

# Tight Moments-Based Bounds for Queueing Systems

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

Takayuki Osogami  
IBM Research, Tokyo  
osogami@jp.ibm.com

## ABSTRACT

We present a new tool to analyze for three queueing systems which have defied exact analysis so far: (i) the classical  $M/G/k$  multi-server system, (ii) queueing systems with fluctuating arrival and service rates, and (iii) the  $M/G/1$  round-robin queue. We argue that rather than looking for exact expressions for the mean response time as a function of the job size distribution, a more fruitful approach is to find distributions which minimize or maximize the mean response time given the first  $n$  moments of the job size distribution.

We prove that for the  $M/G/k$  system in light traffic, and given  $n=2$  and 3 moments, these ‘extremal’ distributions are given by *principal representations* of the moment sequence. Furthermore, if we restrict the distributions to lie in the class of Completely Monotone (CM) distributions, then for all the three queueing systems, for any  $n$ , the extremal distributions under the appropriate “light traffic” asymptotics are hyper-exponential distributions with finite number of phases. We conjecture that the property of *extremality* should be invariant to the system load, and thus our light traffic results should hold for general load as well.

## Categories and Subject Descriptors

G.3 [Probability and Statistics]: Queueing Theory; G.1.2 [Numerical Analysis]: Approximation—*Chebyshev approximation and theory*

## General Terms

Theory, Performance

## Keywords

Moment-based bounds,  $M/G/k$ , Time-varying load, Round-robin, Markov-Krein Theorem, Tchebycheff systems, Light traffic analysis

## 1. INTRODUCTION

Most results in queueing theory are concerned with obtaining explicit expressions for the performance metric of interest (e.g., mean response time) as a function of the distribution of some system parameter (e.g., job size distribution) under suitable assumptions to make the analysis tractable. However, there are many fundamental queueing systems for which such explicit results are not possible. In the absence of exact results, various approximations or bounds are used.

Rather than trying to obtain explicit expressions for the performance metric as a function of the job size distribution, or obtaining

approximations/bounds as functions of some moments of the job size distribution for which no tightness guarantees can be proved, we argue that a more fruitful approach is the following: We first obtain a partial characterization of the job size distribution, say, in terms of the first  $n$  moments. We then look at the set of all distributions which satisfy this partial characterization, and identify those distributions in this set that maximize or minimize the performance metric of interest. Once these extremal distributions are identified, numerical algorithms can be used to obtain **provably tight bounds** on the performance.

In this paper, we take the first step towards obtaining tight bounds on the mean response time of the three queueing systems by analytically investigating suitable asymptotic regimes where the effect of the entire distribution of the system parameter of interest is evident (unlike heavy-traffic asymptotes). Next, rather than using the asymptotic approximations to obtain quantitative behavior (by extrapolating to non-asymptotic regime), we extract qualitative properties by identifying distributions which minimize or maximize the performance metric in the asymptotic regime.

## 2. PRINCIPAL REPRESENTATIONS, AND THE MARKOV-KREIN THEOREM

In this section we will be concerned with random variables with support on  $[0, B]$ . We first introduce the notion of upper and lower principal representations as presented in [2]. Define the function  $f_0(x) = 1, 0 \leq x \leq B$ , and denote the moment space associated with  $\{f_0, f_1, \dots, f_n\}$  as

$$\mathcal{M}^{n+1} = \left\{ \mathbf{m} \in \mathfrak{R}^{n+1} \mid \exists \mu \in \mathcal{D}, m_i = \int_0^B f_i(u) d\mu(u), 0 \leq i \leq n \right\}$$

where  $\mathcal{D}$  is the set of all non-decreasing right continuous functions for which the indicated integrals exist. For a point  $\mathbf{m}^0$  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	$\frac{n}{2}$ mass points in $(0, B)$ , one at $B$	$\frac{n}{2}$ mass points in $(0, B)$ , one at $0$
$n$ odd	$\frac{n-1}{2}$ mass points in $(0, B)$ , one at $0$ , one at $B$	$\frac{n+1}{2}$ mass points in $(0, B)$

We say that functions  $\{h_0, h_1, \dots, h_n\}$  form a Tchebycheff system over  $[a, b]$  provided the determinants

$$U \begin{pmatrix} 0, 1, \dots, n \\ x_0, x_1, \dots, x_n \end{pmatrix} = \begin{vmatrix} h_0(x_0) & h_0(x_1) & \cdots & h_0(x_n) \\ h_1(x_0) & h_1(x_1) & \cdots & h_1(x_n) \\ \vdots & \vdots & & \vdots \\ h_n(x_0) & h_n(x_1) & \cdots & h_n(x_n) \end{vmatrix}$$

are strictly positive whenever  $a \leq x_0 < x_1 < \dots < x_n \leq b$ .

The proof of the following theorem can be found in [4, Chpt. V, Sec. 5]:

**THEOREM 1 (MARKOV-KREIN).** *If  $\{f_0, \dots, f_n\}$  and  $\{f_0, \dots, f_n, g\}$  are Tchebycheff systems on  $[0, B]$ , then*

$$\beta_l \equiv \inf_{\mu_X \in \mathcal{D}} \{\mathbf{E}[g(X)] \mid \Pr[X \in [0, B]] = 1\};$$

$$\mathbf{E}[f_i(X)] = m_i, \quad 0 \leq i \leq n = \int_0^B g(u) d\mu(u)$$

$$\beta_u \equiv \sup_{\mu_X \in \mathcal{D}} \{\mathbf{E}[g(X)] \mid \Pr[X \in [0, B]] = 1\};$$

$$\mathbf{E}[f_i(X)] = m_i, \quad 0 \leq i \leq n = \int_0^B g(u) d\bar{\mu}(u),$$

where  $\mu$  and  $\bar{\mu}$  are the unique lower and upper pr's, respectively, of  $\mathbf{m} = \{1, m_1, \dots, m_n\}$ , and  $\mu_X$  denotes the measure induced by  $X$  on  $\mathfrak{R}$ .

The theorem holds for  $B \rightarrow \infty$  when the corresponding limits exist (see [4]).

**Principal representations within Hyperexponential distributions:** Analogous to the above, we can define pr's within hyperexponential distributions by applying the above theorem to the *spectral density*. We omit the details due to lack of space.

### 3. SUMMARY OF RESULTS

We now briefly describe the three queueing systems, the ‘‘light traffic’’ regime we look at, and our results.

#### The $M/G/k$ multi-server system

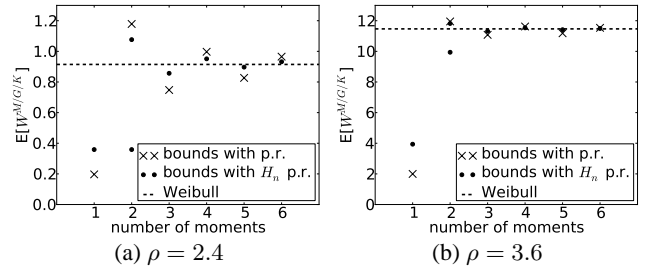
Recall that an  $M/G/k$  system consists of  $k$  identical servers and a FCFS queue. The arrival process is Poisson with rate  $\lambda$ , and the job sizes are assumed to be *i.i.d.* random variables. We will use  $X$  to denote such a generic random variable. We are interested in obtaining bounds on the mean waiting time,  $\mathbf{E}[W^{M/G/k}]$ , as a function of the job size distribution  $X$ . We let the arrival rate  $\lambda \rightarrow 0$ , and look at  $\mathbf{E}[W^{M/G/k}]$  of a random arrival. By exploiting the light-traffic approximation developed by Burman and Smith [1], we can prove the following:

**THEOREM 2.** *Given the first  $n$  ( $n = 2$  or  $3$ ) moments of the job size distribution  $X$ ,  $\mathbf{E}[W^{M/G/k}]$  under light traffic asymptote is extremized by service distributions given by the lower and upper principal representations of the moment sequence.*

**THEOREM 3.** *If the job size distribution is constrained to lie in the CM class, then given the first  $n$  moments of the job size distribution  $X$ ,  $\mathbf{E}[W^{M/G/k}]$  under light traffic is extremized by the lower and upper principal representations of the moment sequence within the hyperexponential class of distributions.*

We conjecture that the above theorems hold for general arrival rates, because, intuitively, increasing the arrival rate to an  $M/G/k$  system should not change the relative performance of two job size distributions.

Finally, we illustrate the utility of our results by presenting numerical results that demonstrate that while two moments of the job size distribution are insufficient for approximating  $\mathbf{E}[W^{M/G/k}]$  for real world heavy-tailed distributions, three moments usually suffice, especially if we add the knowledge of complete monotonicity. Figure 1 shows  $\mathbf{E}[W^{M/G/k}]$  and its bounds obtained with principal representations, when the job size distribution is a Weibull distribution. Notice that the Weibull distribution under consideration is completely monotonic (see [3]), so that a principal representations within hyperexponential distributions give proper bounds.



**Figure 1: Bounding mean delay in an  $M/G/4$  queue when the job size has a Weibull distribution.**

#### The $M/G/1$ round-robin queue

The  $M/G/1$  round-robin queue consists of a single server and an infinite buffer. The arrival process is Poisson with rate  $\lambda$ , and new arrivals join the back of the buffer. Job sizes are assumed to be *i.i.d.*, with  $X$  used to denote a generic job size. Jobs are given  $Q$  units of service at a time (called the quantum size), which for analytical simplicity we assume to be *i.i.d.* samples from an  $\text{Exp}(\nu)$  distribution. We will be interested in obtaining bounds on the mean response time,  $\mathbf{E}[T^{M/G/1/RR}]$ , in terms of moments of  $X$ . We let the arrival rate  $\lambda \rightarrow 0$ , and look at the coefficient of  $\Theta(\lambda)$  in the expression for  $\mathbf{E}[T^{M/G/1/RR}]$ . By deriving the first light-traffic asymptote of  $T^{M/G/1/RR}$  we can prove the following:

**THEOREM 4.** *Given the first  $n$  moments of the job size distribution  $X$  in the CM class,  $\mathbf{E}[T^{M/G/1/RR}]$  under light traffic is extremized by the lower and upper principal representations of the moment sequence within the class of hyperexponential distributions.*

#### Systems with fluctuating arrival and service rates

We analyze an  $M/M/1$  system whose arrival and service rates are controlled by an exogenous environment process with two states: L and H. The durations of stay in the L state during each visit are *i.i.d.* random variables with general distribution; we use  $\tau_L$  to denote such a generic random variable. Similarly, we use  $\tau_H$  to denote a generic random variable for the duration of stay in the H states during each visit. We will be interested in obtaining bounds on the mean number of jobs,  $\mathbf{E}[N]$ , in terms of moments of  $\tau_L$  and  $\tau_H$ . We consider the ‘‘fast-switching’’ asymptote, where we scale  $\tau_L$  and  $\tau_H$  by a parameter  $\alpha$ , and let  $\alpha \rightarrow 0$ . Via a new asymptotic expansion for  $\mathbf{E}[N]$  in terms of  $\alpha$ , we prove the following:

**THEOREM 5.** *If  $\tau_L$  and  $\tau_H$  are constrained to lie in the CM class, then given the first  $n$  moments of  $\tau_L$  and  $\tau_H$ , the mean number of jobs,  $\mathbf{E}[N]$ , under the fast switching asymptote is extremized by the lower and upper principal representations of the moment sequence within the hyperexponential distribution.*

### 4. REFERENCES

- [1] D. Burman and D. Smith. A light-traffic theorem for multi-server queues. *Math. Oper. Res.*, 8:15–25, 1983.
- [2] A. Eckberg Jr. Sharp bounds on Laplace-Stieltjes transforms, with applications to various queueing problems. *Math. Oper. Res.*, 2(2):132–142, 1977.
- [3] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31:245–279, 1998.
- [4] S. Karlin and W. J. Studden. *Tchebycheff systems: With applications in analysis and statistics*. John Wiley & Sons Interscience Publishers, New York, 1966.