M/M/K}\right]$. This can be easily verified for $K = 1$, where the mean waiting time is given by the Pollaczek-Khintchine formula (4).

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 [6] $o(1)$, it suffices to consider only the small jobs.

Lemma 5.4 $\mathbf{E}\left[N_\ell^{M/H_2^{(\epsilon)}/K}\right] = o(1)$

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. Specifically, 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. By $M(a)/M(b)/k$, we mean an $M/M/k$ queue with arrival rate a and service rate b. It is easy to see that

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

The proof is completed by the following lemma:

Lemma 5.5 $\mathbf{E}\left[\overline{N_\ell}^{(\epsilon)}\right] = o(1)$

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 $U^{(\epsilon)}$. 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

[6] We say a function $h(\epsilon)$ is $o(g(\epsilon))$ if

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

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

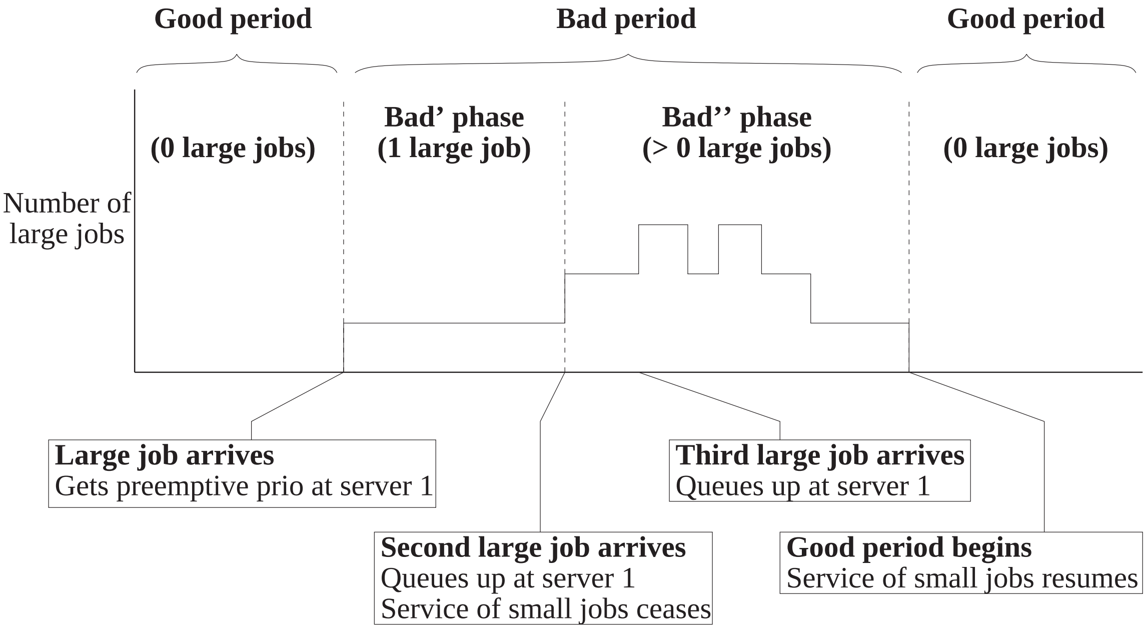


Figure 3: Construction of system $U^{(\epsilon)}$ which upper bounds the number of jobs in an $M/H_2^{(\epsilon)}/K$

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 5.6 *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 $U^{(\epsilon)}$, $N_s^{U^{(\epsilon)}}$.*

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:

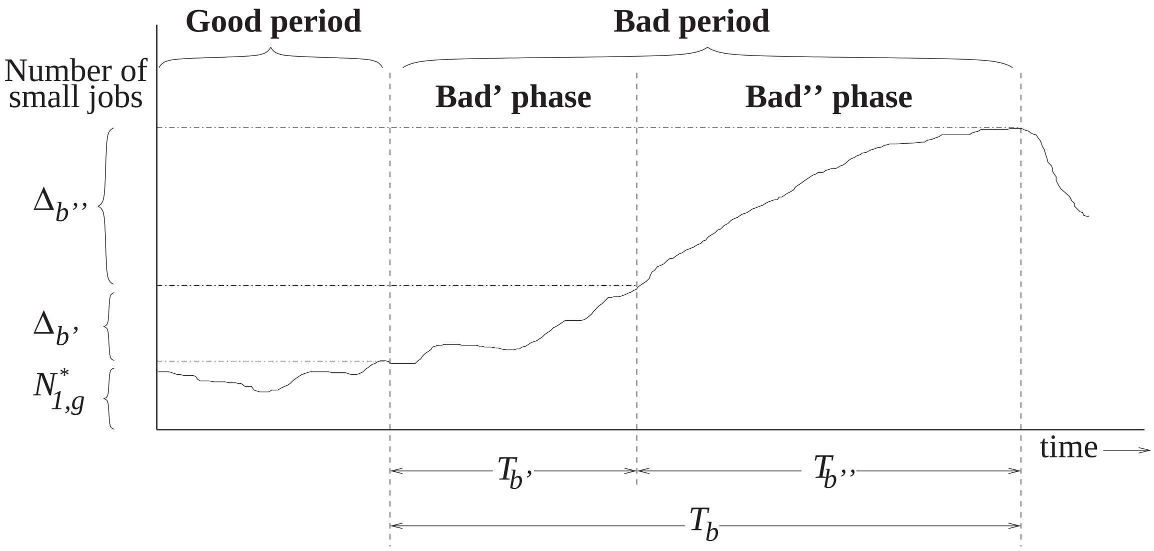


Figure 4: Notation used for analysis of system $U^{(\epsilon)}$

Lemma 5.7 *The upper bounding system, $U^{(\epsilon)}$, is stable for $\epsilon < \epsilon'$ where*

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

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 number of small jobs *during* a good period
- $N_{s,b} \equiv$ the 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,

$$\mathbf{E}[N_s] = \mathbf{E}[N_{s,g}]\mathbf{Pr}[g] + \mathbf{E}[N_{s,b}]\mathbf{Pr}[b] \tag{6}$$

In Section 5.5.1, we derive stochastic upper bounds on $N_{s,g}$ and $N_{s,b}$, which give us an upper bound, (8), on $\mathbf{E}[N_s]$. In Sections 5.5.2 and 5.5.3, we derive expressions for the quantities appearing in (8). These are used to obtain the final upper bound on $\mathbf{E}[N_s]$ at the end of Section 5.5.1.

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:

$$\bar{N}^*_{s,g} \overset{d}{=} \Phi(\bar{N}^*_{s,g} + \Delta_b) \tag{7}$$

Lemma 5.8

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

Proof: 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 [34] (see also [13] for a similar use of conditional PASTA),

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

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 (7) gives a stochastic upper bound on $N^*_{1,g}$. A formal proof of the stochastic inequality is Appendix A. ■

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

Lemma 5.9

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

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: 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 $\mathbf{E}[N_s]$:

$$\mathbf{E}[N_s] \leq \mathbf{E}\left[\bar{N}^*_{s,g}\right]\mathbf{Pr}[g] + \mathbf{E}\left[\bar{N}^*_{s,g} + \Delta_{b'}(0) + \mathbf{I}_{b''|b} A_\lambda\left(T_{b''e}\right)\right]\mathbf{Pr}[b] \tag{8}$$

To complete the proof, we need expressions for each of the quantities in equation (8). 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 5.2 for expressions for $p^{(\epsilon)}$ and $\mu_\ell^{(\epsilon)}$),

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

Further, in the proof of Lemma 5.12, we prove

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

Finally, substituting the expressions for $\mathbf{E}\left[\bar{N}^*_{s,g}\right]$, $\mathbf{E}[\Delta_{b'}(0)]$, $\frac{\mathbf{Pr}[b'']}{\mathbf{Pr}[b]}\mathbf{E}[A_\lambda\left(T_{b''e}\right)]$ and $\mathbf{Pr}[b]$ into equation (8), we get

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

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

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

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

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 5.10 gives the expressions for $\mathbf{E}[\Delta_{b'}(0)]$ and $\mathbf{E}\left[\Delta_{b'}^2(0)\right]$. Lemma 5.12 gives the asymptotic expressions for $\mathbf{E}[\Delta_{b''}]$ and $\mathbf{E}\left[\Delta_{b''}^2\right]$ which will be sufficient for our purposes of obtaining $\mathbf{E}[N_s]$ within $o(1)$.

Lemma 5.10
Case: $\rho < \frac{K-1}{K}$

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

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

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

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 $T \sim \mathrm{Exp}\,(\beta)$ ($\beta = \lambda(1-p) + \mu_\ell$) 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,

$$N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T) \leq_{st} N^{M(\lambda_s)/M(\mu_s)/K-1}(T) \leq_{st} \quad N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T) + (K-1) \tag{9}$$

To see why (9) 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 (9), 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 $\mathbf{E}\left[\Delta_{b'}^2(0)\right]$ 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

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

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 5.11 *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 5.11: See Appendix A.

Substituting in the expressions for μ_s, λ_s and μ_ℓ from Lemma 5.2 and using the inequality (9), we obtain the expressions in the statement of the lemma. ■

Lemma 5.12 *The asymptotics for the first and second moments of $\Delta_{b''}$ are given by:*

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

Proof: See Appendix A. ■

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

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

Lemma 5.13 *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,*

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

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 $\Delta \overset{d}{=} \Delta_b$. Using the expression for $\mathbf{E}[\Delta_b]$ derived in Section 5.5.2, one can check that the condition of Lemma 5.13 is met. Therefore,

$$\mathbf{E}\left[\bar{N}_{s,g}^*\right] = \mathbf{E}\left[N^{M/M/K}\right] + \frac{1}{2}\frac{\lambda(1-p^{(\epsilon)})\mathbf{E}\left[\Delta_b^2\right]}{K\mu - \lambda} + o(1) \tag{10}$$

Substituting $\mathbf{E}\left[\Delta_b^2\right]$ from Section 5.5.2,
Case: $\rho < \frac{K-1}{K}$

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

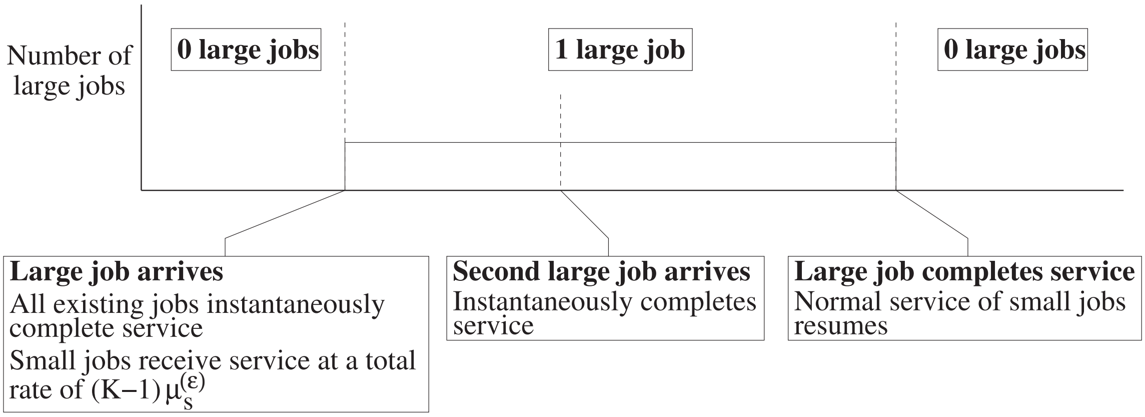

Number of large jobs
0 large jobs
1 large job
0 large jobs
Large job arrives
All existing jobs instantaneously complete service
Small jobs receive service at a total rate of (K−1) μ(ε)s
Second large job arrives
Instantaneously completes service
Large job completes service
Normal service of small jobs resumes

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

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

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

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

Case: $\rho \geq \frac{K-1}{K}$
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 less than $(K-1)$ small jobs in the system during a bad period, they are collectively served at a total rate of $(K-1)\mu_s^{(\epsilon)}$.

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 Exp $\left(\mu_\ell^{(\epsilon)}\right)$ and the length of a good period is distributed as 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.

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, 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 $\frac{1}{\mu_s^{(\epsilon)}}$.

	$C^2 = 19$		$C^2 = 99$	
	$\mathbf{E}[W]$	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 $\rho = 0.9$. All job size distributions have $\mathbf{E}[X] = 1$.

	$C^2 = 19$		$C^2 = 99$	
	$\mathbf{E}[W]$	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 $\rho = 0.6$. All job size distributions have $\mathbf{E}[X] = 1$.

Lemma 5.14 *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. ■

6 Effect of higher moments

In Theorems 1.1 and 1.2, 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 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 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.

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. 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

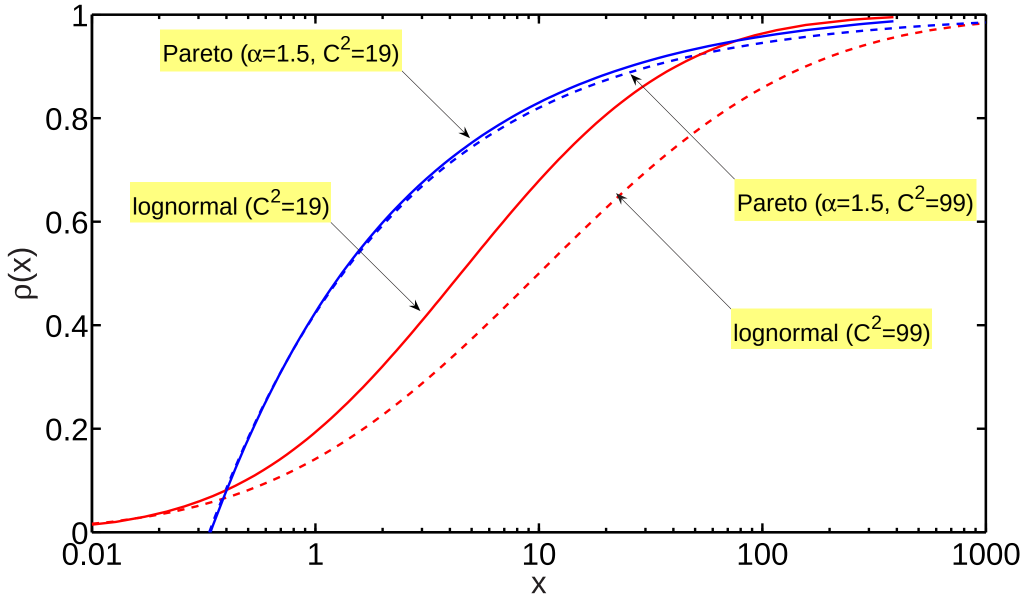


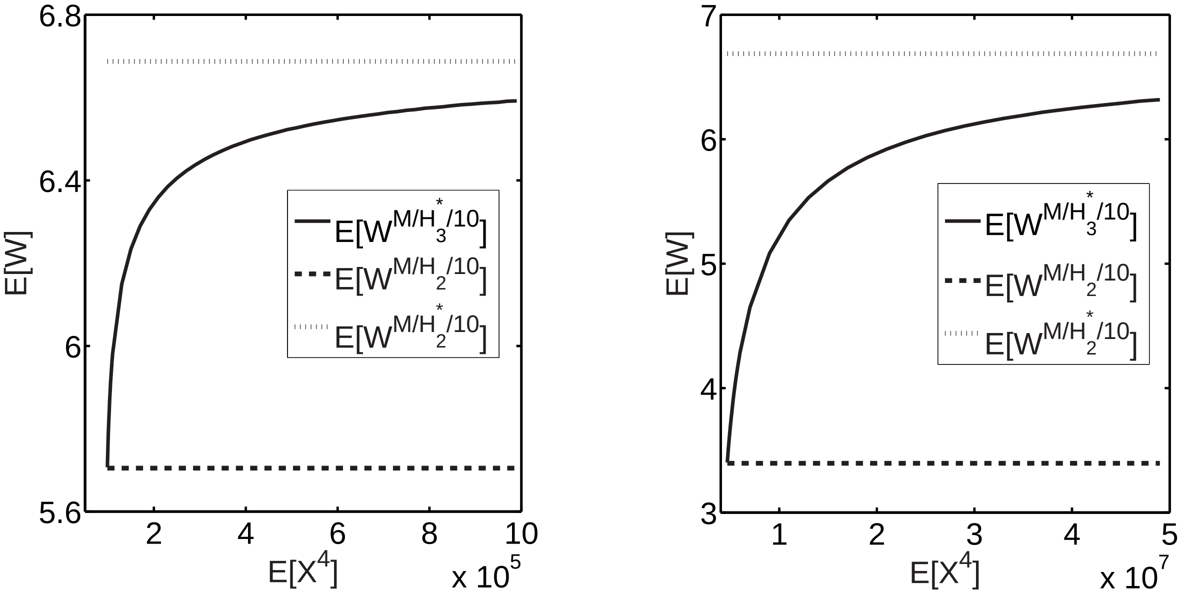
Figure 6: The distribution of load as a function of job size for the lognormal and bounded Pareto ($\alpha = 1.5$) distributions for two values of squared coefficient of variation. Although the lognormal and Pareto distributions have identical first three moments, the distribution of load among different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

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 constituted by larger jobs as compared to the truncated Pareto distribution.

The example of lognormal and Pareto ($\alpha = 1.5$) distributions suggests that even knowledge of 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 mixture of three exponential distributions where mean of one of the phases is 0 (see Definition 3.2). Compared to the H_2 class, the H_3^* class has one more parameter and thus four degrees of freedom, which allow 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. Further, 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



(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).

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. Other characteristics, such as the distribution of load among the small and large job sizes, may lead to more accurate approximations.

7 Conjectures

In this section, we make conjectures on tight bounds on the mean delay of an $M/G/K$ queueing system given first n moments for general n.

7.1 Sharp two-moment bounds

In Theorems 1.1 and 1.2, we proved a lower bound on $W_h^{C^2}$ and an upper bound on $W_l^{C^2}$. Here we make the following conjectures on their exact expressions:

Conjecture 7.1 *For any $\rho < 1$ and finite C^2,*

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

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

Conjecture 7.2 *For any finite C^2,*

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

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

A proof of Conjecture 7.1 might follow these lines: It is easy to prove that $W_h^{C^2} \geq (C^2+1)\mathbf{E}\left[W^{M/D/K}\right]$ by considering the D_2^* distribution (mixture of two point masses with one point mass at 0, see Definition C.2), and following the same argument that we used for the H_2^* distribution in proving Theorem 1.1. However, proving that $W_h^{C^2} \leq (C^2+1)\mathbf{E}\left[W^{M/D/K}\right]$ seems non-trivial, but we provide some justification. First, note that $\mathbf{E}\left[W^{M/D/K}\right] \geq \frac{\mathbf{E}\ W^{M/M/K}}{2}$, and hence the bound in Conjecture 7.1 is indeed tighter than Theorem 1.1. Further, for an $M/G/1$, the mean delay is *exactly linear* in C^2 and one expects the effect of variability to go down as more servers are added. However, we demonstrate a distribution (the D_2^* distribution) which exposes the entire effect of variability - and hence seems to create a worst case scenario. Third, it is known (see Theorem C.3) that given the first two moments, the D_2^* distribution is the unique positive distribution that minimizes all moments higher than the second moment - and therefore extremal.

A proof of Conjecture 7.2 might follow these lines: It should not be too difficult to extend the proof of Theorem 1.2 by defining a $D_2^{(\epsilon)}$ sequence of distributions (parameterized mixture of two point masses, analogous to $H_2^{(\epsilon)}$) to prove that

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

However, proving the tightness of the above bound seems non-trivial.

It is interesting to note that the D_2^* and $D_2^{(\epsilon)}$ distributions were previously used by Whitt [38] as interarrival distributions to obtain extreme values for the mean queue length in the $GI/M/1$ queue.

7.2 Bounds based on higher moments

Just as we have proved (and made stronger conjectures about) the inapproximability of the mean waiting time given the first two moment of the job size distribution by giving the span of the possible values of the mean waiting time, it is useful to know how this span shrinks as we successively know more and more moments. For the third moment, while Figure 1 suggests that within the H_2 class of job size distributions, increasing the third moment causes a drop in the mean waiting time, this statement is too restrictive to be useful. Let $\mathbf{m} = (m_1, m_2, \ldots, m_n) \in \Re^n$ be such that there exists a positive random variable $\mathcal{X}$ with $\mathbf{E}\left[\mathcal{X}^i\right] = m_i,\ i = 1, \ldots, n$. For n odd, define $\mathcal{D}(\mathbf{m})$ to denote the unique $\frac{n+1}{2}$-phase hyperdeterministic distribution (Definition C.1) with moments $(m_1, \ldots, m_n)$. For n even, define $\mathcal{D}^*(\mathbf{m})$ to denote the unique $\left(\frac{n}{2}+1\right)$- phase degenerate hyperdeterministic distribution (Definition C.2) with moments $(m_1, \ldots, m_n)$.

Let

$$W_h(\mathbf{m}) = \sup\left\{\mathbf{E}\left[W^{M/G/K}\right] \left|\mathbf{E}\left[X^i\right] = m_i,\ i = 1, \ldots, n\right.\right\},$$

and

$$W_l(\mathbf{m}) = \inf\left\{\mathbf{E}\left[W^{M/G/K}\right] \left|\mathbf{E}\left[X^i\right] = m_i,\ i = 1, \ldots, n\right.\right\}.$$

We conjecture the following,

Conjecture 7.3 *Let* $\mathbf{m} = (m_1, \ldots, m_n)$*,* $n > 2$*, be a valid moment sequence for positive distributions. Let* $\mathbf{m}' = (m_1, \ldots, m_{n-1})$*. Then,*
Case 1: n ***odd***

(i) $W_h(\mathbf{m}) = \mathbf{E}\left[W^{M/\mathcal{D}^*(\mathbf{m}')/K}\right]$.

(ii) $W_l(\mathbf{m}) = \mathbf{E}\left[W^{M/\mathcal{D}(\mathbf{m})/K}\right]$.

(iii) $W_l(m_1, \ldots, m_{n-1}, x)$ *is strictly decreasing in* x *when* $K > 1$.

Case 2: n ***even***

(i) $W_h(\mathbf{m}) = \mathbf{E}\left[W^{M/\mathcal{D}^*(\mathbf{m})/K}\right]$.

(ii) $W_l(\mathbf{m}) = \mathbf{E}\left[W^{M/\mathcal{D}(\mathbf{m}')/K}\right]$.

(iii) $W_h(m_1, \ldots, m_{n-1}, x)$ *is strictly increasing in* x *when* $K > 1$.

Further, for n *odd,* $W_h(\mathbf{m}) = W_h(\mathbf{m}')$*; and for* n *even,* $W_l(\mathbf{m}) = W_l(\mathbf{m}')$*.*

Implications of Conjectures 7.1, 7.2 and 7.3: Our goal is to estimate $\mathbf{E}\left[W^{M/G/K}\right]$. If we are given only the mean of the job size distribution, we only have enough information to fix a lower bound on $\mathbf{E}\left[W^{M/G/K}\right]$. This lower bound is given by $\mathbf{E}\left[W^{M/D/K}\right]$. Now, if we are told the second moment of the job size distribution, we can fix an upper bound on $\mathbf{E}\left[W^{M/G/K}\right]$. This upper bound is given by $(C^2+1)\mathbf{E}\left[W^{M/D/K}\right]$. (If $\rho \geq \frac{K-1}{K}$ we also refine our lower bound.) By determining the third moment of job size distribution, from Case 1 of Conjecture 7.3, we can *refine* our lower bound to something much tighter (in fact, to $\mathbf{E}\left[W^{M/D_2/K}\right]$) but this *lower bound decreases as the third moment increases*. The upper bound remains unchanged. Therefore, if the θ_3 of the job size distribution is small, the lower bound obtained by considering the first three moments is itself very close to the upper bound (which in turn is close to the approximation in (1)).

Similarly, knowledge of the fourth moment will *refine the upper bound* on the mean waiting time (bring it down), while knowledge of the fifth moment will *refine the lower bound* on the mean waiting time (raise it), and so forth for alternating higher even and odd moments[7]. We conjecture further that these bounds are achieved by the D_n and D_n^* distributions (defined in Definitions C.1 and C.2 as mixtures of point masses, analogous to Definitions 3.1 and 3.2), respectively.

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 moments of 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. We did this by finding the possible range of values for the mean waiting time, given the first two moments of the job size distribution, and showing that the maximum possible value is at least $\left(\frac{C^2+1}{2}\right)$ times the minimum possible value.

Further, we suggest that *moments* are not the ideal job size characteristic on which to base approximations for mean waiting time. Moments of the job size distribution can, at best, provide bounds on the mean waiting time which may be too far to be useful. 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 distributions like the lognormal distribution (all of whose moments are 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

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

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

[7] We may need to impose certain regularity conditions on the job size distribution, such as that the moment sequence uniquely determines the distribution.

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] A.E. Eckberg Jr. Sharp bounds on Laplace-Stieltjes transforms, with applications to various queueing problems. *Math. Oper. Res.*, 2(2):132–142, 1977.

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

[13] 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.

[14] 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.

[15] 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.

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

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

[18] Samuel Karlin and William J. Studden. *Tchebycheff systems: With applications in analysis and statistics*. John Wiley & Sons Interscience Publishers, New York, 1966.

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

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

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

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

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

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

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

[32] 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.

[33] 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.

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

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

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

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

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

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

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

A Proofs

Proof of Claim 4.1: 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:

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

We get 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, the smallest possible value of $\mathbf{E}\left[X^3\right]$ given the first two moments is $\frac{3\mathbf{E}\;X^2\;{}^2}{2\mathbf{E}[X]}$ and is realised 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 5.5: Recall that $N_{\ell,lp}^{(\epsilon)}$ 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. Since this is a geometrically ergodic process, the second moment of the busy period of this queue is 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. [1]. 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. Define $\overline{V_\ell}^{(\epsilon)}$ to be the system time (response time) of large customers in the modified queue. From Adan & Resing [1], we get

$$\mathbf{E}\left[\overline{V_\ell}^{(\epsilon)}\right] = \mathbf{E}[G] + \left(\frac{\rho_G}{1-\rho_G}\right)\frac{\mathbf{E}\left[G^2\right]}{2\mathbf{E}[G]} + \left(\frac{\lambda\mathbf{E}[B_{\lambda,1,K}]}{1+\lambda\mathbf{E}[B_{\lambda,1,K}]}\right)\frac{\mathbf{E}\left[B_{\lambda,1,K}^2\right]}{2\mathbf{E}[B_{\lambda,1,K}]}. \tag{11}$$

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

$$\mathbf{E}[G] = \frac{1}{\mu_\ell^{(\epsilon)}}\left(1 + \frac{\mathbf{E}[B_{\lambda,1,K}]}{\lambda}\right) \tag{12}$$

and that

$$\mathbf{E}\left[G^2\right] = \frac{2}{\left(\mu_\ell^{(\epsilon)}\right)^2}\left(1 + \frac{\mathbf{E}[B_{\lambda,1,K}]}{\lambda}\right)^2 + \frac{1}{\mu_\ell^{(\epsilon)}}\lambda\mathbf{E}\left[B_{\lambda,1,K}^2\right]. \tag{13}$$

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

Proof of Lemma 5.7: 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 already stop serving small jobs. The fraction of time this modified system $U^{(\epsilon)}$ is busy with large jobs is given by $\frac{1-p^{(\epsilon)}}{\mu_\ell^{(\epsilon)}}$. The load of the small jobs is less than ρ. Thus, system $U^{(\epsilon)}$ will be stable if $\rho < 1 - \frac{1-p^{(\epsilon)}}{\mu_\ell^{(\epsilon)}}$.

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

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

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

Now,

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

It is easy to verify that for all $\epsilon < \epsilon'$, the upper bound in the rightmost expression above is smaller than $(1-\rho)$. ■

Proof of Lemma 5.8: 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 $\Psi(A) \leq_{st} A + \Delta_{b'}(0) + \Delta_{b''} \overset{def}{=} A + \Delta_b$. 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:

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

Also,

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

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 5.9: (We suppress the superscript (ϵ) throughout for readability.) Let $N_{s,b'}$ denote the number of 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,

$$N_{s,b'} \leq_{st} \bar{N}^*_{s,g} + \Delta_{b'}(0) \tag{14}$$

Bound for $N_{s,b''}$: To obtain 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 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 $N_{s,b''}$,

$$N_{s,b''} \leq_{st} \bar{N}^*_{s,g} + \Delta_{b'}(0) + A_\lambda\left({T_{b''}}_e\right) \tag{15}$$

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 (14) and (15),

$$N_{s,b} \leq_{st} \bar{N}^*_{s,g} + \Delta_{b'}(0) + \mathbf{I}_{b''|b} A_\lambda\left({T_{b''}}_e\right) \tag{16}$$

■

Proof of Lemma 5.11: The z-transform of $N^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)$ is given by ([13], Theorem 4):

$$\widehat{N}^{M(\lambda_s)/M((K-1)\mu_s)/1}(T)(z) = \frac{\beta z - (K-1)\mu_s(1-z)p_0}{\beta z - ((K-1)\mu_s - \lambda_s z)(1-z)} \tag{17}$$

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 $(1,\infty)$).

By differentiating the transform in (17), we have

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

Factoring the denominator of (17), we can write η as,

$$\eta = 1 + \frac{\beta}{\lambda_s - (K-1)\mu_s} + \Theta(\beta^2)$$

which results in the expressions in the lemma. ■

Proof of Lemma 5.12: Let $\widetilde{T_{b''}}(s)$ denote the Laplace transform for the length of the bad$''$ phase of a bad period. It is easy to show that given $\widetilde{T_{b''}}(s)$, $A_\lambda\widehat{({T_{b''}}_e)}(z)$ is given by:

$$\begin{aligned}
A_\lambda\widehat{({T_{b''}}_e)}(z) &= \frac{1}{\mathbf{E}[T_{b''}]}\int_{t=0}^{\infty} \mathbf{Pr}[T_{b''} \geq t] e^{-\lambda t(1-z)} dt \\
&= \frac{1 - \widetilde{T_{b''}}(\lambda(1-z))}{\lambda(1-z)\mathbf{E}[T_{b''}]}
\end{aligned}$$

The Laplace transform of $T_{b''}$, $\widetilde{T_{b''}}(s)$, is given by

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

where $\widetilde{B}(s)$ is the Laplace transform for the length of busy periods of an $M/M/1$ with arrival rate $\lambda(1-p^{(\epsilon)})$ and service rate $\mu_\ell^{(\epsilon)}$.

Note that the z-transform of $\Delta_{b''}$ is:

$$\widehat{\Delta_{b''}}(z) = \frac{\mu_\ell^{(\epsilon)}}{\mu_\ell^{(\epsilon)} + \lambda(1-p^{(\epsilon)})} + \frac{\lambda(1-p^{(\epsilon)})}{\mu_\ell^{(\epsilon)} + \lambda(1-p^{(\epsilon)})}\widetilde{B}^2(\lambda(1-z))$$

and $\widehat{\Delta_{b''}}(z) \neq \widehat{A\left(T_{b''e}\right)}(z)$ since $T_{b''}$ is not an exponential random variable.

Substituting the values from Section 5.2, we get the following asymptotics which will be sufficient for our purposes:

$$\mathbf{E}[\Delta_{b''}] = O(1) \tag{18}$$

$$\mathbf{E}\left[\Delta_{b''}^2\right] = \Theta\left(\frac{1}{\epsilon}\right) \tag{19}$$

$$\mathbf{E}[A_\lambda\left(T_{b''e}\right)] = \Theta\left(\frac{1}{\epsilon}\right) \tag{20}$$

$$\frac{\mathbf{Pr}[b'']}{\mathbf{Pr}[b]}\mathbf{E}[A_\lambda\left(T_{b''e}\right)] = \frac{\mathbf{E}[T_{b''}]}{\mu_\ell^{(\epsilon)} - \lambda(1-p^{(\epsilon)})}\mathbf{E}[A_\lambda\left(T_{b''e}\right)] \;=\; \Theta(1) \tag{21}$$

■

Proof of Lemma 5.13: 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 $T \sim \text{Exp}\left(\alpha\right)$ 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 $p_i(t)$:

$$\frac{d}{dt}p_0(t) = -\lambda p_0(t) + \mu p_1(t) \tag{22}$$

$$\frac{d}{dt}p_i(t) = \lambda p_{i-1}(t) - (\lambda + i\mu)p_i(t) + (i+1)\mu p_{i+1}(t) \qquad \ldots 1 \leq i \leq K-1 \tag{23}$$

$$\frac{d}{dt}p_i(t) = \lambda p_{i-1}(t) - (\lambda + K\mu)p_i(t) + K\mu p_{i+1}(t) \qquad \ldots i \geq K \tag{24}$$

Let $\widehat{\Pi}(z,t) = \sum_{i=0}^{\infty} p_i(t)z^i$. Using the above differential equations, we have:

$$\frac{\partial}{\partial t}\widehat{\Pi}(z,t) = \widehat{\Pi}(z,t)\left[s\mu\left(\frac{1}{z} - 1\right) + \lambda\left(z-1\right)\right] \tag{25}$$
$$+ \mu\left(1 - \frac{1}{z}\right)\left[Kp_0(t) + (K-1)zp_1(t) + \ldots + z^{K-1}p_{K-1}(t)\right]$$

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:

$$\begin{aligned}\widehat{\Pi}_\alpha(z) = \widehat{\Pi}(z,0) + \frac{\widehat{\Pi}_\alpha(z)}{\alpha}\left[K\mu\left(\frac{1}{z}-1\right) + \lambda\left(z-1\right)\right] \\ + \frac{\mu}{\alpha}\left(1-\frac{1}{z}\right)\left[Kp_{0,\alpha} + (K-1)zp_{1,\alpha} + \ldots + z^{K-1}p_{K-1,\alpha}\right]\end{aligned}$$

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:

$$\widehat{N^{(Int)}}(z) = \frac{\mu\left[Kp_0 + (K-1)zp_1 + \ldots + z^{K-1}p_{K-1}\right]}{(K\mu - \lambda z) - \alpha z\left(\frac{1-\widehat{\Delta}(z)}{1-z}\right)} \tag{26}$$

Since $\widehat{N^{(Int)}}(1) = 1$, we get

$$Kp_0 + (K-1)p_1 + \ldots + p_{K-1} = K - \frac{\lambda}{\mu} - \frac{\alpha}{\mu}\mathbf{E}[\Delta] \tag{27}$$

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

$$\mathbf{E}\left[N^{(Int)}\right] = \frac{d}{dz}\widehat{N^{(Int)}}(z)\Big|_{z=1} \tag{28}$$

$$= \frac{\mu C}{K\mu - \lambda - \alpha\mathbf{E}[\Delta]} + \frac{\lambda + \frac{\alpha}{2}\left(\mathbf{E}\left[\Delta^2\right] + 3\mathbf{E}[\Delta]\right)}{K\mu - \lambda - \alpha\mathbf{E}[\Delta]} \tag{29}$$

where,

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

To calculate C we need the following relations obtained from integrating by parts the differential equations (22)-(23):

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

which yields $p_{i,\alpha} = p_{0,\alpha}\frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i + o(\alpha)$. Combining with (27) and the assumption that $\mathbf{E}[\Delta] = o\left(\frac{1}{\alpha}\right)$, we get $p_i = \pi_i + o(1)$ for $i \le K$, where π_i are the stationary probabilities of an $M/M/K$ system with arrival rate λ and mean job size $\frac{1}{\mu}$. Using this, we have:

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

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, set $\Delta \equiv 0$.) 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 1.3

The proof is trivial for $\rho < (K-1)/K$. For $\rho \geq (K-1)/K$, the inequality $\underline{W_h^{C^2}} > \overline{W_l^{C^2}}$ is equivalent to

$$\frac{C^2-1}{2}\mathbf{E}\Big[W^{M/M/K}\Big] > \frac{1}{1-\rho}\left[\rho - \frac{K-1}{K}\right]\frac{C^2-1}{2}. \tag{30}$$

Recall that we still take $\mathbf{E}[X] = 1$ without loss of generality so that $\rho \geq K/(K-1)$ 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

$$\mathbf{E}\Big[W^{M/M/K}\Big] = \frac{C(K,\lambda)}{K-\lambda}. \tag{31}$$

Therefore, (30) holds if (we have assumed $\frac{C^2-1}{2} > 0$)

$$C(K,\lambda) > [\lambda - (K-1)]\,. \tag{32}$$

It is known that $C(K,\lambda)$ is a strictly convex function in λ on $[0,K]$ (see [23]). Since (32) trivially holds for $\lambda = K-1$, and since the right hand side of (32) has derivative (w.r.t. λ) 1, it suffices to show that

$$\frac{d}{d\lambda}C(K,\lambda)|_{\lambda=K} < 1. \tag{33}$$

Let A_λ be a random variable that is Poisson with rate λ. It is well known that

$$C(K,\lambda) = \frac{1}{\rho + (1-\rho)\frac{P(A_\lambda \leq K)}{P(A_\lambda = K)}}. \tag{34}$$

Using this expression, we find that

$$\frac{d}{d\lambda}C(K,\lambda)|_{\lambda=K} = \frac{1}{K}\frac{P(A_K \leq K-1)}{P(A_K = K)} = \frac{1}{K}\sum_{k=0}^{K-1}\frac{P(A_K = k)}{P(A_K = K)}. \tag{35}$$

Now, note that

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

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

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

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

$$\frac{d}{d\lambda}C(K,\lambda)|_{\lambda=K} = \frac{1}{K}\sum_{k=0}^{K-1}\frac{K^k/k!}{K^K/K!} < 1, \tag{36}$$

which completes the proof of the proposition.

C Hyperdeterministic distributions and their extremal properties

In this section we discuss the utility of hyperdeterministic distributions in obtaining bounds on various metrics based on moments or other partial information of the random variable involved via the theory of Tchebycheff systems. The following discussion is borrowed from the work of Eckberg [11] who applied the theory of Tchebycheff systems to queueing problems. A full treatment of appears in [18].

C.1 Definitions

We first define the n-phase hyperdeterministic distribution, D_n, and the n-phase degenerate hyperdeterministic distribution, D_n^*, in Definitions C.1 and C.2, respectively.

Definition C.1 *Let $0 \leq x_1 < x_2 \ldots < x_n$. Let $p_i > 0$, $i = 1, \ldots, n$, be such that $\sum_{i=1}^{n} p_i = 1$. We define the $n-$phase hyperdeterministic distribution, D_n, with parameters x_i, p_i, $i = 1, \ldots, n$, as:*

$$D_n \sim \begin{cases} x_1 & \textit{with probability } p_1 \\ x_2 & \textit{with probability } p_2 \\ \vdots & \\ x_n & \textit{with probability } p_n. \end{cases}$$

Definition C.2 *Let $0 < x_1 < x_2 \ldots < x_{n-1}$. 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 hyperdeterministic distribution, D_n^*, with parameters p_0, x_i, p_i, $i = 1, \ldots, n-1$, as:*

$$D_n^* \sim \begin{cases} 0 & \textit{with probability } p_0 \\ x_1 & \textit{with probability } p_1 \\ \vdots & \\ x_{n-1} & \textit{with probability } p_{n-1}. \end{cases}$$

C.2 Tchebycheff inequalities and principal representations

The area of Tchebycheff inequalities is concerned with solving problems of the following kind: We are given a partial characterization of a random variable X in terms of generalized moment constraints:

$$\mathbf{Pr}[0 \leq X \leq B] = 1 \tag{37}$$

$$\mathbf{E}[g_i(X)] = m_i, \quad 1 \leq i \leq n. \tag{38}$$

Let $\mathcal{T} = \{X | X \text{ satisfies (37) and (38)}\}$. Given another function f, we wish to determine the bounds

$$\beta_l = \inf\{\mathbf{E}[f(X)] | X \in \mathcal{T}\},$$
$$\beta_u = \inf\{\mathbf{E}[f(X)] | X \in \mathcal{T}\}.$$

Define the function $g_0(x) = 1, 0 \leq x \leq B$, and denote the moment space associated with $\{g_0, g_1, \ldots, g_n\}$ as

$$\mathcal{M}^{n+1} = \left\{ \mathbf{c} \in \Re^{n+1} \mid c_i = \int_0^B g_i(u) d\mu(u),\ 0 \leq i \leq n,\ \text{for some } \mu \in \mathcal{D} \right\}$$

where $\mathcal{D}$ is the set of all non-decreasing right continuous functions for which the indicated integrals exist. For a point $\mathbf{c^0}$ in the interior of $\mathcal{M}^{n+1}$, we define the *unique lower and upper principal representation (pr)* as follows:

	Upper pr ($\bar{\mu}$)	Lower pr ($\underline{\mu}$)
n even	$n/2$ mass points in $(0, B)$, one at B	$n/2$ mass points in $(0, B)$, one at 0
n odd	$(n-1)/2$ mass points in $(0, B)$, one at 0, one at B	$(n+1)/2$ mass points in $(0, B)$

We say that functions $\{g_0, g_1, \ldots, g_n\}$ form a Tchebycheff system over $[a, b]$ provided the determinants

$$U\left(\begin{array}{c} 0, 1, \cdots, n \\ x_0, x_1, \cdots, x_n \end{array}\right) = \left|\begin{array}{cccc} g_0(x_0) & g_0(x_1) & \cdots & g_0(x_n) \\ g_1(x_0) & g_1(x_1) & \cdots & g_1(x_n) \\ \vdots & \vdots & & \vdots \\ g_n(x_0) & g_n(x_1) & \cdots & g_n(x_n) \end{array}\right|$$

are strictly positive whenever $a \le x_0 < x_1 < \cdots < x_n \le b$. The functions $g_0, g_1, \ldots, g_n$ are referred to as a complete Tchebycheff system if $\{g_0, g_1, \cdots, g_r\}$ is a Tchebycheff system for each $r = 0, 1, \cdots, n$. The following theorem describes the random variables that attain the extremal values β_l and β_u:

Theorem C.3 *(Markov-Krein) If $\{g_0, g_1, \ldots, g_n\}$ and $\{g_0, g_1, \ldots, g_n, f\}$ are Tchebycheff systems on $[0, B]$, then*

$$\beta_l = \int f(u) d\underline{\mu}(u)$$
$$\beta_u = \int f(u) d\bar{\mu}(u),$$

where $\underline{\mu}$ and $\bar{\mu}$ are the unique lower and upper principal representations, respectively, of $\mathbf{c} = \{1, m_1, \ldots, m_n\}$.

Note that the upper and lower principal representations belong to the classes D_n or D_n^* for some n. The Markov-Krein Theorem shows that for a large family of moment constraints, and in particular given a few raw moments, random variables with hyperdeterministic distribution maximize or minimize the expected values of a large class of functions. Statements of the form similar to conjectures presented in Section 7 for an $M/G/K$ system with a partial characterization of the job size distribution can be proven for $GI/M/1$ systems with partial characterization of the interarrival distribution by considering $f(x) = e^{-sx}$ (see Whitt [38]). The applicability of Tchebycheff systems in verifying Conjecture 7.3 is a potentially interesting research direction.