Stochastic Models and Analysis for Resource Management in Server Farms

Thesis Oral

VARUN GUPTA
Advantages of server farm architecture

- High compute capacity
- Incremental growth
- Fault-tolerance
- Efficient resource utilization
- Energy efficiency
- High parallelism
Design Choice 1: How many servers to buy? Of what capacity?

Design Choice 2: Which server to assign jobs to?

Design Choice 3: Scheduling policy for backend servers?

Design Choice 4: When to turn servers on/off for energy-efficiency?
Design Choice 1: Provisioning/Dimensioning

Design Choice 2: Load Balancing policy

Front-end load balancer/dispatcher

Design Choice 3: Scheduling policy

Back-end servers

Design Choice 4: Dynamic capacity scaling
Design Choice 2: Load Balancing policy

Design Choice 3: Scheduling policy

Design Choice 4: Dynamic capacity scaling

Design Choice 1: Provisioning/Dimensioning

OPTION 1: Trial and error/Simulations

OPTION 2: Worst-case analysis

OPTION 3: Stochastic Modeling
- have estimates for real workloads
- understanding of “what-if” scenarios
Queueing Theory: The Origins

Manual telephone exchange (< 1900)  Automatic telephone exchange (~1910)

Q: Use observed demand to dimension tel. exchanges

A.K. Erlang
Queueing Theory: The Origins

Congestion ↔ stochastic demand

Q: Use observed demand to dimension tel. exchanges

Manual telephone exchange (< 1900)  Automatic telephone exchange (~1910)

A.K. Erlang
Q: Use observed demand to dimension tel. exchanges

Assumption 1: Call durations are i.i.d. Exponentially distributed random variables

Assumption 2 (Poisson arrivals): Inter-call arrival times are i.i.d. Exponentially distributed
BUT existing queueing models are lacking for computing server farms

I. Workloads
   • Classic models assume low variability in workload

II. Architectures
   • Assume First-Come-First-Served servers
   • Scale of traditional applications much smaller than data centers
   • Dynamic capacity scaling not feasible

NEED new analysis and new models
Part I. Impact of new workloads

- New analysis for a classical multi-server model
- Broader applications of analysis technique

Part II. Impact of new architectures on:

- Concurrency control for servers
- Server management policies for energy-efficiency
- Load balancing
A classic multi-server model

First-Come-First-Serve Buffer

Waiting time ($W$)
The $M/G/k$/FCFS model

First-Come-First-Serve Buffer

Waiting time ($W$)
The $M/G/k$/FCFS model

Waiting time ($W$)

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

First-Come-First-Serve Buffer

$W$  
Waiting time

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

- **Poisson($\lambda$)**
- **First-Come-First-Serve Buffer**
- **k Homogeneous servers**
- **Waiting time ($W$)**

- $\lambda = \text{arrival rate}$
The $M/G/k$/FCFS model

First-Come-First-Serve
Buffer

Poisson($\lambda$)

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

$k$
Homogeneous
servers

Waiting time ($W$)

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

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

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

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

$$C^2 = \frac{\text{var}(S)}{E[S]^2}$$
Sq. Coeff. of Variation (SCV) > 20 for computing workloads

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

$E[W^{M/M/k}]$ via Markov chain

Case: $S \sim \text{General (M/G/k)}$
No exact analysis known

The Gold-standard approximation:

Lee, Longton (1959)

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

\[ \mathbb{E}[W_{M/G/k}] \approx \frac{C^2+1}{2} \mathbb{E}[W_{M/M/k}] \]

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

Can not provision using this approximation!

Graph showing L-L Approximation with various distributions:
- Weibull
- Lognormal
- Pareto (1.1)
- Pareto (1.3)
- Pareto (1.5)

(k=10, \( \rho=6 \), \( C^2=19 \))
2 moments not enough for $E[W^{M/G/k}]$

Tighter bounds via higher moments of job size distribution
Lee, Longton approximation:

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

GOAL: Bounds on approximation ratio
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?
Outline: Part I

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

Tightener bounds via higher moments of job size distribution
GOAL: Identify the "extremal" distributions with given moments

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

IDEA: Light-traffic asymptotics ($\lambda \to 0$)
GOAL: Tight bounds on $E[W^{M/G/k}]$ given $n$ moments of $S$
IDEA: Identify extremal distributions

RELAXATION: Light Traffic

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

**Given:** Moment conditions on random variable $X$ with support $[0,B]$.

- $E[X^0] = m_0$
- $E[X^1] = m_1$
- $E[X^n] = m_n$

---

### Lower p.r.

- $n$ even

- $0$ and $B$ with $1 + \frac{n}{2}$ point masses

### Upper p.r.

- $n$ even

- $0$ and $B$ with $1 + \frac{n}{2}$ point masses
GIVEN: Moment conditions on random variable \( X \) with support \([0, B]\)

Want to bound: \( \mathbb{E}[g(X)] \)

\[
\begin{align*}
\mathbb{E}[X^0] &= m_0 \\
\mathbb{E}[X^1] &= m_1 \\
&\quad \vdots \\
\mathbb{E}[X^n] &= m_n
\end{align*}
\]

**THEOREM [Markov-Krein]:**

If \( \{x^0, \ldots, x^n, g(x)\} \) is a Tchebycheff-system on \([0, B]\), then \( \mathbb{E}[g(X)] \) 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

$\lambda \to 0$

**THEOREM:**
- For $n = 2$ or $3$

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

(contains Weibull, Pareto, Gamma)

**THEOREM:**
- For all $n$. 
Simulation Results (k=4, ρ=2.4)

Approximation Schema:
Refine lower bound via an additional odd moment,
Upper bound via even moment until gap is acceptable
Outline: Part I

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

Incomplete jobs

Poisson(λ) arrivals

Jobs served for q units at a time

THEOREM: Upper and lower p.r. extremize mean response time under $\lambda \to 0$, when $S$ is a mixture of Exponentials.
Example 2: Systems with fluctuating load

**THEOREM:** Upper and lower p.r. extremize mean waiting time under $\alpha \to 0$, when $T_H$, $T_L$ are mixtures of Exponentials.
Part I. Impact of new workloads
  • New analysis for a classical multi-server model
  • Broader applications of analysis technique

Part II. Impact of new architectures on:
  • Concurrency control for servers
  • Dynamic server management for energy-efficiency
  • Load balancing
TRADITIONAL
(e.g., manufacturing, call centers)

A. FCFS servers

NEW
(Computing)

B. Processor sharing servers

Ideal time-sharing

Ideal time-sharing

C. Small + homogeneous farms

Large + heterogeneous farms

D. Dynamic scaling for energy efficiency not feasible

Servers with sleep states for energy efficiency
Application: Concurrency control in database servers

Contribution 1: Heuristic concurrency control algorithm under static arrival rate

Contribution 2: A simple traffic-oblivious heuristic

High variability workload $\Rightarrow$ K* is suboptimal

Current Queue Length

Server speed

# jobs at server

Concurrent level

K* = # jobs at server

Concurrent level

K* = $\# \text{jobs at server}$
**Contribution 1:** Join-the-Shortest-Queue (JSQ) near optimal for homogeneous servers

**Contribution 2:** JSQ is optimal for heterogeneous servers as size $\to \infty$

**Contribution 3:** First closed-form approximation for JSQ in many-servers regime
No existing analysis for multi-server systems with setup delays

**Contribution:** A new traffic-oblivious policy **DELAYEDOFF**

**DELAYEDOFF** also extends to
- Heterogeneous servers
- Virtual Machine management
Stochastic modeling a powerful tool to analyze and optimize computer systems... 

...but need new techniques to handle the new applications

- New workloads ⇒ new analysis

- New architectures ⇒ new models
## References

<table>
<thead>
<tr>
<th>Conference/Year</th>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>[QUESTA’11]</td>
<td>V. Gupta and T. Osogami,</td>
<td>On Markov-Krein characterization of mean sojourn time in M/G/K.</td>
</tr>
</tbody>
</table>

## Other Work

<table>
<thead>
<tr>
<th>Topic</th>
<th>Conference/Year</th>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
</table>