Queueing Models

Introduction

Many systems can be readily modeled as a **queueing model**. Such a model represents the system as a set of one or more customers and one or more servers. A **customer** is defined as any entity that perodically requires service, while a **server** is any entity that provides service. For example, an automated banking system can be represented as a queueing model, in which the customers are patrons who hold accounts with the bank, while the servers consist of a set of ATM machines. Perhaps a less obvious example is a system consisting of machines stationed in various locations throughout a geographical region, and one or more technicians who drive to the locations to make repairs. In this case the machines represent customers, while the technicians represent servers. Henceforth we refer to a **queueing system** as any system that admits a queueing model.

One almost universal aspect of a queueing system is that a customer who requests service at time s may have to wait until time $t > s$ to receive the service. When this occurs we say that a **delay** has occurred, and the customer is placed a **waiting line**. Such a line could have a physical realization (e.g. customers form a line to use an ATM machine) or a virtual one (e.g. the repair technician adds a broken machine to a list of machines that still need repairing). Note that *queue* and *waiting line* are often used interchangeably.

Queueing System Terminology

Calling Population the set of potential customers

System Capacity the number of customers who are allowed to be in the system at any given time.

Arrival Process a stochastic or deterministic process that provides a model of how customers arrive to the system. Examples inlcude the **Poisson process**, where the number of arriving customers during a period of time follows a Poisson distribution, and **scheduled arrivals**, where arrivals are pre-scheduled.

Pending Customer a customer who is currently outside the system

Runtime length of time from a cusotmer's departure from the system to her next arrival

Queue Behavior customer actions while in the queue, including leaving the queue, or moving to another queue

Queue Discipline determines how the next customer is chosen for servicing. Some disciplines include

- **FIFO:** first customer to arrive is the first one served
- **LIFO:** last customer to arrive is the first one served
- **SIRO:** service in random order
- **PR:** service according to priority
- **SPT:** service according to shortest processing time

Service Process a stochastic or deterministic process that provides a model for the duration of time required to serve a customer

Service Mechanism refers to the number of servers and their configuration, and method for determining how a customer is assigned to a server.

Example 1. Describe how the above queueing-system terminology applies to a walk-in restraunt that serves dinner from 6:00-10:00PM. Assume that the customers are patrons, while the servers are tables.

Example 2. Describe how the above queueing-system terminology applies to a computer-programming tutoring lab where students may receive help from one of three tutors working in the lab from 11:00AM to 4:00PM. Assume that students are the customers and the tutors are the servers. Assume that the lab has infinite capacity (i.e. the demand never reaches full capacity).

Queueing Notation

The notation $A/B/c/N/K$ can be used for succinctly describing the more important aspects of a queueing system, where

A is the interarrival distribution,

B is the service distribution,

c is the number of parallel servers,

N is the system capacity, and

K is the size of the calling population.

Note that N and K may be dropped in case they are infinite.

Common symbols for A and B

M: exponential

D: constant or deterministic

N: normal

G, GI: general, general independent

Long-Run Performance Measures for Queueing Systems

We are often interested in the performance of a queueing system after it has been in operation for a long period of time. Here a "long period" could mean years, months, days, minutes, or even seconds, depending on the system of interest. Perhaps the first issue to address is the stability of the system. We say that a queueing system is **stable** iff the customer arrival rate does not reach or exceed the service rate. Otherwise the sytem is said to be **unstable**. For a stable system, one may readily compute reliable long-run performance measures for the system. Such measures are not guaranteed for unstable systems, unless at some point they reach a point of stability (which often takes the form of shutting down altogether). The long-run behavior of a stable system is often referred to as its **steady-state behavior**.

The following includes the most important long-run performance measures of a queueing system, along with statistics that support the computing of the performance measure. For example, one such measure is L, then long-run average number of customers in system at any given time. This may be computed with the help of $L(t)$, the total number of customers in system at time t.

L long-run average number of customers in system at any given time

$L(t)$ total number of customers in system at time t

L_Q long-run average number of customers waiting at any given time

$L_Q(t)$ total number of customers waiting at time t

w long-run average time spent in the system by a customer

w_Q long-run average time spent in the queue by a customer

W_n total time spent in system by n th arriving customer

w_Q long-run average time spent in the queue by a customer

W_i^Q total time spent in queue by i th arriving customer

λ arrival rate

λ_e arrival rate (not including customers who decide not to enter the system)

A_n interarrival time between customer $n-1$ and n

μ service rate

ρ server utilization; i.e. fraction of time server is in use

S_i service time for the i th customer

P_n long-run probability of having n customers in system

$P_n(t)$ probability of having n customers in system at time t

Perhaps a concrete way of thinking about L is that, if we sampled the system at times $t_1, \ldots, t_n$, where the times are far into the future, and $t_{i+1} - t_i$ is large for all $i = 1, \ldots, n-1$, then we would expect the average

$$\frac{1}{n}[L(t_1) + \cdots + L(t_n)]$$

to approach L for increasingly larger values of n.

We now provide an estimator for L and use it as a means for defining L. Suppose a system has been in operation for T units of time, then $\hat{L}$ is defined as the average number of customers in the system during the time interval $[0, T]$, and is computed as

$$\hat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i,$$

where T_i is the duration of time for which there were i customers, for $i = 1, 2, \ldots$.

Propositiion 1. An equivalent definition for $\hat{L}$ is

$$\hat{L} = \frac{1}{T}\int_0^T L(t)dt.$$

Proof of Propositiion 1. Assume that $L(t)$ is a step function that moves up or down at discrete points of time that coincide with the arrival or departure of one or more customers. In this case the area under the graph of $L(t)$ consists of a set of rectangles. Now, let A_i, $i = 1, 2, \ldots$, denote the total area under the graph of $L(t)$ that is attributed to rectangles of height i. Moreover, let T_i denote the total elapsed time for which $L(t) = i$. Then it follows that $A_i = iT_i$. Therefore,

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

Henceforth, we define L mathematically so that

$$L = \lim_{T\to\infty}\frac{1}{T}\sum_{i=0}^{\infty} iT_i,$$

provided this limit exists.

Example 3. The following table shows both the arrival (in minutes after the hour) and service times for customers at an ATM machine. Use the table to compute $\hat{L}$

customer	**arrival** **time (min. after hr.)**	**service** **time (min.)**	**departure** **time (min. after hr.)**
al	0	5	
bo	6	2	
cat	7	2	
du	10	3	
ed	12	2	

Theorem 1 (Little's Conservation Law). Consider a queueing system that has been in operation for T units of time, and has been visited by N customers. Define

$\hat{\lambda} = N/T$ as the average arrival rate, and

$\hat{w} = \frac{1}{N}\sum\limits_{i=1}^{N} W_i$ as the average time a customer has spent in the system, where W_i denotes the time that customer i has spent in the system, $i = 1, \ldots, N$

Then $\hat{L} = \hat{\lambda}\hat{w}$.

Proof of Theorem 1. Note that all the area under graph $L(t)$ is due to the presence of customers in the system at any given time. Moreover, each customer i, $i = 1, \ldots, N$, contributes a total area of $W_i \times 1$, since, when customer i enters the system, the height of $L(t)$ increases by one unit for W_i units of time, upon which customer i departs the system. Therefore,

$$\hat{L} = \frac{1}{T}\int_0^T L(t)dt = \left(\frac{N}{T}\right)\left(\frac{1}{N}\right)\sum_{i=1}^{N} W_i = \hat{\lambda}\hat{w}.$$

Example 4. Use the table from Example 1 to verify Little's Conservation Law using estimators $\hat{L}$, $\hat{\lambda}$, and $\hat{w}$.

Corollary 1. Given the three estimators $\hat{L}$, $\hat{w}$, and $\hat{\lambda}$, if two of them converge, then so does the third.

Theorem 2. Given arrival rate λ and service rate μ, $\frac{\lambda}{\mu}$ is called the **offered load**, and is a measure of the workload imposed on the system. For a queueing system with c parallel servers, the **long-run server utilization** is defined as

$$\rho = \frac{\lambda}{c\mu}.$$

Moreover, if $\rho < 1$, then the system is stable.

Proof of Theorem 2. Since the servers work in parallel, the net service rate of the system is $c\mu > \lambda$, since $\rho = \lambda/c\mu < 1$. Alternatively, one could imagine each server having its own queue, in which case the arrival rate to each queue would be $\lambda/c < \mu$, implying that each server and its respective queues is stable, and hence the entire system is stable.

Example 5. Customers arrive at a school financial-aid office at a rate of $\lambda = 50$ students per hour. If any financial-aid officer can serve students at a rate of $\mu = 13$ students per hour, what is the minimum number of officers needed to stabilize the system?

Steady-State Behavior of $M/G/1$ Queues

The following Theorem is stated without proof.

Theorem 3. For a steady-state $M/G/1$ queue with arrival rate λ, service rate μ, and service variance σ^2

1. $\rho = \frac{\lambda}{\mu}$
2. $L = \rho + \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$
3. $w = \frac{1}{\mu} + \frac{\lambda(1/\mu^2+\sigma^2)}{2(1-\rho)}$
4. $w_Q = \frac{\lambda(1/\mu^2+\sigma^2)}{2(1-\rho)}$
5. $L_Q = \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$
6. $P_0 = 1 - \rho$

Example 6. Two auto mechanics, Alice and Bob, apply for a position at a repair shop where arrivals occur according to a Poisson process at a rate of $\lambda = 1$ per hour. Alice can make a repair in an average time of 35 minutes with a standard deviation of 15 minutes, while Bob can make a repair in an average time of 39 minutes with a standard deviation of 10 minutes. If the average length of the queue is the criterion for hiring, who gets offered the job?

Steady-State Behavior of $M/M/1$ Queues

The following Theorem is stated without proof.

Theorem 4. For a steady-state $M/M/1$ queue with arrival rate λ and service rate μ

1. $\rho = \frac{\lambda}{\mu}$
2. $L = \frac{\lambda}{\mu-\lambda} = \frac{\rho}{1-\rho}$
3. $w = \frac{1}{\mu-\lambda} = \frac{1}{\mu(1-\rho)}$
4. $w_Q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu(1-\rho)}$
5. $L_Q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{1-\rho}$
6. $P_n = (1-\frac{\lambda}{\mu})(\frac{\lambda}{\mu})^n = (1-\rho)\rho^n$

Example 7. Same as Example 6, but now assume that Alice has an exponential service rate of 12/7 repairs per hour, while Bob has an exponential service rate of 20/13 repairs per hour.

Example 8. If arrivals are occurring at rate $\lambda = 10$ per hour, and management has a choice of two servers, one who works at rate $\mu_1 = 11$ customers per hour and the second at at rate of $\mu_2 = 12$ customers per hour; assuming an $M/M/1$ queue, compute ρ and L for each server and make a recommendation.

Steady-State Behavior of $M/M/c$ Queues

The following theorem is stated without proof.

Theorem 5. For a steady-state $M/M/c$ queue with arrival rate λ and service rate μ

1. $\rho = \frac{\lambda}{c\mu}$
2. $$\begin{aligned} P_0 &= \{[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!}] + [(\frac{\lambda}{\mu})^c(\frac{1}{c!})(\frac{c\mu}{c\mu-\lambda})]\}^{-1} \\ &= \{[\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!}] + [(c\rho)^c(\frac{1}{c!})(\frac{1}{1-\rho})]\}^{-1} \end{aligned}$$
3. $$P(L(\infty) \geq c) = \frac{(\lambda/\mu)^c P_0}{c!(1-\lambda/c\mu)} = \frac{(c\rho)^c P_0}{c!(1-\rho)}$$
4. $$L = c\rho + \frac{(c\rho)^{c+1}P_0}{c(c!)(1-\rho)^2} = c\rho + \frac{\rho P(L(\infty) \geq c)}{1-\rho}$$
5. $w = \frac{L}{\lambda}$
6. $w_Q = w - \frac{1}{\mu}$
7. $L_Q = \lambda w_Q$
8. $L - L_Q = \frac{\lambda}{\mu} = c\rho$

Example 9. During weekends at the mall, an $M/M/c$ ice-cream parlor queue has an arrival rate of $\lambda = 2$ customers per minute. Assuming a service rate of 40 seconds, what is the minimum number of servers needed for a stable queue? For this number of servers, compute ρ, L, w, w_Q, and L_Q.

Queueing Networks

A **queueing network** is a directed weighted graph $G = (V, E)$ which has the following properties.

1. There are one more nodes $n \in V$ that is a **source node**. Each of these nodes generates customers according to the nodes prescribed interarrival distribution. Each source node has zero in-degree.
2. There are one or more nodes $n \in V$ that is a **collection node**. Each of these nodes represents a final destination point for customers. Each collection node has zero out-degree.
3. Every other node $n \in V$ represents a queueing system. Such a node is referred to as a **queue** node.
4. an edge $e \in E$ leaving node m and entering node n implies that some customers leaving m have the opportunity to enter n. Moreover, each node m is accompanied with a procedure for determining which outward edge is selected by an exiting customer. Such a procedure is referred to as an **exit strategy**.

Theorem 6. Let G be a queue network which is stable (i.e. each queue node has a long-run finite expected size) with infinite calling population, and no limits on system capacity.

1. If no customers are created or destroyed in any queue node n, then the long-run departure rate out of n is equal to the long-run arrival rate into n.
2. If customers arrive to node i at rate λ_i (or depart node i at rate λ in case i is a source), and a fraction $0 \le p_{ij} \le 1$ are routed to queue j upon departure, then the arrival rate from queue i to queue j is given by $\lambda_i p_{ij}$.
3. The overall arrival rate into queue j is given by
$$\sum_i \lambda_i p_{ij},$$
4. If queue j has $c_j < \infty$ parallel servers, each working at rate u_j, then the long-run utilization of each server is
$$\rho_j = \frac{\lambda_j}{c_j \mu_j}$$

Example 10. At a driver's-license branch office, drivers arrive at a rate of 13 per hour, with exponentially distributed interarrival times. All arrivals must first be serviced at the check-in station where service times are exponentially distributed with a mean of four minutes. At this station, 10% of drivers are turned away and must leave the system. The remaining 90% proceed to the "driving test station" where the service times are exponentially distributed with a mean of thirty minutes. After the driving test, 20% of drivers will leave the system (since they failed the test), while the other 80% proceed to a station that takes photos and issues a temporary license. The service time at this station is exponentially distributed with a mean of eight minutes. After receiving her temporary license, a driver exits the system. Currently, each station has just enough servers to stabilize it. However, the branch manager has just received approval to add a new worker to this queueing network. To which station should the worker be added if the goal is to maximize a decrease in average total time that a driver spends in the network? Assume network capacity and calling population are both infinite. Provide a graphical representation of the queueing network.

Exercises

1. Consider a five mile stretch of highway consisting of a single northbound lane. The highway is modeled as a queueing system with five servers, with each server representing a one-mile stretch of the highway. Answer the following questions.

   a. Define the calling population.

   b. The **three-second rule** states that a vehicle should follow another vehicle by no less than $d = 3s$ feet, where s is the vehicle speed in feet per second (here we assume both vehicles have the same speed). If the average vehicle length is 15 feet, then estimate the entire system capacity assuming that traffic is moving at 60 miles per hour (mph), and all vehicles are obeying the three-second rule.

   c. What constitutes a waiting line for this system?

   d. Determine the average time spent in the system if traffic is moving at 60 mph.

2. Consider the VEC building as a queueing system whose servers are classrooms, and whose customers are the set of courses scheduled in the building.

   a. Determiine the runtime for CECS 552, once class ends at 7:45pm on Tuesday.

   b. What constitutes a waiting line for this system? Under what circumstances will a line have a customer?

   c. Estimate the system capacity.

   d. If courses meet for 3 hours per week, and are scheduled Monday through Thursday from 8:00AM-10:PM, and on Fridays from 8:00AM-5:PM, then use the previous exercise to estimate the maximum possible size of the calling population.

3. Give an example from everyday life of a queueing system for which customers are selected for service in random order.

4. On a stretch of road that is jammed with vehicles, vehicles are entering at a rate of 10 per minute, and each vehicle spends an average of 20 minutes on the stretch. How many vehicles on average are on the stretch at any given time. Explain.

5. The arrival and service times are provided below for the first thirteen customers entering a single-server system with a FIFO queue discipline.

   a. Assuming that at most one customer can be in service at any time, determine the departure times for each of the customers.

Customer	**Arrival Time**	**Service Time**	**Departure Time**
1	12	40	
2	31	32	
3	63	55	
4	95	48	
5	99	18	
6	154	50	
7	198	47	
8	221	18	
9	304	28	
10	346	54	
11	411	40	
12	455	72	
13	537	12	

b. Letting T denote the departure time of Customer 13, compute $\hat{L}$ by i) using an integral, and ii) using Little's Conservation Law.

6. Repeat the previous exercise assuming a *LIFO* queue discipline. Would you expect $\hat{L}$ to change under this new assumption? Explain.

7. Repeat Exercise 5 under the assumption that there are two parallel servers, and that each customer can use either server.

8. Let A_n, S_n, and D_n, $n \geq 1$, denote the respective arrival, service, and departure times for customers entering a single-server queueing system with a FIFO queue discipline. Assuming $D_0 = 0$, then provide a recurrence for D_n.

9. Repeat the previous exercise assuming two servers.

10. At Sweet P's bakery there is a single cake decorator who takes an average of 15 minutes to decorate a cake, with a standard deviation of 10 minutes. Determine the maximum arrival rate (for an exponential interarrival distribution) of cakes that need decorating if the average length of the waiting queue is not to exceed five cakes.

11. Arrivals to a small airport all use the same runway. At a certain time of day the number of arrivals is represented by a Poisson process with $\lambda = 30$ arrivals per hour. The time to land an airplane and move it off the runway is constant at 90 seconds. If a delayed aircraft burns \$5,000 per hour while waiting to land, determine the average cost per aircraft while delayed waiting to land.

12. Prove that $M/G/1$ and $M/M/1$ queueing systems produce the same L, w, L_Q, and w_Q values when G is assumed exponential with rate μ and $\sigma^2 = 1/\mu^2$.

13. Vans Grocery store has adopted the slogan "Three is a Crowd", meaning that it has the goal of keeping the average length of a checkout waiting line to no more than 2 customers. The store experiences its highest Poisson arrival rate of 1 customer per minute on Sunday afternoons. If all cashiers are assumed to have an exponential service distribution, with a mean of 4 minutes per customer, then how many cashiers will be needed to achieve the waiting-line goal? Assume one waiting line per cashier. Hint: for the sake of analysis, you may assume a single waiting

line, and divide its length by the number of servers to achieve the length per cashier waiting line.

14. An $M/M/1$ queueing system within a forklift repair company has mean time-between-arrivals of 4 minutes, and a mean service time of 3 minutes. The arriving cutomers are mecchanics who earn \$23 per hour, and enter the system for the purpose of checking out parts that are needed for repair jobs. The system server earns \$15 dollars per hour. Management is concerned about the money being wasted while mechanics wait in the queue before checking out their needed parts. One solution is to hire a second server at \$15 dollars per hour so that wait times are diminished. If the new server works 8 hours per day, and an average of 120 mechanics are served each day, determine the net change in daily expenditures. Hint: compute w_Q before and after the new server is hired.

15. Consider a queueing network with two queue nodes A and B. Customers arrive to node A with a net arrival rate of λ. After being serviced at A, these customers leave A and enter B for service (note: these are the only customers who enter B). After being serviced at B, a fraction ϵ of the customers return to A for more service (after which they will return to B), while the other $(1-\epsilon)$ fraction of customers leave the network. Determine the net arrival rate of customers entering B.

16. A repair an inspection facility consists of two stations: a repair station with two technicians, and an inspection station with one inspector. Each repair technician works at a rate of 3 items per hour, while the inspector can inspect 8 items per hour. Approximately 10% of all inspected items are sent back to the repair station, and this percentage is independent of the number of times the item was previously inspected. If items arrive at the rate of 5 per hour, then determine the average time that an item spends in the system. Assume that all arrival and service distributions are exponential. Hint: use the results of the previous exercise.

17. Recall the DMV queueing network from Example 8. Management has decided to not hire another employee, but rather control the network arrival rate λ so that at no station will the average waiting time exceed 15 minutes. Determine the maximum allowable value for λ that will be able to support this goal. Hint: it will be helpful to use an R script, along with R's `uniroot` function for finding the root of a function.

Exercise Solutions

1. a) all possible vehicles