Applied Statistics and Queuing Theory Credits: 3

Introduction, mean, standard deviation and measures of dispersion, Moments, Skewness and Kurtosis, Elementary probability theory, Characteristics of distributions, Elementary sampling theory, Estimation, Hypothesis testing and regression analysis.

Probability distribution and expectations, discontinuous probability distribution e.g. binomial, Poission and negative binomial; Continuous probability distributions. e.g. normal and exponential

Stochastic processes, Discrete time Markov Chain and continuous time Markov chain, birth death process in queuing.

Queuing models: M/M/1, M/M/C, M/G/1, M/D/1, G/M/1 solution of network of queue-closed queuing models, approximate solution methods, Application of queuing models in Computer Science.

0904061

CSE-08(3A)

Applied Statistics & Queueing Theory

(*) Ref. Books:

1. Probability, Statistics, and Queueing Theory with Computer Science Applications
   — Arnold O. Allen
2. Probability & statistics with Reliability, Queuing, and Comp. Science Applications
   — Kishor S. Trivedi

(*) Introduction:

Queue: an orderly line of one or more persons or jobs.

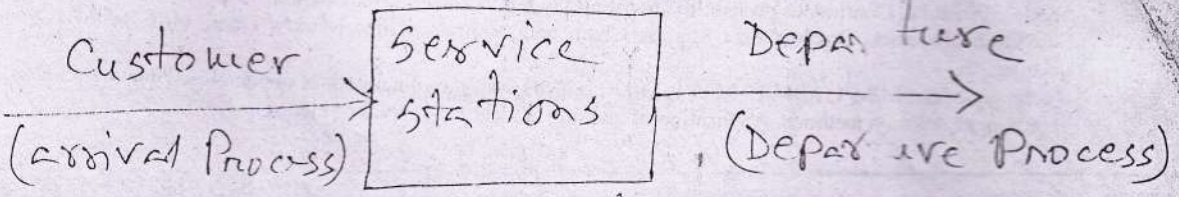
Queue occurs in front of any service system, e.g.,

ATM Booth
Traffic signal light
Ticket Counter
Petrol Pump stations, etc.

Queue are also common in computer systems, e.g.,

interactive/time-sharing Computer System
Data base request processing
I/O requests processing, etc.

Simple Queueing System



Jobs
persons
Aeroplanes
Cars

printer
memory space
runway
Parking area
Airline counter
(service Process)

(*) Also see Fig.- 5.0.1

(*) In Table-5.0.1 we list some typical computer queueing system.

(*) Scope of Queueing Theory

Q. Theory is mainly seen as a branch of applied probability theory. Its applications are in different fields, e.g. communication networks, computer systems, machine plants and so forth.

The subject of Q. Theory can be described as follows:

Fig.- 5.0.1

Queueing Theory tries to answer questions like e.g., the mean waiting time in queue, the mean system response time, mean utilization of service facility, distribution of

the no of customers in
the sys.

(*) Describing a Queueing System

Fig.- 5.1.1

The basic queuing theory definitions and notations are listed in table-5.1.1.

In order to describe a queueing system, analytically, a number of elements of the system must be known.
- the most important ones are -

1. Population or source:
   - finite or infinite

2. Arrival Pattern:
   $\tau_k = t_k - t_{k-1}$ $(k = 1, 2, 3 \cdots)$
   - interarrival times
   - exponential pattern
     $[A(t) = 1 - e^{-\lambda t}]$
   - constant
   - Erlang-K
   - hyper exponential

3. Service Time Distribution:

$$F(t) = P(S \le t) = 1 - e^{-\mu t} = 1 - e^{-t/H_S}$$

- exponential
- Erlang-k
- constant
- hyper exponential

4. Maximum Queueing System Capacity:

- infinite
- loss systems
- k: max^m no of customers

5. Number of servers:

- single server system
- multiserver system
- infinite server system

6. Queue Discipline (service Discipline)

- FCFS / FIFO
- LCFS / LIFO
- RSS / SIRO
- PRI

(*) Notation for describing a queuing system:

- Kendall's notation:

A/B/c/K/m/Z.

A: interarrival time Dist^n

B: service time Dist^n

c: no of servers

K: the system capacity

m: the no in the population or source.

Z: the queue discipline

- Usually the shorter notation A/B/c is used.

- The symbols chosen by Kendall and traditionally used:

A:

M - exponential, D - Deterministic/constant, E_k - Erlangian ($k=1,2,\dots$), G - General, GI - General Independent, U - Uniform

B: - do -

c: 1, 2, 3, ...

K: 1, 2, 3, ... ∞

m: 1, 2, 3, ... ∞

Z: FCFS, LCFS, PRI, RSS, SIRO, etc.

- Example:

$M/E_4/3/20/\infty/SIRO$

(*) Measurement of performance of the Queueing system:

1. Traffic Intensity: (offered load)

$$a = \lambda/E[z]$$
$$= \lambda H_s \quad [\because \lambda = 1/E[z]]$$
$$= \lambda/\mu \quad [\mu = 1/H_s]$$

2. Server Utilization:

$$\rho = a/c = \lambda/(c\mu)$$

(*) Example-5.1.1

$$\lambda < c\mu$$
$$a = \lambda/\mu < c$$

(*) Little's Law:

$$L = \lambda W$$

(*) Example-5.1.2

CSE-02(3A)

Chapter 5

Queueing Theory

queue n (Brit.) an orderly line of one or more persons or jobs.

Stan Kelly-Bootle
The Devil's DP Dictionary

Hurry up and wait.
Old army saying.

Have you ever encountered a queue,
In which Poisson arrivals accrue?
In Statistics, I'm told
This assumption can hold...
... But it sure sounds more fishy than true!

Ben W. Lutek

5.0 Introduction

One of the most fruitful areas of applied probability theory is that of queueing theory or the study of waiting line phenomena (a queue is a waiting line). Waiting in line (queueing) for service is one of the most unpleasant experiences of life on this planet. Barrer [2] says it all in the title of his paper, "Queueing[1] with Impatient Customers and Indifferent Clerks." Barrer says,

[1] I cannot bring myself to spell queueing "queuing" as Barrer did.

> In certain queueing processes a potential customer is considered "lost" if the system is busy at the time service is demanded. The telephone subscriber hangs up when he gets a busy signal. A man trying to get a haircut during his lunch hour does not wait unless a chair is immediately available. Another form of this general situation is that in which customers wait for service, but wait for a limited time only. If not served during this time, the customer leaves the system and is considered lost. Such situations occur in the processing or merchandising of perishable goods. Many types of military engagements are similarly characterized. An attacking airplane engaged by antiaircraft or guided missiles is available for "service," i. e., is within range, for only a limited time.

In spite of the catchy title, which is descriptive of the common feeling about queues, Barrer's paper is an innovative application of queueing theory to the destruction of attacking warplanes, not to general queueing theory.

We must join a queue when we want to get cash from an automatic teller machine (ATM), buy stamps, pay for our groceries, purchase a movie ticket, obtain a table in a crowded restaurant, etc. Larson [38] discusses some of the psychological implications of queues. He says,

> Queues involve waiting, to be sure, but one's attitudes toward queues may be influenced more strongly by other factors. For instance, customers may become infuriated if they experience *social injustice*, defined as violation of first in, first out. *Queueing environment* and *feedback regarding the likely magnitude of the delay* can also influence customer attitudes and ultimately, in many instances, a firm's market share. Even if we focus on the wait itself, the "outcome" of the queueing experience may vary nonlinearly with the delay, thus reducing the importance of average time in queue, the traditional measure of queueing performance. This speculative paper uses personal experiences, published and unpublished cases, and occasionally "the literature" to begin to organize our thoughts on the important attributes of queueing.

Larson discusses some techniques that help to make queues more bearable for humans.

Queues are also common in computer systems. Thus, there are queues of inquiries waiting to be processed by an interactive computer system, queues of data base requests, queues of I/O (input/output) requests, etc.

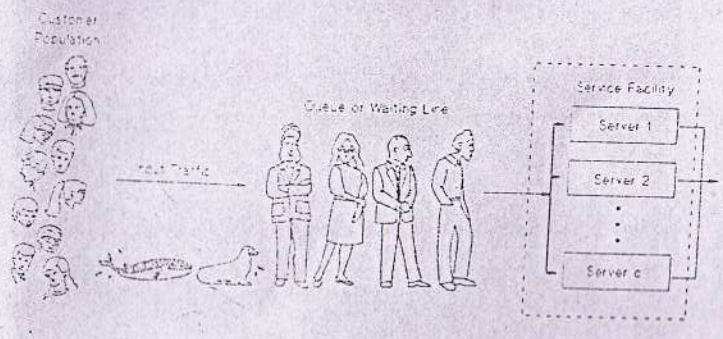


Figure 5.0.1. Elements of a queueing system.

Figure 5.0.1 represents the elements of a basic queueing system, pictorially. We consider a queueing system to be basic if it has only one service facility, although there may be more than one server in the facility. (The reader may note that queueing is spelled "queuing" in some publications (the last *e* is elided) but I prefer "queueing" because (1) that is the way most queueing theory authorities spell it, and (2) it is a delightful and rare word having five consecutive vowels.)

Customers from a *population* or *source* enter a queueing system to receive some type of service. The word *customer* is used in the generic sense and thus may be an inquiry message requiring transmission and processing, a program requiring *I/O* service, a program in a multiprogramming computer system requiring CPU service, etc. The *service facility* of the queueing system has one or more *servers* (sometimes called *channels*). A server is an entity capable of performing the required service for a customer. If all servers in the service center are busy when a customer enters the queueing system, the customer must join the queue until a server is free.

In any system that can be modeled as a queueing system, there are trade-offs to be considered. If the service facility of the system has such a large capacity that queues rarely form, then the service facility is likely to be idle a large fraction of the time so that unused capacity exists. Conversely, if almost all customers must join a queue (wait for service) and the servers are rarely idle, there may be customer dissatisfaction and possibly lost customers as Barrer [2] noted.

In Table 5.0.1 we list some typical computer queueing systems.

Table 5.0.1. Typical Queueing Systems

Queueing System	Customer	Server(s)
Airline reservation system	Traveler wanting information and/or reservations	Agent plus terminal to a computer reservation system
Interactive inquiry system	Inquiry from terminal	Communication line plus a computer
Interactive order entry system	Order	Communication line plus a computer
DASD (direct access storage device) queueing system	Request for records from DASD	Channel plus control unit and DASD
Message buffering system	Message (incoming or outgoing)	Message buffer(s) (all of them together form the service facility)

Queueing theory, in many cases, enables a designer to ensure that the proper level of service is provided in terms of response time requirements (response time is the sum of customer queueing time and service time) while avoiding excessive cost. The designer can do this by considering several alternative systems and evaluating them by analytic queueing theory models. The future performance of an existing system can be predicted so that upgrading of the system can be done on a timely basis. For example, an analytical model of an interactive system may indicate that the expected load a year in the future will swamp the present system; the model may make it possible to evaluate different alternatives for increased capacity, such as adding more main memory, getting a faster CPU, providing more auxiliary storage, replacing some disk drives by drums, etc. We shall give a number of practical examples of how queueing theory can help one explore the alternatives available in an informed way. For very large computer systems, commercial analytical queueing theory modeling packages exist. Most of these are described in Howard and Butler [23].

In this chapter we discuss the elements of queueing theory and study some basic queueing models that are of great utility in the study of com-

puter systems. These basic models can be used to study subsystems of large computer systems, such as the *I/O* subsystem. In the next chapter we show how some of these basic queueing models can be combined to study more complex systems in which the output of one service center may be the input to another (queues in tandem and networks of queues). In such systems there is a slight "abuse of notation" in which we refer to a "queue" to describe a service center plus the associated queue or queues. The modeling packages we mentioned above model networks of queues.

5.1 Describing a Queueing System

Figure 5.1.1 illustrates the primary random variables in a queueing system.

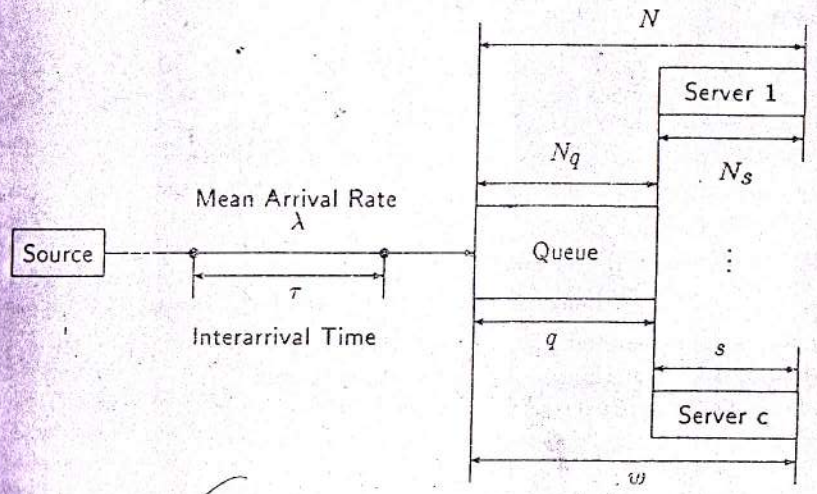


Figure 5.1.1. Queueing theory random variables.

The basic queueing theory definitions and notation are listed in Table 5.1.1. A more complete set of definitions and notation is given in Table I of Appendix C.

Table 5.1.1. Basic Queueing Theory Notation and Definitions

c	Number of identical servers.
L	Expected steady state number of customers in the queueing system, $E[N]$.
L_q	Expected steady state number of customers in the queue, $E[N_q]$. Does not include those receiving service, $E[N_s]$.
L_s	Expected steady state number of customers receiving service, $E[N_s]$.
λ	Mean (average) arrival rate of customers to the system.
μ	Mean (average) service rate per server, that is, the mean rate of service completions while the server is busy.
$N[t]$	Random variable describing the number of customers in the system at time t.
N	Random variable describing the steady state number of customers in the system.
$N_q[t]$	Random variable describing the number of customers in the queue at time t.
N_q	Random variable describing the steady state number of customers in the queue.
$N_s[t]$	Random variable describing the number of customers receiving service at time t.
N_s	Random variable describing the steady state number of customers in the service facility.
$p_n[t]$	Probability there are n customers in the system at time t.
p_n	Steady state probability there are n customers in the system.
q	Random variable describing the time a customer spends in the queue (waiting line) before service begins.
ρ	Server utilization $= \frac{\lambda}{c\mu} = \frac{E[N_s]}{c}$.
s	Random variable describing the service time. $E[s] = \frac{1}{\mu}$.
τ	Random variable describing interarrival time. $E[\tau] = \frac{1}{\lambda}$.
w	Random variable describing the total time a customer spends in the queueing system. $w = q + s$.
W	Expected (mean or average) steady state time a customer spends in the system. $W = E[w] = W_q + W_s$.
W_q	Expected (mean or average) steady state time a customer spends in the queue. $W_q = E[q] = W - W_s$.
W_s	Expected (mean or average) customer service time. $E[s]$.

There are some obvious relations between some of the random variables shown in Figure 5.0.1. With respect to the number of customers in the queueing system, we must have

$$N[t] = N_q[t] + N_s[t]. \tag{5.1}$$

$$N = N_q + N_s. \tag{5.2}$$

In (5.2) we assume the queueing system has reached the *steady state*, that we now describe. When a queueing system is first put into operation and for some time afterwards, the number of customers in the queue and in service depends strongly on both the initial conditions (such as the number of customers queued up waiting for the facility to begin operation) and on how long the system has been in operation (the time parameter t). After the system has been in operation for some time, however, the influences of the initial conditions have "damped out" and state of the system is independent of time—the system is in equilibrium or the steady state. However, N, N_q, and N_s *are* random variables; that is, they are not constant but have probability distributions.

Equation (5.2) of course implies that

$$E[N] = E[N_q] + E[N_s], \tag{5.3}$$

that is often written

$$L = L_q + L_s. \tag{5.4}$$

There are some obvious relationships between the random variables describing time, also; clearly the total time in the queueing system for a customer is the sum of the queueing time (time spent in line waiting for service) and service time; that is,

$$w = q + s, \tag{5.5}$$

and

$$E[w] = E[q] + E[s]. \tag{5.6}$$

Equation (5.6) is often written

$$W = W_q + W_s. \tag{5.7}$$

Some common English words have a special meaning in queueing theory. A customer who refuses to enter a queueing system because the queue is too long is said to be *balking* while one who leaves the queue without receiving service because of excessive queueing time is said to have *reneged*.

Customers may *jockey* from one system to another with a shorter queue for service.

In order to describe a queueing system analytically, a number of elements of the system must be known. We consider the most important of these below.

Population or Source

The primary characterization of the population or source of potential customers is whether it is finite or infinite. An infinite source system is easier to describe mathematically than one with a finite source. The reason for this is that, in a finite source system, the number of customers in the system affects the arrival rate; indeed, if every potential customer is already in the system, the arrival rate drops to zero. For infinite population systems the number of customers in the system has no effect on the arrival pattern. If the customer population is finite but large, we sometimes assume an infinite source to simplify the mathematics.

Arrival Pattern

The ability of a queueing system to provide service for an arriving stream of customers depends not only on the mean arrival rate, λ, but also on the pattern in which they arrive. Thus, if customer arrivals are evenly spaced in time, say every h time units, the service facility can provide better service than if customers arrive in clusters. (The extreme case of clustering in which a number of customers arrive simultaneously is called *bulk arrivals*.) We assume customer arrivals at times

$$0 \le t_0 < t_1 < t_2 < \cdots < t_n < \cdots \tag{5.8}$$

(We always assume that observation of a queueing system begins at time $t = 0$.) The random variables $\tau_k = t_k - t_{k-1}$, $(k = 1, 2, 3, \ldots)$, are called *interarrival times*. We usually assume that $\tau_1, \tau_2, \ldots$ is a sequence of independent identically distributed random variables and use the symbol τ for an arbitrary interarrival time. The usual method of specifying the arrival pattern is to give the distribution function, $A[\cdot]$, of the interarrival time. The arrival pattern most commonly assumed for applied queueing models (because of its pleasant mathematical properties) is the exponential pattern: $A[t] = 1 - e^{-\lambda t}$, where λ is the average arrival rate. (If you are rusty on the exponential distribution, you should review Section 3.2.2. The exponential distribution is of paramount importance in queueing theory.)

Because of the properties of the exponential distribution, summarized in Theorem 3.21, if the interarrival time of customers to a queueing system has

an exponential distribution, the arrival pattern is called a *Poisson arrival pattern* (or *process*) or a *random arrival pattern* (or sometimes just said to be *random*). Other commonly assumed arrival patterns are constant, Erlang-k, and hyperexponential.

The symbol λ is reserved (except for finite queue systems and loss systems) for the mean or average rate into the system; therefore, the average interarrival time, $E[\tau]$, equals $1/\lambda$. (For finite queue systems (Section 5.2.2) and loss systems (examined in Section 5.2.4), λ_a is used for average arrival rate into the system.)

Service Time Distribution

The exponential distribution is often used to describe the service time of a server because of the Markov or "memoryless" property of this distribution (Theorem 3.2.1(d)). Thus, if the service time is exponential, the expected time remaining to complete a customer service is independent of the service already provided. Suppose now that the queueing system has several identical servers, each with an exponential service time with parameter μ, and that n of the servers are now busy. Let T_i be the remaining service time for server i $(i = 1, 2, \ldots, n)$. By the Markov property, each T_i has an exponential distribution with parameter μ. T, the time until the next service completion, is the minimum of $\{T_1, T_2, \ldots, T_n\}$. Hence, by Theorem 3.2.1(h), T has an exponential distribution with parameter $n\mu$; the service facility is performing like a single exponential server with mean service rate $n\mu$. In queueing theory exponentially distributed service time is called *random service* and the distribution function, $W_s[\cdot]$, is given by

$$W_s[t] = P[s \le t] = 1 - e^{-\mu t} = 1 - e^{-t/W_s}. \tag{5.9}$$

Here μ is called the *average service rate*. The average (mean) service rate, $\mu = 1/W_s$, is the average rate at which a server processes customers when the server is busy. This definition is valid for all service time distributions. Other common service time distributions are Erlang-k, constant, and hyperexponential. The hyperexponential distribution is useful to describe a service time distribution with a large variance relative to the mean (see Section 3.2.9).

The *squared coefficient of variation*, C_X^2, defined for a random variable X with $E[X] \neq 0$ by

$$C_X^2 = \frac{\mathrm{Var}[X]}{E[X]^2}, \tag{5.10}$$

is a useful parameter to measure the character of probability distributions used to represent service time or interarrival time.[2] If X is constant,

[2] We have discussed C_X^2 for some special distributions in Chapter 2. It is often used by queueing theory specialists to compare distributions to the exponential distribution.

$C_s^2 = 0$; if X has an exponential distribution, then $C_X^2 = 1$; if X has an Erlang-k distribution, then $C_X^2 = 1/k$; and, if X has a k-stage hyperexponential distribution, then $C_X^2 \geq 1$. (For example, if X has a two-stage hyperexponential distribution with $q_1 = 0.4$, $q_2 = 0.6$, $\mu_1 = 0.5$, and $\mu_2 = 0.01$, then $E[X] = 60$, $E[X^2] = 12,003.2$, $\mathrm{Var}[X] = 8,306.56$, and $C_X^2 = 2.25$.) We conclude that, for C_s^2 close to zero, the service time is almost constant; if C_s^2 is close to one, the service time is approximately exponential; if C_s^2 is close to $1/k$ for some positive integer k, then s can be approximated by an Erlang-k distribution; and, finally, if $C_s^2 > 1$ then s has a great deal of variability and can be approximated by a two-stage hyperexponential distribution. Similarly if C_r^2 is close to zero, the arrival process has a regular pattern; if C_r^2 is close to one, the arrival pattern is nearly random; if C_r^2 is close to $1/k$ for some positive integer k, then r can be approximated by an Erlang-k distribution; while if $C_r^2 > 1$, then the arrivals tend to cluster.

Maximum Queueing System Capacity

In some systems the queue capacity is assumed to be infinite; that is, every arriving customer is allowed to wait until service can be provided. Other queueing systems, called "loss systems," have zero queue capacity; thus, if a customer arrives when the service facility is fully utilized (all the servers are busy), the customer is turned away. For example, some dial-up telephone systems are loss systems. Still other queueing systems, such as a message buffering system, have a positive but not infinite capacity queue. We use K to represent the maximum number of customers allowed in such a system.

Number of Servers

The simplest queueing system, in this sense, is the *single-server system* that can serve only one customer at a time. A *multiserver system* has c (usually) identical servers and can provide service to as many as c customers simultaneously. In an *infinite server system* each arriving customer is immediately provided with a server. Although there cannot actually be infinitely many servers in any system, there are queueing systems that have sufficient servers that they *appear* to have infinitely many.

Queue Discipline (Service Discipline)

This is the rule for selecting the next customer to receive service. The most common queue discipline is "first-come, first-served," abbreviated as FCFS or sometime called "first-in, first-out" and abbreviated FIFO. Other common queue disciplines include "last-come, first-served," LCFS (or "last-in, first-out," LIFO); "random-selection-for-service," RSS (or "service-in-random-order," SIRO), that means that each customer in the queue has the

same probability of being selected for service; or "priority service," PRI. Priority service means that some customers get preferential treatment, just as in George Orwell's *Animal Farm* some animals (the pigs) were "more equal" than others. In a priority queueing system, customers are divided into priority classes with preferential treatment afforded by class. We study priority queueing systems in Section 5.4.

A special notation, called the Kendall notation, after David Kendall [26], its originator, has been developed to describe queueing systems. The notation has the form $A/B/c/K/m/Z$, where A describes the interarrival time distribution, B the service time distribution, c the number of servers, K the system capacity (maximum number of customers allowed in the system), m the number in the population or source, and Z the queue discipline. Usually the shorter notation $A/B/c$ is used, and it is assumed that there is no limit to the length of the queue, the customer source is infinite, and the queue discipline is FCFS. The symbols chosen by Kendall and traditionally used for A and B are

GI	general independent interarrival time
G	general service time
H_k	k-stage hyperexponential interarrival or service time distribution
E_k	Erlang-k interarrival or service time distribution
M	exponential interarrival or service time distribution
D	deterministic (constant) interarrival or service time distribution[3]
U	uniform interarrival or service time distribution

When we say a queueing model, such as M/G/1, has a general *service time distribution*, we mean the equations of the model are valid for general service time distributions (make few assumptions about the service time distribution) and thus, in particular, the equations are valid for the M/M/1 system. However, equations developed specifically to describe an M/M/1 queueing system would give more information than the general equations developed for the M/G/1 model and applied, as a special case, to M/M/1. Similar remarks apply to the phrase *general independent interarrival time distribution*.

An example of the full Kendall notation is $M/E_4/3/20/\infty/SIRO$. For this system the interarrival time is exponential, the service time is Erlang-4 for each of the three servers, the maximum system capacity is 20 (3 in service and 17 in the queue), the source is infinite, and the queue discipline is service in random order.

[3] Kendall used the word deterministic to describe D, but degenerate is more descriptive.

As the Kendall notation suggests, certain properties of a queueing system are assumed known: it is desired to calculate measures of performance of the queueing system from these known parameters. It is usually assumed that the average arrival rate λ (or, equivalently, the average interarrival time, $E[\tau]$) and the average service rate per server μ (or the average service time per server, W_s) are known. It is also assumed that the arrival and service time distributions are known. (For the M/G/1 model, only some of the moments need to be known to compute useful performance information.) One fundamental measure of queueing system performance is the *traffic intensity*[4] $a = W_s/E[\tau]$, also known as the *offered load*. It should be noted that W_s is the average service time per server, while $E[\tau]$ is the average interarrival time for all customers entering the queueing system and not just for the customers who are serviced by a particular server (unless, of course, there is but one server). Since $\lambda = 1/E[\tau]$ and $\mu = 1/W_s$, the traffic intensity can also be written as λW_s or λ/μ. The quantity $\rho = a/c = \lambda/(c\mu)$ is called the *server utilization* because it represents the average fraction of the time that each server is busy (assuming the traffic is evenly distributed to the servers); that is, it is the probability that a given server is busy (as observed by an outside observer).

Example 5.1.1 Consider a D/D/1 queueing system with a constant interarrival time of 20 seconds and a constant service time of 10 seconds. Then the server is busy half of the time, since $\rho = a = 10/20 = 0.5$. If the server is replaced by one that requires exactly 15 seconds to service a customer, then $\rho = 15/20 = 0.75$ and this server is busy three-fourths of the time. Replacing this server by one requiring exactly 30 seconds to service a customer may save some money but the traffic intensity $a = 30/20 = 1.5$. In order to keep up the server must provide 30 seconds of service every 20 seconds! This is impossible. Two servers must be provided. Thus, the traffic intensity a is a measure of the required number of servers and ρ (when it is less than one) is a measure of congestion. In general we can argue that if customers are arriving at the rate λ and the c servers serve them at the rate $c\mu$, then we must have $\lambda < c\mu$ if the servers are to keep up. But this means that $a = \lambda/\mu < c$. □

Although server utilization, ρ, is a measure of congestion, there are some other useful measures of queueing system performance including the following steady state values:

[4] Although traffic intensity is dimensionless, it is often referred to in units of erlangs in honor of the queueing theory pioneer Agner Krarup Erlang.

W	average customer time in the system (queueing for and in service)
W_q	average customer queueing time
$\pi_w[90]$	90th percentile value of w
$\pi_q[90]$	90th percentile value of q
L	average number of customers in system
L_q	average number of customers in the queue
p_n	probability there are n customers in the queueing system

Little's Law

One of the foundations of queueing theory is the formula

$$L = \lambda W. \tag{5.11}$$

The formula (5.11) applies to any system in equilibrium in which customers arrive, spend a certain amount of time, and then depart. In (5.11) we assume that λ is the average arrival rate, W is the average time a customer spends in the system, and L is the average number of customers in the system. The formula goes by a number of names, including "Little's law", "Little's formula", and "Little's theorem." It was first proven by John D. C. Little [42] in the context of a steady state queueing system in which L, λ, and W have the queueing theory definitions. However, (5.11) holds in more general situations that need not have anything to do with queues. Although Little's law is easy to state and intuitively reasonable, the proof is difficult. Little [42] provided the first known proof. However, it is a rather formal, nonintuitive proof using the mathematical concept of metric transitivity. Stidham [56] has published a simpler proof that is quite general and more intuitive than Little's proof. We state Stidham's version of Little's theorem without proof.

Theorem 5.1.1 (*Little's Theorem According to Stidham*) *Let $L(x)$ be the number of customers present at time x. Define L by*

$$L = \lim_{t \to \infty} \frac{1}{t} \int_0^t L(x)dx. \tag{5.12}$$

Define λ by

$$\lambda = \lim_{t \to \infty} \frac{N(t)}{t}, \tag{5.13}$$

where $N(t)$ is the number of customers who arrive in the interval $[0,t]$. Let W_i be the time in the system for the ith customer and define the mean time

in the system W by

$$W = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} W_i. \tag{5.14}$$

If λ and W exist and are finite, then so does L, and

$$L = \lambda W. \tag{5.15}$$

Proof See Stidham [56]. ■

In the following example we show some simple applications of Little's law.[5]

Example 5.1.2 Little's law can be applied to the queue, itself, to prove that

$$L_q = \lambda W_q, \tag{5.16}$$

and to the service center, alone, to prove that

$$L_s = \lambda W_s = a, \tag{5.17}$$

for any number of servers. If λ and W_s are known for any steady state queueing system, then Little's law allows us to calculate all of the primary performance measures L, L_q, W, and W_q, if any one of them is known. For example, if W is known then

$$L = \lambda W, \qquad W_q = W - W_s, \qquad \text{and} \qquad L_q = \lambda W_q. \quad \square$$

Many phenomena encountered in queueing theory are not intuitive. The following example is one such case.

Example 5.1.3 (*A Queueing Theory Paradox*) Taxis pass a certain corner with an average interarrival time of 20 seconds. What is the average time that one would expect to wait for a taxi? (Assume that you are in New York City so you can't telephone for a taxi.)

Solution Intuitively, it would seem that a taxi is just as likely to arrive at one point in time between arrivals as any other; that is, by symmetry, the distribution of arrival time should be uniform on the interval from 0 to 20 seconds. Thus, the average waiting time should be 10 seconds. This is true, however, only when the taxis arrive exactly 20 seconds apart. In fact, it is shown by Takács [59, page 20], if w is the length of time until the next

[5] I asked John D.C. Little which of the appellations, Little's law, Little's formula, or Little's theorem he preferred. Little said that he had no preference, he just hoped his name would be spelled correctly.

arrival of a taxi measured from the time of arrival of the person seeking a cab, then

$$E[w] = \frac{E[\tau]}{2} \{1 + C_\tau^2\}, \tag{5.18}$$

where τ is the taxi interarrival time. Thus, if the interarrival time is exponential one would expect to wait 20 seconds, on the average, for a taxi. It follows from Proposition 4.5.3 that, if τ is exponential, so is w with the same parameters as τ. If the interarrival time is hyperexponential with $C_\tau^2 = 3$, one would expect to wait 40 seconds, on the average! This well-known "waiting time paradox" is discussed by Feller [16, pages 11, 23] (he uses a bus in his example rather than a taxi). Snell [55] also calls it the "bus paradox" and provides a BASIC program to simulate the waiting time for exponential τ and to draw a graph of the result. □

The above example highlights the fact that in queueing theory, intuition is often misleading. One *can* get an appropriate intuitive picture of what is happening in this example by thinking of the taxi arrivals as being appropriately scattered along the time axis and realizing that a randomly chosen point on this axis is more likely to fall in a long interval between two arrivals than in a short one. Also, the larger C_τ^2 is, the more clustered the arrivals are. If arrivals are clustered, then there must be some very large gaps between some of the arrivals to make up for the short interarrival times in the clusters.

5.2 Birth-and-Death Process Models

A number of important queueing theory models fit the birth-and-death process description of Section 4.3. A queueing system based on this process is in state E_n at time t if the number of customers in the system is n, that is, if $N[t] = n$. A birth is a customer arrival and a death occurs when a customer leaves the system after completing service. We consider only steady state solutions to the queueing model. Thus, given the birth rates $\{\lambda_n\}$ and death rates $\{\mu_n\}$, and assuming that

$$S = 1 + C_1 + C_2 + \cdots < \infty, \tag{5.19}$$

where

$$c_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} \qquad n = 1, 2, \ldots, \tag{5.20}$$

we calculate $P_0 = \frac{1}{S}$ (5.21)