

EEC 686/785  
Modeling & Performance Evaluation of  
Computer Systems

---

## Lecture 21

Wenbing Zhao

Department of Electrical and Computer Engineering  
Cleveland State University

wenbing@ieee.org

(based on Dr. Raj jain's lecture notes)

2



## Outline

---

- Midterm #2
- Review of lecture 20
- Analysis of A Single Queue



## Midterm #2 Results

P1	P2	P3	P4	P5	Total	Total (normalized)
20	16	10	20	20	86	107.5
26	16	10	20	10	82	102.5
12	14	10	20	10	66	82.5
24	16	10	16	0	66	82.5
20	16	10	19	0	65	81.25
22	12	10	16	0	60	75
22	16	10	0	0	48	60



## Accumulated Credit (53.3%)

HW1	HW2	HW3	HW4	MT1	MT2	Total (53.3%)	Total (normalized)
10	10	38	38	78	107.5	48.8	91.6
9	10	38	33	78.5	102.5	47.2	88.6
10	10	36	32	64	82.5	40.5	75.9
9	10	38	42	80	60	39.6	74.4
8.5	10	38	31	61.5	81.25	39.3	73.7
8.5	10	38	42	61.5	75	38.8	72.7
4	0	38	33	0	82.5	22.6	42.3



## Introduction to Queueing Theory

---

- Queueing notation
- Rules for all queues
- Little's law
- Types of stochastic processes



## Queueing Notation

---

- Kendall notation:  $A/S/m/B/K/SD$ 
  - A: arrival process
  - S: service time distribution
  - m: number of servers
  - B: number of buffers (system capacity)
  - K: population size
  - SD: service discipline

## Key Variables

- $\tau$  = interarrival time  
= time between two successive arrivals
- $\lambda$  = mean arrival rate =  $1/E[\tau]$
- $s$  = service time per job
- $\mu$  = mean service rate per server =  $1/E[s]$
- $n$  = number of jobs in the system
- $r$  = response time or the time in the system

## Rules for All Queues (G/G/m)

- Stability condition:  $\lambda < m\mu$  ( $m$ : number of servers)
- Number in system versus number in queue:  
$$n = n_q + n_s$$
- Number versus time: if jobs are not lost due to insufficient buffers,  
Mean number of jobs in the system  
= (arrival rate)  $\times$  (mean response time)
- Time in system versus time in queue:  
$$r = w + s$$



## Stochastic Processes

---

- **Process**: function of time
- **Stochastic process**: random variables, which are functions of time
- Types of Stochastic Processes
  - Discrete or continuous state processes
  - Markov processes
  - Birth-death processes
  - Poisson processes



## Poisson Processes

---

- Interarrival times = IID + exponential
  - Number of arrivals  $n$  over a given interval  $(t, t + x)$  has a Poisson distribution
  - The arrival process is referred to as a **Poisson process** or a **Poisson stream**



## Properties of Poisson Processes

---

- **Merging:**  $\lambda = \sum_{i=1}^k \lambda_i$
- **Splitting:** if the probability of a job going to  $i$ th substream is  $p_i$ , each substream is also Poisson with a mean rate of  $p_i \lambda$



## Properties of Poisson Processes

---

- If the arrivals to a single server with exponential service time are Poisson with mean rate  $\lambda$ , the departures are also Poisson with the same rate  $\lambda$  provided  $\lambda < \mu$



## Properties of Poisson Processes

- If the arrivals to a service facility with  $m$  service centers are Poisson with a mean rate  $\lambda$ , the departures also constitute a Poisson stream with the same rate  $\lambda$ , provided  $\lambda < \sum_i \mu_i$ .
  - Here, the servers are assumed to have exponentially distributed service times



## Relationship Among Stochastic Processes

- **Markov process** is broader than **birth-death process** and **birth-death process** is broader than **Poisson process**
- The Poisson process can be modeled as a pure birth process with constant birth rate
- All birth-death processes are Markov processes with the restriction that the transitions are restricted to neighboring states



## Analysis of A Single Queue

---

- Birth-death processes
- M/M/1 queue
- M/M/m queue
- M/M/m/B queue
- Results for other queueing systems (omitted)



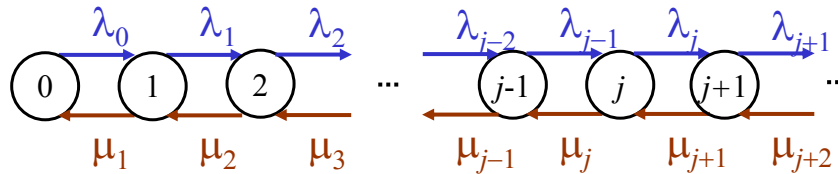
## Birth-Death Processes

---

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

## Birth-Death Processes

- State-transition diagram



- In state  $n$ :
  - New arrivals take place at a rate  $\lambda_n$
  - The service rate is  $\mu_n$
- Both the interarrival times and service times are assumed exponentially distributed

## Theorem: State Probability

- The steady-state probability  $p_n$  of a birth-death process being in state  $n$  (i.e., there are  $n$  jobs in the system) is given by:

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

- Here,  $p_0$  is the probability of being in the zero state (no job in system, i.e., system is idle)



## Proof

- Suppose the system is in state  $j$  at time  $t$
- In the next  $\Delta t$ , the system can move to state  $j-1$  or  $j+1$  with the following probabilities
  - $P\{n(t+\Delta t) = j+1 \mid n(t) = j\}$   
= probability of one arrival in interval  $\Delta t = \lambda_j \Delta t$
  - $P\{n(t+\Delta t) = j-1 \mid n(t) = j\}$   
= probability of one departure in interval  $\Delta t = \mu_j \Delta t$
- If there are no arrivals or departures, the system will stay in state  $j$  and, thus
  - $P\{n(t+\Delta t) = j \mid n(t) = j\} = 1 - \lambda_j \Delta t - \mu_j \Delta t$



## Proof

- $\Delta t = \text{small} \Rightarrow$  zero probability of two events (two arrivals, two departures, or a arrival and a departure) occurring during this interval



## Proof

- $p_j(t)$  = probability of being in state  $j$  at time  $t$

$$p_0(t + \Delta t) = (1 - \lambda_0 \Delta t) p_0(t) + \mu_1 \Delta t p_1(t)$$

$$p_1(t + \Delta t) = \lambda_0 \Delta t p_0(t) + (1 - \mu_1 \Delta t - \lambda_1 \Delta t) p_1(t) + \mu_2 \Delta t p_2(t)$$

$$p_2(t + \Delta t) = \lambda_1 \Delta t p_1(t) + (1 - \mu_2 \Delta t - \lambda_2 \Delta t) p_2(t) + \mu_3 \Delta t p_3(t)$$

...

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

...



## Proof

- The  $j$ th equation above can be written as follows:

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

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

## Proof

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

$$\lim_{t \rightarrow \infty} p_j(t) = p_j \quad \text{and} \quad \lim_{t \rightarrow \infty} \frac{dp_j(t)}{dt} = 0$$

- Substituting these in the  $j$ th equation, we get:

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

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

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

## Proof

- The solution to this set of equations is:

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

- The sum of all probabilities must be equal to 1:

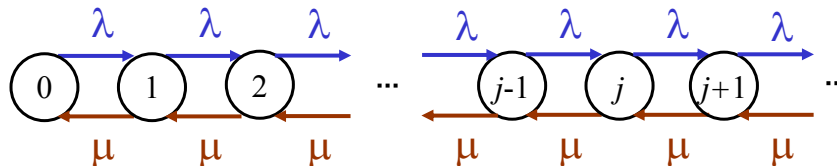
$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \prod_{j=0}^{n-1} \frac{\lambda_j}{\mu_{j+1}}}$$

## 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  $\lambda$  and the mean service rate  $\mu$

## M/M/1 Queue

- State = number of jobs in the system



## Results for M/M/1 Queues

- Birth-death processes with

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

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

- Probability of  $n$  jobs in the system

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

- The quantity  $\lambda/\mu$  is called **traffic intensity** and is usually denoted by symbol  $\rho$ . Thus:

## Results for M/M/1 Queues

$$p_n = \rho^n p_0$$

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

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

$n$  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 Queues

- 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:

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

## Results for M/M/1 Queues

- 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  $\times$  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 Queues

- Cumulative distribution function of the response time:

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

- The response time is exponentially distributed.  
=>  $q$ -percentile of the response time

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

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