

CSC 2232: Topics in Computer System Performance and Reliability

Bianca Schroeder

Department of Computer Science
University of Toronto

WHO AM I

WHAT IS THIS CLASS ABOUT?

System performance & reliability

- Is one fast server better or two slow servers?
- How many data replicas do I need for reliability?
- How do I generate/simulate realistic workloads?
- What is the impact of workload characteristics?

⇒ Methods, tools, back-of-the envelope calculations for system evaluation, simulation & measurement.

⇒ Study of recent papers in the area.

LOGISTICS

- Class time: Friday 11am-1pm
 - Can we move to Wed 2-4pm
 - Does anybody have serious scheduling conflicts?
- Office hours: Wed 4-5pm
 - Plus open door policy
- Class web page
 - www.cs.toronto.edu/~bianca/csc2232.html

GRADING

- 30% class participation
 - Participating in class discussion
 - Class presentation of at least one research paper
- 70% class project
 - I will suggest possible projects
 - You can propose your own
 - Final results: Conference style paper
- No exams!

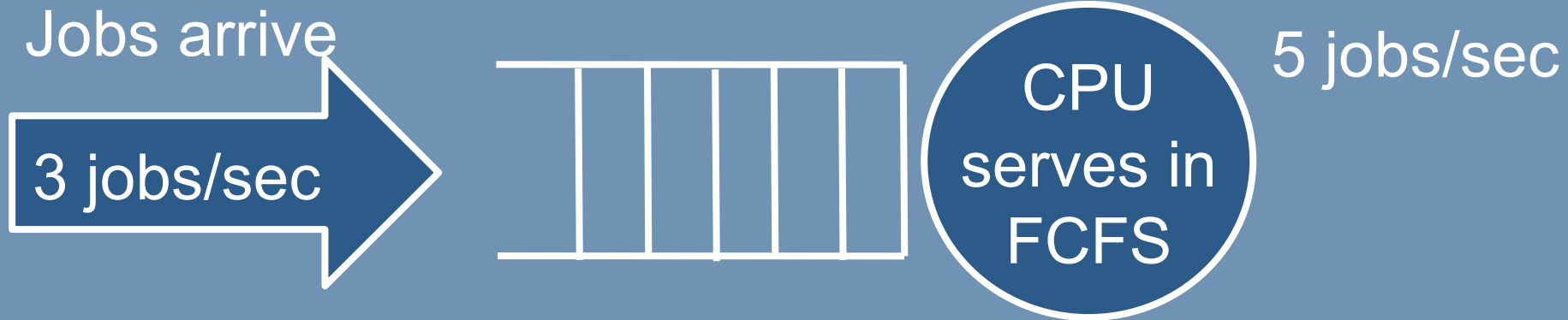
OUTLINE

- Administrivia ✓
- Motivating examples
- Questionnaire

MOTIVATIONAL EXAMPLES / CASE STUDY

- System design is often a *counter-intuitive process*
- Do not expect to understand *everything* in the examples
- Don't worry if you're not familiar with all terminology
- Ask questions!

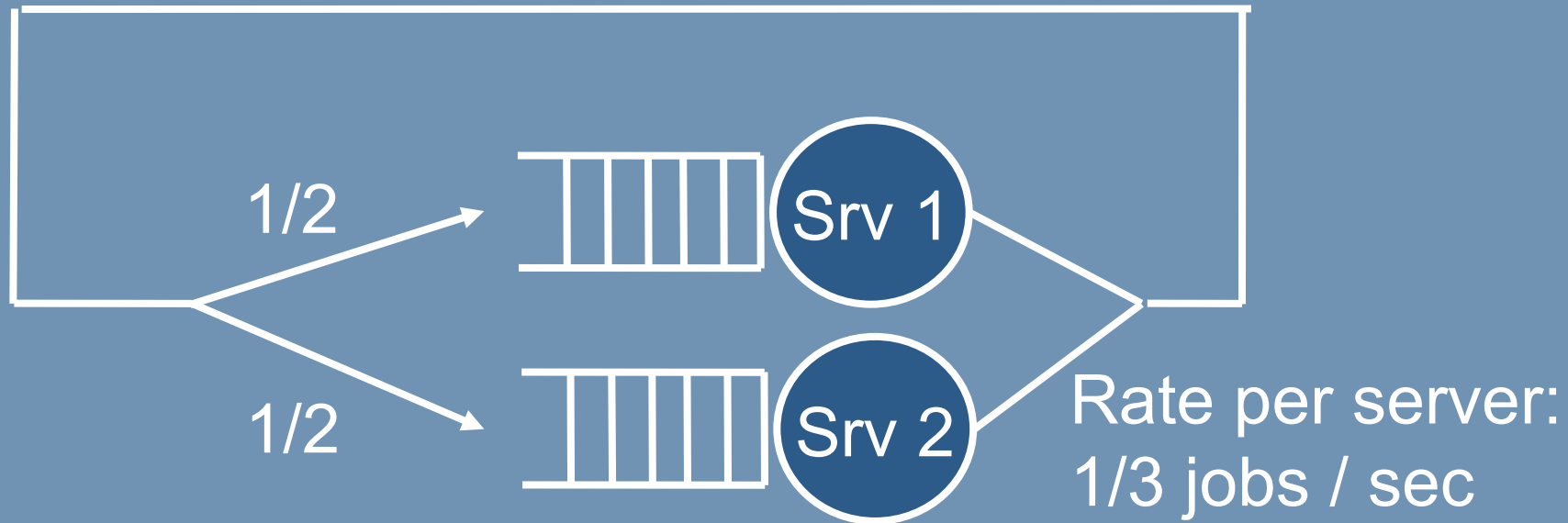
EXAMPLE 1



- Question: Assume the arrival rate doubles. By how much do you have to increase CPU speed to keep mean response times the same?
- Does the answer change with PS scheduling instead of FCFS?

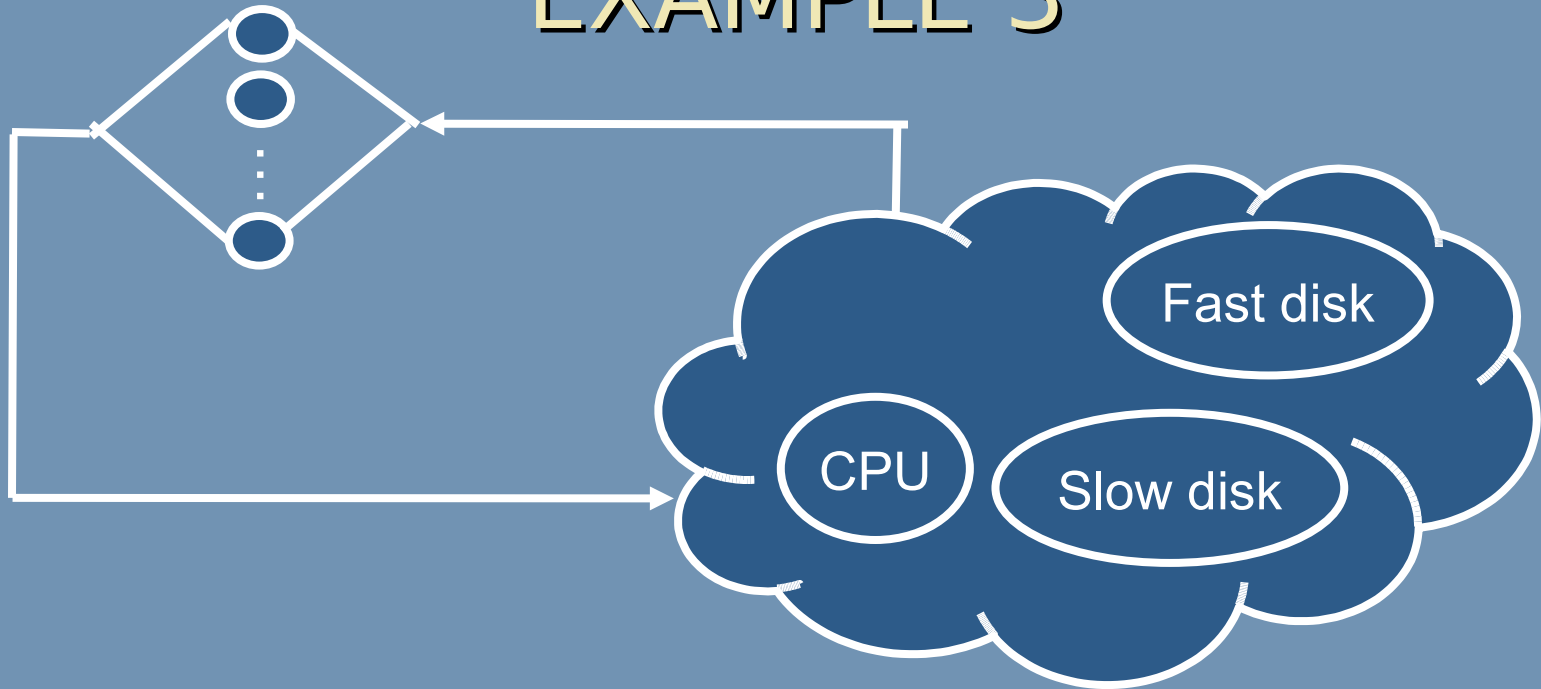
EXAMPLE 2

$N = 6$



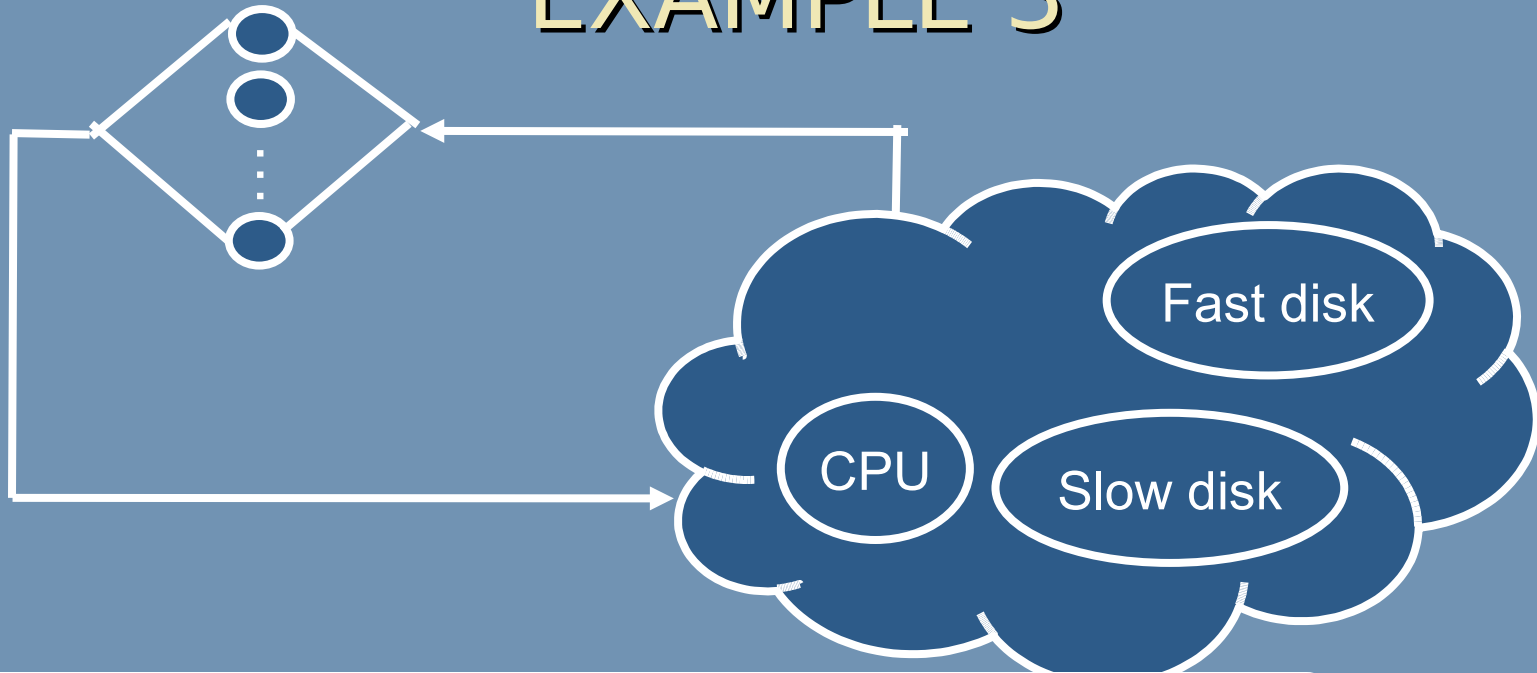
- Question: Server 1 is replaced with a server twice as fast.
- Does this change affect mean response time?
- Does this change affect throughput?

EXAMPLE 3



- Question: Which of the following changes is most effective in increasing throughput?
 - Replace CPU with one twice as fast?
 - Balance load between fast & slow disk?
 - Buy second fast disk?

EXAMPLE 3

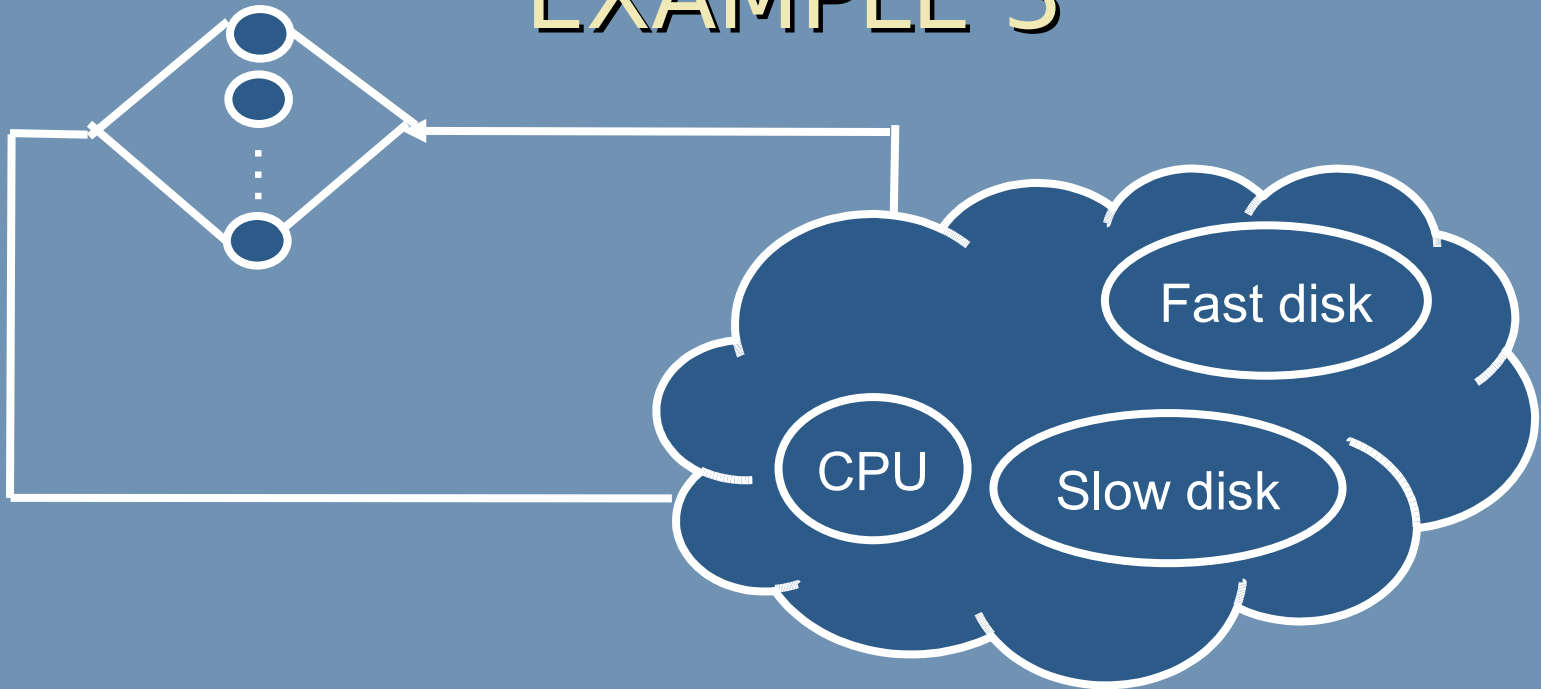


Measurements:

- time each device is busy
- number of completions per device
- total number completions
- mean “think time” between sending jobs

most effective

EXAMPLE 3



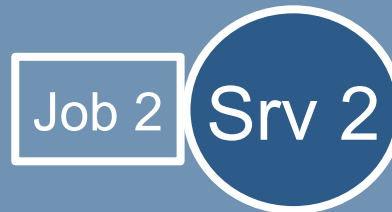
- We can solve this problem with simple back-of-the-envelope bound analysis
 - No math is necessary!
 - No assumptions on distributions
 - No knowledge of full network topology necessary

EXAMPLE 4

\$\$\$\$
New job



Has been running for long time



Has been running for short time

- Question: Which job is closer to completion, job 1 or job 2?

EXAMPLE 5

\$\$\$\$

New job

Srv 1

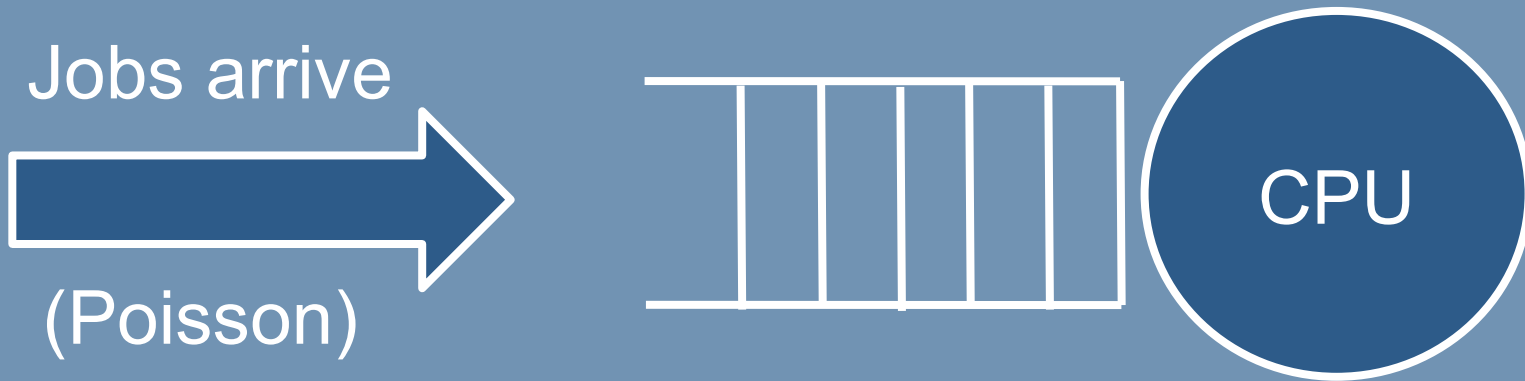
Has been running non-stop for 1 year

Srv 2

Has just crashed and been fixed yesterday

- Question: Which server will fail first?

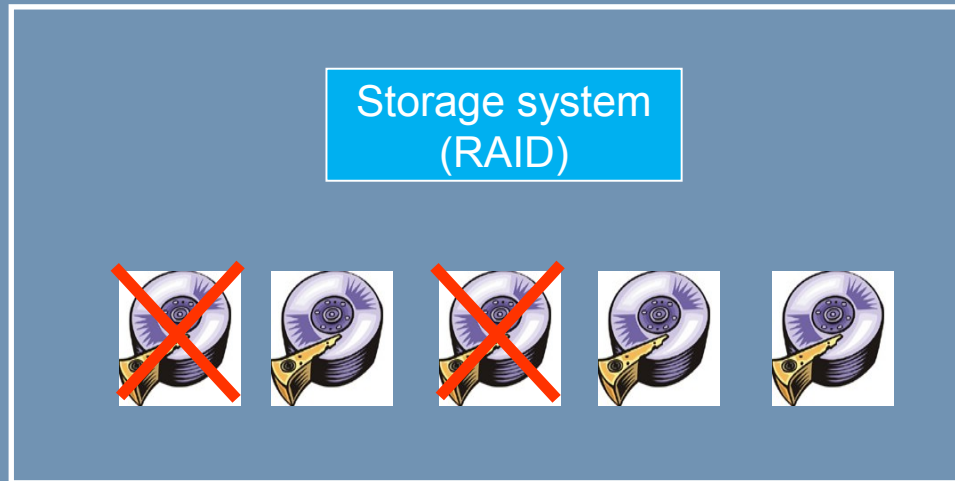
EXAMPLE 6



- Question: Which non-preemptive service order will result in lowest mean response time:
 - FCFS
 - LCFS = Last-Come-First-Serve
 - Random

SOME EXAMPLES FROM MY OWN RESEARCH

Probability of losing data in a RAID?



- Depends on probability that after one drive fails, a second drive fails while reconstructing data.

Estimating probability of data loss

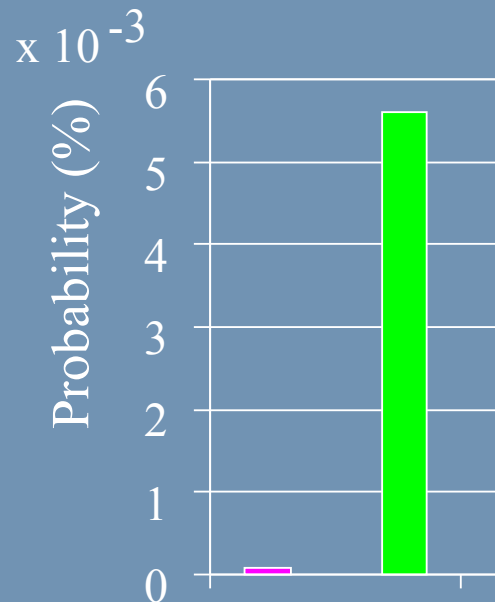
- Need probability of second failure during reconstruction
 - Standard approach: Use datasheet MTTF and exponential distr.



1 hour reconstruction time

Estimating probability of data loss

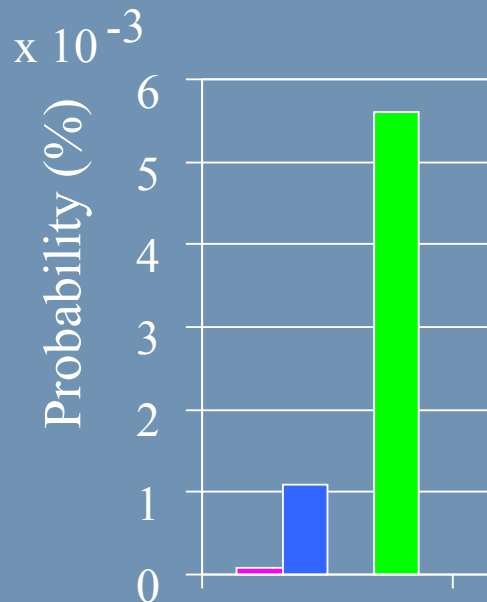
- Need probability of second failure during reconstruction
 - Standard approach: Use datasheet MTTF and exponential distr.
 - Estimate based on data



1 hour reconstruction time

Estimating probability of data loss

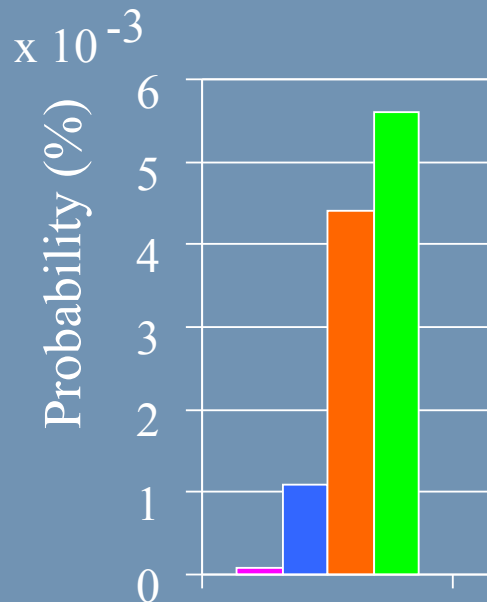
- Need probability of second failure during reconstruction
 - Standard approach: Use datasheet MTTF and exponential distr.
 - Use measured MTTF and exponential distribution
 - Estimate based on data



1 hour reconstruction time

Estimating probability of data loss

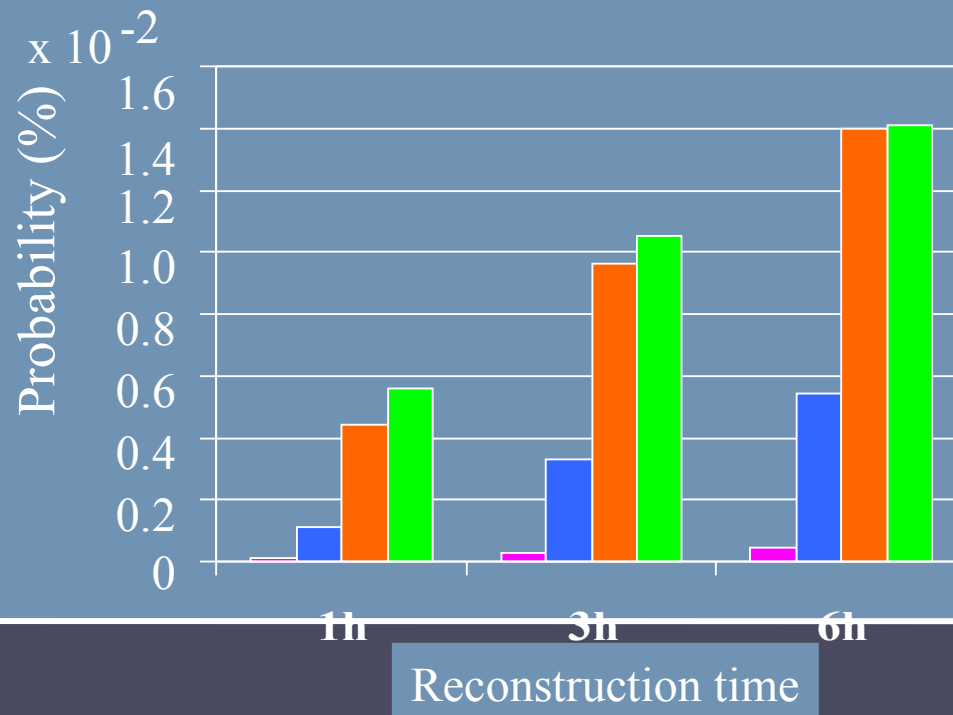
- Need probability of second failure during reconstruction
 - Standard approach: Use datasheet MTTF and exponential distr.
 - Use measured MTTF and exponential distribution
 - Use measured MTTF and Weibull distribution
 - Estimate based on data



1 hour reconstruction time

Estimating probability of data loss

- Need probability of second failure during reconstruction
 - Standard approach: Use datasheet MTTF and exponential distr.
 - Use measured MTTF and exponential distribution
 - Use measured MTTF and Weibull distribution
 - Estimate based on data

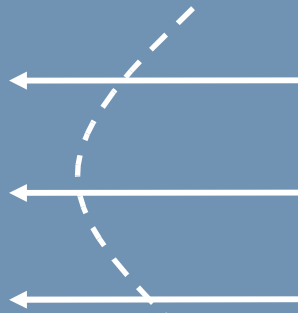


EXAMPLE 2

SCHEDULING STATIC WEB REQUESTS



PS (timesharing)



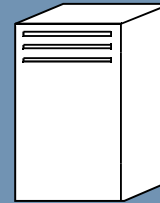
Socket 1



Socket 2



Socket 3



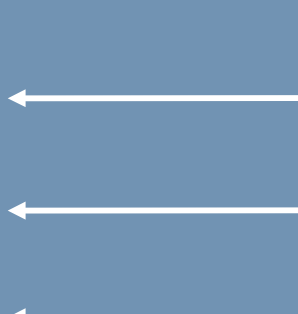
Standard web server



Size-based scheduling for better response times.



SRPT (shortest-remaining-time)



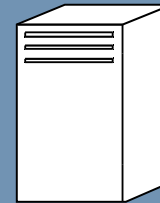
Socket 1



Socket 2



Socket 3



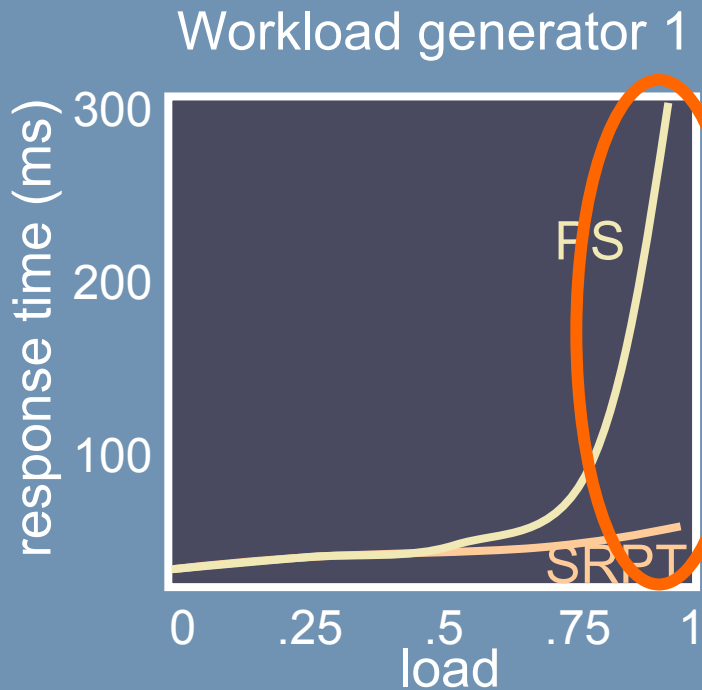
SRPT web server (kernel-level Implementation)

STATIC WEB WORKLOAD APACHE/LINUX

WHY?

- Mean file size
- File size distribution
- Access pattern
- Request rate
- CPU utilization
- Bandwidth
- Network effects

ALL THE SAME!

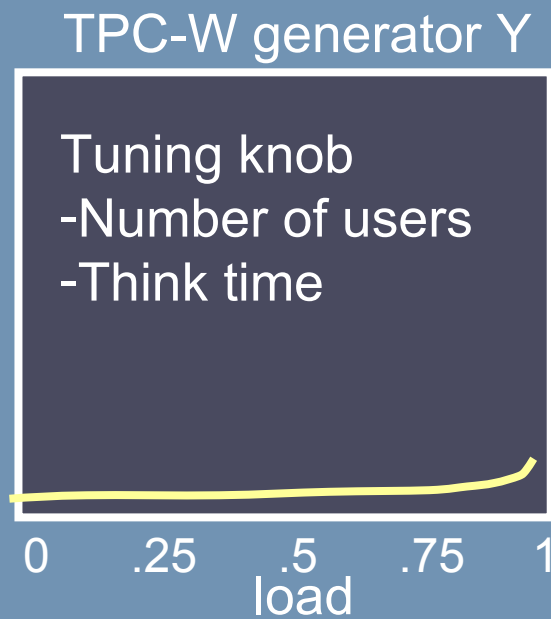
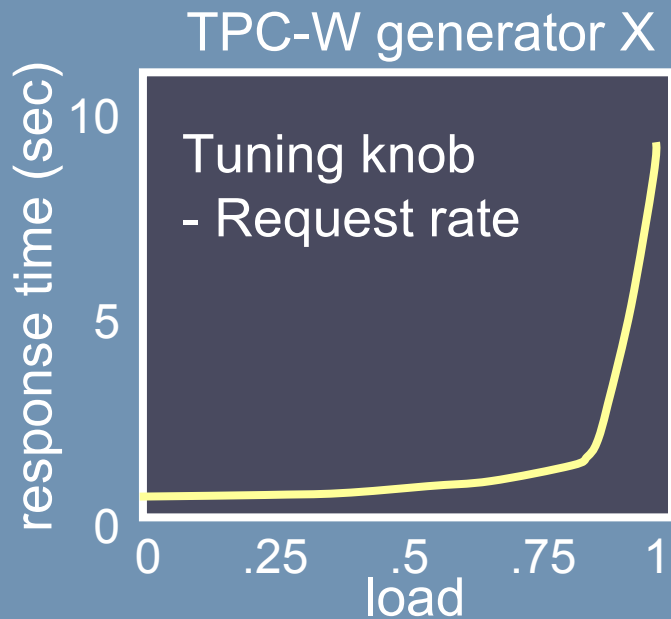
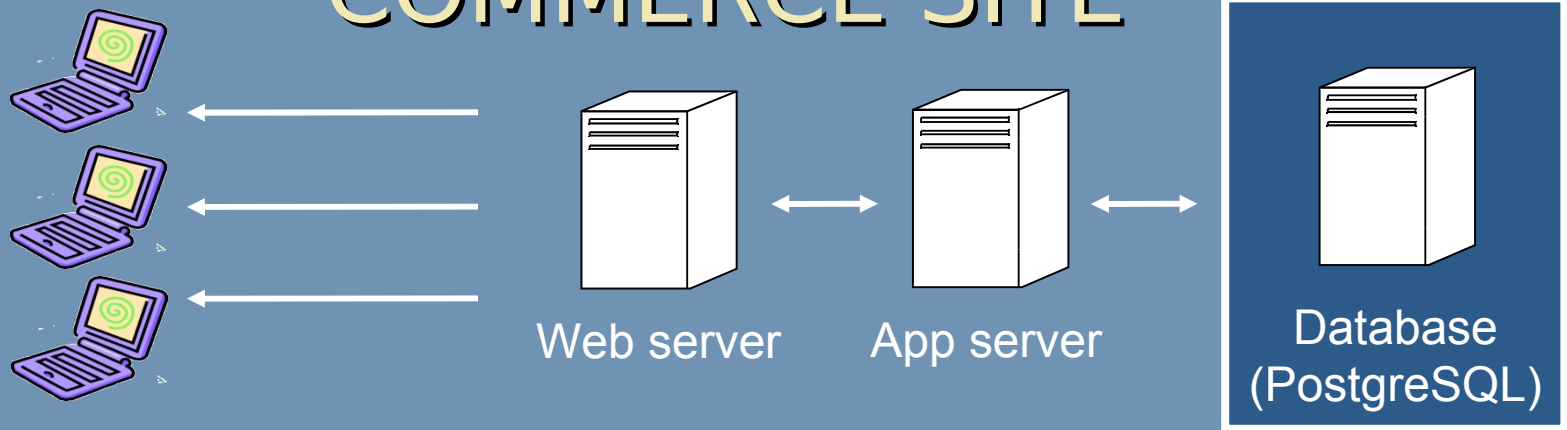


Tuning knob
- Request rate



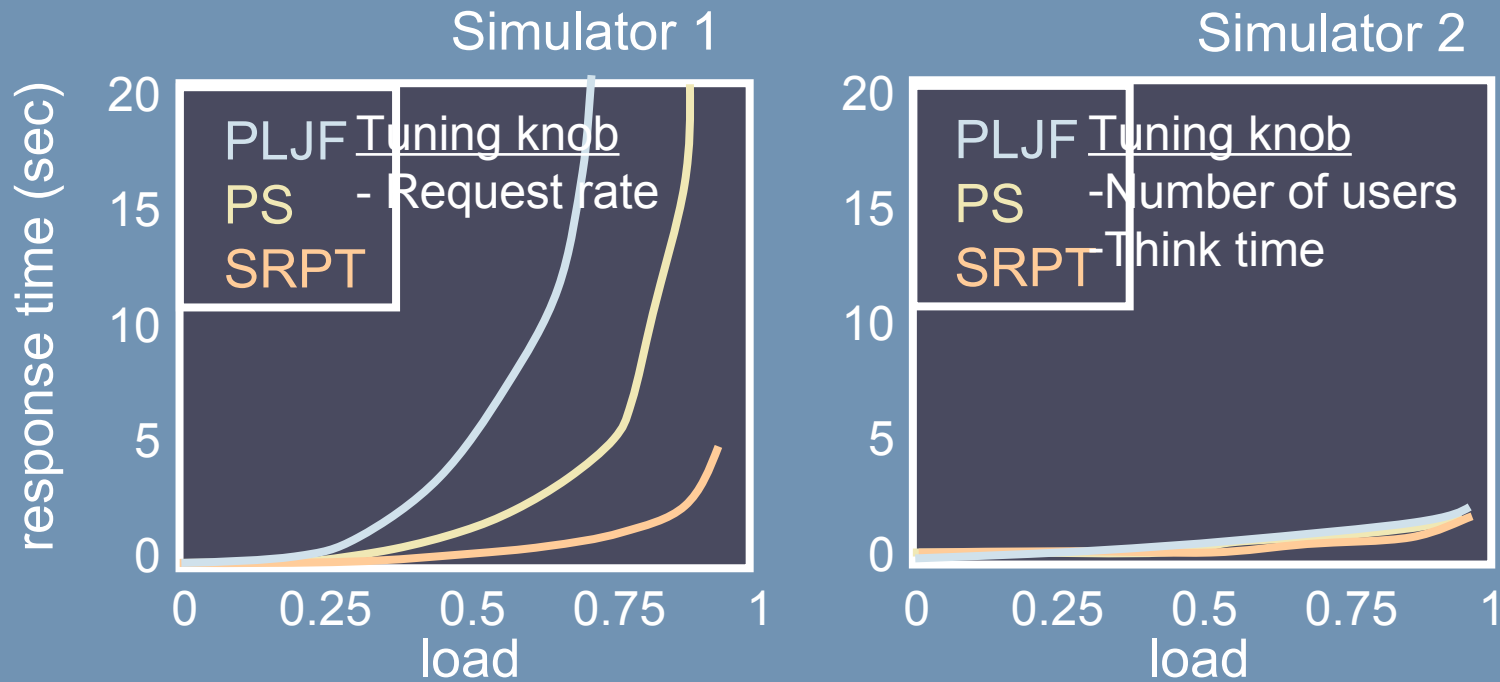
Tuning knob
- Number of users
- Think time

DATABASE BACKEND OF E-COMMERCE SITE

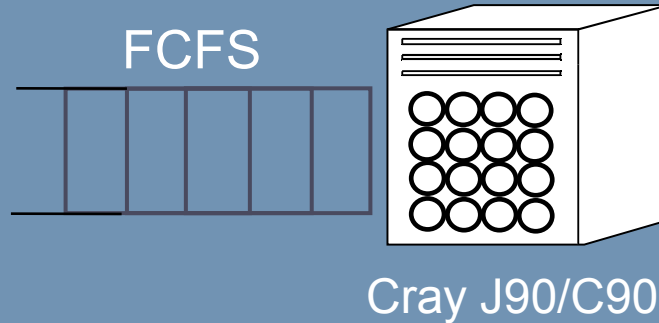


ONLINE AUCTION SITE - SIMULATION

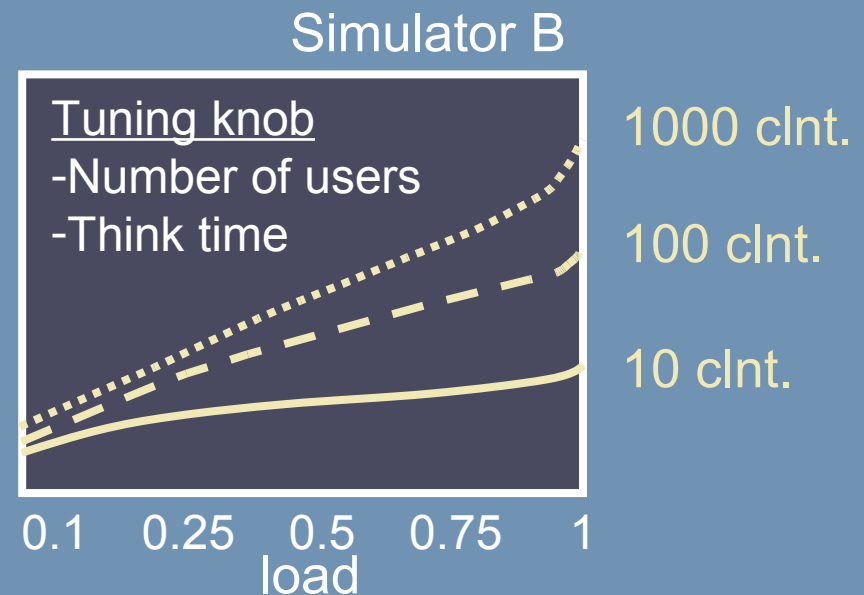
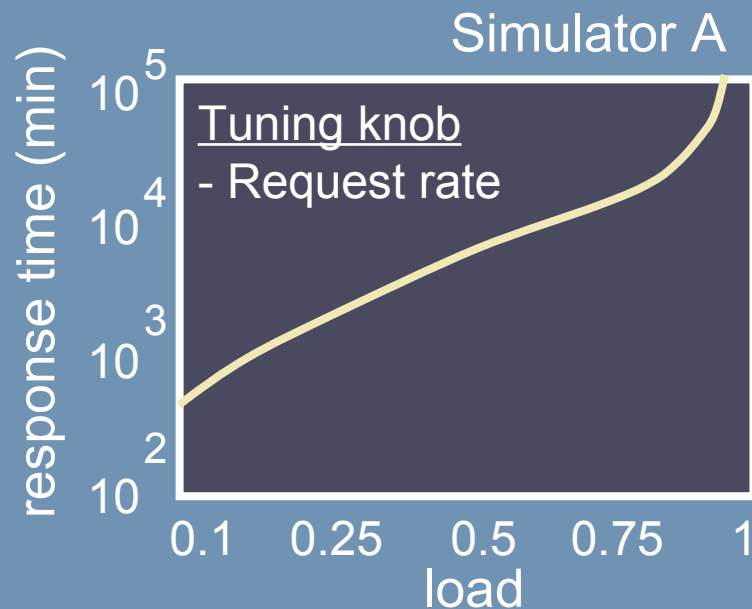
- Based on trace from top-10 online auctioning site.



BATCH JOBS AT A SUPERCOMPUTING SITE



- Simulation based on trace from Pittsburgh Supercomputing Center.



OPEN

CLOSED

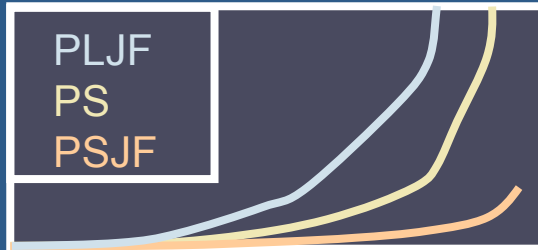
STATIC
WEB



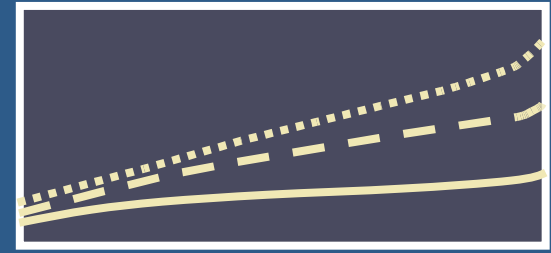
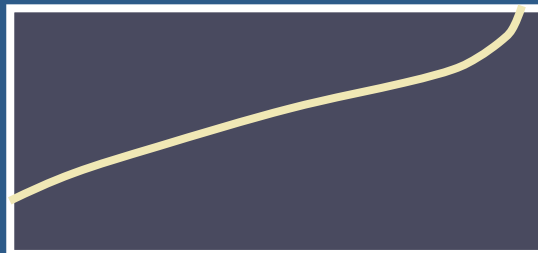
DATABASE
BACKEND



ONLINE
AUCTION



SUPER-
COMPUTING



Tuning knob
- Request rate

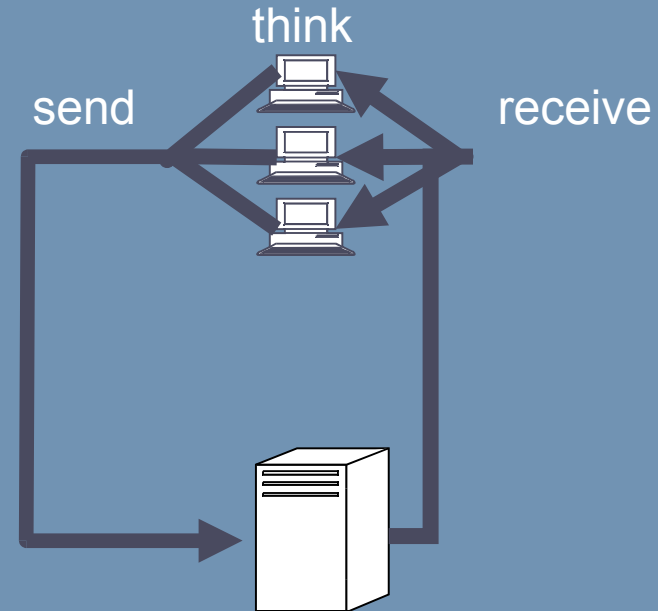
Tuning knob
- Number of users
- Think time

CLOSED SYSTEM MODEL

Model of user behavior

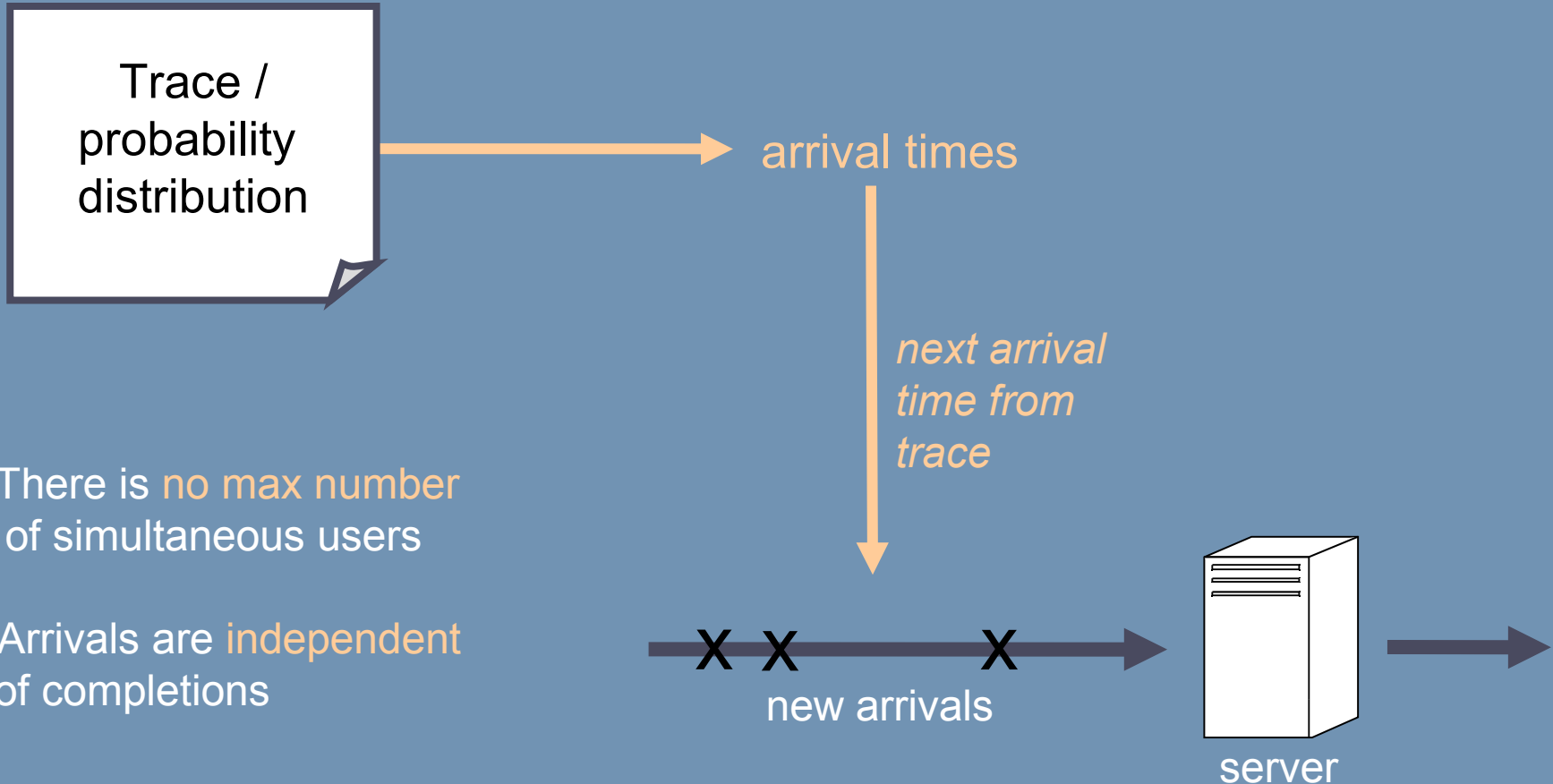


User requests web page, receives page, reads page, clicks on new link



- Fixed number of users, called the Multi-Programming-Level (MPL)
- Arrivals triggered by completions.

OPEN SYSTEM MODEL



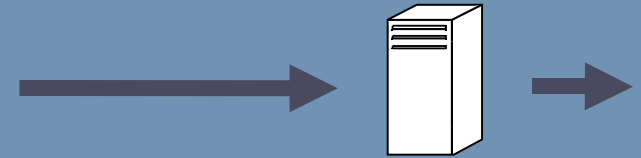
- There is **no max number** of simultaneous users
- Arrivals are **independent** of completions

WHICH MODEL DO WORKLOAD GENERATORS USE?

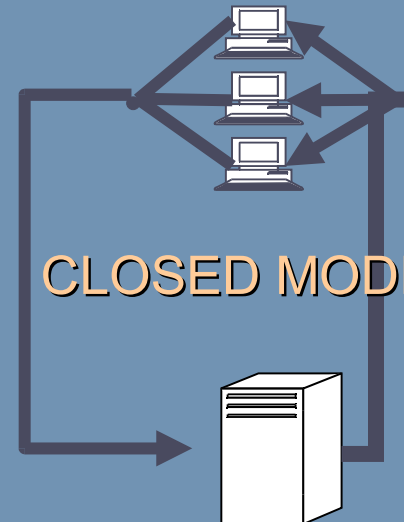
WEB WORKLOAD GENERATORS

- Client purpose
 - Surge
 - SPECWeb
 - TPC-W
- Client machine
 - Sclient
 - RUBiS
 - WebBench
 - Webjamma

OPEN MODEL



CLOSED MODEL



WHAT IS KNOWN IN THE LITERATURE?

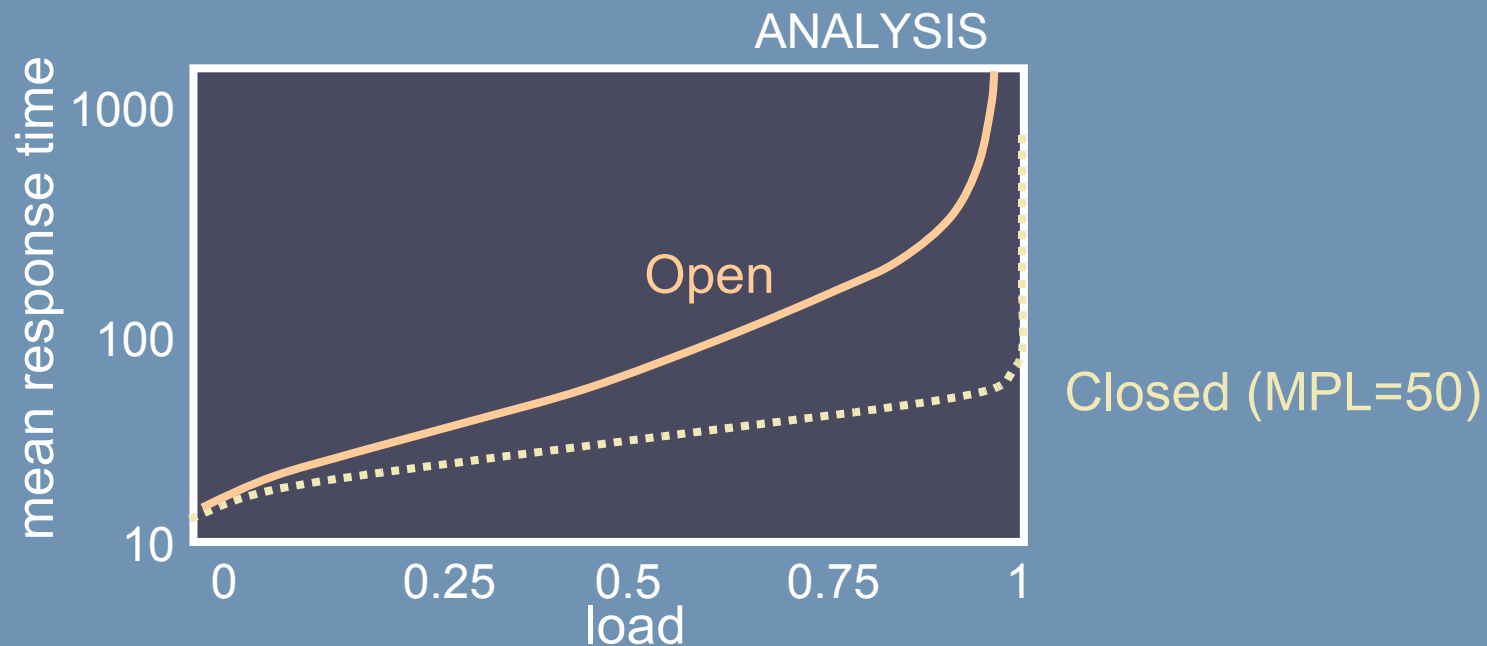
- Very little ...
- Limited to **FCFS single server queue**.
 - Response times under open system higher than under closed [Bondi and Whitt 1986].
 - For $MPL \rightarrow \infty$, closed system converges to open system [Schatte83, Schatte84].

STILL UNANSWERED:

- What is the **magnitude** in difference of response times?
- What is the **speed** of convergence?
- How does **variability** (heavy tails) affect results?
- How are different **scheduling** disciplines affected?
- in practice?

PRINCIPLES FOR OPEN VS. CLOSED

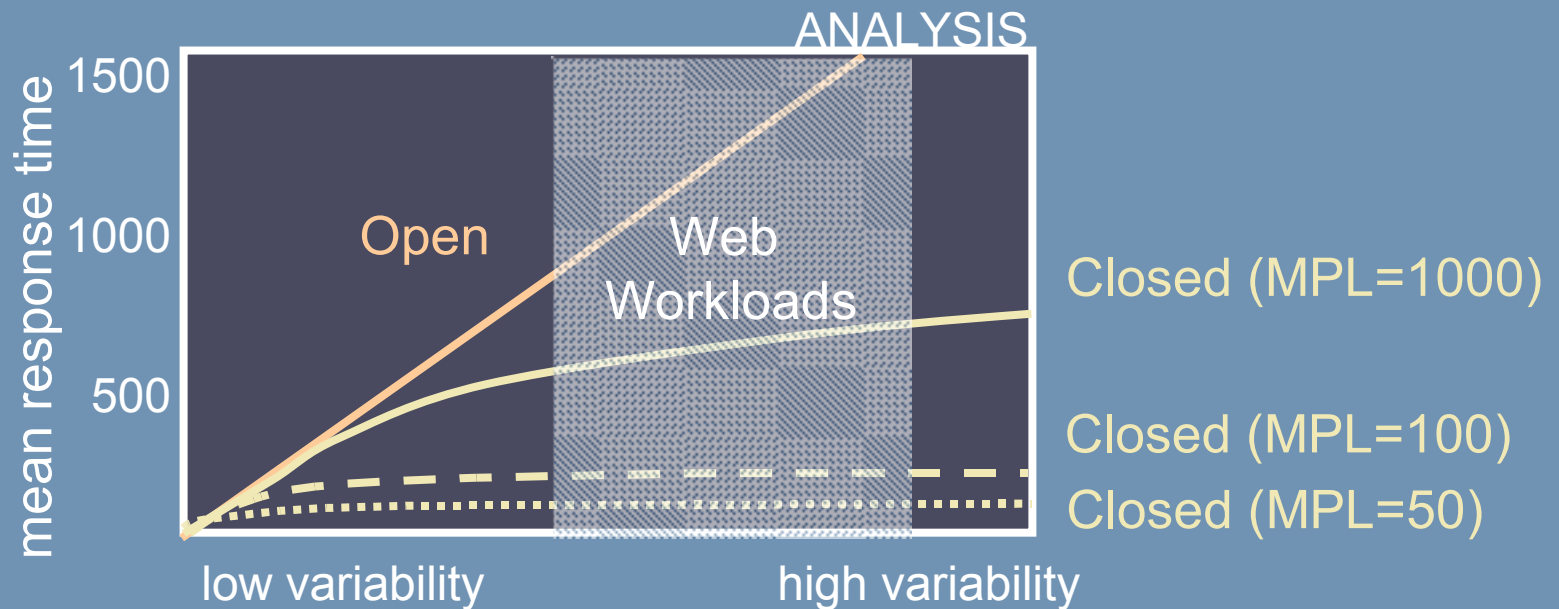
- What is the magnitude in difference of response times?
 - Orders of magnitude!



- Why?
 - *Bounded number of jobs in closed system.*

PRINCIPLES FOR OPEN VS. CLOSED

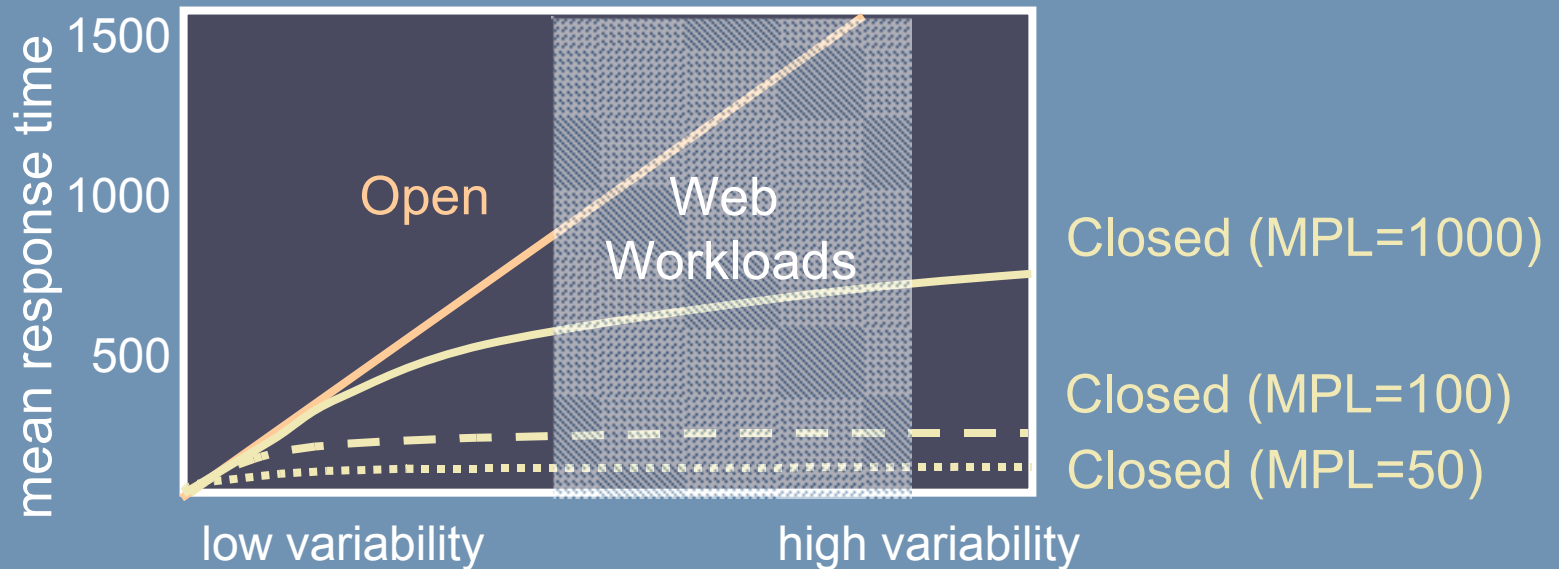
- How does variability affect open/closed response times?
 - Huge effect on open, limited effect on closed system.



- Why?
 - *Dependency between completions and arrivals in closed system reduces burstiness.*

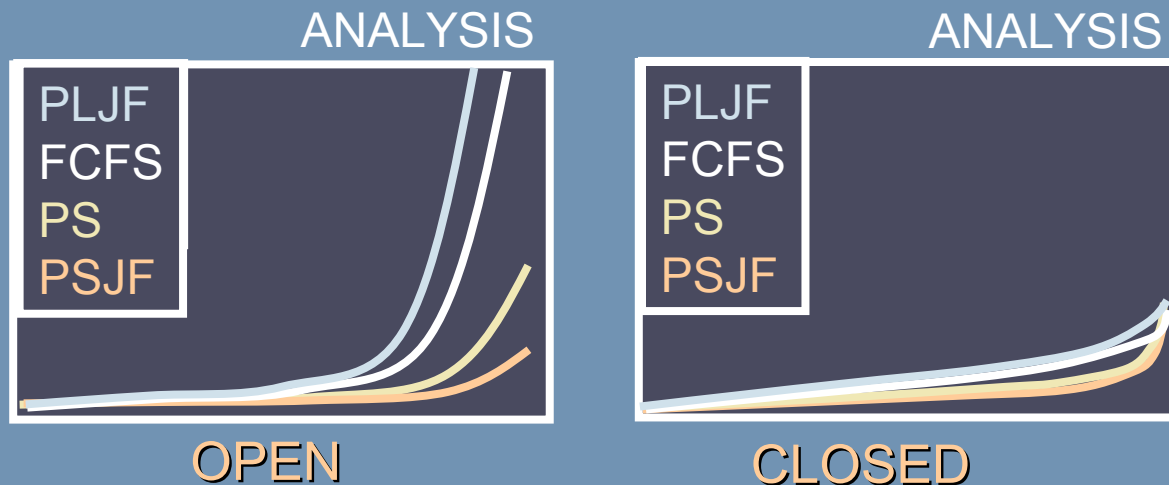
PRINCIPLES FOR OPEN VS. CLOSED

- Can we make closed look like open, by increasing MPL?



PRINCIPLES FOR OPEN VS. CLOSED

- What is the impact of scheduling?
 - Huge in open system, almost none in closed system.



- Why?
 - *Scheduling takes advantage of variability in the system.*
 - *Closed systems reduce the effect of variability.*