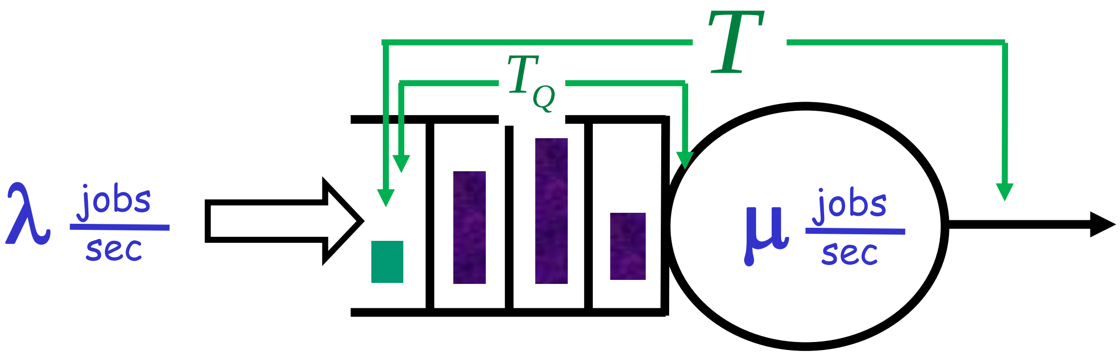
Variability in service time



S: job size

$$E[S] = \frac{1}{\mu}$$

$$\rho = \lambda E[S] = \frac{\lambda}{\mu}$$

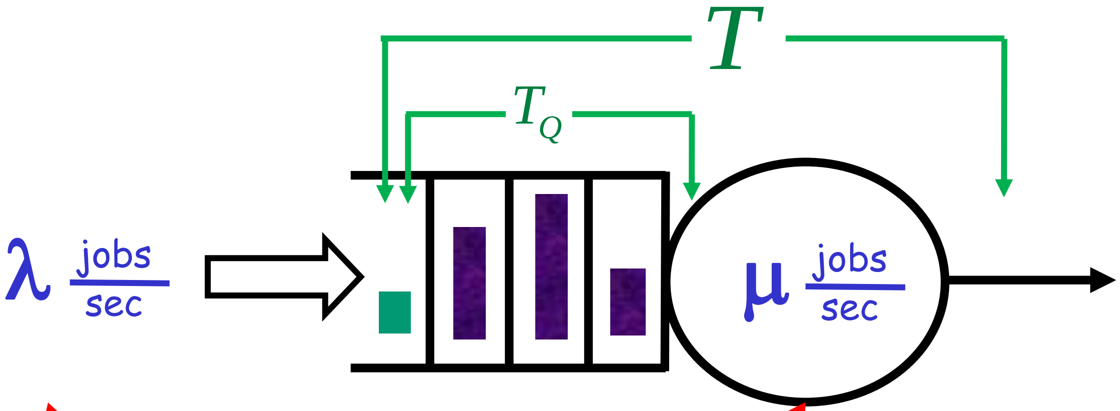
$T =$ response time

$T_Q =$ queueing time (waiting time)

Q: Given that $\lambda < \mu$, what causes wait?

A: <u>Variability</u> in the arrival process & service requirements

Variability



S : job size

$$E[S] = \frac{1}{\mu}$$

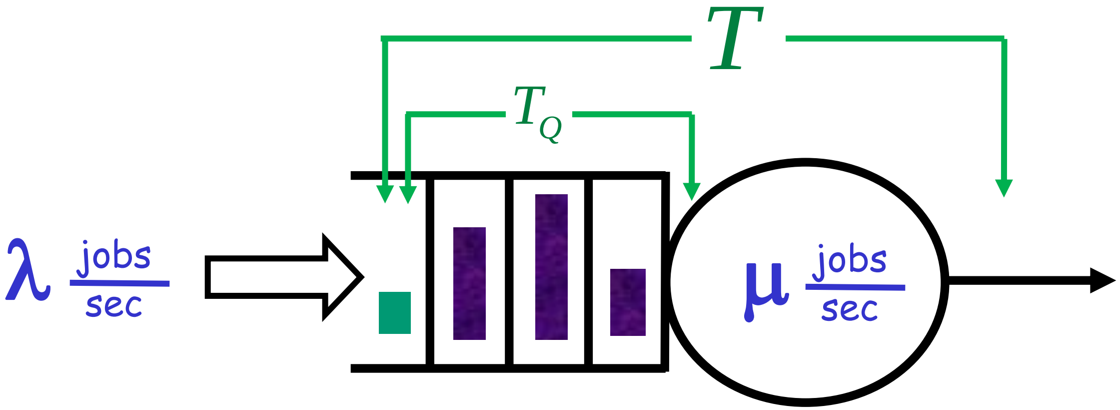
$$\rho = \lambda E[S] = \frac{\lambda}{\mu}$$

Variability in arrival process

Variability in job size, S

$$E[T_Q] = \frac{\rho}{1-\rho} \cdot \frac{E[S^2]}{2E[S]}$$

M/G/1



S : job size

$$E[S] = \frac{1}{\mu}$$

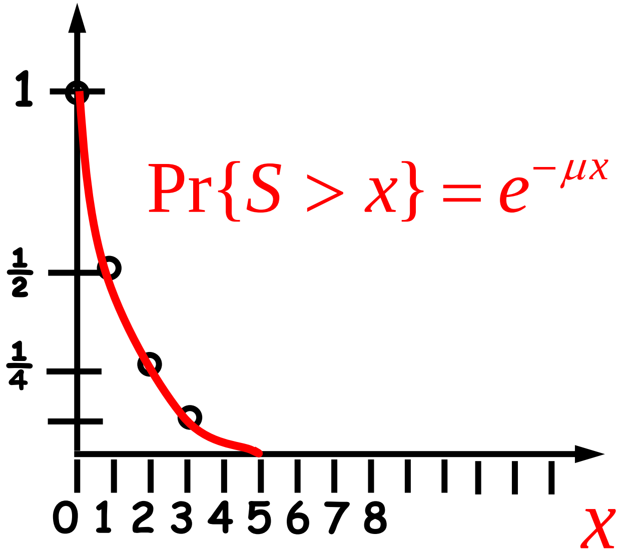
$$\rho = \lambda E[S] = \frac{\lambda}{\mu}$$

$$E[T_Q] = \frac{\rho}{1-\rho} \cdot \frac{E[S^2]}{2E[S]}$$

Job Size Distributions

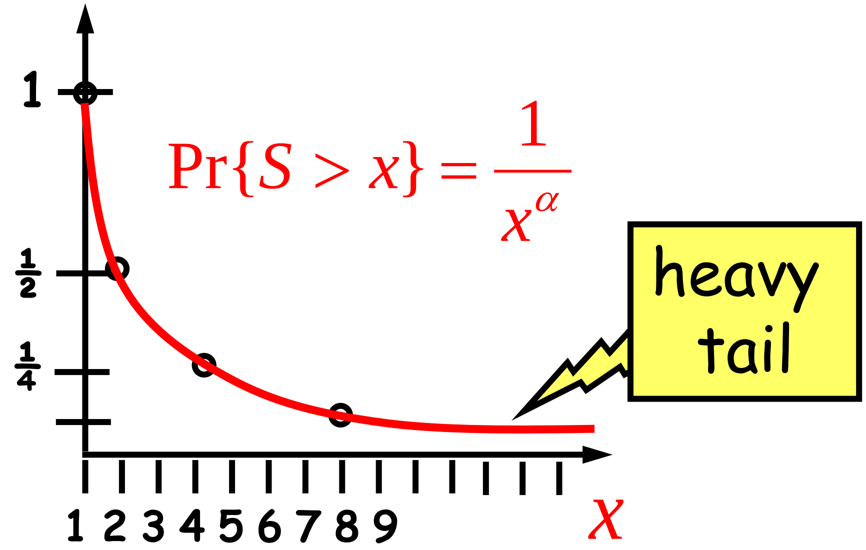
"Most jobs are small; few jobs are large"

$S \sim \text{Exp}(\mu)$



$\Pr\{S > x\} = e^{-\mu x}$

$S \sim \text{Pareto}(\alpha)$



$\Pr\{S > x\} = \dfrac{1}{x^{\alpha}}$

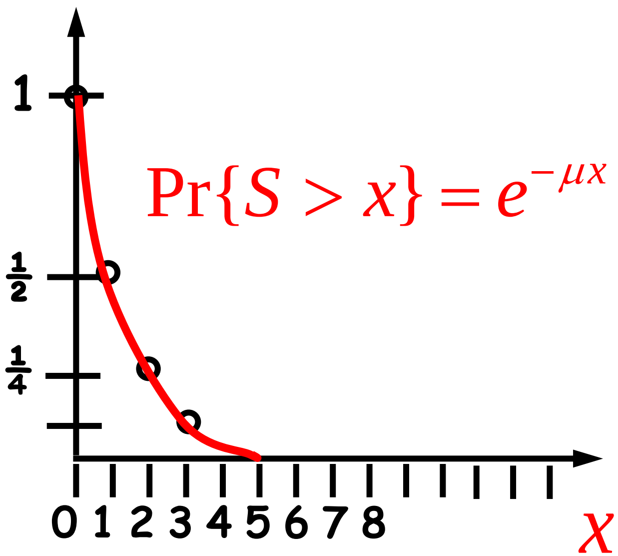
Job Size Distributions

QUESTION: Which best represents UNIX process lifetimes?

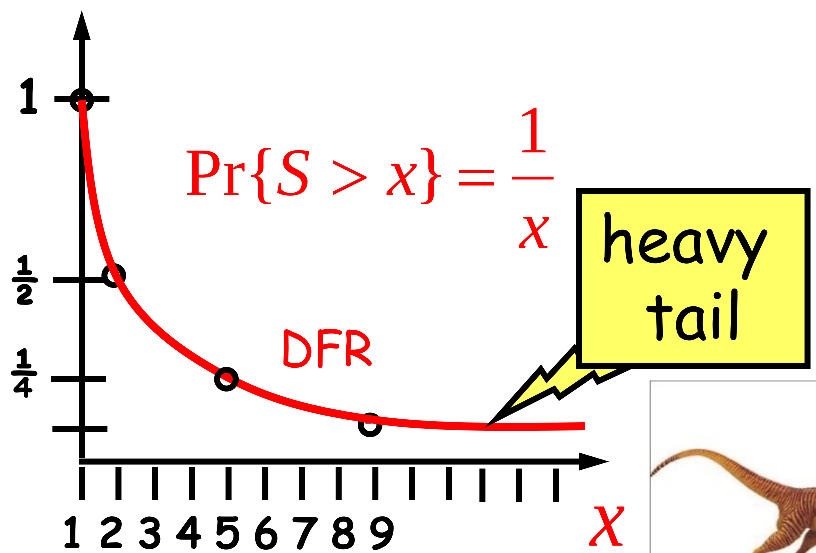
QUESTION: For which do top 1% of jobs comprise 50% of load?

QUESTION: Which distribution fits the saying, "the longer a job has run so far, the longer it is expected to continue to run."

$S \sim \text{Exp}(\mu)$



$S \sim \text{Pareto}(\alpha = 1)$



Pareto Job Size Distribution

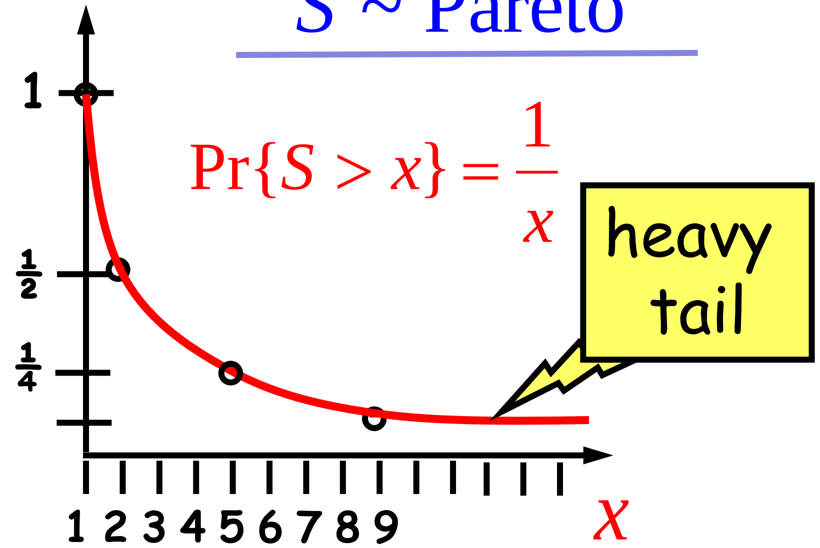
Pareto job sizes are ubiquitous in CS:

- CPU lifetimes of UNIX jobs [Harchol-Balter, Downey 96]
- Supercomputing job sizes [Schroeder, Harchol-Balter 00]
- Web file sizes [Crovella, Bestavros 98], [Barford, Crovella 98]
- IP flow durations [Shaikh, Rexford, Shin 99]
- Wireless call durations [Blinn, Henderson, Kotz 05]

Also ubiquitous in nature:

- Forest fire damage
- Earthquake damage
- Human wealth
  [Vilfredo Pareto '65]

$S \sim$ Pareto



Variability in Job Sizes

$C^2 = 0$ Deterministic

$C^2 \approx .02$ Human IQs

$C^2 = \frac{1}{6}$ Uniform(0,b) - for any b

$C^2 = 1$ Exponential distribution

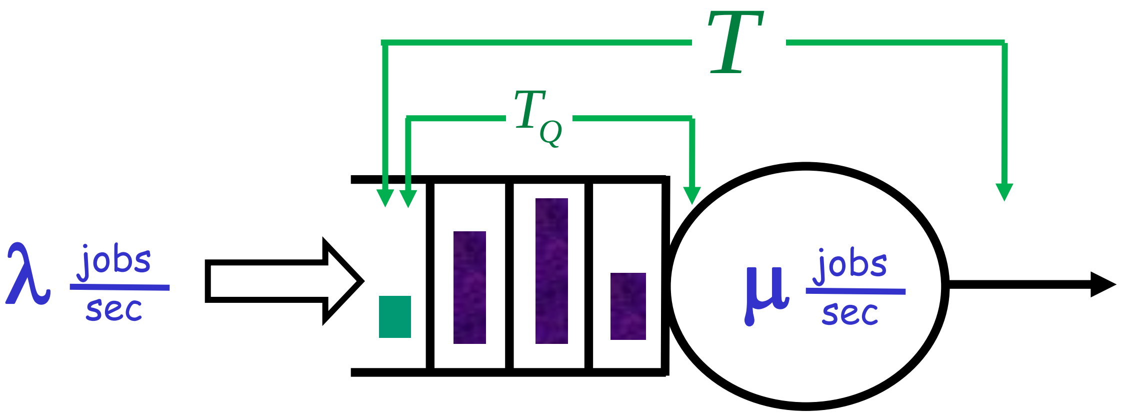
$C^2 = 50 - 100$ Unix process lifetimes

$C^2 = \infty$ Pareto distribution

Squared Coefficient of Variation

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

Single-Server Queue



S : job size

$$E[S] = \frac{1}{\mu}$$

$$\rho = \lambda E[S] = \frac{\lambda}{\mu}$$

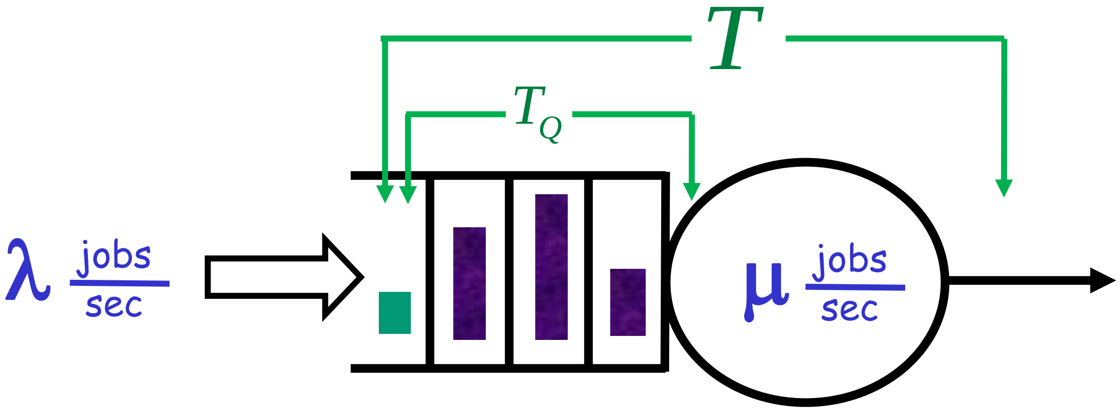
D/D/1 — Deterministic service times

M/M/1 — Exponential inter-arrival times; Exponential service times; 1 server

M/G/1 — General service times

M="memoryless"="Markovian"

Single-Server Queue



S : job size

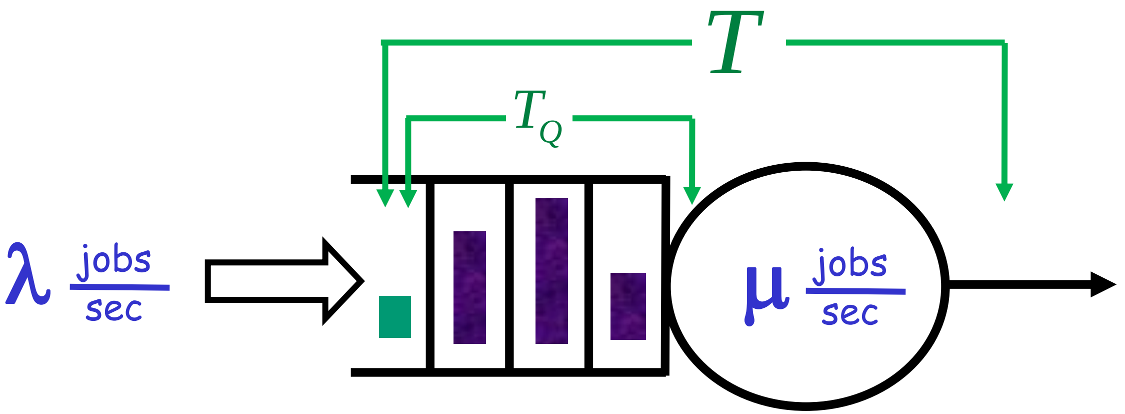
$$E[S] = \frac{1}{\mu}$$

$$\rho = \lambda E[S] = \frac{\lambda}{\mu}$$

D/D/1 **M/M/1** **M/G/1**

Q: Does low ρ → low $E[T_Q]$?

Single-Server Queue



S : job size

$$E[S] = \frac{1}{\mu}$$

$$\rho = \lambda E[S] = \frac{\lambda}{\mu}$$

D/D/1

$$E[T_Q] = 0$$

M/M/1

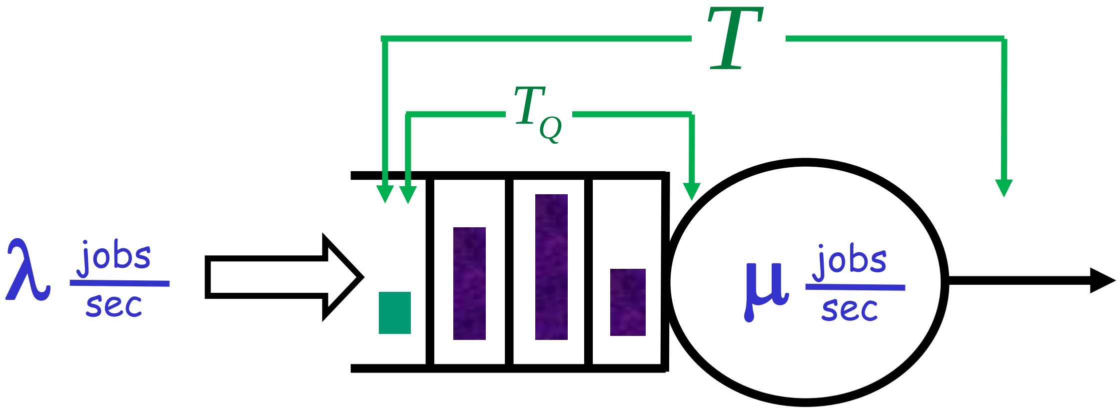
$$E[T_Q] = \frac{\rho}{1-\rho} \cdot E[S]$$

M/G/1

$$E[T_Q] = \frac{\rho}{1-\rho} \cdot \frac{E[S^2]}{2E[S]}$$

related to C^2: variability job size

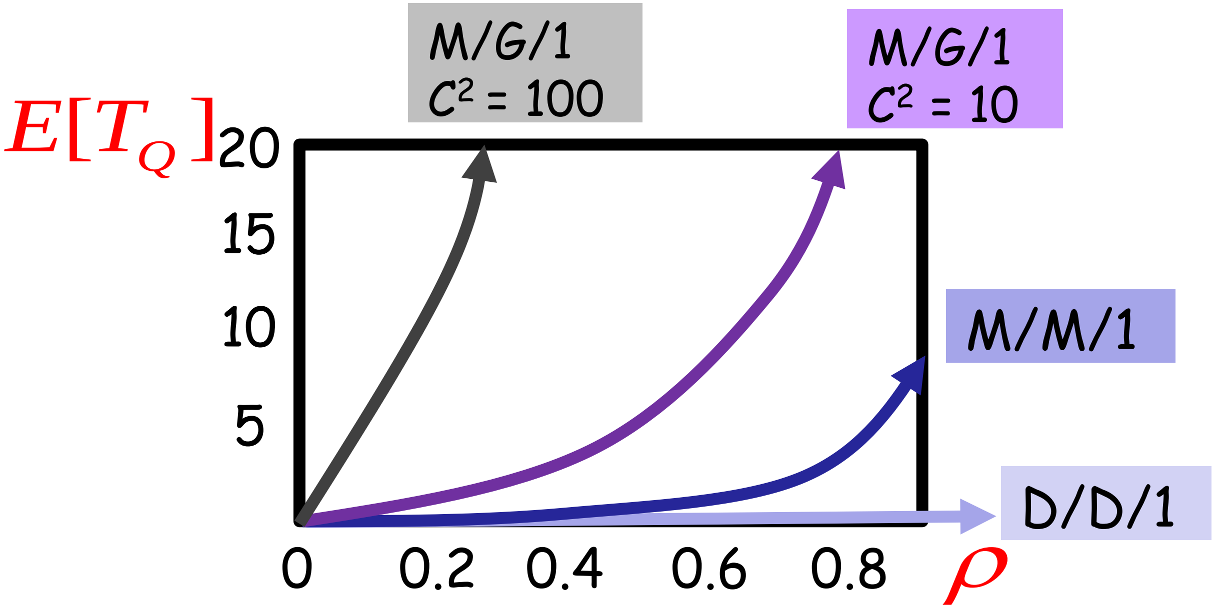
Single-Server Queue



S : job size

$$E[S] = \frac{1}{\mu}$$

$$\rho = \lambda E[S] = \frac{\lambda}{\mu}$$



low load
does NOT imply
low wait

M/G/1

$$E[T_Q] = \frac{\rho}{1-\rho} \cdot \frac{E[S^2]}{2E[S]}$$

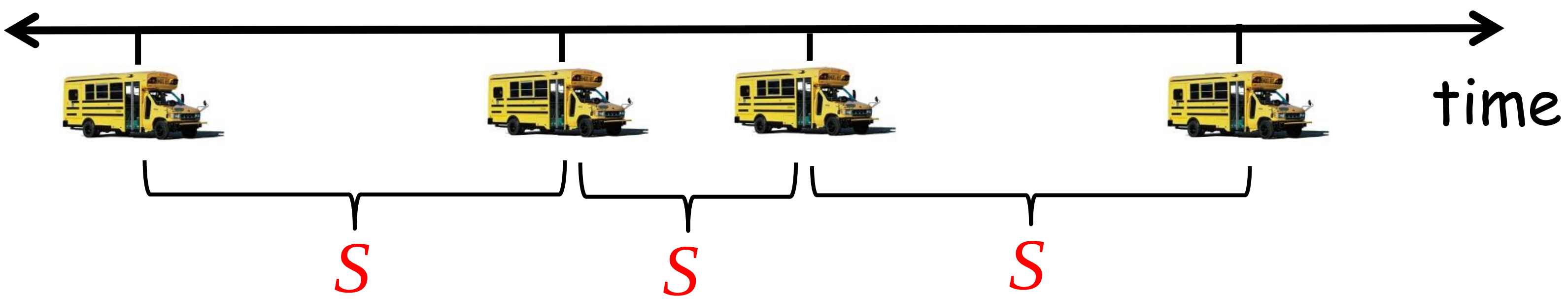
Where is this
coming from?

Waiting for the bus

Waiting for the bus

S: time between buses

$$E[S] = 10\,\mathrm{min}$$



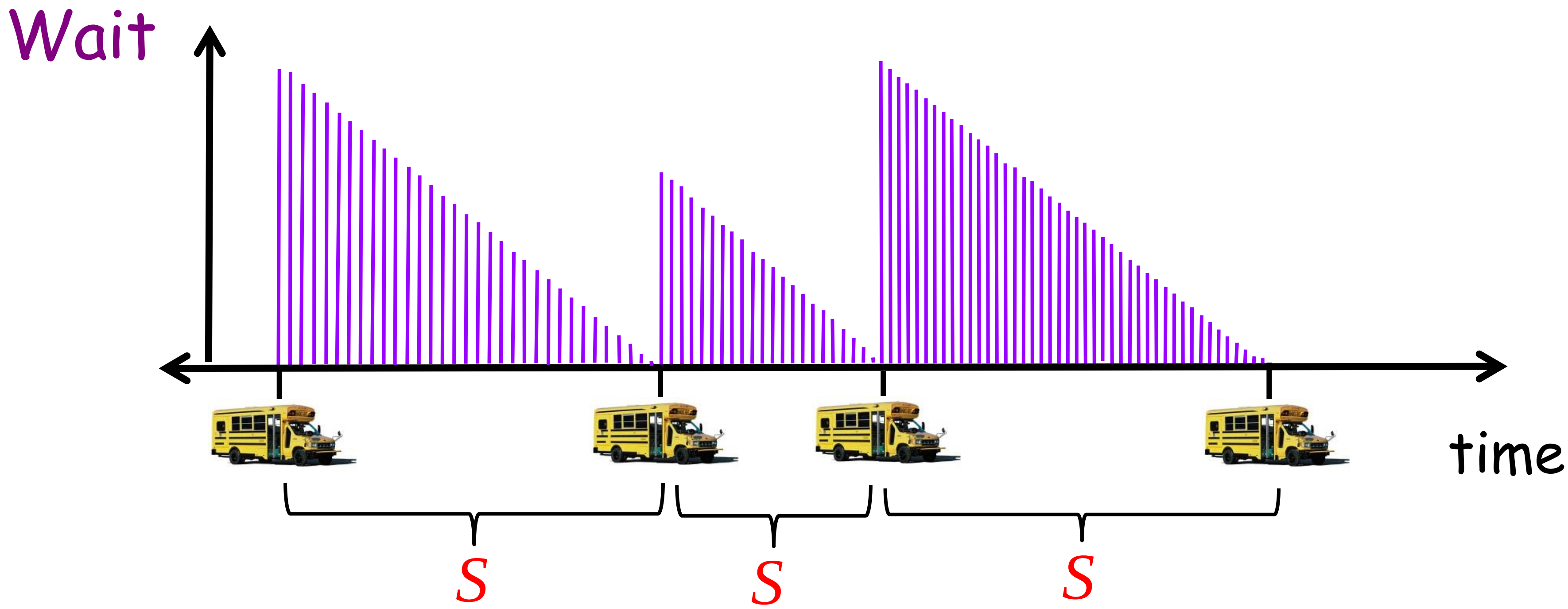
QUESTION:

On average, how long do I have to wait for a bus?

(a) < 5 min
(b) 5 min
(c) 10 min
(d) >10 min

Waiting for the bus

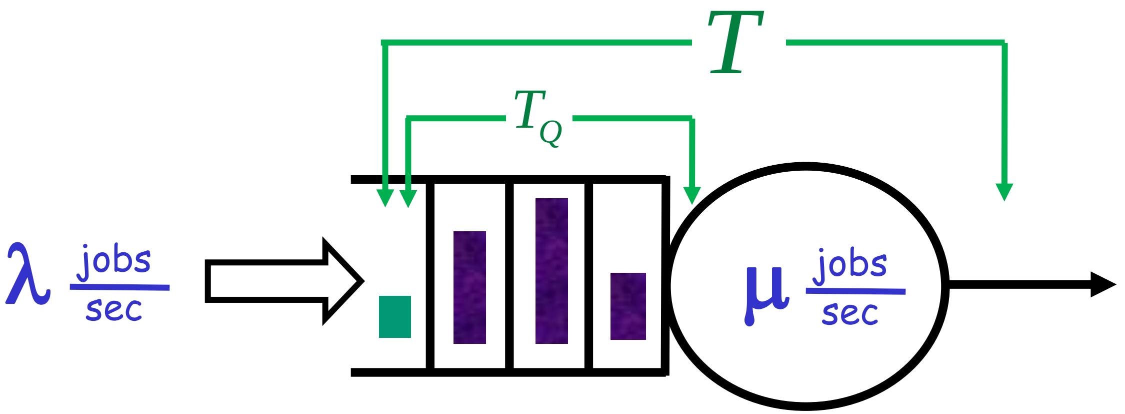
"It is higher the probability to arrive when there is longer interarrival time"



S: time between buses

$$E[\mathrm{Wait}] = \frac{E[S^2]}{2E[S]} >> E[S]$$

Back to Single-Server Queue



S : job size

$$E[S] = \frac{1}{\mu}$$

$$\rho = \lambda E[S] = \frac{\lambda}{\mu}$$

$$E[T_Q]^{M/G/1} = \frac{\rho}{1-\rho} \cdot \frac{E[S^2]}{2E[S]}$$

Low $\rho \not\Rightarrow$ Low $E[T_Q]$

Waiting for the Loo

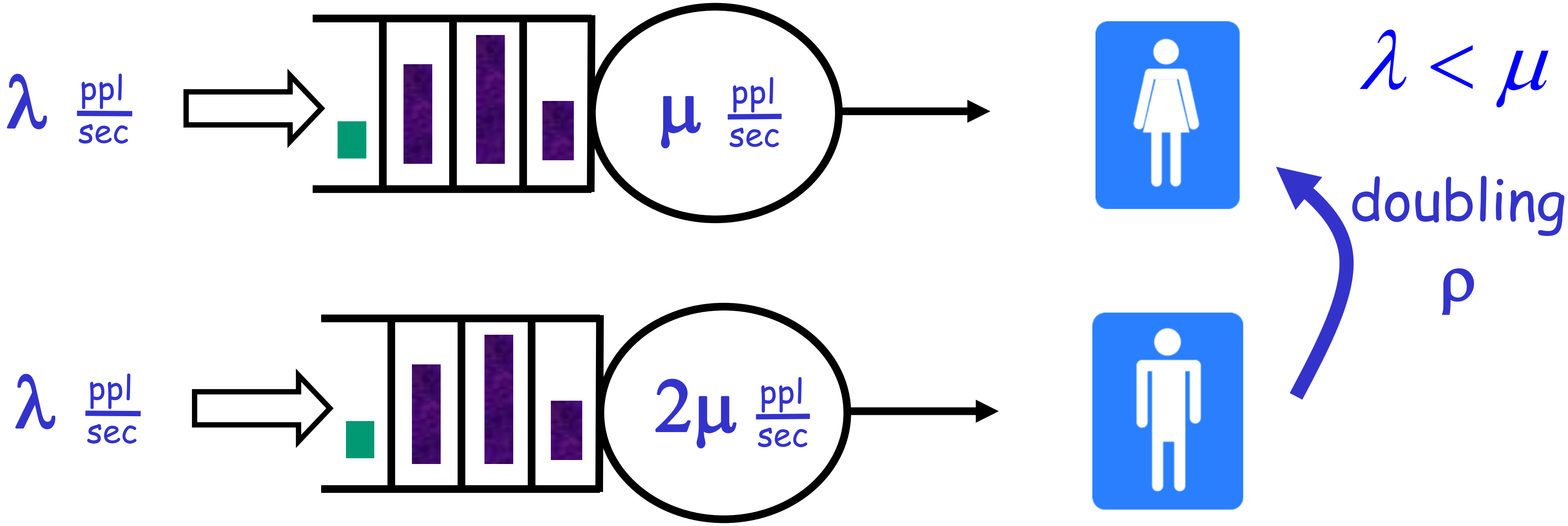
Check out the line for the men's room …

Waiting for the Loo

- On avg, Women spend 88 sec in loo.
- On avg, Men spend 40 sec in loo.

Waiting for the Loo

M/M/1 model



QUESTION:

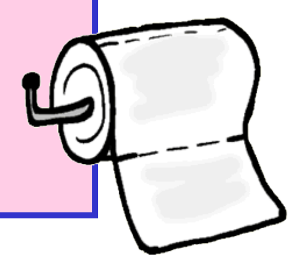
Women take 2X as long. What's the difference in their wait?

(a) factor < 2

(b) factor 2

(c) factor 4

(d) factor > 4



Waiting for the Loo

M/M/1

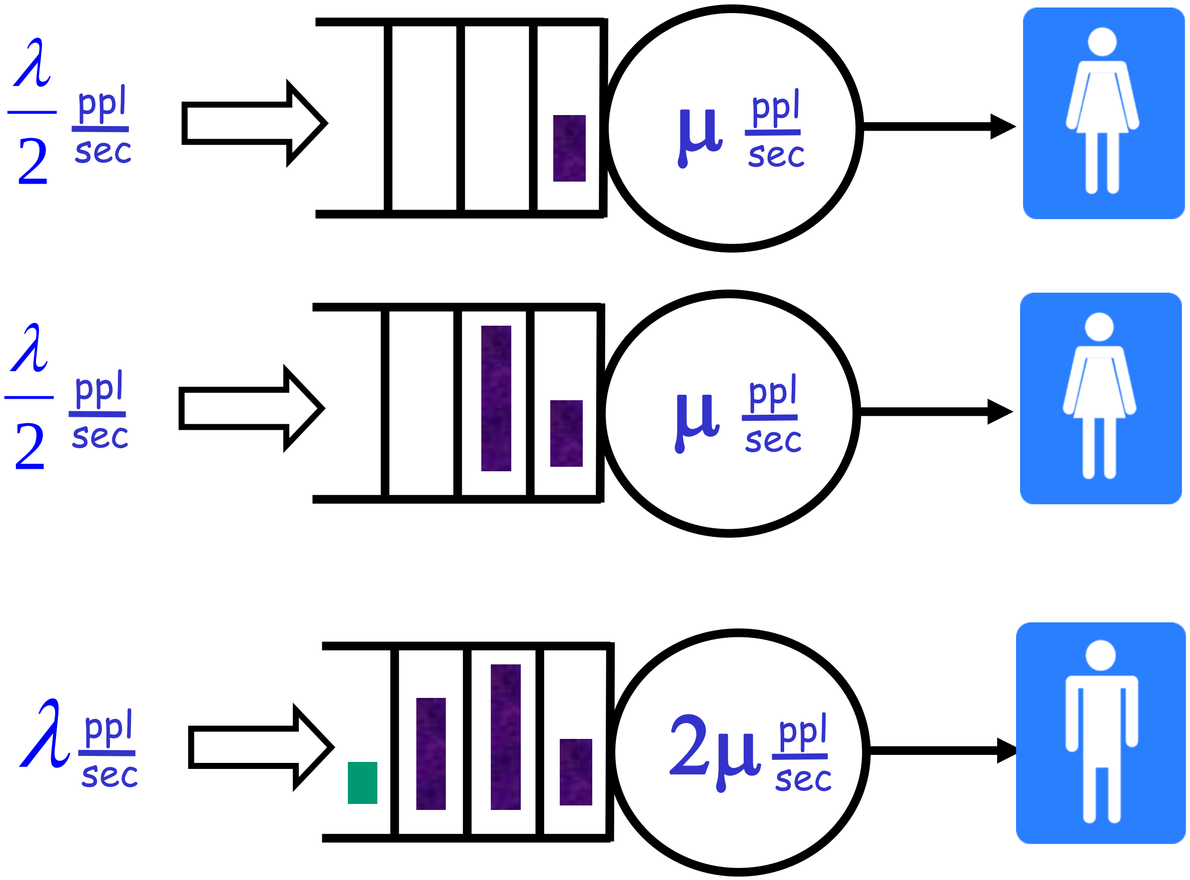
$$E[T_Q] = \frac{\rho}{1-\rho} \cdot E[S]$$

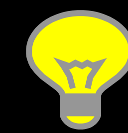
M/G/1

$$E[T_Q] = \frac{\rho}{1-\rho} \cdot \frac{E[S^2]}{2E[S]}$$

Doubling ρ can increase $\mathbf{E[T_Q]}$ to ∞

Equalizing the wait for men & women

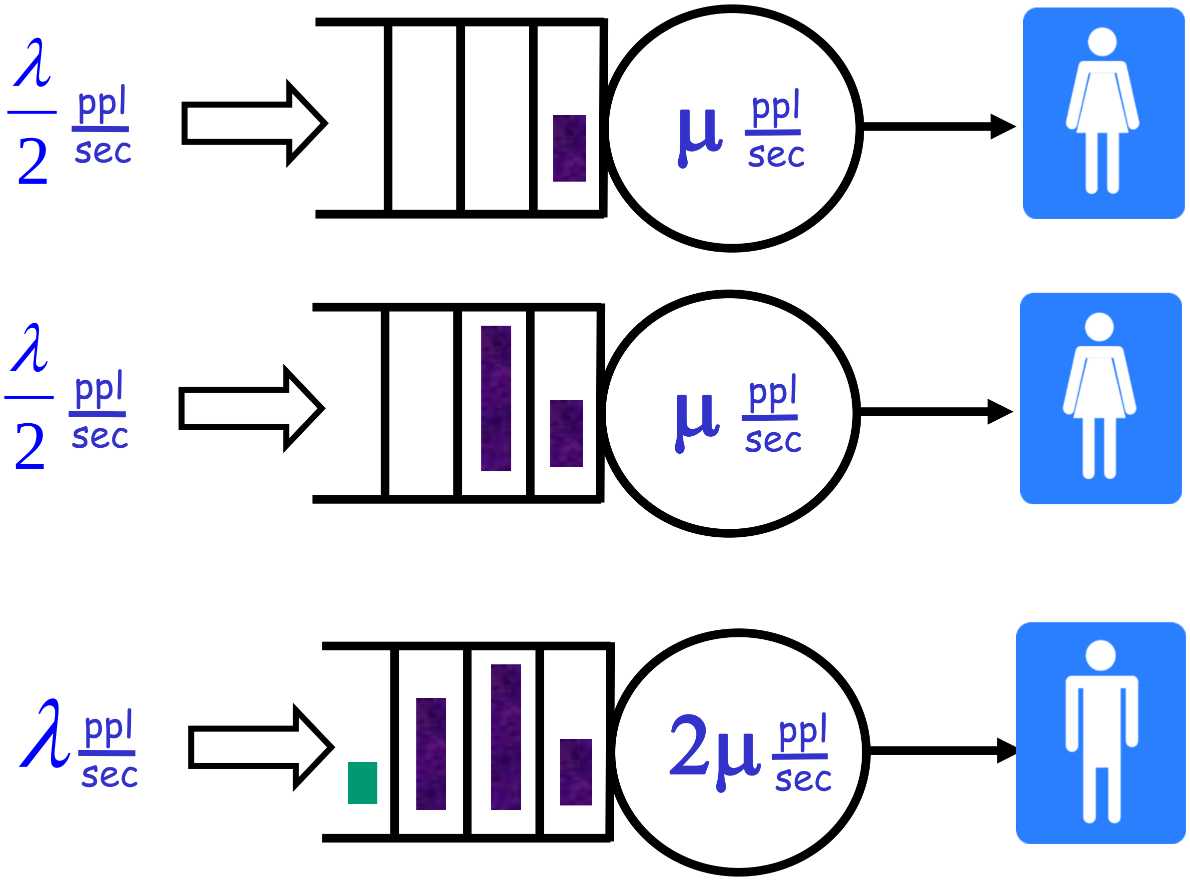




2 Women's rooms for each Men's room.

QUESTION:
Is this (a) insufficient (b) overkill (c) just right

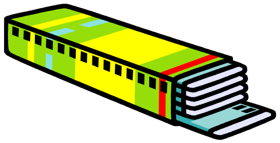
Equalizing the wait for men & women



Insufficient!

Waiting time for women is still factor of 2 higher.

$$E[T_Q] = \frac{\rho}{1-\rho} \cdot E[S]$$



Also true under M/G/1 model.

M/G/1

$$E[T_Q] = \frac{\rho}{1-\rho} \cdot \frac{E[S^2]}{2E[S]}$$

High load leads to high wait

High job size variability leads to high wait

To drop load, we can increase server speed.

Q: What can we do to combat job size variability?

A: Smarter scheduling!

Scheduling in M/G/1

QUESTION:
Which scheduling policy is best for minimizing $\mathrm{E}[T]$?

FCFS (First-Come-First-Served, non-preemptive)

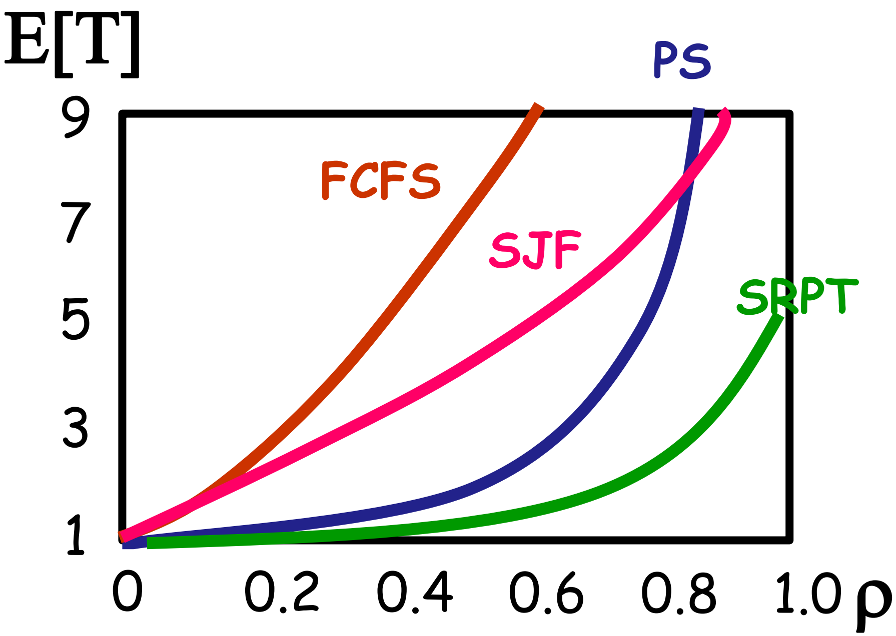
PS (Processor-Sharing, preemptive)

SJF (Shortest-Job-First, non-preemptive)

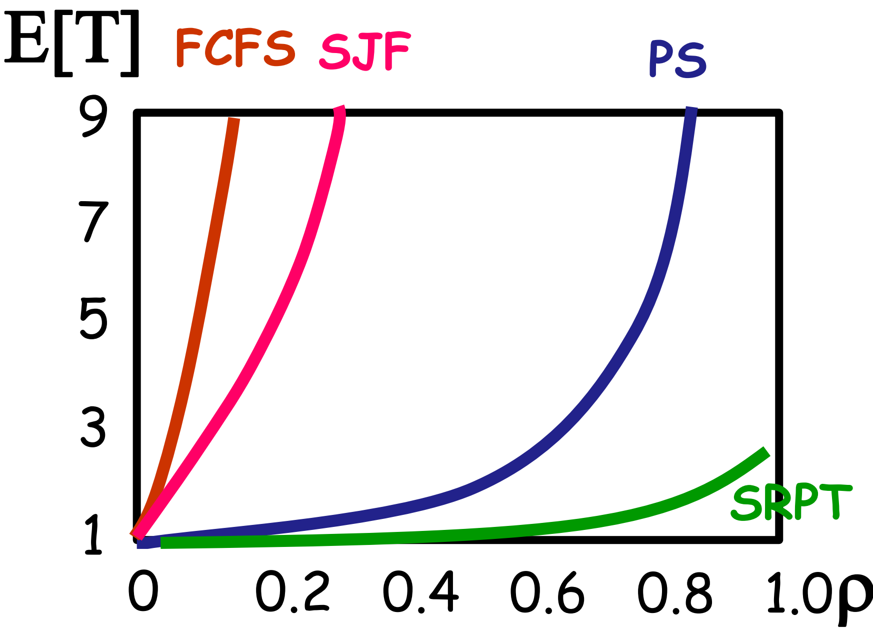
SRPT (Shortest-Remaining-Processing-Time, preemptive)

[Harchol-Balter EORMS 2011]

Scheduling in M/G/1



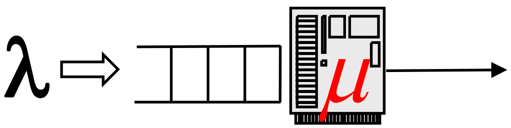
$C^2=10$



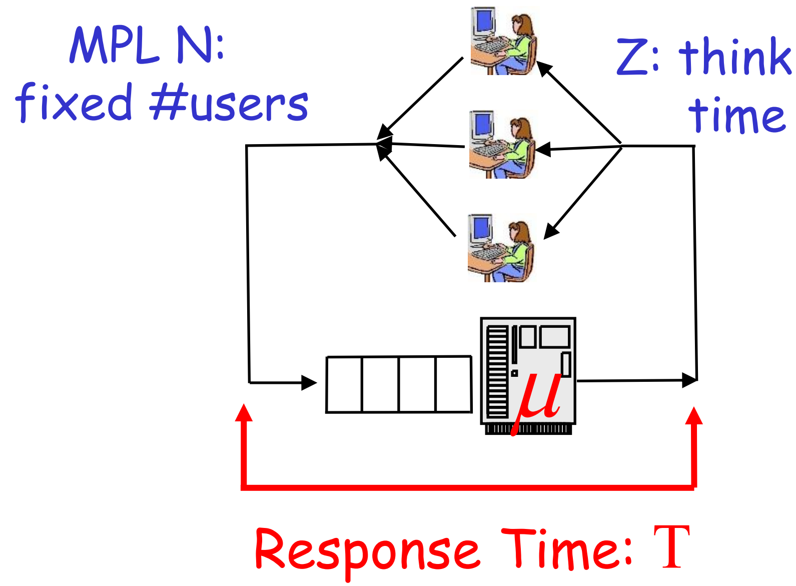
$C^2=100$

Caution: Open versus Closed

<u>Open</u>



<u>Closed</u>

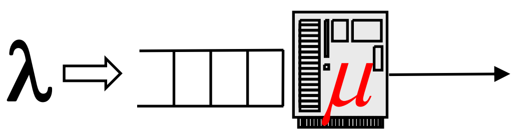


QUESTION: When run with same load ρ, which has higher $E[T]$?

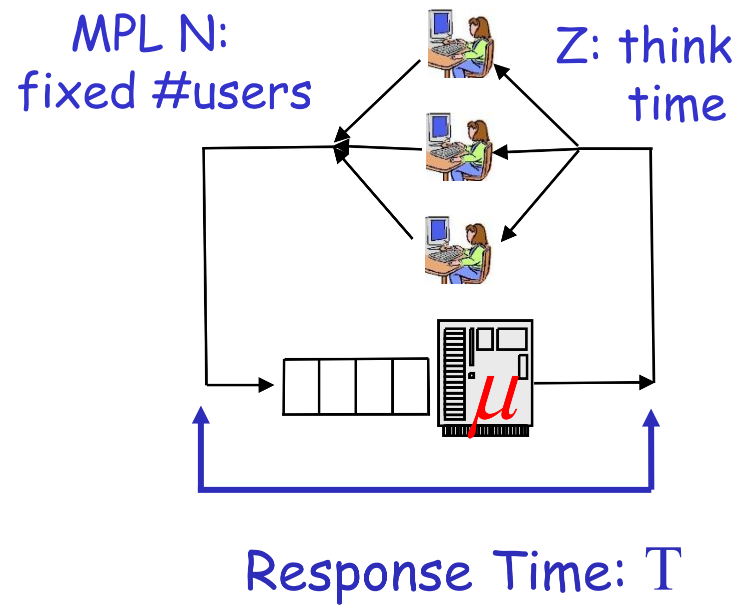
(a) Open
(b) Closed
(c) Same

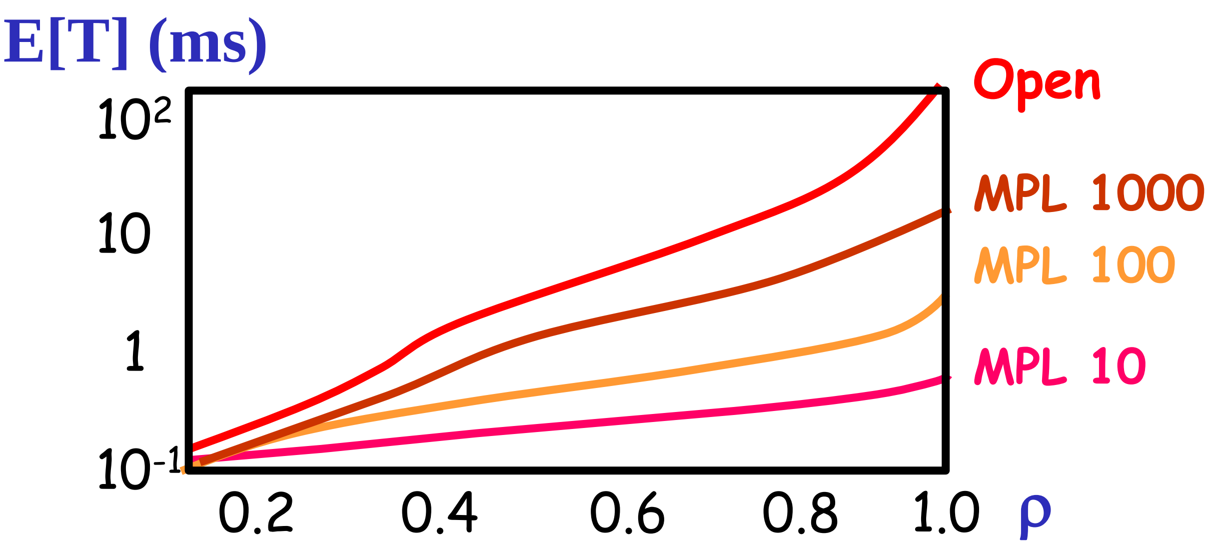
Caution: Open versus Closed

<u>Open</u>



<u>Closed</u>





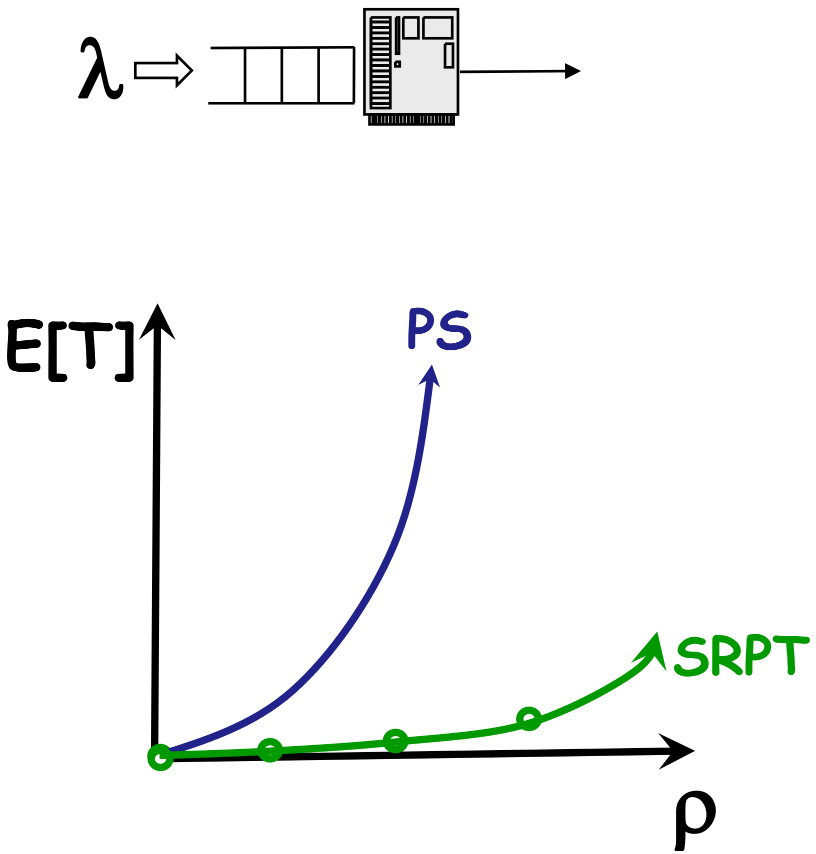
Performance of Auction Site

[Schroeder, Wierman, Harchol-Balter NSDI 2006]

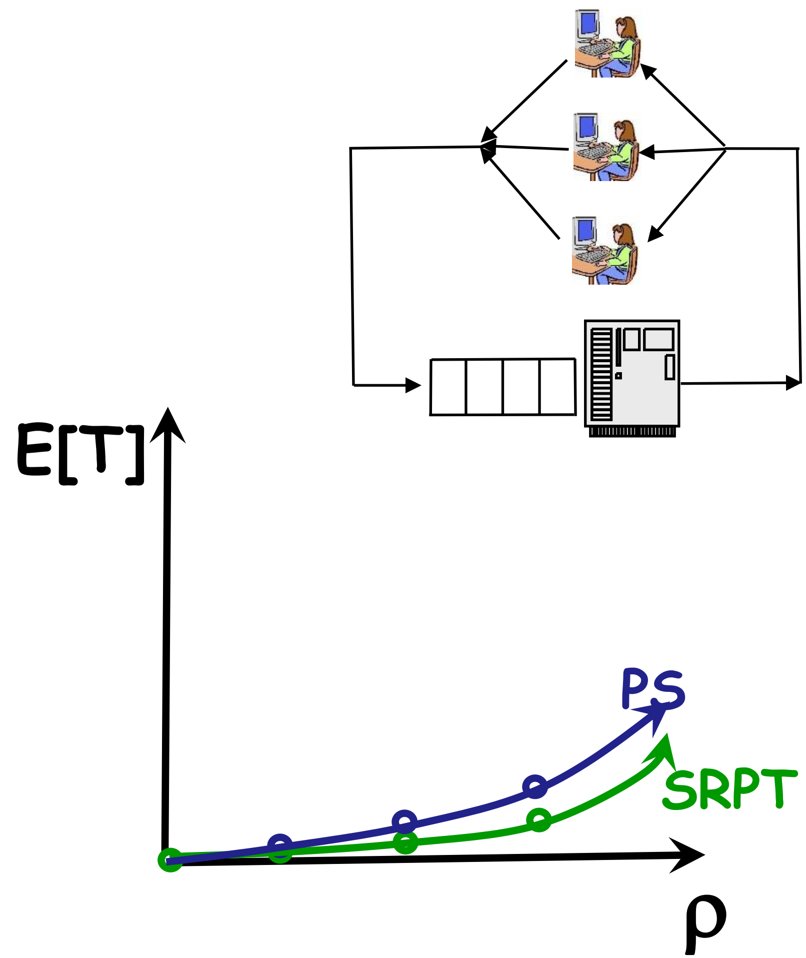
E[T] much lower for closed system w/ same ρ

Caution: Open versus Closed

Open



Closed



Closed & open systems run w/ same job size distribution and same load.

[Schroeder, Wierman, Harchol-Balter, NSDI 06]