Tight Moments-based Bounds for Queueing Systems

VARUN GUPTA

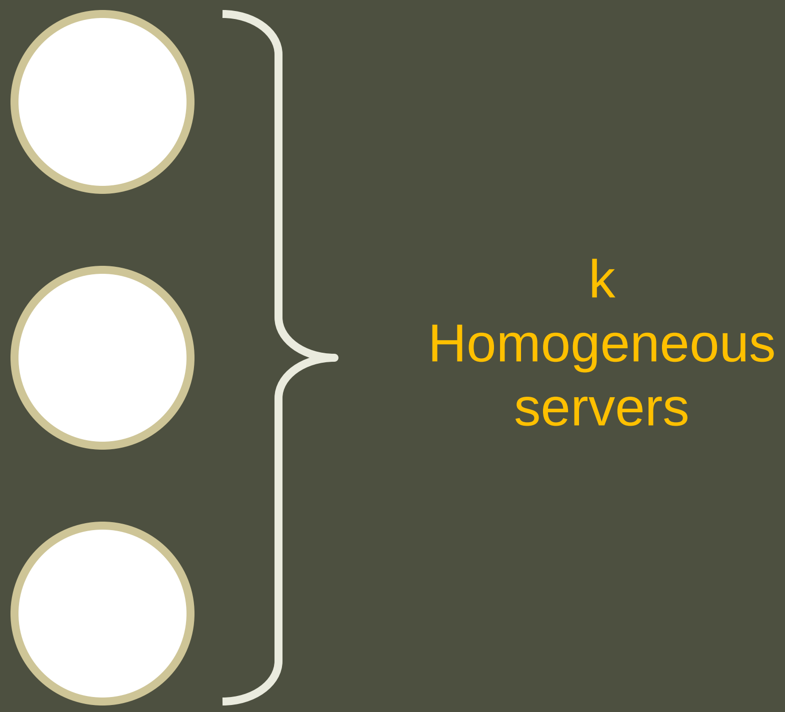
Carnegie Mellon ⟶ Google Research - - → University of Chicago Booth School of Business

With:

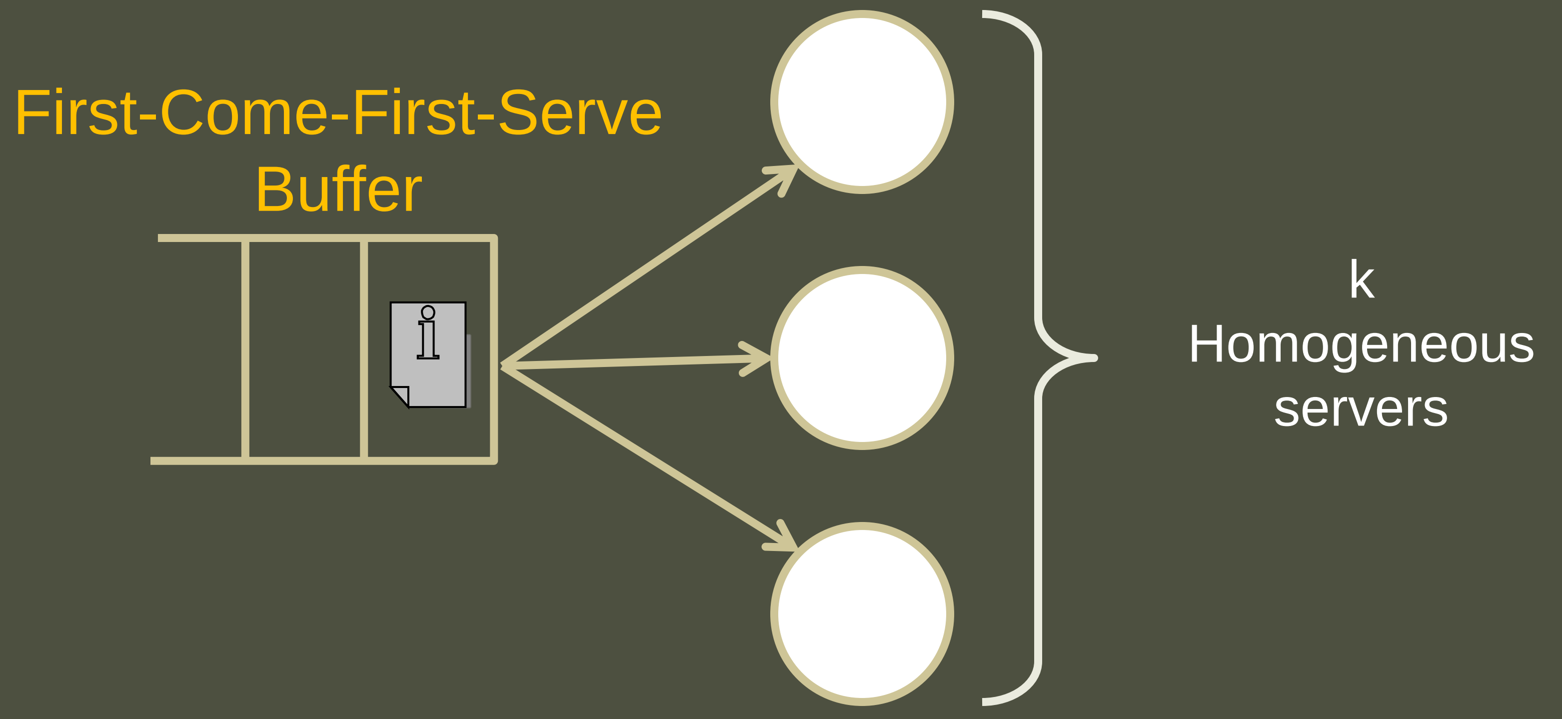
Takayuki Osogami
(IBM Research-Tokyo)

The $M/G/k/$FCFS model

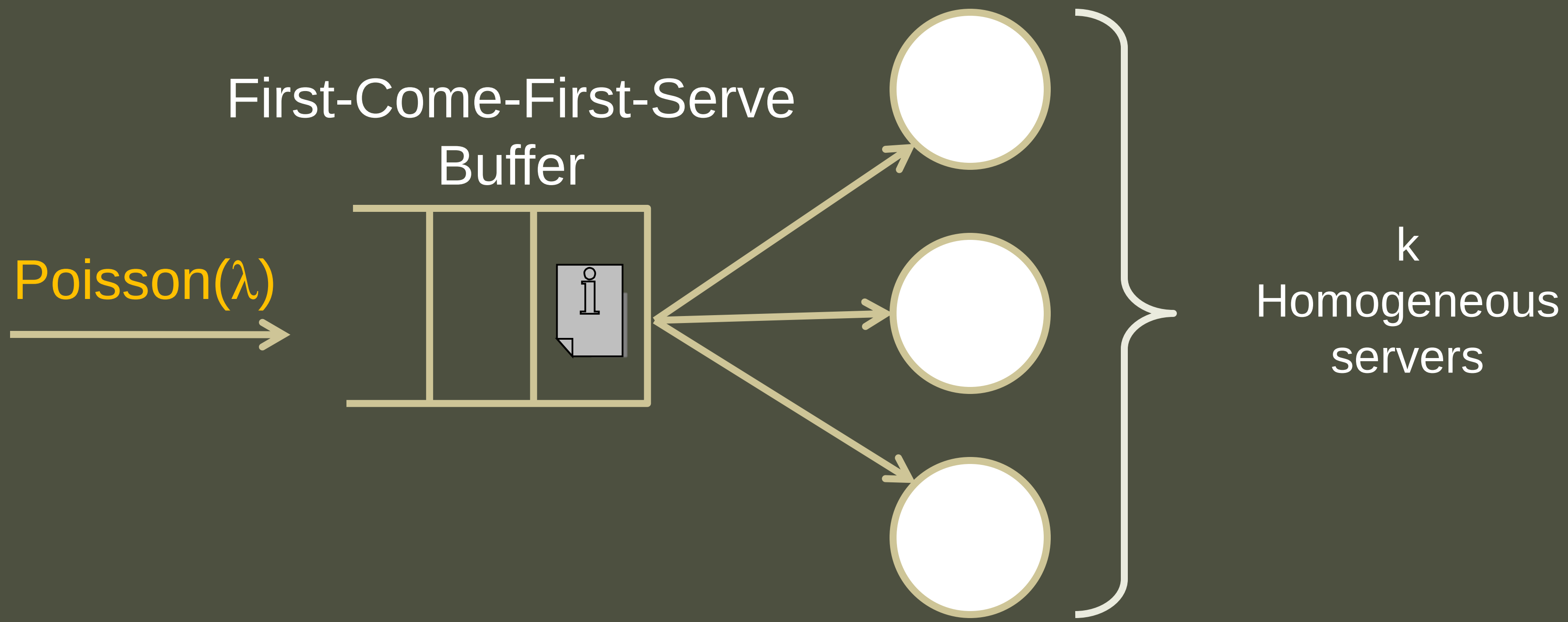
The *M*/*G*/***k***/FCFS model



The *M/G/k/*FCFS model

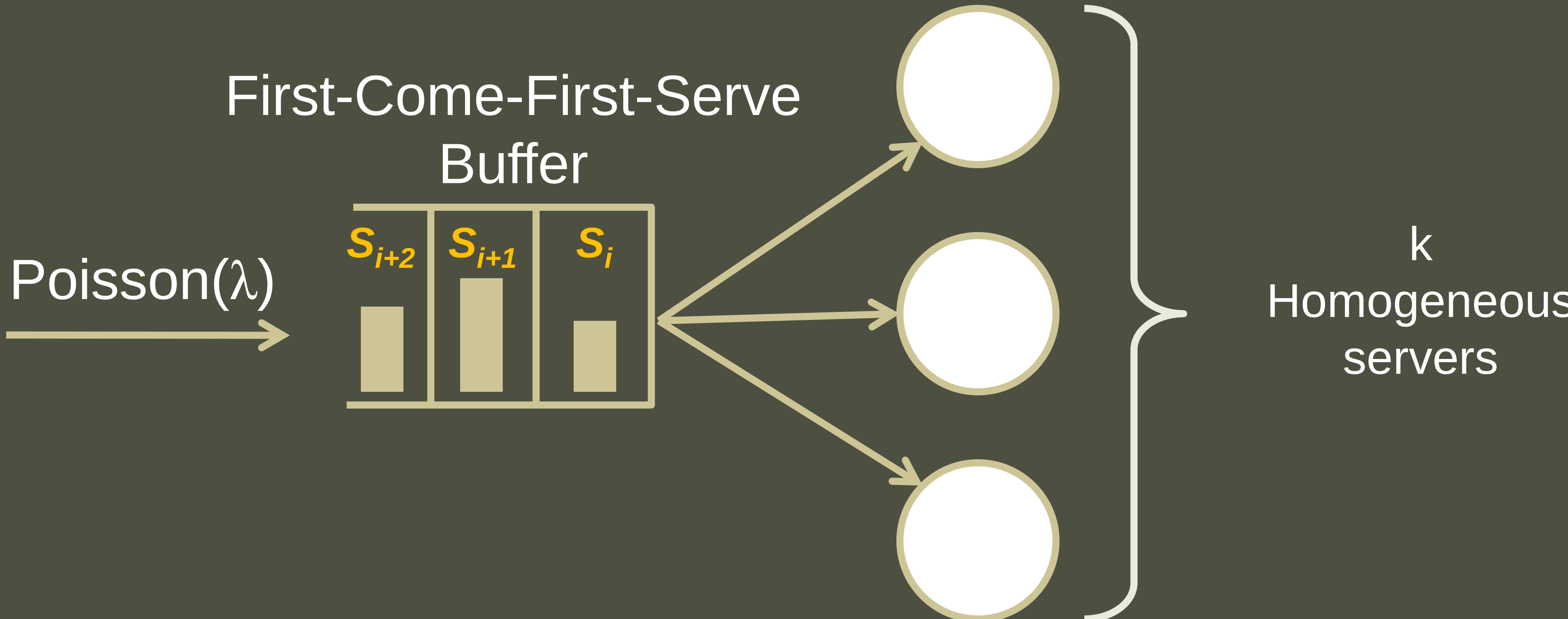


The *M*/G/*k*/FCFS model



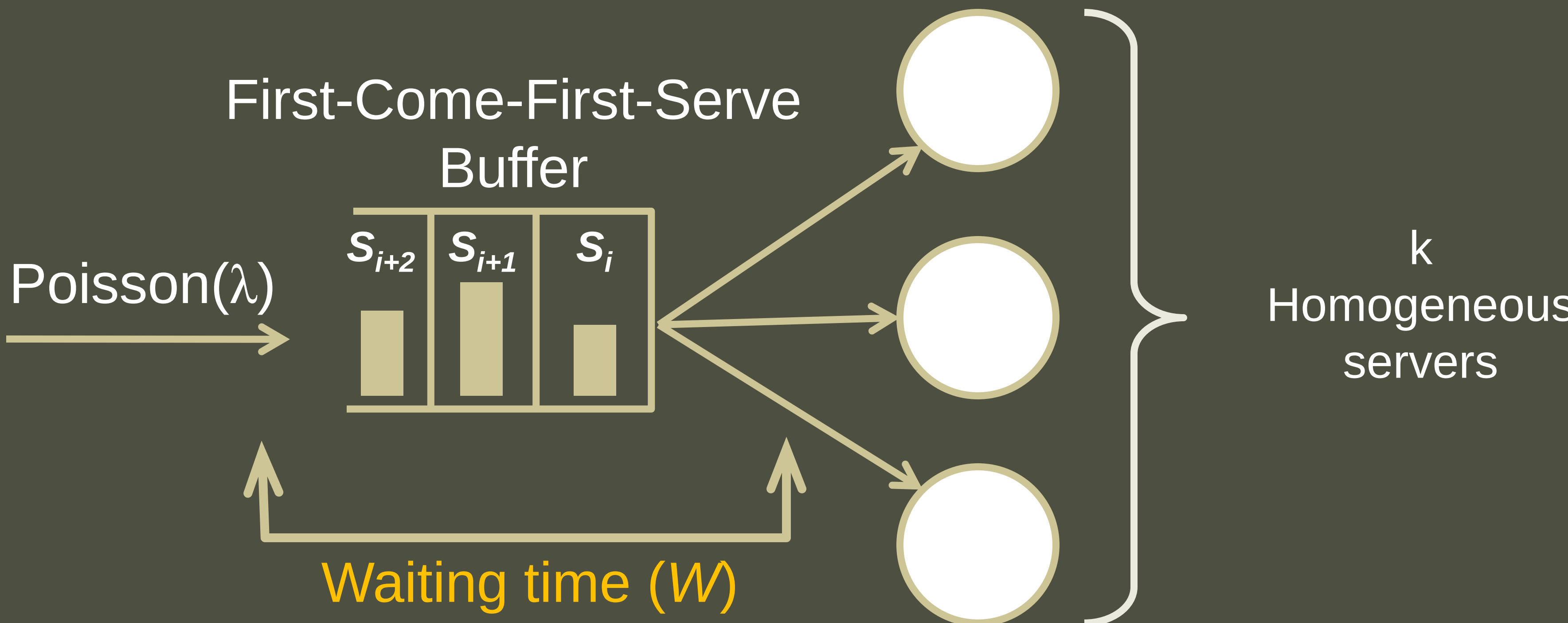
- λ = arrival rate

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



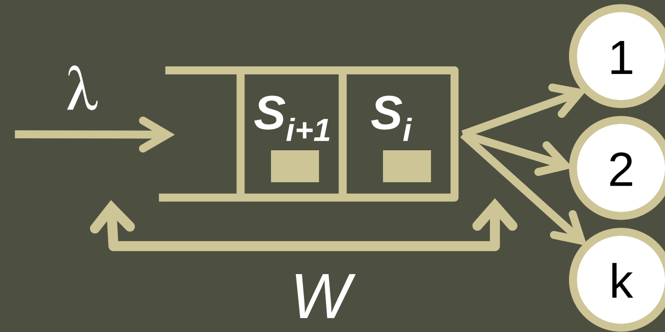
- λ = arrival rate
- job sizes (S_1, S_2, …) i.i.d. samples from S
- “load” $\rho \equiv \lambda\, E[S]$

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



$\rho \equiv \lambda\, \mathrm{E}[S]$

k=1

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

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

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

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

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

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

k>1

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

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

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

No exact analysis known

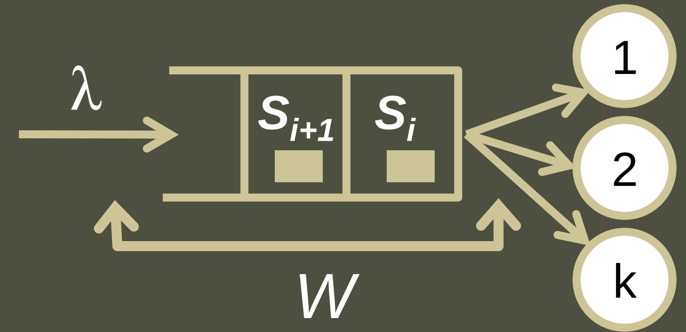
The Gold-standard approximation:

Lee, Longton (1959)

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

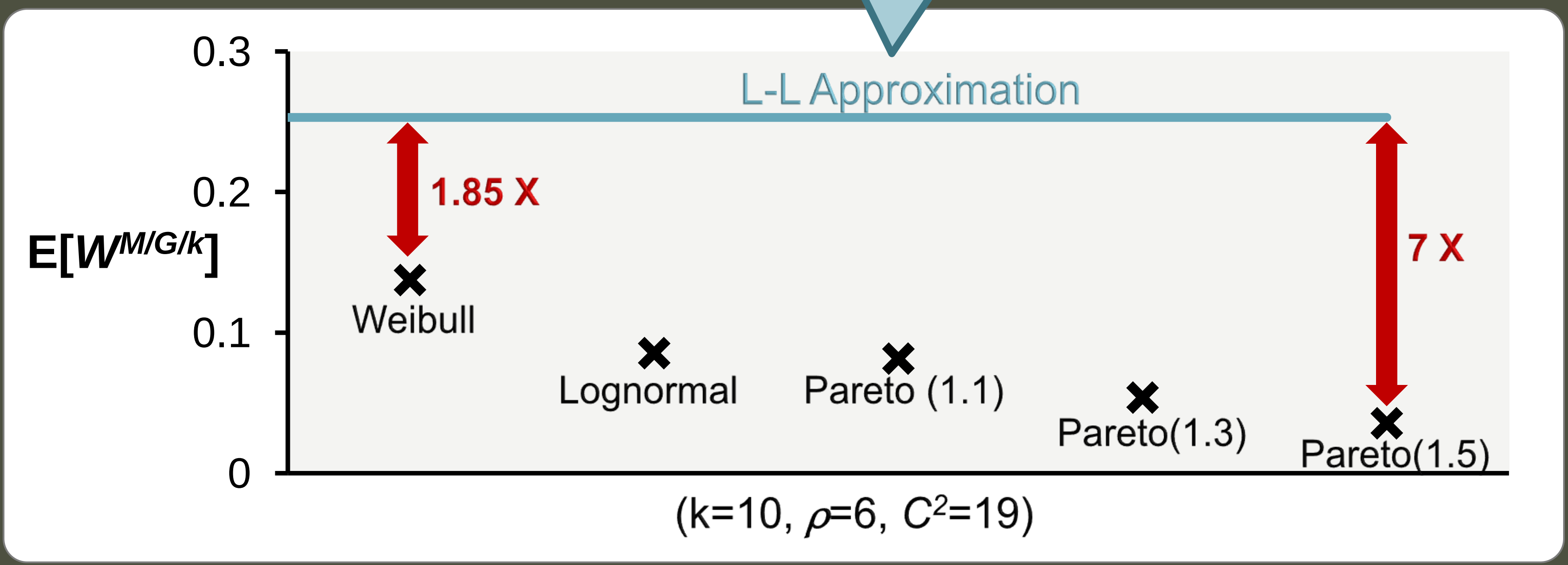
Lee, Longton approximation:

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

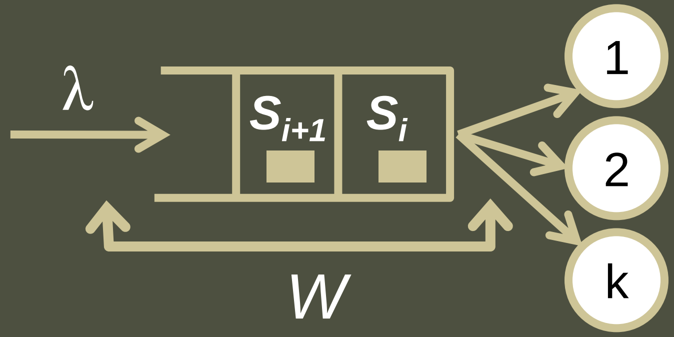


- Simple
- Exact for k=1
- Asymptotically tight as $\rho \rightarrow k$ (entral Limit Thm.)

Can not provision using this approximation!



Outline

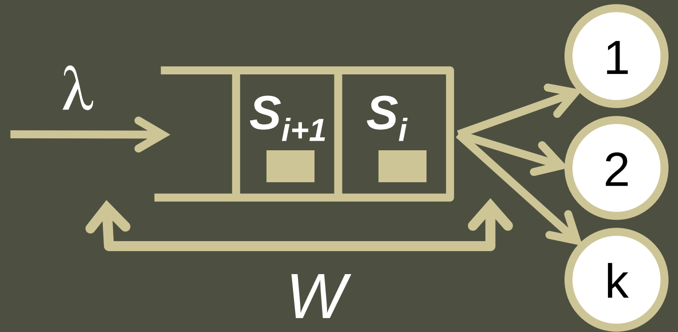


2 moments not enough for $E[W^{M/G/k}]$

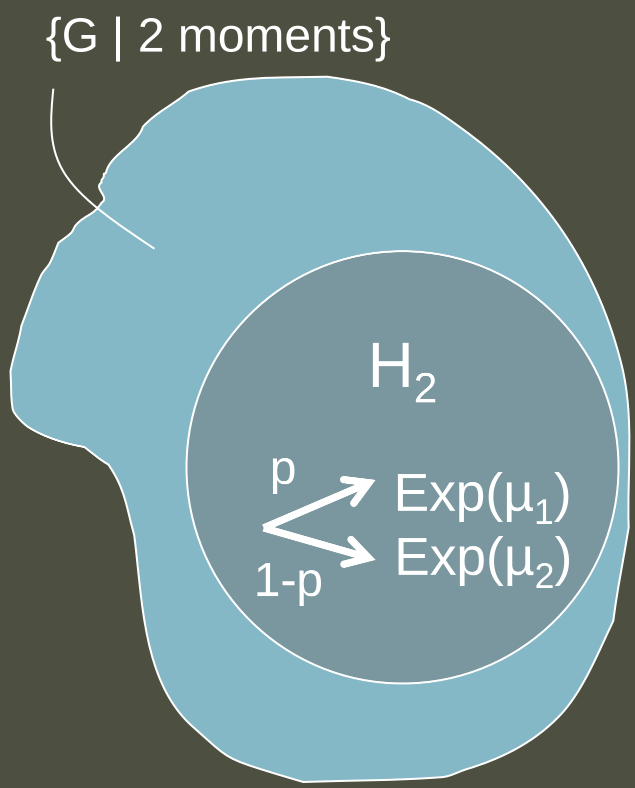
Tighter bounds via higher moments of job size distribution

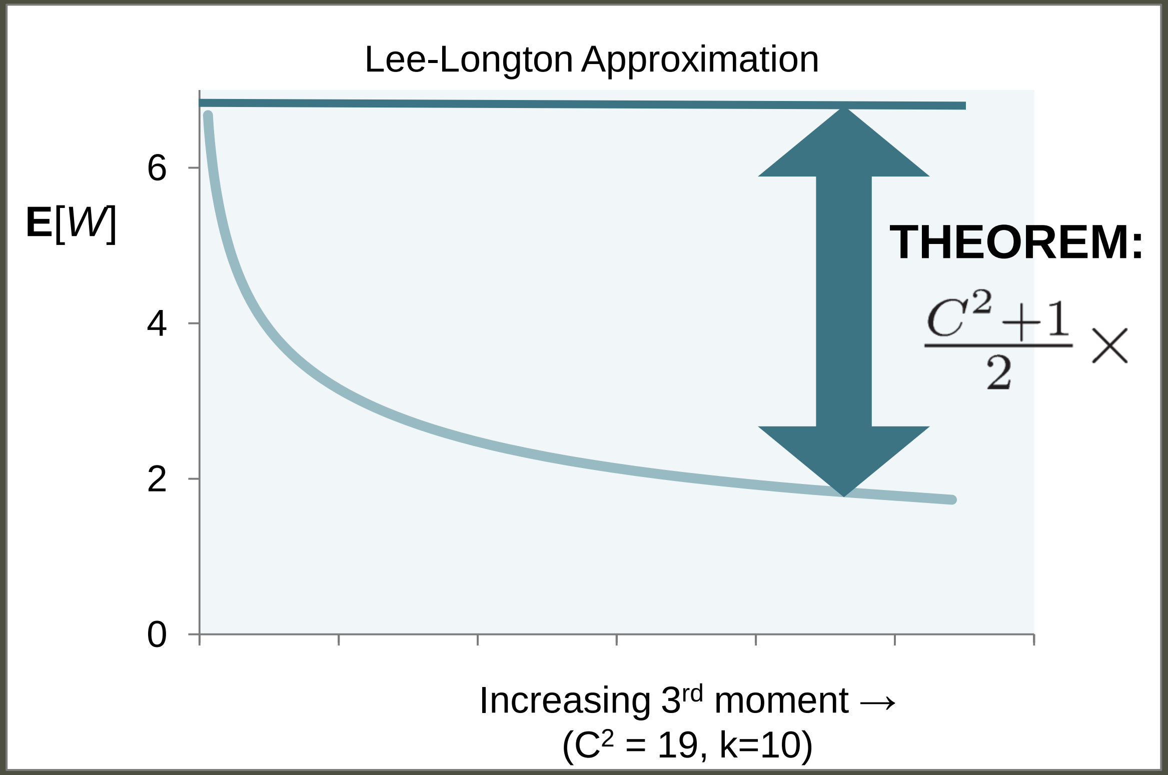
Lee, Longton approximation:

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

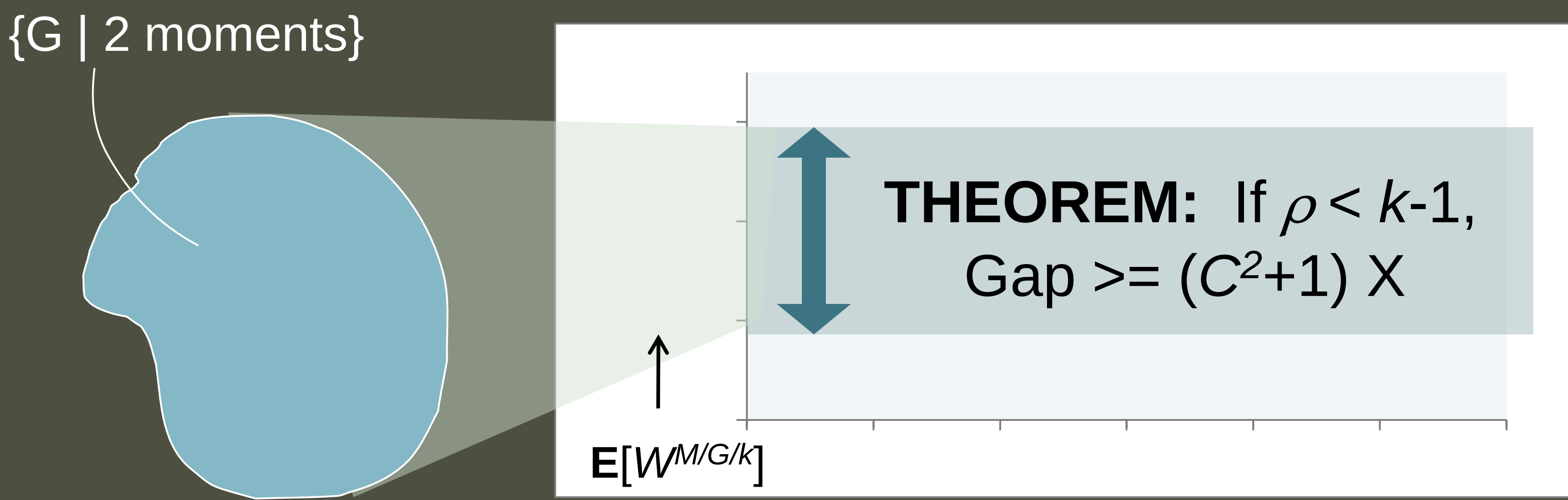


GOAL: Bounds on approximation ratio



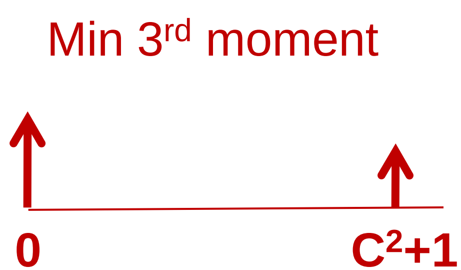


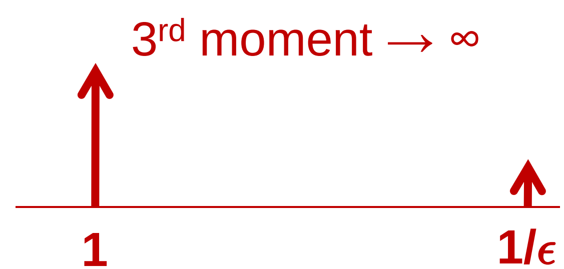
[Dai, G., Harchol-Balter, Zwart]



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 ≡ mixture of 2 points

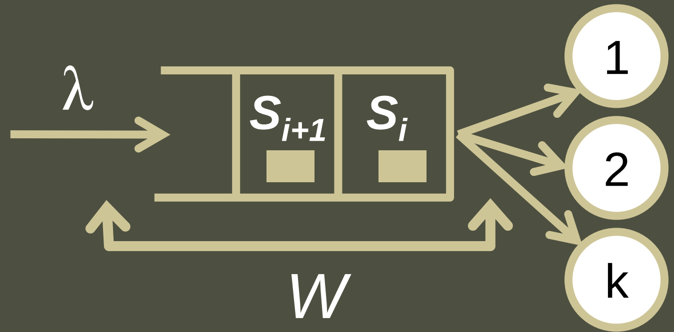




Approximations using higher moments?

[Dai, G., Harchol-Balter, Zwart]

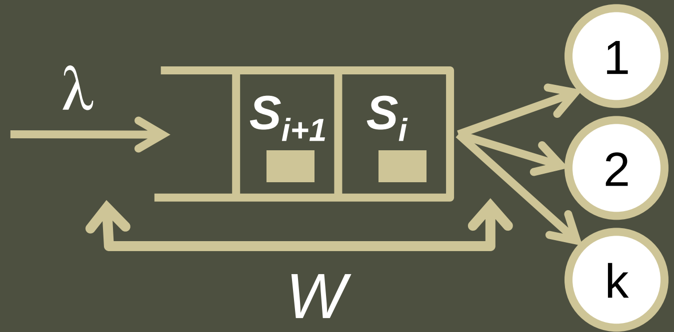
Outline

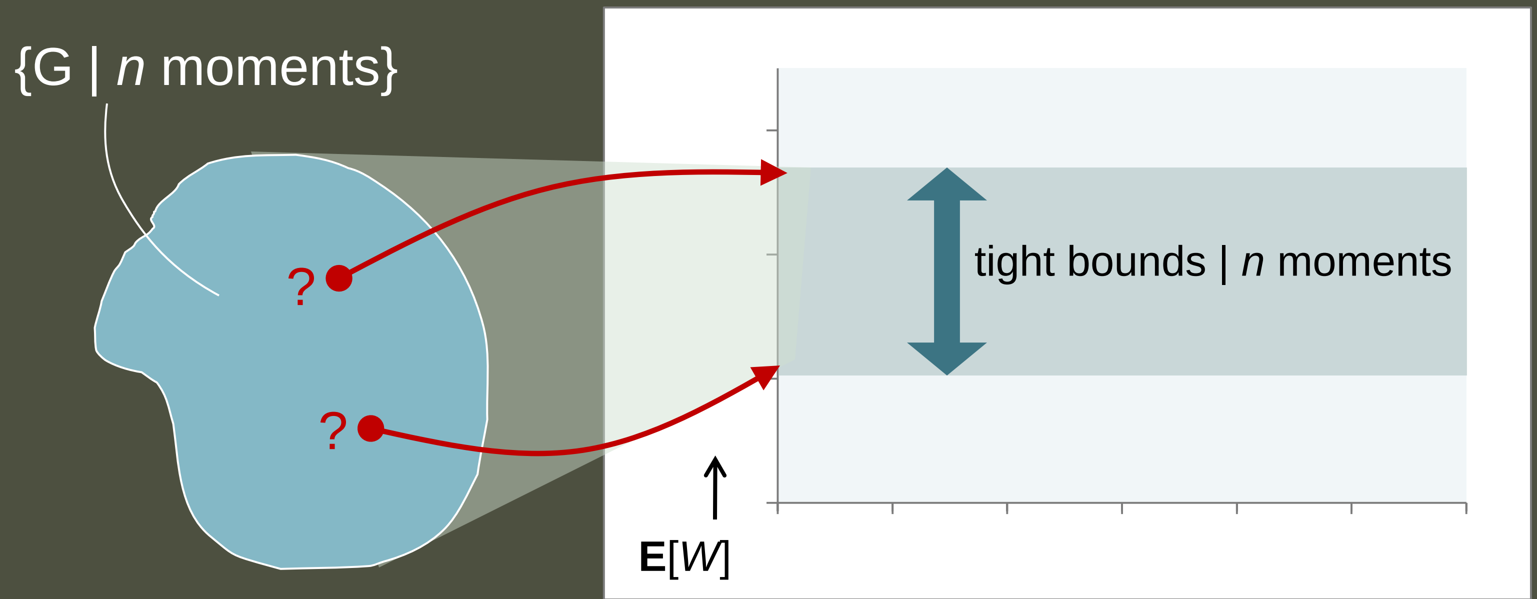


2 moments not enough for E[$W^{M/G/k}$]

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

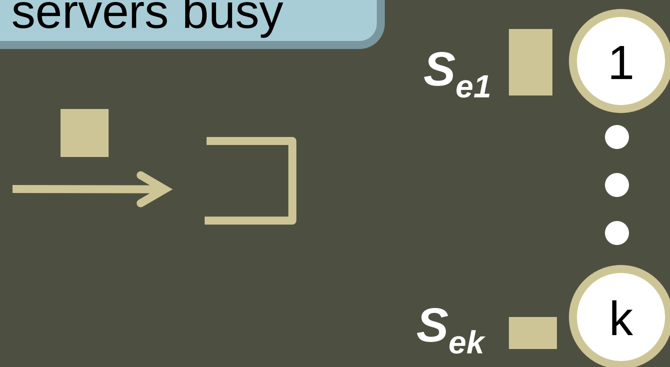
IDEA: Light-traffic asymptotics (λ→0)

RELAXATION: Identify the “extremal” distributions in light traffic

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

$$E[W^{M/G/k}] = \frac{\rho^k}{k!} E[\min\{S_{e_1}, S_{e_2}, \ldots, S_{e_k}\}] + o(\rho^k)$$

Probability of finding all servers busy

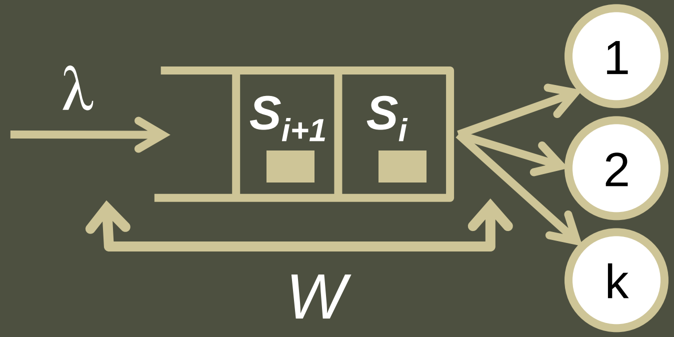


i.i.d. copies of S_e ≡ *equilibrium excess* of S

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

SUBGOAL: Extremal distributions for $E[\min\{S_{e1},\ldots,S_{ek}\}]$
s.t. $E[S^i] = m_i$ for i=1,..,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

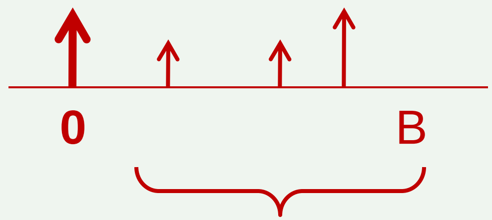
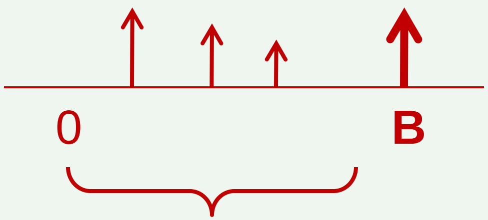
E[min{S_{e1},…,S_{ek}}] s.t. E[S^i] = m_i for i=1,..,n

Principal Representations and Extremal Problems

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

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

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]

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

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

Principal Representations and Extremal Problems

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

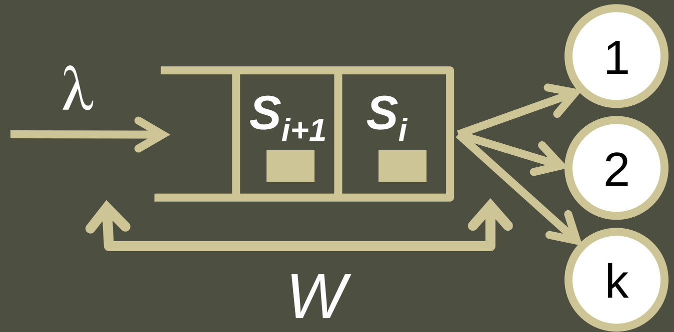
$E[X^0]=m_0$
$E[X^1]=m_1$
…
$E[X^n]=m_n$

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): Extremal distributions for

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

THEOREM:
For n = 2 or 3

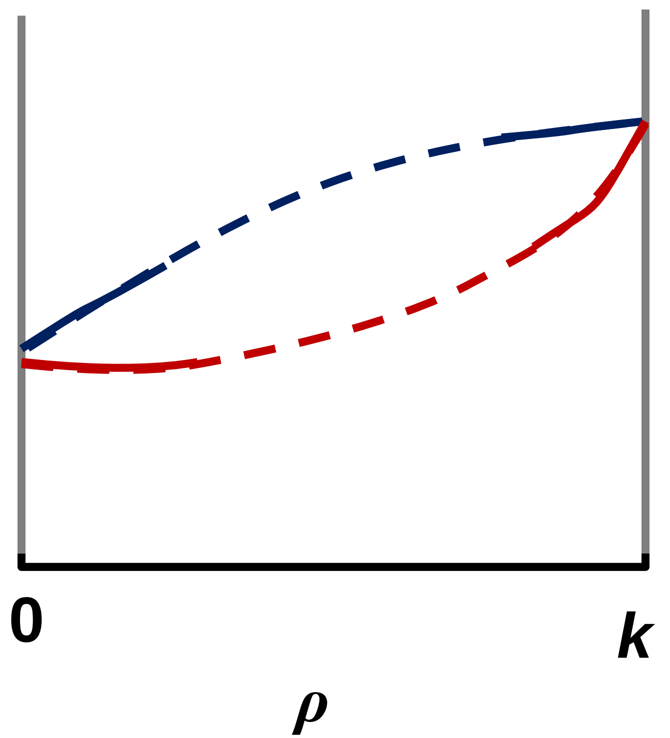
RELAXATION 2: Restrict to Completely Monotone distributions (mixtures of Exponentials)

(contains Weibull, Pareto, Gamma)

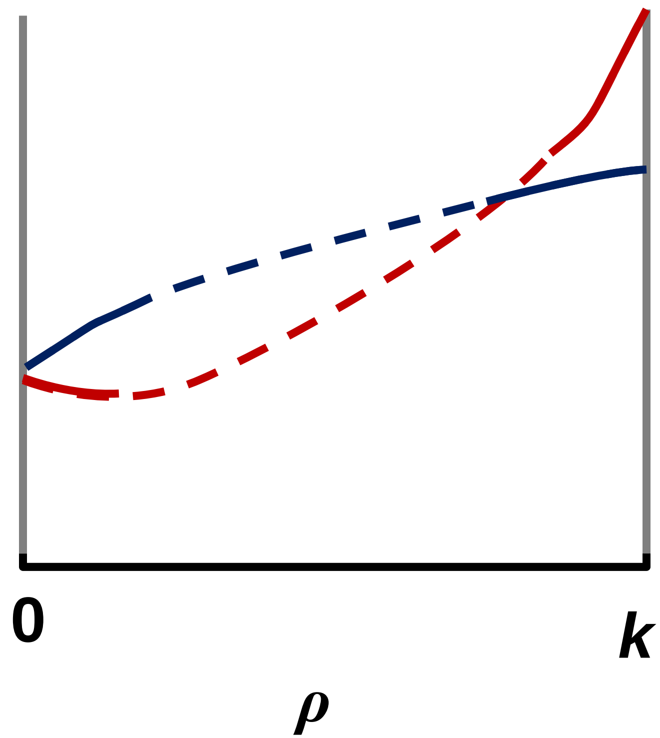
THEOREM:
For all n.

CONJECTURE: P.R.s are extremal for $E[W^{M/G/k}]$ **for all ρ**, for all n, if moment constraints are integral.

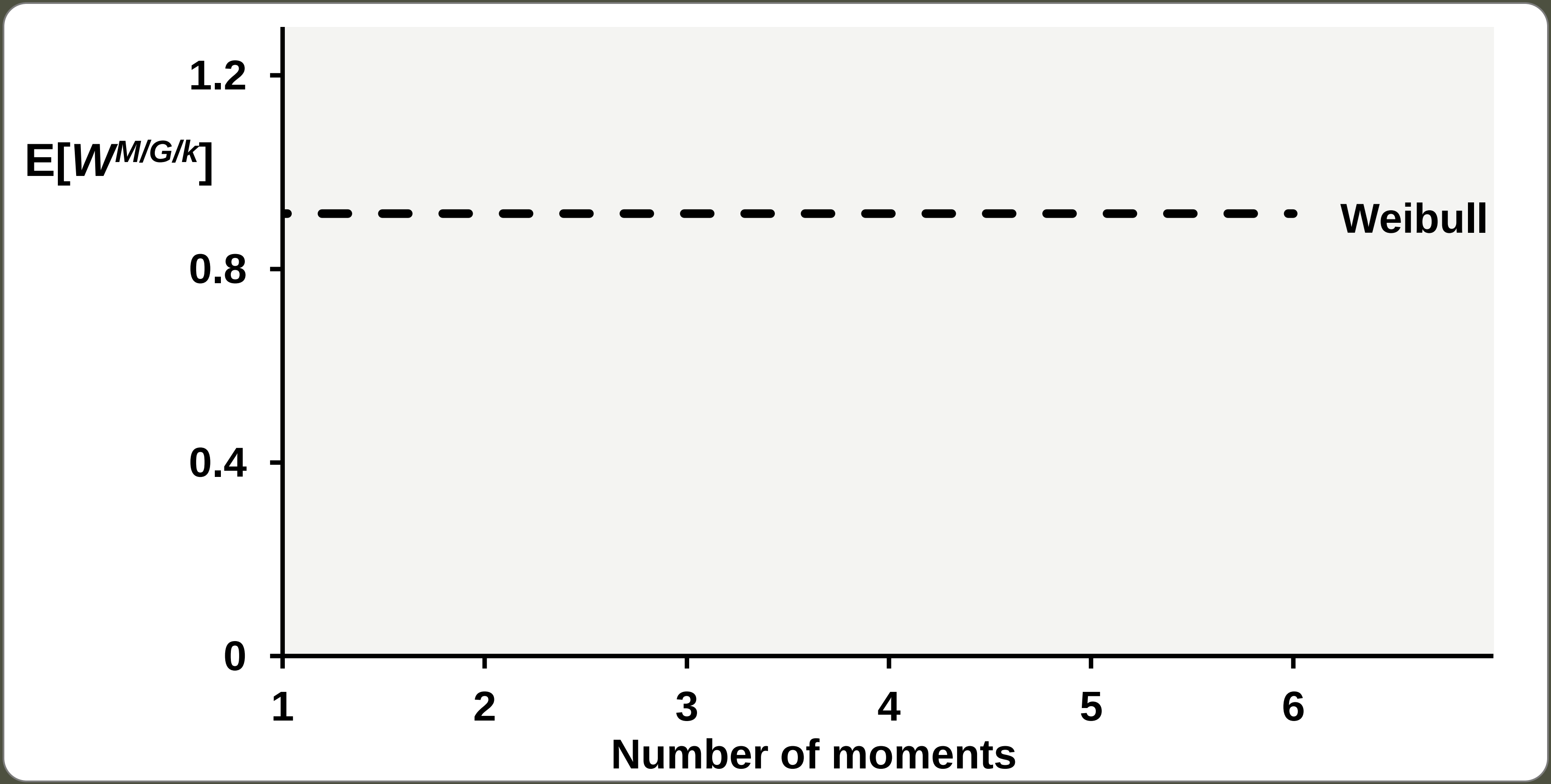
Given at least $E[S]$, $E[S^2]$



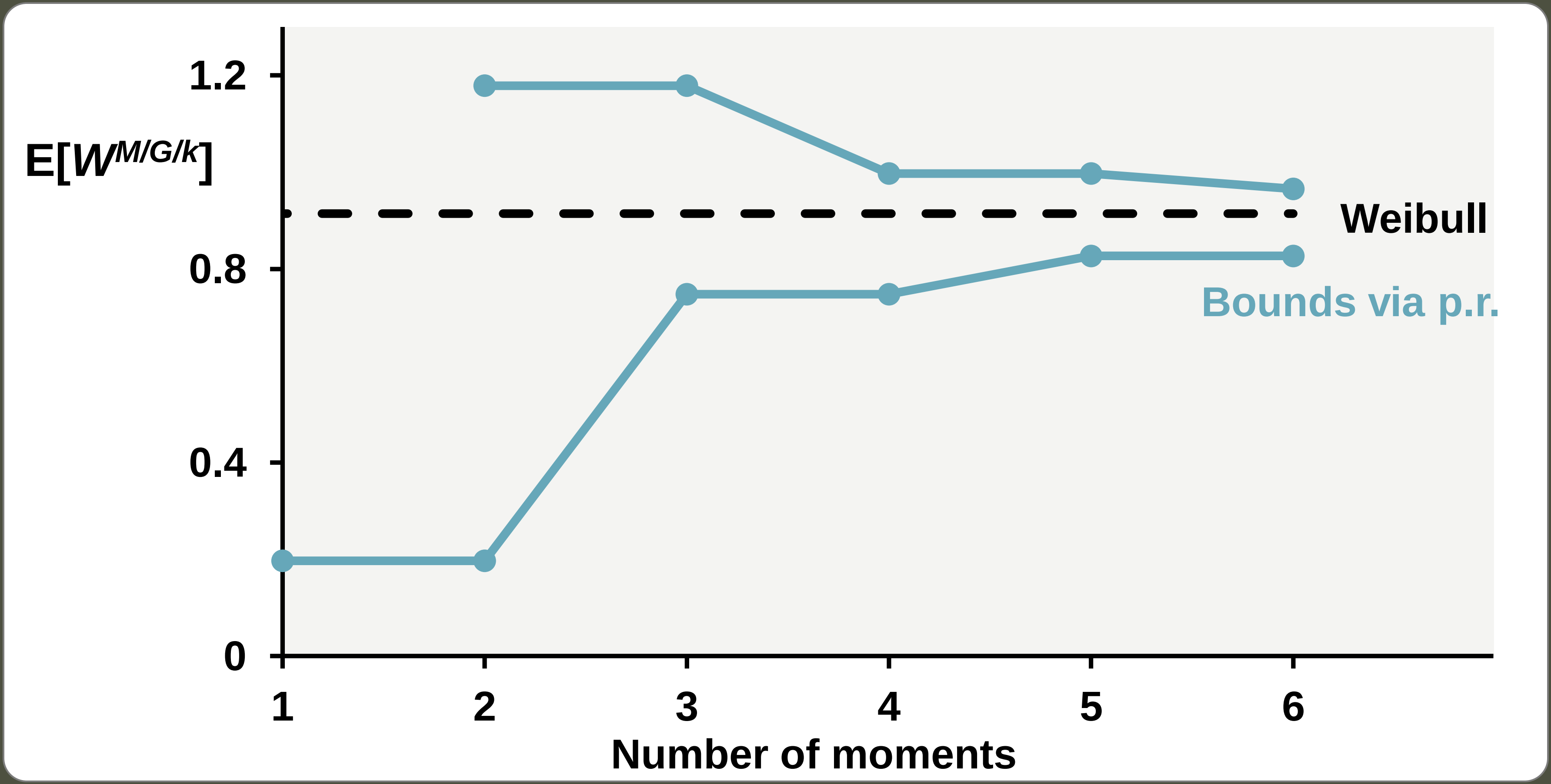
Not given $E[S^2]$, even # of moment constraints in (0,2)



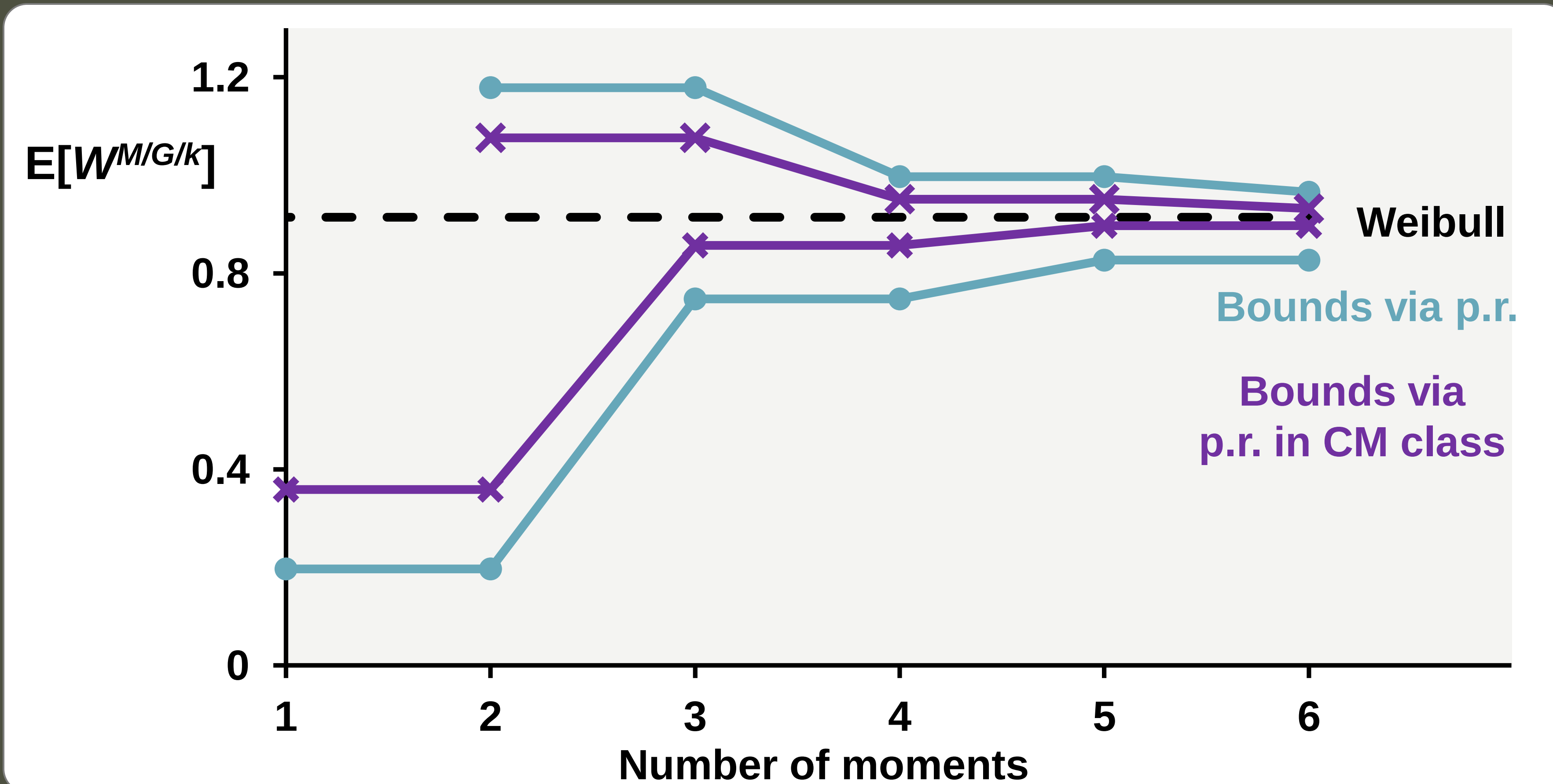
Simulation Results (k=4, ρ=2.4,)



Simulation Results (k=4, ρ=2.4,)



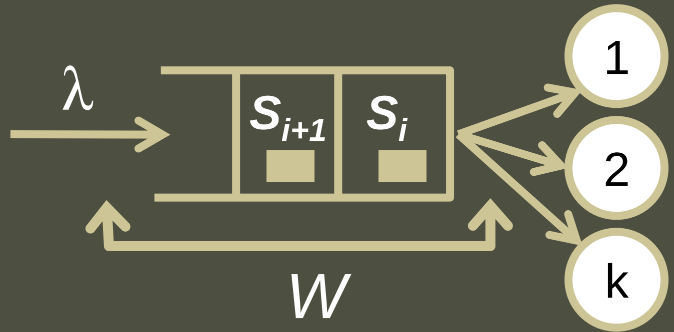
Simulation Results (k=4, ρ=2.4,)



Approximation Schema:

Refine lower bound via an additional odd moment,
Upper bound via even moment until gap is acceptable

Outline



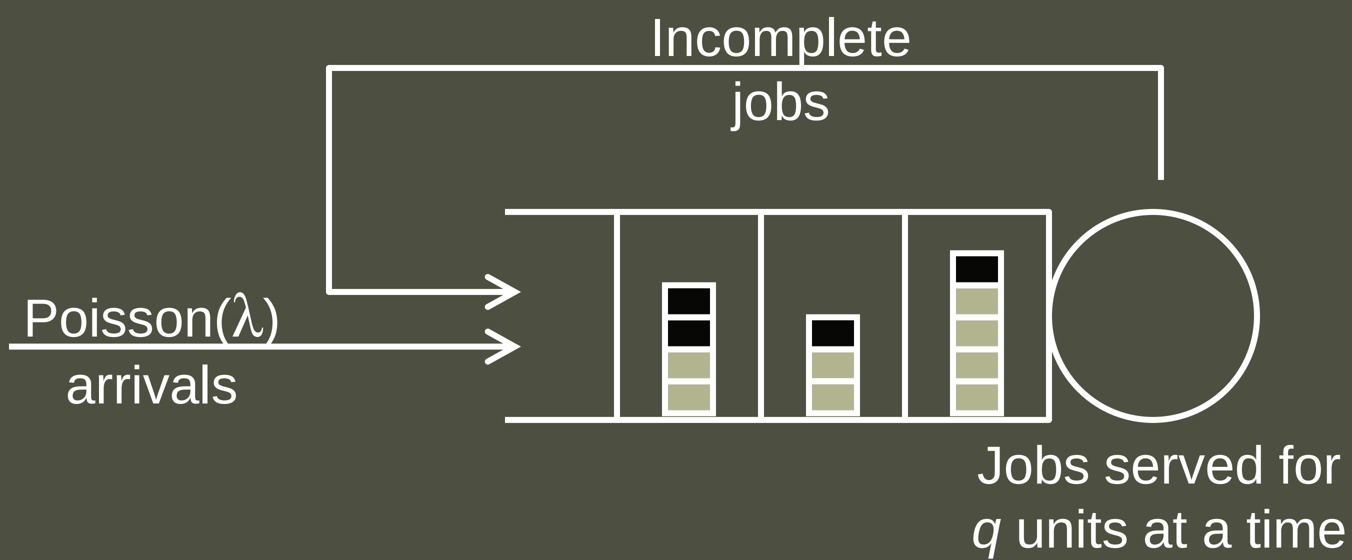
2 moments not enough for E[$W^{M/G/k}$]

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

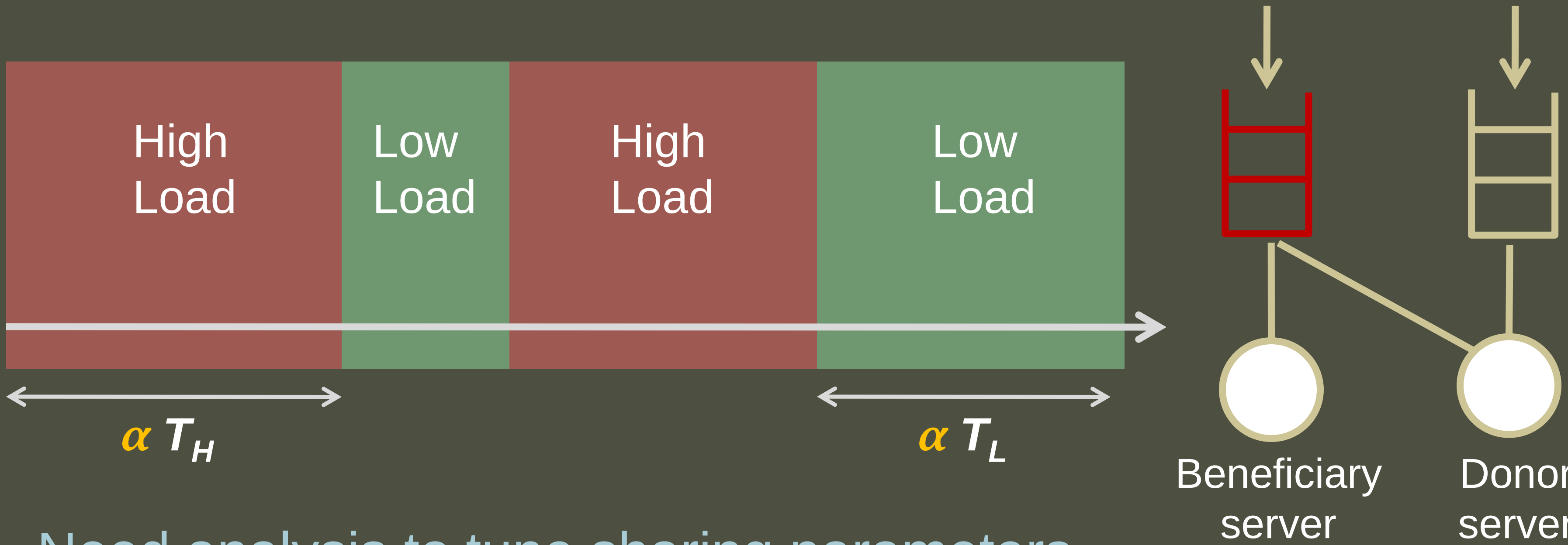


Need analysis to find *q* that balance overheads/performance

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



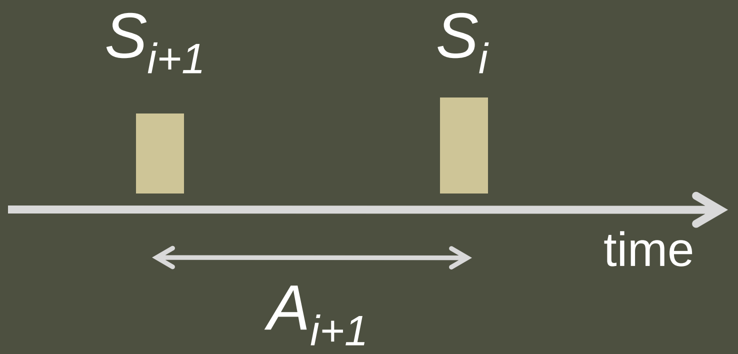
Need analysis to tune sharing parameters

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

Open problem: Markov-Krein characterization of Stochastic Recursive Sequences

Example: Single server system

W_{i+1} = waiting time of S_{i+1}

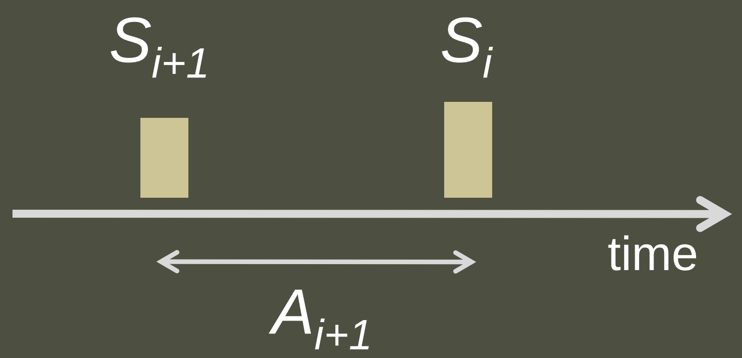


$$W_{i+1} = \Phi(W_i, S_i, A_{i+1})$$

Open problem: Markov-Krein characterization of Stochastic Recursive Sequences

Example: Single server system

W_{i+1} = waiting time of S_{i+1}

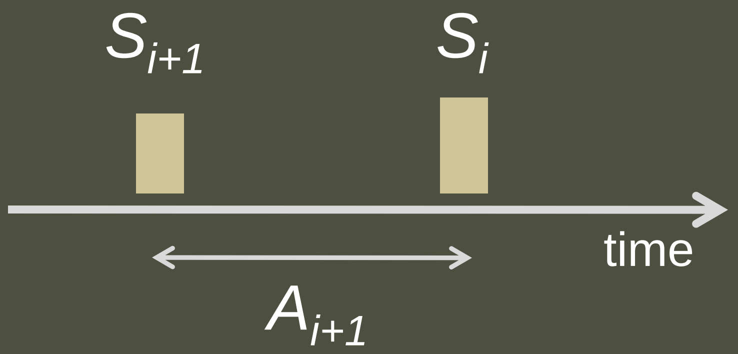


$$W_{i+1} = (W_i + S_i - A_{i+1})^+$$

Open problem: Markov-Krein characterization of Stochastic Recursive Sequences

Example: Single server system

W_{i+1} = waiting time of S_{i+1}



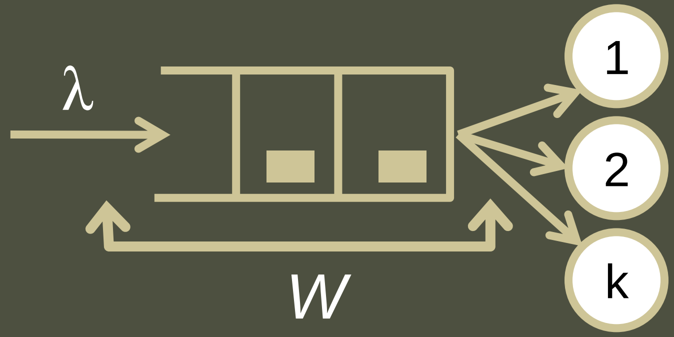
$$W \overset{d}{=} (W + S - A)^+$$

Stationary behavior of a queueing system = Fixed point of a stochastic recursive sequence of the form

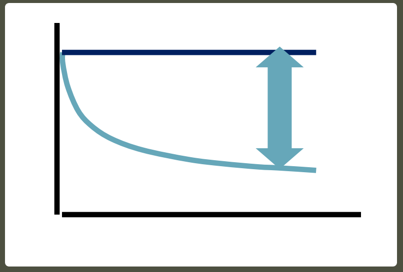
$$W \overset{d}{=} \Phi(W,S)$$

Q: Given moments of S, under what conditions on f, Φ, is E[f(W)] extremized by p.r.s?

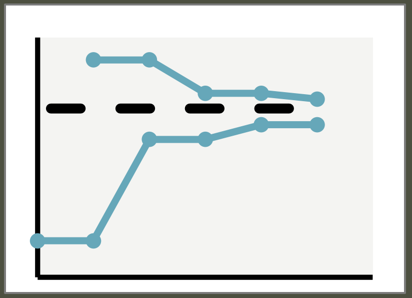
Conclusions



- 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