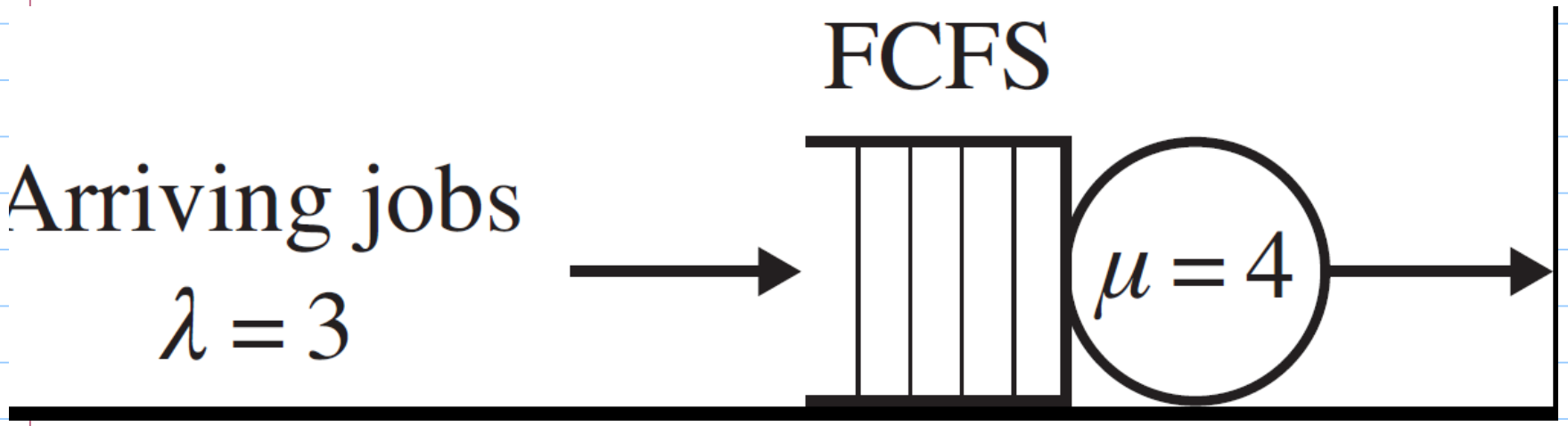
317 2015-01-15

Note Title 2015-01-15




Single-server network.

Parameters associated with the above

- Service order
  By default, FCFS
  (First Come First Served)
- Average Arrival Rate, $\lambda$, in jobs/sec
- Mean Interarrival Time, $\frac{1}{\lambda}$, avg time between successive job arrivals
- Service Requirement Size, $S$, in seconds.
  The time needed to process a job in the absence of queueing
- Mean Service Time, the expected value of $S$, $E[S]$
- Average Service Rate, $\mu = \frac{1}{E[S]}$, in jobs/sec.

# Some performance metrics

$T$, Response Time or Turnaround Time or Time in System or Sojourn Time

$T = t_{depart} - T_{arrive}$

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

$T_Q$: time in queue or delay or waiting time

$Var[T]$ (variance of Response Time)

$P\{T > t\}$ tail behavior of $T$

Number of Jobs in the System, $N$
(includes the jobs in queue and the job being served, if any)

Number of Jobs in Queue, $N_Q$

---

We assume that $\lambda < \mu$
(average arrival rate is less than the average service rate)

Suppose that the arrival and service distributions are deterministic (both constant).

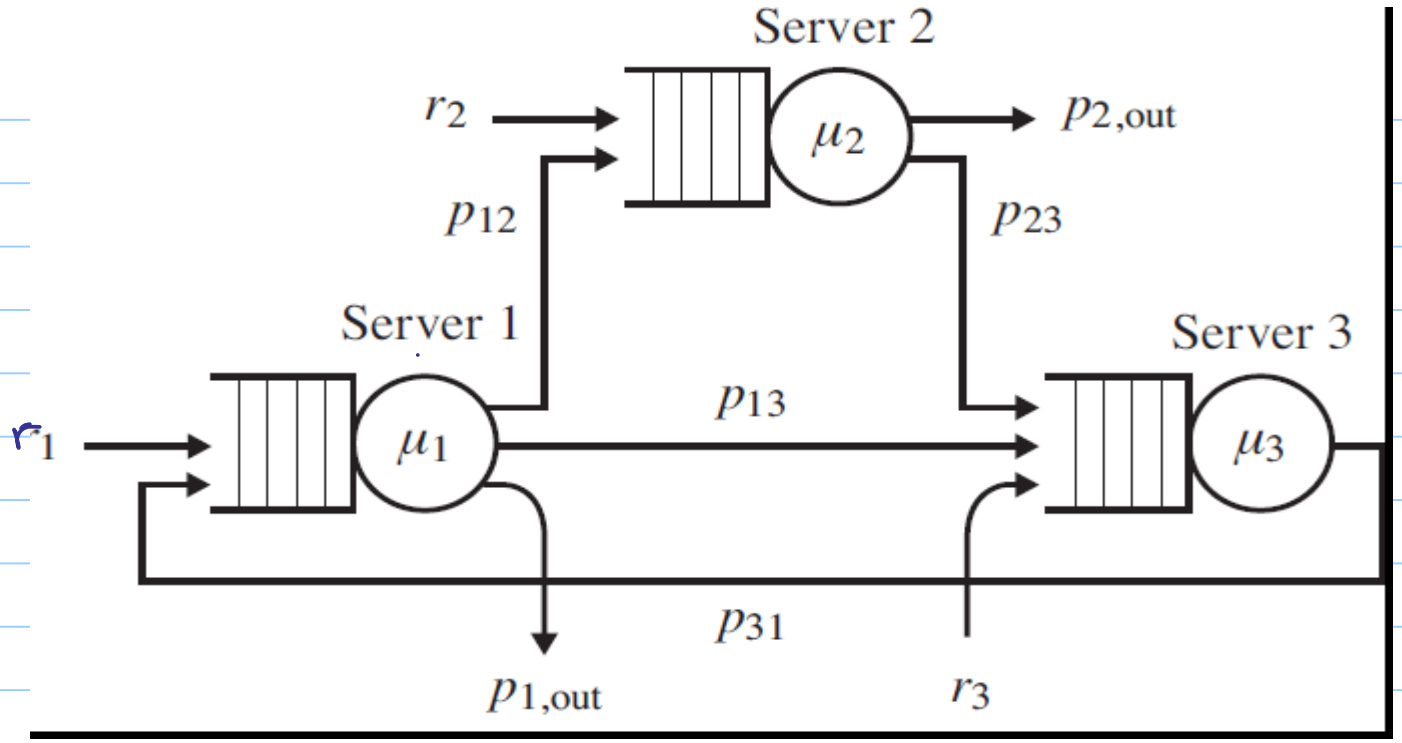
What is $T_Q$ ? What is $T$ ?

Answer: $T_Q = 0$ , $T = S$

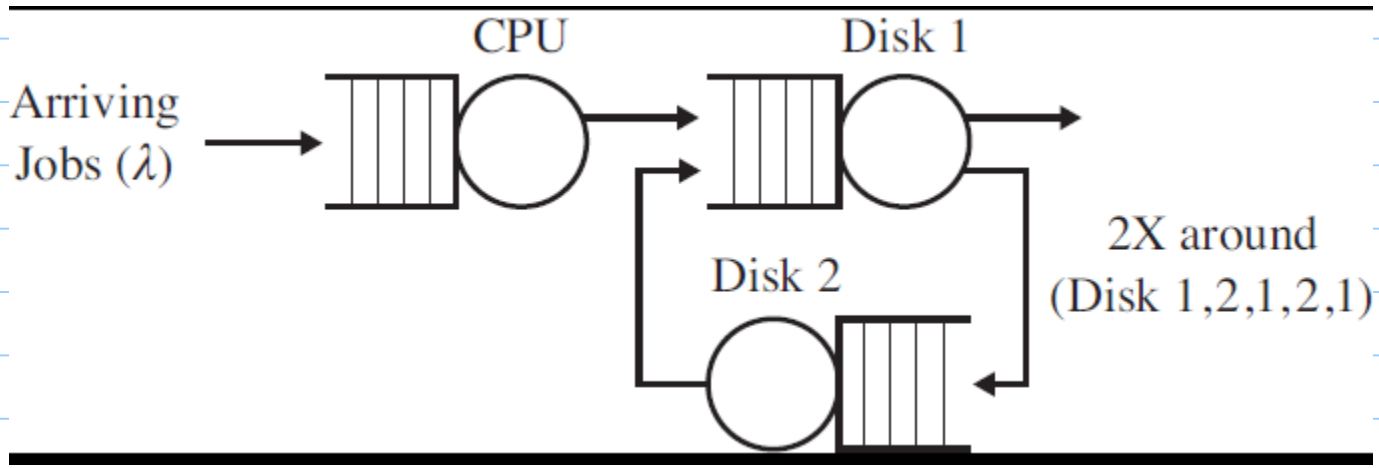
Open networks

1. Single server (see above)

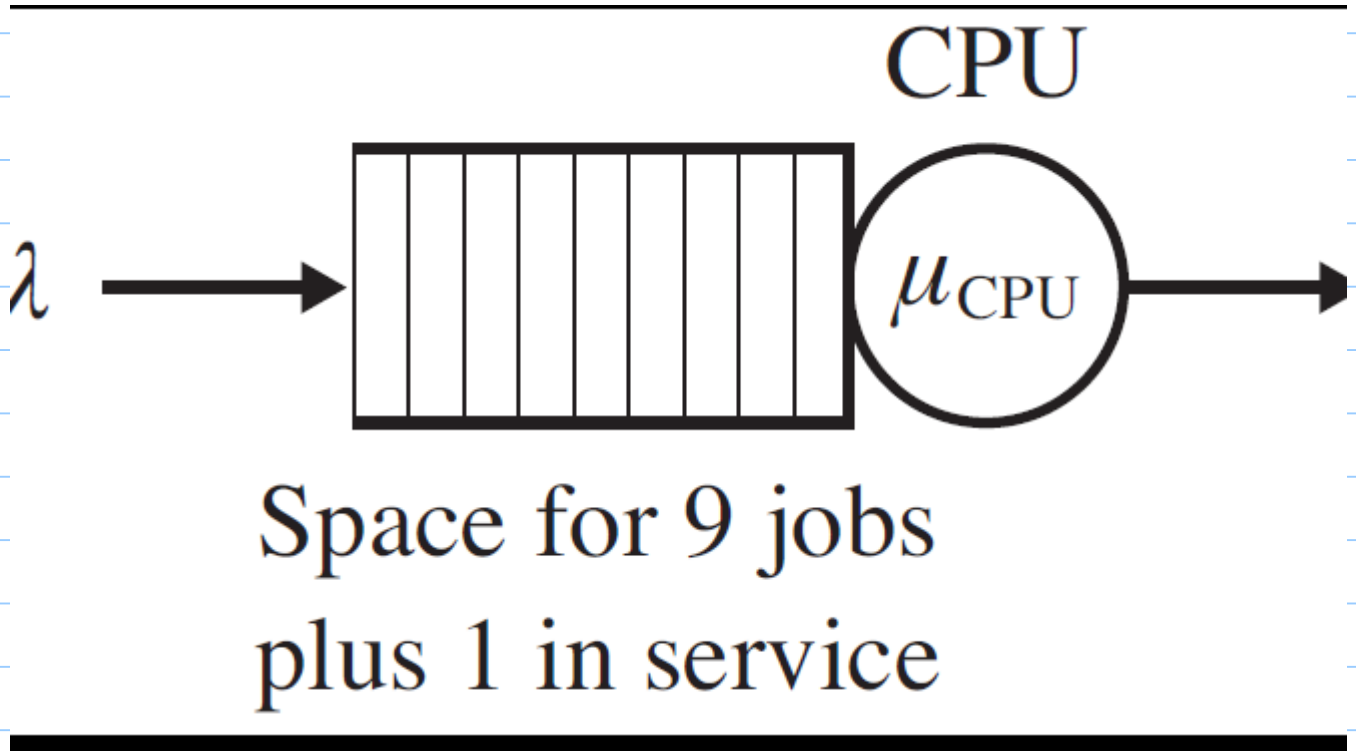
2. Network of Queues w/ Probabilistic Routing


Server 2
$r_2$
$\mu_2$
$p_{2,out}$
$p_{12}$
$p_{23}$
Server 1
Server 3
$p_{13}$
$r_1$
$\mu_1$
$\mu_3$
$p_{31}$
$p_{1,out}$
$r_3$

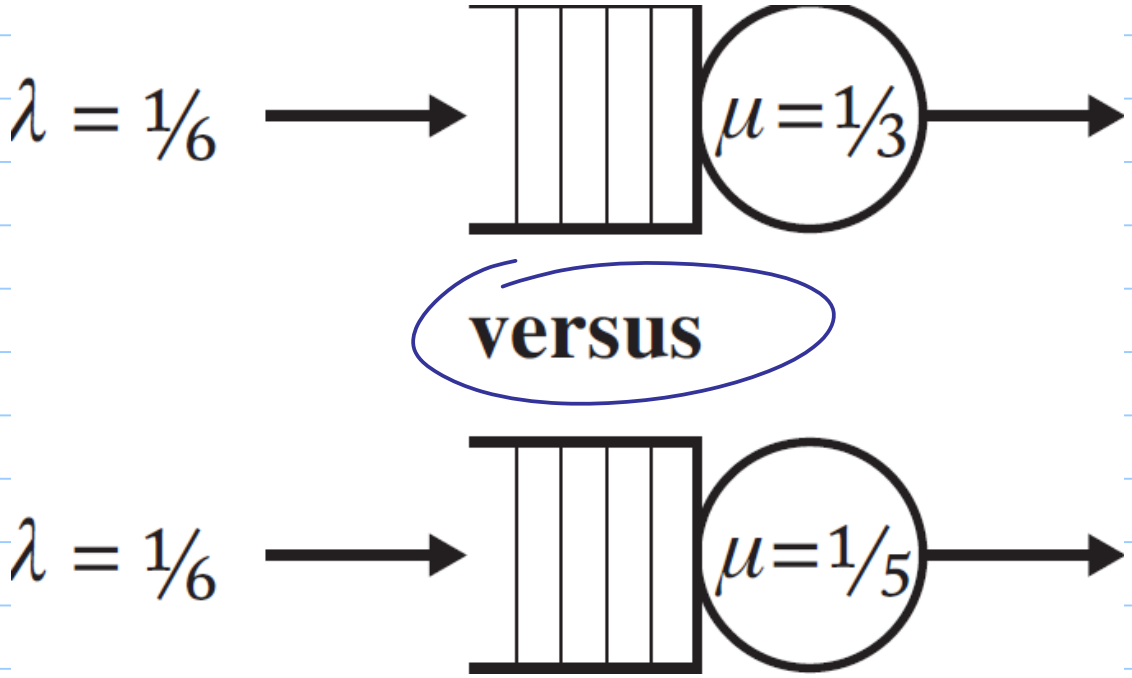
3. Network of Queues with non-Probabilistic Routing



4. Finite Buffer



# (2.5) More Metrics: Throughput & Utilization




Which of these systems has the higher throughput?

## Device

- **Utilization** $\rho_i$ is the fraction of time device $i$ is busy.

$$\rho_i = \frac{B}{\tau}$$

, where $\tau$ is a long time interval and $B$ is the time within $\tau$ that device $i$ is busy.

**Device Throughput** $X_i$ is the rate of completions at device $i$ (jobs/sec)

$$X_i = \frac{C}{\tau}$$

, where $\tau$ is a long time interval and $C$ is the # jobs completed during $\tau$.

$$X_i = \frac{C}{T} = \frac{C}{B} \cdot \frac{B}{T} = \mu_i \cdot \rho_i = \frac{1}{E[S_i]} \cdot \rho_i$$

$\frac{C}{B}$ = avg service rate $\mu_i$

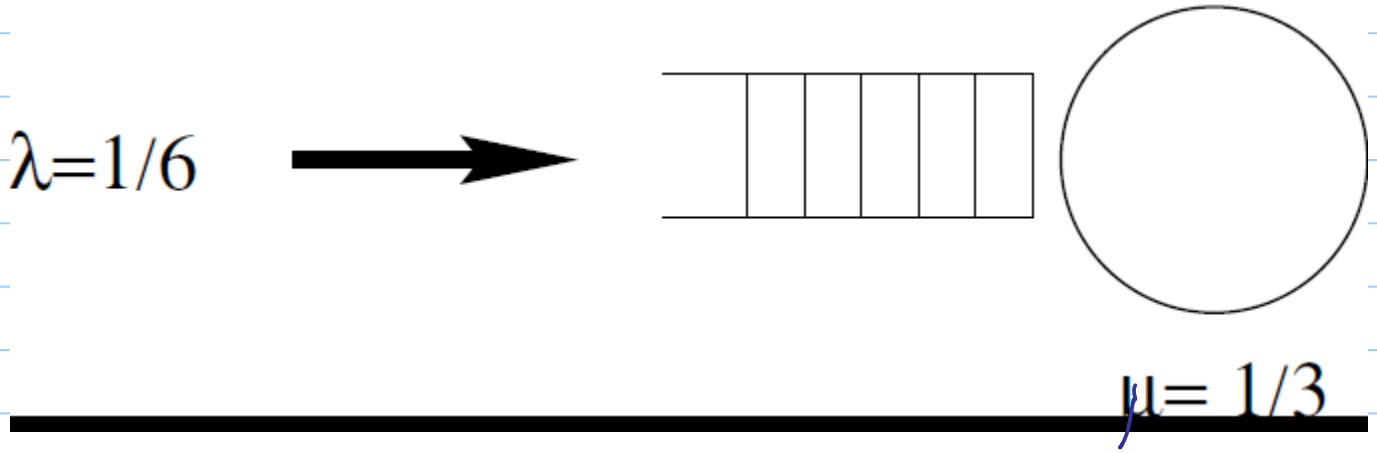
$\frac{B}{T}$ = $\rho_i$

$E[S_i]$ = mean service time

$$\rho_i = X_i \cdot E[S_i] \quad \text{(Utilization Law)}$$

$$\frac{\text{sec}}{\text{sec}} = \frac{\text{jobs}}{\text{sec}} \cdot \frac{\text{sec}}{\text{jobs}}$$

What is the throughput here ?




$X = \rho \cdot \mu$ . We have $\mu$. What is $\rho$ ?

$\rho$ = fraction of the server is busy :

$$= \frac{\text{average service time required by a job}}{\text{average time between arrivals}} =$$

$$= \frac{1/\mu}{1/\lambda} = \frac{\lambda}{\mu} = \frac{\text{average arrival rate}}{\text{average service rate}}$$

[Warning: not a formal proof.
Proof is given on p. 100 – Corollary 6.5]

Substitute $\rho = \frac{\lambda}{\mu}$ in $X = \rho \cdot \mu$, and obtain

$$X = \frac{\lambda}{\mu} \cdot \mu = \lambda \qquad (!!)$$

The throughput does not depend on the service rate!