Analysis of A Single Queue

Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu

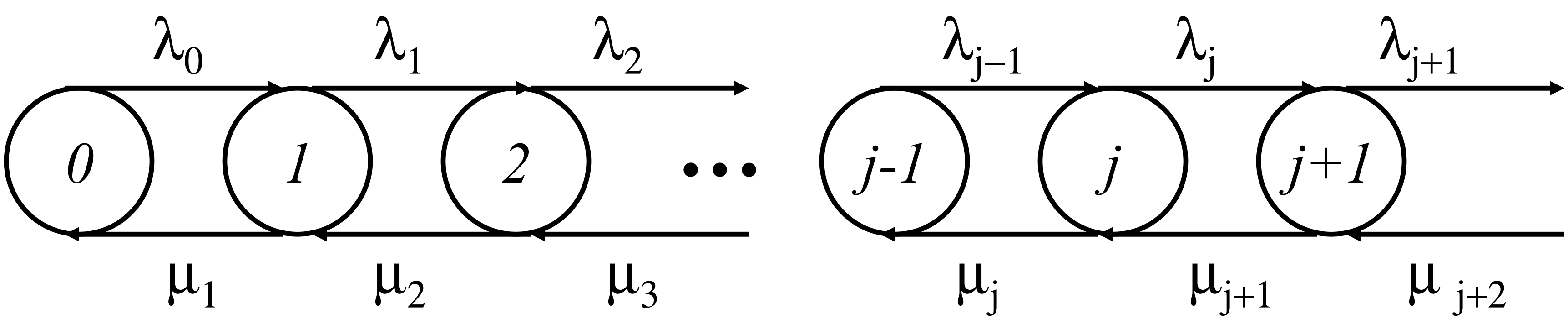
Audio/Video recordings of this lecture are available at:

http://www.cse.wustl.edu/~jain/cse567-13/

- Birth Death Processes
- M/M/1 Queue
- M/M/m Queue
- M/M/m/B Queue with Finite Buffers
- Results for other Queueing systems

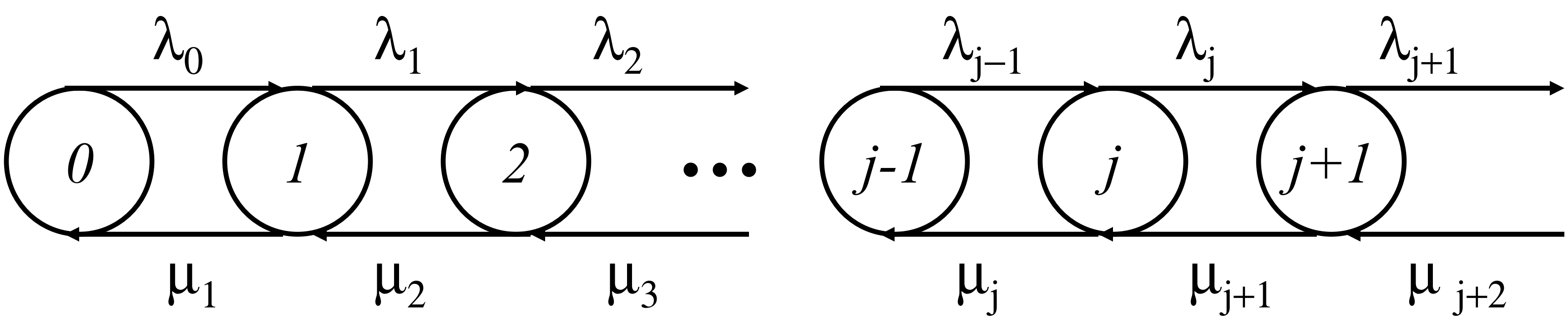
Birth-Death Processes

- Jobs arrive one at a time (and not as a batch).
- State = Number of jobs n in the system.
- Arrival of a new job changes the state to $n+1$ ⇒ birth
- Departure of a job changes the system state to $n-1$ ⇒ Death
- State-transition diagram:



Birth-Death Processes(Cont)

- When the system is in state n, it has n jobs in it.
  - The new arrivals take place at a rate λ_n.
  - The service rate is μ_n.
- We assume that both the inter-arrival times and service times are exponentially distributed.

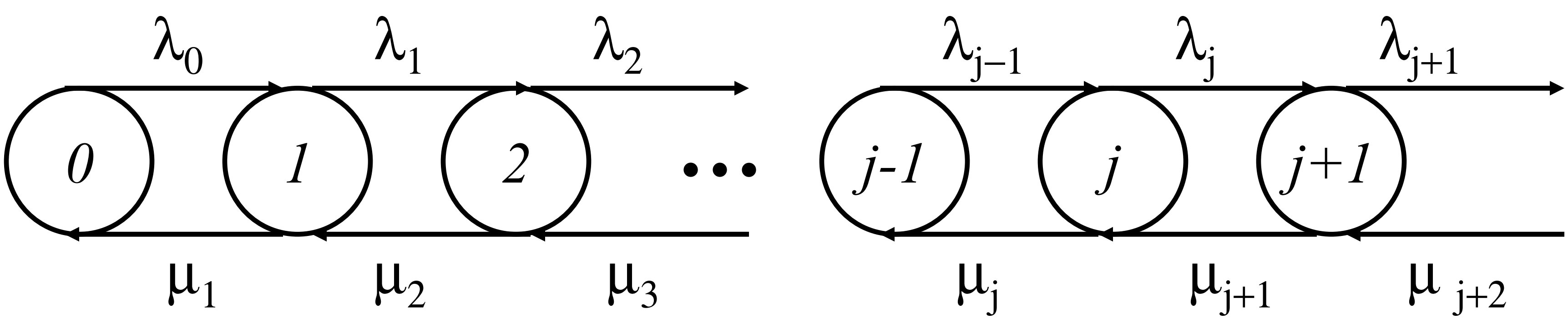


Theorem: State Probability

- The steady-state probability p_n of a birth-death process being in state n is given by:

$$p_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} p_0 \quad n = 1, 2, \ldots, \infty$$

- Here, p_0 is the probability of being in the zero state.



Proof

- Suppose the system is in state *j* at time *t*. There are *j* jobs in the system. In the next time interval of a very small duration Δt, the system can move to state *j-1* or *j+1* with the following probabilities:

$$\begin{aligned} P\{n(t+\Delta t) = j+1 | n(t) = j\} &= \text{Probability of one arrival in interval } \Delta t \\ &= \lambda_j \Delta t \end{aligned}$$

$$\begin{aligned} P\{n(t+\Delta t) = j-1 | n(t) = j\} &= \text{Probability of one departure in interval } \Delta t \\ &= \mu_j \Delta t \end{aligned}$$

$$P\{n(t+\Delta t) = j | n(t) = j\} = 1 - \lambda_j \Delta t - \mu_j \Delta t$$

Proof(Cont)

- If there are no arrivals or departures, the system will stay in state j and, thus:
- Δt = small $\Rightarrow$ zero probability of two events (two arrivals, two departure, or a arrival and a departure) occurring during this interval $p_j(t)$ = probability of being in state j at time t

$$\begin{aligned}
p_0(t+\Delta t) &= (1-\lambda_0\Delta t)p_0(t) + \mu_1\Delta t p_1(t) \\
p_1(t+\Delta t) &= \lambda_0\Delta t p_0(t) + (1-\mu_1\Delta t - \lambda_1\Delta t)p_1(t) + \mu_2\Delta t p_2(t) \\
p_2(t+\Delta t) &= \lambda_1\Delta t p_1(t) + (1-\mu_2\Delta t - \lambda_2\Delta t)p_2(t) + \mu_3\Delta t p_3(t) \\
& \dots \\
p_j(t+\Delta t) &= \lambda_{j-1}\Delta t p_{j-1}(t) + (1-\mu_j\Delta t - \lambda_j\Delta t)p_j(t) + \mu_{j+1}\Delta t p_{j+1}(t) \\
& \dots
\end{aligned}$$

Proof(Cont)

- The j^{th} equation above can be written as follows:

$$\begin{aligned}
\lim_{\Delta t \to 0} \frac{p_j(t+\Delta t) - p_j(t)}{\Delta t} &= \lambda_{j-1} p_{j-1}(t) - (\mu_j + \lambda_j) p_j(t) + \mu_{j+1} p_{j+1}(t) \\
\frac{dp_j(t)}{dt} &= \lambda_{j-1} p_{j-1}(t) - (\mu_j + \lambda_j) p_j(t) + \mu_{j+1} p_{j+1}(t)
\end{aligned}$$

$$\lim_{t \to \infty} p_j(t) = p_j$$

$$\lim_{t \to \infty} \frac{dp_j(t)}{dt} = 0$$

- Under steady state, $p_j(t)$ approaches a fixed value p_j, that is:

$$\begin{aligned}
\lim_{\Delta t \to 0} \frac{p_j(t+\Delta t) - p_j(t)}{\Delta t} &= \lambda_{j-1} p_{j-1}(t) - (\mu_j + \lambda_j) p_j(t) + \mu_{j+1} p_{j+1}(t) \\
\frac{dp_j(t)}{dt} &= \lambda_{j-1} p_{j-1}(t) - (\mu_j + \lambda_j) p_j(t) + \mu_{j+1} p_{j+1}(t)
\end{aligned}$$

Proof(Cont)

- Substituting these in the j^{th} equation, we get:

$$0 = \lambda_{j-1}p_{j-1} - (\mu_j + \lambda_j)p_j + \mu_{j+1}p_{j+1}$$

$$p_{j+1} = \left(\frac{\mu_j + \lambda_j}{\mu_{j+1}}\right)p_j - \frac{\lambda_{j-1}}{\mu_{j+1}}p_{j-1} \quad j = 1, 2, 3, \ldots$$

$$p_1 = \frac{\lambda_0}{\mu_1}p_0$$

- The solution to this set of equations is:

$$\begin{aligned} p_n &= \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}p_0 \\ &= p_0\prod_{j=0}^{n-1}\frac{\lambda_j}{\mu_{j+1}} \quad n = 1, 2, \ldots, \infty \end{aligned}$$

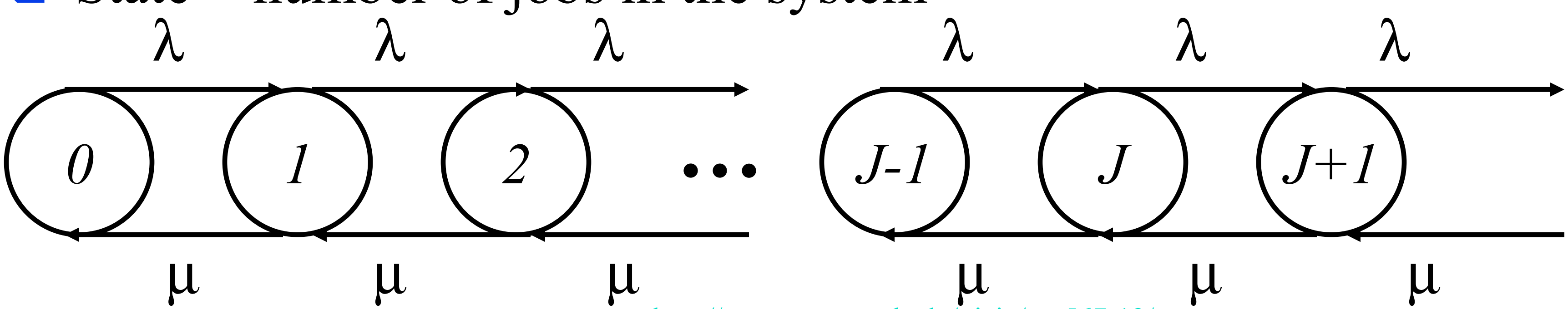
Proof(Cont)

- The sum of all probabilities must be equal to one:

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \prod_{j=0}^{n-1} \frac{\lambda_j}{\mu_{j+1}}}$$

M/M/1 Queue

- *M/M/1* queue is the most commonly used type of queue
- Used to model single processor systems or to model individual devices in a computer system
- Assumes that the interarrival times and the service times are exponentially distributed and there is only one server.
- No buffer or population size limitations and the service discipline is FCFS
- Need to know only the mean arrival rate λ and the mean service rate μ.
- State = number of jobs in the system



Results for M/M/1 Queue

- Birth-death processes with

$$\lambda_n = \lambda \quad n = 0, 1, 2, \ldots, \infty$$

$$\mu_n = \mu \quad n = 1, 2, \ldots, \infty$$

- Probability of n jobs in the system:

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 \quad n = 1, 2, \ldots, \infty$$

- The quantity λ/μ is called **traffic intensity** and is usually denoted by symbol ρ. Thus:

$$p_n = \rho^n p_0 \qquad n = 1, 2, \ldots, \infty$$

Results for M/M/1 Queue(Cont)

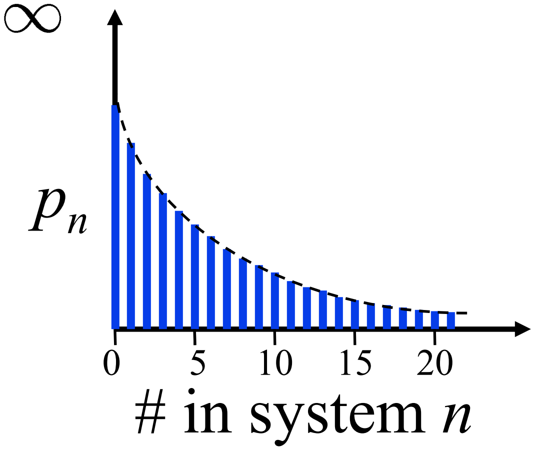
$$p_n = \rho^n p_0$$

$$p_0 + p_1 + p_2 + \cdots + p_n + \cdots \infty = 1$$

$$p_0 = \frac{1}{1 + \rho + \rho^2 + \cdots + \rho^\infty} = 1 - \rho$$

$$p_n = (1-\rho)\rho^n \quad n = 0, 1, 2, \ldots, \infty$$

- *n* is geometrically distributed.



- Utilization of the server
  = Probability of having one or more jobs in the system:

$$U = 1 - p_0 = \rho$$

Results for M/M/1 Queue(Cont)

- Mean number of jobs in the system:

$$E[n] = \sum_{n=1}^{\infty} n p_n = \sum_{n=1}^{\infty} n(1-\rho)\rho^n = \frac{\rho}{1-\rho}$$

- Variance of the number of jobs in the system:

$$\begin{aligned} \mathrm{Var}[n] &= E[n^2] - (E[n])^2 \\ &= \left(\sum_{n=1}^{\infty} n^2 (1-\rho)\rho^n\right) - (E[n])^2 = \frac{\rho}{(1-\rho)^2} \end{aligned}$$

Box 31.1: M/M/1 Queue

1. Parameters:
   λ = arrival rate in jobs per unit time
   μ = service rate in jobs per unit time
2. Traffic intensity: $\rho = \lambda/\mu$
3. Stability condition: Traffic intensity ρ must be less than 1.
4. Probability of zero jobs in the system: $p_0 = 1 - \rho$
5. Probability of n jobs in the system: $p_n = (1-\rho)\rho^n, \qquad n = 0, 1, \ldots, \infty$
6. Mean number of jobs in the system: $E[n] = \rho/(1-\rho)$
7. Variance of number of jobs in the system: $\mathrm{Var}[n] = \rho/(1-\rho)^2$
8. Probability of k jobs in the queue: $P(n_q = k) = \begin{cases} 1-\rho^2 & k = 0 \\ (1-\rho)\rho^{k+1} & k > 0 \end{cases}$
9. Mean number of jobs in the queue: $E[n_q] = \rho^2/(1-\rho)$
10. Variance of number of jobs in the queue: $\mathrm{Var}[n_q] = \rho^2(1+\rho-\rho^2)/(1-\rho)^2$
11. Cumulative distribution function of the response time: $F(r) = 1 - e^{-r\mu(1-\rho)}$
12. Mean response time: $E[r] = (1/\mu)/(1-\rho)$
13. Variance of the response time: $\mathrm{Var}[r] = (1/\mu^2)/(1-\rho)^2$

Box 31.1: M/M/1 Queue (Cont)

14. q-Percentile of the response time $= E[r]\ln[100/(100-q)]$
15. 90-Percentile of the response time $= 2.3E[r]$
16. Cumulative distribution function of waiting time: $F(w) = 1 - \rho e^{-\mu w(1-\rho)}$
17. Mean waiting time: $E[w] = \rho(1/\mu)/(1-\rho)$
18. Variance of the waiting time: $\mathrm{Var}[w] = (2-\rho)\rho)/[\mu^2(1-\rho)^2]$
19. q-Percentile of the waiting time: $\max(0, \frac{E[w]}{\rho}\ln[100\rho/(100-q)])$
20. 90-Percentile of the waiting time: $\max(0, \frac{E[w]}{\rho}\ln[10\rho])$
21. Probability of finding n *or more* jobs in the system: ρ^n
22. Probability of serving n jobs in one busy period: $\frac{1}{n}\begin{pmatrix} 2n-2 \\ n-1 \end{pmatrix}\frac{\rho^{n-1}}{(1+\rho)^{2n-1}}$
23. Mean number of jobs served in one busy period: $\frac{1}{1-\rho}$
24. Variance of number of jobs served in one busy period $= \frac{\rho(1+\rho)}{(1-\rho)^3}$
25. Mean busy period duration: $\frac{1}{\mu(1-\rho)}$
26. Variance of the busy period: $\frac{1}{\mu^2(1-\rho)^3} - \frac{1}{\mu^2(1-\rho)^2}$

Example 31.1

- On a network gateway, measurements show that the packets arrive at a mean rate of 125 packets per second (pps) and the gateway takes about two milliseconds to forward them. Using an M/M/1 model, analyze the gateway. What is the probability of buffer overflow if the gateway had only 12 buffers? How many buffers do we need to keep packet loss below one packet per million?
- Arrival rate $\lambda = 125$ pps
- Service rate $\mu = 1/.002 = 500$ pps
- Gateway Utilization $\rho = \lambda/\mu = 0.25$
- Probability of n packets in the gateway
  $= (1-\rho)\rho^n = 0.75(0.25)^n)$

Example 31.1(Cont)

- Mean Number of packets in the gateway
  $= \rho/(1-\rho) = 0.25/0.75 = 0.33$
- Mean time spent in the gateway
  $= (1/\mu)/(1-\rho) = (1/500)/(1-0.25) = 2.66$ milliseconds
- Probability of buffer overflow

$$
\begin{aligned}
&= \text{P(more than 13 packets in the gateway)} \\
&= \rho^{14} = 0.25^{14} = 3.73 \times 10^{-9} \\
&\approx \text{4 packets per billion packets.}
\end{aligned}
$$

- To limit the probability of loss to less than 10^{-6}:

$$\rho^n \leq 10^{-6}$$

$$n > \log(10^{-6})/\log(0.25) = 9.96$$

We need about nine buffers.

Example 31.1(Cont)

- The last two results about buffer overflow are approximate. Strictly speaking, the gateway should actually be modeled as a finite buffer *M/M/1/B* queue. However, since the utilization is low and the number of buffers is far above the mean queue length, the results obtained are a close approximation.

Quiz 31

- The average response time on a database system is three seconds. During a one minute observation interval, the idle time on the system was measured to be ten seconds. Using an *M/M/1* model for the system, determine the following:

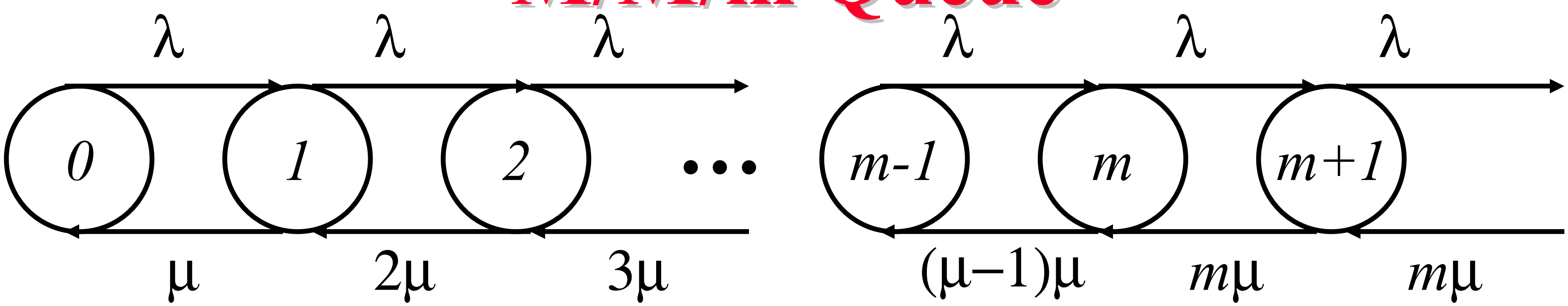
Given: $E[r] =$_________ $\rho =$ _________

1. System utilization $u =$ __________
2. Average service time per query $1/\mu =$ __________
3. Averate arrival rate $\lambda = \mu\rho =$ _________
4. Average number of jobs in the system $E[n] =$ __________

Key:

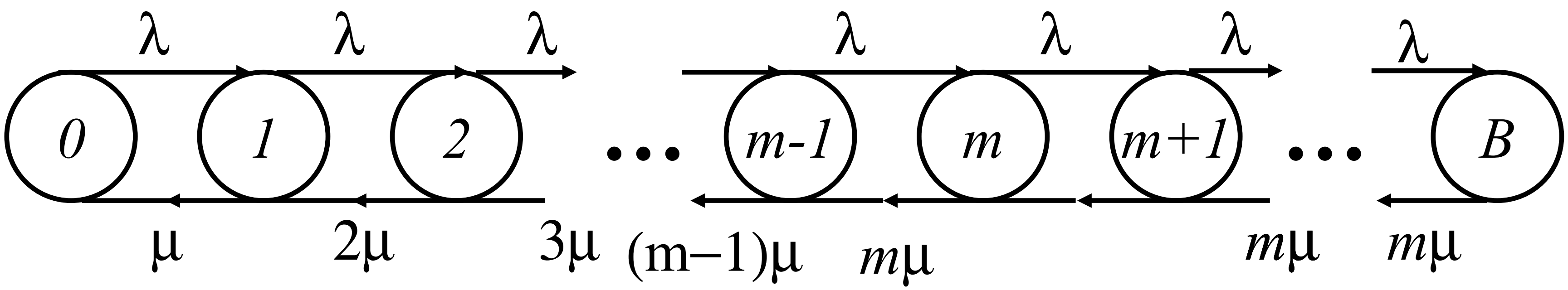
$$\rho = \lambda/\mu$$
$$p_n = (1-\rho)\rho^n$$
$$E[n] = \rho/(1-\rho)$$
$$E[r] = (1/\mu)/(1-\rho)$$

M/M/m Queue



M/M/1	M/M/m
$\rho = \lambda/\mu$	$\rho = \lambda/(m\mu)$
$U = \rho$	$U = \rho$
$p_0 = 1 - \rho$	$p_0 = \left[1 + \frac{(m\rho)^m}{m!(1-\rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!}\right]^{-1}$
$p_n = p_0 \rho^n$	$p_n = \begin{cases} p_0 \frac{(m\rho)^n}{n!} & n < m \\ p_0 \frac{\rho^n m^m}{m!} & n \geq m \end{cases}$
$P(queueing) = \rho$	$\varrho = P(\geq m \text{ jobs}) = \frac{(m\rho)^m}{m!(1-\rho)} p_0$
$E[r] = \frac{1/\mu}{1-\rho}$	$E[r] = \frac{1}{\mu}\left(1 + \frac{\varrho}{m(1-\rho)}\right)$

M/M/m/B Queue



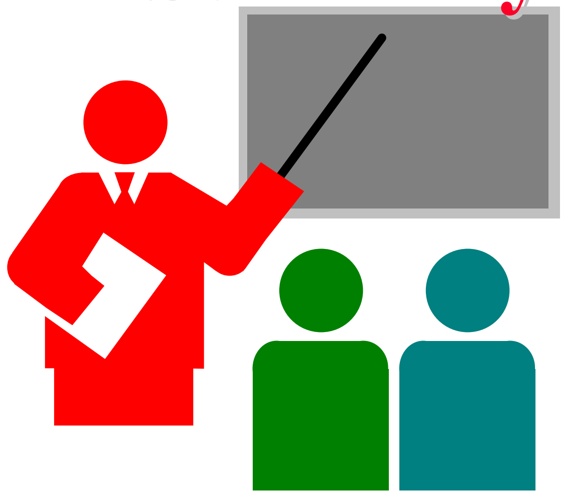
M/M/1	**M/M/m/B**
$\rho = \lambda/\mu$	$\rho = \lambda/(m\mu)$
$p_0 = 1 - \rho$	$p_0 = \left[1 + \frac{(1-\rho^{B-m+1})(m\rho)^m}{m!(1-\rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!}\right]^{-1}$
$p_n = \rho^n(1-\rho)$	$p_n = \begin{cases} \frac{1}{n!}(m\rho)^n p_0 & 0 \le n < m \\ \frac{m^m \rho^n}{m!} p_0 & m \le n \le B \end{cases}$
$U = \rho$	$U = \rho(1 - p_B)$
Loss Rate $= 0$	Loss Rate $= \lambda p_B$

Other Queues

- M/G/1
- M/G/1/∞/∞/Processor Sharing
- M/D/1
- M/G/∞
- G/M/1
- G/G/m

The textbook has results for these queues.

Summary



- Birth-death processes: Compute probability of having n jobs in the system
- M/M/1 Queue: Load-independent => Arrivals and service do not depend upon the number in the system $\lambda_n = \lambda$, $\mu_n = \mu$
- Traffic Intensity: $\rho = \lambda / \mu$
- Mean Number of Jobs in the system = $\rho/(1-\rho)$
- Mean Response Time = $(1/\mu)/(1-\rho)$

Homework 31

- Reconsider the system your team selected for Homework 30 (or any other system for which you know the arrival rates and service rate). Using an *M/M/1* model for the system, determine the following:
  a. System utilization
  b. Average response time per job
  c. Average number of jobs in the system
  d. Probability of number of jobs in the system being greater than *10*
  e. *90*-percentile response time
  f. *90*-percentile waiting time
- Due: Via email to jain@eecs by noon next Monday.

Reading List

- Chapter 31