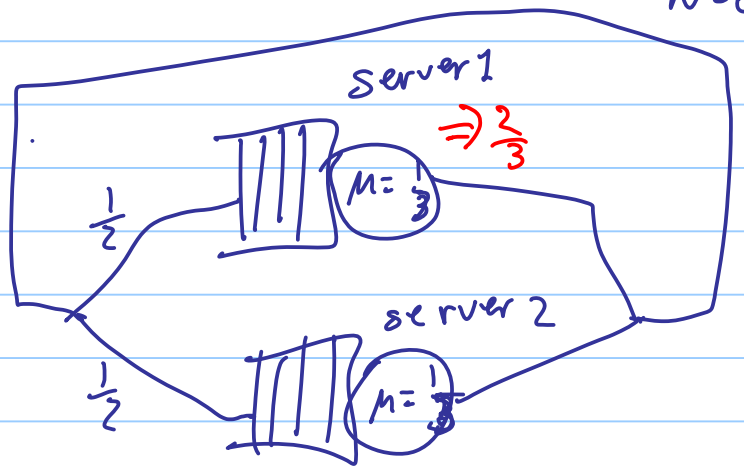
317_20160114

Note Title 2016-01-14

A closed system (Fig. 1.3 [H])



N=6 jobs
(the multiprogramming ratio)
(a batch system)

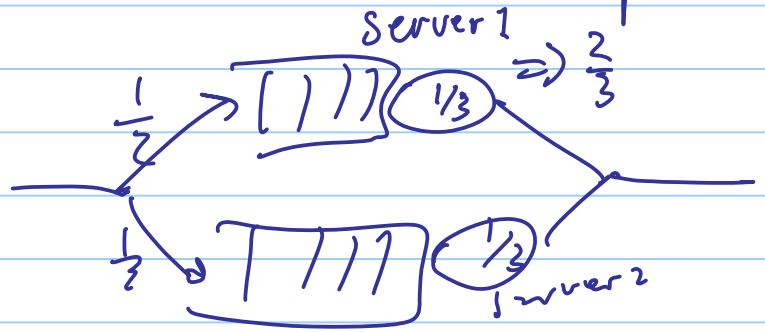
What is the effect of the change? We will see (in Ch. 6)

1. The effect is very small.

2. The effect goes to zero as $N \to \infty$

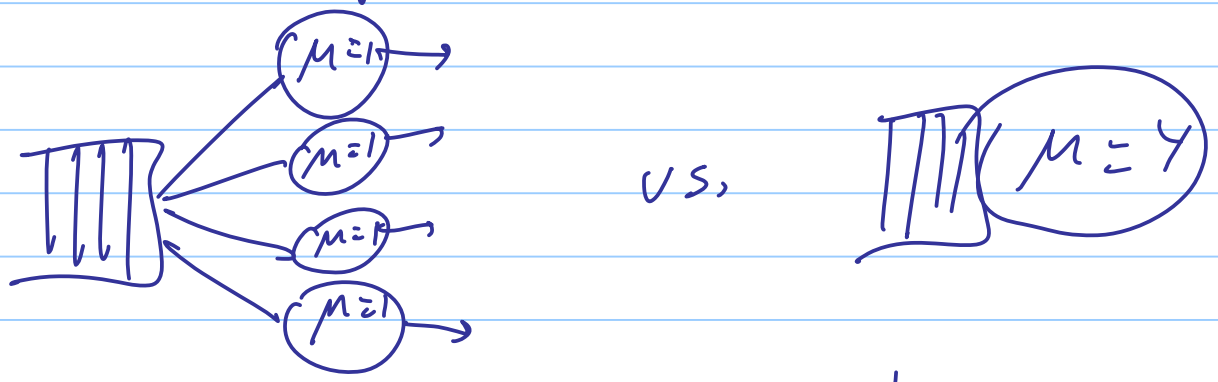
3. There is instead a noticeable improvement in the response time for very small values of N.

On the other hand, in an open system like this,



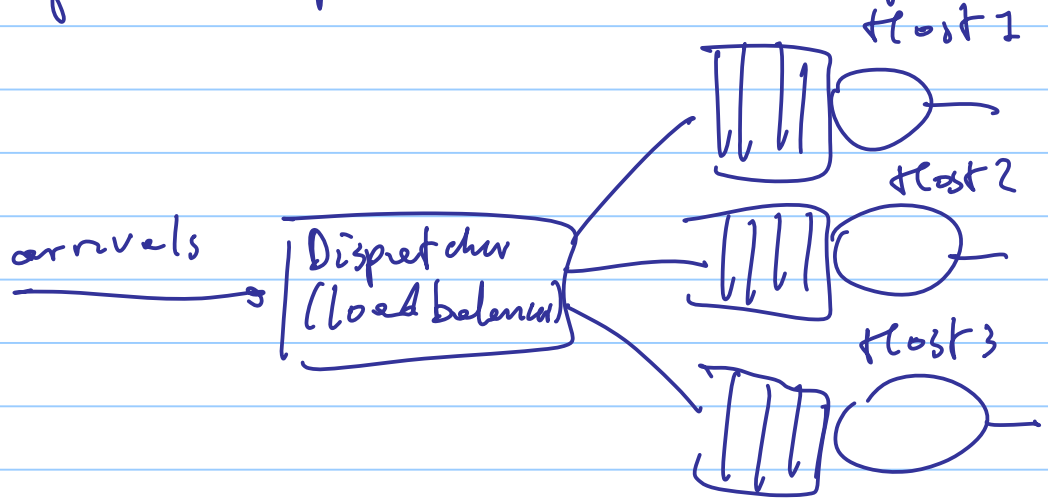
the improvement is substantial.

Design example 3 ("One Machine or Many?")



(Also: power allocation to many servers;
bandwidth allocation.)

Design Example 4 ("Task Assignment in a Server Farm")



Assignment policies:

Random (with a die that has as many sides as there are hosts)

Round robin: Job n goes to host n mod m, where there are m hosts numbered 0 to m-1.

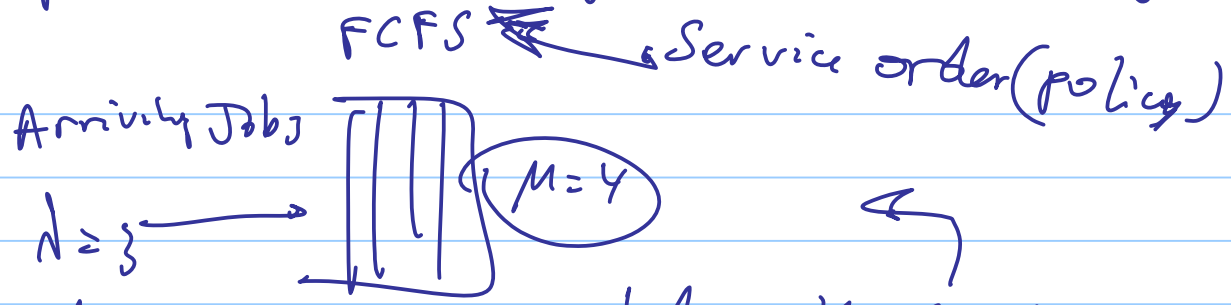
Shortest queue

Size-interval-task-assignment (SITA)

Least-Work-Left (LWL)

Central Queue

Chapter 2 ("Queueing Theory Terminology")



Several parameters associated with this (single server network)

Average Arrival Rate, λ, in jobs/sec

Mean Interarrival Time, $\frac{1}{\lambda}$, avg. time (in seconds) between successive job arrivals

Service Requirement Size, S, in seconds; the time needed to service a job in the absence of queueing.

Mean Service Time, $E[S]$, the expected value of S

Average Service Rate, μ, $\frac{1}{E[S]}$, in jobs/sec

Note: the parameters above are abstractions of time ones that used more commonly (eg., cycles per job, cycles per second)

Some performance metrics:

T, Response Time or Turnaround Time, Time in system, Sojourn Time

$T = t_{departure} - t_{arrival}$ (for a job)

$$E[T] = E[T_Q] + E[S]$$

S ← time in server

T_Q → time in queue, waiting time, delay

$Var[T]$ (variance of response time)

$P\{T > t\}$ tail behavior of the response time

(the probability that the response time is greater than t)

N, number of jobs in the system (includes jobs in queue, possibly plus one)

N_Q, number of jobs in the queue.

We assume $\lambda < \mu$.

Imagine what happens when $\lambda > \mu$. Queue size $\to \infty$

Intuition

$$E[N(t)] = E[A(t)] - E[D(t)] \geq \lambda t - \mu t =$$

$$= (\lambda - \mu) t$$

Now, assume $\lambda < \mu$ (actually $\lambda \leq \mu$ necessary for stability)