On Markov-Krein Characterization of Mean Sojourn Time in Queueing Systems

Varun Gupta[*]
Takayuki Osogami[†]

April 2011
CMU-CS-11-109

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[*]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[†]IBM Research - Tokyo, JAPAN

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

Abstract

We present a new analytical tool 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 asymptote and given first n $(= 2,\ 3)$ moments of the job size distribution, analogous to the classical Markov-Krein Theorem, these 'extremal' distributions are given by the *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, and propose potential strategies for a unified approach to finding moments-based bounds for queueing systems. By identifying the extremal distributions, our results allow numerically obtaining *tight bounds* on the performance of these queueing systems.

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 this paper we consider three such queueing systems: (*i*) the $M/G/k$ First-Come-First-Serve multi-server model, (*ii*) the $M/G/1$ round-robin scheduling model, and (*iii*) systems with time-varying load, and present a fresh approach towards their analysis: via obtaining tight bounds on the performance metric, given a partial characterization of the system parameter in terms of the first n moments.

Motivation

Due to the abundance of work surrounding the $M/G/k$ multi-server model, we use it as an example to motivate our approach. 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 and their service requirements are assumed to be independent, identically distributed (*i.i.d.*) random variables having a general distribution. If an arriving job finds an idle server, it immediately enters service; otherwise it waits in the FCFS queue. When a server becomes idle, it chooses the next job to process from the head of the FCFS queue. We will use ρ to denote the average number of busy servers. Our focus will be on the metric of mean waiting time, 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.

To the best of our knowledge, the first approximation for $\mathbf{E}\left[W^{M/G/k}\right]$ dates back to 1959 and was given by Lee and Longton [19]. Their approximation only involves the first two moments of the job size distribution, is exact for $k = 1$, and was shown to be asymptotically exact in heavy traffic by Köllerström [18]. However, the inaccuracy of the Lee-Longton approximation was realized by many authors who proposed new closed-form approximations, but which still involved at most the first two moments of the job size distribution and for which no tightness guarantees were proved. Recently, it was proved that no approximation based on only the first two moments of the job size distribution can be accurate for all job size distributions with the given moments [14].

Burman and Smith [7] proved a light-traffic approximation for $W^{M/G/k}$ which involves the entire job size distribution, and Boxma et al. [6] used it to obtain tighter approximations for $\mathbf{E}\left[W^{M/G/k}\right]$ for job size distributions with low variance (squared coefficient of variation < 1). This was achieved by interpolating the mean waiting time under deterministic jobs sizes, and under exponentially distributed job sizes, with the Burman-Smith approximation as the weighting function. However, extrapolating the Burman-Smith approximation yields inaccuracies when the job size distribution has high variance as is common in applications in computer science.

Bounds on the mean waiting time for $M/G/k$ queues (and more generally, for $GI/G/k$ queues) have mainly been obtained via two approaches (e.g., see Section 11-7 from Wolff [29]). The first approach is by assuming various orderings (stochastic ordering, increasing convex ordering) on the service distributions (see [25, 20, 26, 27, 8]), but these tend to be very loose as approximations. Moreover, one does not always have the required strong orderings on the service distribution. The second, and more practical, approach that started with the work of Kingman [17] is obtaining bounds on mean waiting time in terms of the first two moments of the inter-arrival and service distributions. The best known bounds of this type for $\mathbf{E}\Big[W^{GI/G/k}\Big]$ are presented by Daley [9]. Scheller-Wolf and Sigman [23] derive bounds on for the case $\rho < \left\lfloor \frac{k}{2} \right\rfloor$ by reducing the $GI/G/k$ waiting time recursion into an equivalent single-server recursion with dependent service times. Foss and Korshunov [12] and Scheller-Wolf and Vesilo [24] use dependent $D/GI/1$ queues to bound a $GI/G/k$ system, and obtain necessary and sufficient conditions under which higher (even fractional) moments of delay are finite.

Another approach used in the literature to establish bounds is by formulating a semidefinite program (SDP) with joint moments of service and inter-arrival time distribution forming the constraint set, and moments of waiting time as the objective function. With this approach, Bertsimas and Popescu [3] prove that the Markov inequality (using only the first moment) is tight, improve the Tchebycheff inequality (using the first two moments) to rediscover the corresponding tight inequality, and establish the analogous tight inequality that involves the first three moments. SDPs have also been used to obtain bounds on performance metrics for several queueing models. Recently, Bertsimas and Natarajan [2] have obtained numerical bounds on the moments $W^{GI/G/K}$ given the information of moments of the service and the inter-arrival time distributions. Although most of the prior work obtains numerical bounds, Osogami and Raymond [22] use SDPs to establish closed-form bounds on the waiting time in a transient $GI/G/1$ queue. However, there are insufficient experimental results on the tightness of the resulting bounds.

Our Approach

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. That is, **the bounds so obtained are the tightest achievable bounds given the partial characterization of the job size distribution**, not just arbitrary approximations or bounds. Our approach has the added benefit that many times the entire job size distribution is not available, while estimating first few moments via sampling is a much easier task. By quantifying the gap between the upper and lower bounds given these first few moments, it can be determined if a more refined characterization, say, in terms of higher order moments, is necessary.

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 (to be made precise later). The asymptotic regimes are chosen so that 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. The intuition behind the validity of this approach is the conjecture that increasing the arrival rate to an $M/G/k$ system, for example, should not change the extremality property the job size distributions, and thus extremal distribution in the asymptotic regime should remain extremal in non-asymptotic regime as well (this is a non-trivial conjecture because there exist examples where the *relative performance* of two job size distributions is sensitive to the arrival rate for $M/G/k$).

The idea of obtaining tight bounds on the performance of a queueing system based on partial characterization of the system parameters was first advocated by Eckberg [10] (and extended in Whitt [28]) for $GI/M/1$ model. However, these authors used the available implicit expressions for the performance of $GI/M/1$ as a function of the Laplace transform of the inter-arrival time distribution. The queueing systems we consider in this paper do not even have such implicit expressions.

Summary of Results

We now briefly describe the three queueing systems, the "light traffic" asymptote we look at, and our results.

The *M/G/k* multi-server system
Model: Recall that an $M/G/k$ system consists of k identical servers and a FCFS queue. The arrival process is Poisson with rate λ, 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, $\mathbf{E}\left[W^{M/G/k}\right]$, as a function of the job size distribution X.
Asymptotic Regime: We let the arrival rate $\lambda \to 0$, and look at $\mathbf{E}\left[W^{M/G/k}\right]$ of a random arrival conditioned on the event that the arrival finds all servers busy. This can be seen as the first term in the Taylor series expansion of $\mathbf{E}\left[W^{M/G/k}\right]$ around $\lambda = 0$.
Results: We start with the Burman-Smith [7] light-traffic approximation, and prove the following:
1. Given the first $n = 2$ or 3 moments of the job size distribution, the extremal distributions are given by the principal representations of the moment sequence (defined in Section 2).
2. If we restrict the job size distribution to lie in the class of completely monotone (CM) distributions, then given the first n moments, the extremal distributions are given by the principal representations

of the moment sequence within the hyperexponential class of distributions (mixtures of approximately $\frac{n}{2}$ exponential distributions; to be made formal in Section 2).

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 $\mathbf{E}\left[W^{M/G/k}\right]$ for real world heavy-tailed distributions, three moments usually suffice, especially if we add the knowledge of complete monotonicity.

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

Model: The $M/G/1$ round-robin queue consists of a single server and an infinite buffer. The arrival process is Poisson with rate λ, 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), and if the job does not finish service, it joins the back of the buffer. For analytical simplicity we assume that quantum sizes are exponentially distributed random variables. That is, each time a job gets to the server, its service quantum is an *i.i.d.* sample from an exponential distribution with rate ν. We will be interested in obtaining bounds on the mean response time, $\mathbf{E}\left[T^{M/G/1/RR}\right]$, in terms of moments of X.

Asymptotic Regime: We let the arrival rate $\lambda \to 0$, and look at the coefficient of $\Theta(\lambda)$ in the expression for $\mathbf{E}\left[T^{M/G/1/RR}\right]$.

Results: 1. We derive the light-traffic approximation for $\mathbf{E}\left[T^{M/G/1/RR}\right]$ when the job size distribution is hyperexponential with finite number of phases.

2. We use our light-traffic result to prove that if the job size distribution is restricted to lie in the class of CM distributions, then given the first n moments, the extremal distributions are given by the principal representations of the moment sequence within the hyperexponential class of distributions.

Systems with fluctuating arrival and service rates

Model: 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 job sizes are exponentially distributed. While in the H state, the arrival process is Poisson with rate λ_H, and server serves jobs at rate μ_H. During the L states, the arrival process is Poisson with rate λ_L, and the server's service rate is μ_L. The durations of stay in the L state during each visit are *i.i.d.* random variables with general distribution; we use τ_L to denote such a generic random variable. Similarly, we use τ_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 τ_L and τ_H. (As mentioned later, we expect our results to hold for systems where evolution during L and H states is governed by arbitrary Markov processes satisfying mild conditions.)

Asymptotic Regime: We consider the "fast-switching" asymptote. In particular, we index our time-varying load system with a parameter α, where in the αth system the durations of stay in L and H states are *i.i.d.* and given by $\alpha\tau_L$ and $\alpha\tau_H$, respectively. We then analyze $\mathbf{E}[N]$ in the limit $\alpha \to 0$. Note that as $\alpha \to 0$ and our systems switches very fast, the zeroth order behavior is given by an $M/M/1$ with the average arrival and service rates. We will be interested in the coefficients of higher order terms in the expansion of $\mathbf{E}[N]$ around $\alpha = 0$.

Results: 1. We derive the first fast-switching asymptote approximation for the time-varying load system when the distributions of τ_L and τ_H are hyperexponential with finite number of phases. In particular, we prove the following interesting result: The coefficient of α^i is a function of only the first $(i+1)$ moments of τ_L and τ_H. Further, this coefficient is linear in $\mathbf{E}\left[\tau_L^{i+1}\right]$ and $\mathbf{E}\left[\tau_H^{i+1}\right]$.

2. The above result immediately implies that if τ_L and τ_H are restricted to lie in the CM class, then given the first n moments, the number of jobs in the system (equivalently, the mean response time) in the fast-switching asymptote is extremized by CM distributions with extremal $(n+1)$st moment. These are again given by principal representations of the moment sequence in the class of hyperexponential distributions.

3. Our light-traffic result, and hence the result on extremal distributions, easily extend to general distributions, but we choose not present them here since the analysis is almost identical but proof ideas are easy to illustrate for CM distributions.

Finally, we illustrate the utility of our results in obtaining provable bounds on the performance of the N model for work-stealing (or the N-sharing system). While the N-sharing system can be modeled by a Markov chain, there are no exact numerical algorithms for solving it since the Markov chain is infinite in two dimensions.

A note on completely monotone class of distributions

A probability density function $f_X(\cdot)$ is said to belong to the class of completely monotone (CM) distributions if all derivatives of f_X exist and $(-1)^n f_X^{(n)}(x) \geq 0$ for all $x > 0$ and $n \geq 1$. It is well known that mixture of exponential distributions are dense in the CM class. That is, for any distribution function F in the CM class, there exist hyperexponential distributions $F^{(n)}$ with n phases such that $F^{(n)} \Rightarrow F$ as $n \to \infty$ [11, Theorem 3.2]. In fact, $F_X(\cdot)$ is a CM probability distribution function if and only if

$$F_X(x) = \int_0^\infty e^{-\mu x} dG(\mu),$$

where G is a proper probability distribution function, and commonly called the spectral distribution of F. It can be shown that this denseness is sufficient to approximate arbitrarily many moments of a CM distribution via mixture of exponential distributions. It has been established that many heavy-tailed distributions used to model computer systems workloads fall in the CM class, e.g., Pareto distributions, Weibull distributions with shape parameter less than 1 (heavier than exponential), and Gamma distributions with shape parameter less than 1 [11]. Further, there are several results on conditions under which the convergence of the inter-arrival and service-time distributions imply convergence in distribution of waiting time (see e.g. Borovkov [5, page 118], Stoyan [25]), although care must be exercised since convergence in distribution does not necessarily imply convergence of moments. To prove results about CM distributions, we will therefore restrict to looking for extremal distributions within hyperexponential distributions.

Outline

We introduce the concepts of Tchebycheff systems of functions and principal representations of moment sequences in Section 2. Section 2 also states the classical Markov-Krein Theorem which we use as a tool to prove our results for CM distributions. In Sections 3, 4 and 5, we present our results on tight moment-based bounds for (*i*) the $M/G/k$ multi-server model, (*ii*) $M/G/1$ round-robin scheduling, and (*iii*) systems with time-varying load, respectively, under "light-traffic" asymptote. We present conjectures on bounds under non-asymptotic regimes for these three queueing systems in Section 6. In Section 7, we present some approaches for proving our conjectures, and introduce a novel moment problem as a unified framework for analyzing the question of moment-based bounds for general queueing systems.

2 Principal Representations, Tchebycheff systems, and the Markov-Krein Theorem

The classic Tchebycheff inequality concerns with bounds on the tail probability of a random variable X, given $\mathbf{E}[X]$ and $\mathbf{E}[X^2]$. In other words, given the expectations of functions $f_1(x) = x$ and $f_2(x) = x^2$, one asks for bounds on the expectation of $g(x) = \mathbf{1}_{|x-\mathbf{E}[X]|>a}$. The theory of Tchebycheff systems [16] generalizes this question by asking for bounds on the expectation of some given function $g(\cdot)$ of a random variable, given a partial characterization of the random variable in terms of generalized moment constraints expressed as expectations of some functions $f_1(\cdot), \ldots, f_n(\cdot)$. In this section we will be concerned with the case $f_i(x) = x^i$. Below we present a special case of the results from this area. We will begin with the case where random variables are restricted to bounded support $[0, B]$ and where the results are easy to state. We then present results for the case where the support is $[0, \infty)$ and details are a little delicate. For a detailed treatment, we refer the reader to [16].

2.1 Random variables with support on $[0, B]$

We first introduce the notion of upper and lower principal representations as presented in [10]. Define the function $f_0(x) = 1, 0 \leq x \leq B$, and denote the moment space associated with $\{f_0, f_1, \ldots, f_n\}$ as

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

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 $\mathcal{M}_B^{n+1}$, we define the *lower and upper principal representation (pr)* to be distributions with a particular number of mass probabilities, some of which are restricted to be at 0 or B, in such a way that the first n moments of these distributions agree with $\mathbf{m^0}$. In particular the constraints are:

	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 $(0, B)$

The upper and lower principal representations are *uniquely determined* when the functions $\{f_0, \ldots, f_n\}$ satisfy certain linear independence constraints mentioned later. To see this, consider the case of upper pr for n even. We have $n+1$ constraints, one each for m_i, $0 \le i \le n$. If we are allowed $\frac{n}{2}+1$ probability masses, then we have $n+2$ degrees of freedom – $\frac{n}{2}+1$ for the locations of the probability masses, and $\frac{n}{2}+1$ for the actual probability values. Since one of the probability mass is constrained to be at B, we lose one degree of freedom, and thus the number of constraints match the degrees of freedom, where these constraints are "linearly independent" in a sense made precise next.

Definition 1 *Functions $\{h_0, h_1, \ldots, h_n\}$ form a Tchebycheff system over $[a, b]$ provided the determinants*

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

are strictly positive whenever $a \le x_0 < x_1 < \cdots < x_n \le b$.

In other words, any non-trivial linear combination of $h_0, \ldots, h_n$ must have at most n zeros in the interval $[0, B]$. Systems of polynomials: $h_i(x) = x^{\alpha_i}$ $(0 \le \alpha_0 < \alpha_1 < \ldots < \alpha_n)$ indeed form Tchebycheff systems.

The proof of the following theorem can be found in [16, Chpt. V, Sec. 5]:

Theorem 1 (Markov-Krein) *If $\{f_0, \ldots, f_n\}$ and $\{f_0, \ldots, f_n, g\}$ are Tchebycheff systems on $[0, B]$, then*

$$\beta_l \equiv \inf_{\mu_X \in \mathcal{D}} \{\mathbf{E}[g(X)] \mid \mathbf{Pr}[X \in [0, B]] = 1; \mathbf{E}[f_i(X)] = m_i,\ 0 \le i \le n\} = \int_0^B g(u) d\underline{\mu}(u)\ ,$$

$$\beta_u \equiv \sup_{\mu_X \in \mathcal{D}} \{\mathbf{E}[g(X)] \mid \mathbf{Pr}[X \in [0, B]] = 1; \mathbf{E}[f_i(X)] = m_i,\ 0 \le i \le n\} = \int_0^B g(u) d\bar{\mu}(u)\ ,$$

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

2.2 Random variables with support on $[0, \infty)$

As before, denote the moment space associated with $\{f_0, f_1, \ldots, f_n\}$ as

$$\mathcal{M}_\infty^{n+1} = \left\{ \mathbf{m} \in \mathbb{R}^{n+1} \middle| \exists \mu \in \mathcal{D}, m_i = \int_0^\infty f_i(u) d\mu(u), 0 \le i \le n \right\}$$

where $\mathcal{D}$ is the set of nonnegative regular measures of bounded variation for which the indicated integrals exist.

The definition of lower pr remains unchanged when we extend the support to $[0, \infty)$ as there are no atoms placed at the upper bound of the support. Hence, for a large enough B, the lower pr of $\mathbf{m^0}$ on $[0, B]$ will coincide with the lower pr on $[0, \infty)$. In particular, for n even, the lower pr will constitute of $\frac{n}{2}$ mass points in $(0, \infty)$ and one mass point at 0; for n odd, there will be $\frac{n+1}{2}$ mass points in $(0, \infty)$.

To define the upper pr, $\bar{\mu}$,we first need another definition.

Definition 2 *Functions $\{h_0, h_1, \ldots, h_n\}$ form a Tchebycheff system of Type II over $[0, \infty)$ provided:*
(i) $\{h_0, \ldots, h_{n-1}\}$ and $\{h_0, \ldots, h_n\}$ are Tchebycheff systems on $[0, \infty)$; and
(ii) there exists $A > 0$ such that $h_n(x) > 0$ for $x \ge A$, and

$$\lim_{x \to \infty} \frac{h_i(x)}{h_n(x)} = 0 \quad \textit{for } i < n.$$

If $\{f_0, \ldots, f_n\}$ is a Tchebycheff system of Type II, then for $\mathbf{m^0}$ in the interior of $\mathcal{M}_\infty^{n+1}$, the upper pr puts one mass at ∞, $\left\lfloor \frac{n}{2} \right\rfloor$ mass points in $(0, B)$, and additionally one at 0 if n is odd. The following example might help readers uncomfortable with the idea of mass at infinity: Consider the case $f_i(x) = x^i$ and $n = 2$. In this case, the upper pr can be seen as a limit $\epsilon \to 0$ of distributions with support $[0, \frac{1}{\epsilon}]$ which put $\Theta(\epsilon^2)$ mass on $\frac{1}{\epsilon}$. Thus, this mass at ∞ is needed to satisfy the constraint corresponding to f_n, but does not contribute to constraints for $f_0, \ldots, f_{n-1}$ when $\{f_0, \ldots, f_n\}$ is a Tchebycheff system of Type II.

Theorem 2 (Markov-Krein) *[16, Theorem V5.1] If $\{f_0, \ldots, f_n\}$ and $\{f_0, \ldots, f_n, g\}$ are Tchebycheff systems on $[0, \infty)$, and $\mathbf{m^0}$ lies in the interior of $\mathcal{M}_\infty^{n+1}$, then there exists*

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

which is achieved uniquely for $\mu_X = \underline{\mu}$, the lower pr of $\mathbf{m^0}$.
The upper bound

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

in general may not be attained, or may be infinite. However, if $\{f_0, \ldots, f_n\}$ is a Tchebycheff system of Type II, and

$$\lim_{x\to\infty} \frac{g(x)}{f_n(x)} = \gamma < \infty,$$

then β_u exists and is "achieved" by the upper pr of $\mathbf{m^0}$.

In the last sentence of the theorem, we say "achieved" to emphasize the fact that the upper pr has a mass point at ∞ and thus it is not a finite measure. However, β_u exists and is achieved as a limit.

The Markov-Krein Theorem has been successfully applied in the context of queueing systems [10, 28]. In particular, for a $GI/M/1$ system, Theorem 1 proves that given the first n moments of the inter-arrival time distribution, the mean number of jobs in the system is extremized by inter-arrival time distributions which correspond to the upper and lower pr's. The proof follows by noting that the mean number of jobs in a $GI/M/1$ queue with *i.i.d.* inter-arrival times given by a random variable A is an increasing function of the Laplace-Stieltjes transform of the inter-arrival time distribution ($\widetilde{A}(s) = \mathbf{E}\left[e^{-sA}\right]$), and the functions $g_s(x) = e^{-sx}$ form Tchebycheff system with $f_i(x) = x^i$.

Principal representations within Hyperexponential distributions Consider the following random variable with an n-phase hyperexponential distribution:

$$X \sim \begin{cases} \text{Exp}\left(\frac{1}{x_1}\right) & \text{with probability } q_1 \\ \vdots \\ \text{Exp}\left(\frac{1}{x_n}\right) & \text{with probability } q_n \end{cases}$$

We can now define another random variable Y with distribution given by the *inverse spectrum* of X:

$$Y \sim \begin{cases} x_1 & \text{with probability } q_1 \\ \vdots \\ x_n & \text{with probability } q_n \end{cases}$$

We have the following straightforward relationship between moments of X and Y: $\mathbf{E}[Y^i] = \frac{\mathbf{E}\left[X^i\right]}{i!}$. We define the upper and lower principal representation for a moment sequence $m_1, m_2, \ldots, m_n$ within the class of hyperexponential distributions as the distributions whose inverse spectrum are the upper and lower principal representations, respectively, for the moment sequence $m_1, \frac{m_2}{2!}, \ldots, \frac{m_n}{n!}$.

3 Bounds for the $M/G/k$ Multi-server Model

In this section we present our results on tight moment-based bounds for the $M/G/k$ model in light traffic. Recall that the arrival process is Poisson with rate λ, and the job sizes are *i.i.d.* according to a random variable X. The load of the system is defined as $\rho = \lambda \mathbf{E}[X]$ and denotes the time average number of busy servers. The waiting time of a job is defined to be the time between when a job arrives to the system and when it enters service, and is denoted by $W^{M/G/k}$. We will analyze $\mathbf{E}\left[W^{M/G/k}\right]$ in the light traffic asymptote $\rho \to 0$ while holding X and k unchanged.

In Section 3.1, we present our results on bounds for general job size distributions given first 2 or 3 moments, and in Section 3.2 for completely monotone distributions given any number of moments. In Section 3.3, we present numerical results on bounds obtained using principal representations for common heavy-tailed job size distributions.

3.1 Bounds for general distribution

We begin with a well-known result on the light traffic approximation for $W^{M/G/k}$.

Theorem 3 (Burman Smith [7]) *Under the assumption that the job size distribution is phase-type, as $\rho \to 0$, the probability that an arrival finds all servers busy is asymptotically given by $\frac{1}{k!}\left(\frac{\rho}{k}\right)^k$, and conditioning on this event, $W^{M/G/k}$ is distributed as the minimum of k independent copies of the stationary excess of X, denoted by X_e. The survival function of X_e is given by $\overline{F}_{X_e}(x) = \mathbf{Pr}[X_e > x] = \frac{\int_{u=x}^{\infty} \mathbf{Pr}[X>u]du}{\mathbf{E}[X]}$.*

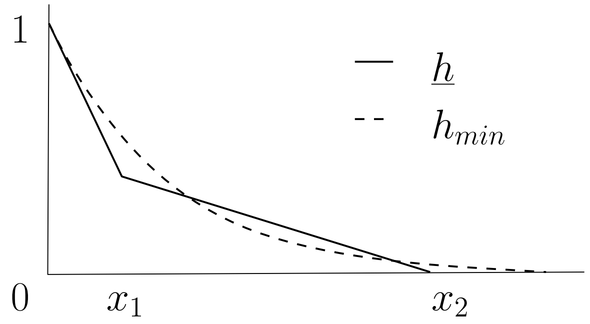
Theorem 4 *Given the first n ($n = 2$ or 3) moments of the job size distribution X, $\mathbf{E}\left[W^{M/G/k}\right]$ under light traffic is extremized by service distributions given by the lower and upper principal representations of the moment sequence.*

Proof: Due to lack of space we present the proof for the case $n = 3$, where the lower pr minimizes, and upper pr maximizes $\mathbf{E}\left[W^{M/G/k}\right]$. Denote $\overline{F}_{X_e} = h$ for succinctness. Since $1-h(x) = \frac{\int_0^x \mathbf{Pr}[X>u]du}{\mathbf{E}[X]}$ is the integral of a bounded, non-negative, decreasing function, $(1 - h(x))$ is a continuous, non-decreasing, concave function. The problem of extremizing $\mathbf{E}\left[W^{M/G/k}\right]$ in the light-traffic asymptote can thus be equivalently formulated as:

$$\begin{aligned}
&\text{min/max} \int_0^\infty h(u)^k du \\
&\text{subject to } h(\cdot) \text{ continuous, non-negative, non-increasing, convex ;} \\
&\qquad h(0) = 1 \ ; \quad |h'(0^+)| = \frac{\mathbf{Pr}[X>0]}{\mathbf{E}[X]} \le \frac{1}{\mathbf{E}[X]} \ ; \\
&\qquad \int_0^\infty h(u)du = \frac{\mathbf{E}[X^2]}{2\mathbf{E}[X]} \ ; \quad \int_0^\infty u \cdot h(u)du = \frac{\mathbf{E}[X^3]}{12\mathbf{E}[X]}.
\end{aligned}$$

(Note that a solution to the above problem exists because $0 \le h(u)^k \le h(u)$, and $\int h(u)du$ is finite.) Let $\underline{h}$ represent the survival functions of X_e corresponding to the lower pr of X for the given moment sequence. Now, suppose that $\underline{h}$ is not the solution to the minimization problem above, and the solution is instead given by h_{min}. For $n = 3$, we have that the lower pr has 2 point masses, say at

$0 < x_1 < x_2 < \infty$, as shown below.



The absolute value of the slope of h_{min} at 0^+ must be at most that of $\underline{h}$ since the lower pr has no mass at 0 for $n = 3$, and because h_{min} is convex, it follows that $\delta = h_{min} - \underline{h}$ satisfies (i) $\int_0^\infty \delta(u)du = 0$ (i.e., areas under $\underline{h}$ and h_{min} are equal), (ii) $\int_0^\infty u \cdot \delta(u)du = 0$ (from moment conditions), and (iii) $\delta(\cdot)$ changes sign exactly twice, and the sequence of signs is $+-+$ (see the figure above). We obtain a contradiction:

$$\begin{aligned} & \int h_{min}(u)^k du - \int \underline{h}(u)^k du \\ = & \int \delta(u) \left[h_{min}(u)^{k-1} + h_{min}(u)^{k-2}\underline{h}(u) + \ldots + \underline{h}(u)^{k-1} \right] du \\ > & \ 0 \end{aligned}$$

To see the last inequality, denote the function in the square brackets by $\ell(\cdot)$, and note that ℓ is convex. Now $\int_0^\infty \delta(u)\ell(u)du = [\delta(u)\ell'(u)]_0^\infty - \int_{u=0}^\infty \ell'(u) \int_{v=0}^u \delta(v)dvdu$. The first term is zero because $\delta(0) = \delta(\infty) = 0$. Now assuming derivatives exist (by approximating by smoothed versions), we find that $\ell'(u)$ is an increasing function, and $\int \delta(v)dv$ is a function that changes sign only once, from $+$ to $-$ and integrates to 0. Thus $\int_{u=0}^\infty \ell'(u) \int_{v=0}^u \delta(v)dvdu < 0$.

The proof for upper pr is identical except that the sequence of signs of δ in this case is $-+-$. For $n = 2$, δ changes sign once. ■

3.2 Bounds for CM job size distributions

Theorem 5 *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,* $\mathbf{E}\left[W^{M/G/k}\right]$ *under light traffic is extremized by the lower and upper principal representations of the moment sequence within the hyperexponential class of distributions.*

Proof: The first step of the proof is to restrict our attention to hyperexponential distributions with finite number of phases as they are dense in the CM class. We will now use the Markov-Krein Theorem to prove the result. However, Theorem 1 does not apply directly to our problem because as Theorem 3 shows, the mean waiting time in light traffic can not be written as $\mathbf{E}[f(X)]$ for any function $f(\cdot)$. Instead, we prove a stronger result.

Consider a tagged arrival that finds all the servers busy. We fix the distribution of the job sizes at the first $k-1$ servers to be exponential with arbitrary parameters (say $\nu_1, \nu_2, \ldots, \nu_{k-1}$). We will now show that given the moments of the job size distribution for the job at the kth server, the hyperexponential distributions that minimize or maximize the time until first departure, and hence the waiting time of the arrival, are given by the pr's *irrespective of the choice of* $\nu_1, \ldots, \nu_{k-1}$. Let the job size distribution of the job at the kth server be:

$$X \sim \begin{cases} \text{Exp}\left(\frac{1}{x_1}\right) & \text{with probability } q_1, \\ \vdots & \\ \text{Exp}\left(\frac{1}{x_n}\right) & \text{with probability } q_n. \end{cases}$$

As defined in Section 2, let Y denote a random variable whose distribution is given by the inverse spectrum of X, and let $M = \sum_{j=1}^{k-1} \nu_j$. The mean waiting time of the tagged arrival, $\mathbf{E}[W^*]$, is then given by:

$$\begin{aligned} \mathbf{E}[W^*] &= \sum_{i=1}^{n} \frac{q_i x_i}{\mathbf{E}[X]} \cdot \frac{1}{M + \frac{1}{x_i}} \\ &= \frac{1}{\mathbf{E}[X]} \sum_{i=1}^{n} q_i \left(\frac{M x_i - 1}{M^2} + \frac{1}{M^2 (M x_i + 1)} \right) \\ &= \frac{1}{M} - \frac{1}{M^2 \mathbf{E}[Y]} + \frac{1}{M^2 \mathbf{E}[Y]} \mathbf{E}\left[\frac{1}{MY + 1} \right] \end{aligned}$$

From Theorem 31 of [15], $\frac{1}{Mx+1}$ forms a Tchebycheff system with the functions $i!x^i$, and hence by Theorem 1, the result follows. ■

Remark: The reader might wonder if we could use a similar proof outline as Theorem 5 to prove the result for general distributions. To be more precise, we can arbitrarily fix the residual sizes of jobs at the first $k-1$ servers as $u_1 \leq \ldots \leq u_{k-1}$. We may then ask the question: for given first n moments, what job size distribution for X extremizes $\mathbf{E}[\min\{X_e, u_1\}]$. The latter expectation can indeed be written as $\mathbf{E}[f(X)]$, where $f(\cdot)$ is a piecewise polynomial function. However, even for $n = 3$, $f(x)$ **does not** form a Tchebycheff system with the moment functions x^0, x^1, x^2 and x^3. Thus, Theorem 4 can in some sense be seen as *breaking the Tchebycheff system barrier.*

3.3 Simulation and numerical evaluation

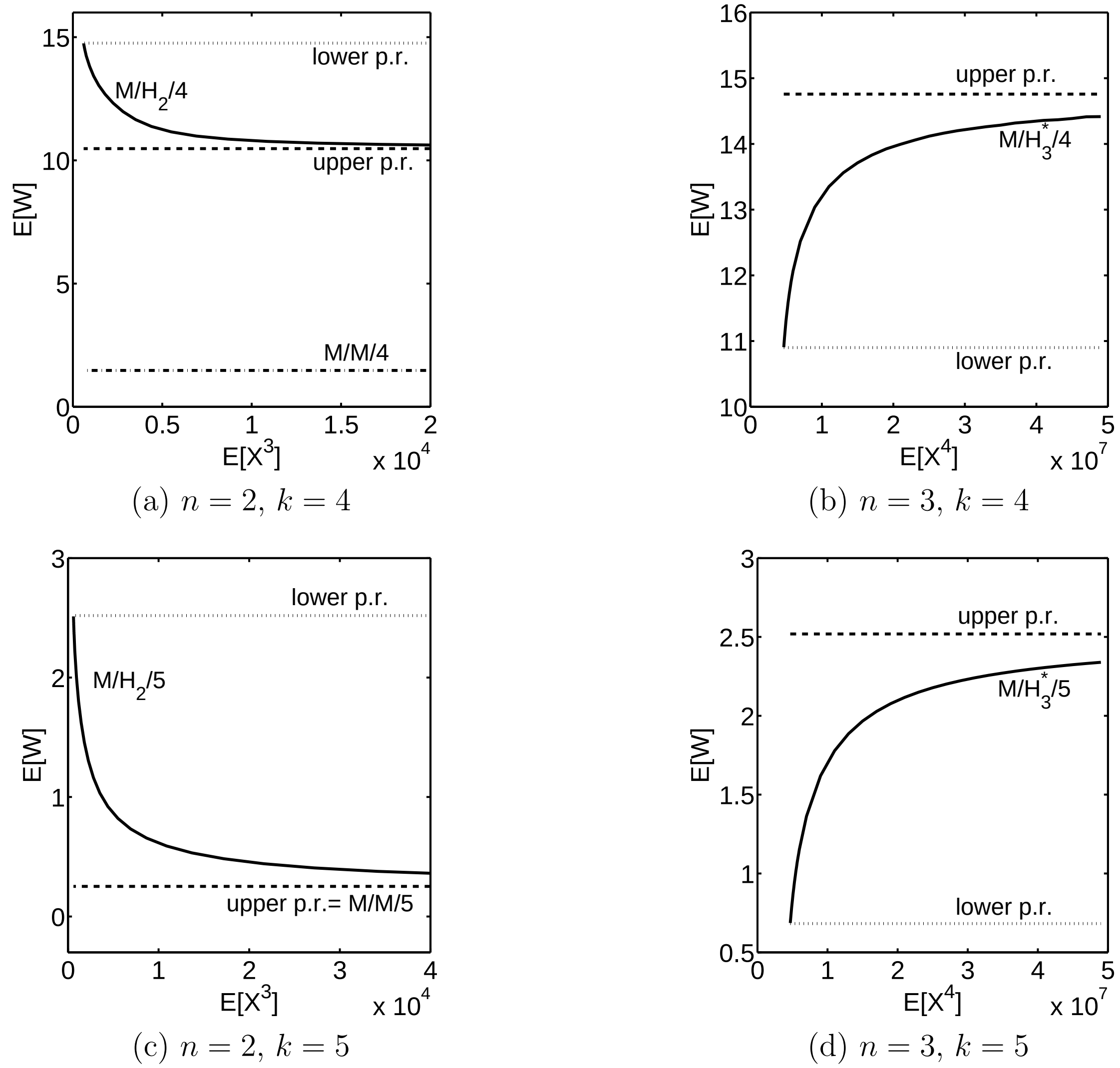
We conjecture that Theorem 4 extends to any number, n, of moments and to general traffic, and Theorem 5 extends to general load. See Section 6 for the specific properties that we conjecture to hold generally for moment-based tight bounds on $\mathbf{E}\left[W^{M/G/k}\right]$. In this section, we provide support for the conjectures and numerically study the quality of bounds obtained with principal representations.

Figure 3.3 provides numerical evidence in support of validity of Theorem 5 for general load. The solid curves in Figure 3.3(a) and Figure 3.3(c) show $\mathbf{E}\left[W^{M/G/k}\right]$ when the job size has a two-phase hyperexponential (H_2) distribution which allows us to vary $\mathbf{E}[X^3]$ while holding the first two moments fixed. The solid curves in Figure 3.3(b) and Figure 3.3(d) show $\mathbf{E}\left[W^{M/G/k}\right]$ when the job size has a degenerate three-phase hyperexponential ($H_3^\star$) distribution, which has two non-zero mean exponential phases and a phase with zero mean. Within $H_3^\star$ distributions, we can vary $\mathbf{E}\left[X^4\right]$, while holding the first three moments fixed. We set the number k of servers as indicated below each figure.

Observe that the solid curves lie between the mean waiting times attained when the job size distributions are given by principal representations within hyperexponential distributions (dashed line and dotted line). The principal representations are determined by the first two moments of the H_2 distribution in Figure 3.3(a) and Figure 3.3(c) and the first three moments of the $H_3^\star$ distribution in Figure 3.3(b) and Figure 3.3(d). Also, observe that the solid curve is decreasing in $\mathbf{E}[X^3]$ and increasing in $\mathbf{E}\left[X^4\right]$. A detail is that, in Figure 3.3(a), the upper principal representation refines the lower bound obtained from an exponential job-size distribution (line labeled with $M/M/4$). However, in Figure 3.3(c), the lower bound obtained with a principal representation coincides with the lower bound obtained from an exponential job-size distribution. See Conjecture 1 in Section 6 for the properties that we conjecture to hold generally for the bounds on $\mathbf{E}\left[W^{M/G/k}\right]$.

Figure 2 shows $\mathbf{E}\left[W^{M/G/k}\right]$ and its bounds obtained with principal representations, when the job size distribution is a Weibull distribution. We fix the parameters of the Weibull distribution such that its probability density function is $f(x) = \frac{1}{2}x^{-1/2}\exp\left(-x^{1/2}\right)$ for $x \geq 0$. We also fix the number of servers, $k = 4$, and vary the load, $\rho \equiv \lambda\mathbf{E}[X]$, as indicated below each figure. The dashed line shows the exact value of $\mathbf{E}\left[W^{M/G/k}\right]$, and the crosses and the dots show bounds on $\mathbf{E}\left[W^{M/G/k}\right]$ obtained with principal representations. Specifically, a bound shown with a cross is the mean delay in the $M/G/k$ system whose job size distribution has a principal representation that is determined by the moments of the Weibull distribution (see Theorem 4). A bound shown with a dot is obtained analogously with a principal representation within hyperexponential distributions (see Theorem 5). Notice that the Weibull distribution under consideration is completely monotonic (see [11]), so that a principal representations within hyperexponential distributions give proper bounds. The horizontal axis indicates the number of moments used to determine the principal representations. The moments of the Weibull distribution under consideration are $\mathbf{E}[X^n] = (2n)!$ for $n = 1, 2, \ldots$.

In all cases, $\mathbf{E}\left[W^{M/G/k}\right]$ and the bounds shown with a cross are obtained via simulations; the bounds shown with a dot are calculated via matrix analytic methods. For each data point, the simulation is run 10 times and the average value of the 10 simulated mean waiting times is plotted. Each run of simulation is continued for 10,000,000 events, where an event is either an arrival or a departure of a job, and waiting times of the departed jobs are recorded (we ignore the first 100,000 departures).



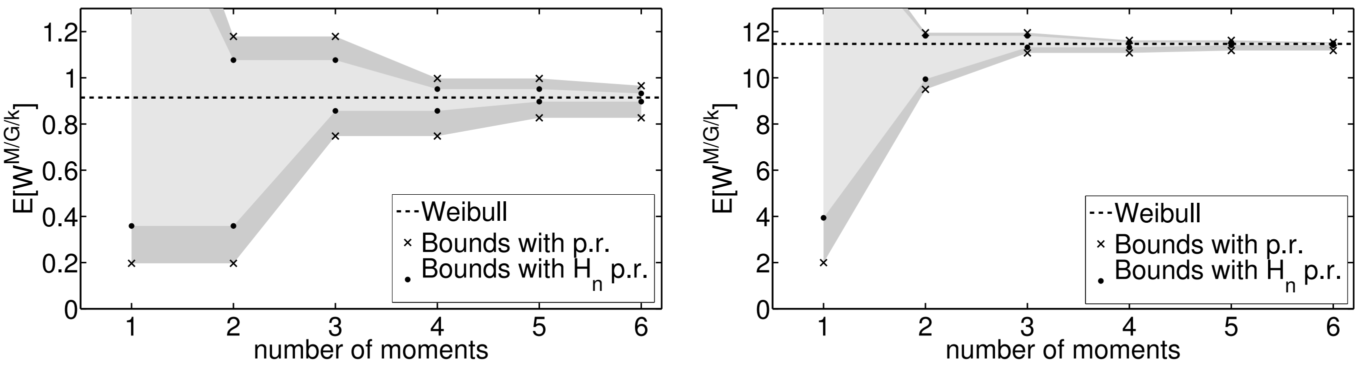
(a) $n = 2$, $k = 4$

(b) $n = 3$, $k = 4$

(c) $n = 2$, $k = 5$

(d) $n = 3$, $k = 5$

Figure 1: Mean delay in an $M/G/k$ system when the job size has a hyperexponential distribution. In (a) and (c), we vary $\mathbf{E}[X^3]$, while keeping $\mathbf{E}[X] = 1$ and $\mathbf{E}[X^2] = 20$. In (b) and (d), we vary $\mathbf{E}\Big[X^4\Big]$, while keeping $\mathbf{E}[X] = 1$, $\mathbf{E}[X^2] = 20$, and $\mathbf{E}[X^2] = 8000$.



(a) $\rho = 2.4$ (b) $\rho = 3.6$

Figure 2: Bounding mean delay in an $M/G/4$ queue when the job size has a Weibull distribution.

Confidence intervals are sufficiently small and not shown.

In Figure 2, notice that, except for $n = 2$, either lower bounds or upper bounds are shown for each n, where n is the number of moments used to determine the principal representation. This is because the lower (respectively, upper) bound obtained with an even (respectively, odd) number n of moments in general does not improve the corresponding lower (respectively, upper) bound obtained with $n-1$ moments. An exception is that the lower bound obtained with $n = 2$ moments improves upon that with $n = 1$ for $\rho = 3.6$ (but not for $\rho = 2.4$). The lower bound with $n = 2$ moments is given by a limiting distribution where one of mass points, B, approaches infinity. This lower bound corresponds to the principal representation with $B = 10^6$. It appears that the mean delay with the principal representation hardly changes in the interval between $B = 10^4$ and $B = 10^6$. For a $B > 10^6$, the analysis of the mean delay suffers from numerical errors.

Observe in Figure 2 that the principal representations within hyperexponential distributions (dot) can give bounds that are significantly better than the corresponding bounds obtained with the standard principal representations (cross). The principal representations within hyperexponential distributions provide bounds that are valid only for (job size) distributions that are completely monotonic. The difference between a dot and a cross show the refinement of the bound that we gain from the knowledge of complete monotonicity. Also observe that, as the number of moments used to determine principal representations grows, the upper and lower bounds approach each other quickly particularly at high load (Figure 2(b)). This makes intuitive sense, because $\mathbf{E}\left[W^{M/G/k}\right]$ becomes insensitive to third and higher moments in heavy-traffic.

3.4 A departure from Markov-Krein

The classical Markov-Krein theorem only enforces the condition that the moment constraint functions $\{f_0, \ldots, f_n\}$ be linearly independent (modulo signs of functions). As mentioned earlier, this condition holds for the power functions $f_i(x) = x^{\alpha_i}$, $0 \leq \alpha_0 < \ldots \leq \alpha_n$. In particular, note that α_i need not be integral. However, here we see a departure in the behavior of $M/G/k$ from the classical Markov-Krein characterization – If the moment constraints involve fractional moments, the relative performance of upper and lower principal representations may flip as the arrival rate increases from light traffic to heavy traffic. Further, the upper and lower pr's may no longer provide bounds.

We will illustrate this point with an example. Consider the moment constraints $m_0 = \mathbf{E}[X^0] = 1$, $m_1 = \mathbf{E}[X^1] = 1$, and $m_{\frac{3}{2}} = \mathbf{E}\Big[X^{\frac{3}{2}}\Big] = 5$, and let us restrict ourselves to the class of hyperexponential distributions (since we have established the light traffic extremality results). The upper pr places almost entire probability mass on the mean, and behaves as an exponential distribution in light traffic. Therefore in light traffic, the mean sojourn time of the upper pr is smaller than the mean sojourn time of the lower pr. However, due to the mass at ∞ in upper pr, the second moment is ∞ whereas the lower pr has all moments finite. Since the mean sojourn time in heavy traffic limit is completely determined by the first two moments, the mean sojourn time of the upper pr in heavy traffic is higher than the mean sojourn time of the lower pr. Further, the mean sojourn times of the upper and lower pr will cross at some arrival rate λ^*, where the mean sojourn time is T^*. Clearly, there are distributions with the given moment constraints with mean sojourn time different than T^* at λ^*. Thus the pr's do not provide bounds in this case. The same behavior is observed whenever the cardinality of moment constraints in the interval $(0, 2)$ is even.

The above discussion, while discomfiting, should be taken as an instructive caution. While we strive to prove a Markov-Krein characterization for $M/G/k$ mean sojourn time, conditions more than those in Theorems 1 and 2 would be needed. We conjecture that the knowledge of the integral moments suffices. However, fractional moments, in general, may not be admissible.

4 Bounds for $M/G/1$ Round-Robin

In this section we prove tight moment-based bounds for the mean response time, $\mathbf{E}\Big[T^{M/G/1/RR}\Big]$, of an $M/G/1$ round-robin queue with exponentially distributed quantum sizes and CM job size distribution in the limit when the arrival rate $\lambda \to 0$. Formally, we consider round-robin scheduling where every time a job gets to the server, the server picks a quantum size i.i.d. from an Exp(ν) distribution..

Lemma 1 *Consider a $M/G/1$ round-robin system with i.i.d.* Exp(ν) *quanta, arrival rate λ and the following H_n job size distribution:*

$$X \sim \begin{cases} \text{Exp}(\gamma_1) & \textit{with probability } q_1 \\ \vdots & \\ \text{Exp}(\gamma_n) & \textit{with probability } q_n \end{cases}$$

As the arrival rate $\lambda \to 0$:

$$\mathbf{E}\Big[T^{M/G/1/RR}\Big] = \mathbf{E}[X]\,(1 + \lambda\mathbf{E}[X]) + \frac{\lambda}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{q_i q_j(\gamma_i - \gamma_j)^2}{\gamma_i\gamma_j(\gamma_i\gamma_j + (\gamma_i + \gamma_j)\nu)} + o(\lambda).$$

Proof: As the arrival rate approaches 0, the coefficient of $\Theta(\lambda)$ term will be dominated by events where (*i*) a job arrives to an empty system and is interrupted at most once during its stay, or (*ii*) a job arrives to a system with a yet uninterrupted job already in service, and there are no more arrivals during its sojourn.

Let us consider the case where an Exp(ξ) job is interrupted by an Exp(χ) job. In this case, the mean residual response time of the interrupted Exp(ξ) job satisfies

$$\begin{aligned} A_{\xi,\chi} &= \frac{1}{\xi+\nu} + \frac{\nu}{\xi+\nu}\left(\frac{1}{\chi+\nu} + \frac{\nu}{\chi+\nu}A_{\xi,\chi} + \frac{\chi}{\chi+\nu}\frac{1}{\xi}\right) \\ &= \frac{1}{\xi}\left(1 + \frac{\xi\nu}{(\xi+\nu)(\chi+\nu) - \nu^2}\right) \end{aligned} \tag{1}$$

Similarly, the mean response time of the interrupting Exp(χ) job is given by:

$$B_{\chi,\xi} = \frac{1}{\chi}\left(1 + \frac{\chi^2 + \chi\nu}{(\xi+\nu)(\chi+\nu) - \nu^2}\right) \tag{2}$$

Returning to our original round-robin system, a tagged class i job arrives to an empty system with probability $(1 - \lambda\mathbf{E}[X])$, and stays there for Exp($\gamma_i + \lambda$) time. With probability $\frac{\lambda}{\lambda+\gamma_i}$, the tagged class i job gets interrupted by another arrival which is of class j with probability q_j and spends additional time A_{γ_i,γ_j}. With probability $\lambda\mathbf{E}[X]$, the class i job arrives to a busy system and interrupts a class j job with probability $\frac{q_j}{\gamma_j\mathbf{E}[X]}$, in which case the response time of the tagged class i job is B_{γ_i,γ_j}. Thus, the overall response time of a class i job is given by:

$$\begin{aligned} \mathbf{E}[T_i] &= (1 - \lambda\mathbf{E}[X])\left(\frac{1}{\gamma_i+\lambda} + \frac{\lambda}{\gamma_i+\lambda}\sum_j q_j A_{\gamma_i,\gamma_j}\right) + \lambda\mathbf{E}[X]\sum_j \frac{q_j}{\gamma_j\mathbf{E}[X]}B_{\gamma_i,\gamma_j} + O(\lambda^2) \\ &= \frac{1 + \lambda\mathbf{E}[X]}{\gamma_i} + \frac{1}{\gamma_i}\sum_{j=1}^{n} q_j \frac{\gamma_i - \gamma_j}{\gamma_j(\gamma_i\gamma_j + (\gamma_i+\gamma_j)\nu)} \end{aligned} \tag{3}$$

Calculating $\mathbf{E}\Big[T^{M/G/1/RR}\Big] = \sum_i q_i\mathbf{E}[T_i]$, we get the expression in the theorem statement. ■

Theorem 6 *Given the first n moments of the job size distribution X in the CM class, $\mathbf{E}\Big[T^{M/G/1/RR}\Big]$ under light traffic is extremized by the lower and upper principal representations of the moment sequence within the class of hyperexponential distributions.*

Proof: We will follow similar steps as in the proof of Theorem 5. The first is to restrict our attention to hyperexponential distributions with finite number of phases as they are dense in the CM class. We will then use the Markov-Krein Theorem to show that $\mathbf{E}\left[T^{M/G/1/RR}\right]$ given in Lemma 1 is extremized by the principal representations within the hyperexponential class of distributions. Let Y denote a random variable with the same distribution as the inverse spectrum of the job size distribution X, and let $x_i = \frac{1}{\gamma_i}$. From Lemma 1 (ignoring $o(\lambda)$ terms):

$$\begin{aligned}
&\mathbf{E}\left[T^{M/G/1/RR}\right]\\
=&\mathbf{E}[X]\left(1+\lambda\mathbf{E}[X]\right)+\frac{\lambda}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{q_iq_j(\gamma_i-\gamma_j)^2}{\gamma_i\gamma_j(\gamma_i\gamma_j+(\gamma_i+\gamma_j)\nu)}\\
=&\mathbf{E}[Y]\left(1+\lambda\mathbf{E}[Y]\right)+\frac{\lambda}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{q_iq_j(x_i-x_j)^2}{(1+(x_i+x_j)\nu)}\\
=&\mathbf{E}[Y]\left(1+\lambda\mathbf{E}[Y]\right)+\frac{\lambda}{2}\sum_{i=1}^{n}q_i\sum_{j=1}^{n}\frac{q_j}{\nu^2}\left(\nu x_j-1+\frac{\nu^2x_i^2+\nu x_i+1}{\nu(x_i+x_j)+1}\right)\\
=&\mathbf{E}[Y]\left(1+\lambda\mathbf{E}[Y]\right)+\frac{\lambda}{2}\sum_{i=1}^{n}q_i\left[\frac{\nu\mathbf{E}[Y]-1}{\nu^2}-\frac{\nu^2x_i^2+\nu x_i+1}{\nu^2}\mathbf{E}[f_i(Y)]\right]
\end{aligned}$$

where $f_k(x) = \frac{1}{\nu(x+x_k)+1}$. From Theorem 31 of [15], each $f_k(x)$ forms a Tchebycheff system with the functions $i!x^i$ (and the same pr minimizes each $\mathbf{E}[f_k(Y)]$, and similarly the other pr maximizes each $\mathbf{E}[f_k(Y)]$), and hence by Theorem 1, the result follows. ■

5 Bounds for systems with time-varying load

In this section we prove tight moment-based bounds for an $M/M/1$ queue with arrival and service rates controlled by a 2-state environment process. However, we believe the results extend to much more general time-varying systems (see remark after Theorem 8). The asymptotic regime we consider is what we call the "fast-switching asymptote": we let the duration of stay in the environment states on each visit approach 0. In Theorem 7, we prove the result when the distributions for the durations of environment states are CM, but our proof extends to generally distributed durations. In Section 5.2 we show an application of our results to obtaining (conjectured) tight bounds on the performance of work-stealing, an exact analysis of which is impossible due to the resulting 2-D infinite Markov chain.

Formally, we consider a system with an exogenous environment process with states L and H. The durations of the H states are i.i.d. according to a random variable τ_H, and those of L states are i.i.d. according to τ_L. The job sizes are i.i.d. exponential with mean 1. However, during the L state, the arrival process is Poisson with rate λ_L and the server's service rate is μ_L. Similarly, during the H states, the arrival process is Poisson with rate λ_H and the server's service rate is μ_H. We define

$\mu_{avg} = \frac{\mu_L \mathbf{E}[\tau_L]+\mu_H \mathbf{E}[\tau_H]}{\mathbf{E}[\tau_L]+\mathbf{E}[\tau_H]}$, $\lambda_{avg} = \frac{\lambda_L \mathbf{E}[\tau_L]+\lambda_H \mathbf{E}[\tau_H]}{\mathbf{E}[\tau_L]+\mathbf{E}[\tau_H]}$, and $\rho = \frac{\lambda_{avg}}{\mu_{avg}}$. We will consider a sequence of systems indexed by a parameter α, where the durations of L and H states in the αth system are i.i.d. as $\alpha\tau_L$ and $\alpha\tau_H$, respectively. We will analyze the mean number of jobs in this sequence of systems, $\mathbf{E}[N_\alpha]$, as $\alpha \to 0$.

5.1 Fast-switching asymptote and bounds

Theorem 7 *Consider a time-varying load system with residence time in L and H states given by $\alpha\tau_L$ and $\alpha\tau_H$, respectively. Further, assume that the distributions of τ_L and τ_H are hyperexponential with finite number of phases. Then the mean number of jobs in the system as $\alpha \to 0$ is given by:*

$$\mathbf{E}[N_\alpha] = \frac{\rho}{1-\rho} + \sum_{i=1}^{\infty} \phi_i \alpha^i \tag{4}$$

where ϕ_i are functions of the first $i+1$ moments of τ_L and τ_H (and μs and λs), and are linear in $\mathbf{E}\left[\tau_H^{i+1}\right]$ and $\mathbf{E}\left[\tau_L^{i+1}\right]$.

Proof: We defer the details to Appendix A. Due to lack of space, we illustrate the main ideas by instead looking at a finite buffer system with a buffer size of 1 (i.e., there can only be either 0 or 1 jobs in the system) with time-varying arrival and service rates. The proof easily extends to the infinite buffer case as well. ■

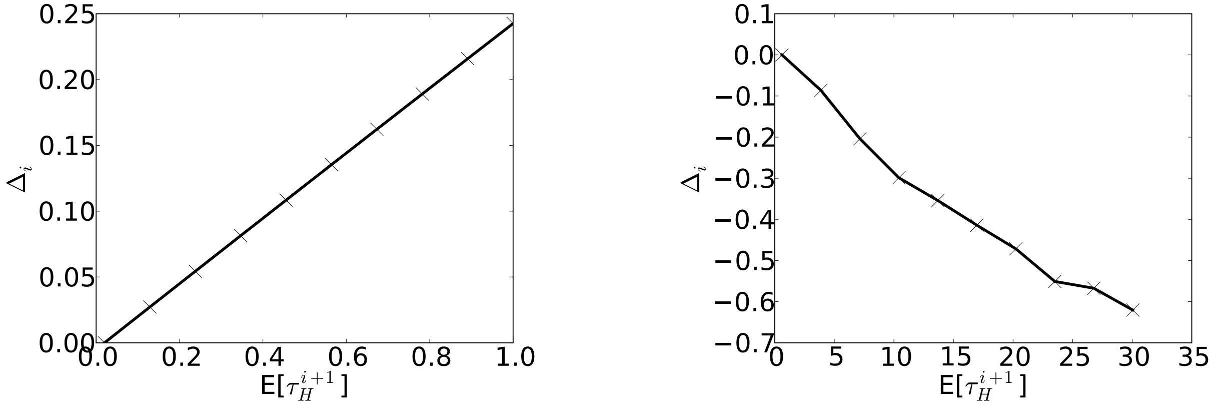
Theorem 8 *If τ_L and τ_H are constrained to lie in the CM class, then given the first n moments of τ_L and τ_H, the mean number of jobs, $\mathbf{E}[N]$, under the fast switching asymptote is extremized by the lower and upper principal representations of the moment sequence within the hyperexponential distribution.*

Proof: Given the first n moments of τ_L and τ_H, the coefficients of α^i for $0 \le i \le n-1$ are already fixed. The distributions which extremize the mean number of jobs will be those that extremize the coefficient of α^n. Since this is linear in the $(n+1)$st moments, and moment functions $f_i(x) = x^i$ form a Tchebycheff system, the theorem follows from Theorem 1. ■

Remark: The result of this section easily extends to the case of general distributions for τ_L and τ_H. The only fact that is needed is that for any finite x, the probability of i arrivals or departures in duration αx is $c_i(\alpha x)^i - d_i(\alpha x)^{i+1} + o(\alpha^i)$ for some constants c_i and d_i – a simple consequence of the Poisson process.

Remark: The results of this section should also extend to more general time-varying systems. For example, during the L and H, the system could evolve according to arbitrary finite-state Markov processes with generators Q_L and Q_H, respectively, as long as the characteristic polynomials of Q_L and Q_H ($\phi_L(s) = det(sI - Q_L)$, $\phi_H(s) = det(sI - Q_H)$) do not have repeated roots.

Remark: Unlike $M/G/k$ and $M/G/1$ round-robin models where the heavy-traffic limits tend to be



(a) $i = 1$

(b) $i = 2$

Figure 3: Dependency of ϕ_i (Theorem 7) on $\mathbf{E}\left[\tau_H^{i+1}\right]$

insensitive to the job size distribution beyond the first or second moments, for the time-varying load model, we actually do have an interesting result in the "slow switching asymptote" ($\alpha \to \infty$): in the special case when $\tau_H \sim \text{Exp}(\gamma)$ and $\lambda_H > \mu_H$. Under transient overload during H states, as the mean durations of H and L states become long, the time-varying load system converges to a fluid system. For the special case mentioned, it is not hard to see that the mean response time of this fluid system can be derived from a $GI/M/1$ system with inter-arrival time distribution given by τ_L. As stated earlier, characterization of bounds for $GI/M/1$ via principal representations is known from the work of Eckberg [10]. We have proved that this characterization also holds under the fast switching asymptote, irrespective of the choice of arrival and service rates, and when both L and H state durations may be generally distributed.

We validate Theorem 7 numerically with Figure 3. Consider a time-varying load system with $\lambda_L = 4$, $\lambda_H = 8$, $\mu_L = \mu_H = 10$. We let τ_L have an exponential distribution with rate 10 and vary τ_H as is specified in the following. In Figure 3(a), τ_H has a two-phase hyperexponential ($H_2^\star$) distribution having a non-zero exponential phase and a phase of zero mean. The $H_2^\star$ allows us to hold the mean at 0.1 and vary the second moment $\mathbf{E}[\tau_H^2]$, which is indicated along the horizontal axis. The vertical axis shows $\Delta_i \equiv (\mathbf{E}[N_\alpha] - \mathbf{E}[N'_\alpha])/\alpha^i$ for $i = 1$, where N'_α indicates N_α with $\mathbf{E}[\tau_H^2] = 0.02$ (the lowest value studied in Figure 3(a)). Throughout we set $\alpha = 10^{-3}$, so that $o(\alpha^{i+1})$ terms are negligible relative to $\Theta(\alpha^i)$ term. Because $\mathbf{E}[\tau_H]$ is fixed, Δ_1 shows (approximately) how ϕ_1 depends on $\mathbf{E}[\tau_H^2]$. Indeed, we find that ϕ_1 grows linearly with $\mathbf{E}[\tau_H^2]$. In Figure 3(b), τ_H has a two-phase hyperexponential (H_2) distribution, which allows us to hold $\mathbf{E}[\tau_H] = 0.1$ and $\mathbf{E}[\tau_H^2] = 0.2$, and vary $\mathbf{E}[\tau_H^3]$, which is indicated along the horizontal axis. The vertical axis in Figure 3(b) shows Δ_2 (here, N'_α represents N_α with $\mathbf{E}[\tau_H^3] = 0.601$ (the lowest value studied in Figure 3(b))), which indicates (approximately) how ϕ_2 depends on $\mathbf{E}[\tau_H^3]$. Although the line in Figure 3(b) is not as straight as Figure 3(a) due to numerical errors, we find that ϕ_2 does decrease linearly with $\mathbf{E}[\tau_H^3]$.

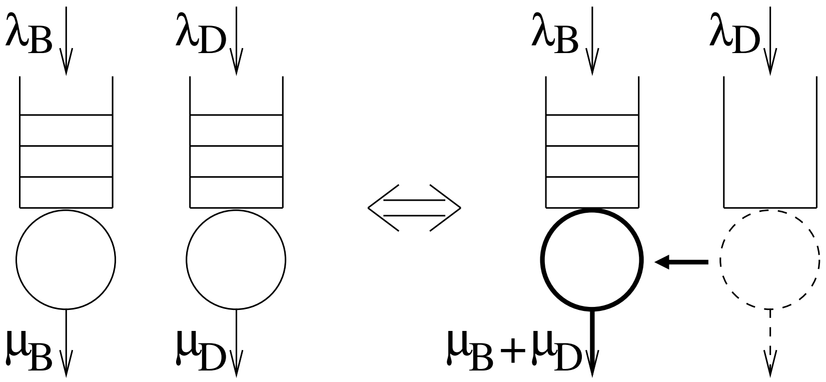


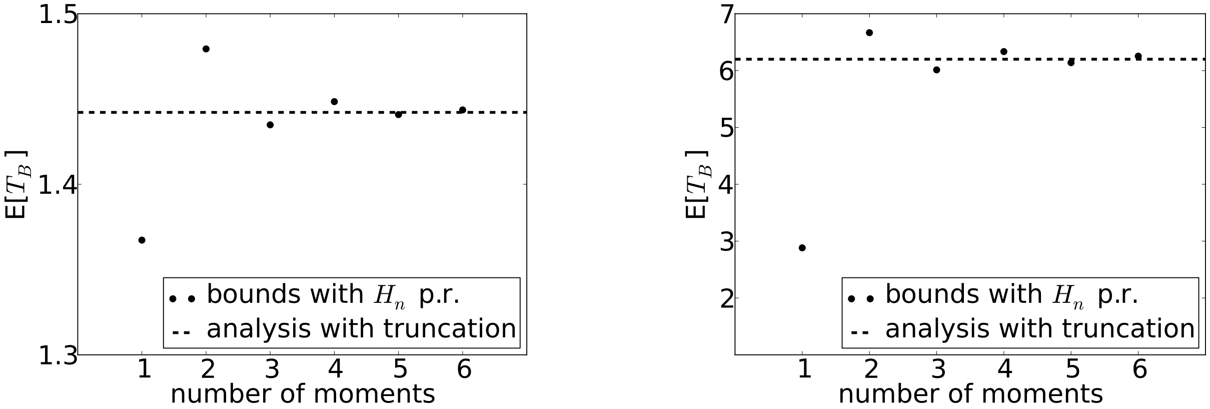
Figure 4: The service rate at the beneficiary queue is $\mu_B + \mu_D$ when the donor queue is empty and μ_B otherwise.

5.2 Application to analysis of N-sharing model

In this section, we apply the analysis of the time-varying load system to a work-stealing system with two $M/M/1$ queues, beneficiary and donor (see Figure 4). The two queues are independent except that the service rate at the beneficiary queue becomes larger when the donor queue is empty. Let λ_B (respectively, λ_D) be the arrival rate at the beneficiary (respectively, donor) queue. Let μ_B (respectively, μ_D) be the service rate at the beneficiary (respectively, donor) queue when the donor queue is nonempty. When the donor queue is empty, the service rate at the beneficiary queue becomes $\mu_B + \mu_D$. We assume that the jobs are preemptive, so that the service rate at the beneficiary queue changes from $\mu_B + \mu_D$ to μ_B immediately after a job arrives at the empty donor queue. The jobs arriving at the donor queue see a standard $M/M/1$ system with arrival rate λ_D and service rate μ_D.

Observe that the jobs arriving at the beneficiary queue, which we refer to as beneficiary jobs, see a time varying system, where $\lambda_H = \lambda_L = \lambda_B$, $\mu_H = \mu_B$, $\mu_L = \mu_B + \mu_D$, τ_L has an Exponential distribution with rate λ_D, and τ_H is the busy period of the $M/M/1$ system with arrival rate λ_D and service rate μ_D. To analyze the response time of beneficiary jobs, we need to consider a Markov chain that is infinite in two dimensions, where one dimension represents the number of beneficiary jobs and the other dimension represents the number of donor jobs. Such a Markov chain cannot be solved exactly, so that the prior work has investigated various approximations (e.g., truncation in [13] and approximating the donor busy period with moment matching in [21]). However, such approximations do not guarantee their accuracy and can be computationally expensive.

Now, because the busy period of an $M/M/1$ system has a hyperexponential distribution with a continuous spectrum [1], our results in Section 5.1 suggest that the stationary mean response time of beneficiary jobs, $\mathbf{E}[T_B]$, is likely to be extremized by lower and upper principal representations, given the first n moments of the busy period for $n = 1, 2, \ldots$. Figure 5 shows $\mathbf{E}[T_B]$ and the bounds obtained with principal representations. We fix $\mu_B = \mu_D = 1.0$ and vary $\rho = \lambda_D = \lambda_B$ as indicated below each figure. The dashed line shows the exact value of $\mathbf{E}[T_B]$, which is obtained by numerically



(a) $\rho = 0.6$

(b) $\rho = 0.9$

Figure 5: Bounding mean response time of beneficiary jobs in the N-sharing model.

analyzing a Markov chain. Here the state space of the Markov chain is truncated so that the number of jobs at the donor queue is at most a threshold, 200, and we verify that increasing the threshold does not change the results of the analysis. The resulting Markov chain is a quasi-birth-and-death (QBD) process that can be analyzed with a matrix analytic method.

A dot in Figure 5 shows a bound on $\mathbf{E}[T_B]$ obtained by replacing the busy period with a principal representation, where the number of moments used to determine the principal representation is shown on the horizontal axis. Here again, the bound is numerically computed by analyzing a QBD process via matrix analytic methods (but due to a small number of levels, the computational cost is much lower than truncation). Observe that a principal representation using odd number (1, 3, and 5 are shown in the figure) of moments gives a lower bound on $\mathbf{E}[T_B]$, while an upper bound is given by a principal representation using even number (2, 4, and 6 are shown in the figure) of moments. When five or six moments are used, the upper bound and the lower bound give nearly exact value (specifically, the two bounds differ by 0.62% in Figure 5(a) and 2.2% in Figure 5(b)).

The results in Figure 5 justify the approximation in [21], where the donor busy period is approximated by matching its first three moments. The lower bound obtained with the first three moments gives a nearly perfect approximation, and using fourth and higher moments do not significantly improve the bound. In determining the principal representations for the busy period, B, we have used the following expression obtained by manipulating the Laplace-Stieltjes transform of B (we omit the details due to lack of space): $\mathbf{E}\left[B^k\right] = \frac{k!}{\mu_D^k(1-\rho_D)^{2k-1}}\xi_k$, where $\rho_D = \lambda_D/\mu_D$, $\xi_1 = \xi_2 = 1$, $\xi_3 = 1 + \rho_D$, $\xi_4 = 1 + 3\rho_D + \rho_D^2$, $\xi_5 = 1 + 6\rho_D + 6\rho_D^2 + \rho_D^3$, and $\xi_6 = 1 + 10\rho_D + 20\rho_D^2 + 10\rho_D^3 + \rho_D^4$.

6 Conjectures on tight bounds for general traffic

Let $\mathbf{m} = (m_0 = 1, m_1, m_2, \ldots, m_n) \in \mathbb{R}_+^n$ be such that there exists a positive random variable $\mathcal{X}$ with $\mathbf{E}[\mathcal{X}^i] = m_i,\ i = 0, \ldots, n$. For n odd, let $\mathcal{D}(\mathbf{m})$ denote the unique lower pr with moments

$\mathbf{m}$ and support $[0,\infty)$ (and therefore has mass at ∞), and let $\mathcal{D}^*_B(\mathbf{m})$ denote the unique upper pr with moments $\mathbf{m}$ and support $[0,B]$. For n even, let $\mathcal{D}^*(\mathbf{m})$ denote the unique lower principal representation with moments $\mathbf{m}$ and support $(0,\infty]$, and let $\mathcal{D}_B(\mathbf{m})$ denote the unique upper pr with moments $\mathbf{m}$ and support $[0,B]$. (The star in the superscript is to emphasize a point mass at 0, and the B in the subscript emphasizes the point mass at the upper bound, B, of the support.)

Let

$$T_h(\mathbf{m}) = \sup_{\mu_X\in\mathcal{D}} \left\{ \mathbf{E}\Big[T^{\mathcal{S}(X)}\Big] \,\Big|\, \mathbf{E}\Big[X^i\Big] = m_i,\ i=0,\dots,n \right\},$$
$$T_l(\mathbf{m}) = \inf_{\mu_X\in\mathcal{D}} \left\{ \mathbf{E}\Big[T^{\mathcal{S}(X)}\Big] \,\Big|\, \mathbf{E}\Big[X^i\Big] = m_i,\ i=0,\dots,n \right\}.$$

where $\mathcal{S}(X)$ represents either the $M/G/k$ ($k\geq 2$) multi-server system, the $M/G/1$ round-robin system, or the time-varying load system, with X as the random variable for the job size distribution for the $M/G/k$ and the $M/G/1$ round-robin models, or the duration of the L or H states for the time-varying load model, and T denotes the response time.

Conjecture 1 *Let $\mathbf{m} = (m_0 = 1, m_1, \dots, m_n)$, $n \geq 1$, be a valid moment sequence for positive distributions. Let $\mathbf{m}' = (m_0, m_1, \dots, m_{n-1})$. Then,*

Case 1:** n **odd

(i) $T_h(\mathbf{m}) = \lim_{B\to\infty} \mathbf{E}\Big[T^{\mathcal{S}(\mathcal{D}^*_B(\mathbf{m}))}\Big]$.
(ii) $T_l(\mathbf{m}) = \mathbf{E}\Big[T^{\mathcal{S}(\mathcal{D}(\mathbf{m}))}\Big]$.
(iii) $T_l(m_1,\dots,m_{n-1},x)$ is strictly decreasing in x.

Case 2:** n **even

(i) $T_h(\mathbf{m}) = \mathbf{E}\Big[T^{\mathcal{S}(\mathcal{D}^*(\mathbf{m}))}\Big]$.
(ii) $T_l(\mathbf{m}) = \lim_{B\to\infty} \mathbf{E}\Big[T^{\mathcal{S}(\mathcal{D}_B(\mathbf{m}))}\Big]$.
(iii) $T_h(m_1,\dots,m_{n-1},x)$ is strictly increasing in x.

Further, for the $M/G/k$ system: for n odd, $T_h(\mathbf{m}) = T_h(\mathbf{m}')$; and for n even (and additionally for the $\rho < (k-1)$ if $n = 2$), $T_l(\mathbf{m}) = T_l(\mathbf{m}')$. [1]

To state in simple language, the conjectures would imply the following for the $M/G/k$ multi-server model: If we are given only the mean of the service distribution, we only have enough information to

[1] Intuitively, as we said before, this is true because the mass at ∞ is only present to satisfy the largest moment constraint. Karlin and Studden write ([16], page 152), "Whenever mass at ∞ is present, this mass may be ignored to obtain a measure representing only the moments $m_0, m_1, \cdots, m_{n-1}$." In the classical Markov-Krein framework, this treatment suffices under some conditions on the function $g(\cdot)$ whose expectation we are extremizing. However for queueing systems, whenever the sup/inf as defined above exist and involve the upper principal representation, we need to be slightly more careful. For example, for the case of $M/G/k$ with $n = 2$ and $\rho \geq (k-1)$ we can not ignore the mass at infinity and must define the sup/inf via the limit of a sequence of systems involving upper pr on finite support. This fact is highlighted via $M/G/1$ where given $n = 2$, the mean sojourn time is completely determined. However, if we ignore the mass at ∞ in the upper pr, we incorrectly obtain $\mathbf{E}\big[T^{M/D/1}\big]$!

fix a lower bound on $\mathbf{E}\Big[W^{M/G/k}\Big]$. This lower bound is given by $\mathbf{E}\Big[W^{M/D/k}\Big]$. If we are additionally given the second moment of the service distribution, we can fix an upper bound on $\mathbf{E}\Big[W^{M/G/k}\Big]$. (It can be shown that this conjectured upper bound is given by $\frac{m_2}{m_1^2}\mathbf{E}\Big[W^{M/D/k}\Big]$ (see e.g., [14]). It can also be shown that when $\rho \geq k-1$, we also refine our lower bound [14], as was also seen in the numerical experiments in Section 3.3.) By determining the third moment of service distribution, we can *refine* (tighten) our lower bound but this *lower bound decreases as the third moment increases*. The upper bound remains unchanged. Similarly, knowledge of the fourth moment will *refine the upper bound* on the mean waiting time (bring it down), and so forth for alternating higher even and odd moments. Further, these bounds are achieved by mixtures of point masses as dictated by the upper and lower pr's.

7 Towards a unified approach for moment-based bounds

While our results offer an intuitive justification for tight moment-based bounds via principal representations for the three queueing systems considered in the paper for general (i.e., non-asymptotic) traffic conditions, we are still quite far from proving the desired result. Further, we believe that similar results are likely to hold for other queueing systems as well. We now discuss some possible lines of attack for proving moment-based bounds for general queueing systems.

One line of approach to proving such results would be similar to what we have tried to do in the present paper. One would first prove the desired result in an "appropriate" asymptotic regime, that is, where the effect of the entire distribution of the parameter of interest (e.g., the job size distribution) is apparent. This is expected to be the easier step, and should offer insights into what distributions are extremal. The remaining open question would then be to prove that the extremality of the conjectured distributions is preserved when we are in non-asymptotic regime. This last step seems very challenging because it is possible to come up with job size distributions whose relative performance flip while going from light to heavy traffic. [2]

While the above approach sounds promising in that obtaining extremal distributions in asymptotic regimes would be tractable, proving such results for every new queueing system *ab initio* would be far from elegant.

A second line of approach could be that of Eckberg [10] for obtaining bounds on the mean response time of the $GI/M/1$ model. As we mentioned earlier, the mean response time of a $GI/M/1$ queue can be written in terms of an implicit quantity that is an increasing function of the Laplace-Stieltjes transform $\mathbf{E}\Big[e^{-sA}\Big]$ of the inter-arrival time duration A. It is well known that the functions e^{-sx}

[2] Indeed, consider moment sequences $\mathbf{m} = (m_1, m_2)$ and $\mathbf{m}' = (m_1', m_2')$ with $m_1 = m_1'$ and $(m_1)^2 < m_2 < m_2'$. The lower pr of $\mathbf{m}$ yields a higher mean sojourn time than the upper pr of $\mathbf{m}'$ in light traffic. However, the mean sojourn time in heavy traffic is completely determined by the first two moments, and hence the lower pr of $\mathbf{m}$ yields a lower mean sojourn time than the upper pr of $\mathbf{m}'$ in heavy traffic.

form a Tchebycheff system with moment functions x^i. Therefore from Theorem 1, the principal representations of the moment sequence would extremize the Laplace-Stieltjes transform point-wise, and hence the mean response time of the $GI/M/1$ queue. Employing a similar approach for the mean response time of queueing systems considered in this paper by expressing these quantities as increasing functions of $\mathbf{E}[f(X)]$ for some function f which forms a Tchebycheff system with $f_i(x) = x^i$, and then directly applying Theorem 1, eludes us (and in light of the discussion in Section 3.4, seems not possible).

To overcome the above shortcomings, we propose a unified framework by posing the following *moment problem*: Observe that the solution to any queueing system can be represented at some level by the fixed point of a stochastic recursive sequence (SRS). That is, there exists Φ such that

$$\mathbf{W} \overset{d}{=} \Phi(\mathbf{W}, X), \tag{5}$$

where $\mathbf{W}$ is the unknown random vector capturing the performance of the system, and $\overset{d}{=}$ denotes equality in distribution. For example, for the $GI/G/1/FCFS$ system, the distribution of the customer average waiting time W is given by the Lindley recursion:

$$W \overset{d}{=} (W + X - A)^+$$

where X is the job size distribution, and A is the inter-arrival time distribution. As another example, for the $GI/G/k/FCFS$ queueing system, let $\mathbf{W} = (W_1, W_2, \ldots, W_k)$ where $W_1 \leq W_2 \leq \ldots \leq W_k$ denote the Kiefer-Wolfowitz workload vector seen by arriving customer (equivalently, the ordered vector of times at which the k servers will idle, assuming the customer arriving at time $t = 0$ has size 0 and and there are no further arrivals). The distribution of $\mathbf{W}$ is then given by:

$$\mathbf{W} \overset{d}{=} \mathscr{R}\left((\mathbf{W} + X \cdot \mathbf{e_1} - A \cdot \mathbf{e})^+\right)$$

where $\mathbf{e_1}$ is a $k-$vector whose first element is 1 and the rest are 0, $\mathbf{e}$ is a k-vector all of whose elements are 1, and $\mathscr{R}$ is a function that reorders the elements of its argument in ascending order.

The final performance metric of interest would be $\mathbf{E}[g(\mathbf{W})]$ for some function g. Our goal is to seek bounds on $\mathbf{E}[g(\mathbf{W})]$, given the first n moments of X. *For what class of probability flows $\Phi(\cdot)$ and functions $g(\cdot)$ can these bounds be characterized along the Markov-Krein Theorem?*

Even partial progress on the above moment problem promises to yield bounds on many interesting queueing systems in a single shot – one only needs to check whether the SRS for the queueing system satisfies certain conditions. Further, an understanding of this problem should give insights into the common thread among queueing systems which share the Markov-Krein characterization property, but are otherwise seemingly very different. For example, what is the fundamental difference between the queueing systems described above and the following queueing system for which the principal representations achieve *identical* mean sojourn time (when n, the number of moment constraints, is even), yet the mean sojourn time is sensitive to the job size distribution?

A queueing system where principal representations are non-extremal Consider a 2-server system where each server follows the ideal Processor Sharing (PS) scheduling discipline. Jobs arrive according to a Poisson process with rate λ, and join the shorter queue on arrival (ties broken randomly, no jockeying between queues). It is easy to see that given any first n moments *with n even*, the job size distributions corresponding to the upper and lower pr's yield identical mean sojourn time. Consider the case $n = 2$ – the mass at ∞ in the upper p.r. does not influence the mean sojourn time; jobs of size 0 in the lower p.r. depart the PS servers instantaneously on arrival. Thus both the upper and lower p.r. systems effectively behave as if the job size distribution is deterministic (albeit, with different means; the arrival process is still Poisson but with different rates). This in turn implies that the distribution for the number of jobs in the upper and lower p.r. systems are identical, and thus by Little's law, so are the mean sojourn time.

While the upper and lower p.r. yield the same mean sojourn time, this system is known to be sensitive to the job size distribution. Bonald and Proutiére [4] have proved that local balance is a necessary and sufficient condition for insensitivity, whereas shortest queue routing with static node capacities violates the local balance condition.

8 Conclusions

In this paper we have taken a significant step towards solving three queueing systems which have not yielded exact analysis so far, one of them being the classical $M/G/k$ multi-server system whose analysis has remained open for more than 50 years. Our approach is different from prior attempts in the literature in that instead of trying to obtain an explicit expression for the mean response time as a function of the job size distribution, we found the job size distributions with given first n moments which minimize or maximize the mean response time, thus obtaining tight lower and upper bounds on the mean response time given a partial characterization of the job size distribution in terms of its moments.

Our approach relied on looking at appropriate tractable asymptotic regimes where the effect of the entire job size distribution is apparent (unlike heavy traffic regimes, for example), and extracting the extremal distributions. For the $M/G/k$ multi-server system, we proved that given the first $n = 2$ or 3 moments, these extremal distributions are given by the principal representations of the moment sequence. If we restrict the job size distribution to lie in the completely monotone class of distributions, then given any first n moments, we proved the extremal distributions are the principal representations within the hyperexponential class of distributions. We proved a similar result for $M/G/1$ round-robin queue with Exponentially distributed quantum sizes, and for systems with fluctuating load (with the parameter being the durations of high and low load states) and presented numerical evidence of the utility of our results.

Finally, analogous to the classical Markov-Krein theorem for scalar functions, we propose exploration of Markov-Krein characterization of solutions of Stochastic recursive equations as a unified approach to identify and study queueing systems permitting moment-based characterization of extrema via principal representations of the moment sequence of the random variables driving them.

References

[1] J. Abate and W. Whitt. Simple spectral representations for the M/M/1 queue. *Queueing Systems*, 3(4):321–345, 1988.

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

[3] D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15:780–804, 2005.

[4] T. Bonald and A. Proutière. Insensitive bandwidth sharing in data networks. *Queueing Syst. Theory Appl.*, 44:69–100, May 2003.

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

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

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

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

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

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

[11] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31:245–279, 1998.

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

[13] L. Green. A queueing system with general use and limited use servers. *Operations Research*, 33(1):168–182, 1985.

[14] V. Gupta, J. Dai, M. Harchol-Balter, and B. Zwart. On the inapproximability of $M/G/K$: why two moments of job size distribution are not enough. *Queueing Systems*, 64(1):5–48, 2010.

[15] M. A. Johnson and M. T. Taaffe. Tchebycheff systems for probabilistic analysis. *American Journal of Mathematical and Management Sciences*, 13(1-2):83–111, 1993.

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

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

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

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

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

[21] T. Osogami, M. Harchol-Balter, and A. Scheller-Wolf. Analysis of cycle stealing with switching costs and thresholds. *Performance Evaluation*, 61(4):347–369, 2005.

[22] T. Osogami and R. Raymond. Semidefinite optimization for transient analysis of queues. *ACM SIGMETRICS Performance Evaluation Review*, 38(1):363–364, 2010.

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

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

[25] D. 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.

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

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

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

[29] R. W. Wolff. *Stochastic Modeling and the Theory of Queues.* Prentice Hall, 1989.

A Proof of Theorem 6

As stated previously, to illustrate the main ideas behind the proof, we will instead consider an $M/M/1/1$ system in the 2-state environment process defined in Section 5. For this case, we only need to analyze the time average idle probability. Let p_L and p_H denote the idle probabilities at the *end* of L and H states, respectively, and let $\overline{p_L}$ and $\overline{p_H}$ be the time average idle probabilities during L and H states, respectively. Our focus is not on deriving the precise coefficients of α^i for all i because our goal is not to propose an approximation by extrapolating the fast-switching asymptote (even though we can do so). Instead, we want to identify sufficient functional dependence of these coefficient on the moments of τ_L and τ_H to be able to conclude that principal representations extremize the performance metric of interest.

Let the distributions of τ_L be given by:

$$\tau_L \sim \begin{cases} \text{Exp}(\gamma_1) & \text{with probability } q_1 \\ \vdots & \\ \text{Exp}(\gamma_n) & \text{with probability } q_n \end{cases}$$

We begin with a simple lemma.

Lemma 2 *Consider an $M/M/1/1$ system with arrival rate λ and service rate μ. Let $\tau \sim \text{Exp}(\gamma)$, and let $p(t)$ denote the idle probability at time t. Then:*

$$p(\tau) = \frac{p(0) + \frac{\mu}{\gamma}}{1 + \frac{\mu+\lambda}{\gamma}} \tag{6}$$

Proof: The Chapman-Kolmogorov equation is given by:

$$\frac{dp(t)}{dt} = -\lambda p(t) + \mu(1 - p(t))$$

Integrating by parts:

$$p(\tau) = \int_0^\infty \gamma e^{-\gamma u} p(u) du = p(0) + \frac{1}{\gamma}\left(\mu - (\lambda + \mu)p(\tau)\right).$$

■

By conditioning on the which of the n phases of the L state duration occurs and using the above lemma, we can obtain p_L in terms of p_H for the αth system as:

$$p_L = \sum_{j=1}^{n} q_j \frac{p_H + \alpha \frac{\mu_L}{\gamma_j}}{1 + \alpha \frac{\mu_L + \lambda_L}{\gamma_j}} \tag{7}$$

$$= p_H \left(1 - \alpha(\mu_L + \lambda_L)\mathbf{E}[\tau_L] + \sum_{k=2}^{i+1} \alpha^k \mathbf{E}\Big[\tau_L^k\Big]\eta_k + \Theta(\alpha^{i+2})\right) + \alpha\mu_L\mathbf{E}[\tau_L] + \sum_{k=2}^{i+1} \alpha^k \mathbf{E}\Big[\tau_L^k\Big]\zeta_k + \Theta(\alpha^{i+2}) \tag{8}$$

where η_k and ζ_k are constants (functions of μ_L and λ_L only). Similarly,

$$\begin{aligned} p_H = p_L &\left(1 - \alpha(\mu_H + \lambda_H)\mathbf{E}[\tau_H] + \sum_{k=2}^{i+1} \alpha^k \mathbf{E}\Big[\tau_H^k\Big]\theta_k + \Theta(\alpha^{i+2})\right) \\ &+ \alpha\mu_H\mathbf{E}[\tau_H] + \sum_{k=2}^{i+1} \alpha^k \mathbf{E}\Big[\tau_H^k\Big]\kappa_k + \Theta(\alpha^{i+2}) \end{aligned} \tag{9}$$

where, again, θ_k and κ_k are constants (functions of μ_H and λ_H only).

Eliminating p_H,

$$\begin{aligned} p_L = p_L \Big(1 - \alpha & \left[(\mu_L + \lambda_L)\mathbf{E}[\tau_L] + (\mu_H + \lambda_H)\mathbf{E}[\tau_H]\right] \\ &+ \sum_{k=2}^{i} \alpha^k \sigma_k + \alpha^{i+1} \left[\mathbf{E}\Big[\tau_L^k\Big]\eta_k + \mathbf{E}\Big[\tau_H^k\Big]\theta_k\right] + \Theta(\alpha^{i+2})\Big) \\ &+ \alpha\left[\mu_L\mathbf{E}[\tau_L] + \mu_H\mathbf{E}[\tau_H]\right] + \sum_{k=2}^{i} \alpha^k \psi_k \\ &+ \alpha^{i+1}\left[\mathbf{E}\Big[\tau_L^{i+1}\Big]\zeta_k + \mathbf{E}\Big[\tau_H^{i+1}\Big]\kappa_k\right] + \Theta(\alpha^{i+2}) \end{aligned} \tag{10}$$

where σ_k and ψ_k for $2 \le k \le i$ involve $\mu_L, \mu_H, \lambda_L, \lambda_H$ and $\mathbf{E}[\tau_L^m]$ and τ_H^m for $1 \le m \le i$ (importantly, not $\mathbf{E}\Big[\tau_L^{i+1}\Big]$, $\mathbf{E}\Big[\tau_H^{i+1}\Big]$, or still higher moments). This gives

$$\begin{aligned} p_L = \frac{\mu_{avg}}{\mu_{avg} + \lambda_{avg}} &\left(1 + \frac{\alpha^i}{\mathbf{E}[\tau_L] + \mathbf{E}[\tau_H]} \left[\frac{\mathbf{E}\Big[\tau_L^{i+1}\Big]\zeta_k + \mathbf{E}\Big[\tau_H^{i+1}\Big]\kappa_k}{\mu_{avg}}\right.\right. \\ &\left.\left. + \frac{\mathbf{E}\Big[\tau_L^{i+1}\Big]\eta_k + \mathbf{E}\Big[\tau_H^{i+1}\Big]\theta_k}{\mu_{avg} + \lambda_{avg}}\right] + \sum_{k=1}^{i} \alpha^k \phi_k\right) + \Theta(\alpha^{i+1}) \end{aligned} \tag{11}$$

where again ϕ_k for $1 \le k \le i$ only involve μ, λ, and the first i moments of τ_L and τ_H. A similar expression holds for p_H. Note that as $\alpha \to 0$, the idle probability of the finite buffer system is indeed given by $\frac{\mu_{avg}}{\mu_{avg}+\lambda_{avg}}$.

Finally, the expression for the time avergae idle probability during L states is obtained as:

$$\overline{p_L} = \frac{1}{\mathbf{E}[\tau_L]} \sum_{j=1}^{n} \frac{\frac{q_j}{\gamma_j}\left(p_H + \alpha\frac{\mu_L}{\gamma_j}\right)}{1 + \alpha\frac{\mu_L+\lambda_L}{\gamma_j}} \tag{12}$$

The contributions to the α^i term in $\overline{p_L}$ are made by $O(\alpha^i)$ terms in p_H, and also from $\alpha\frac{q_j\mu_L}{\gamma_j^2}$ term in the numerator above. It is straightforward to see that the coefficient of the α^i term will again depend on only the first $(i+1)$ moments, and will be linear in the $(i+1)$st moments of τ_L and τ_H.