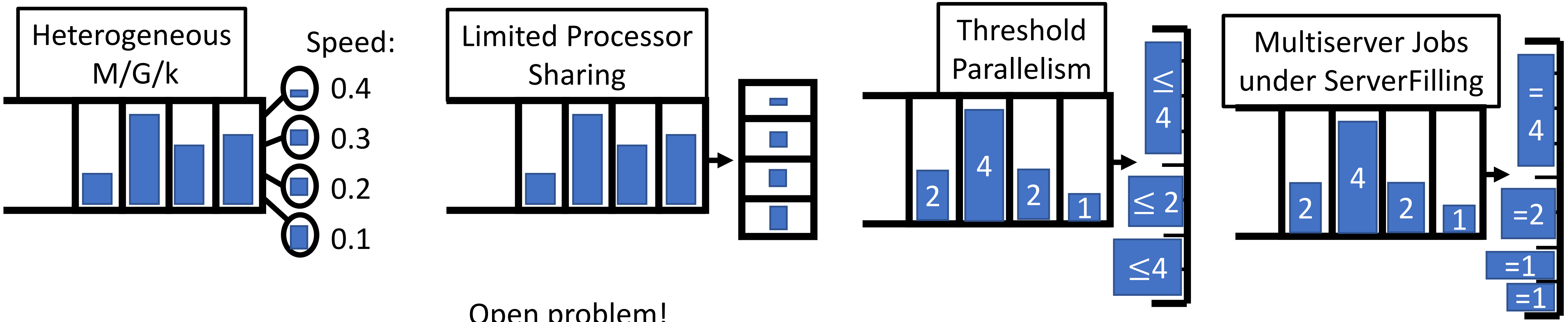
WCFS Queues: A new analysis framework

Isaac Grosof

Mor Harchol-Balter

Alan Scheller-Wolf

Unexpected Similarity between Queues



Open problem!

Open problem!

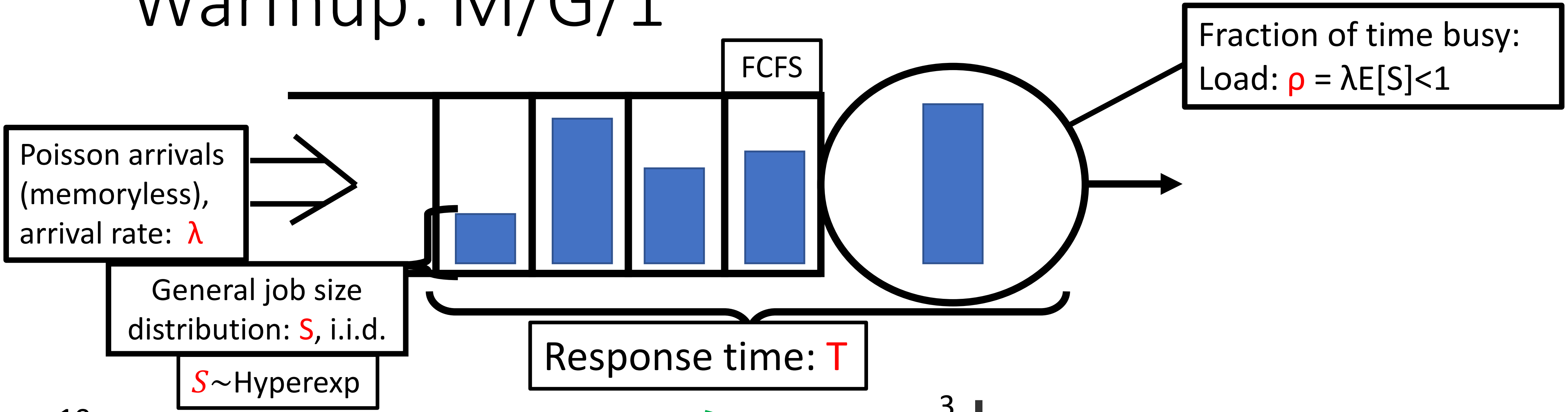
Open problem!

Open problem!

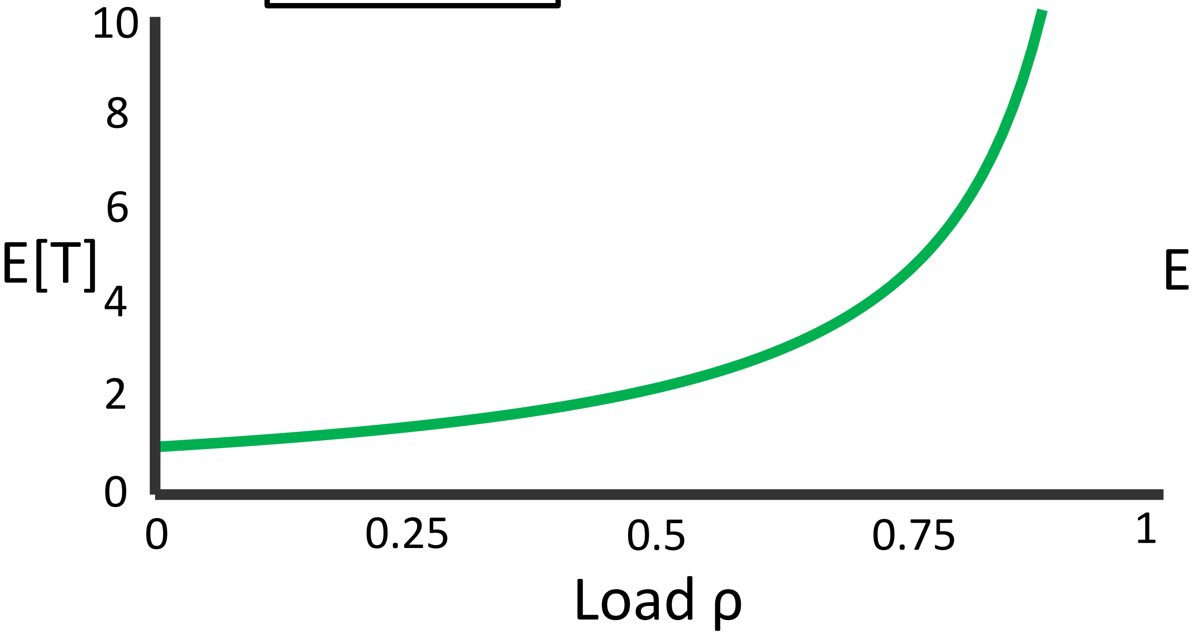
Discover commonality

Use commonality to tightly characterize

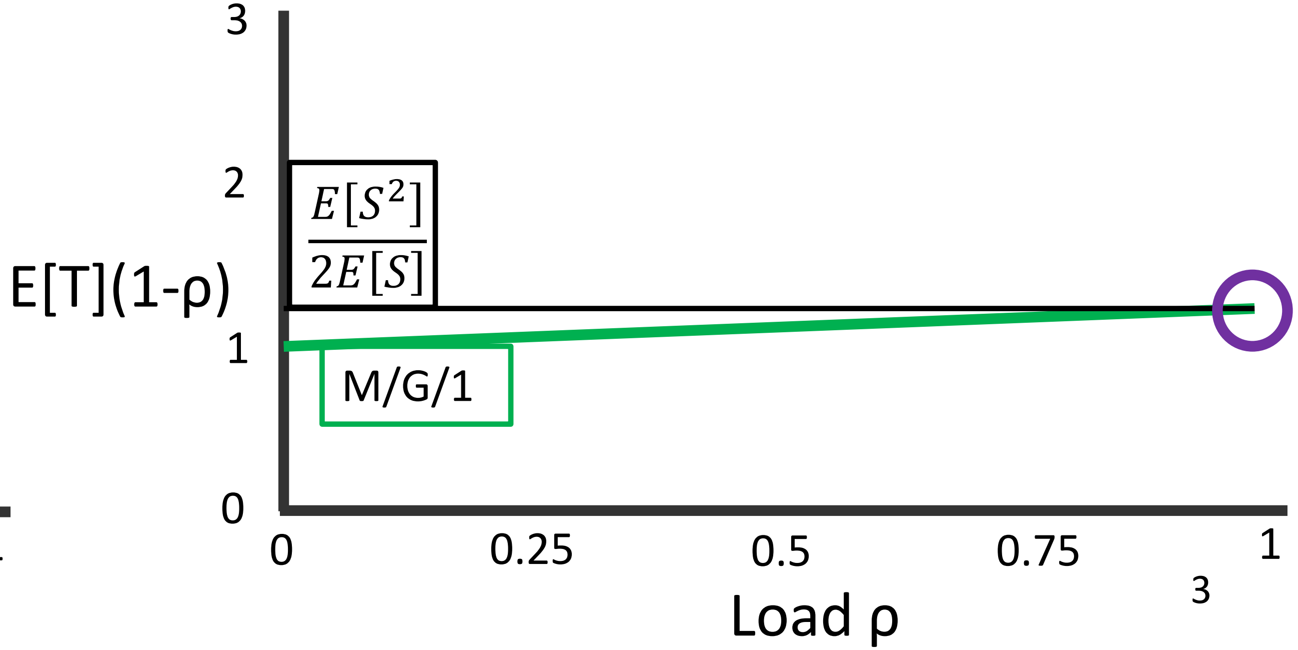
Warmup: M/G/1



Poisson arrivals
(memoryless),
arrival rate: λ
General job size
distribution: S, i.i.d.
S~Hyperexp
FCFS
Fraction of time busy:
Load: ρ = λE[S]<1
Response time: T

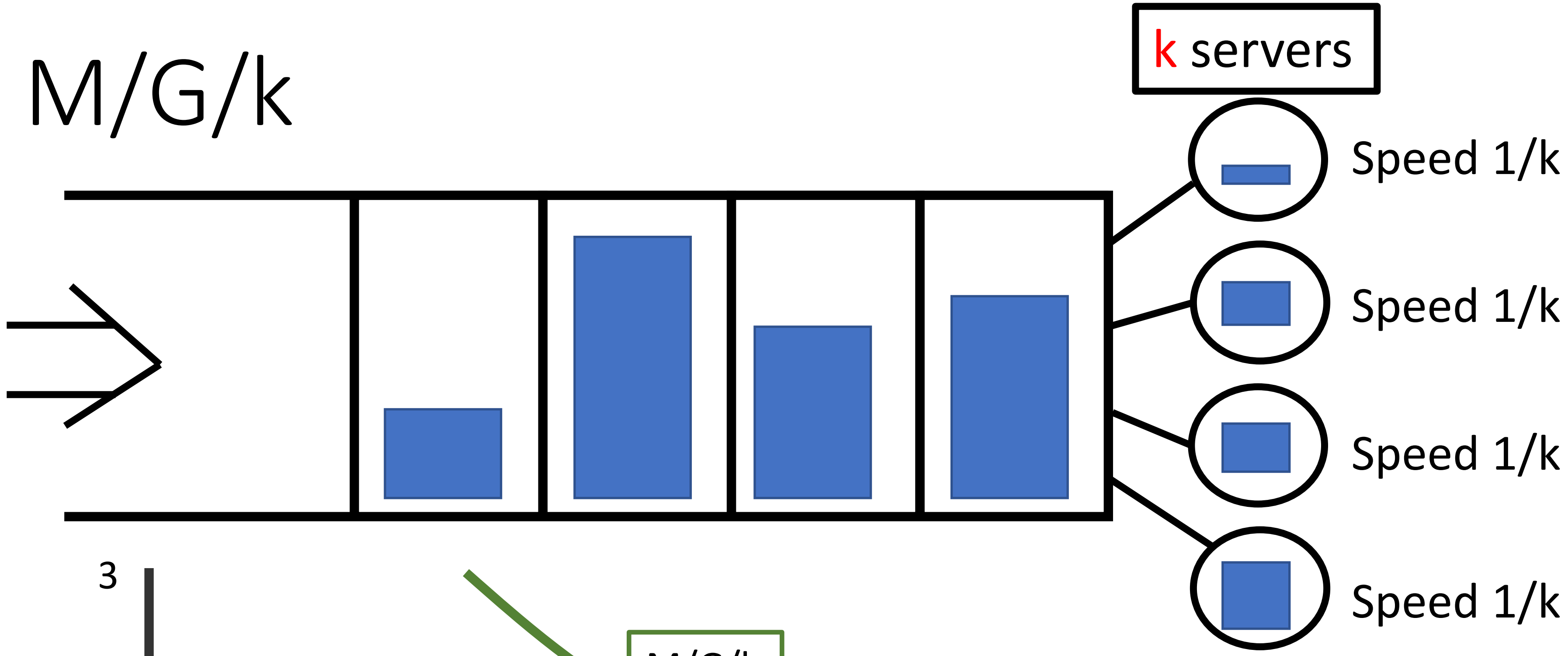


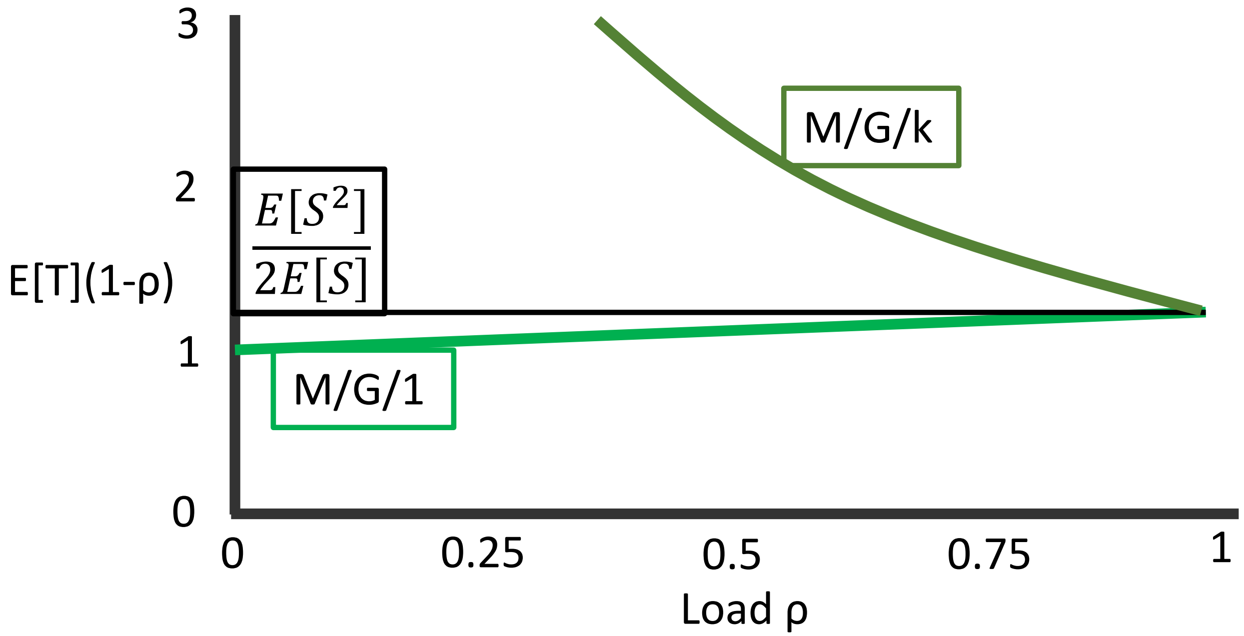
E[T]
10
8
6
4
2
0
0
0.25
0.5
0.75
1
Load ρ



E[T](1-ρ)
3
2
1
0
$\frac{E[S^2]}{2E[S]}$
M/G/1
0
0.25
0.5
0.75
1
Load ρ

M/G/k

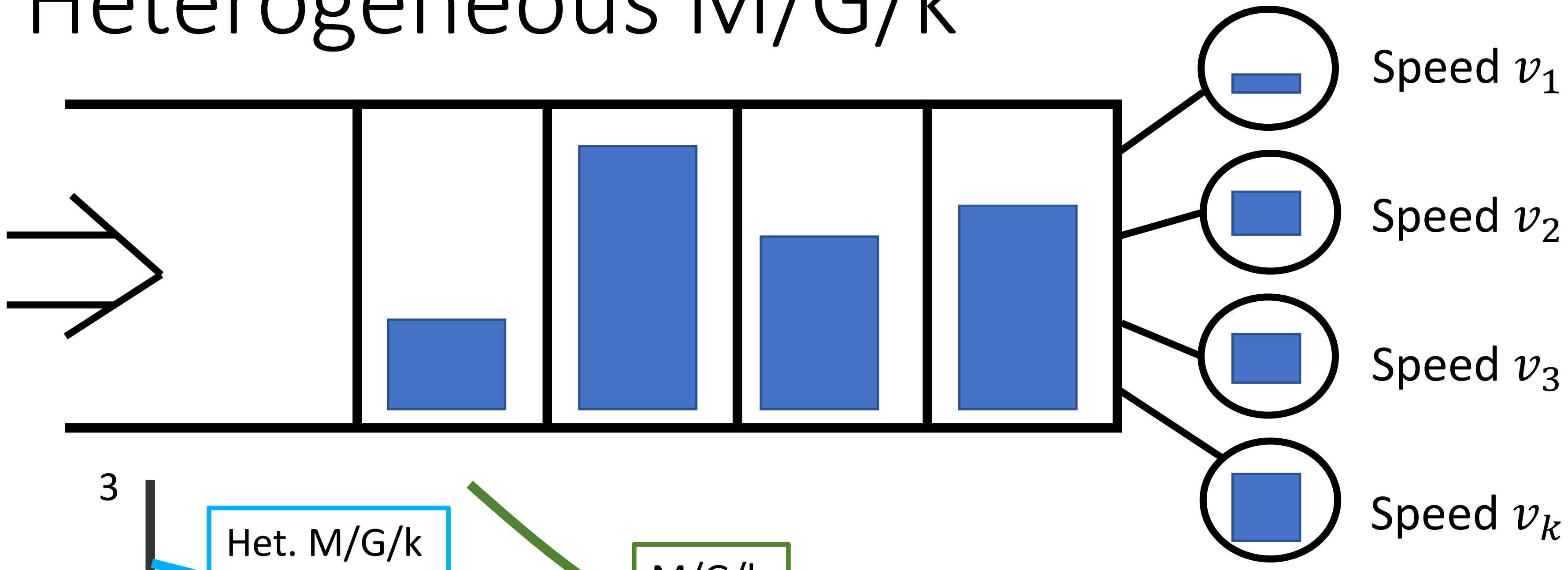


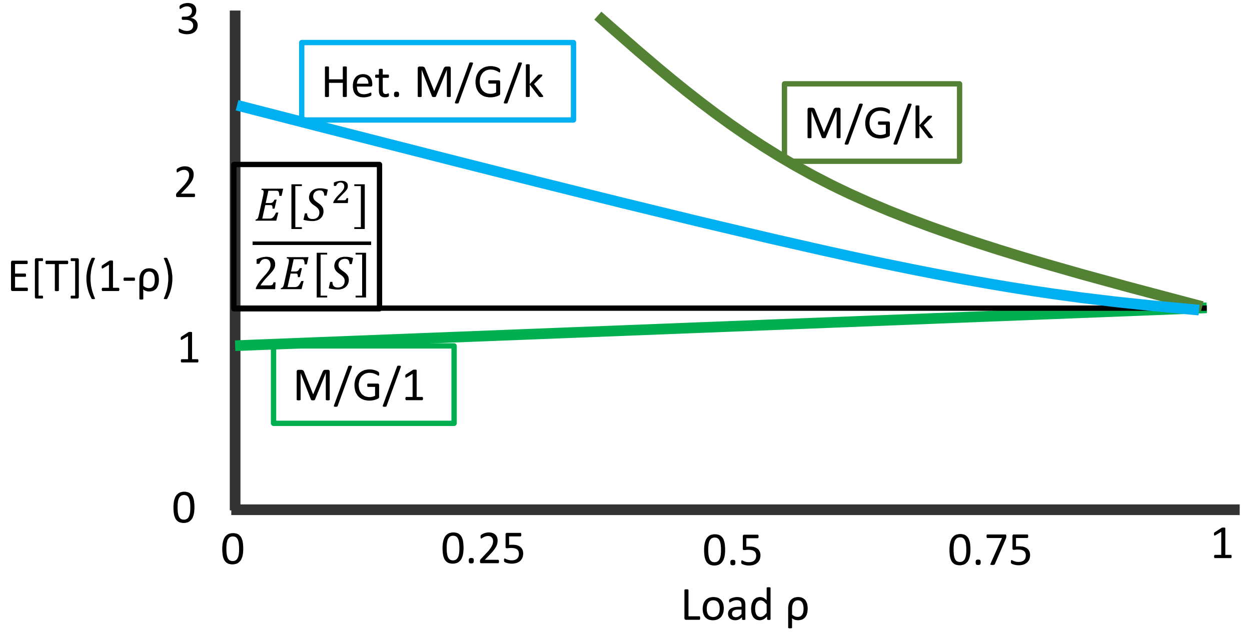


$\rho \to 1$ limit: $\frac{E[S^2]}{2E[S]}$

[Kollerstrom '84]

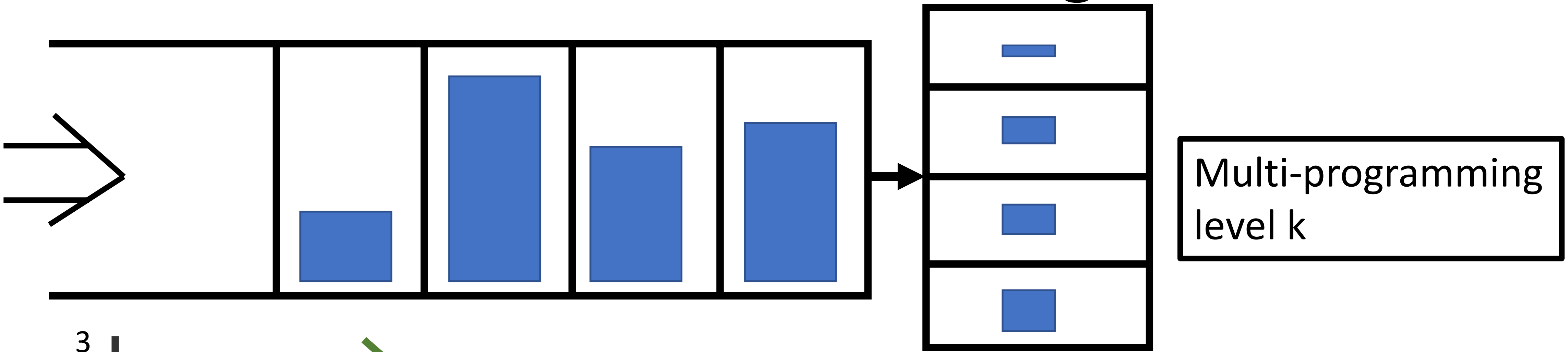
Heterogeneous M/G/k

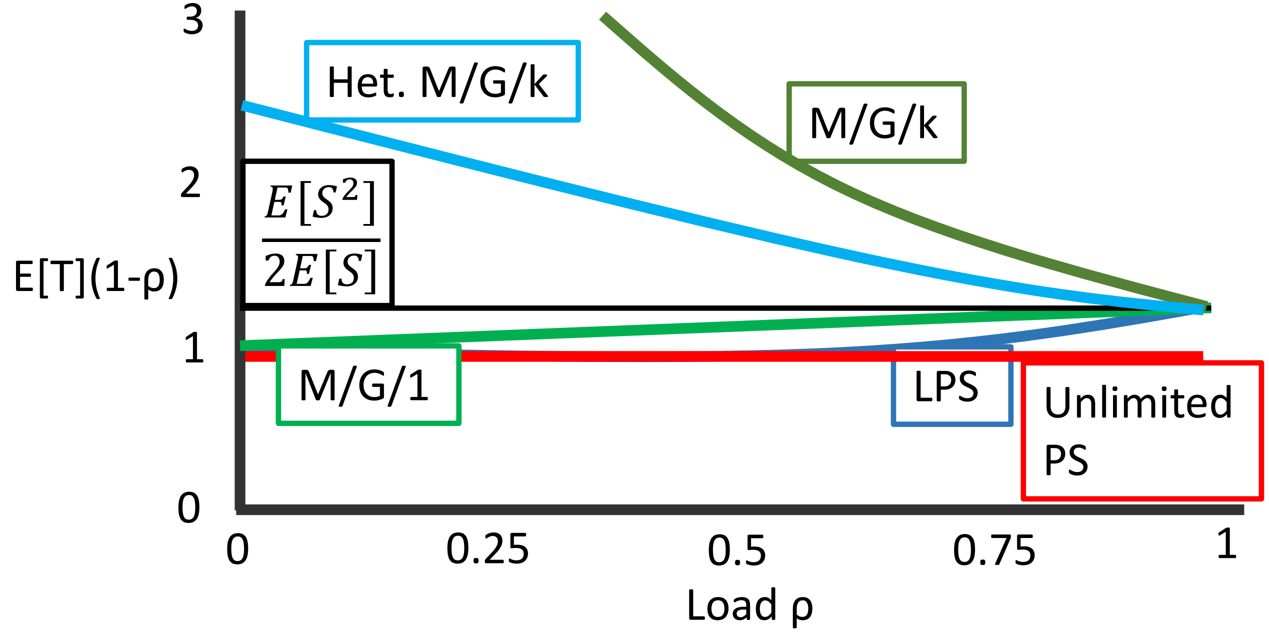




Mean response time behavior:
Open problem!

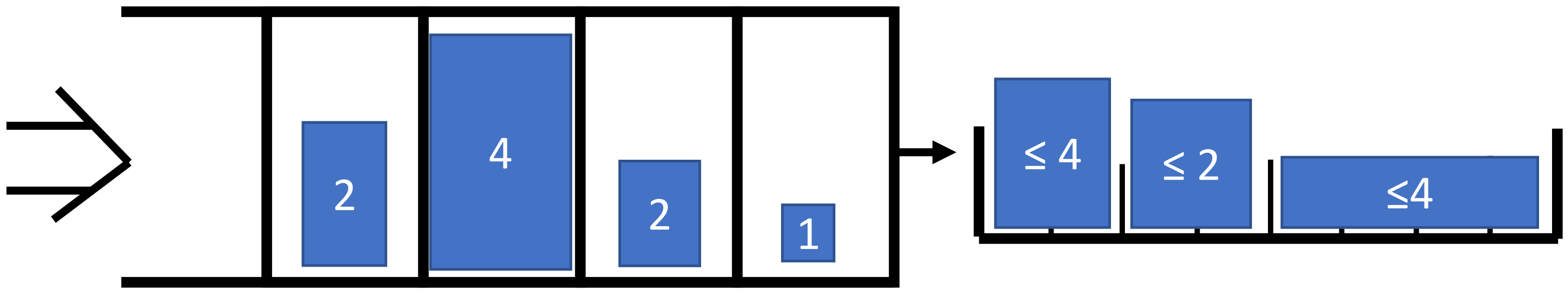
M/G/1/Limited Processor Sharing



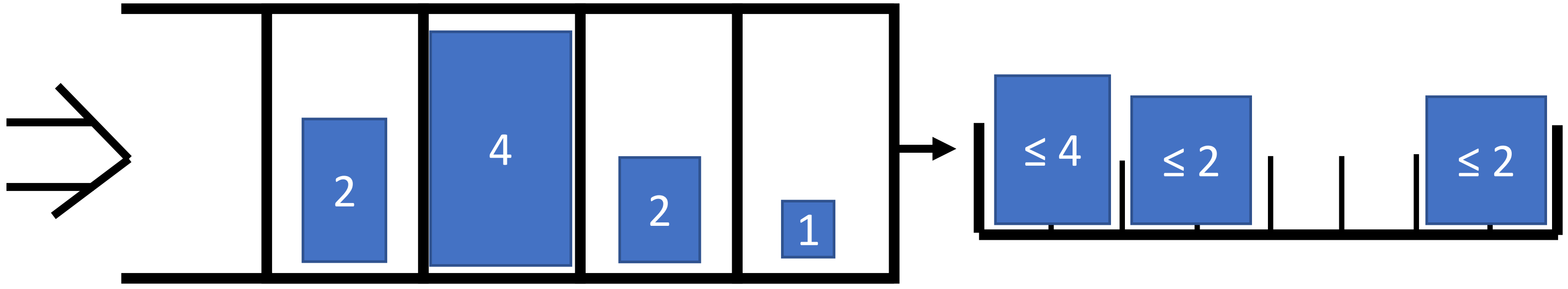


Open problem!
(Unlimited) Processor Sharing:
Known, different limit: E[S]

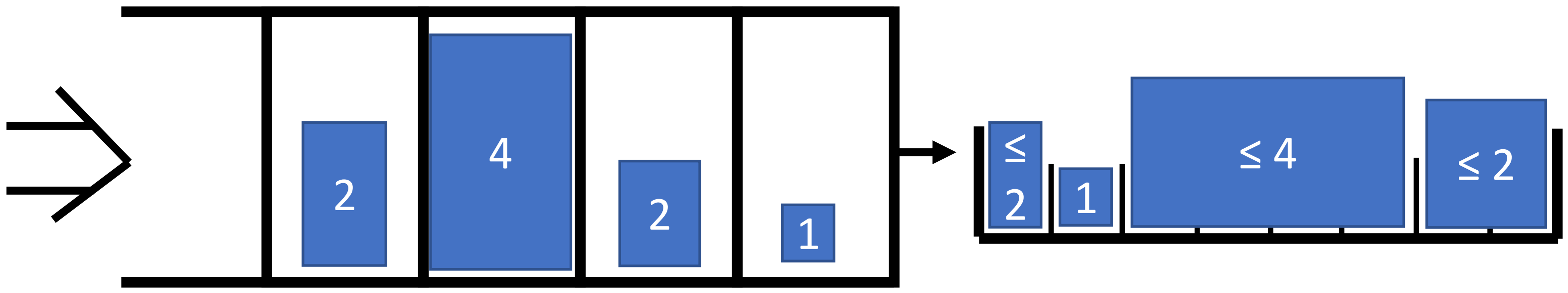
Threshold Parallelism



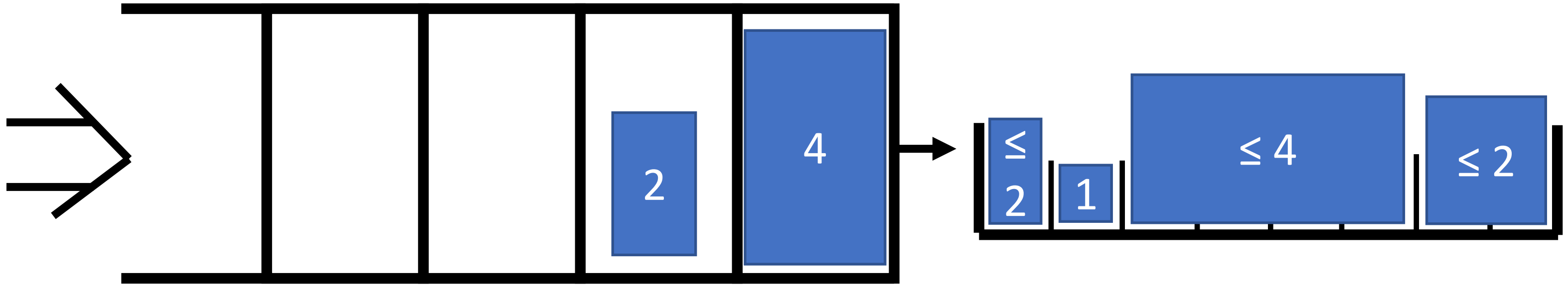
Threshold Parallelism

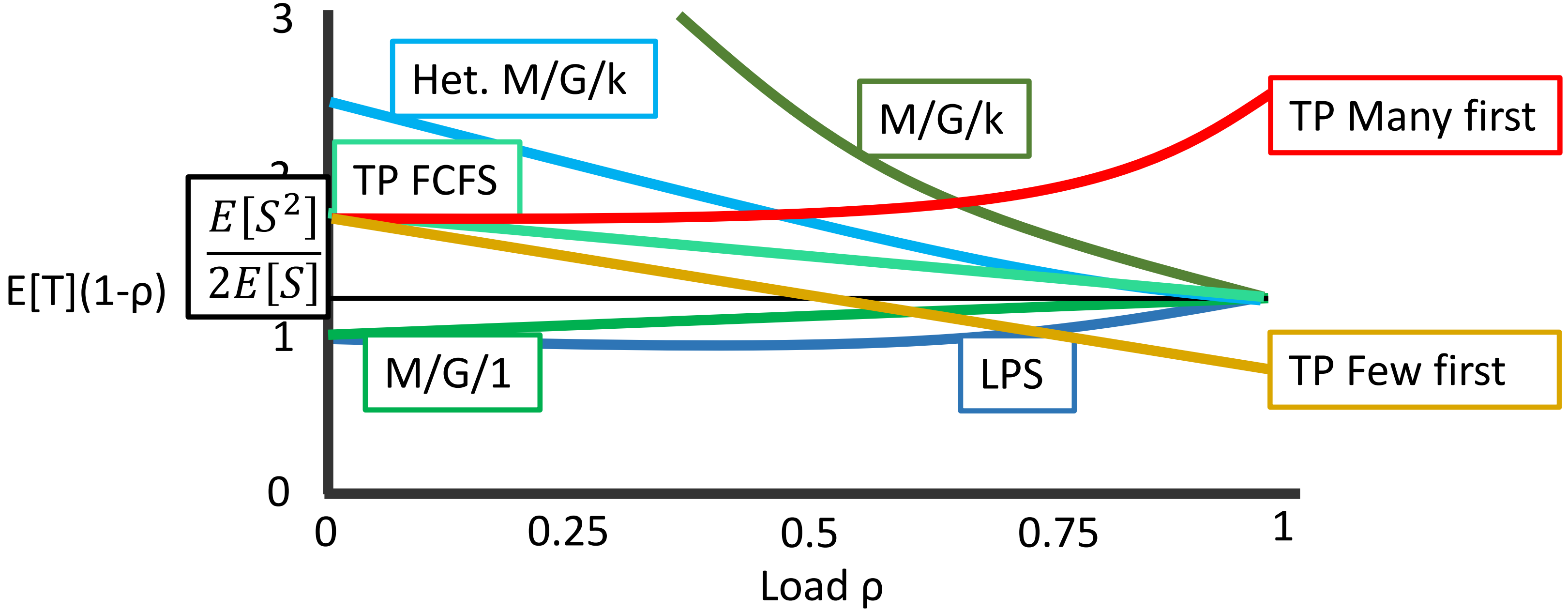


Threshold Parallelism



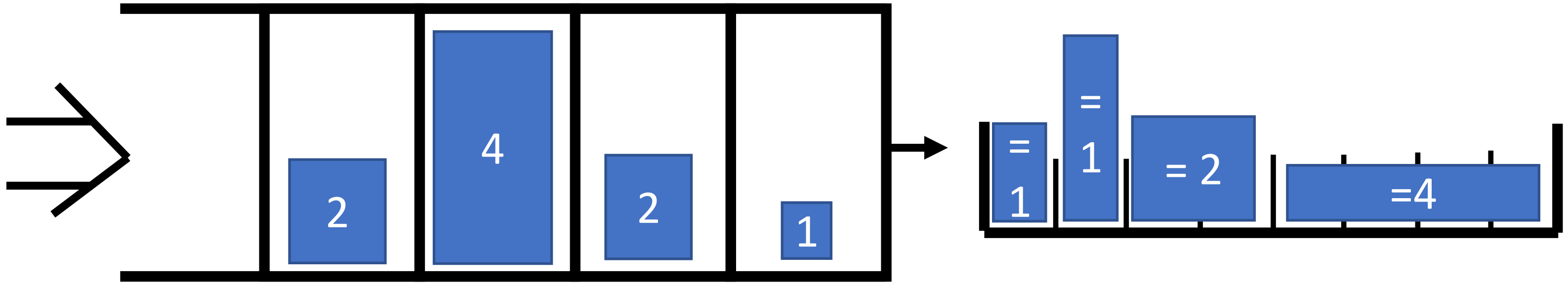
Threshold Parallelism

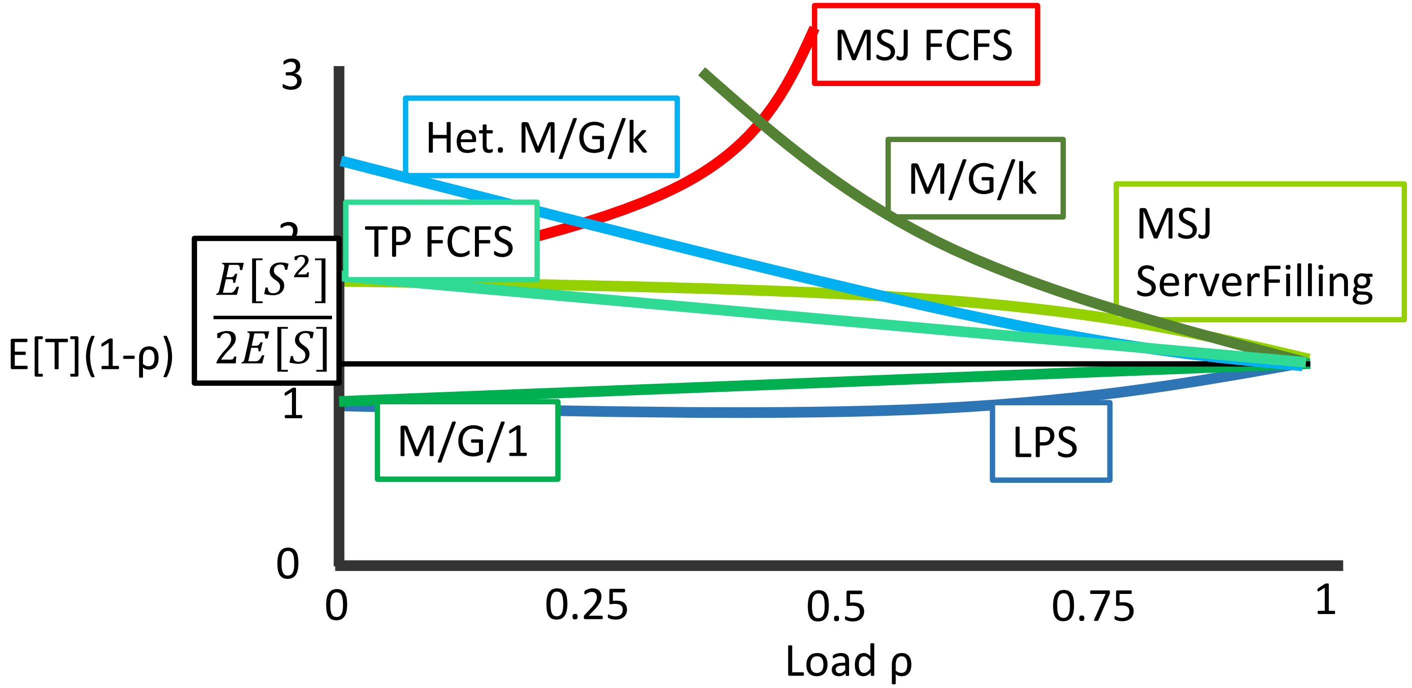




Open problem!
Few first or many first:
Different limits

Multiserver Job Model





All existing policies: open problem!
All different limits, many unstable
We create new policy: ServerFilling

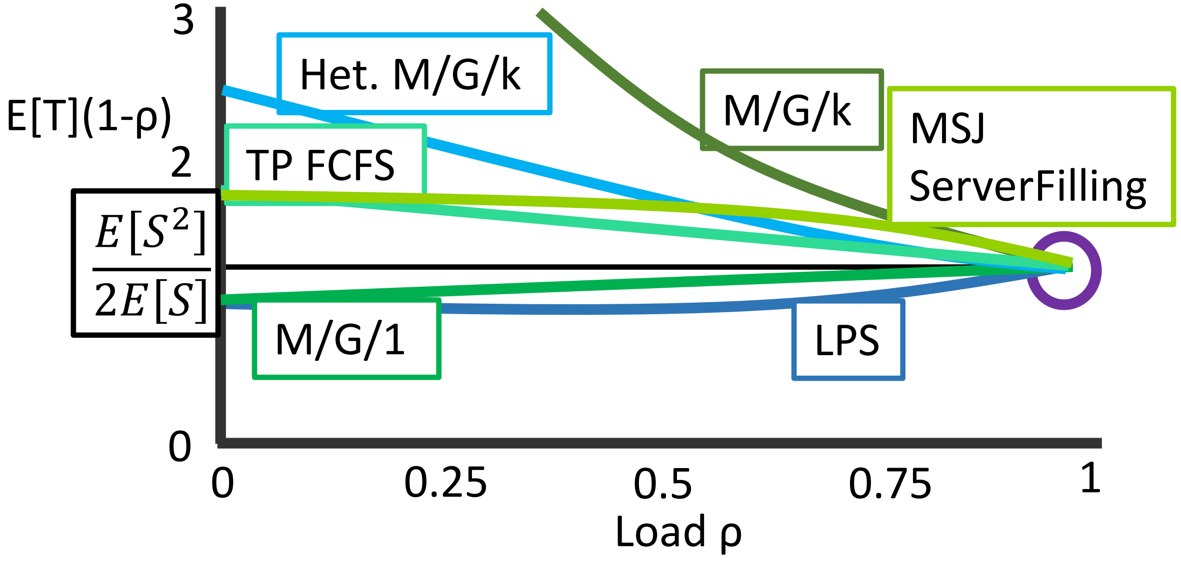
Result: Response time bound

Theorem: All models $\pi \in$ WCFS*:

$$\lim_{\rho \to 1} E[T^\pi](1-\rho) = \frac{E[S^2]}{2E[S]}$$

Even stronger theorem:

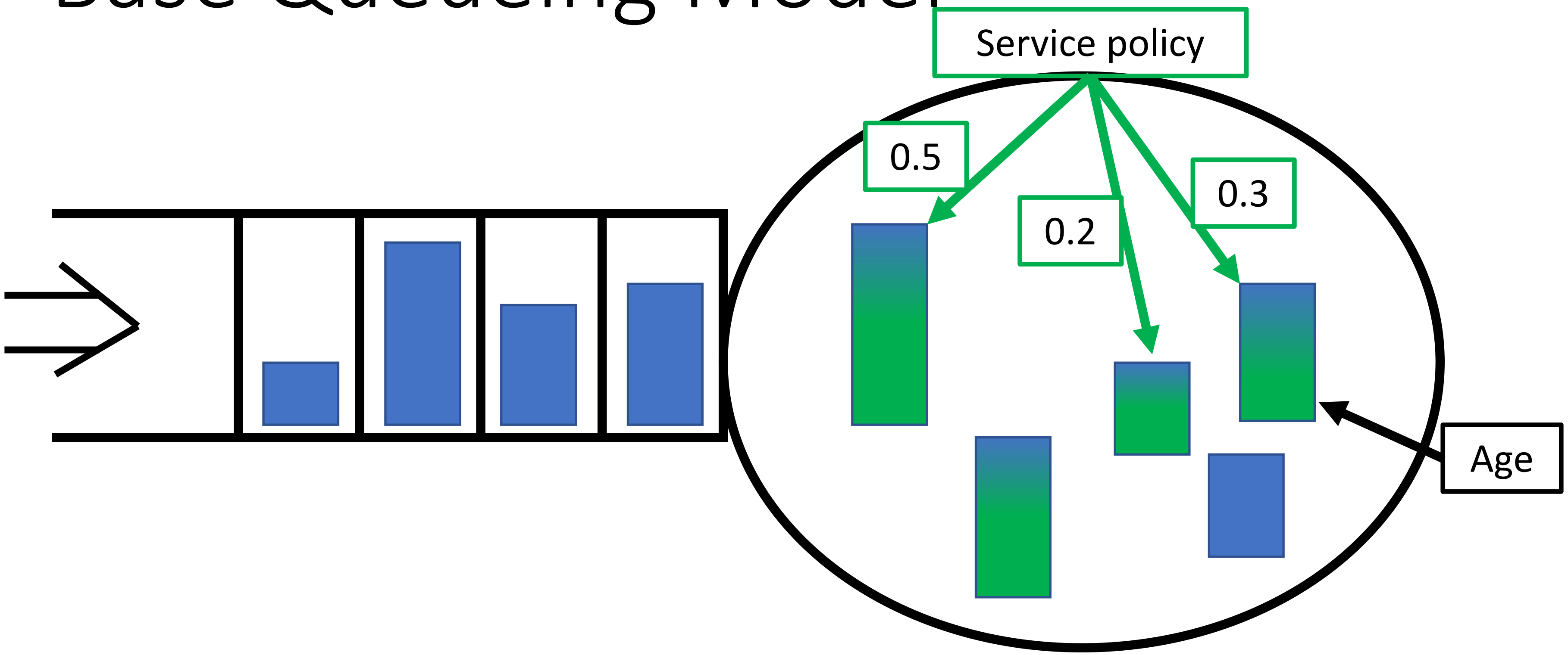
$$E[T^\pi] - E\left[T^{M/G/1}\right] \in [c_l^\pi, c_h^\pi]$$



Goals: Define WCFS, prove result

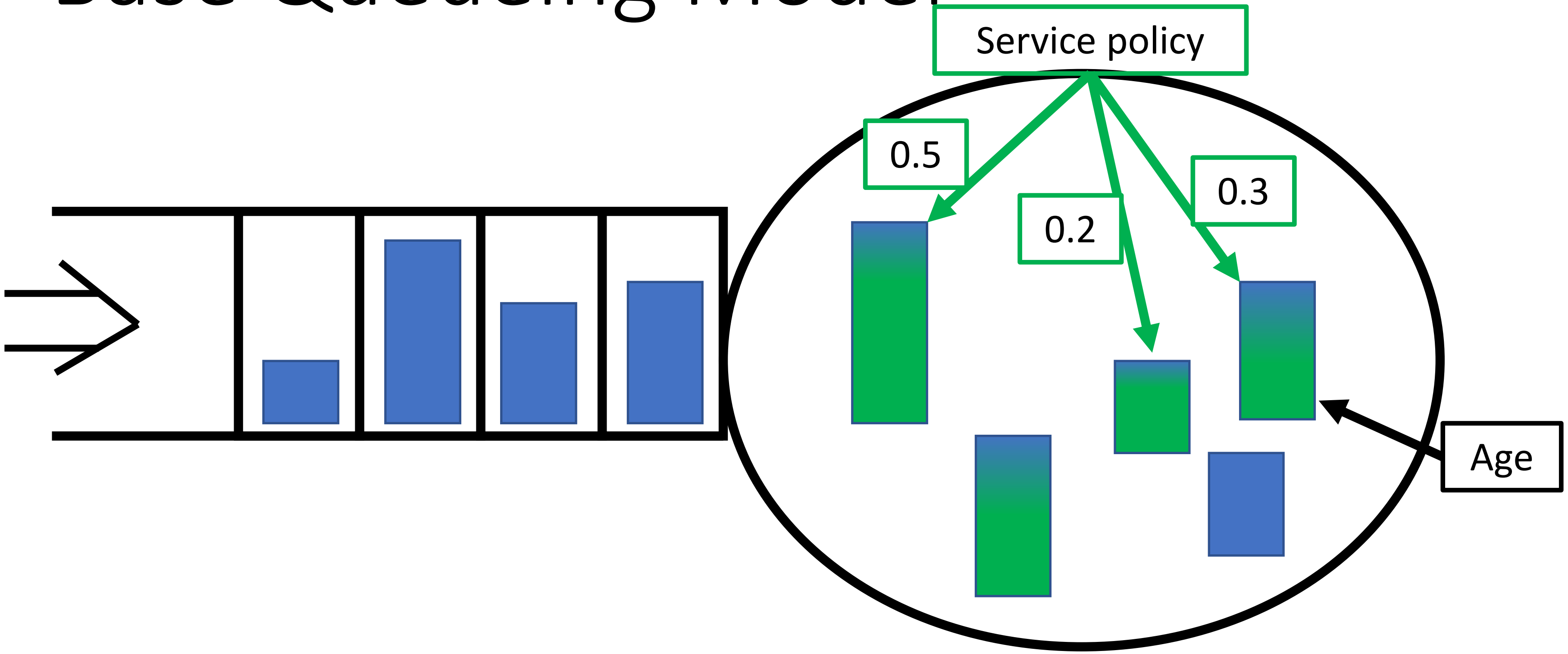
(Subject to minor condition on
job size distribution S)

Base Queueing Model



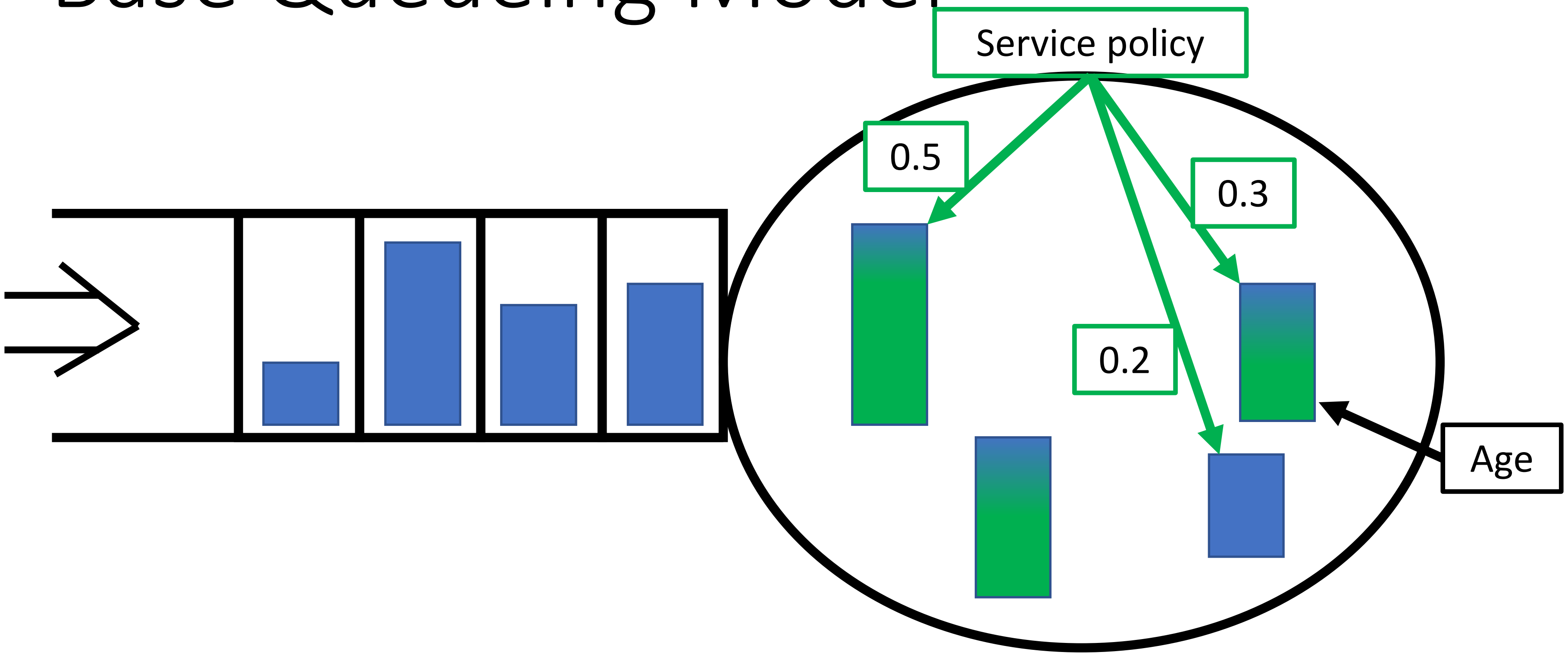
Service policy serves jobs at some service rate
Work completes at service rate, age increases

Base Queueing Model



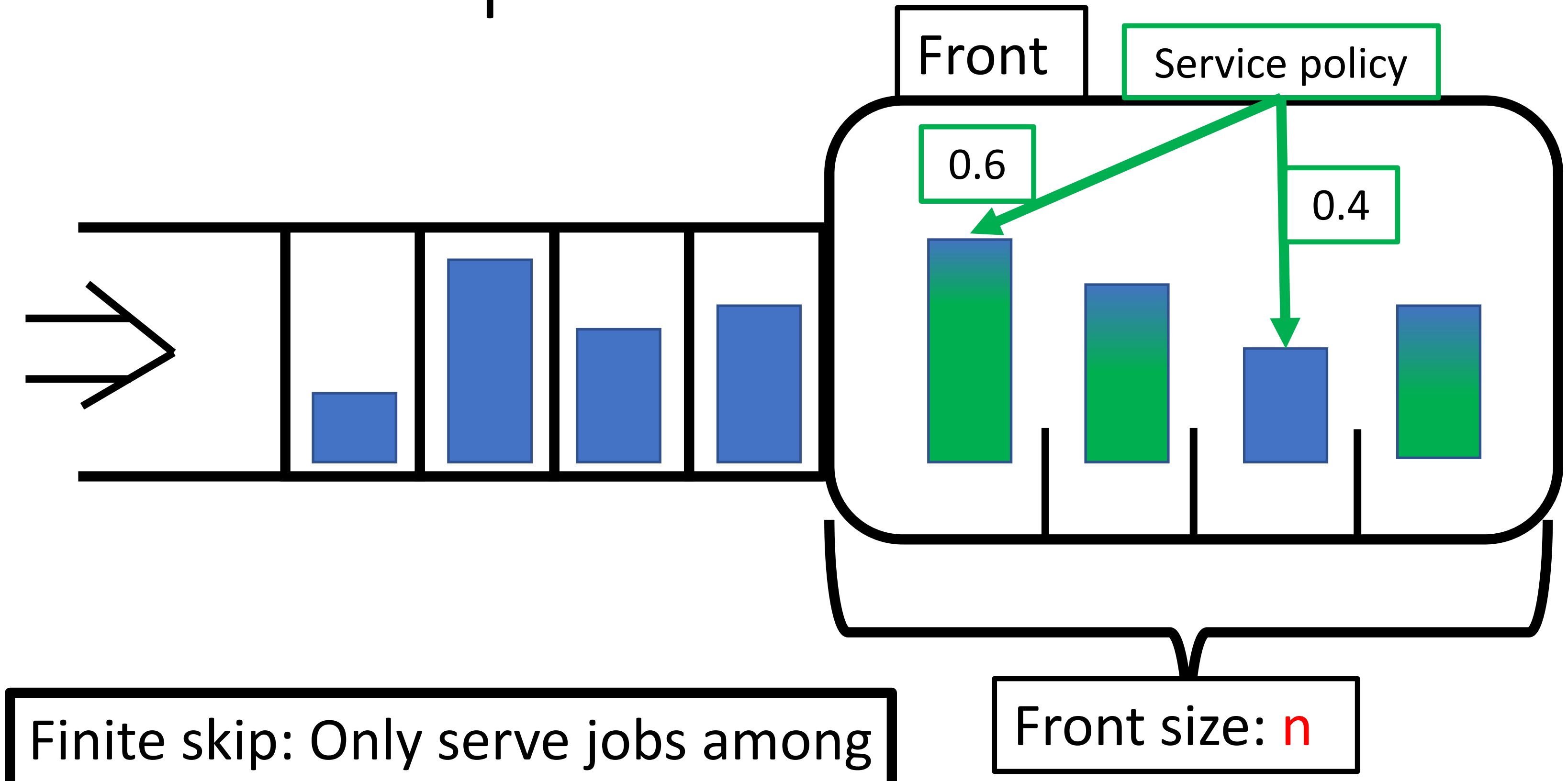
Service policy serves jobs at some service rate
Work completes at service rate, age increases
Job completes when age reaches size

Base Queueing Model



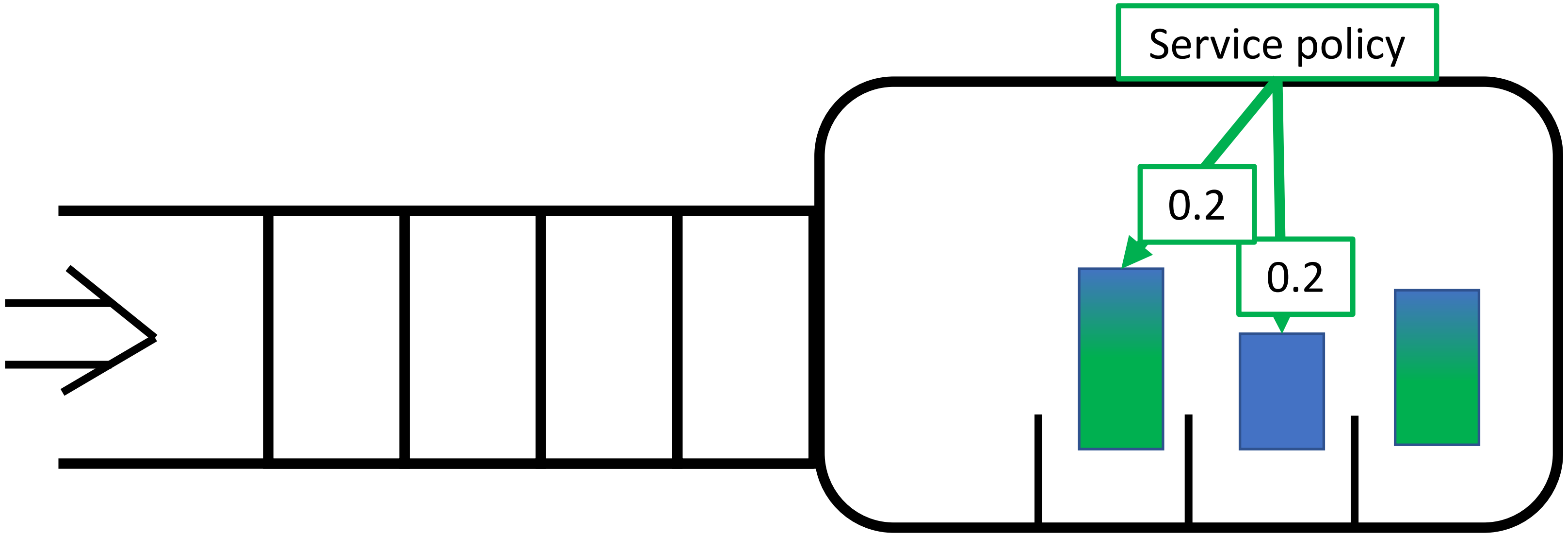
Service policy serves jobs at some service rate
Work completes at service rate, age increases
Job completes when age reaches size
Convention: normalize maximum service rate to 1.

Finite Skip

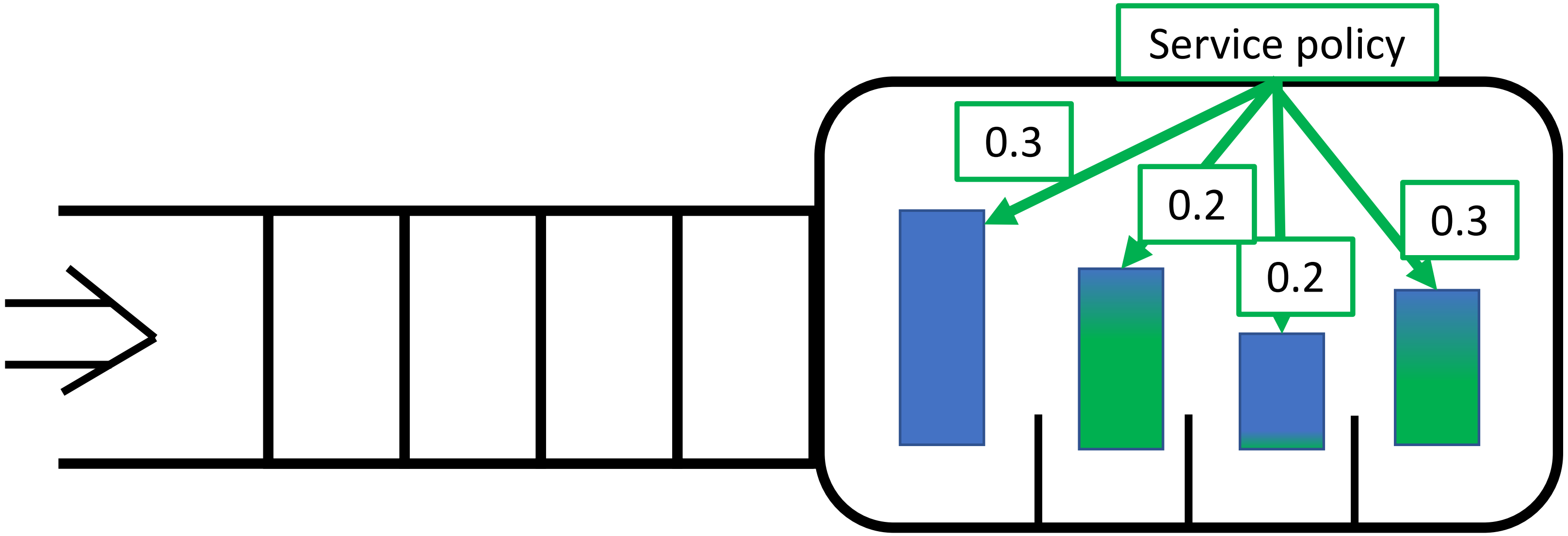


Finite skip: Only serve jobs among the n oldest jobs in arrival order.

Work conserving (for Finite Skip policies)



Work conserving (for Finite Skip policies)



Work conserving: If ≥ n jobs present, total service rate 1 (maximum)

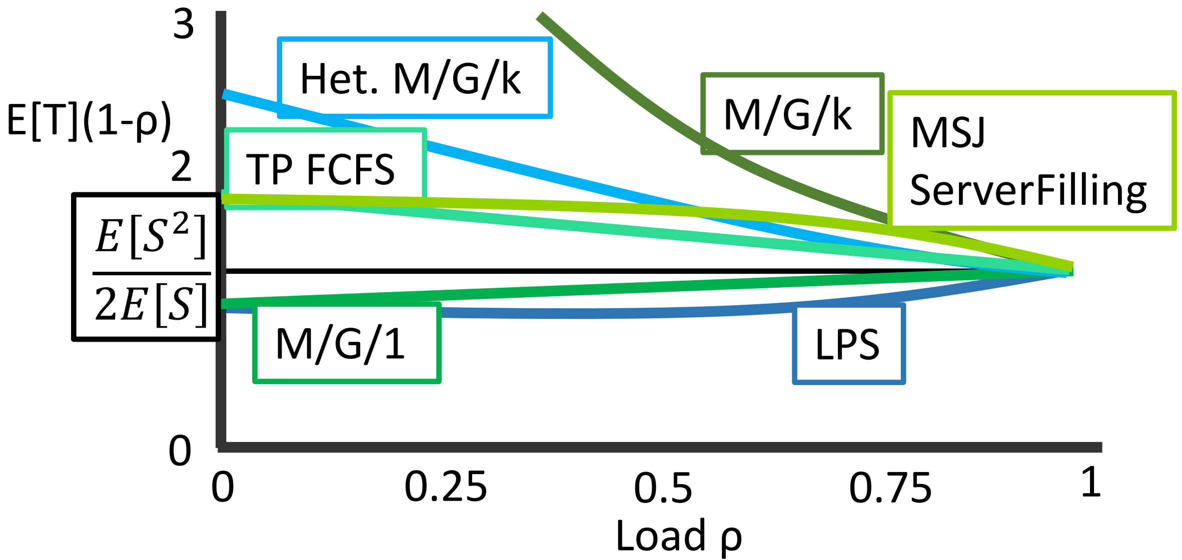
WCFS Policies

Polices that converge:

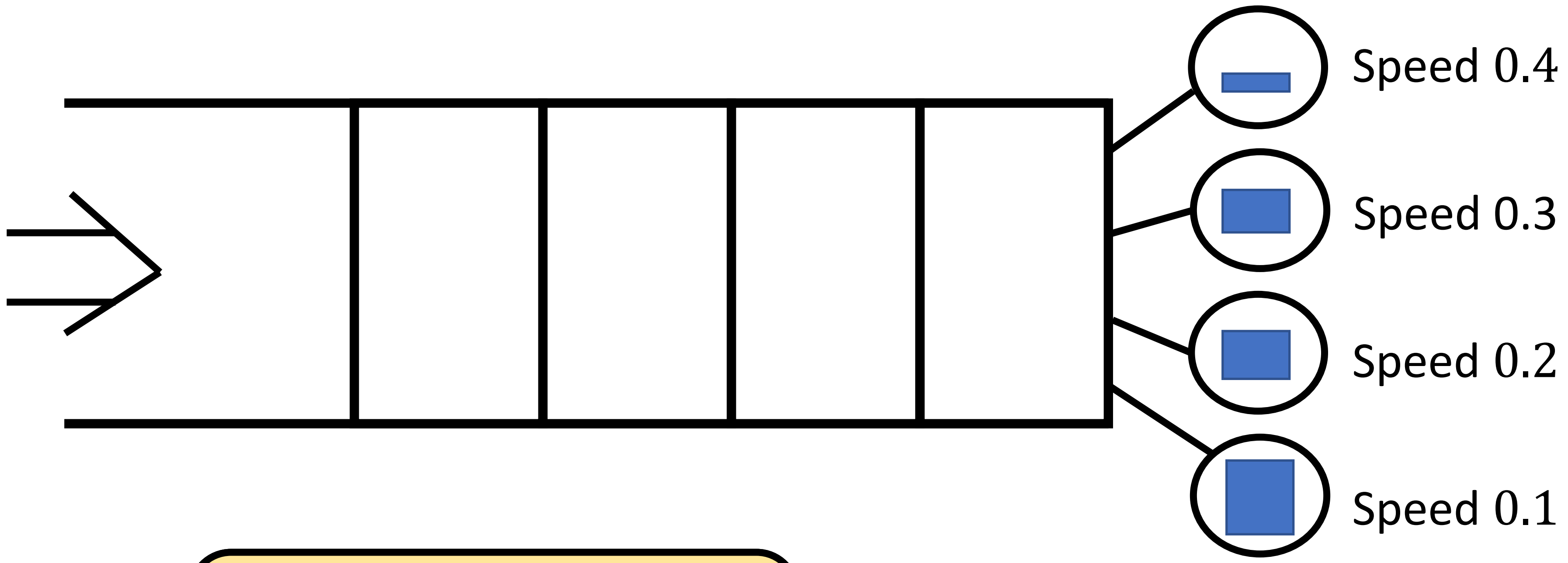
- M/G/1
- M/G/k
- Heterogeneous M/G/k
- Limited Processor Sharing
- Threshold Parallelism FCFS
- Multiserver-job ServerFilling

Polices that don't converge:

- M/G/1/SRPT
- (Unlimited) Processor Sharing
- Threshold Parallelism, most servers first
- Multiserver-job FCFS

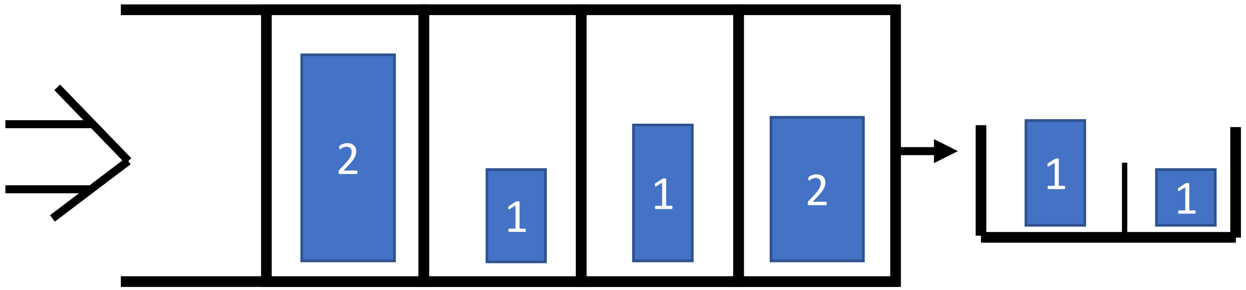


WCFS Example: Heterogeneous M/G/k



Is this model WCFS?
Yes!

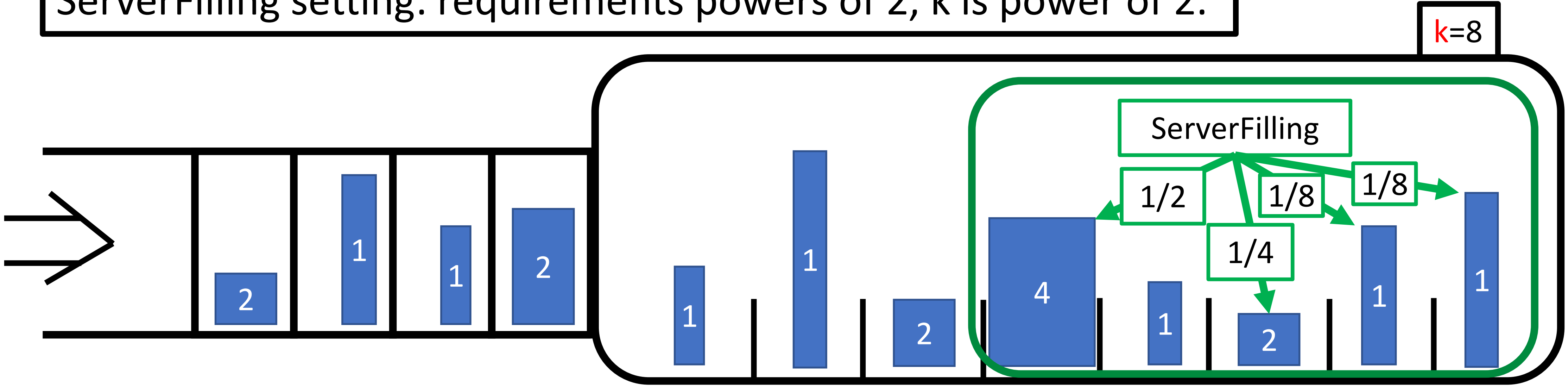
WCFS Example: Multiserver-Job FCFS



Is this model WCFS?
No, not work conserving.

Define ServerFilling for Multiserver-Job model

ServerFilling setting: requirements powers of 2, k is power of 2.



1. Find minimal prefix M containing jobs that require ≥k servers
2. Among prefix M, serve largest requirement first

Finite skip: $|M| \leq k$

Work conserving: Theorem: ServerFilling always fills all k servers.

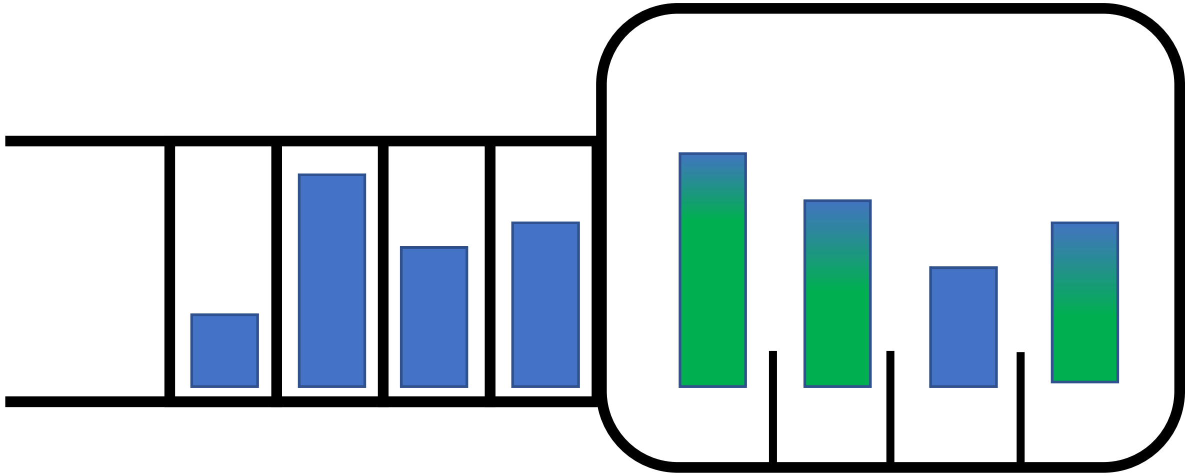
Is this model WCFS?
Yes!

Key ideas behind response time bound

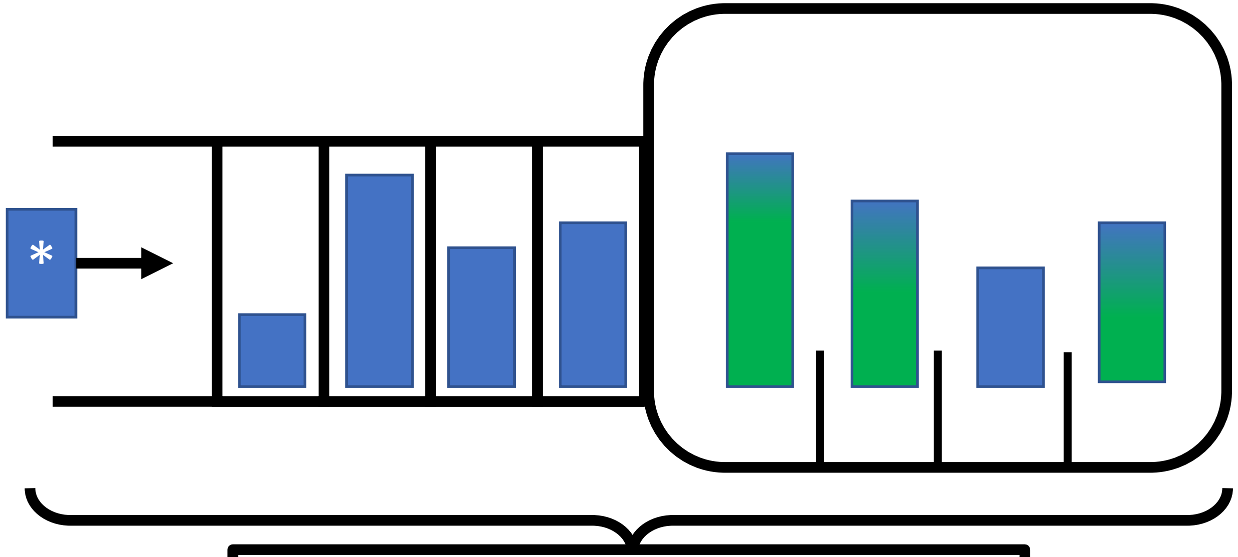
Want to prove: $E[T^{\pi}] - E\left[T^{M/G/1}\right] \in [c_l^{\pi}, c_h^{\pi}]$

Key ideas based on work W.

1. $E[T^{\pi}] \cong E[W^{\pi}]$
2. $E[W^{\pi}] \cong E[W^{M/G/1}]$

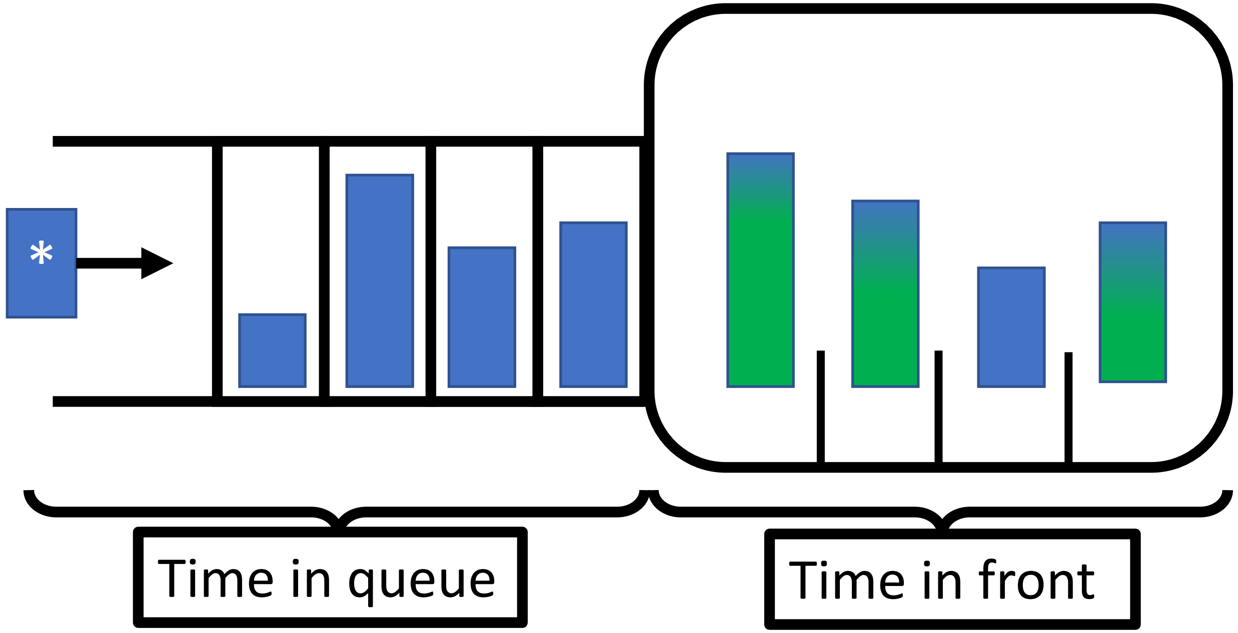


Idea 1: $E[T] \cong E[W]$



Q: Starred job's response time?

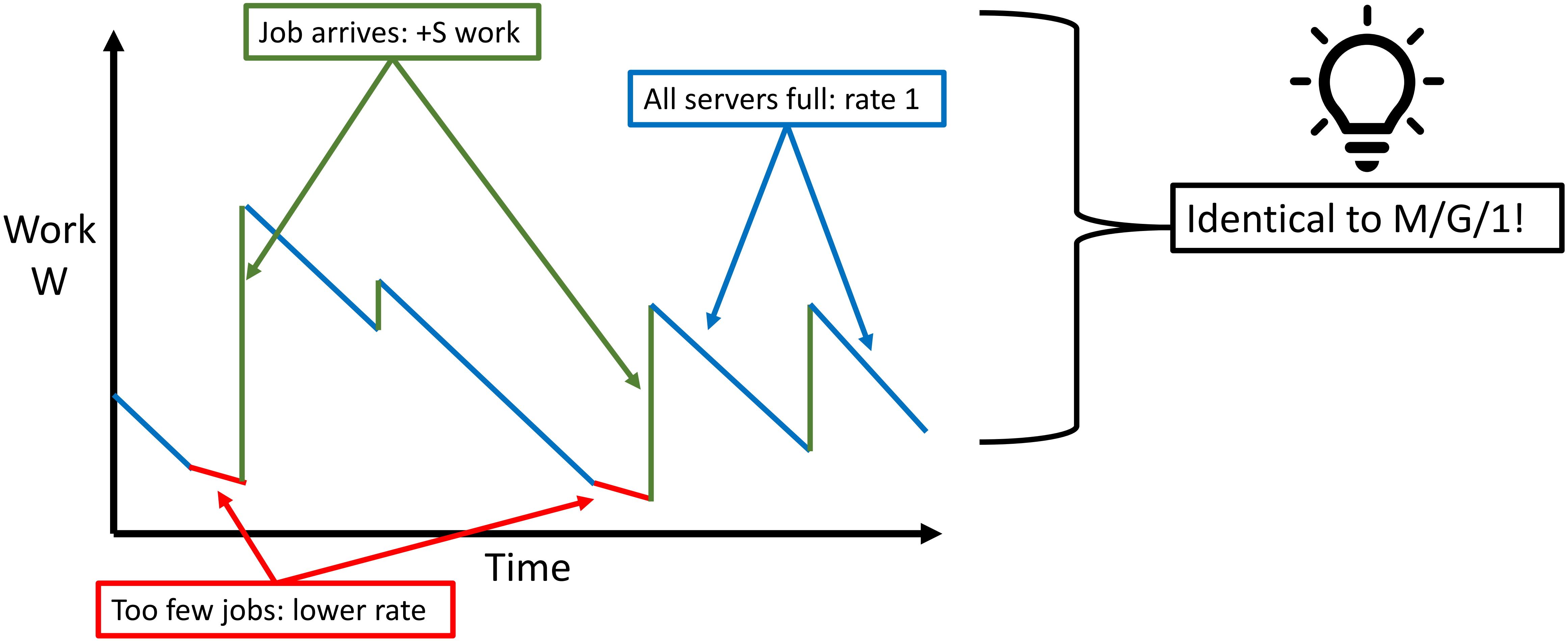
Idea 1: E[T] ≅ E[W]



Thm: Time in front bounded in expectation over all jobs.

Max time in queue: W^A, because completion rate 1.
Min time in queue: $W^A - n$ jobs
Differ by constant in expectation.

Idea 2: $E[W] \cong E\left[W^{M/G/1}\right]$ - Intuition



Proof of $E[W] \cong E\left[W^{M/G/1}\right]$

Consider $W(t)^2$

Consider $\frac{\mathrm{d}}{\mathrm{dt}} W(t)^2$

Expected stationary rate of change: $E\left[\frac{\mathrm{d}}{\mathrm{dt}} W^2\right] = 0$

Increases: Jump up by S, at rate λ.

E[inc.]: $\lambda E[(W+S)^2 - W^2] = 2\lambda E[W]E[S] + \lambda E[S^2] = 2\rho E[S] + \lambda E[S^2]$

Decreases: Work completes at rate B(t). Rate 1 if ≥n jobs.

E[dec.]: $2E[WB] = 2E[W] - 2E[W(1-B)]$

$$E[W] = \frac{\rho}{1-\rho} \frac{E[S^2]}{2E[S]} + \frac{E[W(1-B)]}{1-\rho} = E\left[W^{M/G/1}\right] + \frac{E[W(1-B)]}{1-\rho}$$

Proof of $E[W] \cong E\left[W^{M/G/1}\right]$

$E[W] = E\left[W^{M/G/1}\right] + \frac{E[W(1-B)]}{1-\rho}$

If $B < 1$, W consists of $< n$ jobs, so $W < c$.

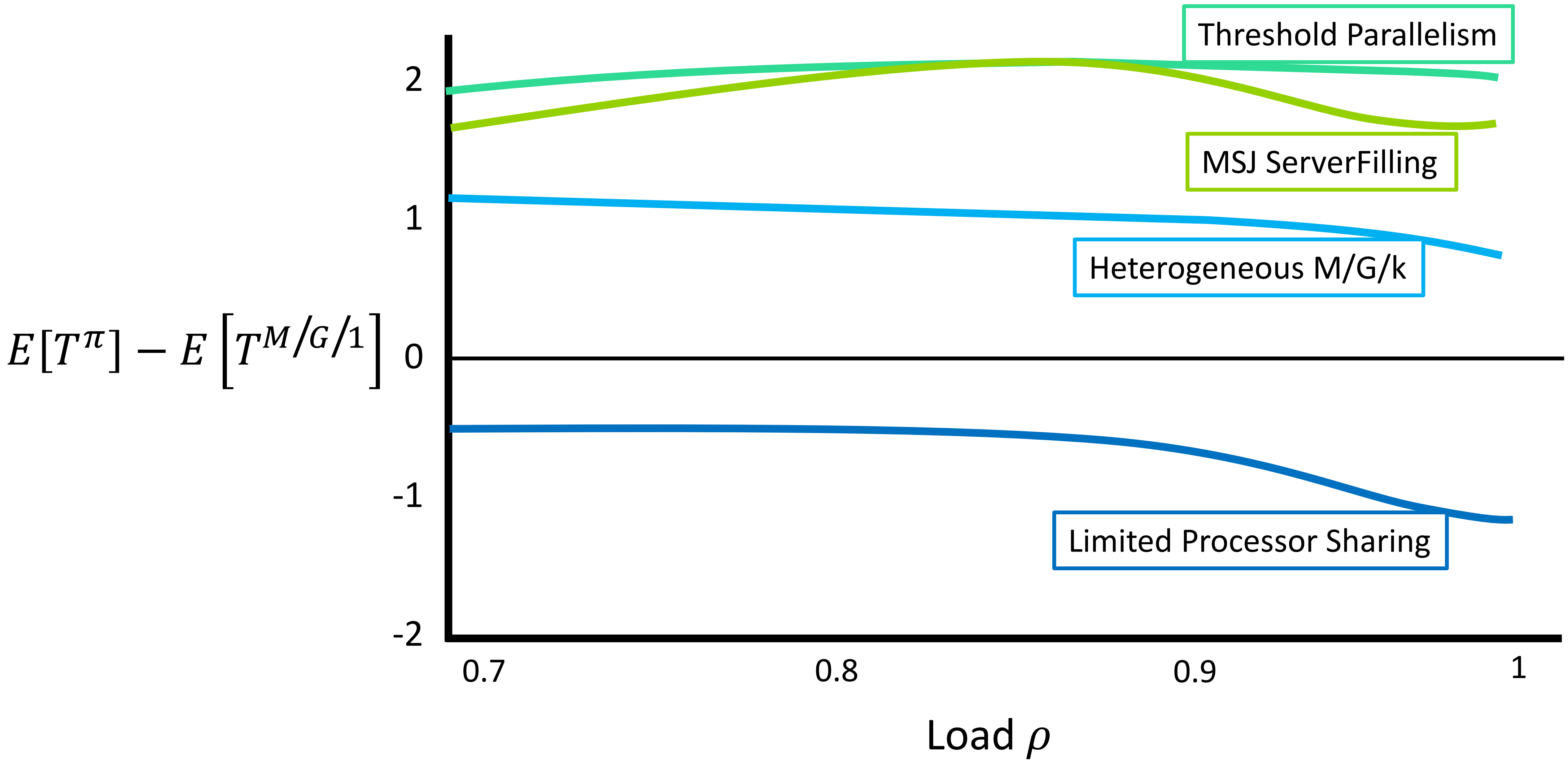
$E[W(1-B)] \leq c\,E[1-B] = c(1-\rho)$

$E[W] \leq E\left[W^{M/G/1}\right] + c$

$E[W] \cong E\left[W^{M/G/1}\right]$

Completes proof that $E[T^{\pi}] - E\left[T^{M/G/1}\right] \in [c_l^{\pi}, c_h^{\pi}]$

Empirical validation



Future questions

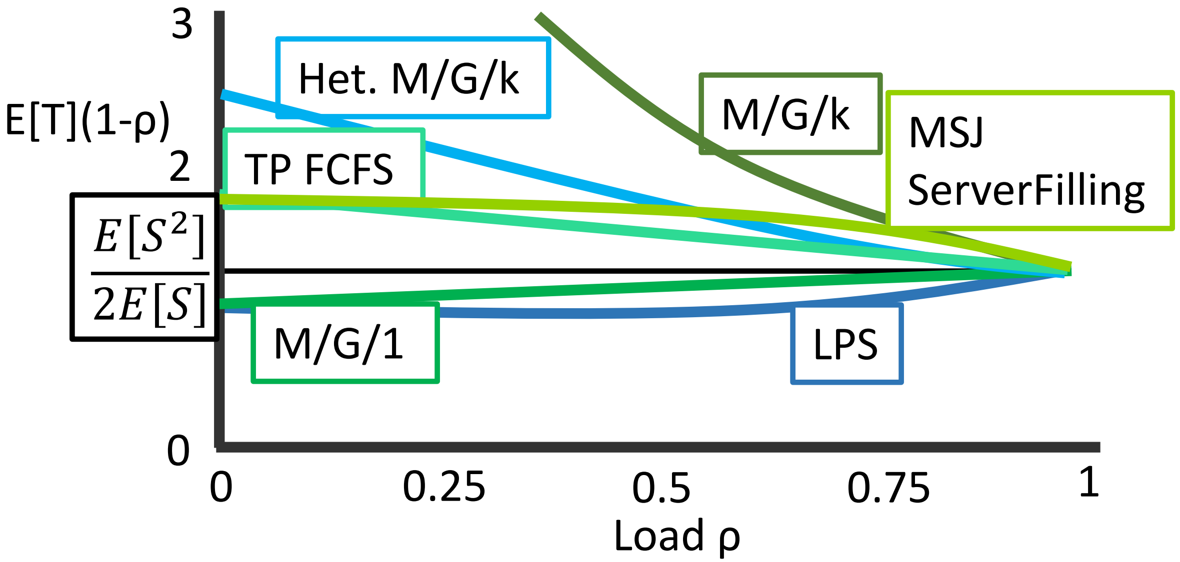
- Work conserving and finite skip, relative to a different job ordering, e.g. SRPT?
- Finite skip but not work conserving?
- Finite skip in expectation?

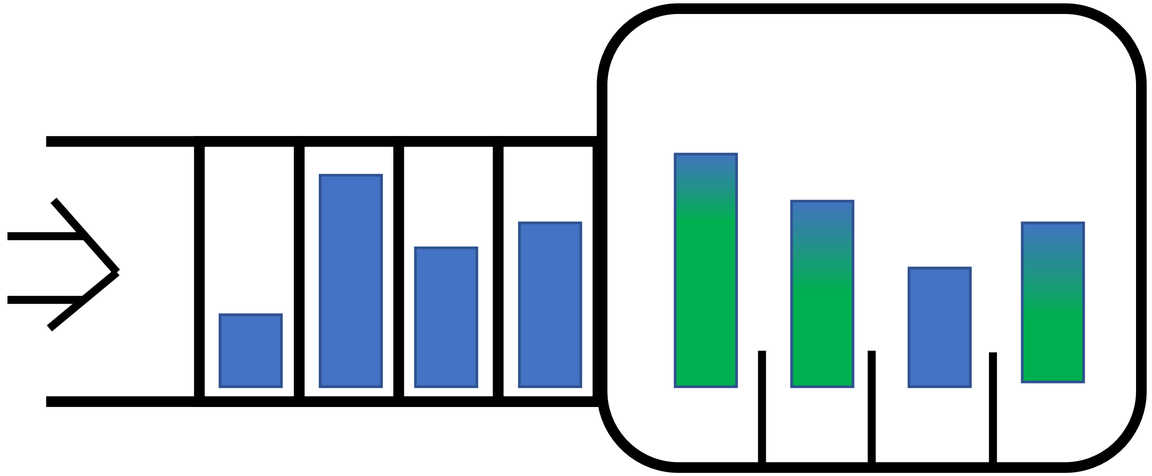
Conclusion

Explained unexpected similarity between queues

Defined work conserving finite skip models

Tightly bounded mean response for all WCFS models.





Extra: Condition on S

Bounded expected remaining size:

Exists constant c such that for all ages a,

$$E[S - a \mid S > a] < c$$

Extra: DivisorFilling

Like ServerFilling, MSJ policy that fills all servers if enough jobs available. In particular, WCFS policy.

DivisorFilling works whenever all requirements divide the number of servers k. If k jobs present, will fill all servers.

If jobs can have requirements that don’t divide k, WCFS not possible.