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:

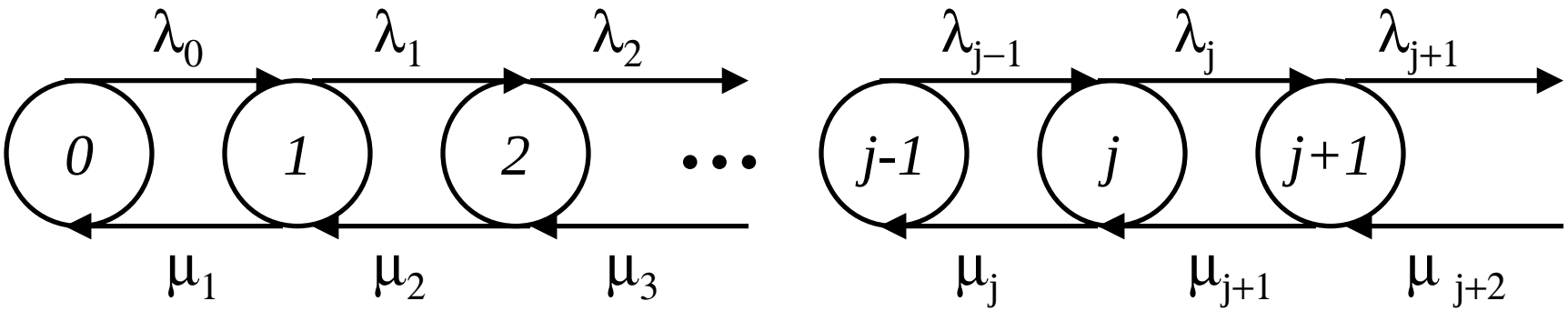
http://www.cse.wustl.edu/~jain/cse567-08/

Overview

- 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 $n+1 \Rightarrow$ birth
- Departure of a job changes the system state to $n-1 \Rightarrow$ 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:

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

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

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

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*

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

Proof(Cont)

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

$$\lim_{\Delta t \leftarrow 0} \frac{p_j(t+\Delta t) - p_j(t)}{\Delta t} = \lambda_{j-1} p_{j-1}(t) - (\mu_j + \lambda_j) p_j(t) + \mu_{j+1} p_{j+1}(t)$$

$$\frac{dp_j(t)}{dt} = \lambda_{j-1} p_{j-1}(t) - (\mu_j + \lambda_j) p_j(t) + \mu_{j+1} p_{j+1}(t)$$

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

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

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

$$\lim_{\Delta t \leftarrow 0} \frac{p_j(t+\Delta t) - p_j(t)}{\Delta t} = \lambda_{j-1} p_{j-1}(t) - (\mu_j + \lambda_j) p_j(t) + \mu_{j+1} p_{j+1}(t)$$

$$\frac{dp_j(t)}{dt} = \lambda_{j-1} p_{j-1}(t) - (\mu_j + \lambda_j) p_j(t) + \mu_{j+1} p_{j+1}(t)$$

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:

$$p_n = \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}p_0$$

$$= p_0\prod_{j=0}^{n-1}\frac{\lambda_j}{\mu_{j+1}} \quad n = 1, 2, \ldots, \infty$$

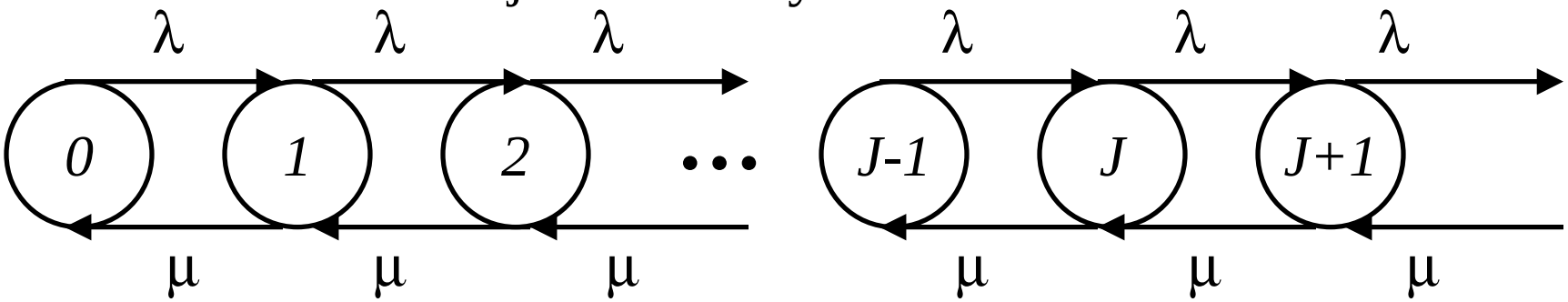
Proof(Cont)

- 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:

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

Results for M/M/1 Queue(Cont)

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

$$p_n = \rho^n p_0$$

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

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

- 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:

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

- Variance of the number of jobs in the system:

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

Results for M/M/1 Queue(Cont)

- Probability of n or more jobs in the system:

$$P(\geq n \text{ jobs in system}) = \sum_{j=n}^{\infty} p_j = \sum_{j=n}^{\infty} (1-\rho)\rho^j = \rho^n$$

- Mean response time (using the Little's law):

Mean number in the system = Arrival rate × Mean response time}

That is:

$$E[n] = \lambda E[r]$$

$$E[r] = \frac{E[n]}{\lambda} = \left(\frac{\rho}{1-\rho}\right)\frac{1}{\lambda} = \frac{1/\mu}{1-\rho}$$

Results for M/M/1 Queue(Cont)

- Cumulative distribution function of the response time:

$$F(r) = 1 - e^{-r\mu(1-\rho)}$$

- The response time is exponentially distributed.
  $\Rightarrow$ q-percentile of the response time

$$1 - e^{-r_q\mu(1-\rho)} = \frac{q}{100}$$

$$r_q = \frac{1}{\mu(1-\rho)} \ln\left(\frac{100}{100-q}\right)$$

Results for M/M/1 Queue(Cont)

- Cumulative distribution function of the waiting time:

$$F(w) = 1 - \rho e^{-w\mu(1-\rho)}$$

- This is a truncated exponential distribution. Its q-percentile is given by:

$$w_q = \frac{1}{\mu(1-\rho)} \ln\left(\frac{100\rho}{100-q}\right)$$

- The above formula applies only if *q* is greater than *100(1-ρ)*. All lower percentiles are zero.

$$w_q = max\left\{0, \frac{E[w]}{\rho} \ln\left(\frac{100\rho}{100-q}\right)\right\}$$

Results for M/M/1 Queue(Cont)

- Mean number of jobs in the queue:

$$E[n_q] = \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty}(n-1)(1-\rho)\rho^n = \frac{\rho^2}{1-\rho}$$

- Idle $\Rightarrow$ there are no jobs in the system
- The time interval between two successive idle intervals
- All results for *M/M/1* queues including some for the busy period are summarized in Box 31.1 in the book.

Example 31.2

- 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 13 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 31.2(Cont)

- Mean Number of packets in the gateway
  = *(ρ/(1-ρ) = 0.25/0.75 = 0.33)*
- Mean time spent in the gateway
  = *((1/μ)/(1-ρ)= (1/500)/(1-0.25)* = *2.66* milliseconds
- Probability of buffer overflow

$$\begin{aligned} &= \quad \text{P(more than 13 packets in the gateway)} \\ &= \quad \rho^{13} = 0.25^{13} = 1.49 \times 10^{-8} \\ &\approx \quad 15 \text{ packets per billion packets.} \end{aligned}$$

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

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

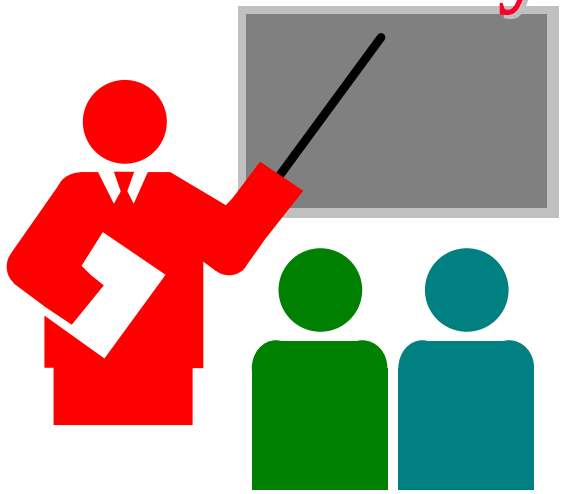
$$n > \log(10^{-6}) / \log(0.25) = 9.96$$

We need about ten buffers.

Example 31.2(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.

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

- Submit answers to modified Exercise 31.3

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

- a. System utilization
- b. Average service time per query
- c. Number of queries completed during the observation interval
- d. Average number of jobs in the system
- e. Probability of number of jobs in the system being greater than *10*
- f. *90*-percentile response time
- g. *90*-percentile waiting time