Fundamentals of Queueing Theory

Prof. Jenhui Chen

E-mail: jhchen@mail.cgu.edu.tw

Computer Science & Information Engineering
Chang Gung University, Taiwan

October 13, 2015

1. Grade Criteria
2. Chapter 1. Introduction
3. Chapter 2. Simple Markovian Queueing Models

Grade Criteria

- Prof. Liu, 25%
- Prof. Chen, 25%
  - 10/13 (1:00–4:00PM), 10/14 (4:00–5:00PM)
  - 10/20 (1:00–4:00PM), 10/21 (4:00–5:00PM)
  - 10/27 (Quiz: 40%), 10/28 (4:00–5:00PM)
  - 11/03 (1:00–4:00PM), 11/04 (4:00–5:00PM)
  - 11/10 (Midterm Exam: 60%)
- Prof. Sahoo, 25%, during November
- Prof. Chang, 25%

Definition of Random Variable

The definition of random variable

- The random variable is the observation.
- The random variable is a function of the observation.
- The random variable is a function of another random variable.

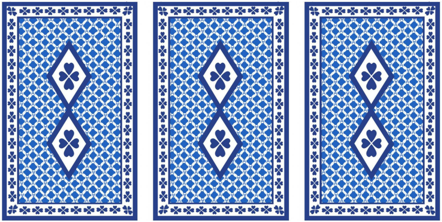
Probability Problems

Q1: The problem of deciding volunteers (a public errand)

- Usually it's common to draw lots to decide who should be the volunteers of a public errand. In this way, we will let all people line up for drawing the made lots. What is the probability of being the volunteers of each person if we need three volunteers out from 10 people?
- What is the probability of the first one to draw the lots?
- Based on the first drawn, what is the probability of the second one to draw the lots? And what is the probability of the third one?

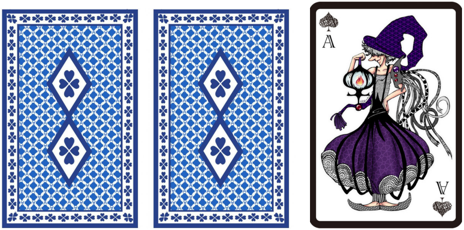
Probability Problems

Q2: The problem of wining your prize.
Among these three cards, there is a card for prize of a luxury car. Now you have the chance to guess which one is the prize. Make a choice!! After your choice is done, the host turns on a 'sorry' card to let you make a new decision (i.e., choose your card again because the win probability becomes 50% which is higher than the beginning $1/3$). Will you change your decision again? Why or why not? What is the probability if you reselect a new card?



Probability Problems

Q2: The problem of wining your prize.
Among these three cards, there is a card for prize of a luxury car. Now you have the chance to guess which one is the prize. Make a choice!! After your choice is done, the host turns on a ‘sorry’ card to let you make a new decision (i.e., choose your card again because the win probability becomes 50% which is higher than the beginning $1/3$). Will you change your decision again? Why and why not? What is the probability if you reselect a new card?



Description of Queueing Problem

- A queueing system can be described as customers arriving for service, waiting for service if it is not immediate, and if having waited for service, leaving the system after being served.
- The term “customer” is used in a general sense and does not imply necessarily a human customer.
- A telephone system is generally characterized by
  - Poisson input, exponential holding (service) times, and multiple channels (servers)
  - Poisson input, constant holding time, and a single channel

Characteristics of Queueing Processes

Six characteristics of queueing processes

1. arrival pattern of customers
2. service pattern of servers
3. queue discipline
4. system capacity
5. number of service channels
6. number of service stages

Notation

A queueing process is described by a series of symbols and slashes such as $A/B/X/Y/Z$, where

- A indicates in some way the interarrival-time distribution
- B the service pattern as described by the probability distribution for service time
  - M: Exponential
  - D: Deterministic
  - E_k: Erlang type k ($k = 1, 2, \ldots$)
  - H_k: Mixture of k exponentials
  - PH: Phase type
  - G: General (Arbitrary)
- X the number of parallel service channels

Notation

A queueing process is described by a series of symbols and slashes such as $A/B/X/Y/Z$, where

- Y the restriction on system capacity
- Z the queu discipline
  - FCFS: First come, first served
  - LCFS: Last come, first served
  - RSS: Random selection for service
  - PR: Priority
  - GD: General discipline
- For example $M/D/2/\infty/\text{FCFS}$

Measuring System Performance

Generally there are three types of system responses of interest:

- Some measure of the waiting time that a typical customer might be forced to endure
- An indication of the manner in which customers may accumulate
- A measure of the idle time of the servers

Some General Results

- $G/G/1$ or $G/G/c$
- Denoting the average rate of customers entering the queueing system as λ, the average rate of serving customers as μ, and a measure of traffic congestion for c-server systems is
- $\rho \equiv \lambda / c\mu$ (often called traffic intensity)
- Three conditions
  - $\rho > 1$ $(\lambda > c\mu)$, as time goes on, the queue to get bigger and bigger, unless, at some point, customers were not allowed to join.
  - $\rho = 1$, unless arrivals and service are deterministic and perfectly scheduled, no steady state exists, since randomness will prevent the queue from ever emptying out and allowing the servers to catch up, thus causing the queue to grow without bound.
  - $\rho = \lambda / c\mu < 1$ is the only condition we consider

- What we most often desire in solving queueing models is to find the probability distribution for the total number of customers in the system at time t, $N(t)$, which is made up of those waiting in queue, $N_q(t)$, plus those in service $N_s(t)$
- Let $p_n(t) = \Pr\{N(t) = n\}$ and $p_n = \Pr\{N = n\}$ in the steady state
- Two expected-value measures of major interest are
  - The mean number in the system

$$L = E[N] = \sum_{n=0}^{\infty} np_n$$

  - The expected number in queue

$$L_q = E[N_q] = \sum_{n=c+1}^{\infty} (n-c)p_n$$

Little's Formulas

- $T = T_q + S$, where S is the service time, and T, T_q, and S are random variables
- Two often used measures of system performance with respect to customers are
  - The mean waiting time in queue
    $$W_q = E[T_q]$$
  - The mean waiting time in the system
    $$W = E[T]$$
  - $E[T] = E[T_q] + E[S]$
- We have the Little's Formulas are

$$L = \lambda W$$

$$L_q = \lambda W_q$$

- Denoting the number of customer as N_c that arrive over the time period $(0, T)$ is 4.

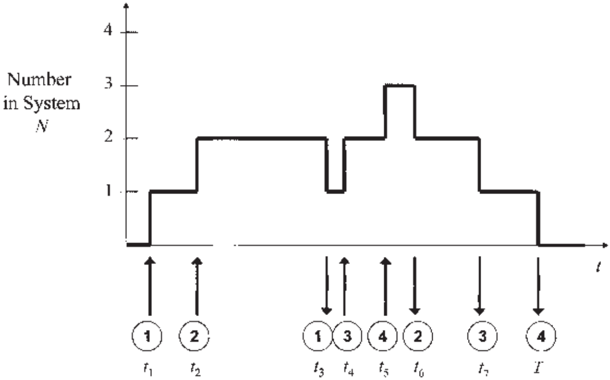


Figure 1.4 Busy-period sample path.

The Calculation for L and W

$$\begin{aligned} L &= [1(t_2 - t_1) + 2(t_3 - t_2) + 1(t_4 - t_3) + 2(t_5 - t_6) \\ &\quad + 3(t_6 - t_5) + 2(t_7 - t_6) + 1(T - t_7)]/T \\ &= (\text{area under curve})/T \\ &= (T + t_7 + t_6 - t_5 - t_4 + t_3 - t_2 - t_1)/T \end{aligned}$$

$$\begin{aligned} W &= [(t_3 - t_1) + (t_6 - t_2) + (t_7 - t_4) + (T - t_5)]/4 \\ &= (T + t_7 + t_6 - t_5 - t_4 + t_3 - t_2 - t_1)/4 \\ &= (\text{area under curve})/N_c \end{aligned}$$

$LT = WN_c$, which yields $L = WN_c/T = W\lambda$

The Calculation for L and W

$$L - L_q = \lambda(W - W_q) = \lambda(1/\mu) = \lambda/\mu$$

- $$L - L_q = E[N] - E[N_q] = E[N - N_q] = E[N_s]$$
- For a single-server system that $r = \rho$ and it follows from simple algebra that

$$L - L_q = \sum_{n=1}^{\infty} np_n - \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty} p_n = 1 - p_0$$

- A simple expected-value argument, we show that $p_b = \rho$

$$r/c = \rho = 0 \cdot (1 - p_b) + 1 \cdot p_b$$

- Because $p_0 = 1 - p_b$, in this case, then

$$p_0 = 1 - \rho = 1 - r = 1 - \lambda/\mu$$

Table 1.2 Summary of General Results

Table: Summary of General Results for $G/G/c$ Queue

$\rho = \lambda/c\mu$	Traffic intensity; offered work load rate to a server
$L = \lambda W$	Little's formula
$L_q = \lambda W_q$	Little's formula
$W = W_q + 1/\mu$	Expected-value argument
$p_b = \lambda/c\mu = \rho$	Busy probability for an arbitrary server
$r = \lambda/\mu$	Expected number of customers in service; offered work load ra
$L = L_q + r$	Combined result — (1.3)
$p_0 = 1 - \rho$	$G/G/1$ empty-system probability
$L = L_q + (1 - p_0)$	Combined result for $G/G/1$

Some Examples

- The Carry Out Curry House, a fast-food Indian restaurant, must decide on how many parallel service channels to provide. They estimate that, during the rush hour, the average number of arrivals per hour will be approximately 40. They also estimate that, on average, a server will take about 5.5 min to serve a typical customer. Using only this information, about how many service channels (clerks) will you recommend they install?

Some Examples

- The Carry Out Curry House, a fast-food Indian restaurant, must decide on how many parallel service channels to provide. They estimate that, during the rush hour, the average number of arrivals per hour will be approximately 40. They also estimate that, on average, a server will take about 5.5 min to serve a typical customer. Using only this information, about how many service channels (clerks) will you recommend they install?

Sol: $\lambda = 40/\text{hr} = 2/3$ per minute; $\mu = 1/5.5$
$\rho = \lambda/c\mu < 1 \rightarrow c > \lambda/\mu = 2/3 \times 5.5 = 3.6667$
Thus, we recommend they install 4 servers.

Some Examples

- Fluffy Air, a small local feeder airline, needs to know how many slots to provide for telephone callers to be placed on hold. They plan to have enough answerers so that the average waiting time on hold for a caller will be 75 seconds during the busiest period of the day. They estimate the average call-in rate to be 3 per minute. How many slots would you advise Fluffy Air to set up?

Some Examples

- Fluffy Air, a small local feeder airline, needs to know how many slots to provide for telephone callers to be placed on hold. They plan to have enough answerers so that the average waiting time on hold for a caller will be 75 seconds during the busiest period of the day. They estimate the average call-in rate to be 3 per minute. How many slots would you advise Fluffy Air to set up?

Sol: $L_q = \lambda W_q = (3/\text{min})([75/60]\text{min}) = 3.75$ or, say 4. The 3.75 number is, of course, the average number in the queue. We may wish to provide 5 or 6 slots to guarantee that most callers get into the queue.

Simple Data Bookkeeping for Queues

- Some expression for expressing the number of customers in the system

$$n(t) = \{\text{number of arrivals in } (0, t]\} - \{\text{number of services completed in } (0, t]\}.$$

- Notice that the notation $(0, t]$ means

$$0 < \text{time} \leq t.$$

Poisson Process and the Exponential Distribution (1/7)

- The most common stochastic queueing models assume that **interarrival times** and **service times** obey the exponential distribution or, equivalently, that the arrival rate and service rate follow a Poisson distribution.
- Consider an arrival counting process $\{N(t), t \geq 0\}$, where $N(t)$ denotes the total number of arrivals up to time t, with $N(0) = 0$, and which satisfies the following three assumptions:

Poisson Process and the Exponential Distribution (2/7)

1. The probability that an arrival occurs between time t and time $t + \Delta t$ is equal to $\lambda \Delta t + o(\Delta t)$. We write this as $\Pr\{\text{arrival occurs between } t \text{ and } t + \Delta t\} = \lambda \Delta t + o(\Delta t)$, where λ is a constant independent of $N(t)$, Δt is an incremental element, and $o(\Delta t)$ denotes a quantity that becomes negligible when compared to Δt as $\Delta t \to 0$; that is,
$$\lim_{\Delta t \to 0} \left(\frac{o(\Delta t)}{\Delta t} \right) = 0$$
2. $\Pr\{\text{more than one arrival between } t \text{ and } t + \Delta t\} = o(\Delta t)$
3. The number of arrivals in nonoverlapping intervals are statistically independent; that is, the process has independent increments.

Poisson Process and the Exponential Distribution (3/7)

- To calculate $p_n(t)$, the probability of n arrivals in a time interval of length t, n being an integer ≥ 0. We will do this by first developing differential-difference equations for the arrival process. For $n \geq 1$ we have

$$
\begin{aligned}
p_n(t+\Delta t) = \Pr\{&n \text{ arrivals in } t \text{ and none in } \Delta t\} \\
&+ \Pr\{n-1 \text{ arrivals in } t \text{ and one in } \Delta t\} \\
&+ \Pr\{n-2 \text{ arrivals in } t \text{ and two in } \Delta t\} + \cdots \\
&+ \Pr\{\text{no arrivals in } t \text{ and } n \text{ in } \Delta t\} \qquad (1.6)
\end{aligned}
$$

- Using assumptions i, ii, and iii, (1.6) becomes

$$p_n(t+\Delta t) = p_n(t)[1-\lambda\Delta t - o(\Delta t)] + p_{n-1}(t)[\lambda\Delta t + o(\Delta t)] + o(\Delta t),$$

where the last term, $o(\Delta t)$, represents the terms $\Pr\{n-j$ arrivals in t and j in $\Delta t; 2 \leq j \leq n\}$.

Poisson Process and the Exponential Distribution (4/7)

- For the case $n = 0$, we have

$$p_0(t + \Delta t) = p_0(t)[1 - \lambda \Delta t - o(\Delta t)] \tag{1.8}$$

- Rewriting (1.7) and (1.8) and combining all $o(\Delta t)$ terms, we have

$$p_0(t + \Delta t) - p_0(t) = -\lambda \Delta t p_0(t) + o(\Delta t) \tag{1.9}$$

and

$$p_n(t+\Delta t) - p_n(t) = -\lambda \Delta t p_n(t) + \lambda \Delta t p_{n-1}(t) + o(\Delta t) \quad (n \geq 1). \quad (1.$$

Poisson Process and the Exponential Distribution (5/7)

- We divide (1.9) and (1.10) by Δt, take the limit as $\Delta t \to 0$, and obtain the differential-difference equations

$$\lim_{\Delta t \to 0}\left[\frac{p_0(t+\Delta t) - p_0(t)}{\Delta t} = -\lambda p_0(t) + \frac{o(\Delta t)}{\Delta t}\right],$$

$$\lim_{\Delta t \to 0}\left[\frac{p_n(t+\Delta t) - p_n(t)}{\Delta t} = -\lambda p_n(t) + \lambda P_{n-1}(t) + \frac{o(\Delta t)}{\Delta t}\right] \quad (n \geq 1$$

which reduce to

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t) \qquad (1.11)$$

and

$$\frac{dp_n(t)}{dt} = -\lambda p_n(t) + \lambda p_{n-1}(t) \quad (n \geq 1). \qquad (1.12)$$

Poisson Process and the Exponential Distribution (6/7)

- We now have an infinite set of linear, first-order ordinary differential equations to solve. Equation (1.11) clearly has the general solution $p_0(t) = Ce^{-\lambda t}$, where the constant C is easily determined to be equal to 1, because $p_0(0) = 1$. Next, let $n = 1$ in (1.12), and we find that

$$\frac{dp_1(t)}{dt} = -\lambda p_1(t) + \lambda p_0(t)$$

or

$$\frac{dp_1(t)}{dt} + \lambda p_1(t) = \lambda p_0(t) = \lambda e^{-\lambda t}.$$

- The solution to this equation is

$$p_1(t) = Ce^{-\lambda t} + \lambda t e^{-\lambda t}.$$

Poisson Process and the Exponential Distribution (7/7)

- Use of the boundary condition $p_n(0) = 0$ for all $n > 0$ yields $C = 0$ and gives

$$p_1(t) = \lambda t e^{-\lambda t}.$$

- Continuing sequentially to $n = 2, 3, \ldots$ in (1.12) and proceeding similarly, we find that

$$p_2(t) = \frac{(\lambda t)^2}{2!} e^{-\lambda t}, \qquad p_3(t) = \frac{(\lambda t)^3}{3!} e^{-\lambda t}, \qquad \cdots (1.13)$$

- From (1.13), we conjecture that the general **Poisson formula** is

$$p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}. \qquad (1.14)$$

Additional Interesting Poisson Properties

- **Memoryless property**
- If we consider the random variable defined as the number of arrivals to a queueing system by time t, this random variable has the Poisson distribution given by (1.14) with a mean of λt arrivals, or a mean arrival rate (arrivals per unit time) of λ.
- Poisson processes have a number of interesting additional properties. One of most important is that the number of occurrences in intervals of equal width are identically distributed (stationary increments). In particular, for $t > s$, the difference $N(t) - N(s)$ is identically distributed as $N(t + h) - N(s + h)$, with probability function

$$p_n(t - s) = \frac{[\lambda(t - s)]^n}{n!} e^{-\lambda(t-s)}.$$

Interarrival Time Follows the Exponential Distribution

- We now show that if the arrival process is Poisson, an associated random variable defined as the time between successive arrivals (interarrival time) follows the *exponential distribution*.
- Let T be the random variable "time between successive arrivals"; then

$$\Pr\{T \geq t\} = \Pr\{\text{no arrivals in time } t\} = p_0(t) = e^{-\lambda t}.$$

- Therefore we see that the cumulative distribution function of T can be written as $A(t) = \Pr\{T \leq t\} = 1 - e^{-\lambda t}$, with corresponding density function

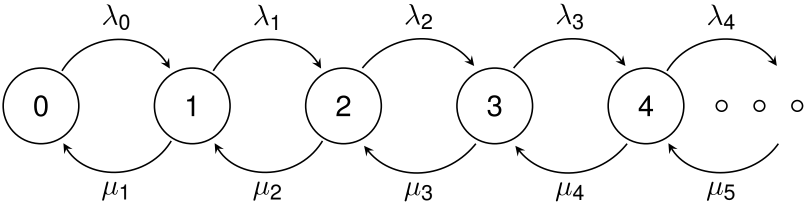
$$a(t) = \frac{dA(t)}{dt} = \lambda e^{-\lambda t}.$$

Thus T has the exponential distribution with mean $1/\lambda$.

Chapter 2. Simple Markovian Queueing Models

- The purpose of this chapter is to develop a broad class of simple queueing models using the theory of birth-death processes.
- The term "**Modeling**":
  - Definition: "Use mathematical symbols, notations, and tools to describe your system." said Jenhui Chen.
- A birth-death process is a specific type of continuous-time Markov chain, the structure of which leads to a straightforward solution for the steady-state probabilities $\{p_n\}$.
- Examples of queues that can be modeled as birth-death processes are $M/M/1$, $M/M/c$, $M/M/c/K$, $M/M/c/c$, $M/M/\infty$, and variations of these queues with state-dependent arrival and service rates.

Birth-Death Process (1/5)



- $0 = -(\lambda_n + \mu_n)p_n + \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1} \quad (n \geq 1)$
- $0 = -\lambda_0 p_0 + \mu_1 p_1,$
- or
- $(\lambda_n + \mu_n)p_n = \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1} \quad (n \geq 1)$
- $\lambda_0 p_0 = \mu_1 p_1.$

Birth-Death Process (2/5)

- Find a solution for (2.1) we first rewrite the equations as

$$p_{n+1} = \frac{\lambda_n + \mu_n}{\mu_{n+1}} p_n - \frac{\lambda_{n-1}}{\mu_{n+1}} p_{n-1} \qquad (n \geq 1)$$

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

- It follows that

$$p_2 = \frac{\lambda_1 + \mu_1}{\mu_2} p_1 - \frac{\lambda_0}{\mu_2} p_0 = \frac{\lambda_1 + \mu_1}{\mu_2} \frac{\lambda_0}{\mu_1} p_0 - \frac{\lambda_0}{\mu_2} p_0 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0$$

$$p_3 = \frac{\lambda_2 + \mu_2}{\mu_3} p_2 - \frac{\lambda_1}{\mu_3} p_1 = \frac{\lambda_2 + \mu_2}{\mu_3} \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0 - \frac{\lambda_1 \lambda_0}{\mu_3 \mu_2} p_0 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} p_0$$

Birth-Death Process (3/5)

- The pattern that appears to be emerging is that

$$p_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1}p_0 \quad (n \geq 1)$$

$$= p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \qquad (2.3)$$

- Apply mathematical induction on (2.3). First, (2.3) is correct for $n = 0$, because $\prod_{i=1}^{n}(\cdot)$ is assumed by default to be 1 when $n = 0$.
- We have shown that (2.3) is correct for $n = 1, 2, 3$
- Assume $n = k$ is also correct

$$p_k = p_0 \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i}$$

Birth-Death Process (4/5)

- Then we have to prove that it is also correct for $n = k + 1$.

$$
\begin{aligned}
p_{k+1} &= \frac{\lambda_k + \mu_k}{\mu_{k+1}} p_k - \frac{\lambda_{k-1}}{\mu_{k+1}} p_{k-1} \\
&= \frac{\lambda_k + \mu_k}{\mu_{k+1}} p_0 \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i} - \frac{\lambda_{k-1}}{\mu_{k+1}} p_0 \prod_{i=1}^{k-1} \frac{\lambda_{i-1}}{\mu_i} \\
&= \frac{p_0 \lambda_k}{\mu_{k+1}} \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i} + \frac{p_0 \mu_k}{\mu_{k+1}} \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i} - \frac{p_0 \mu_k}{\mu_{k+1}} \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i} \\
&= p_0 \prod_{i=1}^{k+1} \frac{\lambda_{i-1}}{\mu_i}
\end{aligned}
$$

- The induction proof is complete.

Birth-Death Process (5/5)

- Since probabilities must sum to 1, it follows that

$$p_0 = \left(1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}\right)^{-1} \tag{2.4}$$

Birth-Death Process (5/5)

- Since probabilities must sum to 1, it follows that

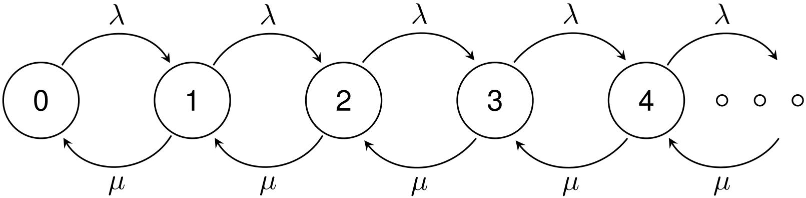
$$p_0 = \left(1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}\right)^{-1} \tag{2.4}$$

- Hint: Since

$$\sum_{n=0}^{\infty} p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} = 1$$

$$p_0 \cdot 1 + p_0 \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} = 1$$

Single-Server Queue $M/M/1$ (1/5)



- Interarrival times and service times, are assumed to be exponentially distributed with density function given, respectively, as

$$a(t) = \lambda e^{-\lambda t},$$
$$b(t) = \mu e^{-\mu t}.$$

Single-Server Queue $M/M/1$ (2/5)

- Let n denote the number of customers in the system. Arrivals can be considered as "births" to the system, and departures can be considered as "deaths." The rate of arrivals λ is fixed, regardless of the number in the system.
- The rate of the server μ is fixed, regardless of the number in the system (provided there is at least one customer in the system).

$$(\lambda + \mu)p_n = \mu p_{n+1} + \lambda p_{n-1} \qquad (n \geq 1),$$
$$\lambda p_0 = \mu p_1.$$

- Alternatively, these can be written as

$$p_{n+1} = \frac{\lambda + \mu}{\mu} p_n - \frac{\lambda}{\mu} p_{n-1} \qquad (n \geq 1),$$
$$p_1 = \frac{\lambda}{\mu} p_0.$$

- Since the $M/M/1$ system is a birth-death process with constant birth and death rates, it follows that

$$p_n = p_0 \prod_{i=1}^{n} \left(\frac{\lambda}{\mu}\right) = p_0 \left(\frac{\lambda}{\mu}\right)^n \qquad n \geq 1$$

- To get p_0, $1 = \sum_{n=0}^{\infty} p_n = p_0 \sum_{n=0}^{\infty} \rho^n \Longrightarrow p_0 = \frac{1}{\sum_{n=0}^{\infty} \rho^n}$.
- Now, $\sum_{n=0}^{\infty} \rho^n$ is a geometric series that converges iff $\rho < 1$. Then, we have

$$\sum_{n=0}^{\infty} \rho^n = \frac{1}{1-\rho} \qquad (\rho < 1),$$

which implies that $p_0 = 1 - \rho \quad (\rho = \lambda/\mu < 1)$.
- Then

$$p_n = (1-\rho)\rho^n \qquad (\rho = \lambda/\mu < 1).$$

Measures of Effectiveness (3/5)

- The steady-state probability distribution for the system size allows us to calculate the system's measures of effectiveness.

$$L = E[N] = \sum_{n=0}^{\infty} np_n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n.$$

- Consider the summation

$$\begin{aligned}\sum_{n=0}^{\infty} n\rho^n &= \rho + 2\rho^2 + 3\rho^3 + \cdots \\ &= \rho(1 + 2\rho + 3\rho^2 + \cdots) \\ &= \rho\sum_{n=1}^{\infty} n\rho^{n-1}.\end{aligned}$$

- Since $\sum_{n=0}^{\infty} \rho^n = 1/(1-\rho)$.

- Since $\sum_{n=0}^{\infty} \rho^n = 1/(1-\rho)$; hence

$$\sum_{n=1}^{\infty} n\rho^{n-1} = 1 + 2\rho + 3\rho^2 + \cdots = \frac{1}{(1-\rho)^2}.$$

- So the expected number in the system at steady state is then

$$L = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = (1-\rho)\rho\sum_{n=1}^{\infty} n\rho^{n-1} = \frac{\rho(1-\rho)}{(1-\rho)^2},$$

or simply

$$L = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}.$$

and

$$L_q = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}.$$

- From L and L_q by using Little's formulas, $L = \lambda W$ and $L_q = \lambda W_q$.

$$W = \frac{L}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu - \lambda}$$

and

$$W_q = \frac{L_q}{\lambda} = \frac{\rho}{\mu(1-\rho)} = \frac{\rho}{\mu - \lambda}.$$

Multiserver Queues $M/M/c$ (1/2)

- The steady-state probabilities p_n:

$$p_n = \begin{cases} \frac{\lambda^n}{n!\mu^n} p_0, & (0 \le n < c), \\ \frac{\lambda^n}{c^{n-c} c! \mu^n} p_0, & (n \ge c). \end{cases}$$

- The initial probability p_0 is

$$p_0 = \left(\frac{r^c}{c!(1-\rho)} + \sum_{n=0}^{c-1} \frac{r^n}{n!} \right)^{-1} \qquad (r/c = \rho < 1).$$

$$L_q = \left(\frac{r^c \rho}{c!(1-\rho)^2} \right) p_0.$$

- To find L, we employ Little's formula to get W_q, then use W_q to find $W = W_q + 1/\mu$, and finally employ Little's formula again to calculate $L = \lambda W$. Thus we get

$$W_q = \frac{L_q}{\lambda} = \left(\frac{r^c}{c!(c\mu)(1-\rho)^2} \right) p_0,$$

$$W = \frac{1}{\mu} + \left(\frac{r^c}{c!(c\mu)(1-\rho)^2} \right) p_0,$$

and

$$L = r + \left(\frac{r^c \rho}{c!(1-\rho)^2} \right) p_0,$$

- Equivalently, the probability that an arriving customer has a *nonzero* wait in queue is

$$C(c, r) \equiv 1 - W_q(0) = \frac{r^c}{c!(1-\rho)} \Bigg/ \left(\frac{r^c}{c!(1-\rho)} + \sum_{n=0}^{c-1} \frac{r^n}{n!} \right).$$

Multiserver Queues $M/M/c$ Example

- Calls to a technical support center arrive according to a Poisson process with rate 30 per hour. The time for a support person to serve one customer is exponentially distributed with a mean of 5 minutes. The support center has 3 technical staff to assist callers. What is the probability that a customer is able to immediately access a support staff, without being delayed on hold? (Assume that customers do not abandon their calls.)
- For this problem, $\lambda = 30$, $\mu = 12$, and $c = 3$. The $r = 2.5$ and $\rho = 5/6$. From (2.38)

$$C(c,r) = \frac{2.5^3}{3!(1-5/6)} / \left(\frac{2.5^3}{3!(1-5/6)} + 1 + \frac{2.5}{1!} + \frac{2.5^2}{2!}\right) \doteq 0.702.$$

  The answer is $1 - C(c,r) = 1 - 0.702 = 0.298$
- Now suppose that the call center wishes to increase the probability of nondelayed calls to 90%. How many servers are needed?