Systems Simulation
Chapter 6: Queuing Models

Fatih Cavdur
fatihcavdur@uludag.edu.tr

April 2, 2014

Introduction

- Simulation is often used in the analysis of queuing models.
- A simple but typical model is the single-server queue system.
- In this model, the term "customer" refers to any type of entity that can be viewed as requesting "service" from a system.
- Some examples are production systems, repair and maintenance facilities, communications and computer systems, transport and material-handling systems etc.

Introduction-cont.

- Queuing models, whether solved analytically or through simulation, provide the analyst with a powerful tool for designing and evaluating the performance of queuing systems.
- Typical measures of system performance are server utilization, length of waiting lines and delays of customers.
- Quite often, the analyst or the decision maker is involved in tradeoffs between server utilization and customer satisfaction in terms of line lengths and delays.

Introduction-cont.

- For relatively simple systems, these performance measures can be computed mathematically - at great savings in time and expense as compared with the use of a simulation model - but, for for realistic models of complex systems, simulation is usually required.
- Nevertheless, analytically tractable models, although usually requiring many simplifying assumptions, are valuable for rough-cut estimates of system performance.
- This chapter will discuss some of the well-known models.

Characteristics of Queuing Systems

- The key elements of queuing systems are the customers and servers.
- The term "customer" can refer to people, parts, trucks, e-mails etc. and the term "server" clerks, mechanics, repairmen, CPUs etc.
- Although the terminology employed will be that of a customer arriving to a server, sometimes the server moves to the customer; for example a repairman moving to a broken machine.

Characteristics of Queuing Systems-cont.

Table : Examples of Queuing Systems

System	*Customers*	*Servers*
Reception Desk	People	Receptionist
Repair Facility	Machines	Repairman
Garage	Trucks	Mechanic
Hospital	Patients	Nurses
Grocery	Shoppers	Checkout Station
Laundry	Dirty Linen	Washing Machines / Dryers
Job Shop	Jobs	Machines, Workers
Computer	Jobs	CPU, Disk, CDs
Telephone	Calls	Exchange

The Calling Population

- The population of potential customers, referred to as the *calling population*, may be assumed to be finite or infinite.
- In systems with large population of potential customers, the calling population is usually assumed to be infinite.
- The main difference between finite and infinite population models is how the arrival rate is defined. In an infinite population model, the arrival rate is not affected by the number of customers who left the calling population and joined the queuing system.

System Capacity

- In many queuing systems, there is a limit to the number of customers that may be in the waiting line or system, such as an automatic car wash.
- In such systems, an arriving customer who finds the system full does not enter and returns to the calling population.
- Some systems may be considered as having unlimited capacity, such as concert ticket sales for students.
- In limited-capacity systems, there is a distinction between the arrival rate and the effective arrival rate.

The Arrival Process

- The arrival process for infinite-population models is usually characterized in terms of inter-arrival times of successive customers.
- Arrivals may occur at scheduled times of at random times.
- The most important model for random arrivals is the Poisson arrival process. If A_n represents the inter-arrival time between customer $n-1$ and customer n, then, for a Poisson arrival process, A_n is exponentially distributed with mean $1/\lambda$ where the arrival rate of the process is λ customers per time unit, and the number of arrivals in a time interval of length t, say $N(t)$, has the Poisson distribution with mean λt customers.

The Arrival Process

Memoryless Property of Exponential Distribution

If X is an exponentially distributed random variable, we can write

$$P\{X > t+h | x > t\} = P\{X > h\}$$

$$\begin{aligned} P\{X > t+h | x > t\} &= \frac{P\{X > t+h, X > t\}}{P\{X > t\}} \\ &= \frac{\int_{t+h}^{\infty} \lambda e^{-\lambda x} dx}{\int_{h}^{\infty} \lambda e^{-\lambda x} dx} \\ &= \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} \\ &= e^{-\lambda h} = P\{X > h\} \end{aligned}$$

The Arrival Process-cont.

Exponential Distribution & Poisson Process

Exponential Distribution:

$$f(x) = \lambda e^{-\lambda x};\ x \geq 0$$

$$E(x) = \frac{1}{\lambda} \text{ and } V(x) = \frac{1}{\lambda^2}$$

Poisson Process:

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!};\ x \geq 0$$

$$E(x) = \lambda \text{ and } V(x) = \lambda$$

The Arrival Process-cont.

Exponential Distribution & Poisson Process

Example: The number of orders per hour follows a Poisson distribution with an average of 30.

- Find the probability that there are exactly 60 orders between 10 PM and 12 AM.

$$P\{X = 60\} = \frac{e^{-\lambda t}(\lambda t)^x}{x!} = \frac{e^{-60}60^{60}}{60!} = 0.051$$

- Find the mean and standard deviation of the number of orders between 9 PM and 1 AM.

$$E(X) = \lambda t = 4 \times 30 = 120 \text{ and } V(X) = \sqrt{120} = 10.95$$

The Arrival Process-cont.

Exponential Distribution & Poisson Process

- Find the probability that the time between two consecutive orders is between 1 and 3 minutes.

Let X be the time in minutes between successive orders. The mean number of orders per minute is exponential with rate 0.5.

$$P\{1 \leq X \leq 3\} = \int_1^3 \lambda e^{-\lambda x dx} = \int_1^3 0.5 e^{-0.5x} dx = 0.38$$

Systems Simulation Chapter 6: Queuing Models
└Characteristics of Queuing Systems

The Arrival Process-cont.

- A second important class of arrivals is scheduled arrivals, such as scheduled airline flight arrivals to an airport.
- In this case, $[A_n, n = 1, 2, \ldots]$ could be either *constant* or *constant plus or minus a small random amount* to represent early or late arrivals.
- A third situation occurs when at least one customer is assumed to always be present in the queue, so that the server is never idle because of a lack of customers.

The Arrival Process-cont.

- For finite-population models, the arrival process is characterized in a completely different fashion.
- Define a customer *pending* when that customer is outside the queuing system and a member of the potential calling population.
- Define a *runtime* of a given customer as the length of time from departure from the queuing system until that customer's next arrival to the queue.
- One important application of finite-population models is the machine-repair problem. Times to failure for a given class of machine have been characterized by the exponential, the Weibull and the gamma distributions.

The Arrival Process-cont.

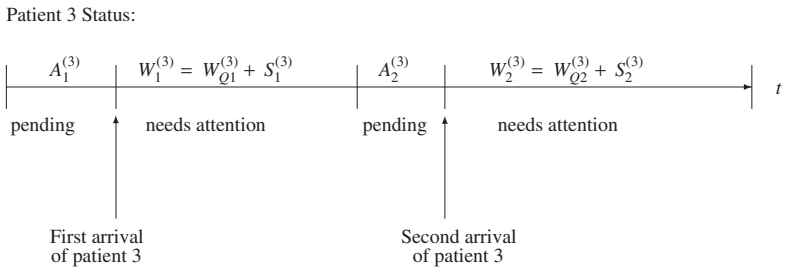

Figure : Arrival Process for a Finite-Population Model

Queue Behavior and Queue Discipline

- Queue behavior refers to the actions of customers while in a queue waiting for service to begin.
- In some situations, customers may balk (leaving when queue is too long), renege (leaving after waiting for a while when queue is moving too slowly) or jockey (moving from one queue to another).
- Queue discipline refers to the logical ordering of customers in a queue and determines which customer will be chosen for service when a server becomes free. Some examples are FIFO, LIFO, SIRO, SPT etc.

Service Times and Service Mechanism

- The service times of successive arrivals are denoted by $S_1, S_2, \ldots$ They may be constant or of random duration.
- In the latter case, $\{S_1, S_2, \ldots\}$ is usually characterized as a sequence of independent and identically distributed (IID) random variables. The exponential, Weibull, gamma, lognormal and truncated normal distributions have all been used successfully as models of service times.
- A queuing system consists of a number of service centers and interconnecting queues. Each service center consists of some number of servers, c, working in parallel, such as single-server ($c = 1$), multiple-server ($1 < c < \infty$) or unlimited servers ($c = \infty$).

Example 6.1

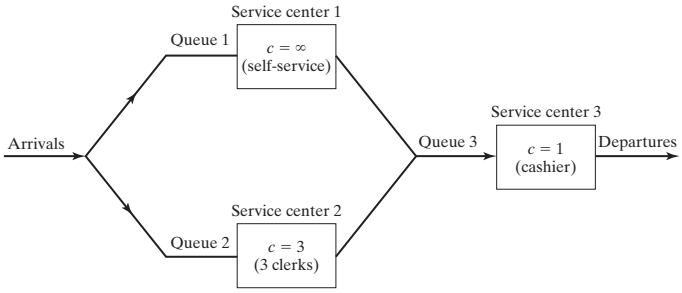

Figure : Example 6.1: Discount Warehouse with Three Service Centers

Example 6.1

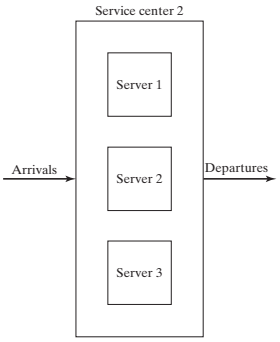

Figure : Service Center 2, with $c = 3$ parallel servers

Example 6.2

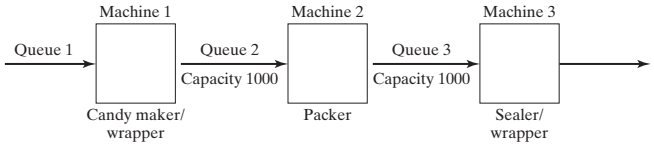

Figure : Candy-Production Line

Queuing Notation

- Based on Kendall's proposition, queuing systems can be characterized as $A/B/c/N/K$, (Kendall-Lee Notation) where
  - A represents the inter-arrival-time distribution
  - B represents the service-time distribution
  - c represents the number of parallel servers
  - N represents the system capacity
  - K represents the size of the calling population

Queuing Notation

- Common symbols for A and B include M (exponential or Markovian), D (constant or deterministic), E_k (Erlang of order k), PH (phase-type), H (hyper-exponential), G (arbitrary or general), and GI (general independent).
- For example, $M/M/1/\infty/\infty$ indicates a single-server system that has unlimited capacity and an infinite population of potential arrivals, and inter-arrival and service times are exponential. When N and K are infinite, they may be dropped from the notation (i.e., $M/M/1$)
- The tire cutting example in the text can be represented by $G/G/1/5/5$.

Queuing Notation-cont.

Table : Queuing Notation for Parallel Server Systems

P_n	steady-state probability of having n customers in system
$P_n(t)$	probability of having n customers in system at time t
λ	arrival rate
λ_e	effective arrival rate
μ	service rate of one server
ρ	server utilization
A_n	inter-arrival time between customers $n-1$ and n
S_n	service time of the nth arriving customer
W_n	total time spent in system by the nth arriving customer
W_n^Q	total time spent in queue by the nth arriving customers
$L(t)$	the number of customers in system at time t
$L_Q(t)$	the number of customers in queue at time t
L	long-run time-average number of customers in system
L_Q	long-run time-average number of customers in queue
w	long-run average time spent in system per customer
w_Q	long-run average time spent in queue per customer

Long-Run Measures of Performance of Queuing Systems

Primary Long-Run Measures of Performance of Queuing Systems

- long-run time-average number of customers in system (L)
- long-run time-average number of customers in queue (L_Q)
- long-run average time spent in system per customer (w)
- long-run average time spent in queue per customer(w_Q)
- server utilization (ρ)

Long-Run Measures of Performance of Queuing Systems

- This section defines the major measures of performance for a general $G/G/c/N/K$ queuing system, discusses their relationships and shows how they can be estimated from a simulation run.
- There are two types of estimators: (i) an ordinary sample average, and (ii) a time-integrated (time-weighted) sample average.

Long-Run Measures of Performance of Queuing Systems

Time-Average Number in System L

- Consider a queuing system over a period of time T, and let $L(t)$ denote the number of customers in the system at time t.
- Let T_i denote the total time during $[0, T]$ in which the system contained exactly i customers. The time-weighted-average number in system is defined by

$$\widehat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = \sum_{i=0}^{\infty} i\left(\frac{T_i}{T}\right)$$

Long-Run Measures of Performance of Queuing Systems

Time-Average Number in System L

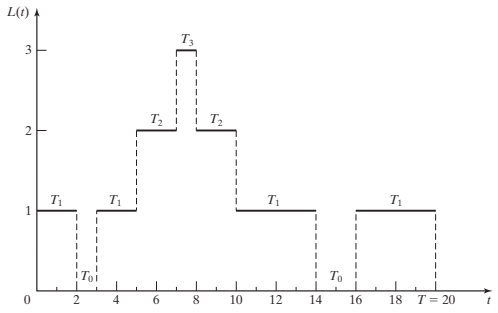

Figure : Number in System, $L(t)$ at time t

Long-Run Measures of Performance of Queuing Systems

Time-Average Number in System L

- For the example problem in the previous slide, $\widehat{L} = [0(3) + 1(12) + 2(4) + 3(1)]/20 = 1.15$ customers.
- It can be seen that the total area under the function $L(t)$ can be decomposed into rectangles of height i and length T_i. It follows that the total area is given by $\sum_{i=0}^{\infty} iT_i = \int_0^T L(t)dt$, and hence,

$$\widehat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = \frac{1}{T}\int_0^T L(t)dt$$

Long-Run Measures of Performance of Queuing Systems

Time-Average Number in System / Queue L

- If we let L be the long-run time-average number in system,

$$\lim_{T\to\infty} \widehat{L} = \lim_{T\to\infty} \frac{1}{T}\int_0^T L(t)dt = L$$

- The above can be applied to any sub-system of a queuing system. If we let L_Q denote the number of customers in line,

$$\widehat{L}_Q = \frac{1}{T}\sum_{i=0}^{\infty} iT_i^Q = \frac{1}{T}\int_0^T L_Q(t)dt \to L_Q \text{ as } T \to \infty$$

Long-Run Measures of Performance of Queuing Systems

Time-Average Number in Queue L

- If the previous figure corresponds to a single-server queue-that is a $G/G/1/N/K$ queuing system ($N \geq 3, K \geq 3$). Then the number of customers waiting in queue is given by $L_Q(t)$

$$\widehat{L}_Q(t) = \begin{cases} 0 & \text{if } L(t) = 0 \\ L(t) - 1 & \text{if } L(t) \geq 1 \end{cases}$$

- It is shown in the next figure. Thus, $T_0^Q = 5 + 10 = 15$, $T_1^Q = 2 + 2 = 4$ and $T_2^Q = 1$. Therefore,

$$\widehat{L}_Q = \frac{(0)(15) + (1)(4) + (2)(1)}{20} = 0.3 \text{ customers}$$

Systems Simulation Chapter 6: Queuing Models
Long-Run Measures of Performance of Queuing Systems
Time-Average Number in System

Long-Run Measures of Performance of Queuing Systems

Time-Average Number in Queue L

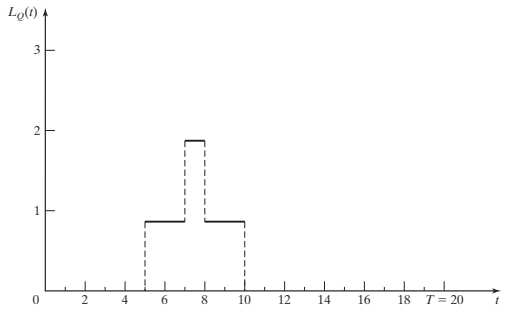

Figure : Number in Queue, $L_Q(t)$ at time t

Long-Run Measures of Performance of Queuing Systems

Average Time Spent in System per Customer w

- If $W_1, W_2, \ldots, W_n$ are the times each customer spends in system during $[0, T]$, where N is the number of arrivals during that time period, the average time spent in system per customer (average system time) is

$$\widehat{w} = \frac{1}{N} \sum_{i=1}^{N} W_i$$

- For stable systems, as $N \to \infty$, $\widehat{w} \to w$. Also,

$$\widehat{w}_Q = \frac{1}{N} \sum_{i=1}^{N} W_i^Q \to \text{ as } N \to \infty$$

Long-Run Measures of Performance of Queuing Systems

Average Time Spent in System per Customer w

- For stable systems, as $N \to \infty$, $\widehat{w} \to w$ with probability 1, where w is called the long-run average system time. Also,

$$\widehat{w}_Q = \frac{1}{N} \sum_{i=1}^{N} W_i^Q \to \text{ as } N \to \infty$$

Long-Run Measures of Performance of Queuing Systems

The Conservation Equation $L = \lambda w$

- For the example system considered previously, there were $N = 5$ arrivals in $T = 20$ time units, and thus, the observed arrival rate was $\widehat{\lambda} = N/T = 1/4$ customers per time unit. Recall that $\widehat{L} = 1.15$ and $\widehat{w} = 4.6$; hence, it follows that

$$\widehat{L} = \widehat{\lambda}\widehat{w}$$

- This is not coincidence. So, if $T \to \infty$ and $N \to \infty$, the above relationship becomes

$$L = \lambda w$$

Long-Run Measures of Performance of Queuing Systems

Server Utilization

- Server utilization is defined as the proportion of time that a server is busy. We have, the observed utilization, $\widehat{\rho}$, defined over $[0, T]$, $\widehat{\rho} \to \rho$ as $T \to \infty$.
- Server Utilization in $G/G/1/\infty/\infty$ Queues
- Server Utilization in $G/G/c/\infty/\infty$ Queues

Long-Run Measures of Performance of Queuing Systems

Server Utilization Example

- From the figure in the next slide and the one we looked for the previous example, and assuming a single server, the server utilization is

$$\widehat{\rho} = \frac{\text{total busy time}}{T} = \frac{\sum_{i=1}^{\infty} T_i}{T} = \frac{T - T_0}{T} = \frac{17}{20}$$

Long-Run Measures of Performance of Queuing Systems

Server Utilization Example

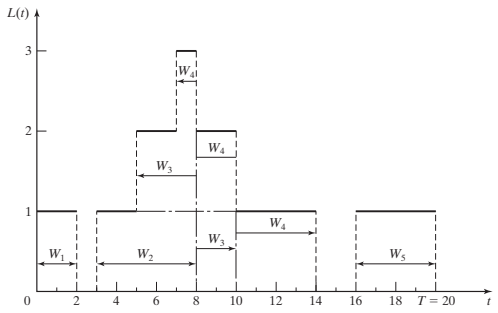

Figure : System Times, W_i for Single-Server FIFO Queuing System

Long-Run Measures of Performance of Queuing Systems

Server Utilization in $G/G/1/\infty/\infty$ Queues

- Let λ customers be the average arrival rate per time unit and $E(S) = 1/\mu$ average service time (when busy, server is working at the rate μ customers per time unit.
- The server alone can be considered as a queuing system itself. If we let λ_s be the average arrival rate to the server, we have, $\lambda_s = \lambda$ for stable systems. If $\lambda_s < \lambda$, then, the queue length tend to increase at an average rate of $\lambda - \lambda_s$ customers per time unit.

Long-Run Measures of Performance of Queuing Systems

Server Utilization in $G/G/1/\infty/\infty$ Queues

- For the server sub-system, we have $w = E(S) = \mu^{-1}$ Hence, the average number in system is

$$\widehat{L}_s = \frac{1}{T}\int_0^T [L(t) - L_Q(t)]dt = \frac{T - T_0}{T}$$

- In this case, $\widehat{L}_s = \widehat{\rho}$. As $T \to \infty$, $\widehat{L}_s = \widehat{\rho} \to L_s = \rho$. Combining this with $L = \lambda w$, we have

$$\rho = \lambda E(S) = \frac{\lambda}{\mu}$$

Long-Run Measures of Performance of Queuing Systems

Server Utilization in $G/G/1/\infty/\infty$ Queues

- That is, the long-run server utilization in a single-server queue is equal to the average arrival rate divided by the average service rate. We should have $\lambda < \mu$ or $\rho = \lambda/\mu < 1$ for this system to be stable.
- For unstable systems ($\lambda > \mu$), the long-run server utilization is 1 and the long-run queue length is infinity.
- For stable systems, the long-run performance measures, such as queue length, number of customers in system, system time, queue time, are well defined.

Long-Run Measures of Performance of Queuing Systems

Server Utilization in $G/G/c/\infty/\infty$ Queues

- The average number of busy servers is given by

$$L_s = \lambda E(S) = \frac{\lambda}{\mu}, \quad 0 \leq L_s \leq c$$

- The long run server utilization is

$$\rho = \frac{L_s}{c} = \frac{\lambda}{c\mu}, \quad 0 \leq \rho \leq 1$$

- The system is stable iif $\lambda < c\mu$.

Long-Run Measures of Performance of Queuing Systems

Server Utilization in $G/G/c/\infty/\infty$ Queues

- Customers arrive at random to a license bureau at a rate of $\lambda = 50$ customers per hour. Currently, there are 20 clerks, each serving $\mu = 5$ customers per hour on the average. Here, the long-run or steady-state average number of busy servers is

$$L_s = \frac{\lambda}{\mu} = \frac{50}{5} = 10$$

- We note that $c > \lambda/\mu$. The long run server utilization is

$$\rho = \frac{\lambda}{c\mu} = \frac{50}{(20)(5)} = 0.5$$

Long-Run Measures of Performance of Queuing Systems

Server Utilization and System Performance

- System performance can vary widely for a given utilization, ρ.
- Consider a $D/D/1$ queue. We have all arrival times are equal to $E(A) = 1/\lambda$, and all service times equal to $E(S) = 1/\mu$.
- By looking at the figure in the next slide, we see that by varying λ and μ, we can obtain any utilization between 0 and 1, and yet, there are no waiting customers.
- What causes the long queues if not a high server utilization?

Long-Run Measures of Performance of Queuing Systems

Server Utilization and System Performance

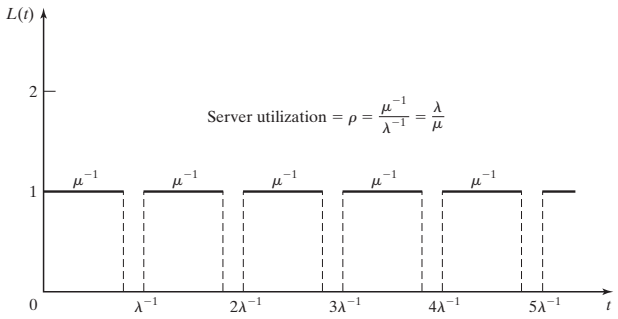

Figure : $D/D/1$ Queue

Cost in Queuing Problems

- We can associate costs with various aspects of the queues and servers. For example, if a cost occurs for each customer in queue, say at a rate of \$ 10 per hour per customer, then, if customer j spends W_j^Q hours in the queue, $\sum_{j=1}^{N}(10W_j^Q)$ is the total cost of the N customers. Thus, the average cost per customer is

$$\sum_{j=1}^{N}\frac{10W_j^Q}{N} = 10\widehat{w}_Q$$

- We can proceed using this.

Steady-State Behavior of Infinite-Population Markovian Models

- A queuing system is said to be in statistical equilibrium or steady state, if the probability that the system is in a given state is not time dependent-that is, $P[L(t) = n] = P_n(t) = P_n$ is independent of time t. For the simple models studied here, the steady-state parameter L, the time-average number of customers in the system, can be computed as

$$L = \sum_{n=0}^{\infty} nP_n$$

where $[P_n]$ are the steady-state probabilities of finding n customers in the system.

Steady-State Behavior of Infinite-Population Markovian Models

- Once L is given, the other steady-state parameters can be computed from Little's equation applied to the whole system and to the queue alone.

$$w = \frac{L}{\lambda}$$

$$w_Q = w - \frac{1}{\mu}$$

$$L_Q = \lambda w_Q$$

Single-Server Queues (Poisson Arrivals & Infinite Capacity)

Table : Steady-State Parameters of the $M/G/1$ Queue

ρ	$\frac{\lambda}{\mu}$
L	$\rho + \frac{\lambda^2(1/\mu^2+\sigma^2)}{2(1-\rho)} = \rho + \frac{\lambda^2(1+\sigma^2\mu^2)}{2(1-\rho)}$
w	$\frac{1}{\mu} + \frac{\lambda(1/\mu^2+\sigma^2)}{2(1-\rho)}$
w_Q	$\frac{\lambda(1/\mu^2+\sigma^2)}{2(1-\rho)}$
L_Q	$\frac{\lambda^2(1/\mu^2+\sigma^2)}{2(1-\rho)} = \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$
P_0	$1-\rho$

Single-Server Queues (Poisson Arrivals & Infinite Capacity)

Example

There are two workers competing for a job; Able and Baker. Able claims an average service time that is faster than Baker's, but Baker claims to be more consistent, even if not as fast. The arrivals occur according to a Poisson process with a rate of $\lambda = 2$ per hour. Able's statistics are an average service time of 24 minutes with a standard deviation of 20 minutes. Baker's statistics are an average service time of 25 minutes with a standard deviation of 2 minutes. If the average queue length is the criterion for hiring, which worker should be hired?

Single-Server Queues (Poisson Arrivals & Infinite Capacity)

Example

For Able,

$$L_Q = \frac{\lambda^2(1/\mu^2 + \sigma^2)}{2(1-\rho)} = \frac{\left(\frac{1}{30}\right)^2 (24^2 + 400)}{2\left(1 - \frac{4}{5}\right)} = 2.711 \text{ customers}$$

For Baker,

$$L_Q = \frac{\lambda^2(1/\mu^2 + \sigma^2)}{2(1-\rho)} = \frac{\left(\frac{1}{30}\right)^2 (25^2 + 4)}{2\left(1 - \frac{5}{6}\right)} = 2.097 \text{ customers}$$

Baker should be hired!

Single-Server Queues (Poisson Arrivals & Infinite Capacity)

Table : Steady-State Parameters of the $M/M/1$ Queue

L	$\frac{\lambda}{\mu-\lambda} = \frac{\rho}{1-\rho}$
w	$\frac{1}{\mu-\lambda} = \frac{1}{\mu(1-\rho)}$
w_Q	$\frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu(1-\rho)}$
L_Q	$\frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{\mu(1-\rho)}$
P_n	$\left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n = (1-\rho)\rho^n$

Single-Server Queues (Poisson Arrivals & Infinite Capacity)

Example

Assume that time between arrivals and service times at a single-chair unisex hair-styling shop have been shown to be exponentially distributed. The values of λ and μ are 2 per hour and 3 per hour, respectively. Compute the server utilization and the probabilities of having 0, 1, 2 and 3 or more customers in the shop. The server utilization is

$$\rho = \frac{\lambda}{\mu} = \frac{2}{3}$$

Single-Server Queues (Poisson Arrivals & Infinite Capacity)

Example

The probabilities of having 0, 1, 2 and 3 or more customers

$$P_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \Rightarrow P_0 = \left(1 - \frac{2}{3}\right)\left(\frac{2}{3}\right)^0 = \frac{1}{3}$$

$$P_1 = \left(1 - \frac{2}{3}\right)\left(\frac{2}{3}\right)^1 = \frac{2}{9}$$

$$P_2 = \left(1 - \frac{2}{3}\right)\left(\frac{2}{3}\right)^2 = \frac{4}{27}$$

$$P\{n \geq 3\} = 1 - \sum_{n=0}^{2} P_n = \frac{8}{27}$$

Single-Server Queues (Poisson Arrivals & Infinite Capacity)

Example

The number of customers in system, in queue and their waiting times

$$L = \frac{\lambda}{\mu - \lambda} = 2 \text{ customers}$$

$$L_Q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{4}{3} \text{ customers}$$

$$w = \frac{1}{\mu - \lambda} = 1 \text{ hour } = \frac{L}{\lambda}$$

$$w_Q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{2}{3} \text{ hour } = w - E(S) = w - \frac{1}{\mu}$$

Multi-Server Queues

Table : Steady-State Parameters of the $M/M/c$ Queue

ρ	$\frac{\lambda}{c\mu}$
P_0	$\left\{\left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!}\right] + \left[\left(\frac{\lambda}{\mu}\right)^c \left(\frac{1}{c!}\right)\left(\frac{c\mu}{c\mu-\lambda}\right)\right]\right\}^{-1}$
$P[L(\infty) \geq c]$	$\frac{(\lambda/\mu)^c P_0}{c!(1-\lambda/c\mu)} = \frac{(c\rho)^c P_0}{c!(1-\rho)}$
L	$c\rho + \frac{(c\rho)^{c+1} P_0}{cc!(1-\rho)^2}$
w	$\frac{L}{\lambda}$
w_Q	$w - \frac{1}{\lambda}$
L_Q	λw_Q
$L - L_Q$	$\frac{\lambda}{\mu} = c\rho$

Multi-Server Queues

Example

Assume Poisson arrivals with $\lambda = 2$ per minute and exponentially distributed service times with mean 40 seconds. Since

$$\frac{\lambda}{\mu} = \frac{2}{3/2} = \frac{4}{3} > 1$$

the system is unstable. We need more servers to make it stable ($c > \lambda/\mu$). We can let $c = 2$. Then,

Multi-Server Queues

Example

The probability that there are no customers in the system

$$\begin{aligned}
P_0 &= \left\{ \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} \right] + \left[\left(\frac{\lambda}{\mu} \right)^c \left(\frac{1}{c!} \right) \left(\frac{c\mu}{c\mu - \lambda} \right) \right] \right\}^{-1} \\
&= \left\{ \left[\sum_{n=0}^{1} \frac{(4/3)^n}{n!} \right] + \left[\left(\frac{4}{3} \right)^2 \left(\frac{1}{2!} \right) \left(\frac{2 \times \frac{3}{2}}{2 \times \frac{3}{2} - 2} \right) \right] \right\}^{-1} \\
&= \left[1 + \frac{4}{3} + \left(\frac{16}{9} \right) \left(\frac{1}{2} \right) (3) \right]^{-1} \\
&= \frac{1}{5}
\end{aligned}$$

Multi-Server Queues

Example

The probability that all servers are busy

$$\begin{aligned} P\{L(\infty) \geq 2\} &= \frac{(\lambda/\mu)^c P_0}{c!(1-\lambda/c\mu)} \\ &= \frac{(4/3)^2(1/5)}{2!\left[1-\frac{2}{2(3/2)}\right]} \\ &= \left(\frac{8}{3}\right)\left(\frac{1}{5}\right) \\ &= \frac{8}{15} \end{aligned}$$

Multi-Server Queues

Example

A grocery store operates with 3 check-out counters. The company uses the following schedule to determine the number of counters in operation based on the number of customers in the store:

# of customers	# of counters
1 to 3	1
4 to 6	2
more than 6	3

Customers arrive in the counters area according to a Poisson distribution with with a mean rate of 10 customers per hour and the average check-out time per customer is exponential with mean 12 minutes. Determine the steady-state probability p_n of n customers in the system.

Multi-Server Queues

Example

We let

$$\lambda_n = \lambda = 10 \text{ customers per hour }, n = 0, 1, \ldots$$

$$\mu_n = \begin{cases} 60/12 \text{ customers per hour}, & n = 0, 1, 2, 3 \\ (2)(60)/12 \text{ customers per hour}, & n = 4, 5, 6 \\ (3)(60)/12 \text{ customers per hour}, & n = 7, 8, \ldots \end{cases}$$

Continue from here.

Multi-Server Queues

Table : Steady-State Parameters of the $M/G/\infty$ Queue

P_0	$e^{-\lambda/\mu}$
w	$\frac{1}{\mu}$
w_Q	0
L	$\frac{\lambda}{\mu}$
L_Q	0
P_n	$\frac{e^{-\lambda/\mu}(\lambda/\mu)^n}{n!}, \quad n = 0, 1, \ldots$

Multi-Server Queues

Table : Steady-State Parameters of the $M/M/c/N$ Queue

P_0	$\left[1 + \sum_{n=1}^{c} \frac{a^n}{n!} + \frac{a^c}{c!} \sum_{n=c+1}^{N} \rho^{n-c}\right]^{-1}$
P_N	$\frac{a^N}{c! c^{N-c}} P_0$
L_Q	$\frac{P_0 a^c \rho}{c!(1-\rho)^2} \left[1 - \rho^{N-c} - (N-c)\rho^{N-c}(1-\rho)\right]$
λ_e	$\lambda(1 - P_N)$
w_Q	$\frac{L_Q}{\lambda_e}$
w	$w_Q + \frac{1}{\lambda}$
L	$\lambda_e w$

Birth and Death Processes

- The number of people present in a queuing system at time t is defined as the state of the system at time t.
- $P_{ij}(t)$ is the probability that there are j people in the system at time t given that there are i people at time 0.
- π_j is defined as the steady-state, long-term or equilibrium probability of state j.
- The behavior of $P_{ij}(t)$ is called as the transient behavior of the queuing system.
- A birth and death process (BDP) is a continuous time stochastic process for which the system's state at any time is a non-negative integer.

Birth and Death Processes-cont.

Laws of Motion for BDP

- With probability $\lambda_j \Delta t + o(\Delta t)$, a birth (arrival) occurs between time t and time $t + \Delta t$. A birth increases the system state by 1 to $j + 1$. The variable λ_j is called the birth (arrival) rate in state j. In most queuing systems, a birth is an arrival.
- With probability $\mu_j \Delta t + o(\Delta t)$, a death (departure) occurs between time t and time $t + \Delta t$. A death decreases the system state by 1 to $j - 1$. The variable μ_j is called the death rate in state j. In most queuing systems, a death is a departure or service completion.
- Births and deaths are independent of each other.

Birth and Death Processes-cont.

Derivation of Steady-State Probabilities for BDP

Table : Steady-State Probabilities for State being j

state (time t)	state (time $t + \Delta t$)	probability
$j - 1$	j	$P_{i,j-1}(t)(\lambda_{j-1}\Delta t + o(\Delta t))$
$j + 1$	j	$P_{i,j+1}(t)(\mu_{j+1}\Delta t + o(\Delta t))$
j	j	$P_{i,j}(t)[1 - \lambda_j \Delta t - \mu_j \Delta t - 2o(\Delta t)]$
$\forall i \neq j, j \pm 1$	j	$o(\Delta t)$

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$$

Birth and Death Processes-cont.

Derivation of Steady-State Probabilities for BDP

$$\begin{aligned} P_{i,j}(t+\Delta t) &= P_{i,j}(t) \\ &+ \Delta t[\lambda_{j-1}P_{i,j-1}(t) + \mu_{j+1}P_{i,j+1}(t) - P_{ij}(t)\mu_j - P_{ij}(t)\lambda_j] \\ &+ o(\Delta t)[P_{i,j-1}(t) + P_{i,j+1}(t) + 1 - 2P_{ij}(t)] \end{aligned}$$

Since the last row is equal to $o(\Delta t)$,

$$P_{i,j-1}(t+\Delta t) - P_{i,j}(t) = \Delta t[\lambda_{j-1}P_{i,j-1}(t) + \mu_{j+1}P_{i,j+1}(t) - P_{ij}(t)\mu_j - P_{ij}(t)\lambda_j] + o(\Delta t)$$

By dividing the last expression by Δt and letting $\Delta t \to 0$, for $\forall i, j$,

$$P'_{ij}(t) = \lambda_{j-1}P_{i,j-1}(t) + \mu_{j+1}P_{i,j+1}(t) - P_{ij}(t)\mu_j - P_{ij}(t)\lambda_j$$

$$j = 0 \Rightarrow P_{i,j-1}(t) = 0 \text{ and } \mu_j = 0 \Rightarrow P'_{i0}(t) = \mu_1 P_{i,1}(t) - \lambda_0 P_{i,0}(t)$$

Birth and Death Processes-cont.

Derivation of Steady-State Probabilities for BDP

- The last expression is an infinite system of differential equations.
- It is very difficult to solve these equations.
- However, we can use the above expressions to obtain the steady-state probabilities π_j's (or P_j's).

$$P_j = \pi_j = \lim_{t\to\infty} P_{ij}(t)$$

- For large t and any initial state i, we can thought of $P_{ij}(t)$ as a constant, and we can assume $P'_{ij}(t) = 0$.

Birth and Death Processes-cont.

Derivation of Steady-State Probabilities for BDP

We can then write, for $j = 1, 2, \ldots,$

$$\lambda_{j-1}P_{j-1} + \mu_{j+1}P_{j+1} - \lambda_j P_j - \mu_j P_j = 0$$

By organizing,

$$\lambda_{j-1}P_{j-1} + \mu_{j+1}P_{j+1} = \lambda_j P_j - \mu_j P_j = 0$$

For $j = 0$,

$$\lambda_0 P_0 = \mu_1 P_1$$

We now have an infinite system of linear equations that can be solved easily!

Birth and Death Processes-cont.

Solution of BDP Equations

$$P_j = \frac{\lambda_0\lambda_1\ldots\lambda_{j-1}}{\mu_1\mu_2\ldots\mu_j}P_0$$

$$\sum_{j=1}^{\infty} P_j = 1 \quad \Rightarrow \quad P_0\left(1 + \sum_{j=1}^{\infty}\frac{\lambda_0\lambda_1\ldots\lambda_{j-1}}{\mu_1\mu_2\ldots\mu_j}\right) = 1$$

$$\Rightarrow \quad P_0 = \frac{1}{1 + \sum_{j=1}^{\infty}\frac{\lambda_0\lambda_1\ldots\lambda_{j-1}}{\mu_1\mu_2\ldots\mu_j}}$$

Birth and Death Processes-cont.

Example

Indiana Bell customer service representatives receive 1,7000 calls per hour. The time between calls follow an exponential distribution. A customer service representative can handle an average of 30 calls per hour. The time to handle a call is also exponential. Indiana Bell can put up to 25 people on hold. If 25 people are on hold, a call is lost to the system. Indiana Bell has 75 representatives.

(a) What fraction of the time are all operators busy?

(b) What fraction of all calls are lost to the system?

Birth and Death Processes-cont.

Example-cont.

We can use MS Excel to compute the desired probabilities. [Click]

(a) The fraction of time all operators are busy is $\sum_{j=75}^{100} P_j \cong 0.013$.

(b) The fraction of all calls that are lost to the system is $P_{100} \cong 0.0000028$.

Others

- More complicated models can be considered.
  - Finite Population Models
  - Networks of Queues

Summary

- Reading HW: Chapter 6.
- Chapter 6 Exercises.