1.5 Queuing Models

Queuing theory is a branch of mathematics that studies and models the act of waiting in lines. The first paper on queuing theory, "The Theory of Probabilities and Telephone Conversations" was published in 1909 by A.K. Erlang, now considered the father of the field. He pondered the problem of determining how many telephone circuits were necessary to provide phone service that would prevent customers from waiting too long for an available circuit.

In developing a solution to this problem, he began to realize that the problem of minimizing waiting time was applicable to many fields, and began developing the theory further. Erlang's **switchboard problem** laid the path for modern queuing theory.

Queueing theory boils down to answering simple questions like the following:

- How likely is it that objects/units/persons will queue up and wait in line?
- How long will the line be?
- How long will the wait be?
- How busy will the server/person/system servicing the line be?
- How much capacity is needed to meet an expected level of demand?

Knowing how to think about these kinds of questions will help clients anticipate **bottlenecks**. As a result, they will build systems and teams to be more efficient and more scalable, to have higher performance and lower costs, and to ultimately provide better service to their customers.

Queueing theory also allows for the quantitative treatment of bottlenecks and effect on performance. For instance, a question such as "how long will the wait be, on average?" will have an answer, but so will other questions concerning the variability of wait times, the distribution of wait times, and the likelihood that a customer someone will receive extremely poor service, and so on [9].

Let us consider a simple example. Suppose a grocery store has a single checkout line and a single cashier. If, on average, one shopper arrives at the line to pay for their groceries every 5 minutes and if scanning, bagging, and paying takes 4.5 minutes, on average, will customers have to wait in line? When the problem is presented this way, our intuition says that there should be no waiting in line, and that the cashier should be idle, on average, 30 seconds every 5 minutes, only being busy 90% of the time. No one ever has to wait before being served!

If you have ever been in grocery store, you know that's not really what happens in reality; many shoppers will be waiting waiting in line, and they will have to wait a long time before being processed. Fundamentally, queueing happens for three reasons:

- **irregular arrivals** – shoppers do not arrive at the checkout line on a regular schedule; they are sometimes spaced far apart and sometimes close together, so they **overlap** (an overlap automatically causes queueing and waiting);
- **irregular job sizes** – shoppers do not all get processed in 4.5 minutes; somebody shopping for a large family will require much more time, for instance (when this happens, overlap is again a problem because new shoppers will arrive and be ready to check out while the existing ones are still in progress), and

- **waste** – lost time can never be regained; shoppers overlap because the second shopper arrived before the first shopper had the time to finish, but looking at it the other way, perhaps it's not the second shopper's fault; perhaps the first shopper should have arrived earlier, but wasted time reading a gossip magazine while the cashier was idle! They missed their chance for quick service and, as a result, made the second shopper have to wait.

In general, **irregular** arrival times and job sizes are guaranteed to cause queueing. The only time there is no queueing is when the job sizes are uniform, the arrivals are timed evenly, and there is little enough work for the cashier to keep up. Even when the cashier is barely busy at all, irregular arrivals or arrivals **in bursts** will cause some queueing.

Queueing gets worse when the following is true of the system:

- **high utilisation** – the busier the cashier is, the longer it takes to recover from wasted time;
- **high variability** – the more variability in arrivals or job sizes, the more waste and the more overlap (queueing) occurs, and
- **insufficient servers** – fewer cashiers means less capacity to absorb arrival spikes, leading to more wasted time and higher utilisation.

1.5.1 Terminology

Queueing theory studies systems and processes in terms of three key concepts:

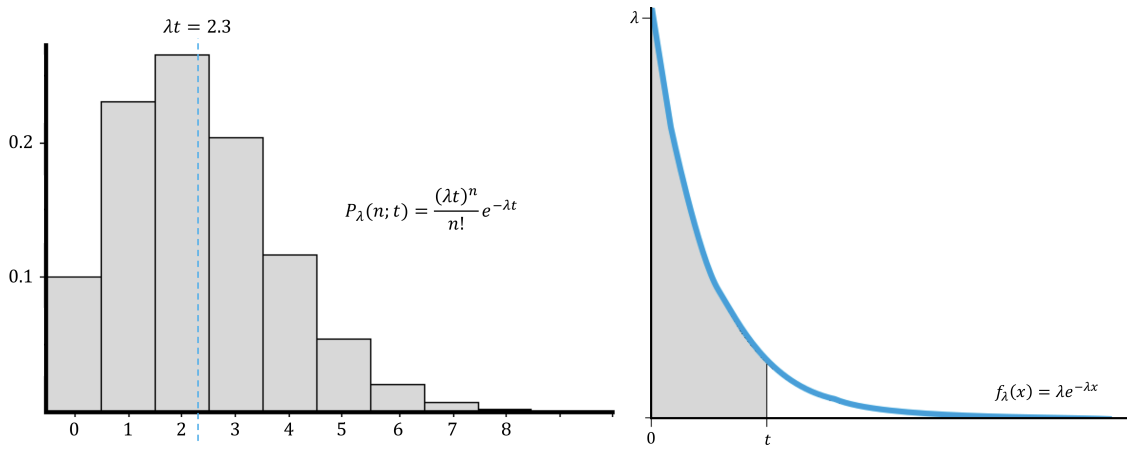
- **customers** are the units of work that the system serves – a customer can be a real person, or it can be whatever the system is supposed to process and complete: a web request, a database query, a part to be milled by a machine, etc.;
- **servers** are the objects that do the processing work – a server might be the cashier at the grocery store, a web server, a database server, a milling machine, etc., and
- **queues** are where the units of work wait if the server is busy and can not start the work as they arrive – a queue may be a physical line, or reside in memory, etc.

To begin understanding and describing queues, we must first have know and understand some useful probability distributions, as well as input and output processes.

Exponential and Poisson Probability Distributions The **Poisson** and **exponential** distributions play a prominent role in queuing theory. The Poisson distribution counts the number of discrete events occurring in a fixed time period; it is closely connected to the exponential distribution, which (among other applications) measures the time between arrivals of the events. The Poisson distribution is a discrete distribution; the random variable can only take non-negative integer values. The exponential distribution can take any (nonnegative) real value.

Consider the problem of determining the probability of n arrivals being observed during a time interval of length t, where the following assumptions are made:

- the probability that an arrival is observed during a small time interval (say of length ν) is proportional to the length of interval; let the proportionality constant be λ, so that the probability is $\lambda\nu$;



(a) Poisson distribution with $\lambda t = 2.3$. **(b)** Exponential distribution with parameter λ.

Figure 1: Poisson and exponential distributions. The shaded area (on the right) represents the probability that a customer will wait up to the length of the time interval t.

- the probability of two or more arrivals in a small interval is zero, and
- the number of arrivals in any time interval is independent of the number in non-overlapping time interval (for example, the number of arrivals occurring between times 5 and 25 does not provide information about the number of arrivals occurring between times 30 and 50).

Now, let $P(n;t)$ be the probability of observing n arrivals in a time interval of length t. Then, for some $\lambda > 0$,

$$P_\lambda(n;t) = \frac{(\lambda t)^n}{n!}e^{-\lambda t},\ n = 0,1,2,\cdots$$

is the p.m.f. of the **Poisson distribution** for the discrete random variable n – the number of arrivals – for a given length of time interval t (see Figure 1a). In a queueing system, such arrivals are referred to as **Poisson arrivals**.

The time between successive arrivals is called the **inter-arrival time**. If the number of arrivals in a given time interval follows a Poisson distribution with parameter λt, the inter-arrival times follow an exponential distribution with p.d.f.

$$f_\lambda(t) = \lambda e^{-\lambda t},\ \text{for } t > 0,$$

and the probability $P(W \leq t)$ that a customer's waiting time W is smaller than the length of the time interval t is

$$P(W \leq t) = 1 - e^{-\lambda t} \quad \text{(see Figure 1b)}.$$

In general, if the arrival rate is **stationary**, if **bulk** arrivals (two or more simultaneous arrivals) cannot occur, and if past arrivals do not affect future arrivals, then inter-arrival times follow an exponential distribution with parameter λ, and the number of arrivals in any interval of length t is Poisson with parameter λt.

One of the most attractive features of using the exponential distribution to model inter-arrival times is that it is **memoryless** – if a random variable X follows an exponential distribution, then for all non-negative values of t and h,

$$P(X \geq t + h | X \geq t) = P(X \geq h). \tag{1}$$

No other density function satisfies (1) [2]. The memoryless property of the exponential distribution is important because it implies that the probability distribution of the time until the next arrival is independent of the time since the last arrival – imagine if that was the case when waiting for public transportation!

For instance, if we know that at least t time units have elapsed since the last arrival, then the distribution of time h until the next arrival is independent of t. If $h = 4$, say, then (1) yields

$$P(X > 9|X > 5) = P(X > 7|X > 3) = P(X > 4|X > 0) = e^{-4\lambda}.$$

Erlang Distribution The exponential distribution is not always an appropriate model of inter-arrival times (perhaps the process should not be memoryless, say). A common alternative is to use the **Erlang** distribution $\mathscr{E}(R,k)$, a continuous random variable with **rate** and **shape** parameter $R > 0$ and $k \in \mathbb{Z}^+$, respectively, whose p.d.f. is

$$f_{R,k}(t) = \frac{R(Rt)^{k-1}e^{-Rt}}{(k-1)!},\ t \geq 0.$$

When $k = 1$, the Erlang distribution reduces to an exponential distribution with parameter R. It can be shown that if $X \sim \mathscr{E}(k\lambda, k)$, then $X \sim X_1 + X_2 + \cdots + X_k$, where each X_i is an independent exponential random variable with parameter $k\lambda$.

When we model the inter-arrival process as an Erlang $\mathscr{E}(k\lambda, k)$ distribution, we are really saying that it is equivalent to customers going through k **phases** (each of which is memoryless) before being served. For this reason, the shape parameter is often referred to as the number of phases of the Erlang distribution [12].

Input/Arrival Process The input process is usually called the **arrival process**. Arrivals are called **customers**. In the models that we discuss, we assume that arrivals cannot be simultaneous (this might be unrealistic when modeling a restaurant, say). If simultaneous arrivals can occur, we say that **bulk arrivals are allowed**.

Usually, we assume that the arrival process is **unaffected by the number of customers** in the system. In the context of a bank, this would imply that whether there are 500 or 5 people at the bank, the process governing arrivals remains unchanged.

There are two common situations in which the arrival process may depend on the number of customers present. The first occurs when arrivals are drawn from a small population – the so-called **finite source models** – if all members of the populations are already in the system, there cannot be another arrival!

Another such situation arises when the rate at which customers arrive at the facility decreases when the facility becomes too crowded. For example, when customers see that a restaurant's parking lot is full, they might very well decide to go to another restaurant or forego eating out altogether. If a customer arrives but fails to enter the system, we say that the customer has **balked**.

Output/Service Process To describe the output process (often called the **service process**) of a queuing system, we usually specify a probability distribution – the **service time distribution** – which governs the customers' service time.

In most cases, we assume that the service time distribution is independent of the number of customers present in the system. This implies, for example, that the server does not work faster when more customers are present.

We can distinguish two types of servers: **in parallel** and **in series**. Servers are in parallel if they all provide the same type of service and a customer only needs to pass through one of them to complete their service. For example, the tellers in a bank are usually arranged in parallel; typically, customers only need to be serviced by one teller, and any teller can perform the desired service. Servers are in series if a customer must pass through several servers before their service is complete. An assembly line is an example of such a queuing system.

Input and output processes occur in a variety of situations:

- **situation:** purchasing Blue Jays tickets at the Rogers Centre
  input: baseball fans arrive at the ticket office
  output: tellers serve the baseball fans;
- **situation:** pizza parlour
  input: requests for pizza delivery are received; *output*: pizza parlour prepares and bakes pizzas, and sends them to be delivered;
- **situation:** government service centre
  input: citizen/residents enter the service centre
  output: receptionist assigns them to a specific queue based on their needs
    input: citizen/residents enter a specific queue based on their needs
    output: public servant addresses their needs;
- **situation:** hospital blood bank
  input: pints of blood arrive
  output: patients use up pints of blood;
- **situation:** garage
  input: cars break-down and are sent to the garage for repairs
  output: cars are repaired by mechanics and sent back on the streets.

The computations are fairly easy to execute, as the following examples demonstrate.

- On average, 4.6 customers enter a coffee shop each hour. If the arrivals follow a Poisson process, the probability that at most two customers will enter in a 30 minute period is

$$\begin{aligned} P_{\lambda=4.6}(n \leq 2; t = 0.5) &= P_{4.6}(0, 0.5) + P_{4.6}(1, 0.5) + P_{4.6}(2, 0.5) \\ &= e^{-4.6 \cdot 0.5} \left[\frac{(4.6 \cdot 0.5)^0}{0!} + \frac{(4.6 \cdot 0.5)^1}{1!} + \frac{(4.6 \cdot 0.5)^2}{2!} \right] \\ &\approx 0.5960; \end{aligned}$$

  the corresponding Poisson distribution is shown in Figure 1a.

- In a fast food restaurant, a cashier serves on average 9 customers in a one-hour time period. If the service time follows an exponential distribution, 89.5% and 2.4% of customers will be served in 15 minutes or less and after 25 minutes, respectively. Indeed,

$$P(W \leq 15/60) = 1 - e^{-9 \cdot 15/60} \approx 0.8946 \quad \text{and} \quad P(W > 25/60) = e^{-9 \cdot 25/60} \approx 0.0235.$$

Queue Discipline To describe a queuing system completely, we must also describe the **queue discipline** and the manner in which customers **join lines**. The queue discipline describes the method used to determine the order in which customers are served:

- the most common queue discipline is the **first come, first served** (FCFS) discipline, in which customers are served in the order of their arrival, as one would expect to see in an Ottawa coffee shop;
- under the **last come, first served** (LCFS) discipline, the most recent arrivals are the first to enter service; for example, if we consider exiting from an elevator to be the service, then a crowded elevator illustrates such a discipline;
- sometimes the order in which customers arrive has no effect on the order in which they are served; this would be the case if the next customer to enter service is randomly chosen from those customers waiting for service, a situation referred to as **service in random order** (SIRO) discipline; when callers to an inter-city bus company are put on hold, the luck of the draw often determines which caller will next be serviced by an operator,
- finally, **priority** discipline classifies each arrival into one of several categories, each of which is assigned a priority level (a **triage** process); within each priority level, customers enter the queue on a FCFS basis; such a discipline is often used in emergency rooms to determine the order in which customers receive treatment, and in copying and computer time-sharing facilities, where priority is usually given to jobs with shorter processing times.

Method Used by Arrivals to Join Queue Another important factor for the behaviour of the queuing system is the **method** used by customers to determine which line to join. For example, in some banks, customers must join a single line, but in other banks, customers may choose the line they want to join.

When there are several lines, customers often join the shortest line. Unfortunately, in many situations (such as at a supermarket), it is difficult to define the shortest line. If there are several lines at a queuing facility, it is important to know whether or not customers are allowed to **switch**, or jockey, between lines. In most queuing systems with multiple lines, jockeying is permitted, but jockeying at a custom inspection booth would not be recommended, for instance.

1.5.2 Queueing Theory Framework

There is a standard notation that is used to describe large families of queueing systems.

Kendall-Lee Notation Queuing systems can be described by six characteristics [8]:

$$x_1/x_2/x_3/x_4/x_5/x_6.$$

The first characteristic x_1 specifies the nature of the **arrival process**. The following standard abbreviations are used:

M = inter-arrival times are independent, identically distributed (iid) exponentials
D = inter-arrival times are iid and deterministic
E_k = inter-arrival times are iid Erlangs with shape parameter k
G = inter-arrival times are iid and governed by some general distribution.

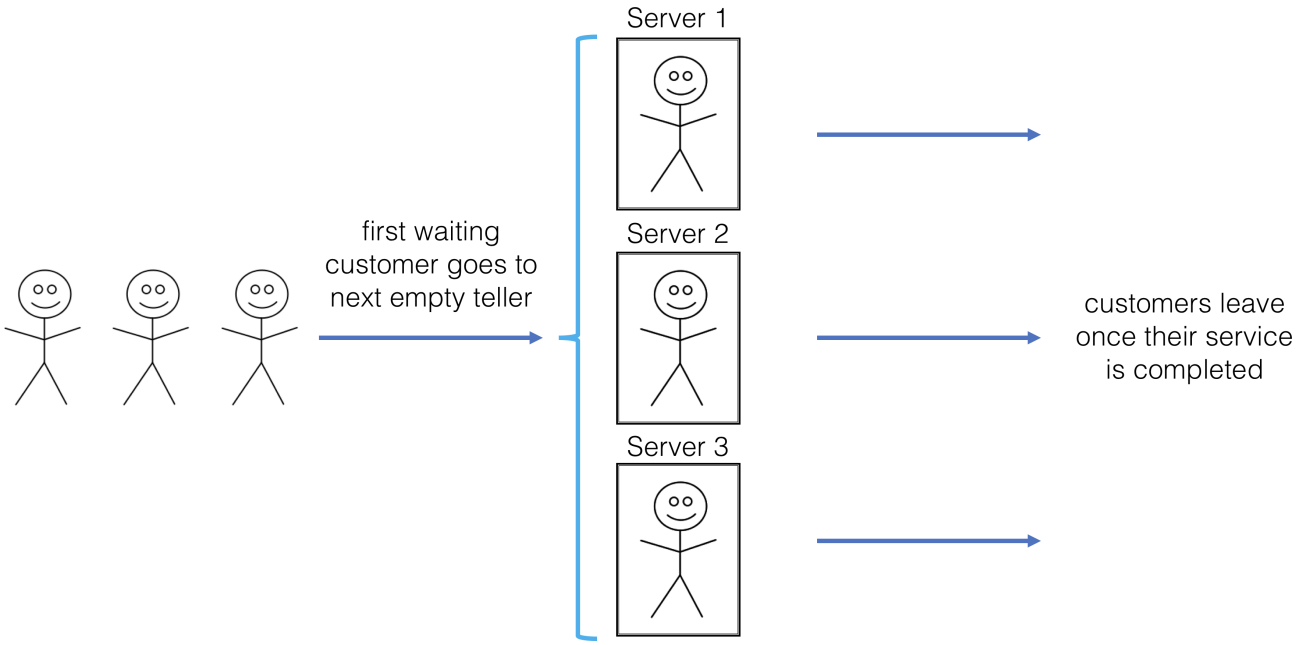


Figure 2: Single line at bank with three tellers – $M/M/3/\text{FCFS}/20/\infty$.

The second characteristic x_2 specifies the nature of the **service times**:

M = service times are iid and exponential
D = service times are iid and deterministic.
E_k = service times are iid Erlang with shape parameter k
G = service times are iid and follow some general distribution.

The third characteristic x_3 is the **number of parallel servers**. The fourth characteristic x_4 describes the **queue discipline**:

FCFS = first come, first served
LCFS = last come, first served
SIRO = service in random order
GD = general queue discipline.

The fifth characteristic x_5 specifies the **maximum allowable number of customers in the system** (including customers who are waiting and customers who are in service). The sixth characteristic x_6 gives the **size of the population** from which customers are drawn. Unless the number of potential customers is of the same order of magnitude as the number of servers, the population size is considered to be infinite.

In many important models $x_4/x_5/x_6$ is $\text{GD}/\infty/\infty$; when this is the case, these characteristics are often omitted. For example, $M/M/3/\text{FCFS}/20/\infty$ could represent a bank with 3 tellers, exponential arrival times, exponential service times, a "first come, first served" queue discipline, a total capacity of 20 customers, and an infinite population pool from which to draw. The situation is partly illustrated in Figure 2.

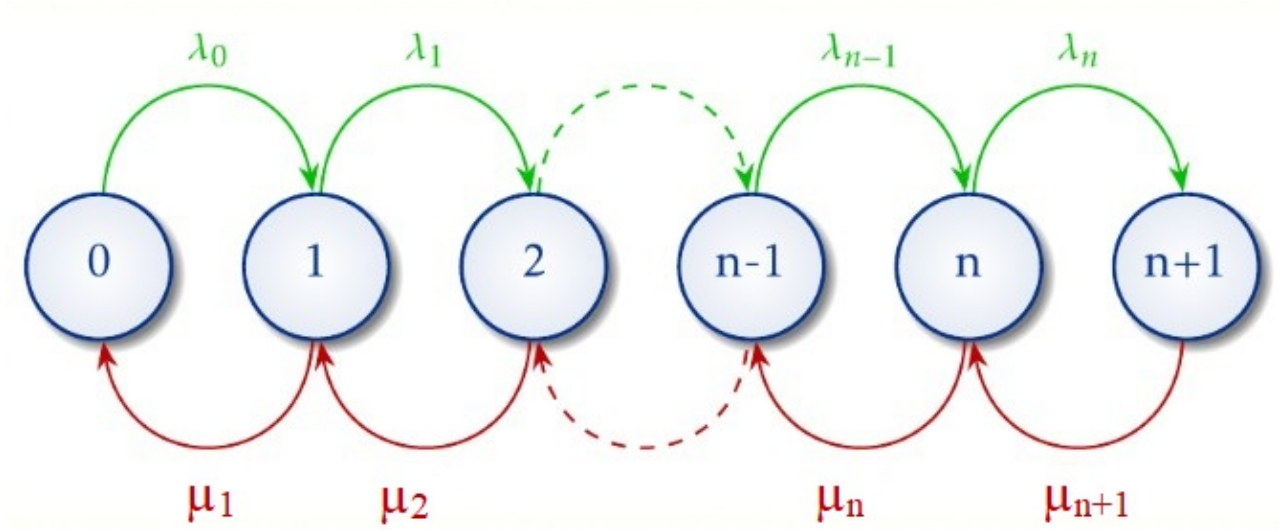


Figure 3: Birth-death Process; queueing states indexed by integers; birth rates and death rates indicated by λ_n and μ_m, respectively.

Birth-Death Processes The **state of a queueing system** at time t is defined to be the number of customers in the queuing system, either waiting in line or in service, at time t. At $t = 0$, the state of the system is the initial number of customers in the system. This state is noteworthy because it clearly affects the state at future t.

Knowing this, we define $P_{ij}(t)$ as the probability that the state at time t is j, given that the state at $t = 0$ was i. For large t, $P_{ij}(t)$ becomes independent of i and approaches a limit π_j. This limit is known as the **steady-state** of state j.

It is generally incredibly difficult to determine the steps of arrivals and services that lead to a steady-state π_j. Likewise, starting from a small t, it is difficult to determine exactly when a system will reach its steady state π_j, if such a state even exists.

For simplicity's sake, when a queuing system is studied, we begin by assuming that the steady-state has already been reached.

A **birth-death process** is a Markov process in which states are indexed by non-negative integers, and transitions are only permitted between "neighbouring" states. After a "birth", the state increases from n to $n+1$; after a "death", the state decreases from m to $m-1$. Typically, we denote the set of birth rates and death rates by λ_n and μ_m, respectively (see Figure 3). **Pure birth** processes are those for which $\mu_m = 0$ for all m; **pure death** processes those for which $\lambda_n = 0$ for all n. The **steady-state solution** of a birth-death process, i.e. the probability p_n of being in state n *can* actually be computed:

$$p_n = p_0 \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}, \quad \text{for } n = 1, 2, \cdots, \tag{2}$$

where p_0 is the probability of being in state 0. It can further be shown [9] that:

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \prod_{j=0}^{n-1} \frac{\lambda_j}{\mu_{j+1}}}.$$

Little's Queuing Formula It is often the case that a client is interested in the amount of time that a typical customer spends in the queuing system. Let W be the **expected waiting time** spent in the queuing system, including time in line plus time in service, and W_q be the **expected time a customer spends waiting in line**. Both W and W_q are computed under the assumption that the steady state has been reached. By using a powerful result known as **Little's queuing formula**, W and W_q are easily related to the number of customers in the queue and those waiting in line.

For any queuing system (or any subset of a queuing system), consider the following quantities:

- λ = average number of arrivals entering the system per unit time;
- L = average number of customers present in the queuing system;
- L_q = average number of customers waiting in line;
- L_s = average number of customers in service;
- W = average time a customer spends in the system;
- W_q = average time a customer spends in line, and
- W_s = average time a customer spends in service.

But customers in the system can only either be found in the queue or in service, so that $L = L_q + L_s$ as well $W = W_q + W_s$. In these definitions, all averages are steady-state averages. For most queuing systems in which a steady-state exists, Little's queuing formula can be summarized as

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

For instance, if 46 customers enter a restaurant each hour it is opened (on average), and if they spend 10 minutes waiting to be served (on average), then we should expect $46 \cdot 1/6 \approx 7.7$ customers in the queue at all time (on average).

1.5.3 The $M/M/1$ Queuing System

An $M/M/1/\text{GD}/\infty/\infty$ system has exponential inter-arrival times, exponential service times, and a single server. It is the simplest non-trivial queueing system to analyse as it can be modeled as a birth-death process with

$$\begin{aligned} \lambda_j &= \lambda, \ j = 0, 1, 2, \ldots \\ \mu_0 &= 0 \\ \mu_j &= \mu, \ j = 1, 2, 3, \ldots \end{aligned}$$

Substituting these rates in (2) yields

$$\pi_j = \frac{\lambda^j \pi_0}{\mu^j} = \rho^j \pi_0,$$

where $\rho = \lambda/\mu$ is the **traffic intensity** of the system. Since the system has to be in exactly one of the states at any given moment, the sum of all probabilities is 1, or

$$\pi_0 + \pi_1 + \pi_2 + \cdots = \pi_0(1 + \rho + \rho^2 + \cdots) = 1.$$

If $0 \leq \rho < 1$ the infinite series converges to $\frac{1}{1-\rho}$ from which we derive

$$\pi_0 \cdot \frac{1}{1-\rho} = 1 \implies \pi_0 = 1-\rho \implies \pi_j = \rho^j \pi_0 = \rho^j(1-\rho)$$

as the **steady-state probability of state** j. If $\rho \geq 1$, the infinite series diverges and no steady-state exists. Intuitively, this happens when $\lambda \geq \mu$ – if the arrival rate is greater than the service rate, then the state of the system grows without bounds and the queue is never cleared.

From this point on, we assume $\rho < 1$ to guarantee that the steady-state probabilities π_j exist, from which we can determine several quantities of interest. Assuming that the steady state has been reached, it can be shown that L, L_s, and L_q are given respectively by:

$$\begin{aligned} L &= \frac{\lambda}{\mu-\lambda} \\ L_s &= \rho \\ L_q &= \frac{\rho^2}{1-\rho}. \end{aligned}$$

Using Little's queuing formula, we can also solve for W, W_s, and W_q by dividing each of the corresponding L values by λ. Notice that, as expected, both W and W_q when $\rho \to 1$. On the other hand, $W_q \to 0$ and $W \to \frac{1}{\mu}$ (the **mean service time**) as $\rho \to 0$.

(The following example is based on [1]) An average of 10 cars arrive at a single-server drive-in teller every hour. If the average customer is served in 4 minutes, service time for each customer is 4 minutes, and both inter-arrival times and service times are exponential, then:

(a) What is the probability that the teller is idle?
(b) Excluding the car that is being served, what is the average number of cars waiting in line at the teller?
(c) What is the average amount of time a drive-in customer spends in the bank parking lot (including time in service)?
(d) On average, how many customers per hour will be served by the teller?

By assumption, we are dealing with an $M/M/1/\text{GD}/\infty/\infty$ queuing system for which $\lambda = 10$ cars/hr and $\mu = 15$ cars/hr, and as such $\rho = 10/15 = 2/3$.

(a) The teller is idle one third of the time on average because $\pi_0 = 1-\rho = 1/3$.
(b) There are $L_q = \rho^2/(1-\rho) = 4/3$ cars waiting in line for the teller.
(c) We know that $L = \lambda/(\mu-\lambda) = 10/(15-10) = 2$, and so $W = L/\lambda = 0.2$ hr $= 12$ min.
(d) If the teller were always busy, it would serve an average of $\mu = 15$ customers per hour. From (a), we know that the teller is only busy two-thirds of the time, thus during each hour, the teller serves an average of $15 \cdot 2/3 = 10$ customers. This is reasonable since, in a steady-state, 10 customers are arriving each hour and 10 customers must leave the system every hour.

(This next example is based on [15]) Suppose that all car owners fill up when their tanks are exactly half full. At the present time, an average of 7.5 customers arrive every hour at a single-pump gas station. It takes an average of 4 minutes to fuel a car. Assume that inter-arrival times and service times are both exponential.

(a) What are the values of L and W in this scenario?
(b) Suppose that a gas shortage occurs and panic buying takes place. To model this phenomenon, assume that all car owners now purchase gas when their tanks are exactly three-quarters full. Since each car owner is now putting less gas into the tank during each visit to the station, we assume that the average service time has been reduced to 10/3 minutes. How has panic buying affected the values of L and W?

By assumption, we have again an $M/M/1/\text{GD}/\infty/\infty$ queuing system, this time with $\lambda = 7.5$ cars/hr and $\mu = 60/4 = 15$ cars/hr. Thus, $\rho = 7.5/15 = 1/2$.

(a) By definition, $L = \lambda/(\mu-\lambda) = 7.5/(15-7.5) = 1$ and $W = 1/7.5 \approx 0.13$ hr $= 7.8$ min. Hence, in this situation, everything is under control, and long lines appear to be unlikely.
(b) Under the panic buying scenario, $\lambda = 2(7.5) = 15$ cars/hr as each car owner now fills up twice as often, and $\mu = 60 \cdot 3/10 = 18$ cars/hr. Then, $\rho = \lambda/\mu = 5/6$. In that scenario,

$$L = \frac{\rho}{1-\rho} = 5 \text{ cars} \quad \text{and} \quad W = \frac{L}{\lambda} = \frac{5}{15} = \frac{1}{3} \text{ hr } = 20 \text{ min.}$$

Thus, panic buying has more than doubled the wait time in line.

In a $M/M/1$ queueing system, we have

$$L = \frac{\rho}{1-\rho} = -1 + \frac{1}{1-\rho},$$

and it is easy to see that $L \to \infty$ as $\rho \to 1$. The 5-fold increase in L when ρ jumps from 1/2 to 5/6 (with accompanying jumps in W) illustrate that fact.

Limited Capacity In the real world, queues never become infinite – they are limited due to requirements of space and/or time, or service operating policy. Such a queuing model falls under the purview of **finite queues**.

Finite queue models restrict the number of customers allowed in the service system. Let N represent the maximum allowable number of customers in the system. If the system is at **capacity**, the arrival of a $(N+1)^{\text{th}}$ customer results in a failure to enter the queue – the customer is assumed to depart without seeking service.

Finite queues can also be modeled as a birth-death process, but with a slight modification in its parameters: with these parameters:

$$\begin{aligned} \lambda_j &= \lambda, \ j = 0, 1, 2, \ldots, N-1 \\ \lambda_N &= 0, \ \mu_0 = 0 \\ \mu_j &= \mu, \ j = 1, 2, 3, \ldots, N \end{aligned}$$

The restriction $\lambda_N = 0$ is what sets this model apart from the $M/M/1/\infty$. It makes it impossible to reach a state greater than N. Because of this restriction, a steady-state always exist because even if $\lambda \geq \mu$, there can never be more than N customers in the system.

Mathematically, this has the effect of replacing the infinite series linking the π_j's by a finite geometric series, which always converges:

$$\pi_0 + \pi_1 + + \cdots + \pi_N = \pi_0(1 + \rho + \cdots + \rho^N) = 1,$$

from which we can derive

$$\pi_0 \cdot \frac{1-\rho^{N+1}}{1-\rho} = 1 \implies \pi_0 = \frac{1-\rho}{1-\rho^{N+1}} \implies \pi_j = \begin{cases} \rho^j \frac{1-\rho}{1-\rho^{N+1}} & \text{for } j = 0, \ldots, N \\ 0 & \text{for } j > N \end{cases}$$

Since $L = \sum_{j=0}^{N} j \cdot \pi_j$,

$$L = \frac{\rho[1 + N\rho^{N+1} - (N+1)\rho^N]}{(1-\rho)(1-\rho^{N+1})}$$

when $\lambda \neq \mu$.

As in the $M/M/1/\infty$ queue, $L_s = 1 - \pi_0$, and $L_q = L - L_s$. It is somewhat trickier to compute W and W_q because, in a finite capacity model, only $\lambda - \lambda\pi_N = \lambda(1-\pi_N)$ arrivals per unit time actually enter the system on average (λ arrive, but $\lambda\pi_N$ find the system full). With this fact,

$$W = \frac{L}{\lambda(1-\pi_N)} \quad \text{and} \quad W_q = \frac{L_q}{\lambda(1-\pi_N)}.$$

What does that look like in practice? Consider a one-man barber shop with a total of 10 seats. Assume, as has always been the case so far but need not be, that inter-arrival times are exponentially distributed with an average of 20 prospective customers arriving each hour at the shop. Those customers who find the shop full do not enter – perhaps they do not like standing? The barber takes an average of 12 minutes to cut each customer's hair; assume that haircut times are also exponentially distributed.

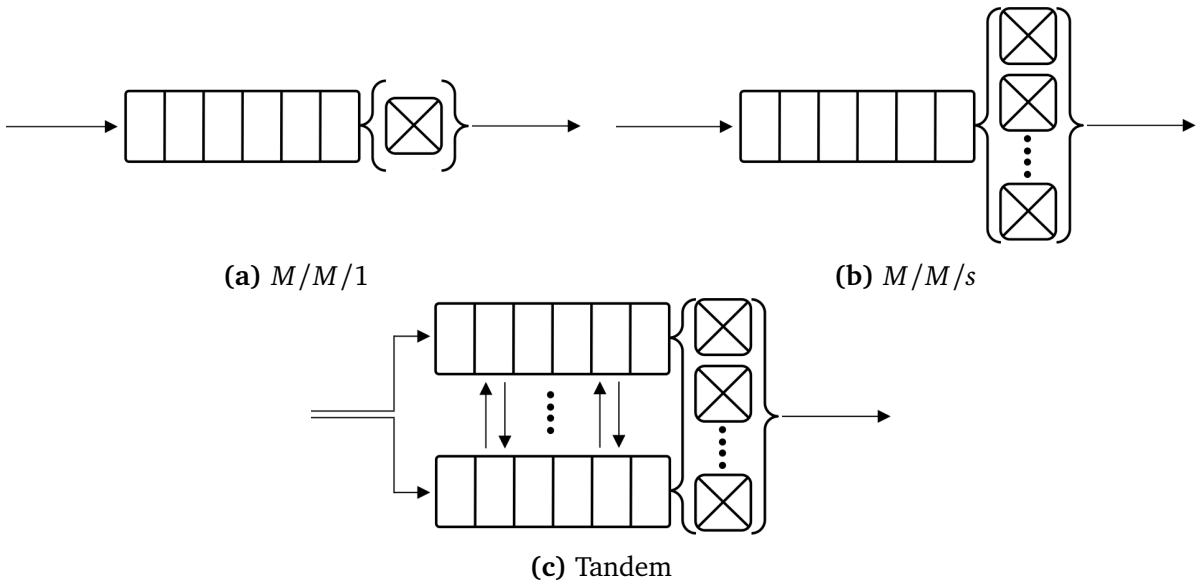
(a) On average, how many haircuts per hour will the barber complete?
(b) On average, how much time will be spent in the shop by a customer who enters?

There is not much to say. Let's dive in!

(a) A fraction π_{10} of all arrivals will find the shop is full. Thus, an average of $\lambda(1-\pi_{10})$ will actually enter the shop each hour. All entering customers receive a haircut, so the barber will give an average of $\lambda(1-\pi_{10})$ haircuts per hour. In this scenario, $N = 10$, $\lambda = 20$ customers/hr, and $\mu = 60/12 = 5$ customers/hr. Thus $\rho = 20/5 = 4$ and we have

$$\pi_0 = \frac{1-\rho}{1-\rho^{N+1}} = \frac{1-4}{1-4^{11}} \quad \text{and} \quad \pi_{10} = 4^{10}\pi_0 = \frac{3}{4}.$$

Thus, an average of $20(1-3/4) = 5$ customers per hour will receive haircuts. This means that an average of $20-5 = 15$ prospective customers per hour will not enter the shop.



(a) $M/M/1$ **(b)** $M/M/s$

(c) Tandem

Figure 4: Schematics of various queueing systems; customers arrive from the left, enter the queue and progress through it until they are served, at which point the exit the queue.

(b) To determine W, we must first compute

$$L = \frac{4[1+(10)4^{11}-(11)4^{10}]}{(1-4)(1-4^{11})} = 9.67.$$

Using the formulas described above, we obtain

$$W = \frac{L}{\lambda(1-\pi_{10})} = \frac{9.67}{5} = 1.93 \text{ hr}.$$

This barber shop is crowded – the barber would be well-advised to hire at least one more barber!

But what *would* be the effect of hiring a second barber? In order to answer this question, let us study $M/M/s$ queueing systems.

1.5.4 The $M/M/c$ Queuing System

An $M/M/c/\text{GD}/\infty$ queueing system also has exponential inter-arrival and service times, with rates λ and μ, respectively. What sets this system apart is that there are now c servers willing to serve from a single line of customers, perhaps like one would find in a bank (see Figure 4).

If $j \leq c$ customers are present in the system, then every customer is being served and there is no wait time; if $j > c$ customers are in the system, then c customers are being served and the remaining $j-c$ customers are waiting in the line.

To model this as a birth-death process, we have to observe that the death rate is dependent on how many servers are actually being used.

ρ	$c=2$	$c=3$	$c=4$	$c=5$	$c=6$	$c=7$
.10	.02	.00	.00	.00	.00	.00
.20	.07	.02	.00	.00	.00	.00
.30	.14	.07	.04	.02	.01	.00
.40	.23	.14	.09	.06	.04	.03
.50	.33	.24	.17	.13	.10	.08
.55	.39	.29	.23	.18	.14	.11
.60	.45	.35	.29	.24	.20	.17
.65	.51	.42	.35	.30	.26	.21
.70	.57	.51	.43	.38	.34	.30
.75	.64	.57	.51	.46	.42	.39
.80	.71	.65	.60	.55	.52	.49
.85	.78	.73	.69	.65	.62	.60
.90	.85	.83	.79	.76	.74	.72
.95	.92	.91	.89	.88	.87	.85

Table 1: Probabilities $P(n \geq c)$ that all servers are busy in an $M/M/c$ system for $c = 2, \ldots, 7$ and values of ρ between 0.1 and 0.95. [1, p.1088].

If each server completes service at a rate of μ (which may not be the case in practice as there might be variations in servers... at least for human servers), then the **actual death rate** is $\mu\times$ the number of customers actually being served. The parameters for this process are

$$\begin{aligned}\lambda_n &= \lambda,\ n = 0, 1, 2, \ldots \\ \mu_n &= n\mu,\ n = 0, 1, 2, \ldots, c \\ \mu_n &= c\mu,\ n = c+1, c+2, \ldots\end{aligned}$$

The traffic intensity for the $M/M/c$ system is $\rho = \lambda/(c\mu)$ and the steady-state solution is

$$\pi_n = \begin{cases} \frac{(c\rho)^n}{n!}\pi_0 & , 1 \leq n \leq c \\ \frac{c^c\rho^n}{c!}\pi_0 & , n \geq c \end{cases} \qquad \text{where } \pi_0 = \left[1 + \frac{(c\rho)^c}{c!(1-\rho)} + \sum_{n=1}^{c-1}\frac{c\rho^n}{n!}\right]^{-1}.$$

Note that, as was the case in a $M/M/1$ system, if $\rho \geq 1$, there can be no steady state – in other words, if the arrival rate is at least as large as the maximum possible service rate $\lambda \geq c\mu$, then the system "blows up".

From the client's point of view, there might be a desire to ensure that customers do not wait in line an inordinate amount of time, but there might also be a desire to minimise the amount of time for which at least one of the server is idle. In a $M/M/c$ queueing system, this steady-state probability is given by

$$P(n \geq c) = \frac{(c\rho)^c}{c!(1-\rho)}\pi_0.$$

Table 1 shows $P(n \geq c)$ for a variety of situations depending on s and ρ. Cumbersome calculations, using $W_s = \frac{1}{\mu}$, yield

$$L_q = P(n \geq c)\frac{\rho}{1-\rho}, \quad W_q = \frac{L_q}{\lambda}, \quad W = \frac{1}{\mu} + W_q, \quad \text{and} \quad L = \frac{\lambda}{\mu} + L_q.$$

Consider, for instance, a bank with two tellers. An average of 80 customers arrive at the bank each hour and wait in a single line for an idle teller. For this specific bank, the average service is 1.2 minutes. Assume that inter-arrival times and service times are exponential. Determine:

(a) The expected number of customers in the bank.
(b) The expected length of time a customer spends in the bank.
(c) The fraction of time that a particular teller is idle.

We are dealing with an $M/M/2$ system with $\lambda = 80$ customers/hr and $\mu = 50$ customers/hr. Thus, $\rho = \frac{80}{2 \cdot 50} = 0.80 < 1$ and the steady-state exists.

(a) From the above table, $P(n \geq 2) = 0.71$, from which we compute

$$L_q = P(n \geq 2) \cdot \frac{.8}{1-.8} = 2.84 \text{ customers}$$
$$L = \frac{80}{50} + L_q = 4.44 \text{ customers.}$$

(b) We know that $W = \frac{L}{\lambda} = \frac{4.44}{80} = 0.055 \text{ hr } = 3.3 \text{ min.}$
(c) To determine the fraction of time that a particular server is idle, note that tellers are idle during all moments when $n = 0$, and half the time (by symmetry) when $n = 1$. The probability that a server is idle is thus given by $\pi_0 + 0.5\pi_1$. But

$$\pi_0 = \left[1 + \frac{(2 \cdot .8)^2}{2!\,(1-.8)} + \sum_{n=1}^{2-1} \frac{2 \cdot .8^n}{n!} \right]^{-1} = \frac{1}{9} \quad \text{and} \quad \pi_1 = \frac{1.6}{1!}\pi_0 = 0.176$$

and so the probability that particular teller is idle is $0.11 + 0.5(0.176) = 0.198$.

IMPORTANT NOTE: other queueing models are not understood to the same extent, and their given performance measurements may only be approximate and highly-dependent on the specifics of the problem at hand. For this reason, $M/M/c$ models are sometimes used even when their use is not supported by the data (the situation is not unlike the wide use of the normal distribution). In various applications, the empirical distributions of arrivals and service times are nearly Poisson and exponential, respectively, so that the assumption is not entirely missing the mark, but numerical simulations should not be eschewed when departures from the $M/M/c$ model are too pronounced.

1.5.5 Case Study: Wait Time Impact Model at Canadian Airports

By providing efficient and effective **pre-board screening** (PBS), the *Canadian Air Transport Security Authority* (CATSA) ensures the safety of all passengers and crew aboard flights departing Canadian airports while maintaining an appropriate balance between staffing and the wait time experienced by passengers.

The number of active screening stations and the number of passengers affect the wait times, and, as a result, budget cuts have a strong impact on the system, both in Canada and in the United States.

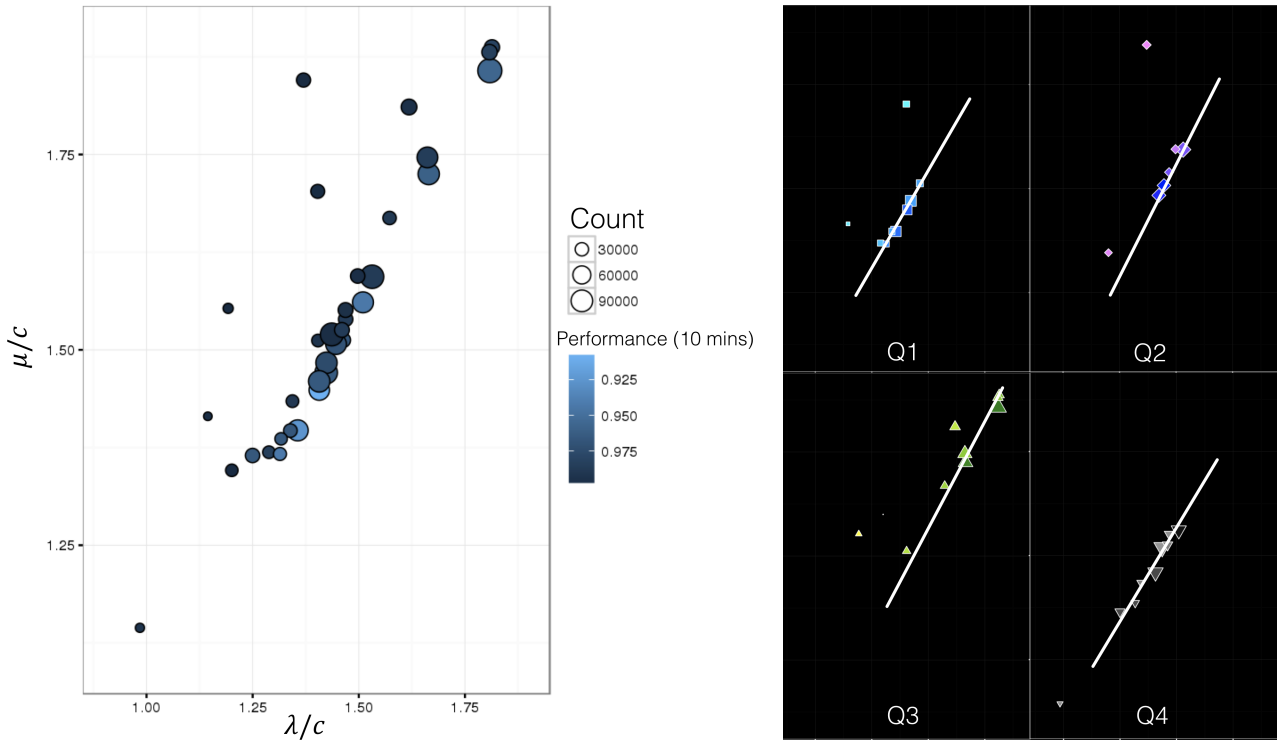


Figure 5: Visualisation of a specific checkpoint's queueing parameters – λ, μ, $\bar{c}$, passenger count, and performance (percentage of travellers waiting less than 15 minutes to be screened); the relationship between $\lambda/\bar{c}$ and $\mu/\bar{c}$ is practically linear (left), which is easier to see at the quarter level (right).

Numerous factors influence the wait time at pre-board screening checkpoints at Canadian airports: the schedule intensity of departing flights, the volume of passengers on these flights, the number of servers and processing rates at a given checkpoint, etc.

One of CATSA's goals is to ensure that the pre-board screening experience at Canadian airports is made as efficient as possible by minimizing the waiting time at checkpoints. With this in mind, the **Wait-Time Impact Model** (WTIM) was designed to achieve the following tasks:

1. provide estimates of the passenger arrival rates λ, the processing rates μ and the number of servers c at each checkpoints, using available field data;
2. calculate the Quality of Service (QoS) level (p_x, x) and determine what service level can be achieved at each checkpoint (i.e. the percentage p of passengers which will wait less than x minutes, for x fixed) for a given arrival rate λ, processing rate μ, number of servers c;
3. provide the average number of servers c^* required to achieve a prescribed QoS level (p_x, x), given an arrival profile λ^*;
4. provide quality of service (QoS) level curves $(p_x(x), x)$ (i.e. cumulative distribution curves) under various arrival rate and number of active servers for each checkpoint (where x is allowed to vary).

The queueing structure leads to some interesting insights (see Figure 5). The prediction's quality are seen in Figure 6. Details are available in the Project Summary, as well as in the Final Report (extract).

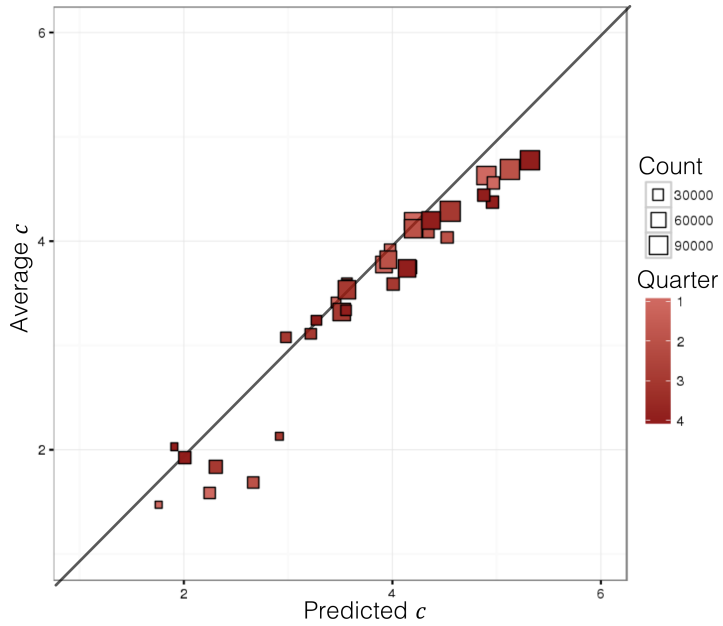


Figure 6: Predicted average number of server against actual number of server required to maintain prescribed performance, with passenger count, by quarter. The perfect prediction line is added for ease of comparison.

References

[1] Winston, W.L. [2004], Operations Research: Applications and Algorithms, 2nd ed., PWS-Kent Publishing, Boston.

[2] Ross, S.M. [2014], Introduction to Probability Models, 11th ed., Academic Press.

[3] **https://nptel.ac.in/courses/110106046/Module%209/Lecture%204.pdf**

[4] **https://nptel.ac.in/courses/110106046/Module%209/Lecture%205.pdf**

[5] **http://web.engr.illinois.edu/ dmnicol/ece541/slides/queueing.pdf**

[6] **https://www.wisdomjobs.com/e-university/quantitative-techniques-for-management-tutorial-297/poisson-and-exponential-distributions-9931.html**

[7] **https://www.percona.com/live/17/sites/default/files/the-essential-guide-to-queueing-theory.pdf**

[8] Kendall, D.G. [1953], "Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain", Ann.Math.Stat. 24 (3): 338.

[9] Kleinrock, L. [1975], Queueing Systems, vol. 1, Wiley.

[10] Gross, D., Shortle, J.F., Thompson, J.M., Harris, C.M. [2008], Fundamentals of Queueing Theory, 4th ed., Wiley.

[11] Burke, P.J. [1956], The Output of a Queueing System, Operations Reserach vol 4 (6): 699704.

[12] Newell, G.F. [1971], Applications of Queueing Theory, Chapman and Hall.

[13] Walrand, J. [1983], A probabilistic look at networks of quasi-reversible queues, IEEE Transactions on Information Theory, vol 29 (6): 825831.

[14] **https://en.wikipedia.org/wiki/Queueing_theory**

[15] Erickson, W. [1973], Management Science and the Gas Shortage, Interfaces 4:47âĂŞ51.