

Single Server Queueing Systems

22nd April 2007

Contents

1	Introduction - What Is A Queue?	2
1.1	Input	2
1.2	Queue Discipline	2
1.3	Service Mechanism	2
1.4	Cost Structure	3
2	Basic Queueing Identities	3
2.1	Steady-State Probabilities	4
3	Exponential Queueing Models	5
3.1	A Single Server Exponential Queueing System	5
3.2	A Single Server Exponential Queueing System with finite capacity	9
4	Application - Profit Maximisation	11
5	Poisson Processes More Formally	12
5.1	Alternative Method for Deriving Equatons 3.1	15
6	Appendices	17
6.1	Appendix 1	17
6.2	Appendix 2	18
6.3	Appendix 3	19
6.4	Appendix 4	20
6.5	Appendix 5	20
6.6	Appendix 6	21

1 Introduction - What Is A Queue?

[1]

A queue occurs when potential customers arrive at a system that offers a certain facility or service that the customers wish to use. For example, Aircraft waiting to land at or depart from an Airport form queues, as do cutomers awaiting service at supermarket checkouts, or thrillseekers waiting to board a ride at a theme park. Clearly, there are very many different forms such a system can take, defined principally by the following four factors:

1.1 Input

The Input is the manner in which customers join the system, either as individuals or in groups. A *waiting system* is one in which every arriving customer joins the system. Is the system only has *finite capacity*, a customer may not be able to join the system if he/she arrives to find it already full.

1.2 Queue Discipline

The Queue Discipline is how the queue is formed, and the behaviour of customers whilst they are waiting in it. A simple example is the *First Come, First Served* system, whereby customers are served in the exact same order in which they joined the system in. The reverse of this is to have a *Last Come, First Served* discipline. Systems could have multiple servers in parallel, giving customers a choice as to which queue they which join (where each service counter has it's own queue, subject to it's own individual discipline). Customers can be divided into classes, with a specific discipline for each class (such as in a hospital Accident & Emergency department, where more severely injured patients are treated sooner than those with minor injuries).

Differing customer behaviour can also be a factor, with impatient customers leaving (*reneging*) the system if they have been waiting too long, or in systems with multiple servers in parallel, customers could switch (*jockey*) between different queues in order to try and be served as fast as possible.

1.3 Service Mechanism

The way in which customers are served is called the *Service Mechanism*. In a system with a finite number of servers s , customers will be served in a specified order. For example, in a First Come, First Served system, the customer who has waited the longest without being served is attended by the first one of the s servers to become free.

If the s servers are arranged in *series*, this means that customers are serviced at each of the s counters once, in a set order, leaving the system once they have recieved service at the sth server. There may be waiting provision before a particular counter, but if this becomes full (or there is no waiting provision at all), then *blocking* may occur, whereby the customer at the kth counter, having

finished in service, is unable to proceed to the $(k + 1)$ st counter (or its queue), due to the ongoing service of a customer at the $(k + 1)$ st counter, so counter k will be blocked until counter $k + 1$ becomes available.

A *queueing network* is comprised of multiple different systems, with their own individual input, queue discipline and service mechanism, with customers moving between these systems in a specified manner.

1.4 Cost Structure

The *cost structure* is comprised of the revenue the operator receives from each customer who is served (the *reward*), and the operating costs of the system, made up of the cost of providing the servers and a holding cost for waiting facilities (a function of the number of customers present or the delay they endure, for example, providing tea and coffee for customers waiting in a hairdressers). Usually, it is in the operator's interest to find the service level that will maximise profit that is made.

2 Basic Queueing Identities

The following quantities are of interest for queueing models:

L := The average number of customers in the system.

L_Q := The average number of customers waiting in a 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 a queue.

Let

N_t := The Number of Customers who have arrived by time t ,

and;

$$\lambda_a = \lim_{t \rightarrow \infty} \frac{N_t}{t}$$

Then λ_a is the average rate of arriving customers. Hence, if we assume that each arriving customer pays an 'entrance fee' to join the system, then we deduce the following identity:

Theorem 2.1. *Let R = Average Rate at which system earns,
 P = Average Amount Entering Customer Pays. Then:-*

$$R = \lambda_a P \tag{2.1}$$

Note: The following proof is only a heuristic argument, and is taken from the given reference.

Proof. [3] Let T be a large fixed number. In two different ways, we will compute the average amount of money the system has earned by time T . 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 . 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 ($\approx \lambda_a T$). Hence, both sides of Equation 2.1 when multiplied by T are approximately equal to the average amount earned by T . The result then follows by letting $T \rightarrow \infty$. \square

Useful formulas can now be derived via special cases of equation 2.1. In one instance, the rate at which the system earns can be expressed as sL , if customers pay $\pounds s$ per unit time whilst in the system. Also, a customer who has been in the system for a length of time W in the system will pay $\pounds sW$ for the this time in the system, so the rate at which the system earns by this argument is $s\lambda_a W$. Hence, equating the two gives;

$$sL = s\lambda_a W$$

And letting $s = 1$ (i.e. customers pay $\pounds 1$ per unit time in the system) gives rise to Little's Formula: [3]

$$L = \lambda_a W \tag{2.2}$$

If we repeat the above reasoning, but only considering the customers in a queue, we can also say:

$$L_Q = \lambda_a W_Q \tag{2.3}$$

Furthermore, if we define:

$$E[S] = \text{average time a customer spends in service}$$

It follows that:

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

Which is obtained from Equation 2.1, by assuming customers pay $\pounds 1$ per unit time whilst in service (and hence, on average, a customer will pay $E[S]$), and the average number of customers in service is equal to the rate at which the system earns.

2.1 Steady-State Probabilities

Let

$$X(t) = \text{Number of customers in system at time } t$$

and, for $n > 0$,

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

Hence, P_n describes the proportion of time that there are exactly n customers in the system in the long-run. For example, if $P_3 = 0.4$, then there will be exactly 3 customers in the system for about 40% of the time, over a long period of time.¹

¹Provided queueing process is regenerative

Two other useful limiting probabilities are $\{a_n : n \geq 0\}$ and $\{d_n : n \geq 0\}$, defined as:

a_n = Proportion of customers who find n customers when they arrive.

d_n = Proportion of departures who leave with n customers in the system behind them.

In general, it is not true that $a_n = d_n = P_n$, as demonstrated by the following example:

Example 1. *Suppose we have a system where every customer has service time t , and customers arrive at intervals of $2t$ with probability 1. Then, it is clear that customer k 's service has finished (and they have left the system) before customer $k + 1$ arrives. Thus, every arriving customer finds the system empty and every departing customer leaves the system empty, so:*

$$a_0 = d_0 = 1$$

$$a_k = d_k = 0 \text{ for } k \geq 1$$

but $P_0 \neq 1$ and $P_1 \neq 0$ all of the time

Proposition 2.2. *For Poisson arrivals,*

$$P_n = a_n$$

Proof. Take T to be a large number. Then, for any $n \in \mathbb{N}$, the length of time the system has been in state n (the times when there are n customers in the system) during the time interval $[0, T]$ is approximately $P_n T$. As a Poisson(λ) process is memoryless, the rate of arrival of customers to the system is independent of the current state of the system, and thus arrivals always occur at rate λ . Hence, approximately $\lambda P_n T$ customers arriving in the interval $[0, T]$ find n already in the system. Therefore, in the long run, customers finding the system in state n arrive at rate λP_n . Since λ is the overall arrival rate, the proportion of customers who arrive to find the system in state n (i.e. a_n) is:

$$\frac{\lambda P_n}{\lambda} = P_n$$

and thus:

$$P_n = a_n$$

□

3 Exponential Queueing Models

3.1 A Single Server Exponential Queueing System

In this section, I will consider a system (with one server) where customers arrive according to a Poisson process with rate λ (i.e. on average, λ customers

arrive per some fixed time interval). Equivalently, customers arrive with inter-arrival times (the time between customer k arriving and customer $k+1$ arriving) distributed independently according to exponential random variables with mean $\frac{1}{\lambda}$.

In this model, an arriving customer enters service immediately if the server is available, or joins the rear of the queue if the server is busy (or forms the start of the queue if no queue exists). When one customer's service has been completed, the person who is currently at the front of the queue enters service immediately, if there is such a person. Otherwise, the server becomes vacant until another customer enters the system. Additionally, service times are assumed to be independent exponential random variables with mean $\frac{1}{\mu}$ (hence customers are served according to a Poisson process with rate μ). I will also assume that there is no maximum capacity in the system, so any state $n \in \mathbb{N} \cup \{0\}$ is possible.

To examine this type of queueing system, I will consider the limiting probabilities $\{P_n | n \in \mathbb{N}\}$.

Suppose that in a given time period, $[0, T]$ say, there are k 'sub-periods', $S_k | k \in \{1, \dots, n\}$ (in time order S_1, S_2, \dots) during each of which there are n customers in the system, throughout the sub-period. There is no intersection at all between any of these sub-periods and there are never exactly n customers in the system outside of these sub-periods. If we now consider the interval $[0, T]$, either of the following must be true:

- $t = 0 \in S_1$ and $t = T \in S_k$. Here we leave state n $(k - 1)$ times, and enter state n $(k - 1)$ times.
- $t = 0 \in S_1$ and $t = T \notin S_k$. In this case we leave state n k times and enter state n $(k - 1)$ times.
- $t = 0 \notin S_1$ and $t = T \in S_k$. We leave state n $(k - 1)$ times and enter state n k times.
- $t = 0 \notin S_1$ and $t = T \notin S_k$. We leave state n k times and enter state n k times.

Clearly, in all of the above cases, we can see that the number of times we enter state n is equal to within one the number of times we leave state n . Hence, in the long run, *the rate at which the system enters state n is equal to the rate at which the system leaves state n* . Henceforth, I will refer to this as the *rate-equality principle*.

Now we need to consider the various states of the system, using our rate-equality principle:

n=1

Clearly, we can leave state 1 only via an arrival to take the system to state 2 (since when there are 0 customers in the system initially, it is impossible for anyone to leave). As customers arrive to the system at rate λ , and the proportion

of time spent in state 0 is P_0 , the rate at which we leave state 0 is λP_0 . We can only enter state 0 via a departure from state 1. (Arrivals to take us to state 0 cannot occur due to there being no state $n = -1$ defined). Departures occur in accordance to the service process, at rate μ , and the proportion of time spent in state 1 is P_1 . Hence, we enter state 0 at a rate μP_1 . Thus, by the rate-equality principle, we obtain,

$$\lambda P_0 = \mu P_1$$

$\mathbf{n} | \mathbf{n} \in \mathbb{N}$

We can enter state i either via an arrival from state $i-1$ or a departure from state $i+1$. By the same reasoning as before, these occur at rates λP_{i-1} and μP_{i+1} . Hence, we enter state i at an exponential rate $\lambda P_{i-1} + \mu P_{i+1}$.² Furthermore, we can leave state i via either a departure to take the system to state $i-1$, or an arrival to take the system to state $i+1$, which occur at exponential rates μP_i and λP_i respectively. Thus, by the rate-equality principle, we obtain,

$$(\mu + \lambda)P_i = \lambda P_{i-1} + \mu P_{i+1}$$

This leads to the following system of equations:

$$\begin{aligned} P_1 &= \frac{\lambda}{\mu} P_0 \\ P_{i+1} &= P_i + \frac{\lambda}{\mu} P_i - \frac{\lambda}{\mu} P_{i-1} \end{aligned}$$

So, solving for each P_n ,

$$\begin{aligned} P_0 &= P_0 \\ P_1 &= \frac{\lambda}{\mu} P_0 \\ P_2 &= P_1 + \frac{\lambda}{\mu} P_1 - \frac{\lambda}{\mu} P_0 = \frac{\lambda}{\mu} P_0 + \left(\frac{\lambda}{\mu}\right)^2 P_0 - \frac{\lambda}{\mu} P_0 = \left(\frac{\lambda}{\mu}\right)^2 P_0 \\ P_3 &= P_2 + \frac{\lambda}{\mu} P_2 - \frac{\lambda}{\mu} P_1 = \left(\frac{\lambda}{\mu}\right)^2 P_0 + \left(\frac{\lambda}{\mu}\right)^3 P_0 - \left(\frac{\lambda}{\mu}\right)^2 P_0 = \left(\frac{\lambda}{\mu}\right)^3 P_0 \end{aligned}$$

So, inductively, we obtain:

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0 \tag{3.1}$$

Clearly, $\sum_{n=0}^{\infty} P_n = 1$, for $\lambda < \mu$, and if this is true, then:

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

²See Appendix 1 for proof of this statement

So we get that:

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

Note the necessity that $\lambda < \mu$, i.e. the arrival rate is less than the departure rate, otherwise:

$$\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n P_0 = \infty \text{ and } P_n = 0 \forall n$$

Henceforth, I will assume that $\lambda < \mu$.

Since P_n is the long-run probability that there are n customers in the system, we can calculate the average number of customers L in the system as follows:

$$L = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) = 3 \frac{\lambda}{\mu - \lambda}$$

Now, by equation 2.2, we obtain the average time a customer spends in the system, and also their average queuing time:

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

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

And, by equation 2.3, we obtain:

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

Example 2. *At Lostock Parkway railway station, Greater Manchester, there were 44849 station entries [4] (i.e. people starting a train journey at the station) in the year 2002/03. Making the assumption that half of these (which I will round down to 22424) wished to buy a ticket from the ticket office prior to boarding the train (i.e. they were about to start the outward leg of a single or return journey, with the half I am ignoring already holding a ticket for a return journey leg or a season ticket), and that demand for tickets is spread evenly over the ticket office opening hours 06:20 - 22:50 daily (16.5 hours per day, 364 days per year)⁴, meaning customers arrive at a Poisson rate of one every 16 minutes ($\lambda = \frac{1}{16}$), and assuming that we have an exponential service time of one every 2 minutes ($\mu = \frac{1}{2}$), what are L , W , L_Q and W_Q ?*

$$L = \frac{\frac{1}{16}}{\frac{1}{2} - \frac{1}{16}} = \frac{1}{7}$$

³See Appendix 2 for proof of this identity

⁴In reality, footfall is not as evenly spread as these assumptions suggest, but, for simplicity, I will use them

$$W = \frac{1}{\frac{1}{2} - \frac{1}{16}} = \frac{16}{7}$$

$$W_Q = \frac{\frac{1}{16}}{\frac{1}{2}(\frac{1}{2} - \frac{1}{16})} = \frac{2}{7}$$

$$L_Q = \frac{(\frac{1}{16})^2}{\frac{1}{2}(\frac{1}{2} - \frac{1}{16})} = \frac{1}{56}$$

This tells us that on average there are $\frac{1}{7}$ customers in the system, customers spend on average $\frac{16}{7}$ minutes (about 2 minutes and 17 seconds) in the system. Of this time, an average of $\frac{2}{7}$ minutes (about 17 seconds) in a queue (unsurprising as the average service time is 2 minutes and expectations transform linearly), and the queue contains, on average, $\frac{1}{56}$ people.

3.2 A Single Server Exponential Queueing System with finite capacity

In reality, almost all forms of queueing systems will be subject to some sort of capacity constraint, in that we are limited to a maximum of N customers in the system at any one time. Again, if we assume customers arrive at a Poisson rate λ and are served at a Poisson rate μ , and we will use the same rate equality principle as previously. Hence, by similar reasoning we get the following for state 0:

$$\lambda P_0 = \mu P_1$$

and for states $1, \dots, N-1$:

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

And, considering state N , we can only leave via a departure (at Poisson rate μ), as state $N+1$ is not defined, and we can only enter state N via an arrival from state $N-1$ (at Poisson rate λ) as we can never arrive from state $N+1$. This gives rise to the final equation:

$$\mu P_N = \lambda P_{N-1}$$

So, as before, we get:

$$P_1 = \frac{\lambda}{\mu} P_0$$

$$\vdots$$

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

$$P_N = \frac{\lambda}{\mu} P_{N-1} = \left(\frac{\lambda}{\mu}\right)^N P_0$$

Notice that $\sum_{n=0}^N P_n = 1$, so

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

Thus

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

$$P_n = \left(\frac{\lambda}{\mu}\right)^n \left(\frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}} \right) \text{ for } n \in [0, 1, \dots, N]$$

Notice that it is not necessary that $\frac{\lambda}{\mu} < 1$, since we are only summing a finite number of terms.

So, similarly to before, we have the average number of customers in the system:

$$L = \sum_{n=0}^N n P_n$$

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

$$= \frac{1 + N \left(\frac{\lambda}{\mu}\right)^{N+1} - (N+1) \left(\frac{\lambda}{\mu}\right)^N}{(\mu - \lambda) \left(1 - \left(\frac{\lambda}{\mu}\right)^N \right)}$$

In the following, let us assume that a customer is counted if they actually enter the queuing system, thus ignoring any potential customers who arrive to find the system full, and are thus unable to enter the system due to there being insufficient capacity. This will only occur when the system is in state N , which occurs with probability P_N . Hence, the proportion of arrivals who enter the system is:

$$1 - P_N = 1 - \frac{\left(\frac{\lambda}{\mu}\right)^N \left(1 - \frac{\lambda}{\mu} \right)}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}}$$

⁵See Appendix 3 for Proof of this identity

As customers arrive to the system at Poisson rate λ , the rate at which customers join the system, λ_a , is:

$$\lambda_a = \lambda(1 - P_N) = \lambda - \frac{\lambda \left(\frac{\lambda}{\mu}\right)^N \left(1 - \frac{\lambda}{\mu}\right)}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}} = \frac{\lambda - \lambda \left(\frac{\lambda}{\mu}\right)^{N+1} - \lambda \left(\frac{\lambda}{\mu}\right)^N \left(1 - \frac{\lambda}{\mu}\right)}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}}$$

and we can deduce the average time a customer spends in the system:

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

4 Application - Profit Maximisation

What I have done already can be used in industry to calculate the optimal level of service needed in, for example, a shop, in order to maximise total profit. In the following, suppose that, in order to provide service at rate μ , there is an associated cost of $\mathcal{L}c\mu$ per hour, and each customer whose service is fully completed generates $\mathcal{L}A$ revenue, with, again, a maximum system capacity of N customers in the system at any one time. I am aiming to find the rate of service, μ , which maximises the total profit made in a single server exponential queueing system with finite capacity N .

As explained previously, customers join the system at an hourly Poisson rate of $\lambda(1 - P_n)$, where λ is the hourly rate of arrival of potential customers. As each customer generates $\mathcal{L}A$ of gross profit, the rate at which gross profit is made is $\mathcal{L}\lambda A(1 - P_n)$ per hour, with the system costing $\mathcal{L}c\mu$ to run for the hour at the hourly service rate μ . This gives us:

$$\begin{aligned} \text{Profit Per Hour} &= \lambda A(1 - P_n) - c\mu \\ &= \lambda A \left[1 - \frac{\left(\frac{\lambda}{\mu}\right)^N \left(1 - \frac{\lambda}{\mu}\right)}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}} \right] - c\mu = \frac{\lambda A - \lambda A \left(\frac{\lambda}{\mu}\right)^{N+1} - \lambda A \left(\frac{\lambda}{\mu}\right)^N \left(1 - \frac{\lambda}{\mu}\right)}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}} - c\mu \\ &= \frac{\lambda A \left(1 - \left(\frac{\lambda}{\mu}\right)^N\right)}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}} - c\mu \end{aligned} \quad (4.1)$$

It is my aim to find the value(s) of μ that maximises this function.

Example 3. [3] Suppose we are considering a car wash, where, due to size limitations, a maximum of 2 cars can be present on site at any one time (one being washed, the other waiting, with queuing on adjacent roads not permitted), so $N = 2$. Cars arrive at an according to an exponential distribution with mean 1 per hour (so $\lambda = \frac{1}{1} = 1$). Each customer pays $\mathcal{L}5$ for a wash ($A = 5$), and each service incurs a cost to the car wash of $\mathcal{L}2$ ($c = 2$). We wish to find the

service rate μ which maximises total profit. Substituting $N = 2, \lambda = 1, A = 5$ and $c = 2$ into equation 4.1 gives:

$$\begin{aligned} \text{Profit per hour} &= \frac{5 \left(1 - \left(\frac{1}{\mu}\right)^2\right)}{1 - \left(\frac{1}{\mu}\right)^3} - \mu = \frac{5 - 5\frac{1}{\mu^2}}{1 - \frac{1}{\mu^3}} - \mu = \frac{5\mu^3 - 5\mu}{\mu^3 - 1} - \mu = f(\mu) \\ \Rightarrow f'(\mu) &= \frac{d}{d\mu} \left(\frac{5\mu^3 - 5\mu}{\mu^3 - 1} - \mu \right) = \frac{(\mu^3 - 1)(15\mu^2 - 5) - (5\mu^3 - 5\mu)3\mu^2}{(\mu^3 - 1)^2} - 1 \\ &= \frac{10\mu^3 - 15\mu^2 + 5}{(\mu^3 - 1)^2} - 1 \end{aligned}$$

Hence, the value of μ that maximises total profit can be found by solving the equation $f'(\mu) = 0$ numerically, then substituting the values of μ found back into the equation $f(\mu)$ to find the one that gives the highest total profit. In this example, if we equate $f'(\mu) = 0$ then it can be found that there is a solution at $\mu \approx 1.4$. Plotting the graph of f (see end of Appendices), it can be seen that this is the only maximum in the valid range ($\mu > 0$), thus our optimal service rate is $\mu = 1.4$. Thus we have $N = 2, \lambda = 1$ and $\mu = 1.4$, so,

$$\begin{aligned} L &= \frac{1 + 2 \left(\frac{1}{1.4}\right)^3 - (3) \left(\frac{1}{1.4}\right)^2}{(1.4 - 1) \left(1 - \left(\frac{1}{1.4}\right)^2\right)} \approx 1.012 \\ \lambda_a &= \frac{1 - \left(\frac{1}{1.4}\right)^3 - \left(\frac{1}{1.4}\right)^2 \left(1 - \frac{1}{1.4}\right)}{1 - \left(\frac{1}{1.4}\right)^3} \approx 0.178 \end{aligned}$$

and

$$W \approx \frac{1.012}{0.178} \approx 5.685$$

So, on average there are 1.012 customers in the system, each spending an average of 5.685 minutes (about 5 minutes 41 seconds) in the system.

5 Poisson Processes More Formally

[2][5] In this section, I am going to return to looking at how all the rates (i.e. λ, μ , etc.) come about, and I will prove some relationships more formally.

The Poisson process that I have described for arrival to the system, satisfies the following assumptions:

- $\lim_{t \rightarrow 0} P[1 \text{ arrival in time interval } (t, t + h)] = \lambda h$, where $\lambda \in \mathbb{R}$ is constant.
- $\lim_{h \rightarrow 0} P[\text{More than 1 arrival in time interval } (t, t + h)] = 0$ (i.e. $P[> 1 \text{ arrival in } (t, t + h)] = o(h)$)

- The above two properties are independent of t , so Poisson processes are *stationary*, and probabilities associated with the number of arrivals in $(t, t+h)$ are independent of n , the number of people that have entered the system already, so they are *memoryless*.

If we let $p_n(t) = P[n \text{ arrivals in interval } (0, t)]$ we can see (using the first two assumptions above) that,

$$\begin{aligned} & \lim_{h \rightarrow 0} P[0 \text{ arrivals in } (t, t+h)] \\ &= 1 - \lim_{h \rightarrow 0} P[1 \text{ arrival in } (t, t+h)] - \lim_{h \rightarrow 0} P[> 1 \text{ arrival in } (t, t+h)] \\ &= 1 - \lambda h - 0 = 1 - \lambda h \end{aligned}$$

and

$$\lim_{h \rightarrow 0} P[1 \text{ arrival in } (t, t+h)] = \lambda h$$

If we suppose that n people ($n \geq 1$) have arrived by time $t+h$. Therefore at time t either one of the following two mutually exclusive events could have been the case (since, for small h there is no probability of 2 or more arrival in a time interval of length h):

- n people have arrived to the system, with no further arrival in the time interval $(t, t+h)$, which occurs with probability $p_n(t)(1 - \lambda h)$.⁶
- $n-1$ people have arrived, with 1 more arrival in $(t, t+h)$, which occurs with probability $p_{n-1}(t)\lambda h$.

Together these give,

$$p_n(t+h) = p_n(t)(1 - \lambda h) + p_{n-1}(t)\lambda h \quad (\text{for } n \geq 1) \quad (5.1)$$

If 0 people have arrived by time $t+h$, clearly, we were also in state 0 at time t , with no arrivals in $(t, t+h)$, as state $n = -1$ is not defined. Hence, the probability that we are in state 0 at time $t+h$ is equal to the probability that the system was in state 0 at time t , and that there were no subsequent arrivals in $(t, t+h)$, which occurs with probability $p_0(t)(1 - \lambda h)$, due to time independence of arrivals and memorylessness of Poisson random variables. So:

$$p_0(t+h) = p_0(t)(1 - \lambda h) \quad (5.2)$$

Equation 5.1 can be rewritten as follows:

$$\begin{aligned} p_n(t+h) &= p_n(t)(1 - \lambda h) + p_{n-1}(t)\lambda h = p_n(t) - \lambda h p_n(t) + \lambda h p_{n-1}(t) \\ &\iff p_n(t+h) - p_n(t) = \lambda h p_{n-1}(t) - \lambda h p_n(t) \end{aligned}$$

⁶This probability (and that of the alternative case) are given by multiplying the probability of n ($n-1$) arrivals by time t with the probability of 0 (1) arrivals in $(t, t+h)$. We can do this since Poisson arrivals are memoryless.

$$\iff \frac{p_n(t+h) - p_n(t)}{h} = \lambda p_{n-1}(t) - \lambda p_n(t)$$

and similarly for 5.2:

$$\begin{aligned} p_0(t+h) &= p_0(t) - \lambda h p_0(t) \\ \iff p_0(t+h) - p_0(t) &= -\lambda h p_0(t) \\ \iff \frac{p_0(t+h) - p_0(t)}{h} &= -\lambda p_0(t) \end{aligned}$$

As h is small, the left-hand sides in both 5.1 and 5.2 can be replaced by derivatives:

$$\frac{dp_n(t)}{dt} = \lambda p_{n-1}(t) - \lambda p_n(t) \quad (5.3)$$

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t) \quad (5.4)$$

And hence, we have a set of differential equations. If we solve 5.4 first, we get:⁷

$$p_0(t) = e^{-\lambda t}$$

Using this, we can compute p_1 since:

$$\begin{aligned} \frac{dp_1(t)}{dt} &= -\lambda p_1(t) + \lambda e^{-\lambda t} \\ \Rightarrow \frac{dp_1(t)}{dt} + \lambda p_1(t) &= \lambda e^{-\lambda t} \\ \Rightarrow p_1(t) &= {}^8 \lambda t e^{-\lambda t} \end{aligned}$$

Or inductively;

$$p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

By the third assumption, this holds for any time interval, so it depends only on the length of the interval, not on its starting time. Using this, we can now compute the expected number of arrivals to the system in the time interval $(0, t)$. Using that $E(N_t) = \sum_{n=0}^{\infty} n p_n(t)$, we get:

$$\begin{aligned} E(N_t) &= \sum_{n=0}^{\infty} \frac{n \lambda^n t^n}{n!} e^{-\lambda t} = \sum_{n=1}^{\infty} \frac{(\lambda t)^n}{(n-1)!} e^{-\lambda t} = \lambda t e^{-\lambda t} \sum_{n=1}^{\infty} \frac{(\lambda t)^{n-1}}{(n-1)!} \\ &= {}^9 \lambda t e^{-\lambda t} e^{\lambda t} = \lambda t \end{aligned}$$

This shows that the average number of arrivals per unit time (i.e. $t = 1$) is λ , thus λ is the arrival rate.

⁷See Appendix 4

⁸See Appendix 5

⁹Since $\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x$

We can also examine the intervals between 2 successive arrivals. Let:

$$F(x) = P(\text{There is at least 1 arrival in a random time interval } x) = 1 - p_0(x) = 1 - e^{-\lambda x}$$

, so

$$f(x) = F'(x) = \lambda e^{-\lambda x}$$

where $f(x)$ is the probability density function calculated by taking the derivative of the distribution function $F(x)$. $f(x)$ is an expression for the probability of there being a certain gap between two successive arrivals. Hence,

$$E(\text{Time interval between successive arrivals}) = \int_0^{\infty} t \lambda e^{-\lambda t} dt$$

(Substituting $u = t$ and $\frac{dv}{dt} = e^{-\lambda t} \Rightarrow \frac{du}{dt} = 1, v = \frac{-e^{-\lambda t}}{\lambda}$)

$$= \lambda \left[\frac{-te^{-\lambda t}}{\lambda} \right]_0^{\infty} + \int_0^{\infty} e^{-\lambda t} dt = \left[\frac{-e^{-\lambda t}}{\lambda} \right]_0^{\infty} = \frac{1}{\lambda}$$

Hence, the average inter-arrival time is $\frac{1}{\lambda}$.

If we repeat all these calculations for the exponential service random variable (with $\mu \in \mathbb{R}$ constant in assumption 1 in place of λ), μ is the average number of services completed per unit time, with $\frac{1}{\mu}$ the average duration of service - in other words, exactly as I have shown, but with λ in place of μ .

5.1 Alternative Method for Deriving Equatons 3.1

If we continue to use the first two assumptions stated earlier, letting:

$$P_n(t) = P(n \text{ customers in the system at time } t)$$

$$\lim_{h \rightarrow 0} P(0 \text{ arrivals in } (t, t+h)) = 1 - \lambda h$$

$$\lim_{h \rightarrow 0} P(1 \text{ arrival in } (t, t+h)) = \lambda h$$

$$\lim_{h \rightarrow 0} P(0 \text{ departures in } (t, t+h)) = 1 - \mu h$$

$$\lim_{h \rightarrow 0} P(1 \text{ departure in } (t, t+h)) = \mu h$$

Clearly, we can have exactly one of the following possibilities of occurrences in the time interval $(t, t+h)$ if there are n ($n \geq 1$) people in the system at time $t+h$:

- n people in the system at time t , with one arrival and one departure.
- n people in the system at time t , with no arrival or departure.
- $n - 1$ people in the system at time t , with one arrival and no departures.
- $n + 1$ people in the system at time t , with one departure and no arrival.

Due to the time independence and memorylessness properties of Poisson random variables, these occur with probabilities $P_n(t)(1 - \lambda h)(1 - \mu h)$, $P_n(t)\lambda h\mu h$, $P_{n-1}(t)\lambda h(1 - \mu h)$ and $P_{n+1}(t)(1 - \lambda h)\mu h$ respectively. Therefore, the probability of being at state n at time $t + h$ is the sum of these,

$$P_n(t + h) =$$

$$P_n(t)(1 - \lambda h)(1 - \mu h) + P_n(t)\lambda h\mu h + P_{n-1}(t)\lambda h(1 - \mu h) + P_{n+1}(t)(1 - \lambda h)\mu h$$

and for $n = 0$, we could have either been at state $n = 0$ at time t with no arrival or departure, or at state 1 with a departure in $(t, t + h)$, so by similar reasoning,

$$P_0(t + h) = P_0(t)(1 - \lambda h) + P_1(t)(1 - \lambda h)\mu h$$

These two equations can be rearranged to:

$$\frac{P_n(t + h) - P_n(t)}{h} = \lambda P_{n-1}(t) + \mu P_{n+1}(t) - (\mu + \lambda)P_n(t)$$

and

$$\frac{P_0(t + h) - P_0(t)}{h} = \mu P_1(t) - \lambda P_0(t)$$

As h is small, the left hand sides can be replaced with derivatives:

$$\frac{dP_n(t)}{dt} = \lambda P_{n-1}(t) + \mu P_{n+1}(t) - (\mu + \lambda)P_n(t)$$

$$\frac{dP_0(t)}{dt} = \mu P_1(t) - \lambda P_0(t)$$

The following is taken from [2]: "These equations represent a set of differential equations called *Kolmogorov Differential Equations*. Their solution is a set of equations showing how each probability changes with time... Because of the third assumption, after the some transition period the system will become stable. Of course the state will permanently change, but the probabilities of various numbers of customers in the system will be constant, so the functions $P_n(t)$ become constants P_n ." This means that in the two differential equations above, $P_i(t)$ can be replaced with P_i , and since $\lim_{t \rightarrow \infty} \frac{dP_i(t)}{dt} = 0$ by the asymptotic stability of the system, we can let $\frac{dP_i(t)}{dt} = 0$, giving:

$$\mu P_{n+1} - (\lambda + \mu)P_n + \lambda P_{n-1} = 0$$

and

$$P_1 = \frac{\lambda}{\mu} P_0$$

so, as before, we find inductively that,

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

and, as $\sum_{n=0}^{\infty} P_n = 1$, we get (as before):

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

Finally, we can derive some other useful probabilities: Note: as λ and μ are constant, let $\rho = \frac{\lambda}{\mu}$, where ρ is called the traffic density, the ratio of arrival rate to service rate.

$$P(\text{Service Idle}) = P(0 \text{ customers in the system}) = P_0 = 1 - \rho$$

$$P(\text{Service Busy}) = 1 - P(\text{Service Idle}) = \rho$$

$$P(n \text{ or more customers in the system}) = \sum_{m=n}^{\infty} P_m = \sum_{m=n}^{\infty} \rho^m (1 - \rho)$$

$$= \sum_{k=0}^{\infty} \rho^{n+k} (1 - \rho) = \rho^n (1 - \rho) \sum_{k=0}^{\infty} \rho^k = \frac{\rho^n (1 - \rho)}{1 - \rho} = \rho^n$$

$$P(\text{Less than } n \text{ customers in the system}) = 1 - \rho^n$$

Example 4. *Suppose that I am running an ice cream van in a busy seaside resort on a hot summer's day. I will assume the queueing system to have infinite capacity (as the queue is formed in a large open area). Customers arrive at an average of one every 30 seconds ($\lambda = \frac{1}{30}$), and I only serve single cornets, which I have had much practise in producing, taking me only 20 seconds to produce one ($\mu = \frac{1}{20}$). Calculate the probabilities above.*

The values of λ and μ give $\rho = \frac{2}{3}$. This gives us:

$$P(\text{Service Idle}) = 1 - \frac{2}{3} = \frac{1}{3}$$

$$P(\text{Service Busy}) = \frac{2}{3}$$

$$P(n \text{ or more customers in the system}) = \left(\frac{2}{3}\right)^n$$

$$P(\text{Less than } n \text{ customers in the system}) = 1 - \left(\frac{2}{3}\right)^n$$

6 Appendices

6.1 Appendix 1

Theorem 6.1. *Suppose we have two events, Event 1 and Event 2, occurring at Poisson rates λ and μ respectively. Then together they occur at a Poisson rate $\lambda + \mu$.*

Proof. [3] Let T_i be the length of time until event i occurs. As the intermediate times between successive occurrences are distributed according to exponential random variables with means λ and μ respectively, we have that:

$$P(T_1 \leq t) = \int_0^t \lambda e^{-\lambda x} dx = \lambda \left[\frac{-e^{-\lambda x}}{\lambda} \right]_0^t = -[e^{-\lambda x}]_0^t = 1 - e^{-\lambda t}$$

and, similarly,

$$P(T_2 \leq t) = 1 - e^{-\mu t}$$

The time until *either* the first occasion when event 1 occurs *or* until the first occasion when event 2 occurs (whichever is sooner) is given by $\min(T_1, T_2) = T$

$$\Rightarrow P(T \leq t) = 1 - P(T > t) = 1 - P(\min(T_1, T_2) > t)$$

$$\min(T_1, T_2) > t \iff T_1, T_2 > t$$

$$\begin{aligned} \Rightarrow P(T \leq t) &= 1 - P(T_1, T_2 > t) = {}^{10}1 - P(T_1 > t)P(T_2 > t) \\ &= 1 - e^{-\lambda t}e^{-\mu t} = 1 - e^{-(\lambda+\mu)t} \end{aligned}$$

So T has an exponential distribution with mean $\lambda + \mu$, so events 1 and 2 together occur according to an exponential distribution with mean $\lambda + \mu$ \square

6.2 Appendix 2

Theorem 6.2.

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

Proof. [6] Consider:

$$y = 1 + x + x^2 + \dots = \sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

$$\Rightarrow \frac{dy}{dx} = 1 + 2x + 3x^2 + \dots = \frac{1}{(1-x)^2}$$

So

$$\begin{aligned} \sum_{n=0}^{\infty} nx^n &= x + 2x^2 + 3x^3 + \dots = x \frac{dy}{dx} = \frac{x}{(1-x)^2} \\ \Rightarrow \sum_{n=0}^{\infty} n \left(\frac{\lambda}{\mu} \right)^n &= \frac{\frac{\lambda}{\mu}}{\left(1 - \frac{\lambda}{\mu} \right)^2} = \frac{\frac{\lambda}{\mu}}{1 - \frac{2\lambda}{\mu} + \left(\frac{\lambda}{\mu} \right)^2} = \frac{\frac{\lambda\mu}{\mu^2}}{\frac{\mu^2 - 2\lambda\mu + \lambda^2}{\mu^2}} = \frac{\lambda\mu}{(\mu - \lambda)^2} \\ &\Rightarrow \left(1 - \frac{\lambda}{\mu} \right) \sum_{n=0}^{\infty} n \left(\frac{\lambda}{\mu} \right)^n = \frac{\lambda\mu(\mu - \lambda)}{\mu(\mu - \lambda)^2} = \frac{\lambda}{\mu - \lambda} \end{aligned}$$

\square

¹⁰As events 1 and 2 are independent

6.3 Appendix 3

Theorem 6.3.

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

Proof. [6] Consider

$$\begin{aligned} y &= 1 + x + x^2 + \dots + x^N = \sum_{n=0}^N x^n = \frac{x^{N+1} - 1}{x - 1} \\ \Rightarrow \frac{dy}{dx} &= \frac{(x-1)(N+1)x^N - x^{N+1} + 1}{(x-1)^2} \\ \Rightarrow x \frac{dy}{dx} &= \frac{(x-1)(N+1)x^{N+1} - x(x^{N+1} - 1)}{(x-1)^2} \end{aligned}$$

and

$$\begin{aligned} \frac{dy}{dx} &= 1 + 2x + 3x^2 + \dots + Nx^{N-1} \Rightarrow x \frac{dy}{dx} = x + 2x^2 + \dots + Nx^N = \sum_{n=0}^N nx^n \\ &\Rightarrow \sum_{n=0}^N nx^n = \frac{(x-1)(N+1)x^{N+1} - x(x^{N+1} - 1)}{(x-1)^2} \\ &= \frac{(N+1)x^{N+2} - (N+1)x^{N+1} - x^{N+2} + x}{(x-1)^2} = \frac{Nx^{N+2} - (N+1)x^{N+1} + x}{(x-1)^2} \\ &\Rightarrow \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu} \right)^{N+1}} \sum_{n=0}^N n \left(\frac{\lambda}{\mu} \right)^n = \frac{\left(1 - \frac{\lambda}{\mu} \right) \left(N \left(\frac{\lambda}{\mu} \right)^{N+2} - (N+1) \left(\frac{\lambda}{\mu} \right)^{N+1} + \frac{\lambda}{\mu} \right)}{\left(1 - \left(\frac{\lambda}{\mu} \right)^{N+1} \right) \left(\left(\frac{\lambda}{\mu} \right)^2 - \frac{2\lambda}{\mu} + 1 \right)} \\ &= \frac{N \left(\frac{\lambda}{\mu} \right)^{N+2} - (N+1) \left(\frac{\lambda}{\mu} \right)^{N+1} + \frac{\lambda}{\mu} - N \left(\frac{\lambda}{\mu} \right)^{N+3} + (N+1) \left(\frac{\lambda}{\mu} \right)^{N+2} - \left(\frac{\lambda}{\mu} \right)^2}{\left(1 - \left(\frac{\lambda}{\mu} \right)^{N+1} \right) \left(\left(\frac{\lambda}{\mu} \right)^2 - \frac{2\lambda}{\mu} + 1 \right)} \\ &= \frac{(2N+1) \left(\frac{\lambda}{\mu} \right)^{N+2} - (N+1) \left(\frac{\lambda}{\mu} \right)^{N+1} + \frac{\lambda}{\mu} - N \left(\frac{\lambda}{\mu} \right)^{N+3} - \left(\frac{\lambda}{\mu} \right)^2}{\left(\frac{\lambda}{\mu} \right)^2 - \frac{2\lambda}{\mu} + 1 - \left(\frac{\lambda}{\mu} \right)^{N+3} + 2 \left(\frac{\lambda}{\mu} \right)^{N+2} - \left(\frac{\lambda}{\mu} \right)^{N+1}} \\ &= \frac{\frac{\lambda}{\mu} \left[(2N+1) \left(\frac{\lambda}{\mu} \right)^{N+1} - (N+1) \left(\frac{\lambda}{\mu} \right)^N + 1 - N \left(\frac{\lambda}{\mu} \right)^{N+2} - \frac{\lambda}{\mu} \right]}{\left(1 - \frac{\lambda}{\mu} \right) \left(1 - \frac{\lambda}{\mu} - \left(\frac{\lambda}{\mu} \right)^{N+1} + \left(\frac{\lambda}{\mu} \right)^{N+2} \right)} \end{aligned}$$

$$\begin{aligned}
&= \frac{\lambda \left[(2N+1) \left(\frac{\lambda}{\mu}\right)^{N+1} - (N+1) \left(\frac{\lambda}{\mu}\right)^N + 1 - N \left(\frac{\lambda}{\mu}\right)^{N+2} - \frac{\lambda}{\mu} \right]}{(\mu - \lambda) \left(1 - \frac{\lambda}{\mu} - \left(\frac{\lambda}{\mu}\right)^{N+1} + \left(\frac{\lambda}{\mu}\right)^{N+2} \right)} \\
&= \frac{\lambda \left(1 - \frac{\lambda}{\mu} \right) \left(1 - (N+1) \left(\frac{\lambda}{\mu}\right)^N + N \left(\frac{\lambda}{\mu}\right)^{N+1} \right)}{(\mu - \lambda) \left(1 - \frac{\lambda}{\mu} \right) \left(1 - \left(\frac{\lambda}{\mu}\right)^{N+1} \right)} \\
&= \frac{\lambda \left[1 - (N+1) \left(\frac{\lambda}{\mu}\right)^N + N \left(\frac{\lambda}{\mu}\right)^{N+1} \right]}{(\mu - \lambda) \left(1 - \left(\frac{\lambda}{\mu}\right)^{N+1} \right)}
\end{aligned}$$

□

6.4 Appendix 4

$$\frac{dp_0(t)}{dt} + \lambda p_0(t) = 0$$

Solve with integrating factors:

$$\begin{aligned}
P(t) &= \exp\left(\int \lambda dt\right) = e^{\lambda t} \\
\Rightarrow \frac{d}{dt}(e^{\lambda t} p_0(t)) &= 0 \\
\Rightarrow e^{\lambda t} p_0(t) &= c \\
\Rightarrow p_0(t) &= e^{-\lambda t} \text{ is a solution}
\end{aligned}$$

6.5 Appendix 5

$$\frac{dp_1(t)}{dt} + \lambda p_1(t) = \lambda e^{-\lambda t}$$

This can be solved using the integrating factor $P(t) = \exp(\int \lambda dt) = e^{\lambda t}$.

$$\begin{aligned}
\Rightarrow \frac{d}{dt}(e^{\lambda t} p_1(t)) &= e^{\lambda t} \lambda e^{-\lambda t} = \lambda \\
\Rightarrow e^{\lambda t} p_1(t) &= \lambda t + c
\end{aligned}$$

where $c \in \mathbb{R}$, so

$$p_1(t) = \lambda t e^{-\lambda t} \text{ is a solution}$$

6.6 Appendix 6

References

- [1] Prabhu, N.U.: *Foundations of Queueing Theory*, Nuwer Academic, Boston, London (1997)
- [2] <http://staff.um.edu.mt/jskl1/simweb/mm1.htm>
- [3] Ross, S.: *Introduction to Probability Models*, Eighth Edition, Academic Press Inc., U.S. (2003)
- [4] http://www.rail-reg.gov.uk/upload/xls/stat_usage.xls
- [5] Jones, P.W, Smith, P.: *Stochastic Processes: An Introduction*, Arnold Texts in Statistics, London (2001)
- [6] <http://www.nrich.maths.org/>