Tight Moments-based Bounds for Queueing Systems

VARUN GUPTA

Carnegie Mellon  →  Google Research  ←  University of Chicago Booth School of Business

With:

Takayuki Osogami
(IBM Research-Tokyo)
The $M/G/k/\text{FCFS}$ model
The $M/G/k$/FCFS model

$k$
Homogeneous servers
The $M/G/k/\text{FCFS}$ model

First-Come-First-Serve Buffer

$\{\text{k Homogeneous servers}\}$
The $M/G/k/FCFS$ model

First-Come-First-Serve Buffer

Poisson($\lambda$)

$\lambda = \text{arrival rate}$

$k$ Homogeneous servers
Poisson($\lambda$)

First-Come-First-Serve Buffer

$S_{i+2}$ $S_{i+1}$ $S_i$

• $\lambda =$ arrival rate
• job sizes ($S_1$, $S_2$, …) i.i.d. samples from $S$
• “load” $\rho \equiv \lambda \ E[S]$

The $M/G/k$/FCFS model

k Homogeneous servers
The $M/G/k$/FCFS model

- $\lambda = \text{arrival rate}$
- job sizes $(S_1, S_2, \ldots)$ i.i.d. samples from $S$
- “load” $\rho \equiv \lambda \ E[S]$

**GOAL** : $E[W^{M/G/k}]$
\[ \lambda \rightarrow S_{i+1} \quad S_i \rightarrow \rho \equiv \lambda \mathbb{E}[S] \]

### k=1

**Case: \( S \sim \text{Exponential} (M/M/1) \)**
Analyze \( \mathbb{E}[W^{M/M/1}] \) via Markov chain (easy)

**Case: \( S \sim \text{General} (M/G/1) \)**

\[
\mathbb{E}[W^{M/G/1}] = \frac{C^2 + 1}{2} \mathbb{E}[W^{M/M/1}]
\]

\[
C^2 = \frac{\text{var}(S)}{\mathbb{E}[S]^2}
\]

Sq. Coeff. of Variation (SCV) > 20 for computing workloads

### k>1

**Case: \( S \sim \text{Exponential} (M/M/k) \)**

\( \mathbb{E}[W^{M/M/k}] \) via Markov chain

**Case: \( S \sim \text{General} (M/G/k) \)**

No exact analysis known

The Gold-standard approximation:

Lee, Longton (1959)

\[
\mathbb{E}[W^{M/G/k}] \approx \frac{C^2 + 1}{2} \mathbb{E}[W^{M/M/k}]
\]
Lee, Longton approximation:
\[ E[W^{M/G/k}] \approx \frac{C^2 + 1}{2} E[W^{M/M/k}] \]

- Simple
- Exact for \( k=1 \)
- Asymptotically tight as \( \rho \to k \) (think Central Limit Thm.)

Can not provision using this approximation!
Outline

2 moments not enough for $E[W^{M/G/k}]$

Tighter bounds via higher moments of job size distribution
Lee, Longton approximation:
\[ \mathbb{E}[W^{M/G/k}] \approx \frac{C^2+1}{2} \mathbb{E}[W^{M/M/k}] \]

GOAL: Bounds on approximation ratio

\{G \mid 2 \text{ moments}\}

{Dai, G., Harchol-Balter, Zwart}
COR.: No approx. for $E[W_{M/G/k}]$ based on first two moments of job sizes can be accurate for all distributions when $C^2$ is large

PROOF: Analyze limit distributions in $D_2 \equiv$ mixture of 2 points

Approximations using higher moments?

THEOREM: If $\rho < k-1$, Gap $\geq (C^2+1) \times$

[Dai, G., Harchol-Balter, Zwart]
2 moments not enough for $E[W_{M/G/k}]$

Tighter bounds via higher moments of job size distribution
Exploiting higher moments

GOAL: Identify the “extremal” distributions with given moments

RELAXED GOAL: Extremal distributions in some “non-trivial” asymptotic regime

IDEA: Light-traffic asymptotics ($\lambda \to 0$)
RELAXATION: Identify the “extremal” distributions in light traffic

Light traffic theorem for $M/G/k$ [Burman Smith]:

$$
E[W^{M/G/k}] = \frac{\rho^k}{k!} E[\min\{S_{e1}, S_{e2}, \ldots, S_{ek}\}] + o(\rho^k)
$$

Probability of finding all servers busy

i.i.d. copies of $S_e \equiv equilibrium excess$ of $S$

pdf of $S_e$: $f_{S_e}(x) = \frac{\text{Prob}[S \geq x]}{E[S]}$

SUBGOAL: Extremal distributions for $E[\min\{S_{e1}, \ldots, S_{ek}\}]$

s.t. $E[S_i] = m_i$ for $i=1,\ldots,n$
Where we are...

**GOAL:** Tight bounds on $E[W^{M/G/k}]$ given $n$ moments of $S$

**IDEA:** Identify extremal distributions

**RELAXATION (Light Traffic):** Extremal distributions for

$$E[\min\{S_{e1},\ldots,S_{ek}\}] \text{ s.t. } E[S] = m_i \text{ for } i=1,\ldots,n$$
GIVEN: Moment conditions on random variable $X$ with support $[0,B]$

$E[X^0] = m_0$
$E[X^1] = m_1$

...  
$E[X^n] = m_n$

**Principal Representations (p.r.)** on $[0,B]$ are distributions satisfying the moment conditions, and the following constraints on the support

- **Lower p.r.**
  - $n$ even:
    - 0
    - $B$
    - 1 + $n/2$ point masses

- **Upper p.r.**
  - 0
  - $B$
  - 1 + $n/2$ point masses
GIVEN: Moment conditions on random variable $X$ with support $[0,B]$

Want to bound: $E[g(X)]$

$E[X^0]=m_0$
$E[X^1]=m_1$
$\ldots$
$E[X^n]=m_n$
Principal Representations and Extremal Problems

GIVEN: Moment conditions on random variable $X$ with support $[0,B]$

Want to bound: $E[g(X)]$

**THEOREM [Markov-Krein]:**

If $\{x^0,\ldots,x^n,g(x)\}$ is a Tchebycheff-system on $[0,B]$, then $E[g(X)]$ is extremized by the unique lower and upper principal representations of the moment sequence $\{m_0,\ldots,m_n\}$. 
GOAL: Tight bounds on $E[W^{M/G/k}]$ given $n$ moments of $S$

IDEA: Identify extremal distributions

RELAXATION (Light Traffic): Extremal distributions for

$$E[\min\{S_{e1},\ldots,S_{ek}\}] \text{ s.t. } E[S_i] = m_i \text{ for } i=1,\ldots,n$$

RELAXATION 2: Restrict to Completely Monotone distributions (mixtures of Exponentials)

(contains Weibull, Pareto, Gamma)

THEOREM: For all $n$.

THEOREM: For $n = 2$ or $3$.
CONJECTURE: P.R.s are extremal for $E[W^{M/G/k}]$ for all $\rho$, for all $n$, if moment constraints are integral.

Given at least $E[S]$, $E[S^2]$:

Not given $E[S^2]$, even # of moment constraints in $(0,2)$.
Simulation Results (k=4, $\rho=2.4$)

$$E[W^{M/G/k}]$$

Number of moments

Weibull
Simulation Results (k=4, $\rho=2.4,$)

The graph shows the expected number of moments, $E[W^{M/G/k}]$, for different numbers of moments. The Weibull bounds and bounds via p.r. are indicated on the graph.
Simulation Results \((k=4, \rho=2.4,\)\)

Approximation Schema:
- Refine **lower bound** via an additional **odd moment**,  
- Upper bound via **even moment** until gap is acceptable
2 moments not enough for $E[W^{M/G/k}]$

Tighter bounds via higher moments of job size distribution

Many other “hard” queueing systems fit the approximation schema
Example 1: M/G/1 Round-robin queue

THEOREM: Upper and lower p.r. extremize mean response time under $\lambda \to 0$, when $S$ is a mixture of Exponentials.
Other queuing systems exhibiting Markov-Krein characterization

Example 2: Systems with fluctuating load

THEOREM: Upper and lower p.r. extremize mean waiting time under $\alpha \rightarrow 0$, when $T_H$, $T_L$ are mixtures of Exponentials.
Example: Single server system

\[ W_{i+1} = \text{waiting time of } S_{i+1} \]

\[ W_{i+1} = \Phi(W_i, S_i, A_{i+1}) \]
Example: Single server system

\[ W_{i+1} = \text{waiting time of } S_{i+1} \]

\[ W_{i+1} = (W_i + S_i - A_{i+1})^+ \]
Q: Given moments of $S$, under what conditions on $f, \Phi$, is $E[f(W)]$ extremized by p.r.s?
Conclusions

- All existing analytical approx for performance based on 2 moments, but 2 moments inadequate
- Provide evidence for tight $n$-moments based bounds via asymptotics for M/G/k and other queuing systems
- A new problem in analysis: Markov-Krein characterization of stochastic fixed point equations