Chapter 2
Queueing Models for Healthcare Operations

Diwakar Gupta

1 Introduction

Queues form when entities that request service, typically referred to as *customers*, arrive at a *service facility* and cannot be served immediately upon arrival. In healthcare delivery systems, patients are typically the customers and either outpatient clinics or diagnostic imaging centers or hospitals are the service facilities. There are also many atypical examples of customers and service facilities, as shown below.

Customers	*Service facility*
Diagnostic images	Radiology department
Doctors' notes	Coding department (for billing purposes)
Prescriptions	Mail-order pharmacy
Transplant candidates	Organ procurement organization

A service facility may consist of one or more service stations where customers are served. Further, each service station may consist of one or more servers. For example, the processing of diagnostic images may require two types of servers—radiologists who read the images and transcribers who input radiologists' dictated notes into patient charts. Servers are often grouped by their expertise to form service stations, although other configurations (e.g., multiple specialty teams of doctors and nurses) are also prevalent. A common feature of the vast majority of queueing models is that customers are discrete, and the number of customers waiting in the service facility is integer valued.

D. Gupta (✉)
University of Minnesota, 111 Church Street S. E., Minneapolis, MN 55455, USA
e-mail: guptad@me.umn.edu

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_2,

Queues are ubiquitous, particularly in healthcare delivery systems. At the same time, queues are undesirable because delay in receiving needed services can cause prolonged discomfort and economic loss when patients are unable to work and possible worsening of their medical conditions that can increase subsequent treatment costs and poor health outcomes. In extreme cases, long queues can delay diagnosis and/or treatment to the extent that death occurs while a patient waits. For example, there is a severe shortage of organs in the USA and many patients die while waiting for suitable organs for transplant.

Given the negative consequences of queues in healthcare delivery systems, the following questions naturally arise. Why do queues form? Why must customers wait to be served? Which features of system design affect queueing and by how much? What trade-offs must be considered by a service system architect when choosing system parameters? This chapter attempts to provide answers to questions such as these.

Although queues have existed as far back as historical records are available, mathematical study of queues, called queueing theory, has been around since the early 1900s. Works on the theory and applications of queueing systems have grown exponentially since the early 1950s. It is neither possible nor the intent to provide a summary of this vast body of literature in this chapter. For that, there are many excellent books, both at the introductory and advanced levels—see, for example, Bhat (2008), Cohen (1969), Cox and Smith (1961), Gross and Harris (1985), Morse (1958), Newell (1982), Takacs (1962) and Wolff (1989). A review of papers that attempts to tackle real queueing problems can be found in Worthington (2009). This chapter provides a review of a few basic queueing models and discusses their implications for healthcare operations management.

Borrowing terminology from the queueing literature, we shall henceforth use the terms *queueing system* and *service facility* interchangeably. A queueing system has the following elements:

1. Servicestations or workstations, their configuration, and routing protocols that determine flow of customers from one station to another.
2. Number of servers at each station.
3. Service protocol at each station—a commonly used protocol is first in first out (FIFO) because it is deemed to be fair (Larson 1987). However, it is not at all uncommon to give higher priority to certain types of customers (often referred to as *classes* in the queueing literature). For example, patients whose condition is deemed critical by medical professionals generally bypass queues.
4. Service time distribution by customer class, server, and station.
5. Arrival process—the distributions of inter-arrival times, number of arrivals at each arrival epoch, and arrival location.
6. Size of waiting room at each station. When waiting room is limited, either customers are turned away or congestion at a downstream station causes *blocking* at an upstream station.

7. Protocols governing server absences and distributions of server vacations. Vacation refers to a period of time when a server is not available, which could happen for a whole host of reasons including activities such as attending to other tasks and taking a break.

Suppose we observe arrivals and departures from a queueing system that starts empty. The system may consist of an arbitrary number of stations with an arbitrary number and configuration of servers at each station, customer classes, service protocols, and sizes of waiting rooms. We treat the entire system as a black box. An arrival to this system is either turned away on account of a full waiting room or the arrival enters the system. The stream of arrivals that enter the queueing system is characterized by arrival times $a_1 \leq a_2 \leq \cdots \leq a_j \leq \cdots$, where a_j denotes the arrival epoch of the jth arrival in sequence, and corresponding service times $(s_1, \ldots, s_j, \ldots)$. Each customer is served either as soon as it arrives or according to some service protocol. Given this basic setup, queueing models are frequently used to characterize the following stochastic processes:

$$N_q(t) = \text{number of customers in queue at epoch } t, \tag{2.1}$$

$$N(t) = \text{number of customers in the queueing system at epoch } t. \tag{2.2}$$

Clearly, $N(t) = N_q(t) +$ the number of customers receiving service at time t. Similarly, for the jth customer, the quantities of interest are:

$$W_j = \text{time in queue of the } j\text{th customer}, \tag{2.3}$$

$$D_j = W_j + S_j = \text{total delay of the } j\text{th customer}. \tag{2.4}$$

Note that S_j in the above expression denotes the random service time of the jth customer. For queueing systems with finite waiting rooms, we are also interested in the probability that an arrival is turned away.

The purpose of mathematical models of queues is to obtain closed-form or recursive formulae that allow system designers to calculate performance metrics such as average queue length, average waiting time, and the proportion of customers turned away. We say that mathematical models are tractable when closed-form or recursive formulae can be obtained, and in such cases the resulting expressions for the performance metrics are referred to as "analytical results." Note that it is always possible to write equations that describe how the number of customers in each queue in the queueing system of interest changes over time. Such equations can be used to simulate a queueing system's performance. In this chapter, the simulation-based results are also referred to as numerical solutions.

In the vast majority of cases, analytical results are possible only for limiting behavior (called steady state) of the above-mentioned performance metrics and in particular for time-average or customer-average metrics, when such averages exist. Specifically, steady-state analogs of $N_q(t)$, $N(t)$, W_j, and D_j are obtained when

either $t \to \infty$ or $j \to \infty$ and the limiting random variables exist. Loosely speaking, steady-state performance refers to the performance of a system with time-stationary parameters that has been in operation for a sufficiently long time such that time t no longer affects the distributions of number in system, number in different queues, waiting times, and total delay. In contrast, transient queues arise when either system parameters are not time-stationary (therefore a steady state does not exist) or the queueing system does not remain in operation long enough to reach a steady state. If the purpose of the analysis is to obtain performance measures related to transient queues, then that often requires numerical analysis. Many healthcare facilities, such as outpatient clinics, are open for a fixed amount of time during the day and experience time-varying customer arrival patterns. Emergency departments, on the other hand, have demand that varies by the time of day, day of week, and month to month. In such instances, a steady-state may not exist. Still, analysis of steady state behavior can provide useful guidelines for making operational decisions.

Queueing systems in healthcare operations are complex. An example of patient flows through various units of a particular hospital is shown in Fig. 2.1. In this diagram, "out1" denotes the point of entry into the hospital and "out2" denotes the departure point. Ovals represent service stations, each of which is either an inpatient unit or a service department. The numbers shown in the ovals are ward numbers for inpatient units. The rest of the labels can be explained as follows. CL denotes the cath lab, DA refers to direct admits, ED is the emergency department, IR is the interventional radiology department, and PACU is the postanesthesia care unit. The numbers on the connecting arcs are the annual percent of patients that flow in and out of each service station, and arrows show the direction of flow. Each service station provides service to multiple customer classes with different service time distributions (referred to as lengths of stay among inpatient units), different service protocols, and different number of resources (e.g., beds, nurses, and physicians).

Queueing models for systems such as those shown in Fig. 2.1 are intractable unless one makes a number of simplifying assumptions. For these reasons, queueing systems as complex as those shown in Fig. 2.1 are not typically analyzed with the help of mathematical models. Instead, discrete-event simulation, where a computer samples values from different probability distributions to schedule events such as patient arrivals or service completions and keeps track of relevant statistics, is used to analyze such systems and obtain performance metrics. Discrete-event simulation techniques are discussed in Chapter 3 of this book. We focus on relatively simpler models that are tractable and provide useful insights for healthcare operations managers.

The organization of the rest of this chapter is as follows. Basic notation and terminology is introduced in Sect. 2. Single-station models are presented in two sections: Sect. 3 considers models in which there is a single server at each station, whereas Sect. 4 allows multiple servers. Basic results for queueing networks are presented in Sect. 5, and priority queues are discussed in Sect. 6. We conclude the chapter in Sect. 7.

Fig. 2.1 Patient flows through a general hospital

2 Basics

Let $N_q := \lim_{t\to\infty} N_q(t)$, $N := \lim_{t\to\infty} N(t)$, $W := \lim_{j\to\infty} W_j$, and $D := \lim_{j\to\infty} D_j$ denote steady-state distributions of quantities introduced earlier. It is assumed that such limits hold with probability 1. Additionally, we define

$$L = E[N], \tag{2.5}$$

$$L_q = E[N_q], \tag{2.6}$$

$$w = E[W], \qquad \text{and} \tag{2.7}$$

$$d = E[D] \tag{2.8}$$

as time or customer averages. We also define $A(t) =$ the number of customer arrivals during $(0,t]$ and $\lambda = \lim_{t\to\infty} A(t)/t$ as the mean arrival rate. Then, a key result in queueing theory, known as Little's law, is the following relationship:

$$L = \lambda w. \tag{2.9}$$

Little's law is extremely useful for carrying out rough-cut capacity calculations. Consider the following example. Suppose that an emergency department (ED) of a hospital receives on average 50 new patients in each 24 h period. Of the 50 patients, 22 are discharged after examination and treatment in the ED. The remaining 28 are admitted to the hospital as inpatients for further observation and treatment. The average length of inpatient stay is 3 days. Given this information, Little's law allows us to estimate that on average 84 inpatient beds would be needed to serve the needs of patients that are admitted via the ED. This comes from observing that $w = 3$ days, $\lambda = 28$ inpatients per day, and therefore, $L = \lambda w = 84$. Very few assumptions are made when arriving at this result. For example, Little's law remains valid regardless of the priority of service of arriving patients and differences in their service times in the ED.

Single-station queueing systems are often referred to by their four-part shorthand notation $A/B/m/K$, where A and B describe the inter-arrival and service time distributions, $m \in \{1,\ldots,\}$ is the number of servers, and $K \geq m$ is the size of the waiting room including customers in service. The fourth descriptor K is omitted if there is no limit on the size of the queue. Typical values of A and B are as follows: M for exponential, D for deterministic, E_k for Erlang with k phases, PH for phase type, GI for general independent, and G for general. In this notation, $M/G/2/20$ refers to a queueing system consisting of two servers at a single station, exponential inter-arrival times, general service times, and a waiting room capacity of 20.

Literature on queueing systems can be broadly divided into two categories: (1) models that focus on steady-state behavior, that is, stationary distributions N_q, N, D, and W, and (2) models that attempt to characterize transient behavior, that is time-dependent distributions of the number of customers and their waiting times. As mentioned earlier, the vast majority of analytical results pertain to steady-state behavior. Queueing models may be further classified into single- or multiple-station (network) models and those with single or multiple customer classes. In the remainder of this chapter, we provide a summary of key results pertaining to steady-state performance evaluation of single station (with single and multiple servers) and network models. In both types of models, we assume that there is a single customer class. Models with multiple customer classes and service priority are discussed briefly in Sect. 6.

3 Single-Station, Single-Server Models

Single-server queueing models with infinite queues are the workhorses of queueing theory, and among this class of models, the most commonly studied models are the $M/M/1$, $M/G/1$, and $GI/G/1$ models. The popularity of the first two models in this list is in part due to the fact that they are mathematically tractable, which in turn comes from the presence of Markovian property (the M in the model descriptor), and in part from the fact that exponential distribution is a good fit for customer inter-arrival times in many real systems. We describe key results for each of these systems below. Details of analyses that lead to these results can be found in one of several queueing theory books cited earlier. All of these models assume independent and identically distributed inter-arrival and service times. We use the following notation for reporting the key results:

$p(n) = P(N = n),\ n = 0, 1, \ldots,$ the probability distribution of number in system
$F_S(x) = P(S \leq x),\ x \geq 0,$ the CDF of service time distribution
$\mu = 1/E[S],$ the mean service rate
$\rho = \lambda/\mu,$ the server utilization rate
$F_W(x) = P(W \leq x),\ x \geq 0,$ the probability distribution of customer wait time
$F_D(x) = P(D \leq x),\ x \geq 0,$ the probability distribution of customer delay
$F^*(s) = \int_0^\infty \mathrm{e}^{-sx}\mathrm{d}F(x),\ s \geq 0,$ the Laplace–Stieltjes transform (LST) of CDF $F(\cdot)$
$P(z) = \sum_{n=0}^{\infty} z^n p(n),$ the z-transform or probability generating function of $p(n)$

When waiting room is infinite and $\rho \geq 1$, queues can continue to grow over time. If $\rho > 1$, this happens because for each unit of time that the server is available, the average amount of work brought by new arrivals exceeds 1 unit. If $\rho = 1$, queues can still continue to grow because randomness in inter-arrival and service times can cause periods of server idling and the server can never make up for the lost work time. Note that in this instance, for each unit of time that the server is available, the average amount of work brought by new arrivals is exactly 1 unit. The effect of periods of idleness is cumulative and queues continue to grow. In such cases, stationary distributions of N and W do not exist. Therefore, we henceforth assume $\rho < 1$ in all models with infinite waiting room. Finally, we need to specify the service protocol in order to calculate the waiting time distribution. Throughout this section and in Sects. 4 and 5, we assume the first-in-first-out (FIFO) protocol.

3.1 Models with Infinite Waiting Room

In this section, we discuss $M/M/1$, $M/G/1$, and $GI/G/1$ queueing models.

The M/M/1 Model

In healthcare settings, the $M/M/1$ model may prove to be useful either because it fits reality well or because it serves as a reasonable approximation for first-pass analysis. For example, it may be a reasonable choice for modeling walk-in clinics, pharmacy operations, and patient check-in and registration services at hospitals. Similarly, even in situations where customer arrival and service rates vary over time, $M/M/1$ models may be used to estimate capacity requirements to keep peak-period congestion within tolerable limits.

The following distributions and mean performance metrics can be calculated for $M/M/1$ queues from either Chapman–Kolmogorov equations, or an analysis of the embedded Markov chain at customer arrival and/or departure epochs:

$$p(n) = (1-\rho)\rho^n,\ n = 0, 1, \ldots, \tag{2.10}$$

$$E[N] = \frac{\rho}{1-\rho}, \tag{2.11}$$

$$F_D(x) = 1 - \mathrm{e}^{-\mu(1-\rho)x},\ x \geq 0, \tag{2.12}$$

$$E[D] = \frac{1}{\mu(1-\rho)}, \text{ and} \tag{2.13}$$

$$E[W] = \frac{\rho}{\mu(1-\rho)}. \tag{2.14}$$

The analysis also utilizes an important property called PASTA—Poisson arrivals see time averages. Loosely speaking, this property ensures that if system state were observed at moments that coincide with Poisson arrivals, then system properties calculated from these observations are also time average system properties. The PASTA property greatly simplifies the analysis of Markovian queues.

Another important property of $M/M/1$ queues is that the distribution of the number of departures from such queues is also Poisson with parameter λ. It should be clear from the law of conservation of entities that the mean departure rate must equal the mean arrival rate. It is interesting to find that the distribution of departures is also identical to the distribution of arrivals. This property leads to a class of tractable queueing network models called Jackson networks (see Sect. 5).

From expressions (2.11), (2.13), and (2.14), we observe that the expected number in the system, the expected delay, and the expected waiting time are highly sensitive to the server utilization rate. In particular, all three quantities are increasing in ρ at an increasing rate, and as $\rho \to 1$, all three quantities $\to \infty$. This helps explain the fundamental trade-off in queueing systems design. When service times and/or inter-arrival times are random, higher server utilization (efficiency) comes at the cost of increased congestion and customer waiting. High utilization rate may not be economical in situations where waiting cost is very high, and conversely, excess capacity may not be economical when resource cost is significantly higher than

waiting cost. Upon knowing the cost of customer waiting and the cost of providing service resources, designers can find the economic balance between congestion and efficiency.

The M/G/1 Model

In many healthcare settings, it is not appropriate to assume exponentially distributed service times. For example, service times for flu shots, lab services (blood draws), and magnetic resonance imaging (MRI) may not vary much from one patient to another. In the case of surgery practices, it is often found that the lognormal distribution provides a good fit for surgery durations. In such cases, service times may be better modeled by a distribution other than exponential, and the $M/G/1$ model may be more appropriate for calculating queue length and waiting time statistics. For the $M/G/1$ model, the following results have been established:

$$P(z) = \frac{(1-\rho)(z-1)F_S^*(\lambda-\lambda z)}{z-F_S^*(\lambda-\lambda z)}, \tag{2.15}$$

$$E[N] = \frac{\lambda^2 E[S^2]}{2(1-\rho)} + \rho, \tag{2.16}$$

$$F_D^*(s) = \frac{(1-\rho)sF_S^*(s)}{s-\lambda(1-F_S^*(s))}, \tag{2.17}$$

$$E[D] = \frac{\lambda E[S^2]}{2(1-\rho)} + E[S], \tag{2.18}$$

$$E[W] = \frac{\lambda E[S^2]}{2(1-\rho)}. \tag{2.19}$$

Higher moments of the distribution of the number in system and customer delay can be obtained by differentiating $P(z)$ and $F_D^*(s)$. Also, distributions of N and D can be computed numerically by utilizing recent developments in the area of numerical inversions of transforms (see, e.g., Abate and Whitt 1992). However, closed-form expressions for the distribution of N and D are difficult to obtain except for some specific service time distributions.

Expressions (2.16), (2.18), and (2.19) can be recast by expressing $E[S^2]$ in terms of C_s^2, the squared coefficient of variation of service times. Note that the squared coefficient of variation of a random variable is the ratio of its variance to square of its mean. In particular, the expected waiting time expression for $M/G/1$ queues is

$$E[W] = \frac{\rho(1+C_s^2)}{2\mu(1-\rho)}.$$

Upon comparing expressions (2.14) and (2.19), one can better understand the effect of service time variability. When $C_s^2 = 1$, we recover the expected waiting

time in $M/M/1$ queues given in (2.14). All other parameters of the queueing system remaining unchanged, the mean waiting time increases linearly in C_s^2 at rate $\rho/(2\mu(1-\rho))$. This means that the negative effect of C_s^2 on customer waiting time is magnified nonlinearly as ρ increases.

The GI/G/1 Model

The $GI/G/1$ model requires the fewest assumptions about the shape of inter-arrival and service time distributions among the three models we discuss in this section. As such, it is useful in many settings. For example, in addition to examples mentioned before, appointment systems for non-urgent office visits can be modeled as queues in which clinics choose the inter-arrival time of patients. Queueing theory-based approaches for modeling appointment systems utilize either $D/M/1$ or $D/G/1$ queueing models; see surveys of literature in Cayirli and Veral (2003) and Gupta and Denton (2008). We present an application of such models for retail health clinics later in this section.

Analysis of $GI/G/1$ queues requires solving Lindley's integral equation (Lindley 1952). Closed-form solutions, which would be of interest to those interested in the design of healthcare delivery systems, are difficult to obtain except for some specific distributions of inter-arrival and service times. Therefore, many papers have studied approximate methods. Common approaches fall into two categories—(1) approximate either *GI* or *G* by a specific distribution leading to a tractable model and (2) assume a structural form of the distribution of *N* and estimate its parameters. In the first group of methods, commonly studied approximations include Erlang, phase type, and generalized hyperexponential distributions (see, e.g., Neuts (1981, 1989), Li (1997) for details). In the sequel, we present an example of the second approach because it requires knowledge of only the first two moments of inter-arrival and service time distributions and provides greater insights into the key drivers of congestion.

Suppose we can estimate the first two moments of the inter-arrival and service time distributions, but the precise form of these distributions is unknown. In particular, we assume knowledge of C_s^2 and C_a^2, the squared coefficients of variation of service and inter-arrival times. A variety of approximations have been proposed for calculating the mean number in system in $GI/G/1$ queues given C_a^2 and C_s^2. The following is an example of a commonly used expression (see Buzacott and Shanthikumar 1993, p. 75):

$$E[N] \approx \left(\frac{\rho^2(1+C_s^2)}{1+\rho^2 C_s^2}\right)\left(\frac{C_a^2+\rho^2 C_s^2}{2(1-\rho)}\right)+\rho. \tag{2.20}$$

It is easy to verify that when $C_a^2 = 1$, the above expression reduces to the expression we obtained in (2.16) for an $M/G/1$ system. From (2.20), one can also derive expressions for mean waiting time and mean delay with the help of Little's law.

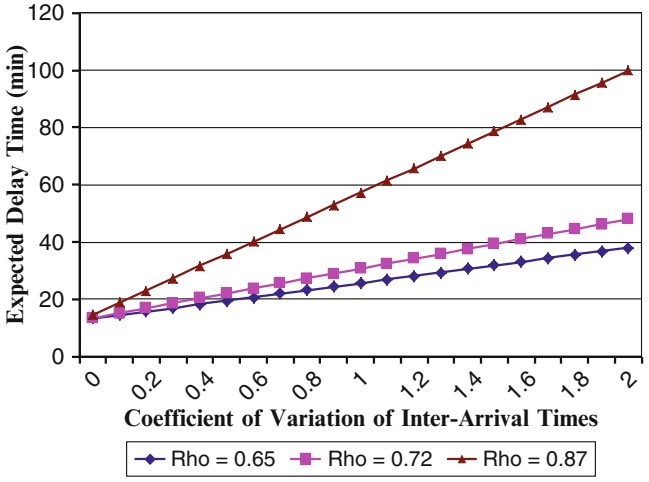

Fig. 2.2 Effect of C_a^2 on expected delay. This figure shows that greater inter-arrival time variability causes expected customer delay to increase much more rapidly when the server has a higher workload (i.e., greater value of ρ)

The expression for $E[N]$ in (2.20) shows that both inter-arrival and service time variability contribute to the congestion in the system and that the effect of variability is magnified by server utilization—the higher the utilization, the greater the effect of variability on congestion. We illustrate the importance of this observation in healthcare operations with the help of an example next.

Retail healthcare clinics, such as MinuteClinic, RediClinics, and Target Clinics, promise to serve patients with routine diagnoses such as ear infections, flu, and minor injuries in a short amount of time without requiring appointments. The success of such clinics depends on providing timely service to a stream of customers that arrive randomly. The study of $GI/G/1$ queueing model can help shine light on the potential benefit to such service providers from reducing inter-arrival time variability by broadcasting time-of-day-dependent estimated waiting times on the web and other promotional media and encouraging customers to time their arrivals when the clinics are not too busy. For example, Target Clinics advise potential visitors about waiting times as follows: "... we are likely to be busiest before and after the average work/school day (8–10 a.m. and 5–7:30 p.m.)" (see response to the frequently asked question "How long will I have to wait to be seen?" at http://sites.target.com/site/en/spot/clinic_faqs.jsp#4). Such practices have the effect of making arrivals more uniform throughout the day, thereby reducing C_a^2. The author obtained data from a retail healthcare clinic chain and studied the effect of variability of arrival pattern on customer waiting times. The results are shown in Fig. 2.2 for three levels of server utilization commonly observed in the data at different clinics. This analysis also shows that if inter-arrival time variability cannot be reduced, clinics need to operate in a range where server utilization would be low in order to keep waiting times from becoming very long.

Without specifying the distributions of inter-arrival and service times, it is not possible to obtain expressions for the distributions of N and D. Therefore, a variety of approximations have been proposed in the literature. In one such approximation, it is assumed that $p(n) = k\sigma^{n-1}$ for $n \geq 1$, where k is a constant, and $p(0) = (1-\rho)$. Note that this was the form of the distribution of number in the system in $M/M/1$ queues. From the requirement that total probability must equal 1, we obtain that $k = \rho(1-\sigma)$. Furthermore, σ can be estimated by equating the calculated mean of the approximate distribution, which equals $\rho/(1-\sigma)$, with the approximate value of $E[N]$ calculated in (2.20). Details of the accuracy of this approximation can be found in Buzacott and Shanthikumar (1993, Sect. 3.3.4).

3.2 Models with Finite Waiting Room

Overcrowding in urgent care clinics and emergency departments is quite common. When waiting rooms become full, patients may leave without receiving service or the service facility may temporarily stop accepting new arrivals. To model such situations, we next consider models in which the maximum number of customers in the system is restricted to K, including the customer in service. When the waiting room limit is reached, one of two possibilities is typically modeled. Either additional arrivals are discouraged until the waiting room is no longer full, or additional arrivals continue to occur but depart immediately upon observing a full waiting room. The latter is identified in the literature as the lost sales model. It turns out that whether arrivals are discouraged or lost makes no difference when arrivals are Poisson. However, the pattern of arrivals when waiting room limit is reached does matter for queues with non-Poisson arrivals. We discuss each of the three basic models next and provide two examples of the usefulness of the $M/M/1/K$ model in healthcare setting. Note that $\rho < 1$ is no longer required for stability of queues. Stability is guaranteed because queue size cannot exceed the size of the waiting room K.

The M/M/1/K Model

The following results are well known for $M/M/1/K$ queueing systems:

$$p(n) = \begin{cases} \dfrac{(1-\rho)\rho^n}{1-\rho^{K+1}} & n = 0, 1, \ldots, K, \\ 0 & \text{otherwise}, \end{cases} \tag{2.21}$$

$$E[N] = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}, \tag{2.22}$$

$$E[D] = E[N]/\lambda. \tag{2.23}$$

The expected waiting time can be calculated from (2.23) and the fact that an average customer spends $E[S]$ in service. Upon comparing (2.22) and (2.23) to their counterparts in (2.11) and (2.13), where the latter allow infinite waiting room, it is easy to see that when all parameters of the two types of queueing systems are identical, both the mean number in the system and the mean delay are smaller in situations where waiting room is limited. This should not come as a surprise because $\lambda p(K)$ fraction of arrivals is not served when waiting room is limited.

The $M/M/1/K$ model can be utilized to make capacity choices for emergency departments. One capacity parameter concerns the number of medical staff, which determines the service rate μ. A second parameter concerns the number of ED beds, which determines the mean waiting time and the number of potential ED arrivals turned away. The latter is sometimes called ambulance diversion. Both types of capacities give rise to different fixed and operating expenses for the hospital. In addition, there are different implications for patient wait times.

For a fixed level of medical staff, that is, fixed service rate, if a hospital increases the number of ED beds, then this would result in greater mean waiting times for patients, but fewer ambulance diversions. Longer wait times can increase a hospital's risk from possible adverse events (e.g., poorer health outcomes and even deaths), and turning away more patients can lower a hospital's revenue because of reduced patient volume, giving rise to the trade-offs we discuss in detail below. Note that the model presented in this chapter is a highly stylized model. EDs are served by teams of multiple doctors and nurses working in parallel, diversions can be caused by a whole host of reasons including lack of availability of inpatient beds, and a variety of regulations may affect a hospital's decision (ability) to go on ambulance diversion. We smooth out such complexities in the discussion that follows.

For each fixed level of ρ, ED managers can develop trade-off curves between ambulance diversions and mean patient waiting times to identify the right combination of capacity parameters, as shown in Fig. 2.3. In this example, it is assumed that if patients are diverted on account of all ED beds being full, then they are able to find appropriate care at other hospitals located in geographical proximity to the hospital in question. If this were not the case, then a network model with strategic capacity choices by administrators of different hospitals will be required (Deo and Gurvich 2011). We do not discuss such models as they are beyond the scope of this chapter.

We consider an ED processing rate of ten patients per hour and two different peak-load scenarios, one with average patient arrival rate of 9 per hour and the other with average patient arrival rate of 9.9 per hour. These scenarios give rise to $\rho = 0.9$ and 0.99, respectively. Figure 2.3 shows that adding more beds (increasing K) increases delay, but reduces the average number of patients turned away, which is denoted as "loss" in Fig. 2.3b. Whereas $E[D]$ increases almost linearly in K, the benefit of having more beds in terms of reduction in expected loss exhibits diminishing returns. That is, each additional bed serves to reduce the expected number of patients turned away by a smaller amount. Service effectiveness can be improved by using triage (prioritizing patients) to identify and serve more critical patients first. We briefly discuss priority queues in Sect. 6.

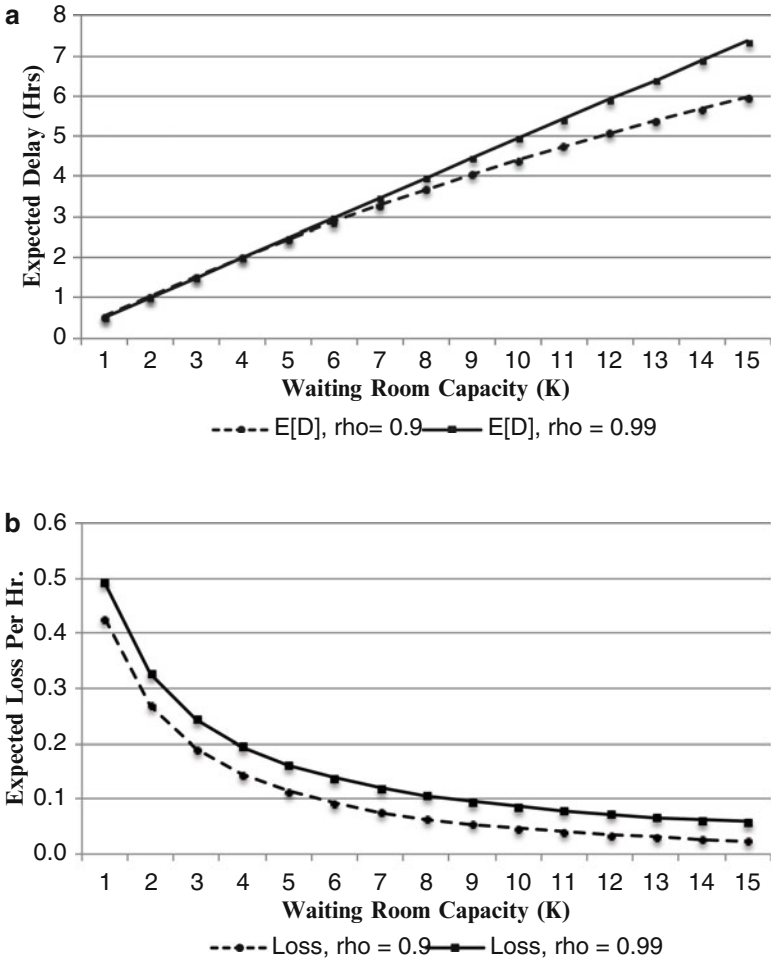

Fig. 2.3 Trade-offs in ED capacity choices

In yet another example, the $M/M/1/K$ model can be used to perform first-cut capacity calculations for physician panel sizes. A panel refers to the list of patients who choose a particular primary care provider as their preferred provider. Suppose a physician implements the *Advanced Access* approach to servingpatients (Chap. 8). In this approach, patients are offered appointments on the day they call, eliminating queues. The physician can serve on average K patients per day and arrival rate is λM where M is the panel size and λ is the incidence rate per patient. Assuming a national average visit rate of 3.356 office visits per patient per year (Hsiao et al. 2010), and

Table 2.1 Panel size (P) and service capacity (K)

K	16	17	18	19	20	21	22	23	24
P	1,208	1,296	1,386	1,463	1,540	1,617	1,701	1,779	1,857

260 working days per year (52 weeks, 5 days per week), we obtain $\lambda = 0.01291$. Suppose the physician is willing to accept the possibility that 5% of the patients who call on any given day would not be accommodated that day, that is, the overflow rate should not exceed 0.05. From the analysis of $M/M/1/K$ model, above, we know that the overflow rate is simply $p(K) = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}}$. Setting this quantity equal to 0.05, we can calculate ρ and subsequently M because $\mu = K$ patients per day. Upon performing these calculations, we obtain estimates of maximum panel sizes for different values of K as shown in Table 2.1.

The modeling approach described above can be refined to include patient no-shows and advance-book appointments (see, e.g., Green and Savin 2008; Robinson and Chen 2010).

The M/G/1/K Model

The classical analysis of $M/G/1/K$ queues relies on the embedded Markov chain observed at service completion epochs. This results in a series of equations relating state probabilities, which can be solved numerically along with the normalization equation (probabilities must sum to 1) to obtain the steady-state distribution of the number in the system. Closed-form expressions are often difficult to obtain. An alternative is to use the method of transforms, which was utilized in the section dealing with $M/G/1$ queues. That method is also primarily a numerical approach. Because our goal in this chapter is to shine light on the insights that queueing models provide for healthcare operations managers, we focus on a subclass of $M/G/1/K$ queueing models in which $\rho < 1$. Recall that $\rho < 1$ is not required for the existence of steady-state distributions of queue congestion and customer waiting times when waiting rooms are finite.

Given $\rho < 1$, let $p_\infty(n)$ and $\bar{P}_\infty(K)$ denote, respectively, the distribution of number in system and the CCDF of this distribution at K in an $M/G/1$ queue. The subscript "∞" emphasizes the fact that these quantities refer to the case in which there is no limit on the size of the waiting room. Then, the distribution of N can be obtained as follows (see Buzacott and Shanthikumar 1993, p. 109 for details):

$$p(n) = \begin{cases} \dfrac{p_\infty(n)}{1-\rho\bar{P}_\infty(K)}, & n = 0, 1, \ldots, K-1, \\ \dfrac{(1-\rho)\bar{P}_\infty(K)}{1-\rho\bar{P}_\infty(K)} & n = K. \end{cases} \tag{2.24}$$

Expected number in system and mean delay can be calculated from the above expression. Remarkably, when $\rho < 1$, the distribution of number in system in an $M/G/1/K$ queueing system is proportional to the number in system in an $M/G/1/\infty$ system. The proportionality constant is $\frac{1}{1-\rho\tilde{P}_\infty(K)}$ for $n = 0, 1, \ldots, K-1$, and $\frac{(1-\rho)}{1-\rho\tilde{P}_\infty(K)}$ for $n = K$.

The GI/G/1/K Model

The $GI/G/1/K$ model is more difficult to analyze, except when inter-arrival and service time distributions have some specific forms. Therefore, papers dealing with the analysis of $GI/G/1/K$ queueing systems propose a variety of approximations. A useful approximation that imposes the relationship between $M/M/1$ and $M/M/1/K$ models onto the relationship between $GI/G/1$ and $GI/G/1/K$ models is presented in Buzacott and Shanthikumar (1993, pp. 110–116). We omit the details in the interest of brevity.

4 Single-Station, Multiple-Server Models

In all of the examples mentioned in Sect. 3, for example, walk-in clinics, pharmacy and lab services, and emergency departments, situations involving multiple servers arise naturally. That is, service facilities in a healthcare setting often have multiple servers taking care of customers who queue up for similar services. In this section, we focus on queueing systems with unlimited waiting room, constant arrival rate λ, and m identical servers. Servers process one customer at a time, and each server can process customers at rate μ when busy. In this case, the overall service rate is a function of the number in system. In particular, if there are n customers in the system, then μ_n, the state-dependent service rate is

$$\mu_n = \begin{cases} n\mu & \text{if } 0 \le n \le m, \\ m\mu & \text{otherwise.} \end{cases} \tag{2.25}$$

The overall server utilization in this instance is $\rho = \frac{\lambda}{m\mu}$, and stationary distributions of N, N_q, D, and W exist if $\rho < 1$. Because it is more difficult to obtain closed-form expressions for performance metrics of multiple-server queueing systems, we focus in this section on $M/M/m$ models, that is, on situations in which the inter-arrival and service times are exponentially distributed. For $M/M/m$ models, it can be shown that

$$p(n) = \begin{cases} p(0)\dfrac{(m\rho)^n}{n!} & \text{if } n \le m, \\ p(0)\dfrac{\rho^n m^m}{m!} & n \ge m, \end{cases} \tag{2.26}$$

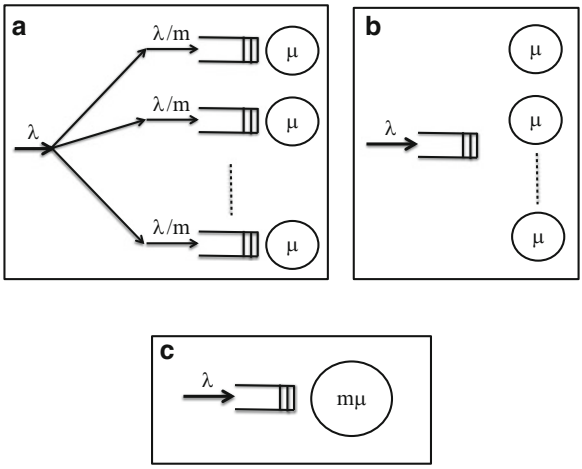

Fig. 2.4 Comparison of server and queue configurations

where

$$p(0) = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \left(\frac{(m\rho)^m}{m!}\right)\left(\frac{1}{1-\rho}\right)\right]^{-1}, \tag{2.27}$$

$$E[D] = \left(\frac{1}{m\mu(1-\rho)}\right)\left(\frac{(m\rho)^m}{m!(1-\rho)}\right)p(0) + \frac{1}{\mu}. \tag{2.28}$$

Expressions for mean delay in the system shown in (2.13) and (2.28) can be utilized to obtain a key result in the design of queueing systems—that combining queues and pooling servers reduces system delay. To demonstrate this, we show an example next in which three systems are compared. A schematic of these systems is shown in Fig. 2.4. Systems are labeled A, B, and C. System A consists of m parallel queues each served by a single server. When a customer arrives, it is routed by a Markovian router that sends the customer to any one of the m queues with equal probability $1/m$. Each server's processing rate is μ. In system B, the queues are combined into a single queue. Customers wait for the next available server upon joining the queue. Each server's processing rate is μ. In system C, queues are combined into a single queue, and the m servers are replaced by a super server with a faster processing rate of $m\mu$. Note that the overall utilization rate remains $\rho = \lambda/m\mu$ in all three systems.

Choices similar to those depicted in Fig. 2.4 can arise in a number of different contexts in the healthcare setting. For example, the three configurations could represent choices for setting up patient registration and check-in counters at a hospital or clinic. System A represents a case in which each server specializes in serving a particular type of patient arrivals. In system B, each server can serve any arrival, and finally in system C, technology may be employed to assist a server,

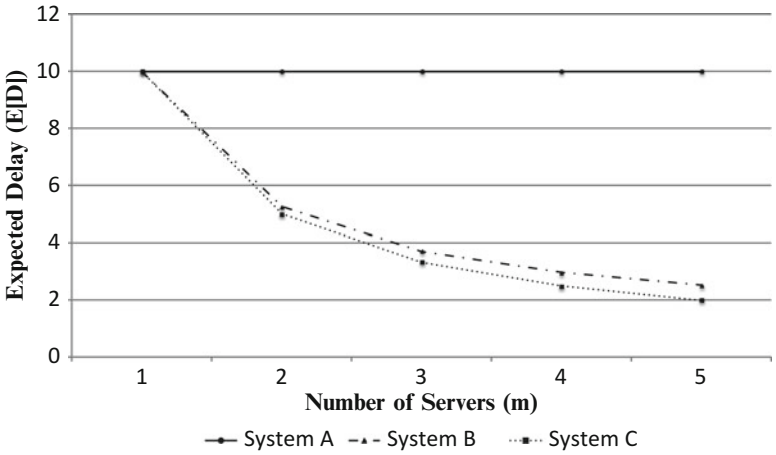

Fig. 2.5 Effect of server/queue pooling

making that server faster. Alternatively, servers in systems A and B could represent the choice between specialized and general-purpose beds in an inpatient unit. In each example, the choice of configuration affects labor and capital costs as well as patient wait times. Queueing models can help compare the impact of these configurations on patient wait times.

In system A, an arbitrary arrival joins one of the m separate queues with equal probability. Therefore, an arbitrary arrival's expected delay (using (2.13)) can be written as follows:

$$E[D^{(A)}] = \frac{\rho}{\mu(1-\rho)} + \frac{1}{\mu}. \tag{2.29}$$

The delay experienced by an arbitrary arrival in system B is as shown in (2.28). Finally, in system C, the expected delay equals

$$E[D^{(C)}] = \frac{\rho}{m\mu(1-\rho)} + \frac{1}{m\mu}. \tag{2.30}$$

To compare the mean delay in the three systems, consider an example in which m is systematically varied from 1 through 5 while keeping $\rho = 0.9$ fixed. This can be achieved by setting $\mu = 1$ and varying λ as m varied. In particular, with $\mu = 1$, set $\lambda = m\rho$ to maintain a fixed ρ. Next, suppose we use the expressions derived above to calculate mean delay and plot the values in Fig. 2.5. Note that the delay in system A remains invariant in m. This is expected because customer waiting occurs in one of the m queues, each of which is independent of m. We observe a significant improvement from combining queues, and a further improvement (though much smaller) from creating a single faster server. Most of the benefits

appear to come from combining queues. This example serves to highlight a general principal for the design of healthcare service systems; namely, combining queues improves efficiency.

A common reason for creating separate queues is that each queue is served by a group of servers who possess special expertise to serve a particular type of customers. Combining queues requires that servers be trained to serve all types of customers. This can be expensive, and such considerations have motivated the study of partial pooling arrangements. Reasons why combining queues is not always beneficial can be found in Rothkopf and Rech (1987), and general principles of work design and pooling have been discussed in Buzacott (1996) and Mandelbaum and Reiman (1998).

Before closing this section, we briefly discuss the Erlang loss formula, which can be used to calculate overflow probability in $M/M/m/m$ systems. These are multiple-server queueing systems in which $K = m$, that is, waiting room size equals the number of servers. In such cases, a customer is lost if no server is available to serve this customer immediately. Erlang loss formula is of great interest in telephony, where it is used to calculate the number of telephone lines needed to accommodate a desired fraction of incoming calls. Erlang loss, or the probability of finding all m servers busy, is given by

$$p(m) = \frac{\rho^m}{m!(\sum_{k=0}^{m} \rho^k/k!)}. \tag{2.31}$$

Erlang loss formula has been used to model capacity requirements of EDs where each ED bay (bed plus care team) is treated as a server.

5 Network Models

Hospitals and specialized treatment facilities for particular medical conditions (e.g., cancer, cardiovascular, or neurological services) perform a range of services, each with its own resources (servers) and queues. Such facilities are best modeled as networks of queues. The simplest network model from the viewpoint of obtaining analytical results is the multi-station (network) analog of the $M/M/1$ queueing system with J service stations, each with a single server. Customers may arrive either from outside or move from one station to another. Suppose server i's service rate is μ_i, customers are routed from station i to j according to probability r_{ij}, and exogenous arrival rate at station i is γ_i. Then, the effective arrival rate at station i is

$$\lambda_i = \gamma_i + \sum_{1 \le j \le J} \lambda_j r_{ji}, \qquad \text{for each } 1 \le i \le J. \tag{2.32}$$

Similarly, the server utilization rate is $\rho_i = \lambda_i/\mu_i$, for each i. A network is called *open* if it allows exogenous arrivals and departures. Departures can be modeled by

designating a particular station to serve as a sink, say station indexed J, such that $r_{Ji} = 0$ for all i. Similarly, a network is called *closed* if no customers can enter or leave the network. In this case $\gamma_i = 0$ and there is no sink.

The state of the number in different stations of a network is a vector $n = (n_1, \ldots, n_J)$. Let $p_i(n_i)$ denote the distribution of number in system of a $M/M/1$ queue with parameters λ_i and μ_i. Then, a key result in the analysis of Markovian open networks is that the stationary distribution $p(n)$ has a product form. In particular,

$$p(n) = p_1(n_1) \times p_2(n_2) \cdots \times p_J(n_J), \tag{2.33}$$

where $p_i(n_i) = (1 - \rho_i)\rho_i^{n_i}$. Networks for which the distribution of number in the system has a product form are also called product-form or Jackson networks (see Jackson (1954, 1957)). The product-form structure remains intact when there are m_i servers at each station, that is, we have a network of $M/M/m_i$, $1 \leq i \leq J$, queues. Because of the existence of a product form, the results from the analysis of $M/M/1$ and $M/M/m$ queues can be applied directly to such networks. For example,

$$E[N] = \sum_{i=1}^{J} \frac{\rho_i}{1 - \rho_i}, \tag{2.34}$$

$$E[D] = \frac{1}{\lambda} \sum_{i=1}^{J} \frac{\rho_i}{1 - \rho_i}. \tag{2.35}$$

Queueing network models have been studied extensively (Walrand 1988), and there are numerous manufacturing applications of these models (Buzacott and Shanthikumar 1993). Many of these models are not directly applicable to health systems design. Each model is specific to a particular type of system (e.g., transfer lines with limited buffer) and typically requires either special techniques or approximations to derive system performance measures. Therefore, we focus only on papers that utilize queueing network methodology for modeling healthcare operations.

Whereas there are many attempts to represent networks of healthcare facilities as networks of queues, these networks are typically not analyzed using queueing-theoretic approaches. Instead, a common approach is to use computer simulation to obtain performance metrics of interest; for examples, see Taylor and Keown (1980), Harper and Shahani (2002), and Feck et al. (1980). Papers that use an analytic approach include Hershey et al. (1981) and Weiss and McClain (1987). Hershey et al. (1981) present a methodology for estimating expected utilization and service level for a class of capacity-constrained service network facilities operating in a stochastic environment. In this paper, queues are not allowed to form, that is, waiting room size equals the number of servers at each facility, and the authors use the Erlang loss formula to approximate the probability of overflow. They show that their calculation is exact for two cases and recommend its use as an approximation in the general case.

Weiss and McClain (1987) model the transition of care from acute to extended care (e.g., a nursing home or community care center). Inadequate capacity at

downstream service facilities can lead to extra wait in the acute care facility, which is often referred to as "administrative days." The authors use a queueing-analytic approach to describe the process by which patients await placement. They model the situation using a state-dependent placement rate for patients backed up in the acute care facility. Using data from seven hospitals in New York State, the study also derives policy implications.

6 Priority Queues

There are many variants of the basic models described in the preceding sections. These may consider different types of service priority (Jaiswal 1968; Takagi 1986, 1990, 1994; Gupta and Gunalay 1997), server vacations (Tian and Zhang 2006), and bulk arrivals and batch service (Chaudhry and Templeton 1986). The literature on each of these topics is vast. In this section, we discuss an assortment of results from priority queues and discuss their implications for healthcare operations.

Suppose in a single-server queueing system, there are k customer classes, indexed by $\ell = 1, \ldots, k$. Type ℓ customers arrive according to an independent Poisson process with rate $\lambda^{(\ell)}$, and their service time distribution is $S^{(\ell)}$. An arrival observes N^{ℓ} type-ℓ customers in the system upon arrival. Therefore, the total expected work in the system at an arbitrary arrival epoch is $\sum_{\ell=1}^{k} E[N^{(\ell)}]E[S^{(\ell)}] - \sum_{\ell=1}^{k} \frac{1}{2}\lambda^{(\ell)}E[(S^{(\ell)})^2]$, where the second term is the amount of work already completed on the customer in service, if any. After some simplification and using Little's law, the total expected work can be expressed as $\sum_{\ell=1}^{k} \rho^{(\ell)}E[D^{(\ell)}] - \sum_{\ell=1}^{k} \frac{1}{2}\lambda^{(\ell)}E[(S^{(\ell)})^2]$, where $\rho^{(\ell)} = \lambda^{(\ell)}E[S^{(\ell)}]$. If the service protocol is work conserving, that is, it neither creates nor destroys work, then the expected total work must be constant. This immediately implies that

$$\sum_{\ell=1}^{k} \rho^{(\ell)}E[D^{(\ell)}] = \text{constant} \tag{2.36}$$

because $\sum_{\ell=1}^{k} \frac{1}{2}\lambda^{(\ell)}E[(S^{(\ell)})^2]$ is independent of service protocol. The queue is stable so long as $\sum_{\ell=1}^{k} \rho^{(\ell)} < 1$, which we assume throughout this section.

The above relationship establishes an important property of priority queues. When service protocol is work conserving, a particular priority scheme may affect delays experienced by different customer classes, but reduction in the expected delay of one customer class is realized at the expense of increase in the expected delay of another class. The work conservation principle is violated when switching from one customer class to another requires setup or switchover time, and/or some amount of work is lost if service of one customer type is preempted by a higher priority customer. Because switchover or setup times are common when a server attends to customers with different service requirements, pseudo work conservation laws have been derived for queues with switchover times and certain service protocols, for example, cyclic priority (see Takagi 1986,

1990, 1994; Gupta and Gunalay 1997). Although queues with switchover times are not work conserving, the amount of additional work created by certain switching protocols can be fully characterized. The expected total work in the system is then the sum of two components—work associated with customer arrivals, which is independent of service protocol, and work associated with switching regime, which is dictated by the priority scheme. For this reason, such queues are said to satisfy pseudo conservation laws.

In healthcare applications, one finds examples of both preemptive and non-preemptive priority. For example, ED physicians often serve the most critical patients preemptively in order to save lives. In the outpatient setting, specialists reserve certain appointment slots for high-priority patients, but once a low-priority patient books an appointment, he or she is rarely preempted during service. In many situations, service protocols may not be work conserving because priority rules may increase service providers' walking time to reach patients located in different inpatient units. Finally, service protocol may be static or dynamic. In a static protocol, each customer class has a fixed priority, and its members receive a strictly higher priority over all lower-ranked classes. In contrast, in dynamic priority protocols, customers of a particular class may be higher ranked at one time and lower ranked at another. A common dynamic priority protocol is one in which customers of each class form a separate queue, the server moves from one queue to another, and when serving a particular queue, all customers of that class (previously waiting or new arrivals) have higher priority. Such a queueing protocol is observed in healthcare operations when a physician travels to different community-based clinics on different days of week.

Next, we provide some basic results that allow an operations manager to quickly calculate mean delay experienced by customers of different classes in a system with a single server, Poisson arrivals, and work-conserving, non-preemptive, and static-priority service protocol. Suppose customers classes are arranged in the order of priority, that is, class-1 customers have the highest priority, followed by class 2, and so on until class k. Then, an arbitrary class-1 customer waits only for class-1 customers already in the queue when it arrives. Moreover, because of PASTA property, we have

$$E[W^{(1)}] = E[N_q^{(1)}]E[S^{(1)}] + \frac{\lambda E[S^2]}{2}, \tag{2.37}$$

where $\frac{\lambda E[S^2]}{2} = \sum_{\ell=1}^{k} \frac{1}{2}\lambda^{(\ell)} E[(S^{(\ell)})^2]$ is the amount of work remaining to complete the service of the customer in service. Introducing new notation $w_0 = \frac{\lambda E[S^2]}{2}$, $\lambda_r = \sum_{\ell=1}^{r} \lambda^{(\ell)}$, and $\rho_r = \sum_{\ell=1}^{r} \rho^{(\ell)}$, and using the fact that $E[W^{(1)}] = \frac{E[N_q^{(1)}]}{\lambda^{(1)}}$, we can rearrange (2.37) to obtain

$$E[N_q^{(1)}] = \frac{\lambda^{(1)} w_0}{(1-\rho_1)}. \tag{2.38}$$

Consider next an arbitrary class-2 arrival. This customer will be served only after all earlier-arriving class-1 and class-2 customers are served, which causes an initial

delay of $E[N_q^{(1)}]E[S^{(1)}] + E[N_q^{(2)}]E[S^{(2)}] + w_0$. In addition, it must also wait until all those type-1 customers who arrive during $E[N_q^{(1)}]E[S^{(1)}] + E[N_q^{(2)}]E[S^{(2)}] + w_0$, and additional class-1 arrivals during the service time of those, and so on, are served. It turns out that the corresponding delay has mean duration

$$(1/(1-\rho_1))\left\{E\left[N_q^{(1)}\right]E[S^{(1)}] + E\left[N_q^{(2)}\right]E[S^{(2)}] + w_0\right\}.$$

That is, $E[W^{(2)}] = (1/(1-\rho_1))\{E[N_q^{(1)}]E[S^{(1)}] + E[N_q^{(2)}]E[S^{(2)}] + w_0\}$. Upon using $E[N_q^{(2)}] = \lambda^{(2)}E[W^{(2)}]$ and simplifying, we obtain

$$E\left[N_q^{(2)}\right] = \frac{\lambda^{(2)}w_0}{(1-\rho_1)(1-\rho_2)}. \tag{2.39}$$

Continuing in the same fashion, we obtain for $\ell = 1, \ldots, k$,

$$\begin{aligned} E[N_q^{(\ell)}] &= \frac{\lambda^{(\ell)}w_0}{(1-\rho_{\ell-1})(1-\rho_\ell)} \\ &= \frac{\lambda^{(\ell)}\lambda E[S^2]}{2(1-\rho_{\ell-1})(1-\rho_\ell)}, \end{aligned} \tag{2.40}$$

where $\rho_0 = 0$. Note the similarity between the above expression and the mean number in the queue for $M/G/1$ systems, where the latter can be calculated from (2.16).

Suppose w_ℓ is the cost of making a type-ℓ customer wait for service. What is an optimal priority rule that minimizes total waiting cost? This question has been addressed in the queueing literature, and it has been shown that class priority should be proportional to $w\mu$, that is priority index should be such that $w_1\mu_1 \geq w_2\mu_2 \geq \cdots \geq w_k\mu_k$. This means that customers with higher waiting cost per unit time and shorter mean processing time should be given higher priority. If all customer classes have the same per-unit time waiting cost, then customers with shorter mean processing time would be processed first. This is also called the shortest-processing-time-first rule.

7 Concluding Remarks

Many types of healthcare service systems are characterized by random demand (in timing and type of services required); time-varying and uncertain availability of service resources due to preferred work patterns, work rules, and planned or unplanned absences; and service protocols that assign different priorities to different customer classes (e.g., urgent versus nonurgent patients). These are precisely the types of environments in which queueing theory can be brought to bear to obtain

useful insights for system design and for developing operating principles for service delivery systems. It is no surprise that queueing theory has been used extensively in healthcare operations. The following is a list of key topic areas and some recent papers on each of these topics:

1. Capacity calculations (matching supply and demand)
   (a) Panel size determination—See the discussion in Sect. 3.2, Green and Savin (2008), Robinson and Chen (2010), Gupta and Wang (2011).
   (b) ED beds—See the discussion in Sect. 3.2, Deo and Gurvich (2011) and references therein.
   (c) Network capacity—See Hershey et al. (1981) and Weiss and McClain (1987).
   (d) Nurse staffing—See Yankovic and Green (2011) and references therein.
2. Scheduling arrivals (appointment systems)—See Gupta and Denton (2008) for a review of queueing-analytic approaches
3. Priority queues (allocation of organs to transplant candidates)—See Su and Zenios (2004) and references therein.

Some of the above-mentioned problems were not discussed in this chapter because of the specialized institutional background necessary to introduce the key operations management challenges.

Notwithstanding the success of queueing models for addressing important questions in the delivery of healthcare services, there remain significant opportunities for new models and analytic tools. For example, hospitals can benefit from insightful network models of patient flow (recall Fig. 2.1). Hospitals have limited capacity within each inpatient unit, which leads to blocking at upstream units. This situation may be resolved by keeping patients longer in some units, placing patients in less-than-ideal units, transferring patients to other hospitals, or refusing admissions. Such decisions can affect lengths of stay and health outcomes (Rincon et al. 2011; Sinuff et al. 2004). Clearly, hospitals could benefit from knowing the performance implications of such practices and having access to models that allow them to factor nursing units' flexibility into capacity calculations. In the manufacturing setting, there are numerous dynamic job shop models that address similar problems; see, for example, Chap. 7 in Buzacott and Shanthikumar (1993). However, the number of similar models for hospital operations is quite limited and represents an opportunity for future research.

References

Abate J, Whitt W (1992) The fourier-series method for inverting transforms of probability distributions. Queueing Syst 10:5–88

Bhat UN (2008) An introduction to queueing theory modeling and analysis in applications. Springer, Boston [distributor]

Buzacott JA (1996) Commonalities in reengineered business processes: Models and issues. Manag Sci 42:768–782
Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice Hall, Englewood Cliffs
Cayirli T, Veral E (2003) Outpatient scheduling in health care: A review of literature. Prod Oper Manag 12(4):519–549
Chaudhry ML, Templeton JGC (1986) Bulk queues. Department of Mathematics, McMaster University, Hamilton
Cohen JW (1969) The single server queue. North-Holland, Amsterdam
Cox DR, Smith WL (1961) Queues. Chapman and Hall, London; Distributed in the USA by Halsted Press, London
Deo S, Gurvich I (2011) Centralized vs. decentralized ambulance diversion: A network perspective. Manag Sci 57:1300–1319
Feck G, Blair EL, Lawrence CE (1980) A systems model for burn care. Med Care 18(2):211–218
Green LV, Savin S (2008) Reducing delays for medical appointments: A queueing approach. Oper Res 56(6):1526–1538
Gross D, Harris CM (1985) Fundamentals of queueing theory. Wiley, New York
Gupta D, Denton B (2008) Appointment scheduling in health care: Challenges and opportunities. IIE Trans 40:800–819
Gupta D, Gunalay Y (1997) Recent advances in the analysis of polling systems. In: Balakrishnan N (ed) Advances in combinatorial methods and applications to probability and statistics. Statistics in industry and technology series. Birkhauser, Boston
Gupta D, Wang WY (2011) Patient appointments in ambulatory care. In: Hall RW (ed) Handbook of healthcare system scheduling: Delivering care when and where it is needed. Springer, New York, chap 4
Harper PR, Shahani AK (2002) Modelling for the planning and management of bed capacities in hospitals. J Oper Res Soc 53(1):11–18
Hershey JC, Weiss EN, Cohen MA (1981) A stochastic service network model with application to hospital facilities. Oper Res 29(1):1–22
Hsiao CJ, Cherry DK, Beatty PC, Rechtsteiner EA (2010) National ambulatory medical survey report: 2007 summary. National Health Statistics Reports, Number 27. Available on the web at http://www.cdc.gov/nchs/data/nhsr/nhsr027.pdf. Cited 7 March 2011
Jackson JR (1957) Networks of waiting lines. Oper Res 5:518–521
Jackson RRP (1954) Queueing systems with phase-type service. Oper Res Q 5:109–120
Jaiswal NK (1968) Priority queues. Academic, New York
Larson RC (1987) Perspectives on queues: Social justice and the psychology of queueing. Oper Res 35(6):895–905
Li J (1997) An approximation method for the analysis of GI/G/1 queues. Oper Res 45(1):140–144
Lindley DV (1952) On the theory of queues with a single server. Proc Camb Philos Soc 48:277–289
Mandelbaum A, Reiman MI (1998) On pooling in queueing networks. Manag Sci 44(7):971–981
Morse PM (1958) Queues, inventories, and maintenance; the analysis of operational systems with variable demand and supply. Wiley, New York
Neuts MF (1981) Explicit steady-state solutions in stochastic models: An algorithmic approach. The Johns Hopkins University Press, Baltimore
Neuts MF (1989) Structured stochastic matrices of M/G/1 type and their applications. Marcel Dekker, New York
Newell GF (1982) Applications of queueing theory. Chapman and Hall, London
Rincon F, Morino T, Behrens D, Akbar U, Schorr C, Lee E, Gerber D, Parrillo J, Mirsen T (2011) Association between out-of-hospital emergency department transfer and poor hospital outcome in critically ill stroke patients. J Crit Care 26(6):620–625
Robinson LW, Chen RR (2010) A comparison of traditional and open-access policies for appointment scheduling. Manuf Serv Oper Manag 12:330–346
Rothkopf MH, Rech P (1987) Perspectives on queues: Combining queues is not always beneficial. Oper Res 35:906–909

Sinuff T, Kahnamoui K, Cook DJ, Luce JM, Levy MM (2004) Rationing critical care beds: A systematic review. Crit Care Med 32(7):1588–1597
Su X, Zenios S (2004) Patient choice in kidney allocation: the role of the queueing discipline. Manuf Serv Oper Manag 6:280–301
Takacs L (1962) Introduction to the theory of queues. Oxford University Press, New York
Takagi H (1986) Analysis of polling systems. MIT, Cambridge
Takagi H (1990) Queueing analysis of polling models: An update. In: Takagi H (ed) Stochastic analysis of computer and communication systems. North-Holland, Amsterdam, pp 267–318
Takagi H (1994) Queueing analysis of polling models: Progress in 1990–93. Institute of Socio-Economic Planning, University of Tsukuba, Japan
Taylor BW III, Keown AJ (1980) A network analysis of an inpatient/outpatient department. J Oper Res Soc 31(2):169–179
Tian N, Zhang ZG (2006) Vacation queueing models. Springer, New York
Walrand J (1988) An introduction to queueing networks. Prentice Hall, Englewood Cliffs
Weiss EN, McClain JO (1987) Administrative days in acute care facilities: A queueing-analytic approach. Oper Res 35(1):35–44
Wolff RW (1989) Stochastic modeling and the theory of queues. Prentice Hall, Englewood Cliffs
Worthington D (2009) Reflections on queue modelling from the last 50 years. J Oper Res Soc 60:s83–s92
Yankovic N, Green LV (2011) Identifying good nursing levels: a queuing approach. Oper Res 59(4):942–955

http://www.springer.com/978-1-4614-5884-5

Handbook of Healthcare Operations Management
Methods and Applications
Denton, B.T. (Ed.)
2013, X, 536 p. 72 illus., 53 illus. in color., Hardcover
ISBN: 978-1-4614-5884-5