4. Queueing Theory

4.1 Introduction

- In this Lecture note we will study a class of models in which customers arrive in some random manner at a service facility. Upon arrival they are made to wait in queue until it is their turn to be served. Once served they are generally assumed to leave the system.

- For such models we will be interested in determining, among other things, such quantities as the average number of customers in the system (or in the queue) and the average time a customer spends in the system (or spends waiting in the queue).

4.2 Prelimnaries

In this section we will derive certain identities that are valid in the great majority of queueing models.

4.2.1 Cost Equations

Some fundamental quantities of interest for queueing models are

- L: the average number of customers in the system;
- L_Q: the average number of customers waiting in queue;
- W: the average amount of time a customer spends in the system;
- W_Q: the average amount of time a customer spends waiting in queue.

A large number of interesting and useful relationships between the preceding and other quantities of interest can be obtained by making use of the following idea: Imagine that entering customers are forced to pay money (according to some rule) to the system. We would then have the following basic cost identity:

average rate at which the system earns

$= \lambda_a \times$ **average amount an entering customer pays**

where λ_a is defined to be average arrival rate of entering customers. That is, if $N(t)$ denotes the number of customer arrivals by time t, then

$$\lambda_a = \lim_{t \to \infty} \frac{N(t)}{t}.$$

Heuristic Proof of the equation above: Let T be a fixed large number

- In two different ways, we will compute the average amount of money the system has earned by time T.

1. On one hand, this quantity approximately can be obtained by multiplying the average rate at which the system earns by the length of time T .

2. On the other hand, we can approximately compute it by multiplying the average amount paid by an entering customer by the average number of customers entering by time T (this latter factor is approximately $\lambda_a T$).

- Hence, both sides of the equation above when multiplied by T are approximately equal to the average amount earned by T . The result then follows by letting $T \to \infty$

By choosing appropriate cost rules, many useful formulas can be obtained as special cases of the above equation, and those formulas are valid for almost all queueing models regardless of the arrival process, the number of servers, or queue discipline.

- By supposing that each customer pays $\$1$ per unit time while in the system, it yields the so-called Little's formula,

$$L = \lambda_a W$$

- Similarly if we suppose that each customer pays $\$1$ per unit time while in queue, then it yields

$$L_Q = \lambda_a W_Q$$

- By supposing the cost rule that each customer pays $\$1$ per unit time while in service we obtain that the

$$\text{average number of customers in service} = \lambda_a \mathrm{E}[S]$$

where $\mathrm{E}[S]$ is defined as the average amount of time a customer spends in service.

4.2.2 Steady-State Probabilities

Let $X(t)$ denote the number of customers in the system at time t and define $P_n, n \geq 0$, by

$$P_n = \lim_{t \to \infty} P\{X(t) = n\}$$

where we assume the preceding limit exists. In other words, P_n is the limiting or long- run probability that there will be exactly n customers in the system.

- It is sometimes referred to as the *steady-state probability* of exactly n customers in the system.

- It also usually turns out that P_n equals the (long-run) proportion of time that the system contains exactly n customers.

- Two other sets of limiting probabilities are $\{a_n, n \geq 0\}$ and $\{d_n, n \geq 0\}$, where

a_n = proportion of customers that find n in the system when they arrive,

d_n = proportion of customers leaving behind n in the system when they depart

That is, P_n is the proportion of time during which there are n in the system; a_n is the proportion of arrivals that find n; and d_n is the proportion of departures that leave behind n. That these quantities need not always be equal is illustrated by the following example.

Example 4.1. Consider a queueing model in which all customers have service times equal to 1, and where the times between successive customers are always greater than 1 (for instance, the interarrival times could be uniformly distributed over (1, 2)). Hence, as every arrival finds the system empty and every departure leaves it empty, we have

$$a_0 = d_0 = 1$$

However,

$$P_0 \neq 1$$

as the system is not always empty of customers.

*** It was, however, no accident that a_n equaled d_n in the previous example. That arrivals and departures always see the same number of customers is always true as is shown in the next proposition.

Proposition 4.1. In any system in which customers arrive and depart one at a time

the rate at which arrivals find n = the rate at which departures leave n

and

$$a_n = d_n$$

Hence, on the average, arrivals and departures always see the same number of customers. However, as Example 4.1 illustrates, they do not, in general, see time averages. One important exception where they do is in the case of Poisson arrivals.

Proposition 4.2. Poisson arrivals always see time averages. In particular, for Poisson arrivals,

$$P_n = a_n$$

*** Consider an arbitrary Poisson arrival. If we knew that it arrived at time t, then the conditional distribution of what it sees upon arrival is the same as the unconditional distribution of the system state at time t (Since the Poisson process has independent increments). Hence, an arrival would just see the system according to the limiting probabilities.

*** The result that Poisson arrivals see time averages is called the **PASTA principle**.

Example 8.2. People arrive at a bus stop according to a Poisson process with rate λ. Buses arrive at the stop according to a Poisson process with rate μ, with each arriving bus picking up all the currently waiting people. Let W_Q be the average amount of time that a person waits at the stop for a bus. Because the waiting time of each person is equal to the time from when they arrive until the next bus, which is exponentially distributed with rate μ, we see that

$$W_Q = 1/\mu$$

Using $L_Q = \lambda_a W_Q$, now shows that L_Q, the average number of people waiting at the bus stop, averaged over all time, is

$$L_Q = \lambda/\mu$$

If we let X_i be the number of people picked up by the ith bus, then with T_i equal to the time between the $(i-1)$st and the ith bus arrival,

$$\mathrm{E}[X_i|T_i] = \lambda T_i$$

which follows because the number of people that arrive at the stop in any time interval is Poisson with a mean equal to λ times the length of the interval. Because T_i is exponential with rate μ, it follows upon taking expectations of both sides of the preceding that

$$\mathrm{E}[X_i] = \lambda \mathrm{E}[T_i] = \lambda/\mu$$

Thus, the average number of people picked up by a bus is equal to the time average number of people waiting for a bus, an illustration of the PASTA principle. That is, because buses arrive according to a Poisson process, it follows from PASTA that the average number of waiting people seen by arriving buses is the same as the average number of people waiting when we average over all time.

Exponential Models

4.3.1 A Single-Server Exponential Queueing System

- Suppose that customers arrive at a single-server service station in accordance with a Poisson process having rate λ. That is, the times between successive arrivals are independent exponential random variables having mean $1/\lambda$.

- Each customer, upon arrival, goes directly into service if the server is free and, if not, the customer joins the queue.

- When the server finishes serving a customer, the customer leaves the system, and the next customer in line, if there is any, enters service. The successive service times are assumed to be independent exponential random variables having mean $1/\mu$.

- The preceding is called the **M/M/1** queue. The two Ms refer to the fact that both the interarrival and the service distributions are exponential (and thus memoryless, or Markovian), and the 1 to the fact that there is a single server.

To analyze it, we shall begin by determining the limiting probabilities P_n , for $n = 0, 1, \ldots$. To do so, think along the following lines.

- Suppose that we have an infinite number of rooms numbered $0, 1, 2, \ldots$, and suppose that we instruct an individual to enter room n whenever there are n customers in the system.

- Now suppose that in the long run our individual is seen to have entered room 1 at the rate of ten times an hour. Then at what rate must he have left room 1? Clearly, at this same rate of ten times an hour. For the total number of times that he enters room 1 must be equal to (or one greater than) the total number of times he leaves room 1.

- This sort of argument thus yields the general principle that will enable us to determine the state probabilities. Namely, for each $n \geq 0$, *the rate at which the process enters state n equals the rate at which it leaves state n*.

- Hence, from our rate-equality principle above we get our equations,

$$\lambda P_0 = \mu P_1$$

$$(\lambda + \mu) P_n = \lambda P_{n-1} + \mu P_{n+1}, n \geq 1$$

- The equation above, which balance the rate at which the process enters each state with the rate at which it leaves that state are known as *balance equations*.
- In order to solve the equation above, we rewrite them to obtain

$$\begin{aligned} P_1 &= \frac{\lambda}{\mu} P_0, \\ P_{n+1} &= \frac{\lambda}{\mu} P_n + \left(P_n - \frac{\lambda}{\mu} P_{n-1} \right) \end{aligned}$$

Solving in terms of P_0 yields

$$P_{n+1} = \frac{\lambda}{\mu} P_n + \left(P_n - \frac{\lambda}{\mu} P_{n-1} \right) = \frac{\lambda}{\mu} P_n = \left(\frac{\lambda}{\mu} \right)^{n+1} P_0$$

- To determine P_0 we use the fact that the P_n must sum to 1, and thus

$$1 = \sum_{n=0}^{\infty} P_n = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n P_0 = \frac{P_0}{1 - \lambda/\mu}$$

or

$$P_0 = 1 - \frac{\lambda}{\mu},\ P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right), n \geq 1$$

- Notice that for the preceding equations to make sense, it is necessary for λ/μ to be less than 1. For otherwise $\sum_{n=0}^{\infty}(\lambda/\mu)^n$ would be infinite and all the P_n would be 0.

Remark

- In solving the balance equations for the M/M/1 queue, we obtained as an intermediate step the set of equations

$$\lambda P_n = \mu P_{n+1}, n \geq 0$$

  These equations could have been directly argued from the general queueing result (shown in Proposition 4.1) that the rate at which arrivals find n in the system-namely λP_n-is equal to the rate at which departures leave behind n-namely, μP_{n+1}.

- We can also prove that $P_n = (\lambda/\mu)^n(1 - \lambda/\mu)$ by using a queueing cost identity. Suppose that, for a fixed $n > 0$, whenever there are at least n customers in the system the nth oldest customer (with age measured from when the customer arrived) pays 1 per unit time.

  Letting X be the steady state number of customers in the system, because the system earns 1 per unit time whenever X is at least n, it follows that

$$\text{average rate at which the system earns} = P\{X \geq n\}$$

- Also, because a customer who finds fewer than $n-1$ in the system when it arrives will pay 0, while an arrival who finds at least $n-1$ in the system will pay 1 per unit time for an exponentially distributed time with rate μ,

$$\text{average amount a customer pays} = \frac{1}{\mu} P\{X \geq n-1\}$$

Therefore, the queueing cost identity yields

$$P\{X \geq n\} = (\lambda/\mu) P\{X \geq n-1\}, n > 0$$

Iterating this gives

$$\begin{aligned} P\{X \geq n\} &= (\lambda/\mu) P\{X \geq n-1\} = (\lambda/\mu)^2 P\{X \geq n-2\} \\ &= \cdots = (\lambda/\mu)^n P\{X \geq 0\} = (\lambda/\mu)^n. \end{aligned}$$

Therefore

$$P\{X = n\} = P\{X \geq n\} - P\{X \geq n+1\} = (\lambda/\mu)^n (1 - \lambda/\mu).$$

- The quantities W, W_Q, and L_Q now can be obtained with the help of equations shown before. That is, since $\lambda_a = \lambda$, we have from the value of L that

$$L = \sum_{n=0}^{\infty} nP_n = \sum_{n=1}^{\infty} n(\lambda/\mu)^n(1-\lambda/m)^n = \frac{\lambda}{\mu - \lambda}$$

$$W = \frac{L}{\lambda} = \frac{1}{\mu - \lambda}$$

$$W_Q = W - \mathrm{E}[S] = W - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu - \lambda)}$$

$$L_Q = \lambda W_Q = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

Example 4.3. Suppose that customers arrive at a Poisson rate of one per every 12 minutes, and that the service time is exponential at a rate of one service per 8 minutes. What are L and W?

Example 4.4. Suppose customers arrive to a two server system according to a Poisson process with rate λ, and suppose that each arrival is, independently, sent either to server 1 with probability α or to server 2 with probability $1-\alpha$. Further, suppose that no matter which server is used, a service time is exponential with rate μ. Letting $\lambda_1 = \lambda\alpha$ and $\lambda_2 = \lambda(1-\alpha)$, then because arrivals to server i follow a Poisson process with rate λ_i , it follows that the system as it relates to server $i, i = 1, 2$, is an M/M/1 system with arrival rate λ_i and service rate μ.

Hence, provided that $\lambda_i < \mu$, the average time a customer sent to server i spends in the system is $W_i = \frac{1}{\mu-\lambda_i}, i = 1, 2$. Because $\mu - \lambda_i$ the fraction of all arrivals that go to server 1 is α and the fraction that go to server 2 is $1 - \alpha$, this shows that the average time that a customer spends in the system, call it $W(\alpha)$, is

$$W(\alpha) = \alpha W_1 + (1-\alpha)W_2 = \frac{\alpha}{\mu - \lambda\alpha} + \frac{1-\alpha}{\mu - \lambda(1-\alpha)}$$

Suppose now that we want to find the value of α that minimizes $W(\alpha)$.

Remarks:

- We have used the fact that if one event occurs at an exponential rate λ, and another independent event at an exponential rate μ, then together they occur at an exponential rate $\lambda + \mu$.

- Given that an M/M/1 steady-state customer–that is, a customer who arrives after the system has been in operation a long time–spends a total of t time units in the system, let us determine the conditional distribution of N, the number of others that were present when that customer arrived. That is, letting W^* be the amount of time a customer spends in the system, we will find $P\{N = n|W^* = t\}$.

- Another argument as to why W^* is exponential with rate $\mu - \lambda$ is as shown in the appendix lecture note.

Example 4.5. For an M/M/1 queue in steady state, what is the probability that the next arrival finds n in the system?

A Single-Server Exponential Queueing System Having Finite Capacity

In the previous model, we assumed that there was no limit on the number of customers that could be in the system at the same time. However, in reality there is always a finite system capacity N, in the sense that there can be no more than N customers in the system at any time. By this, we mean that if an arriving customer finds that there are already N customers present, then he does not enter the system.

As before, we let $P_n, 0 \leq n \leq N$, denote the limiting probability that there are n customers in the system. The rate-equality principle yields the following set of balance equations:

State	Rate at which the process leaves = rate at which it enters
0	$\lambda P_0 = \mu P_1$
$1 \leq n \leq N-1$	$(\lambda + \mu) P_n = \lambda P_{n-1} + \mu P_{n+1}$
N	$\mu P_N = \lambda P_{N-1}$

- We could now either solve the balance equations exactly as we did for the infinite capacity model, or we could save a few lines by directly using the result that the rate at which departures leave behind $n-1$ is equal to the rate at which arrivals find $n-1$. Invoking this result yields

$$\mu P_n = \lambda P_{n-1}, n = 1, \ldots, N$$

giving

$$P_n = \frac{\lambda}{\mu} P_{n-1} = \left(\frac{\lambda}{\mu}\right)^2 P_{n-2} = \cdots = \left(\frac{\lambda}{\mu}\right)^n P_0, \ n = 1, \ldots, N$$

- By using the fact that $\sum_{n=0}^{N} P_n = 1$ we obtain

$$1 = P_0 \sum_{n=0}^{N} \left(\frac{\lambda}{\mu}\right)^n = P_0 \left[\frac{1-(\lambda/\mu)^{N+1}}{1-\lambda/\mu}\right]$$

or

$$P_0 = \frac{1-\lambda/\mu}{1-(\lambda/\mu)^{N+1}}$$

and hence

$$P_n = \frac{(\lambda/\mu)^n(1-\lambda/\mu)}{1-(\lambda/\mu)^{N+1}}$$

-

$$L = \sum_{n=0}^{N} nP_n = \frac{(1-\lambda/\mu)}{1-(\lambda/\mu)^{N+1}} \sum_{n=0}^{N} n \left(\frac{\lambda}{\mu}\right)^n$$

which after some algebra yields

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

In deriving W , the expected amount of time a customer spends in the system, we must be a little careful about what we mean by a customer.

- Specifically, are we including those "customers" who arrive to find the system full and thus do not spend any time in the system?

- Or, do we just want the expected time spent in the system by a customer who actually entered the system?

- In the first case, we have $\lambda_a = \lambda$; whereas in the second case, since the fraction of arrivals that actually enter the system is $1 - P_N$, it follows that $\lambda_a = \lambda(1 - P_N)$. Once it is clear what we mean by a customer, W can be obtained from

$$W = \frac{L}{\lambda_a}$$

Example 4.6. Suppose that it costs $c\mu$ dollars per hour to provide service at a rate μ. Suppose also that we incur a gross profit of A dollars for each customer served. If the system has a capacity N, what service rate μ maximizes our total profit?

An exponential queueing system in which the arrival rates and the departure rates depend on the number of customers in the system is known as a birth and death queueing model.

- Let λ_n denote the arrival rate and let μ_n denote the departure rate when there are n customers in the system.

- Loosely speaking, when there are n customers in the system then the time until the next arrival is exponential with rate λ_n and is independent of the time of the next departure, which is exponential with rate μ_n

- Equivalently, and more formally, whenever there are n customers in the system, the time until either the next arrival or the next departure occurs is an exponential random variable with rate $\lambda_n + \mu_n$ and, independent of how long it takes for this occurrence, it will be an arrival with probability $\frac{\lambda_n}{\lambda_n+\mu_n}$.

- **The M/M/1 Queueing System**
  Because the arrival rate is always λ, and the departure rate is μ when the system is nonempty, the M/M/1 is a birth and death model with

$$\lambda_n = \lambda, n \geq 0; \mu_n = \mu, \mu_n \geq \mu, n \geq 1$$

- **The M /M /1 Queueing System with Balking**
  Consider the M/M/1 system but now suppose that a customer that finds n others in the system upon its arrival will only join the system with probability α_n. (That is, with probability $1 - \alpha_n$ it balks at joining the system.) Then this system is a birth and death model with

$$\lambda_n = \lambda\alpha_n, n \geq 0; \mu_n = \mu, n \geq 1$$

  The M/M/1 with finite capacity N is the special case where

$$\alpha_n = 1, \text{if } n < N, \text{ else } \alpha_n = 0, n \geq N.$$

- **The M /M /k Queueing System**

  Consider a k server system in which customers arrive according to a Poisson process with rate λ.

  - An arriving customer immediately enters service if any of the k servers are free.
  - If all k servers are busy, then the arrival joins the queue.
  - When a server completes a service the customer served departs the system and if there are any customers in queue then the one who has been waiting longest enters service with that server.
  - All service times are exponential random variables with rate μ. Because customers are always arriving at rate λ, $\lambda_n = \lambda, n \geq 0$

  The M/M/k is a birth and death queueing model with arrival rates

$$\lambda_n = \lambda, n \geq 0$$

  and departure rates

$$\mu_n = n\mu, \text{if } n \leq k \text{ else } \mu_n = k\mu, n \geq k.$$

To analyze the general birth and death queueing model, let P_n denote the long-run proportion of time there are n in the system.

- Then, either as a consequence of the balance equations given by

State	Rate at which process leaves = rate at which process enters
$n = 0$	$\lambda_0 P_0 = \mu_1 P_1$
$n \geq 1$	$(\lambda_n + \mu_n) P_n = \lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1}$

- or by directly using the result that the rate at which arrivals find n in the system is equal to the rate at which departures leave behind n, we obtain

$$\lambda_n P_n = \mu_{n+1} P_{n+1}, \text{or } P_{n+1} = \frac{\lambda_n}{\mu_{n+1}} P_n, n \geq 0$$

- Thus

$$P_0 = P_0, P_1 = \frac{\lambda_0}{\mu_1} P_0, P_2 = \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0, P_3 = \frac{\lambda_2}{\mu_3} P_2 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} P_0$$

- In general

$$P_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} P_0, n \geq 1$$

- Using that $\sum_{n=0}^{\infty} P_n = 1$ show that

$$1 = P_0 \left[1 + \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} \right]$$

Hence

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}}, \text{and } P_n = \frac{\frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}}{1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}}, n \geq 1$$

- The necessary and sufficient conditions for the long-run probabilities to exist is that the denominator in the preceding is finite. That is, we need have that

$$\sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} < \infty$$

Example 4.7 For the M/M/k system

$$\frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n} = \begin{cases} \frac{(\lambda/\mu)^n}{n!} & \text{if } n \le k \\ \frac{\lambda^n}{\mu^n k! k^{n-k}} & \text{if } n > k \end{cases}$$

Example 4.8. Find the average amount of time a customer spends in the system for an M/M/2 system.

Example 4.9. (M/M/1 Queue with Impatient Customers). Consider a single-server queue where customers arrive according to a Poisson process with rate λ and where the service distribution is exponential with rate μ, but now suppose that each customer will only spend an exponential time with rate α in queue before quitting the system. Assume that the impatient times are independent of all else, and that a customer who enters service always remains until its service is completed

Remark. As illustrated in the previous example, often the easiest way of determining the proportion of all events that are of a certain type A is to determine the rates at which events of type A occur and the rate at which all events occur, and then use that

$$\text{proportion of events that are type A} = \frac{\text{rate at which type A events occur}}{\text{rate at which all events occur}}$$

For instance, if people arrive at rate λ and women arrive at rate λ_w , then the proportion of arrivals that are women is λ_w/λ.

- To determine W , the average time that a customer spends in the system, for the birth and death queueing system, we employ the fundamental queueing identity $L = \lambda_a W$. Because L is the average number of customers in the system,

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

- Also, because the arrival rate when there are n in the system is λ_n and the proportion of time in which there are n in the system is P_n, we see that the average arrival rate of customers is

$$\lambda_a = \sum_{n=0}^{\infty} \lambda_n P_n$$

Consequently

$$W = \frac{\sum_{n=0}^{\infty} nP_n}{\sum_{n=0}^{\infty} \lambda_n P_n}.$$

- Now consider an equal to the proportion of arrivals that find n in the system. Since arrivals are at rate λ_n whenever there are n in system it follows that the rate at which arrivals find n is $\lambda_n P_n$. Hence, in a large time T approximately $\lambda_n P_n T$ of the approximately $\lambda_a T$ arrivals will encounter n. Letting T go to infinity shows that the long-run proportion of arrivals finding n in the system is

$$a_n = \frac{\lambda_n P_n}{\lambda_a}$$

4.4. Network of Queues

Consider a two-server system in which customers arrive at a Poisson rate λ at server 1. After being served by server 1 they then join the queue in front of server 2. We suppose there is infinite waiting space at both servers. Each server serves one customer at a time with server i taking an exponential time with rate μ_i for a service, $i = 1, 2$. Such a system is called a *tandem or sequential system*

let us define the state by the pair (n, m)–meaning that there are n customers at server 1 and m at server 2. The balance equations are

State	Rate that the process leaves = rate that it enters
$0, 0$	$\lambda P_{0,0} = \mu_2 P_{0,1}$
$n, 0; n > 0$	$(\lambda + \mu_1) P_{n,0} = \mu_2 P_{n,1} + \lambda P_{n-1,0}$
$0, m; m > 0$	$(\lambda + \mu_2) P_{0,m} = \mu_2 P_{0,m+1} + \mu_1 P_{1,m-1}$
$n, m; nm > 0$	$(\lambda + \mu_1 + \mu_2) P_{n,m} = \mu_2 P_{n,m+1} + \mu_1 P_{n+1,m-1} + \lambda P_{n-1,m}$

- We first note that the situation at server 1 is just as in an M/M/1 model. It follows that what server 2 faces is also an M/M/1 queue. Hence the probability that there are n customers at server 1 is

$$P\{n \text{ at server } 1\} = \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right)$$

and, similary

$$P\{m \text{ at server } 2\} = \left(\frac{\lambda}{\mu_2}\right)^m \left(1 - \frac{\lambda}{\mu_2}\right)$$

- Now, if the numbers of customers at servers 1 and 2 were independent random variables, then it would follow that

$$P_{n,m} = \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right) \left(\frac{\lambda}{\mu_2}\right)^m \left(1 - \frac{\lambda}{\mu_2}\right)$$

- $P_{n,m}$, as given by the equation above, satisfy all of the balance equations

- From the preceding we see that L, the average number of customers in the system, is given by

$$\begin{aligned} L &= \sum_{n,m}(n+m)P_{n,m} = \sum_n n\left(\frac{\lambda}{\mu_1}\right)^n\left(1-\frac{\lambda}{\mu_1}\right) + \sum_m \left(\frac{\lambda}{\mu_2}\right)^m\left(1-\frac{\lambda}{\mu_2}\right) \\ &= \frac{\lambda}{\mu_1-\lambda} + \frac{\lambda}{\mu_2-\lambda} \end{aligned}$$

- and from this we see that the average time a customer spends in the system is

$$W = \frac{L}{\lambda} = \frac{1}{\mu_1-\lambda} + \frac{1}{\mu_2-\lambda}$$

Example 4.10. Consider a system of two servers where customers from outside the system arrive at server 1 at a Poisson rate 4 and at server 2 at a Poisson rate 5. The service rates of 1 and 2 are respectively 8 and 10. A customer upon completion of service at server 1 is equally likely to go to server 2 or to leave the system (i.e., $P_{11} = 0$, $P_{12} = \frac{1}{2}$); whereas a departure from server 2 will go 25 percent of the time to server 1 and will depart the system otherwise (i.e., $P_{21} = \frac{1}{4}, P_{22} = 0$). Determine the limiting probabilities, L, and W .

4.5. The System M/G/1

Preliminaries: Work and Another Cost Identity

For an arbitrary queueing system, let us define the work in the system at any time t to be the sum of the remaining service times of all customers in the system at time t. For instance, suppose there are three customers in the system–the one in service having been there for three of his required five units of service time, and both people in queue having service times of six units. Then the work at that time is $2 + 6 + 6 = 14$. Let V denote the (time) average work in the system.

- Now recall the fundamental cost equations, which states that the

  average rate at which the system earns $= \lambda_a \times$average amount a customer pays

  and consider the following cost rule: *Each customer pays at a rate of y/unit time when his remaining service time is y*, whether he is in queue or in service.

- Thus, the rate at which the system earns is just the work in the system; so the basic identity yields

$$V = \lambda_a \mathrm{E}[\text{amount paid by a customer}]$$

- Now, let S and W_Q^* denote respectively the service time and the time a given customer spends waiting in queue. Then, since the customer pays at a constant rate of S per unit time while he waits in queue and at a rate of $S - x$ after spending an amount of time x in service, we have

$$\mathrm{E}[\text{amount paid by a customer}] = \mathrm{E}\left[SW_Q^* + \int_0^S (S - x)dx\right]$$

and thus

$$V = \lambda_a \mathrm{E}[SW_Q^*] + \frac{\lambda_a \mathrm{E}[S^2]}{2}$$

- In addition, if a customer's service time is independent of his wait in queue (as is usually, but not always the case),5 then we have from the equation above that

$$V = \lambda_a \mathrm{E}[S] W_Q^* + \frac{\lambda_a \mathrm{E}[S^2]}{2}$$

Application of Work to M/G/1 The M/G/1 model assumes

(i) Poisson arrivals at rate λ;

(ii) a general service distribution; and

(iii) a single server.

- In addition, we will suppose that customers are served in the order of their arrival. Now, for an arbitrary customer in an M/G/1 system,

  customer's wait in queue = work in the system when he arrives

  Hence by taking expectations of both sides yields,

  $$W_Q = \text{average work as seen by an arrival}$$

- But, due to Poisson arrivals, the average work as seen by an arrival will equal V , the time average work in the system. Hence, for the model M/G/1,

  $$W_Q = V = \lambda \mathrm{E}[S] W_Q + \frac{\lambda \mathrm{E}[S^2]}{2}$$

  yield the so called *Pollaczek-Khintchine formula*.

  $$W_Q = \frac{\lambda \mathrm{E}[S^2]}{2(1 - \lambda \mathrm{E}[S])}$$

The quantities L, L_Q, and W can be obtained from the above equation as

$$L_Q = \lambda W_Q = \frac{\lambda^2 \mathrm{E}[S^2]}{2(1 - \lambda \mathrm{E}[S])},$$

$$W = W_Q + \mathrm{E}[S] = \frac{\lambda \mathrm{E}[S^2]}{2(1 - \lambda \mathrm{E}[S])} + \mathrm{E}[S],$$

$$L = \lambda W = \frac{\lambda^2 \mathrm{E}[S^2]}{2(1 - \lambda \mathrm{E}[S])} + \lambda \mathrm{E}[S]$$

Example 4.11. Suppose that customers arrive to a single server system in accordance with a Poisson process with rate λ, and that each customer is one of r types. Further, type i with probability $\alpha_i, \sum_{i=1}^{r} \alpha_i = 1$. Also, suppose that the amount of time it takes to serve a type i customer has distribution function F_i , with mean μ_i and variance σ_i.

(a) Find the average amount of time a type j customer spends in the system, $j = 1, \ldots, r$.

(b) Find the average number of type j customers in the system, $j = 1, \ldots, r$.

Busy Periods

The system alternates between idle periods (when there are no customers in the sys- tem, and so the server is idle) and busy periods (when there is at least one customer in the system, and so the server is busy).

- Let I and B represent, respectively, the length of an idle and of a busy period.

- Because I represents the time from when a customer departs and leaves the system empty until the next arrival, it follows, since arrivals are according to a Poisson process with rate λ, that I is exponential with rate λ and thus

$$\mathrm{E}[I] = \frac{1}{\lambda}$$

- The long-run proportion of time the system is empty is equal to the ratio of $\mathrm{E}[I]$ to $\mathrm{E}[I] + \mathrm{E}[B]$. That is

$$P_0 = \frac{\mathrm{E}[I]}{\mathrm{E}[I] + \mathrm{E}[B]}$$

- We note that

$$\text{average number of customers in service} = \lambda \mathrm{E}[S]$$

  where $\mathrm{E}[S]$ is defined as the average amount of time a customer spends in service. However, as the left-hand side of the preceding equals $1 - P_0$, we have

$$P_0 = 1 - \lambda \mathrm{E}[S], \text{and then } \mathrm{E}[B] = \frac{\mathrm{E}[S]}{1 - \lambda \mathrm{E}[S]}$$

- Another quantity of interest is C, the number of customers served in a busy period. The mean of C can be computed by noting that, on the average, for every E[C] arrivals exactly one will find the system empty (namely, the first customer in the busy period). Hence,

$$a_0 = \frac{1}{\mathrm{E}[C]}$$

  and, as $a_0 = P_0 = 1 - \lambda \mathrm{E}[S]$ because of Poisson arrivals, we see that

$$\mathrm{E}[C] = \frac{1}{1 - \lambda \mathrm{E}[S]}$$

Variations on the M/G/1

- The M/G/1 with Random-Size Batch Arrivals

  Suppose that, as in the M/G/1, arrivals occur in accordance with a Poisson process having rate λ. But now suppose that each arrival consists not of a single customer but of a random number of customers. As before there is a single server whose service times have distribution G.

- Priority Queues

  Priority queueing systems are ones in which customers are classified into types and then given service priority according to their type. Consider the situation where are two types of customers, which arrive according to independent Poisson processes with respective rates λ_1 and λ_2, and have service distributions G_1 and G_2. We suppose that type 1 customers are given service priority, in that service will never begin on a type 2 customer if a type 1 is waiting. However, if a type 2 is being served and a type 1 arrives, we assume that the service of the type 2 is continued until completion. That is, there is no preemption once service has begun.

- M/G/1 Optimization example.

  Consider a single-server system where customers arrive according to a Poisson process with rate λ, and where the service times are independent and have distribution function G. Let $\rho = \lambda \mathrm{E}[S]$, where S represents a service time random variable, and suppose that $\rho < 1$. Suppose that the server departs whenever a busy period ends and does not return until there are n customers waiting. At that time the server returns and continues serving until the system is once again empty. If the system facility incurs costs at a rate of c per unit time per customer in the system, as well as a cost K each time the server returns, what value of $n, n \geq 1$, minimizes the long-run average cost per unit time incurred by the facility, and what is this minimal cost?

- The M/G/1 Queue with Server Breakdown

  Consider a single server queue in which customers arrive according to a Poisson process with rate λ, and where the amount of service time required by each customer has distribution G. Suppose, however, that when working the server breaks down at an exponential rate α. That is, the probability a working server will be able to work for an additional time t without breaking down is e^{-at}. When the server breaks down, it immediately goes to the repair facility. The repair time is a random variable with distribution H. Suppose that the customer in service when a breakdown occurs has its service continue, when the server returns, from the point it was at when the breakdown occurred. (Therefore, the total amount of time a customer is actually receiving service from a working server has distribution G.)

4.6. The Model G/M/1

The model G/M/1 assumes that the times between successive arrivals have an arbitrary distribution G. The service times are exponentially distributed with rate μ and there is a single server.

- The immediate difficulty in analyzing this model stems from the fact that the number of customers in the system is not informative enough to serve as a state space.

- For in summarizing what has occurred up to the present we would need to know not only the number in the system, but also the amount of time that has elapsed since the last arrival (since G is not memoryless).

- To get around this problem we shall only look at the system when a customer arrives; and so let us define $X_n, n \geq 1$, by

$$X_n = \text{the number in the system as seen by the } n\text{th arrival}$$

It is easy to see that the process $\{X_n, n \geq 1\}$ is a Markov chain.

- To compute the transition probabilities P_{ij} for this Markov chain let us first note that, as long as there are customers to be served, the number of services in any length of time t is a Poisson random variable with mean μ_t. Hence

$$P_{i,i+1-j} = \int_0^\infty e^{-\mu t}\frac{(\mu t)^j}{j!}dG(t), j = 0, 1, \ldots, i$$

- The formula for P_{i0} is a little different (it is the probability that at least $i+1$ Poisson events occur in a random length of time having distribution G) and can be obtained from

$$P_{i0} = 1 - \sum_{j=0}^{i} P_{i,i+1-j}$$

- The limiting probabilities $\pi_k, k = 0, 1, \ldots,$ can be obtained as the unique solution of

$$\pi_k = \sum_{i=0}^{\infty}\pi_i P_{ik} = \sum_{i=k-1}^{\infty}\pi_i \int_0^\infty e^{-\mu t}\frac{(\mu t)^{t+1-k}}{(i+1-k)!}dG(t), k \geq 1, \sum_{k=0}^{\infty}\pi_k = 1$$

- By some mathematical techniques, we can obtain that

$$\pi_k = (1-\beta)\beta^k, k = 0, 1, \ldots$$

where β is the solution of the following equation

$$\beta = \int_0^\infty e^{-\mu t(1-\beta)} dG(t).$$

As π_k is the limiting probability that an arrival sees k customers, it is just the a_k as defined in Section 4.2. Hence,

$$a_k = (1-\beta)\beta^k, k \geq 0$$

- We can obtain W by conditioning on the number in the system when a customer arrives. This yields

$$\begin{aligned} W &= \sum_k \mathrm{E}[\text{time in system}|\text{arrival sees } k](1-\beta)\beta^k \\ &= \sum_k \frac{k+1}{\mu}(1-\beta)\beta^k = \frac{1}{\mu(1-\beta)} \end{aligned}$$

- Then

$$\begin{aligned} W_Q &= W - \frac{1}{\mu} = \frac{\beta}{\mu(1-\beta)} \\ L &= \lambda W = \frac{\lambda}{\mu(1-\beta)} \\ L_Q &= \lambda W_Q = \frac{\lambda\beta}{\mu(1-\beta)} \end{aligned}$$

where λ is the reciprocal of the mean interarrival time. That is,

$$\frac{1}{\lambda} = \int_0^\infty x dG(x)$$

To obtain the P_k we first note that the rate at which the number in the system changes from $k-1$ to k must equal the rate at which it changes from k to $k-1$. Notice that

- rate number in system goes from $k-1$ to $k = \lambda a_{k-1}$.
- rate number in system goes from k to $k-1 = P_k \mu$.

Then these rates yields

$$P_k = \frac{\lambda}{\mu} a_{k-1} = \frac{\lambda}{\mu}(1-\beta)\beta^{k-1}$$

and as $P_0 = 1 - \sum_{k=1}^{\infty} P_k$, we otain

$$P_0 = 1 - \frac{\lambda}{\mu}$$

The G/M/1 Busy and Idle Periods

- Suppose that an arrival has just found the system empty-and so initiates a busy period and let N denote the number of customers served in that busy period. Since the Nth arrival (after the initiator of the busy period) will also find the system empty, it follows that N is the number of transitions for the Markov chain (of Section above) to go from state 0 to state 0. Hence, $1/\mathrm{E}[N]$ is the proportion of transitions that take the Markov chain into state 0

- Hence, $1/\mathrm{E}[N]$ is the proportion of transitions that take the Markov chain into state 0; or equivalently, it is the proportion of arrivals that find the system empty. Therefore,

$$\mathrm{E}[N] = \frac{1}{a_0} = \frac{1}{1-\beta}$$

- The sum of a busy and idle period can be expressed as the sum of N interarrival times. Thus, if T_i is the ith interarrival time after the busy period begins, then

$$\mathrm{E}[\text{Busy}] + \mathrm{E}[\text{Idle}] = \mathrm{E}\left[\sum_{i=1}^{N} T_i\right] = \mathrm{E}[N]\mathrm{E}[T] = \frac{1}{\lambda(1-\beta)}$$

- For a second relation between $\mathrm{E}[\text{Busy}]$ and $\mathrm{E}[\text{Idle}]$, we can use the same argument that

$$1 - P_0 = \frac{\mathrm{E}[\text{Busy}]}{\mathrm{E}[\text{Idle}] + \mathrm{E}[\text{Busy}]}$$

and since $P_0 = 1 - \lambda/\mu$, we obtain that

$$\begin{aligned} \mathrm{E}[\text{Busy}] &= \frac{1}{\mu(1-\beta)}, \\ \mathrm{E}[\text{Idle}] &= \frac{\mu - \lambda}{\mu(1-\beta)}. \end{aligned}$$

4.7. Multiserver Queues

By and large, systems that have more than one server are much more difficult to analyze than those with a single server.

Erlang's Loss System

- A loss system is a queueing system in which arrivals that find all servers busy do not enter but rather are lost to the system.

- The simplest such system is the M/M/k loss system in which customers arrive according to a Poisson process having rate λ, enter the system if at least one of the k servers is free, and then spend an exponential amount of time with rate μ being served. The balance equations for this system are

State	Rate leave = Rate enter
0	$\lambda P_0 = \mu P_1$
1	$(\lambda + \mu)P_1 = 2\mu P_2 + \lambda P_0$
2	$(\lambda + 2\mu)P_2 = 3\mu P_3 + \lambda P_1$
$i, 0 < i < k$	$(\lambda + i\mu)P_i = (i+1)\mu P_{i+1} + \lambda P_{i-1}$
k	$k\mu P_k = \lambda P_{k-1}$

- Rewriting gives

$$\begin{aligned} P_1 &= \frac{\lambda}{\mu} P_0, \\ P_2 &= \frac{\lambda}{2\mu} P_1 = \frac{(\lambda/\mu)^2}{2} P_0, \\ P_3 &= \frac{\lambda}{3\mu} P_1 = \frac{(\lambda/\mu)^3}{3!} P_0, \\ &\vdots \\ P_k &= \frac{\lambda}{k\mu} P_1 = \frac{(\lambda/\mu)^k}{k!} P_0, \end{aligned}$$

and using $\sum_{i=0}^{k} P_i = 1$, we obtain

$$P_i = \frac{(\lambda/\mu)^i / i!}{\sum_{j=0}^{k} (\lambda/\mu)^j / j!}, i = 0, 1, \ldots, k$$

- Since $\mathrm{E}[S] = 1/\mu$, where $\mathrm{E}[S]$ is the mean service time, the preceding can be written as

$$P_i = \frac{(\lambda \mathrm{E}[S])^i / i!}{\sum_{j=0}^{k} (\lambda \mathrm{E}[S])^j / j!}, i = 0, 1, \ldots, k$$

- Consider now the same system except that the service distribution is general– that is, consider the M/G/k with no queue allowed. This model is sometimes called the Erlang loss system. It can be shown (though the proof is advanced) that the equation above (which is called *Erlang's loss formula*) remains valid for this more general system.

- It is easy to see that equation above is valid when k $= 1$. For in this case, $L = P_1, W = \mathrm{E}[S]$,and $\lambda_a = \lambda P_0$ .Using that $L = \lambda_a W$ gives

$$P_1 = \lambda P_0 \mathrm{E}[S]$$

which implies, since $P_0 + P_1 = 1$, that

$$P_0 = \frac{1}{1 + \lambda \mathrm{E}[S]}, \; P_1 = \frac{\lambda \mathrm{E}[S]}{1 + \lambda \mathrm{E}[S]}.$$

The M/M/k Queue

The M/M/k infinite capacity queue can be analyzed by the balance equation technique. We obtain that

$$P_i = \begin{cases} \dfrac{(\lambda/\mu)^i/i!}{\sum_{i=0}^{k-1} \frac{(\lambda/\mu)^i}{i!} + \frac{(\lambda/\mu)^k}{k!}\frac{k\mu}{k\mu-\lambda}}, & i \leq k \\ \dfrac{(\lambda/k\mu)^i k^k}{k!} P_0, i > k \end{cases}$$

where we need to impose the condition $\lambda < k\mu$.

The G/M/k Queue

In this model we again suppose that there are k servers, each of whom serves at an exponential rate μ. However, we now allow the time between successive arrivals to have an arbitrary distribution G. To ensure that a steady-state (or limiting) distribution exists, we assume the condition $1/\mu_G < k\mu$ where μ_G is the mean of G.

- Let X_n be the number in the system at the moment of the nth arrival, then $\{X_n, n \geq 0\}$ is a Markov chain.

- To derive the transition probabilities of the Markov chain, it helps to first note the relationship

$$X_{n+1} = X_n + 1 - Y_n, n \geq 0$$

  where Y_n denotes the number of departures during the interarrival time between the nth and $(n+1)$st arrival. The transition probabilities P_{ij} can now be calculated as follows:

Case 1. $j > i + 1$.
In this case it easily follows that $P_{ij} = 0$.

Case 2. $j \leq i + 1 \leq k$.
Conditioning on the length of this interarrival time yields,

$$\begin{aligned} P_{ij} &= P\{i+1-j \text{ of } i+1 \text{ services are completed in an interarrival time}\} \\ &= \int_0^\infty P\{i+1-j \text{ of } i+1 \text{ are completed}|\text{interarrival time is } t\} dG(t) \\ &= \int_0^\infty \binom{i+1}{j} (i - e^{-\mu t})^{i+1-j} (e^{-\mu t})^j dG(t) \end{aligned}$$

where the last equality follows since the number of service completions in a time t will have a binomial distribution.

Case 3. $i+1 \geq j \geq k$.

To evaluate P_{ij} in this case we first note that when all servers are busy,the departure process is a Poisson process with rate $k\mu$. Hence, again conditioning on the interarrival time we have

$$\begin{aligned} P_{ij} &= P\{i+1-j \text{ departures}\} \\ &= \int_0^\infty P\{i+1-j \text{ departures in time } t\} dG(t) \\ &= \int_0^\infty e^{-k\mu t} \frac{(k\mu t)^{i+1-j}}{(i+1-j)!} dG(t) \end{aligned}$$

Case 4. $i+1 \geq k > j$.

In this case since when all servers are busy the departure process is a Poisson process, it follows that the length of time until there will only be k in the system will have a gamma distribution with parameters $i+1-k$, $k\mu$ (the time until $i+1-k$ events of a Poisson process with rate $k\mu$ occur is gamma distributed with parameters $i+1-k, k\mu$). Conditioning first on the interarrival time and then on the time until there are only k in the system (call this latter random variable T_k) yields

$$
\begin{aligned}
P_{ij} &= \int_0^\infty P\{i+1-j \text{ departures in time } t\} dG(t) \\
&= \int_0^\infty \int_0^t P\{i+1-j \text{ departures in } t | T_k = s\} k\mu e^{-k\mu s} \frac{(k\mu s)^{i-k}}{(i-k)!} ds dG(t) \\
&= \int_0^\infty \int_0^t \binom{k}{j} (1-e^{-\mu(t-s)})^{k-j} (e^{-\mu(t-s)})^j k\mu e^{-k\mu s} \frac{(k\mu s)^{i-k}}{(i-k)!} ds dG(t)
\end{aligned}
$$

where the last equality follows since of the k people in service at time s the number whose service will end by time t is binomial with parameters k and $1-e^{-\mu(t-s)}$.

We now can verify either by a direct substitution into the equations $\pi_j = \sum_i \pi_i P_{ij}$, , that the limiting probabilities of this Markov chain are of the form

$$\pi_{k-1+j} = c\beta^j, j = 0, 1, \ldots$$

Substitution into any of the equations $\pi_j = \sum_i \pi_i P_{ij}$ when $j > k$ yields that β is given as the solution of

$$\beta = \int_0^\infty e^{-k\mu t(1-\beta)} dG(t)$$

The values $\pi_0, \pi_1, \ldots, \pi_{k-2}$ can be obtained by recursively solving the first $k-1$ of the steady-state equations, and c can then be computed by using $\sum_{i=0}^{\infty} \pi_i = 1$.

The M/G/k Queue

Consider the M/G/k system in which customers arrive at a Poisson rate λ and are served by any of k servers, each of whom has the service distribution G.

- If we attempt to mimic the analysis presented in the section before for the M/G/1 system, then we would start with the basic identity

$$V = \lambda \mathrm{E}[S] W_Q + \lambda \mathrm{E}[S^2]/2$$

  and then attempt to derive a second equation relating V ad W_Q.

- Now if we consider an arbitrary arrival, then we have the following identity:

  work in system when customer arrives $= k \times$time customer spends in queue$+R$

  where R is the sum of the remaining service times of all other customers in service at the moment when our arrival enters service.

- The foregoing follows because while the arrival is waiting in queue, work is being processed at a rate k per unit time (since all servers are busy). Thus, an amount of work $k\times$time in queue is processed while he waits in queue. Now, all of this work was present when he arrived and in addition the remaining work on those still being served when he enters service was also present when he arrived–so we obtain the equation above.

- For an illustration, suppose that there are three servers all of whom are busy when the customer arrives. Suppose, in addition, that there are no other customers in the system and also that the remaining service times of the three people in service are 3, 6, and 7. Hence, the work seen by the arrival is $3 + 6 + 7 = 16$. Now the arrival will spend 3 time units in queue, and at the moment he enters service, the remaining times of the other two customers are $6 - 3 = 3$ and $7 - 3 = 4$. Hence, $R = 3 + 4 = 7$ and as a check of the equation above we see that 16$= 3 \times 3 + 7$.

- Taking expectations of the equation above and using the fact that Poisson arrivals see time averages, we obtain

$$V = kW_Q + \mathrm{E}[R]$$

which, along with the equation shown before, would enable us to solve for W_Q if we could compute $\mathrm{E}[R]$. However there is no known method for computing $\mathrm{E}[R]$ and in fact, there is no known exact formula for W_Q.

- The following approximation for W_Q was obtained in the reference by using the foregoing approach and then approximating $\mathrm{E}[R]$:

$$W_Q \approx \frac{\lambda^k \mathrm{E}[S^2](\mathrm{E}[S])^{k-1}}{2(k-1)!(k-\lambda\mathrm{E}[S])^2\left[\sum_{n=0}^{k-1}\frac{(\lambda\mathrm{E}[S])^n}{n!} + \frac{(\lambda\mathrm{E}[S])^k}{(k-1)!(k-\lambda\mathrm{E}[S])}\right]}$$

- The preceding approximation has been shown to be quite close to W_Q when the service distribution is gamma. It is also exact when G is exponential.