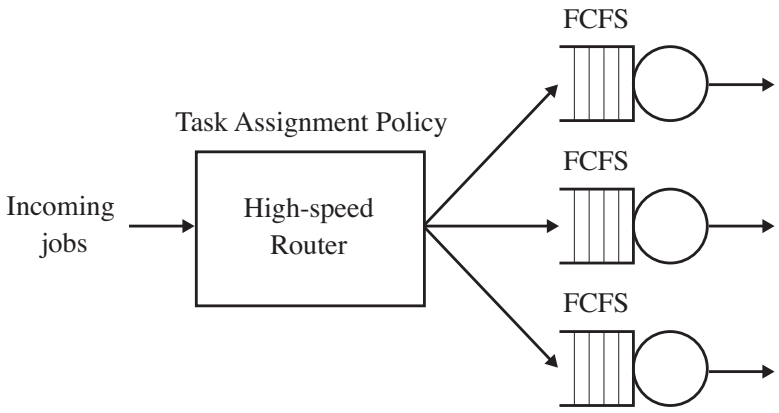
Server Farm Model



Question for lecture: What is a good policy for assigning jobs to hosts? (often called the task assignment policy)

ASSUMPTIONS:

1. A job is assigned to a single server, where it waits in a FCFS queue

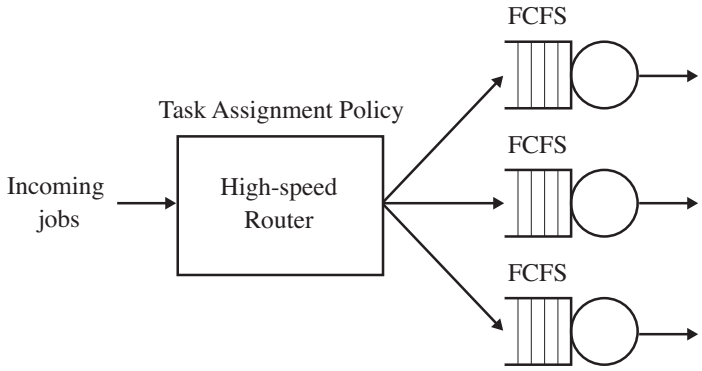
2. One job at a time runs on a server, and the job cannot be preempted when it runs.

3. Job size distribution, S, is highly variable.
   S represents service time on a single server.

4. Poisson arrival process with rate λ.

5. Goal: Minimize $\mathbf{E}[T]$.

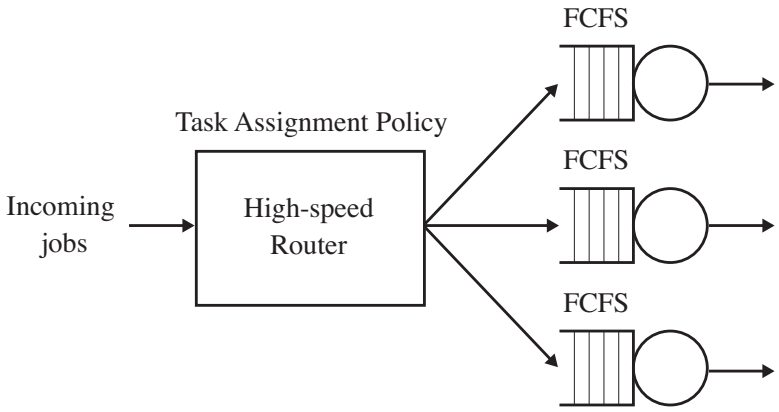
List as many task assignment policies as you can:



Policies which assume no knowledge of job size:

Policies which assume knowledge of job size:

Ranking policies in terms of $\mathbf{E}[T]$:



Question: Privately rank all policies in terms of $\mathbf{E}[T]$.

You're looking for a ranking of the form of

$$\mathbf{E}[T]^A \geq \mathbf{E}[T]^B \geq \mathbf{E}[T]^C = \mathbf{E}[T]^D \geq \mathbf{E}[T]^E \geq \mathbf{E}[T]^F, \text{ etc.}$$

Dynamic versus Static

DEFINITION: A **dynamic policy** adapts based on changes to the state of the system (e.g., state of the queues or servers). A **static policy** does not.

Question: Which assignment policies are dynamic?

Question: Which policies are static?

Random

Question: Assume k servers. Write an expression for $\mathbf{E}[T]^{Random}$.

Random versus Round-Robin

Question: What kind of queue do jobs experience under Random?

Question: What kind of queue do jobs experience under Round-Robin?

Question: Which of Random and Round-Robin has lower $\mathbf{E}[T]$?

JSQ

Question: Which of JSQ, Round-Robin, and Random are load balancing policies?

Question: What is the difference between what JSQ does and what Round-Robin and Random do?

Question: Which is better: JSQ or (central-queue) M/G/k? Why?

LWL versus (central-queue) M/G/k

Question: Which of LWL and (central-queue) M/G/k is dynamic?

Question: Which of LWL and (central-queue) M/G/k requires knowledge of job size?

Question: Which of LWL and (central-queue) M/G/k has lower $\mathbf{E}[T]$?

SITA – Size-Interval-Task-Assignment

Mor Harchol-Balter, Mark Crovella, and Cristina Murta, "On Choosing a Task Assignment Policy for a Distributed Server System," *IEEE Journal of Parallel and Distributed Computing (JPDC)*, vol. 59, no. 2, pp. 204-228, Nov 1999.

Question: Is SITA static or dynamic?

Question: How would we analyze $\mathbf{E}[T]^{SITA}$?

Question: Which of LWL and SITA has lower $\mathbf{E}[T]$?

Question: How do we pick the size cutoffs under SITA?

Final Ranking

Question: What's the final ranking of assignment policies?

Question: Suppose we want to do SITA, but we don't know job sizes?

Harchol-Balter, "Task Assignment with Unknown Duration," In *20th International Conference on Distributed Computing Systems (ICDCS '00)*, Taipei, Taiwan, April 2000, pp. 214-223.