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

Varun Gupta
Carnegie Mellon University
varun@cs.cmu.edu

Mor Harchol-Balter [*]
Carnegie Mellon University
harchol@cs.cmu.edu

Jim Dai
Georgia Institute of Technology
dai@gatech.edu

Bert Zwart
Georgia Institute of Technology
bertzwart@gatech.edu

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Queueing Theory;
C.4 [**Performance of Systems**]: Performance attributes

General Terms

Performance

1. INTRODUCTION

The $M/G/s$ queueing system is the oldest and most classical example of multiserver systems. Such multiserver 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 $M/G/s$ system consists of s identical servers and a First-Come-First-Serve (FCFS) queue. The jobs (or customers) arrive according to a Poisson process with rate λ and their service requirements (job sizes) are assumed to be independent and identically distributed according to a random variable S. 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 $M/G/s$ system as $\rho = \frac{\lambda \mathbf{E}[S]}{s}$. We will focus on the metric of mean waiting time in this work, denoted as $\mathbf{E}[T_Q]$ and defined as the time from the arrival of a customer to the time it enters service.

Even though the the $M/G/s$ queue has received a lot of attention in the the queueing literature, an exact analysis for even simple metrics like mean waiting time for the case $s \geq 2$ still eludes researchers. To the best of our knowledge, the first approximation for the mean waiting time for an $M/G/s$ queue was given by Lee and Longton [1] nearly half a century ago:

$$\mathbf{E}\left[T_Q^{M/G/s}\right] \approx \left(\frac{C^2+1}{2}\right)\mathbf{E}\left[T_Q^{M/M/s}\right] \qquad (1)$$

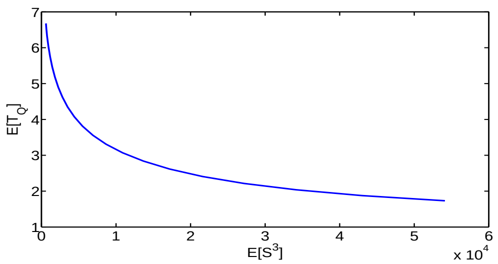


Figure 1: Illustration of the effect of 3rd moment of the job size distribution on mean waiting time of an $M/H_2/10$ system. The parameters of the job size distribution were held constant at $\mathbf{E}[S] = 1$ and $C^2 = 19$ with load $\rho = 0.9$.

where $\mathbf{E}\left[T_Q^{M/M/s}\right]$ is the mean waiting time with exponentially distributed job sizes with the same mean, $\mathbf{E}[S]$, as in the $M/G/s$ system and C^2 is the squared coefficient of variation[1] (SCV) of S. This approximation is very simple, involving only the first two moments, $\mathbf{E}[S]$ and $\mathbf{E}\left[S^2\right]$, of the job size distribution and is exact for a certain class of distributions. Many other authors have also proposed simple approximations for the mean waiting time, but all such closed-form approximations involve only the first two moments of the job size distribution.

In this work, we propose using the third moment of the job size distribution ($\mathbf{E}\left[S^3\right]$) to help in understanding the mean waiting time of an $M/G/s$ system. Contrary to existing approximations, we show that not only does the third moment impact the mean waiting time but it has a huge effect! Moreover, we find that a higher third moment can cause the mean waiting time to drop, and can potentially nullify the effect of variability if the load is not very high. However, the fourth moment influences the mean waiting time in the opposite manner. We make these observations by restricting our attention to subsets of the hyperexponential distributions which also lead to insights into the observed behavior and naturally lead to approximations for $\mathbf{E}[T_Q]$ involving the first three (and even four) moments of the job size distribution.

2. INCREASING THE THIRD MOMENT DECREASES MEAN WAITING TIME

[*] Supported by NSF Career Grant CCR-0133077, NSF Theory CCR-0311383, NSF ITR CCR-0313148.

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

We begin by numerically evaluating an $M/G/s$ system for job size distributions with the same mean and second moment but varying third moment. For this, we choose the job size distributions from the family of 2-phase hyperexponential (H_2) distributions defined below.

Definition 1. A random variable S distributed according to the H_2 distribution with parameters p, μ_1 and μ_2 is given by:

$$S = \begin{cases} \exp(\mu_1), & \text{with probability } p, \\ \exp(\mu_2), & \text{with probability } 1-p. \end{cases}$$

In Figure 1, we show the mean waiting time of an $M/H_2/10$ system as a function of the 3rd moment. The mean of the job size distribution was fixed to $\mathbf{E}[S] = 1$, SCV to $C^2 = 19$ and system load to $\rho = 0.9$. As can be seen in the figure, the mean waiting time drops by more than a factor of 3 as the 3rd moment is increased. Based on this observation, and a great many similar graphs not shown, we make the following conjecture:

CONJECTURE 1. *For a given finite $s \geq 1$, load $\rho < 1$, mean service requirement $\mathbf{E}[S]$ and SCV C^2, the mean waiting time of an $M/H_2/s$ system decreases as the third moment of the job size distribution increases.*

We provide some intuition for why our conjecture should be true: A system with an H_2 job size distribution is equivalent to a system with two job classes. Class 1 jobs are, say, small ones with mean $1/\mu_1$, and class 2 jobs are big ones with mean $1/\mu_2$. As the third moment increases, the size of the big jobs increases but their arrival rate goes down. Further, the load of these jobs can be shown to converge to 0. In fact, the H_2 distribution converges in distribution to an exponential distribution as the third moment approaches ∞. Therefore, the probability that there is one big job in system decreases as the third moment increases. If the load is not very high, the occupancy of one server by this big job still leaves $(s-1)$ servers which can handle the load of the small jobs. This reduces the overall effect of variability causing a decrease of the overall mean waiting time. In an $M/G/1$, while again the probability that there is one big job in the system decreases, this event creates an overloaded system causing large buildup of queues. Therefore, the mean waiting time in an $M/G/1$ system exhibits insensitivity to the third moment of job size distribution.

3. BOUNDS ON THE IMPACT OF THIRD MOMENT

While it is possible to give good approximations for the mean waiting time involving the first three moments of job size distribution by concentrating on a certain class of distributions, these may not be accurate for all classes of distributions. Rather than attempting to find one approximation that works for all classes of distributions, we instead turn to the idea of looking at extremes to provide strong guarantees on the mean waiting time as a function of the first two moments. We look at the entire space of distributions with given first two moments, and find job size distributions which would maximize or minimize the mean waiting time. This gives us the upper bound (Theorem 1) and lower bound (Theorem 2) on the mean waiting time of an $M/G/s$ system, thus quantifying the effect of the third moment, given the first two moments of job size distribution. While these bounds are stated as inequalities, and hence only imply a lower bound on the upper bound and an upper bound on the lower bound, we make stronger conjectures later.

THEOREM 1. *Let $\{G|C^2\}$ be the set of positive distributions with mean $\mathbf{E}[S]$ and SCV C^2. For any given number of servers s and load $\rho < 1$, we have the following lower bound on the upper bound for the mean waiting time in an $M/G/s$ system (with equality applying for $s = 1$):*

$$\sup_{\{G|C^2\}} \mathbf{E}\left[T_Q^{M/G/s}\right] \geq \left(\frac{C^2+1}{2}\right) \mathbf{E}\left[T_Q^{M/M/s}\right]$$

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

The bound in Theorem 1 is obtained by looking at the H_2 distribution with $\mu_1 \to \infty$.

THEOREM 2. *Let $\{G|C^2\}$ be the set of positive distributions with mean $\mathbf{E}[S]$ and SCV C^2. For any given s and load $\rho < 1$, we have the following upper bound on the lower bound for the mean waiting time in an $M/G/s$ system (with equality applying for $s = 1$):*

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

$$\inf_{\{G|C^2\}} \mathbf{E}\left[T_Q^{M/G/s}\right] \leq \mathbf{E}\left[T_Q^{M/M/s}\right]$$

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

$$\inf_{\{G|C^2\}} \mathbf{E}\left[T_Q^{M/G/s}\right] \leq \mathbf{E}\left[T_Q^{M/M/s}\right] + \frac{\mathbf{E}[S]}{1-\rho}\left[\rho - \frac{s-1}{s}\right]\frac{C^2-1}{2}$$

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

The bound in Theorem 2 is obtained by looking at the H_2 distribution with $\mu_2 \to 0$.

Note that if the load is not too large, the ratio of the upper bound to the lower bound grows linearly in C^2. Hence the impact of the third moment is accentuated by the SCV of the job size distribution. While Theorems 1 and 2 only give us a lowerbound on the range of $\mathbf{E}\left[T_Q^{M/G/s}\right]$ given the first two moments of the job size distribution, we conjecture the following strict bounds.

CONJECTURE 2. *For any given number of servers s and load $\rho < 1$, we have the following upper bound for the mean waiting time in an $M/G/s$ system:*

$$\sup_{\{G|C^2\}} \mathbf{E}\left[T_Q^{M/G/s}\right] = \left(C^2+1\right) \mathbf{E}\left[T_Q^{M/D/s}\right]$$

where $\mathbf{E}\left[T_Q^{M/D/s}\right]$ is the mean waiting time when the job size distribution is deterministic with mean $\mathbf{E}[S]$.

CONJECTURE 3. *For any given s and load $\rho < 1$, we have the following lower bound for the mean waiting time in an $M/G/s$ system:*

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

$$\inf_{\{G|C^2\}} \mathbf{E}\left[T_Q^{M/G/s}\right] = \mathbf{E}\left[T_Q^{M/D/s}\right]$$

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

$$\inf_{\{G|C^2\}} \mathbf{E}\left[T_Q^{M/G/s}\right] = \mathbf{E}\left[T_Q^{M/D/s}\right] + \frac{\mathbf{E}[S]}{1-\rho}\left[\rho - \frac{s-1}{s}\right]\frac{C^2}{2}$$

where $\mathbf{E}\left[T_Q^{M/D/s}\right]$ *is the mean waiting time when the job size distribution is deterministic with mean* $\mathbf{E}[S]$.

4. EFFECT OF HIGHER MOMENTS

Given the large effect of the third moment of the job size distribution on mean waiting time, it is only natural to ask the following questions: Do even higher moments of the job size distribution have an equally great impact? Is the qualitative effect of the 4th and higher moments similar to the effect of 3rd moment or is it the opposite? In this section we touch upon these interesting, and largely open, questions by observing the effect of $\mathbf{E}[S^4]$ on $\mathbf{E}[T_Q]$, again by looking at a simple class of distributions.

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 distributions.

Definition 2. A random variable X distributed according to the H_3^* distribution with parameters p_0, p_1, μ_1 and μ_2 is given by:

$$X = \begin{cases} 0 & w.p.\ p_0 \\ exp(\mu_1) & w.p.\ p_1 \\ exp(\mu_2) & w.p.\ 1-p_0-p_1 \end{cases}$$

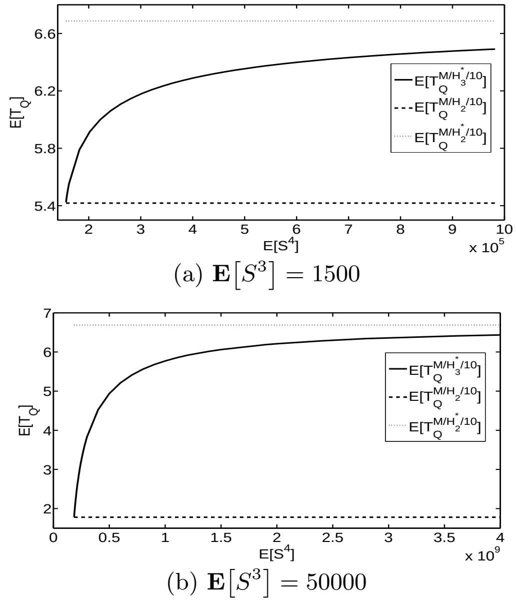
Note that the H_3^* class of distributions has one more parameter than the H_2 class, which allows us control over the 4th moment while holding the first three moments fixed. The subset of H_3^* where $p_0 + p_1 = 1$ is simply called the *degenerate hyperexponential* distribution and denoted as H_2^*.

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, $\mathbf{E}[S] = 1$ and $C^2 = 19$. We choose two values of $\mathbf{E}[S^3]$ and plot the $\mathbf{E}[T_Q]$ curves as a function of $\mathbf{E}[S^4]$ in Figure 2. 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 the H_2^* distribution (with the same first two moments as H_3^*).

As is evident from Figure 2, the fourth moment can have an equally significant impact on the mean waiting time. Moreover, (1) as the fourth moment is increased, the mean waiting time increases, reaching $\mathbf{E}\left[T_Q^{M/H_2^*/s}\right]$ and (2) as the fourth moment is decreased, the mean waiting time drops, reaching $\mathbf{E}\left[T_Q^{M/H_2/s}\right]$.

The following Theorem explains the approach of mean waiting time under the H_3^* job size distribution to that under an H_2^* job size distribution as fourth moment becomes high.

Theorem 3. *Let X_ϵ be a family of random variables distributed according to the H_3^* distribution with a fixed given mean, γ_1, second moment, γ_2, and third moment, γ_3, where $\mathbf{E}[X_\epsilon^4] = 1/\epsilon$. As $\epsilon \to 0$, the distribution of X_ϵ converges to an H_2^* distribution with mean γ_1 and γ_2.*



(a) $\mathbf{E}[S^3] = 1500$

(b) $\mathbf{E}[S^3] = 50000$

Figure 2: 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 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 $\mathbf{E}[S] = 1$ and $C^2 = 19$ with load $\rho = 0.9$.

Further, it can be shown that among all hyperexponential distributions with some given first two moments, the H_2^* distribution has the lowest third moment. As we have observed in Section 2, in this case the effect of variability of the job size distribution is most pronounced causing a higher mean waiting time. Thus a high fourth moment *nullifies* the putative reduction in mean waiting time caused by a high third moment.

To explain the drop in mean waiting time under the H_3^* job size distribution to that of an H_2 job size distribution as the fourth moment becomes small, recall that in Section 2, while looking at H_2 job size distributions, we saw that high third moment caused an appreciable drop in the mean waiting time. Indeed, as Theorem 4 states, when we fix the first three moments of an H_3^* distribution and set the fourth moment to its minimum possible value, the distribution looks like an H_2 distribution thus exposing the benefits of a higher third moment on the mean waiting time.

Theorem 4. *Among all distributions in the class of hyperexponential distributions with given first three moments, the distribution belonging to the class of 2-phase hyperexponential distributions, H_2, has the lowest 4th moment.*

5. REFERENCES

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