# Stochastic Models and Analysis for Resource Management in Server Farms

Thesis Oral

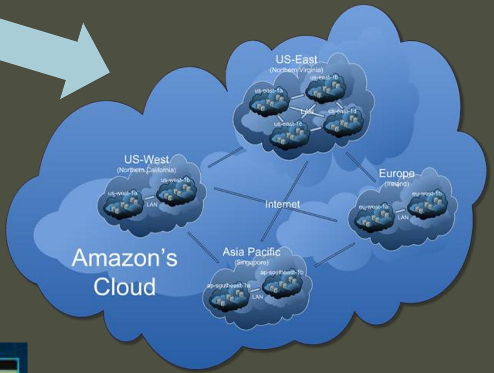
VARUN GUPTA

# Advantages of server farm architecture


Data center pods

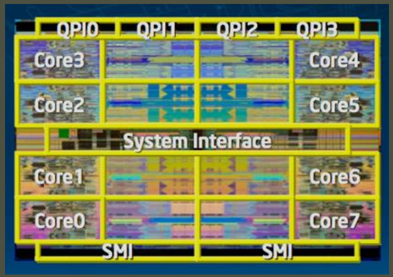
Supercomputers

+ high compute capacity
+ incremental growth
+ fault-tolerance
+ efficient resource utilization
+ energy efficiency
+ high parallelism



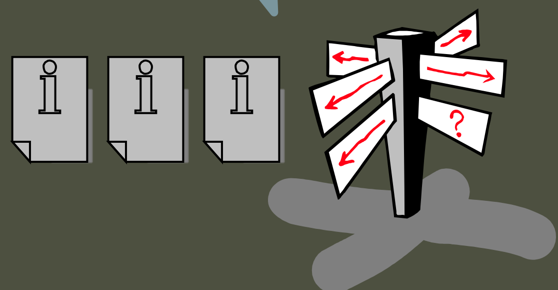

Cloud computing

Array-of-Wimpy-Nodes

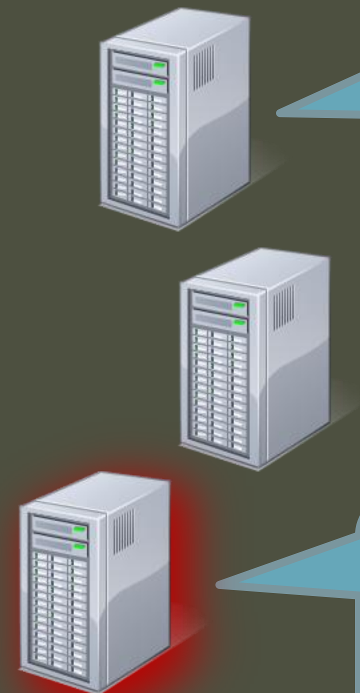



Multi-core chips

Design Choice 2:
Which server to assign jobs to?



Front-end load balancer/dispatcher

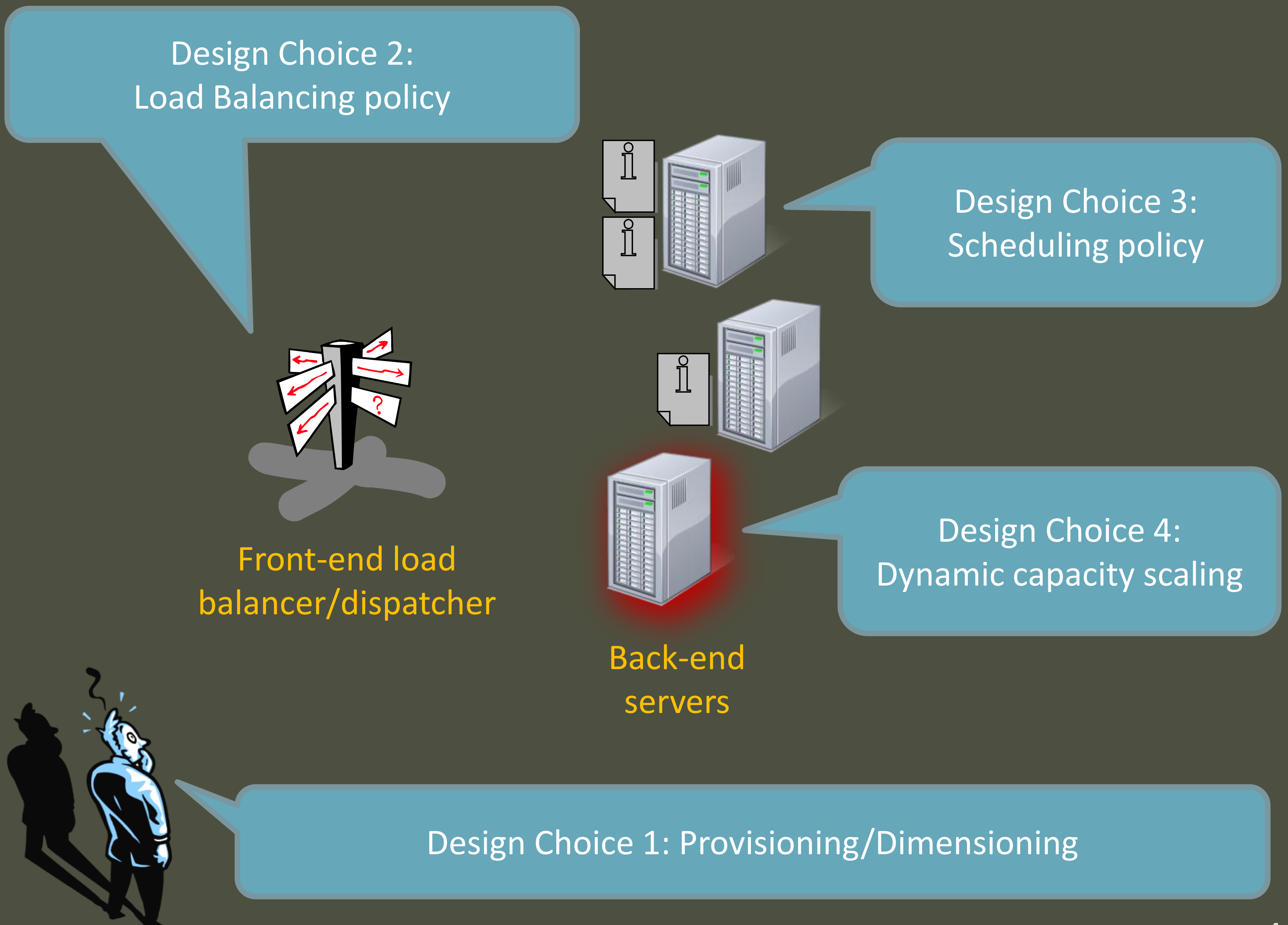


Design Choice 3:
Scheduling policy for backend servers?

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

Back-end servers

Design Choice 1: How many servers to buy? Of what capacity?


Design Choice 2:
Load Balancing policy
Design Choice 3:
Scheduling policy
Front-end load
balancer/dispatcher
Design Choice 4:
Dynamic capacity scaling
Back-end
servers
Design Choice 1: Provisioning/Dimensioning

Design Choice 2:
Load Balancing policy

Design Choice 3:
Scheduling policy

**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**

Design Choice 4:
Dynamic capacity scaling

Design Choice 1: Provisioning/Dimensioning

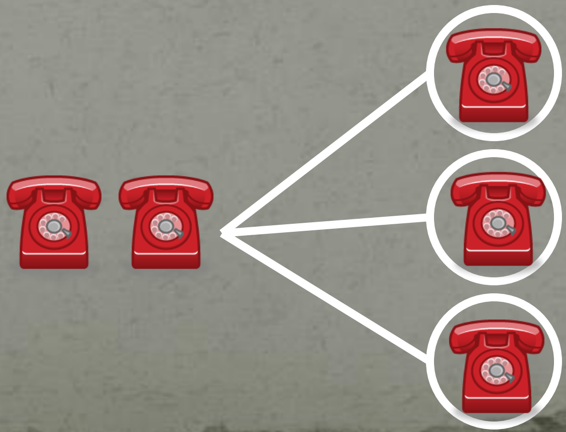
# Queueing Theory : The Origins

Manual telephone exchange (< 1900)

Automatic telephone exchange (~1910)

**Q: Use observed demand to dimension tel. exchanges**

A.K. Erlang



# Queueing Theory : The Origins

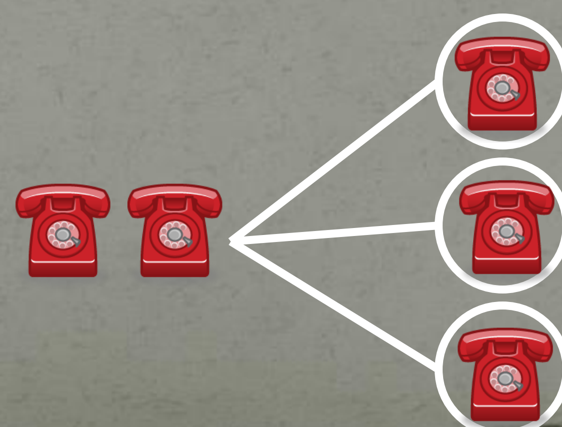
Congestion $\Leftarrow$ stochastic demand

Manual telephone exchange (< 1900)

Automatic telephone exchange (~1910)

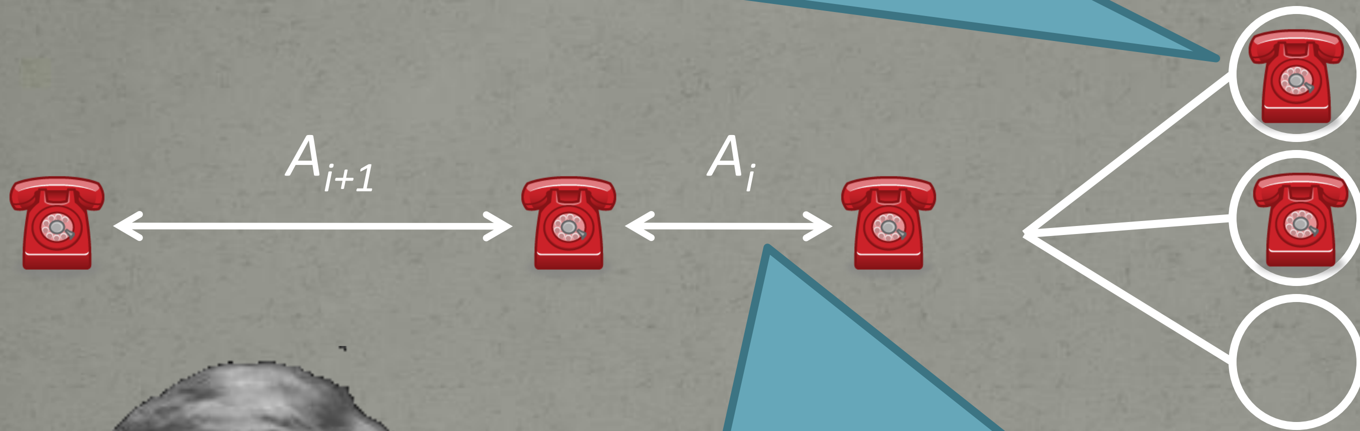
A.K. Erlang

**Q: Use observed demand to dimension tel. exchanges**



# Queueing Theory : The Origins

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

A.K. Erlang

**Q: Use observed demand to dimension tel. exchanges**

# Queueing Theory : The Origins

Assu

**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**

es

A.K. Erlang

# Outline

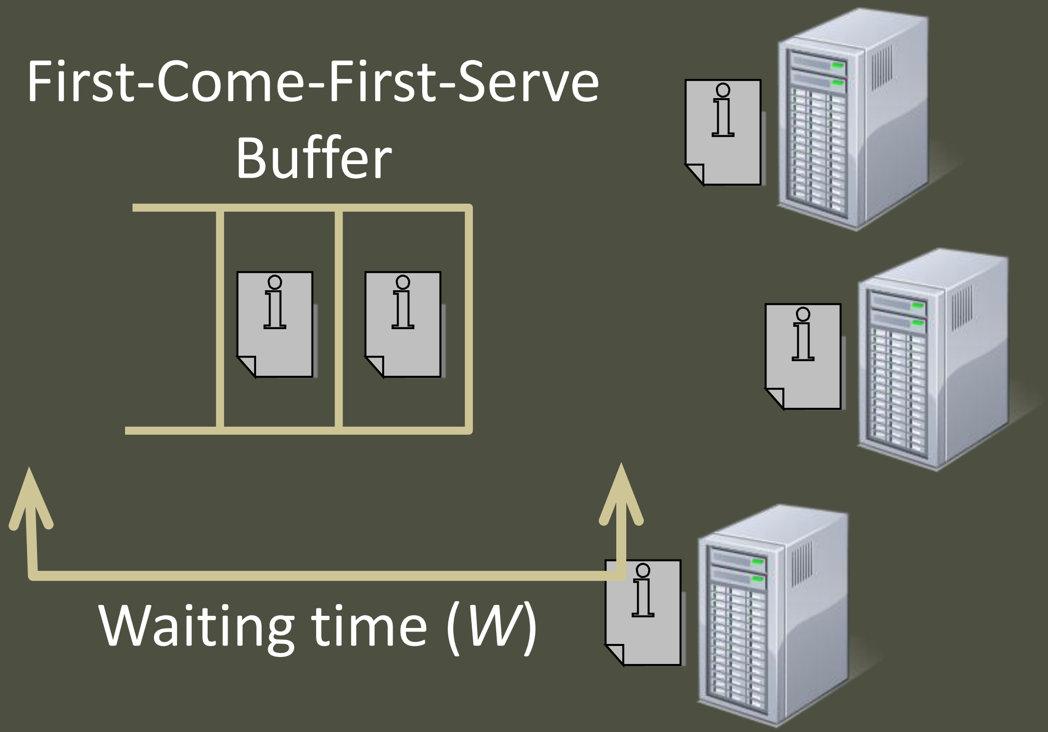
**Part I.** Impact of new workloads

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

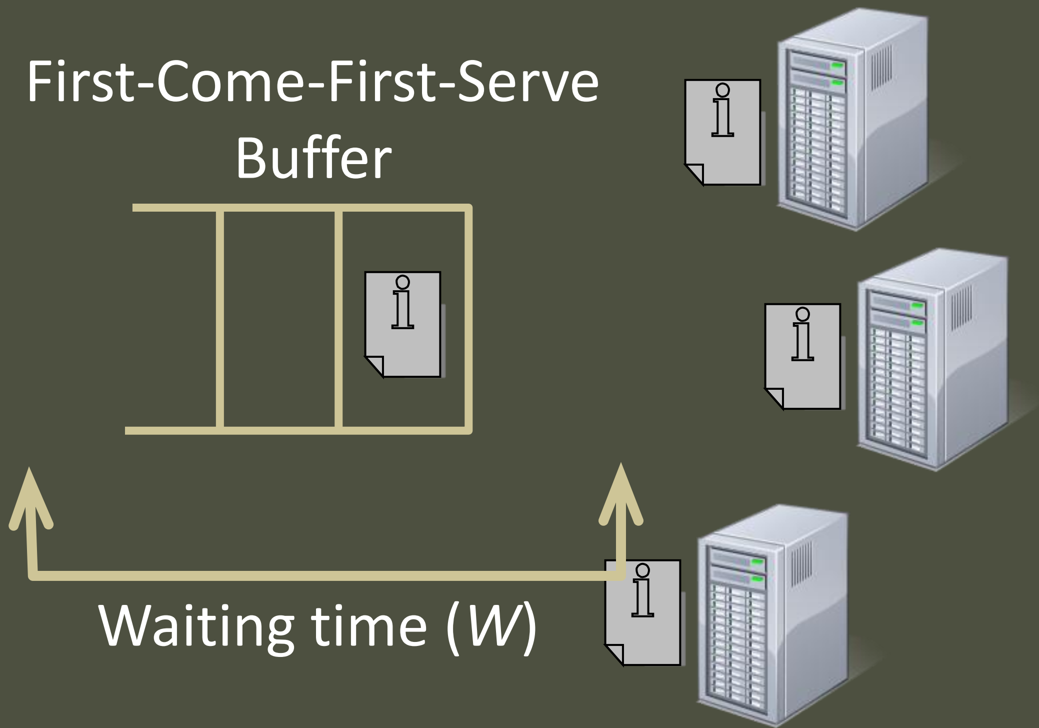
**Part II.** Impact of new architectures on:

- Concurrency control for servers
- Server management policies for energy-efficiency
- Load balancing

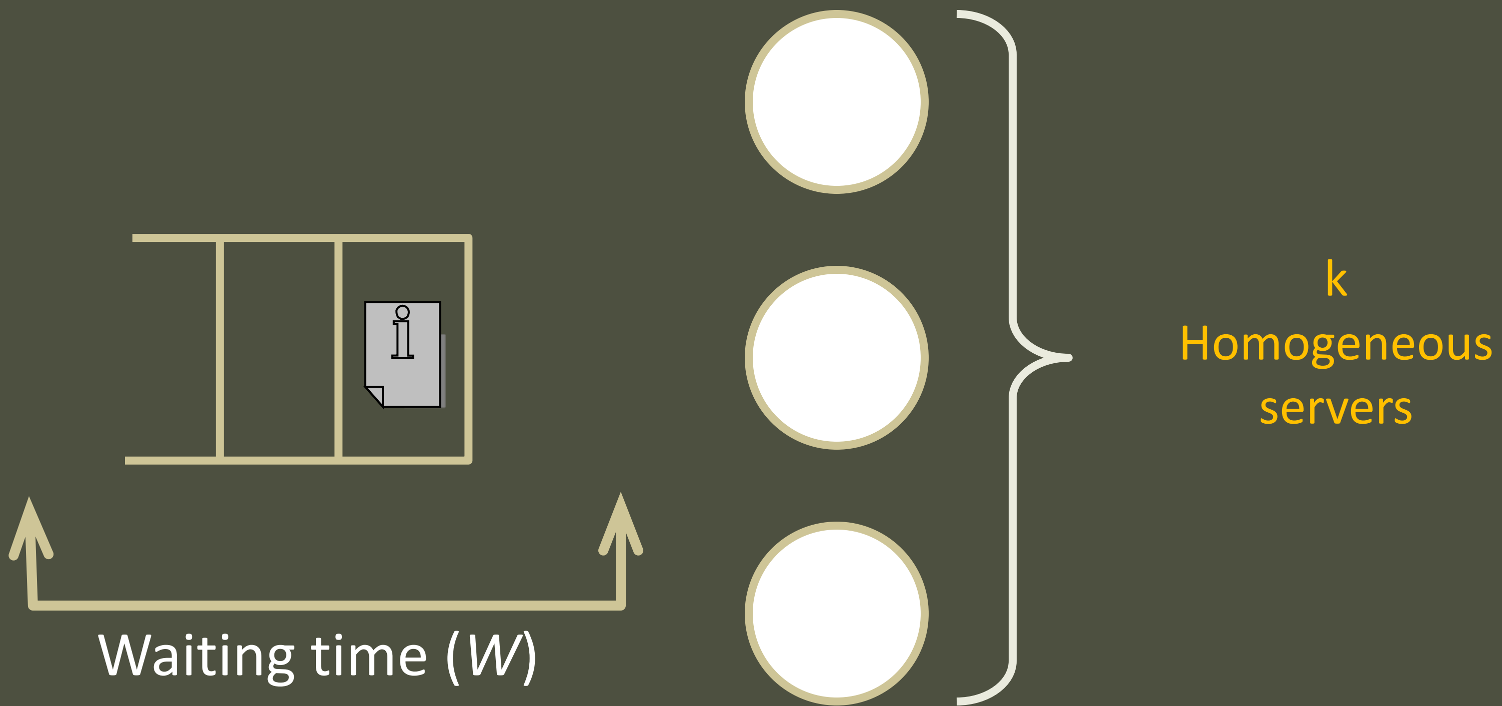
# A classic multi-server model



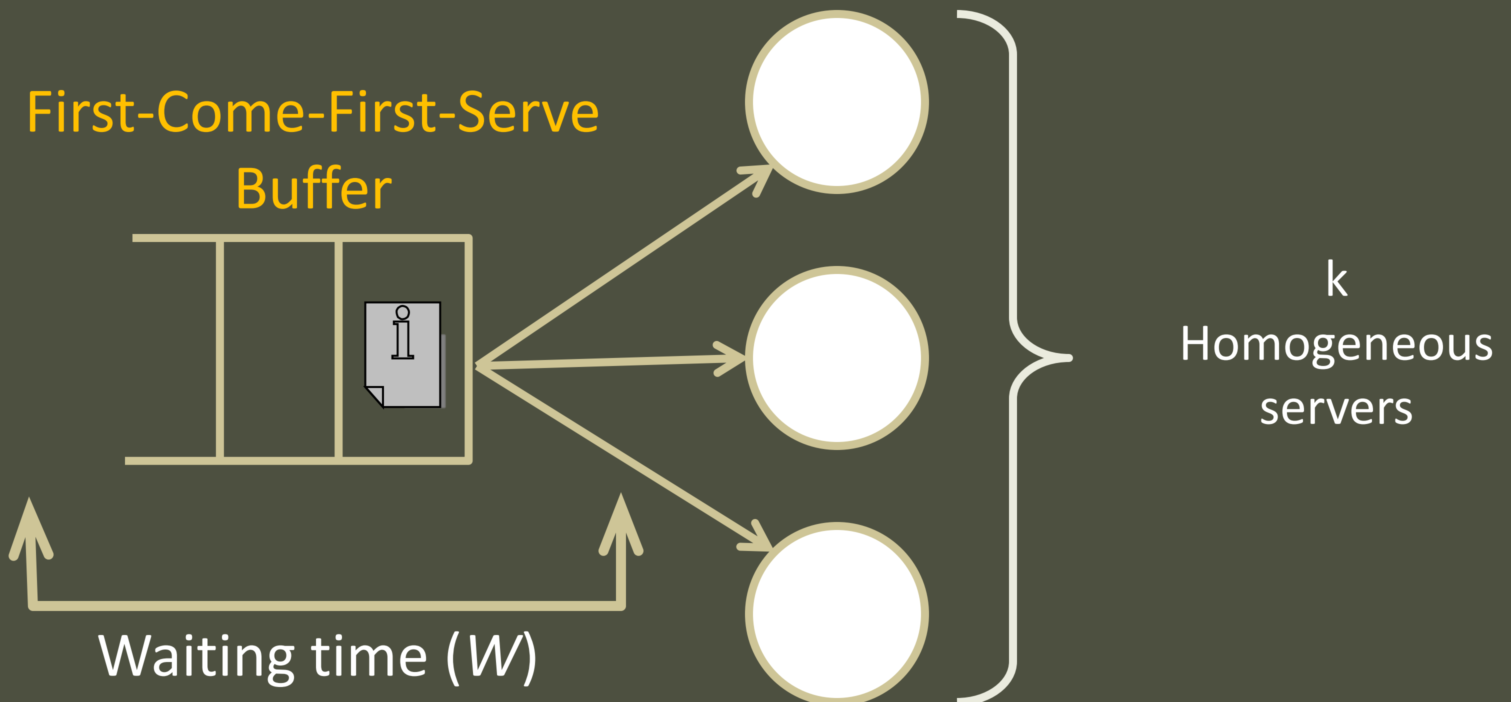
# The *M/G/k/*FCFS model



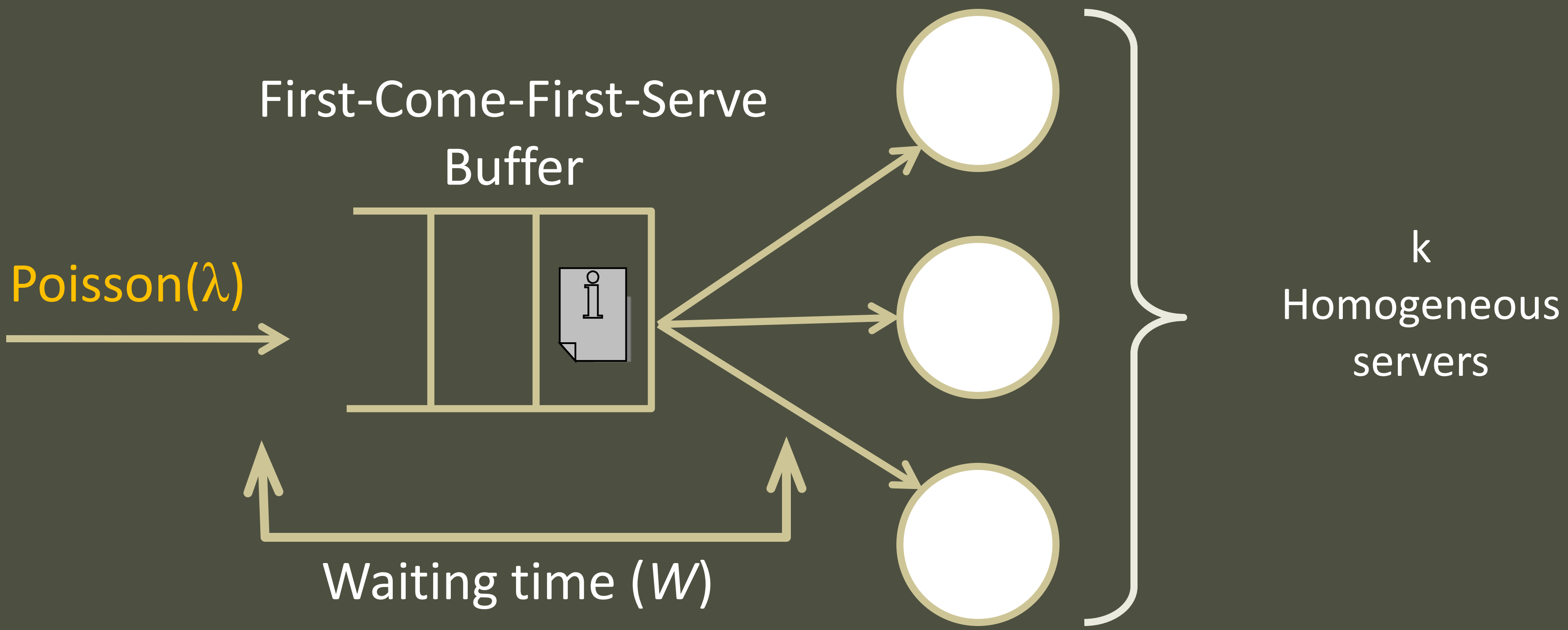
# The *M*/*G*/***k***/FCFS model



# The *M/G/k*/FCFS model

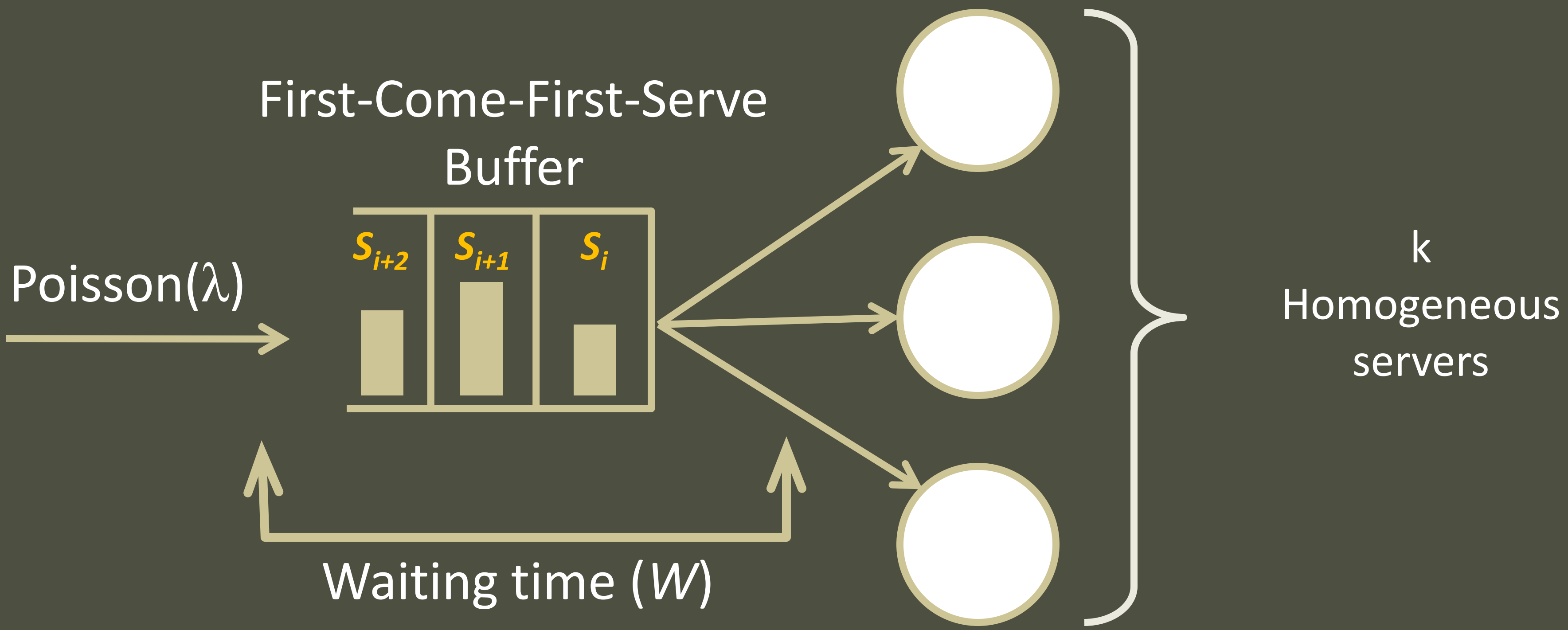


# The *M*/*G*/*k*/FCFS model



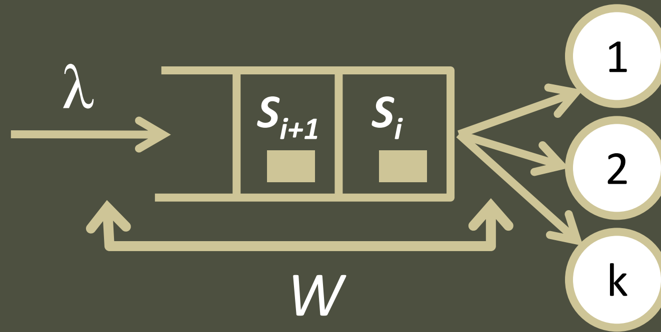

- λ = arrival rate

# The *M*/*G*/*k*/FCFS model




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

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




$\rho \equiv \lambda\, E[S]$

## k=1

**Case : *S* ~ Exponential (*M/M/1*)**

Analyze $E[W^{M/M/1}]$ via Markov chain (easy)

**Case: *S* ~ General (*M/G/1*)**

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

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

Sq. Coeff. of Variation (SCV)

> 20 for computing workloads

## k>1

**Case : S ~ Exponential (*M/M/k*)**

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

**Case: S ~ General (*M/G/k*)**

No exact analysis known

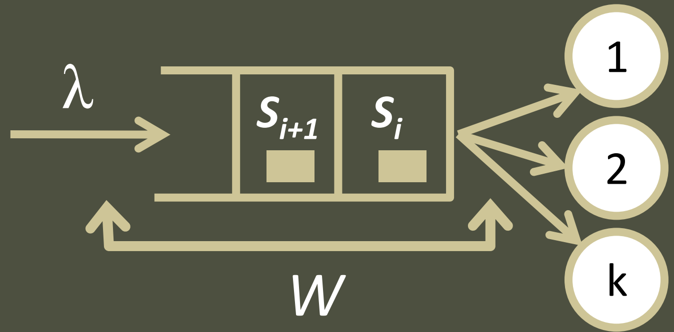
The Gold-standard approximation:

Lee, Longton (1959)

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

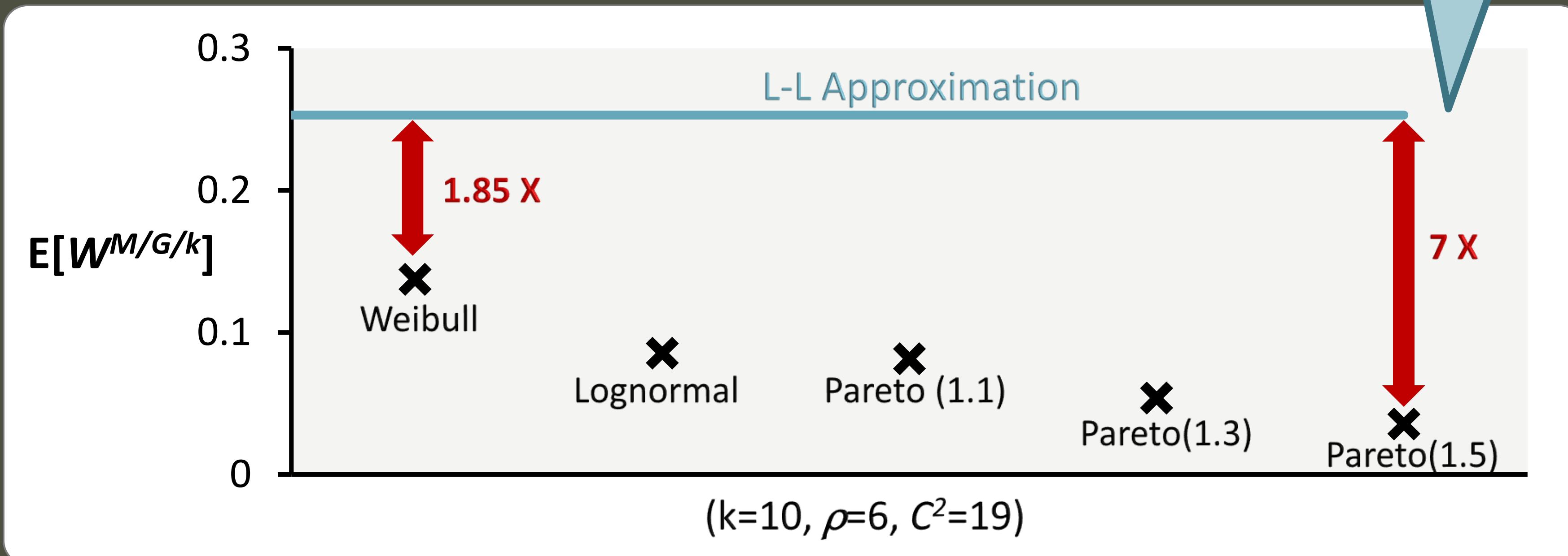
Lee, Longton approximation:

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

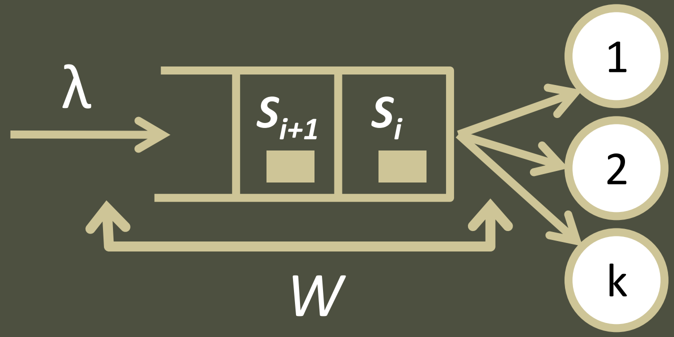



- Simple
- Exact for $k$=1
- Asymptotically tight as $\rho \rightarrow k$ (think Central Limit Thm.)

Can not provision using this approximation!



# Outline: Part I

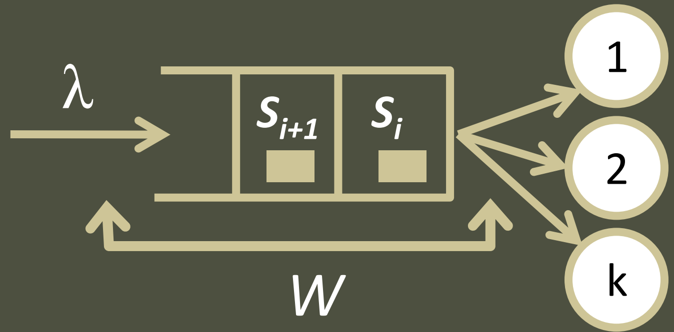



2 moments not enough for E[$W^{M/G/k}$]

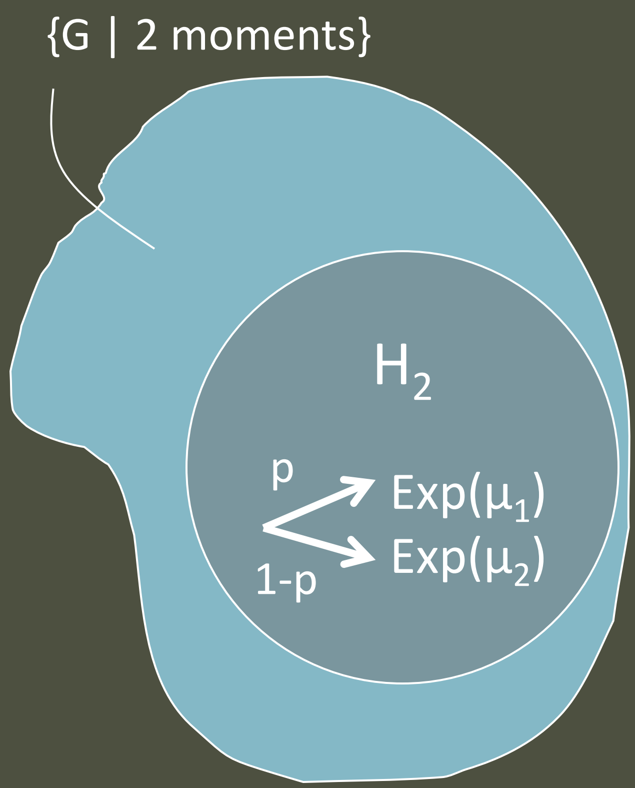
Tighter bounds via higher moments of job size distribution

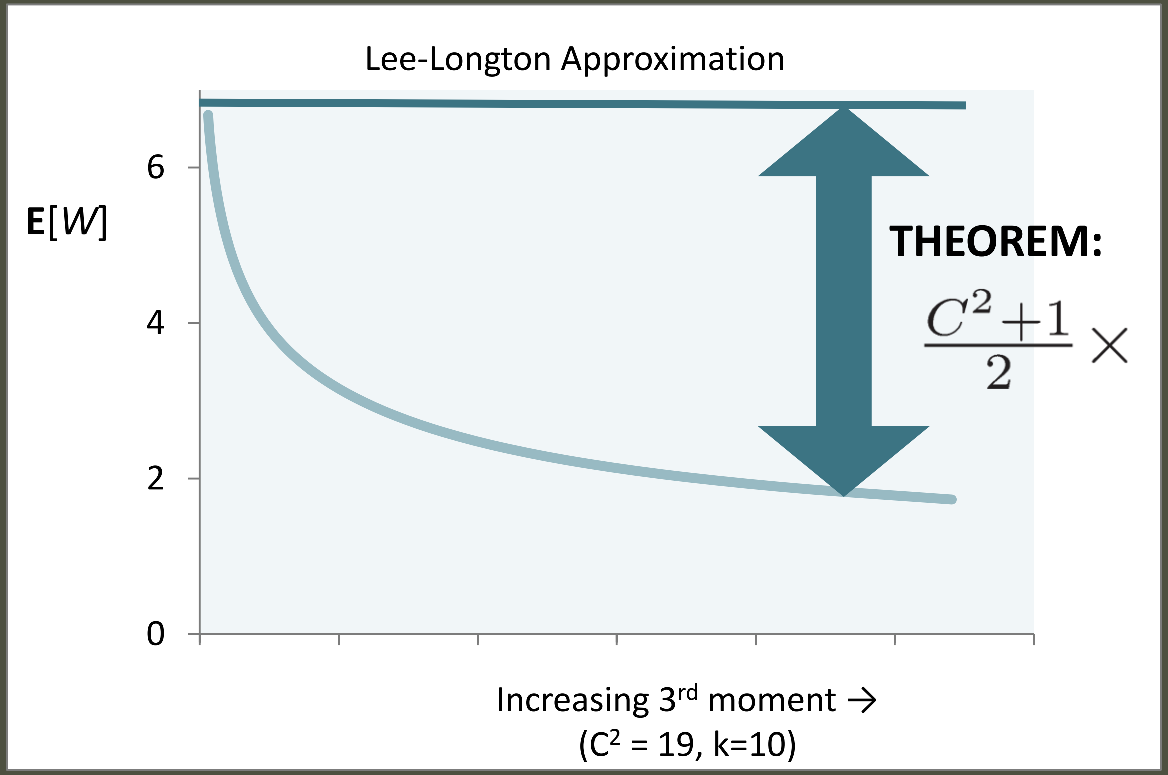
Lee, Longton approximation:

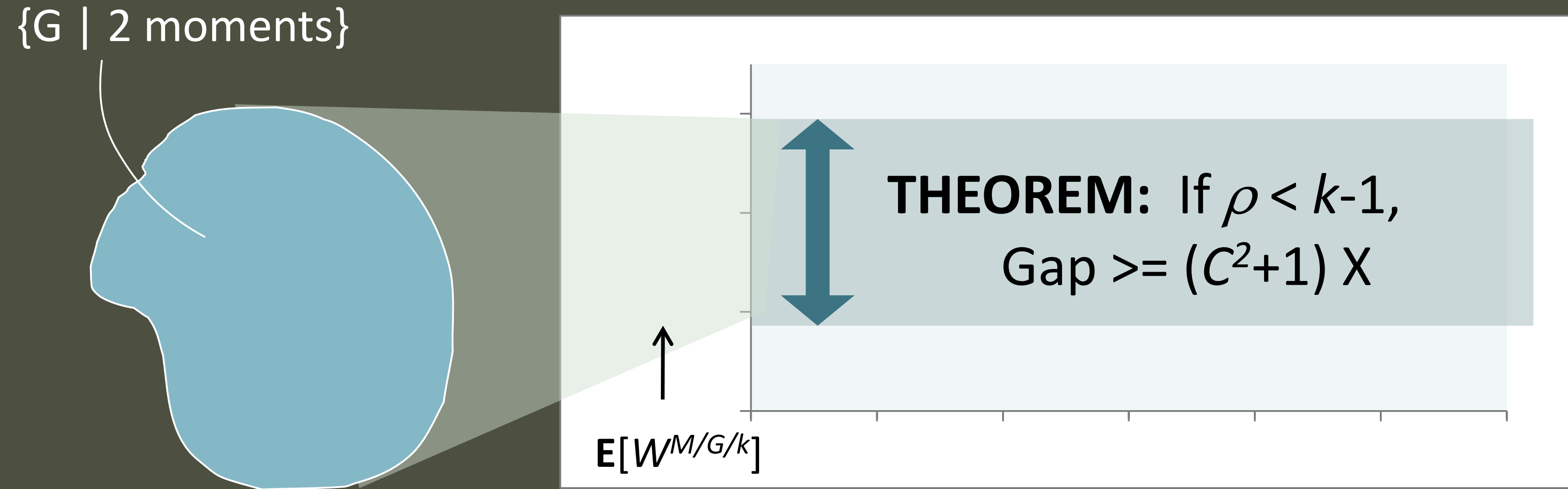
$$\mathrm{E}[W^{M/G/k}] \approx \frac{C^2+1}{2}\mathrm{E}[W^{M/M/k}]$$




GOAL: Bounds on approximation ratio



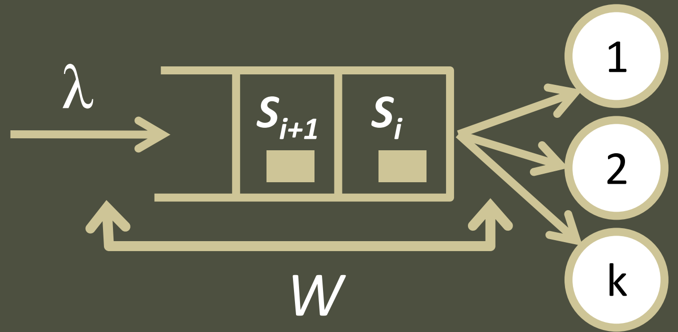



**COR.:** No approx. for E[$W^{M/G/k}$] based on first two moments of job sizes can be accurate for all distributions when $C^2$ is large

**PROOF:** Analyze limit distributions in $D_2$ ≡ mixture of 2 points


**Approximations using higher moments?**

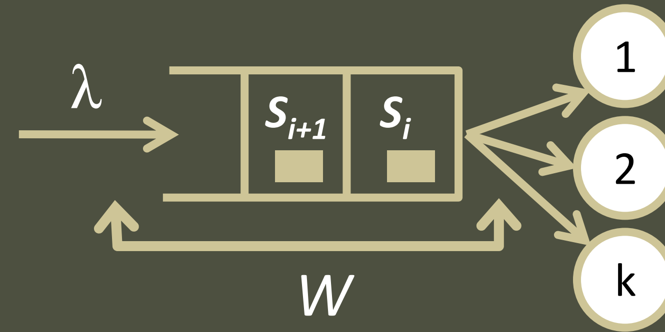
# Outline: Part I

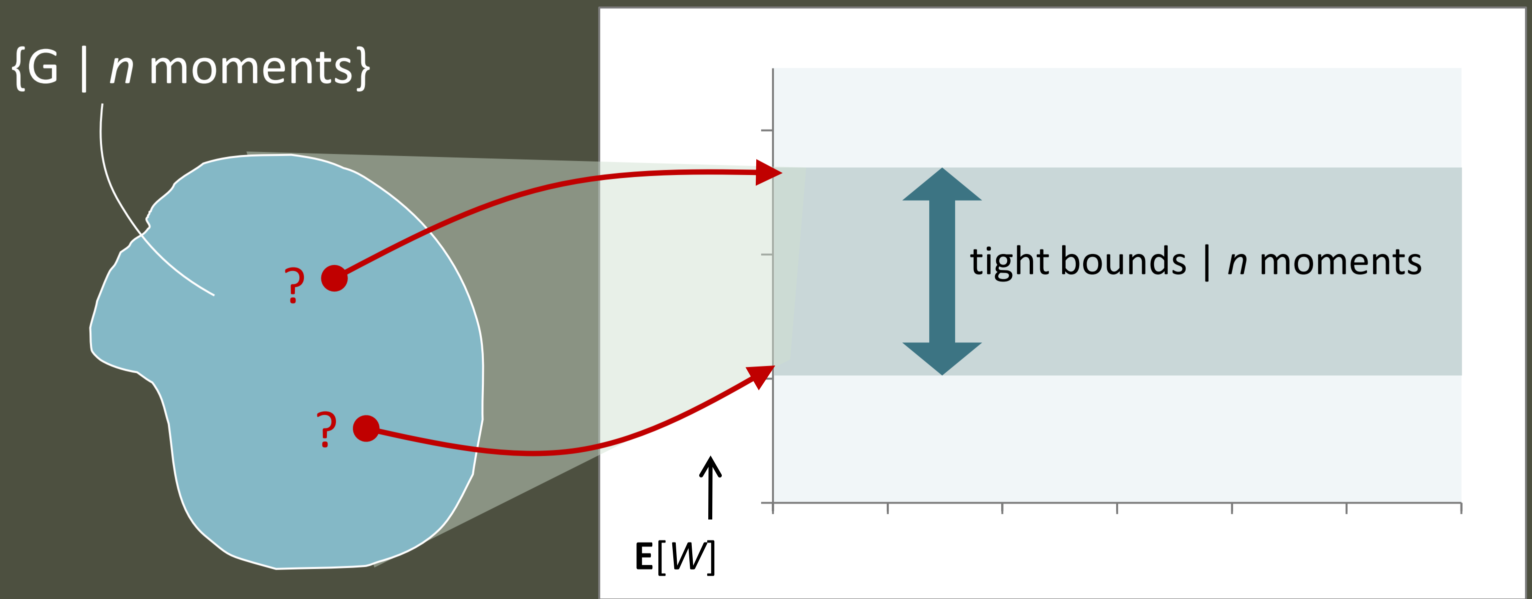



2 moments not enough for $E[W^{M/G/k}]$

Tighter bounds via higher moments of job size distribution

# Exploiting higher moments



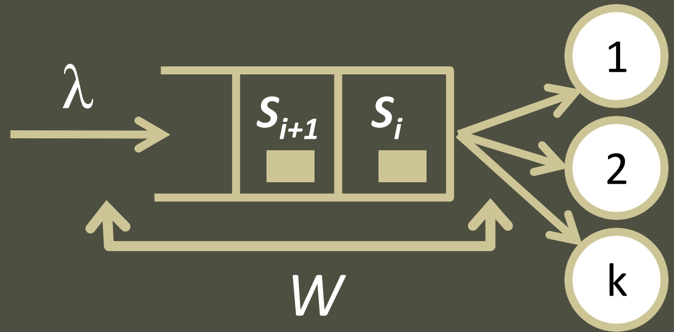




**GOAL:** Identify the "extremal" distributions with given moments

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

**IDEA:** Light-traffic asymptotics ($\lambda \to 0$)

# Where we are…




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

**IDEA:** Identify extremal distributions

**RELAXATION: Light Traffic**

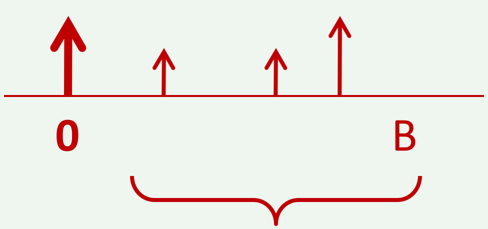
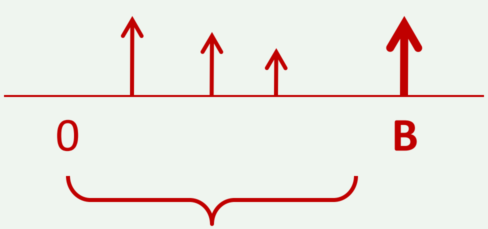
$\lambda \to 0$

# Principal Representations and Extremal Problems

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. | Upper p.r. |
|---|---|---|
| *n* even |   **1** + n/2 point masses |   **1** + n/2 point masses |

# Principal Representations and Extremal Problems

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

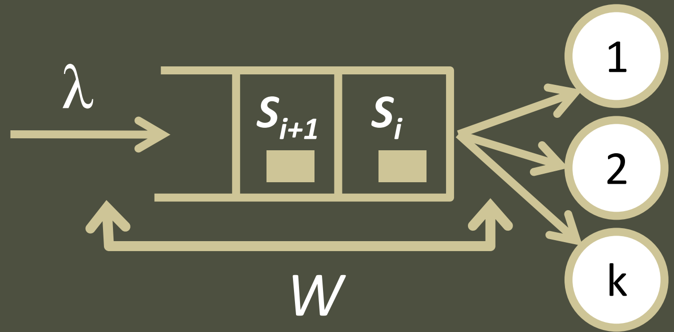
$E[X^0]=m_0$
$E[X^1]=m_1$
…
$E[X^n]=m_n$

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 $\{m_0,\ldots,m_n\}$.

# Where we are...




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

**IDEA:** Identify extremal distributions

**RELAXATION: Light Traffic**

$\lambda \to 0$

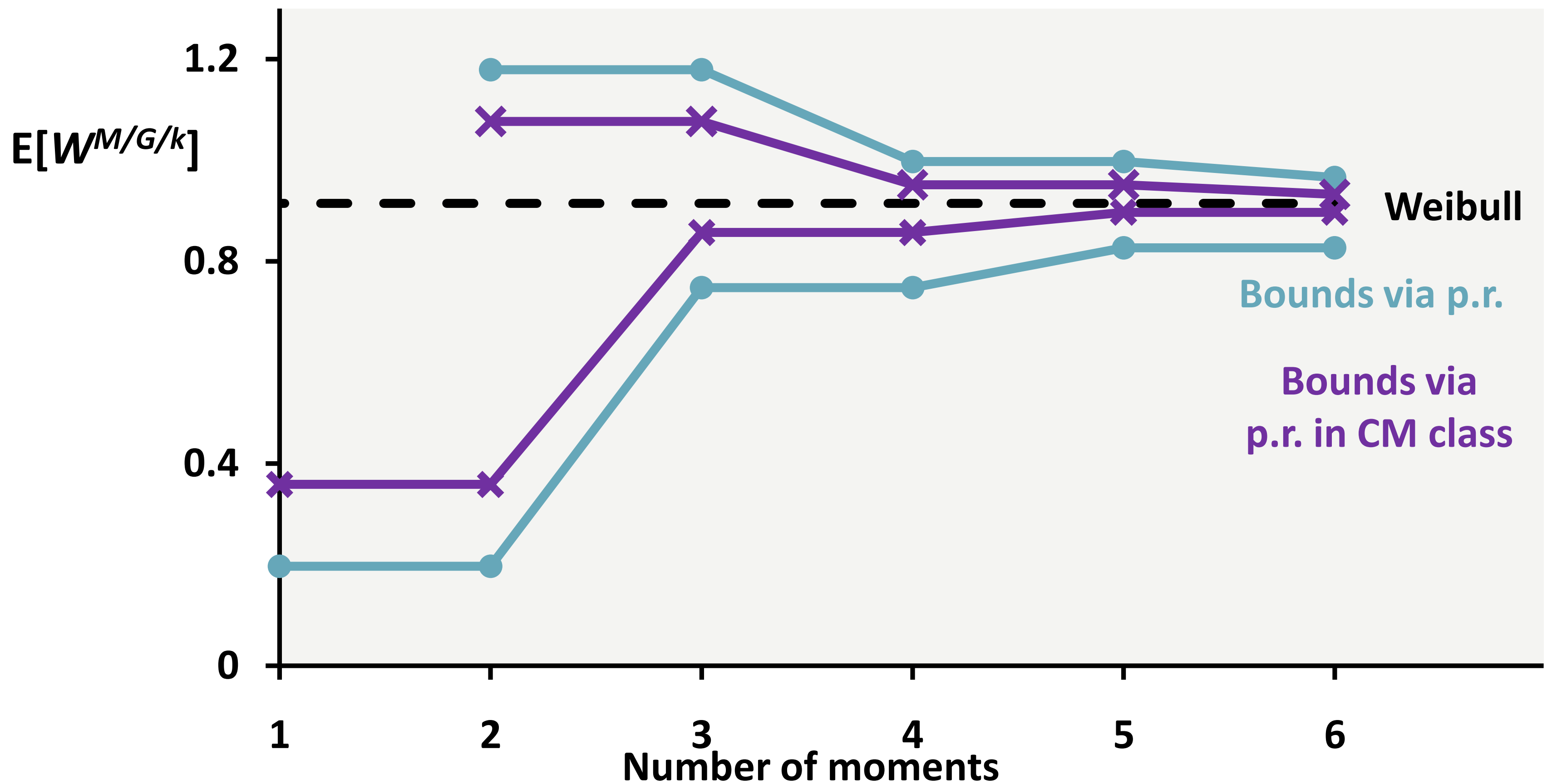
**THEOREM:**
For *n* = 2 or 3

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

(contains Weibull, Pareto, Gamma)

**THEOREM:**
For all *n*.

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

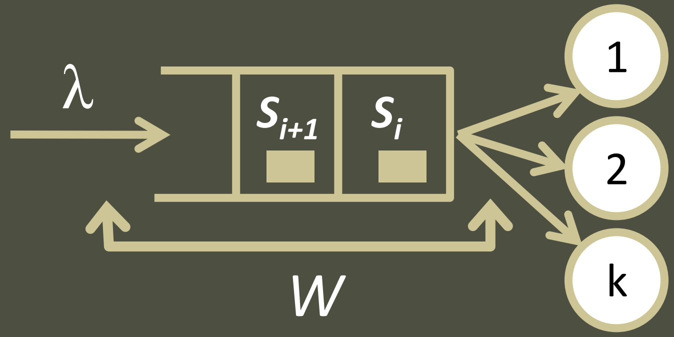



**Approximation Schema:**

Refine lower bound via an additional odd moment,

Upper bound via even moment until gap is acceptable

# Outline: Part I



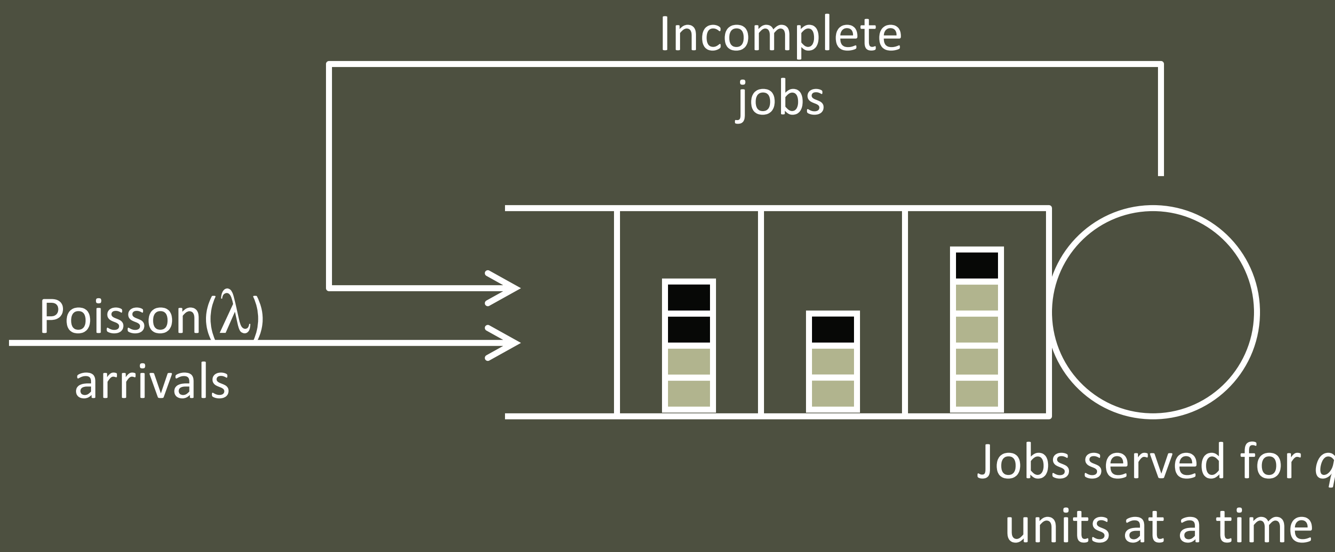

2 moments not enough for E[$W^{M/G/k}$]

Tighter bounds via higher moments of job size distribution

Many other "hard" queueing systems fit the approximation schema

# Other queuing systems exhibiting Markov-Krein characterization

Example 1: M/G/1 Round-robin queue

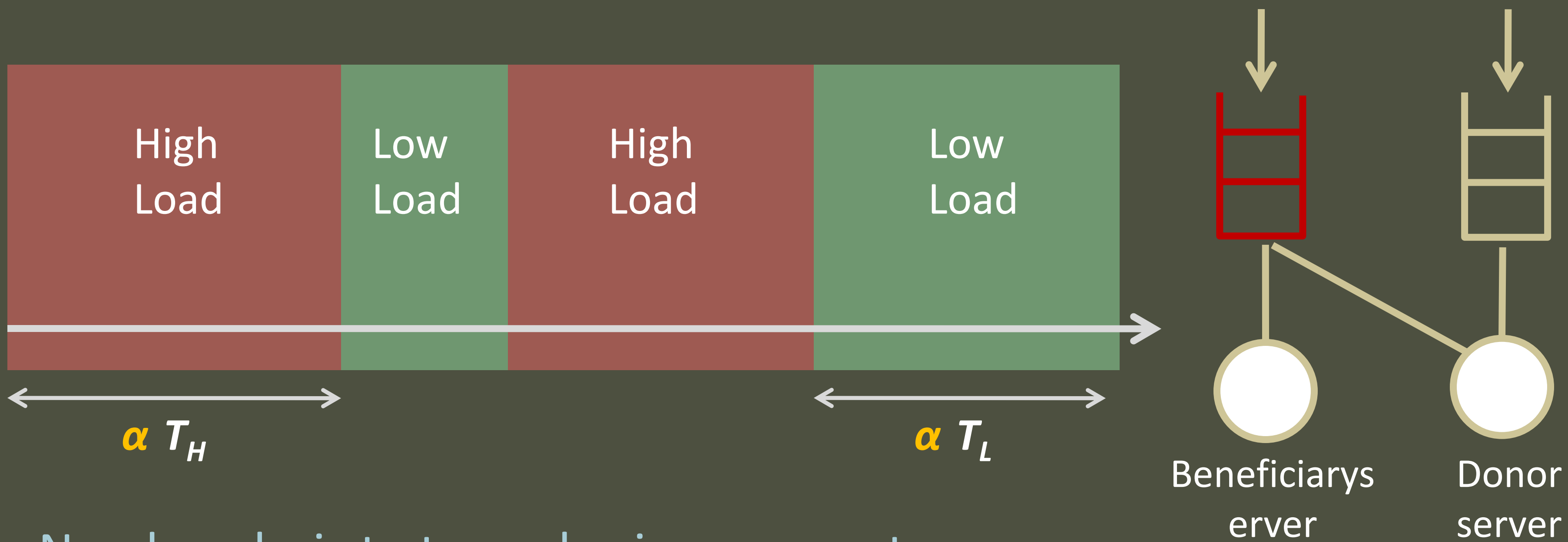



Need analysis to find *q* that balance overheads/performance

**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




Need analysis to tune sharing parameters

**THEOREM:** Upper and lower p.r. extremize mean waiting time under $\alpha \to 0$, when $T_H$, $T_L$ are mixtures of Exponentials.

# Outline

**Part I.** Impact of new workloads

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

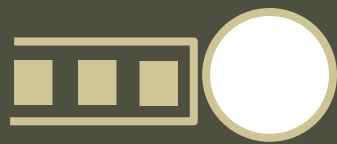
**Part II.** Impact of new architectures on:

- Concurrency control for servers
- Dynamic server management for energy-efficiency
- Load balancing

## TRADITIONAL
## (e.g.,manufacturing, call centers)

## NEW
## (Computing)

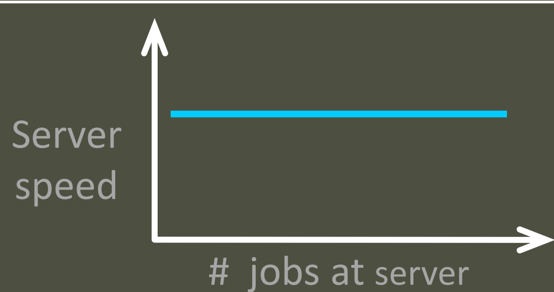
A.



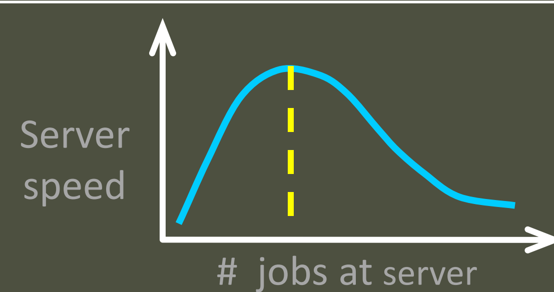
FCFS servers

Processor sharing servers

B.

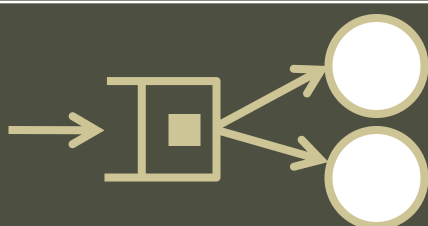



Ideal time-sharing

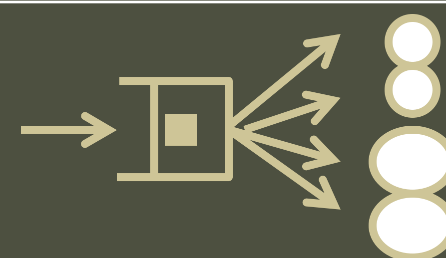



"Thrashing"

C.



Small + homogeneous farms



Large + heterogeneous farms

D.

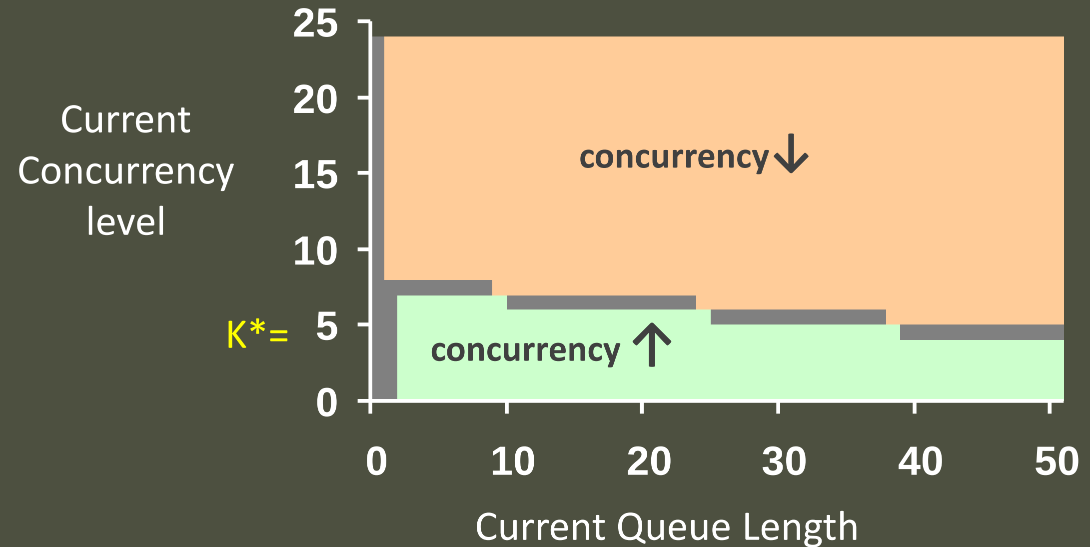
Dynamic scaling for energy efficiency not feasible

Servers with sleep states for energy efficiency

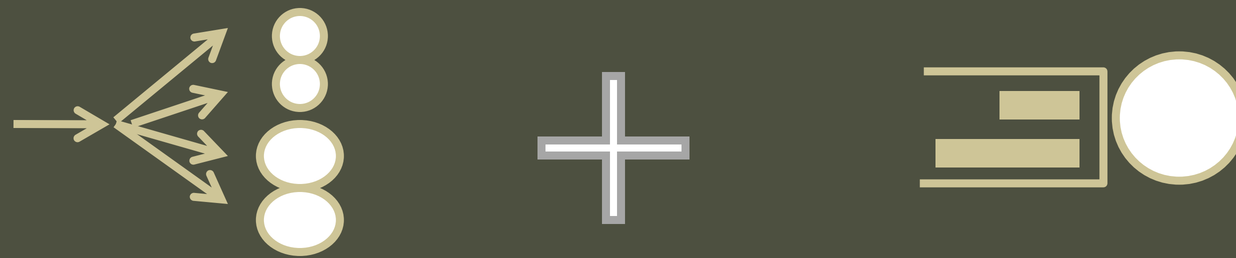
# Application: Concurrency control in database servers


**Contribution 1:** Heuristic concurrency control algorithm under static arrival rate

**Contribution 2:** A simple traffic-oblivious heuristic



# Application: Load Balancing in web server farms



Large server farms

Processor sharing servers

**Contribution 1:** Join-the-Shortest-Queue (JSQ) near optimal for homogeneous servers

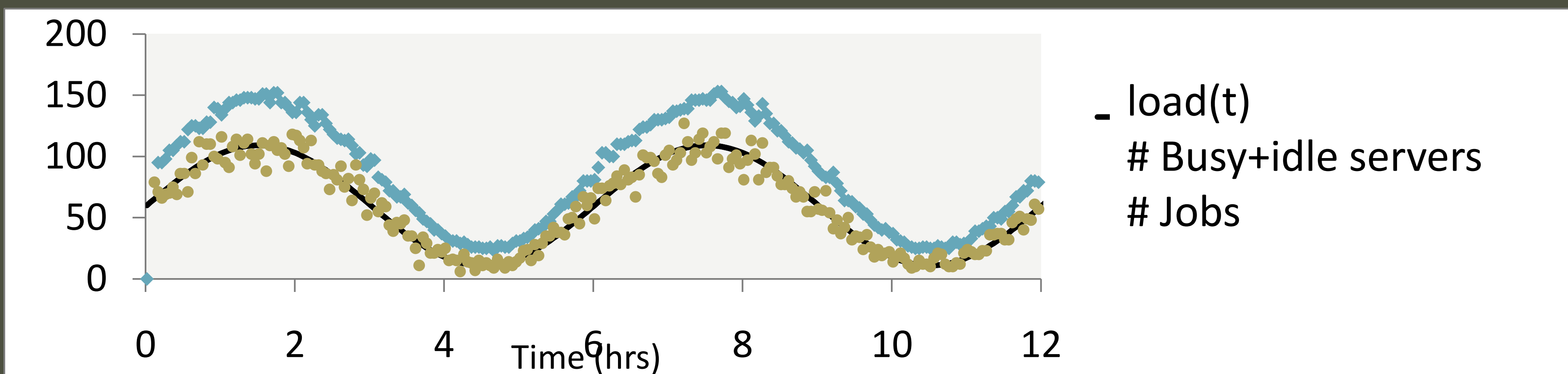
**Contribution 2:** JSQ is optimal for heterogeneous servers as size $\rightarrow \infty$

**Contribution 3:** First closed-form approximation for JSQ in many-servers regime

# Application: Dynamic capacity scaling for enery-efficiency


No existing analysis for multi-server systems with setup delays

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




**DELAYEDOFF** also extends to

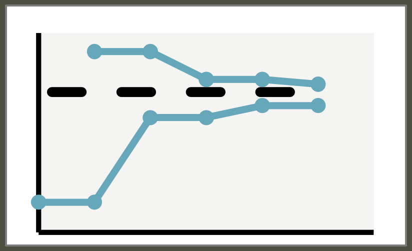
- Heterogeneous servers
- Virtual Machine management

# Summary

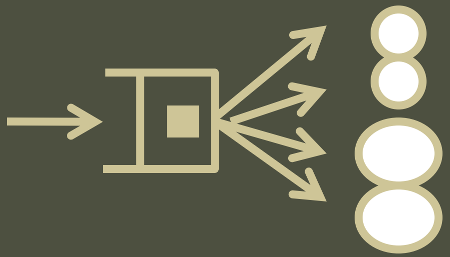
Stochastic modeling a powerful tool to analyze and optimize computer systems…

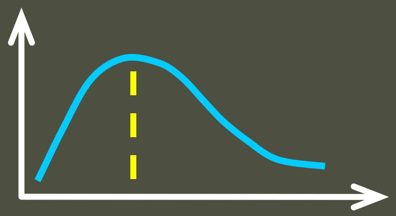
…but need new techniques to handle the new applications

- New workloads ⇒ new analysis



- New architectures ⇒ new models





# References


| | |
|---|---|
| [Performance'07] | V. Gupta, M. Harchol-Balter, K. Sigman, and W. Whitt. Analysis of join-the-shortest-queue routing for web server farms. PERFORMANCE 2007 |
| [Performance'10] | V. Gupta, A. Gandhi, M. Harchol-Balter, and M. Kozuch. Optimality analysis of energy-performance trade-off for server farm management. PERFORMANCE 2010. |
| [QUESTA'10] | 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, Vol 64, 2010. |
| [TR'08] | V. Gupta, J. Dai, M. Harchol-Balter, and B. Zwart. The effect of higher moments of job size distribution on the performance of an M/G/K queueing system. Technical Report ,CMU, 2008. |
| [Sigmetrics'09] | V. Gupta and M. Harchol-Balter. Self-adaptive admission control policies for resource-sharing systems. SIGMETRICS 2009. |
| [QUESTA'11] | V. Gupta and T. Osogami, On Markov-Krein characterization of mean sojourn time in M/G/K. |


# Other Work


| | | |
|---|---|---|
| Time-varying systems | [Sigmetrics'06] | V. Gupta, M. Harchol-Balter, A. Scheller-Wolf, and U. Yechiali. Fundamental characteristics of queues with fluctuating load. Sigmetrics 2006 |
| | [MAMA'08a] | V. Gupta, and P. Harrison. Fluid level in a reservoir with ON-OFF source. MAMA 2008. |
| Single Server Scheduling | [MAMA'08b] | V. Gupta. Finding the optimal quantum size: Sensitivity analysis of the M/G/1 round-robin queue. MAMA 2008. |
| | [Performance'10b] | V. Gupta, M. Burroughs, and M. Harchol-Balter. Analysis of scheduling policies under correlated job sizes. Performance 2010. |
| Distributed Data placement | [INFOCOM'10] | S. Borst, V. Gupta, and A. Walid. Distributed caching algorithms for Content Distribution Networks. INFOCOM 2010. |
| | [SOCC'10] | H. Amur, J. Cipar, V. Gupta, M. Kozuch, G.Ganger, and K. Schwan, Robust and flexible power-proportional storage. Symposium on Cloud Computing, 2010. |
| Epidemics | [INFOCOM'08] | M. Vojnovic, V. Gupta, T. Karagiannis, and C. Gkantsidis. Sampling strategies for epidemic-style information dissemination. INFOCOM 2008. |
| Stability analysis | | A. Busic, V.Gupta, and J. Mairesse. Stability of the bipartite matching model. Under Submission |