Research Topic 1:
Queueing Networks

L. M. Howell

Overview

Queueing is a concept many of us are familiar with. It can be the cause of much tension in our day to day lives and provides several challenges when it comes to mathematical modelling.

A general way of describing a queue is as follows: a customer arrives seeking a service and, if not served immediately, has to wait for a while, potentially behind some others also waiting for the service. Once they reach the front of the queue and the server is free, they are served and then leave the system. This means every queue has some arrival process, a distribution of service times and a corresponding wait time. Whilst we often think of queues in a human sense, this setup applies to many situations. For example, virtual queues are often used in online settings when a product is in high demand such as buying concert tickets on release day. It also describes products waiting to be packaged before being sent out (in this situation, the 'customer' is the product itself) or a train waiting to be allowed to leave a platform (in which the train is the 'customer').

One of the first hurdles for modelling queues is the randomness involved. Customers arrive at random, and the amount of time it takes to serve a customer is random. Most commonly the randomness of customer arrivals is modelled with a Poisson process, though this immediately presents complications: what if the average number of customers arriving per time unit changes? A Poisson process is also inappropriate if there is a reservation based system within the queue, such as at a restaurant. Further complications arise when various behavioural aspects are considered. What if a customer decides to leave the queue if they have been waiting a long time? Or not join the queue in the first place if the queue is too long?

Some of the more interesting applications come from queueing *networks* - nodes of interacting queues with some routeing process through the system. These models are often applied to telecommunications networks where a mobile phone call is associated with the mobile phone tower it is closest to. The nodes of the network are phone towers; calls originate associated with a specific tower, but as a person moves their call could change towers, and so they move through the network. In this situation the 'service time' of a customer is how long the mobile phone is closest to that tower; the customer is the call itself. This model uses infinite server queues. Whilst there is technically a limit to the number of calls a cell tower can handle, in practice this is rarely an issue and so the modelling can be made easier by assuming that there are infinite servers and thus no wait time in the queue (as the 'customer' is immediately 'served'). The interest is instead in the distribution of the service times and the routing through the network, as this can help predict demand. Infinite server networks have also seen applications within biology in the study of population processes.

1 Introduction

A queue is something very familiar to us, and a common problem in modern life. Modelling queues mathematically can provide excellent insight into not just ways of making queues shorter, but also accurately models situations beyond just that of a human waiting in line to pay for their shopping.

In this report we give a brief introduction to basic queueing theory, with some examples and comparisons. This is then followed by an extension into queueing networks.

Descriptions of what defines a queue varies in the literature (for example Gross et al. (2008, pg. 3) use six defining characteristics). We choose to define a queue with five main points.

Definition 1.1 *A queueing system has five characteristics;*

1. *Customers entering the system at some 'rate'.*
2. *The customers are served at some 'rate'.*
3. *Customer behaviour.*
4. *Server behaviour.*
5. *The capacity of the system.*

The first characteristic describes the way customers enter the system, often using a Poisson process but not always. This of course can be stationary or non-stationary. Characteristic two refers to the rate that customers are served; this often uses exponential service times. Customer behaviour refers to the discipline within the queue. This includes variations such as;

- Balking; where a customer arrives to join a queue, but upon seeing the length, chooses to leave instead,
- Reneging; where a customer chooses to leave a queue after deeming that they have waited too long,
- Jockeying; whether customers cut in line.

Server behaviour describes the way in which customers are served. The most basic and common approach is first-come-first-served (FCFS) but other options include last-come-first-served, random, and priority. The capacity is the number of customers allowed in the system at one time.

Notation 1.2 *A queue can be described using Kendall notation $A/B/X/Y/Z$ (Gross et al., 2008; Kendall, 1953), detailed in Table 1.*

Note that "General" means unspecified. Often when the system capacity is infinite or the server behaviour is FCFS, these are dropped from the notation. The specification for X being the number of *parallel* servers is also important; parallel means that the servers are independently serving customers at the same time, which is distinct from the situation in which there are multiple service stages, which introduces dependence.

Characteristic	Symbol	Meaning
A: Interarrival dist.	$M/D/G$	Exponential/ Deterministic/ General
B: Service time dist.	$M/D/G$	Exponential/ Deterministic/ General
X: Number of servers (in parallel)	$1, 2, \ldots \infty$	
Y: System Capacity	$1, 2, \ldots \infty$	
Z: Server behaviour	FCFS/LCFS/ RSS/PR/G	First-Come-First-Served/ Last-Come-First-Served/ Random Service Selection/ Priority/General

Table 1: A table showing how the Kendall notation for queues works.

This provides a framework for *how* to study queues, but not *why*. There are three main statistics of interest; a measure of the amount of time a customer has to wait in the queue, a measure of the amount of time a server spends idle and a way of describing how customers accumulate (congestion). Often this involves looking at the steady-state of a queue, if it exists.

Example 1.3 (M/M/1 Queue) *An $M/M/1$ queue is a queue with exponential interrarrival times (equivalently this can be defined by saying customers arrive according to a Poisson process); exponential service times and 1 server. The server serves customers according to first come first serve and there is infinite capacity in the queue. This queue is equivalent to a birth-death process.*

Definition 1.4 (Traffic Intensity) *For a $G/G/c$ queue where the customers are arriving at rate λ and being served at rate μ then the traffic intensity is defined as $\rho = \frac{\lambda}{c\mu}$ (Gross et al., 2008, pg. 9).*

This is intuitive; the amount of traffic in the system is how many customers are entering the system against how many customers are leaving the system; with λ being the entrance rate and $c\mu$ being the exit rate. Therefore when $\rho > 1$, customers are entering the system faster than they are leaving, and the queue length increases.

1.1 Performance Measures

Denote the number of customers in the system at time t by the random variable $N(t)$. Therefore $N(t) = N_q(t) + N_s(t)$ with N_q denoting the number of customers in the queue and N_s being the number of customers being served. This means that in the steady state, $N = N_q + N_s$. Call $p_n(t)$ the probability that the number of customers in the system is n at time t, or $p_n(t) = P(N(t) = n)$. Similarly in the steady state call $p_n = P(N = n)$.

Consider a queue in steady state with c servers. Then define the average number of customers

in the system

$$L = \mathbb{E}(N) = \sum_{n=0}^{\infty} np_n \tag{1}$$

and the average number of of customers in the queue with

$$L_q = \mathbb{E}(N_q) = \sum_{n=c+1}^{\infty} (n-c)p_n.$$

Also use the random variable T to denote the amount of time a customer spends in the system, so $T = T_q + T_s$, time in the queue plus time being served. These are all random variables. Then define $W = \mathbb{E}(T), W_q = \mathbb{E}(T_q)$. This provides the necessary set up to talk about one of the most important results in queueing theory: Little's Law.

Theorem 1.5 (Little's Law) *Let L be the expected number of customers in a queueing system, W the expected amount of time spent in the system by a customer and λ the rate at which customers arrive. Then*

$$L = \lambda W \tag{2}$$

Furthermore,

$$L_q = \lambda W_q$$

(Little, 1961; Stewart, 2009).

This result is important as it shows that in the steady state of a queueing system, the average number of customers in the system is not influenced by details such as the arrival distribution or the service distribution.

Example 1.6 ($M/M/1$ cont.) *To calculate the average number of customers in the system, L, use Equation 1. However this requires knowing p_n so to calculate that, it is helpful to use the fact that an $M/M/1$ queue can also be modelled as a birth death process (See Stewart (2009, pg,402) for more). Therefore the probability of n customers being in the system can be calculated as*

$$p_n = p_0 \prod_{i=1}^{n} \frac{\lambda}{\mu} = p_0 \left(\frac{\lambda}{\mu}\right)^n.$$

Calculating p_0 can be done using the fact that the sum of all of the p_n's must be 1. Thus

$$1 = \sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} p_0 \left(\frac{\lambda}{\mu}\right)^n \implies p_0 = \frac{1}{\sum_{n=0}^{\infty}(\frac{\lambda}{\mu})^n} = \frac{1}{\frac{1}{1-\frac{\lambda}{\mu}}} = 1 - \frac{\lambda}{\mu}$$

as we can use the sum of a geometric series as the steady state is only valid if $\frac{\lambda}{\mu} < 1$. Now $p_n = (1 - \frac{\lambda}{\mu})\left(\frac{\lambda}{\mu}\right)^n = (1-\rho)\rho^n$. This can then be used to calculate L;

$$L = \sum_{n=0}^{\infty} np_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = (1-\rho)\rho \sum_{n=0}^{\infty} n\rho^{n-1}.$$

Now use the fact that $\rho < 1$ again with

$$\sum_{n=0}^{\infty} n\rho^{n-1} = \frac{d}{d\rho}\left(\sum_{n=0}^{\infty}\rho^n\right) = \frac{d}{d\rho}\frac{1}{1-\rho} = \frac{1}{(1-\rho)^2}$$

which altogether gives

$$L = (1-\rho)\rho\frac{1}{(1-\rho)^2} = \frac{\rho}{1-\rho}.$$

Now using Little's Law:

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

It's also possible to calculate the average number of customers waiting in the queue, L_q. As there is only one server, only one person in the system can be served at any time which gives;

$$L_q = \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty} np_n - \sum_{n=1}^{\infty} p_n = L - (1-p_0) = L - (1-(1-\rho)) = L - \rho.$$

2 Comparison of the $M/M/1$ and $M/M/c$ queues

Compare the following three scenarios; in the case where there are two servers, is it better to have a shared queue, or two individual queues for each server? How does this compare with a single queue with the server process going twice as fast?

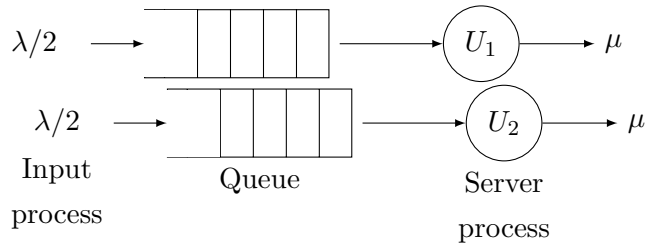

Figure 1: Two servers with separate queues.

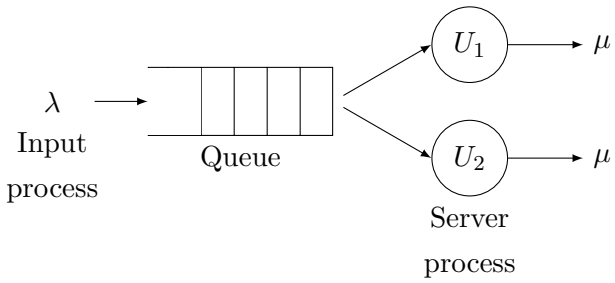

Figure 2: A shared queue with two servers.

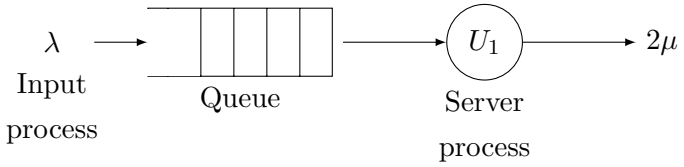

Figure 3: A single queue and single server working at twice the rate.

In maths terms, compare an $M/M/2$ queue with two independant $M/M/1$ queues and an $M/M/1$ queue with twice the rate. Through reasoning similar to that of Example 1.6 the following results are achieved. For an in depth look at the maths, see Stewart (2009, pg. 423).

	Two Indep. $M/M/1$		$M/M/2$		$M/M/1$ with 2μ
L	$\frac{\rho}{1-\rho}$	$\geq$	$\frac{\rho}{1-\rho} \cdot \frac{1}{1+\rho}$	$\geq$	$\frac{\rho}{1-\rho}$
W	$\frac{2}{2\mu-\lambda}$	$\geq$	$\frac{1}{\mu(1-\rho^2)}$	$\geq$	$\frac{1}{2\mu-\lambda}$

Table 2: Comparison of the average number of customers in a system and average amount of time spent in the system for the three different queue setups.

This shows that having a shared queue with two servers is more efficient that two separate queues.

3 Non-stationarity

The goal for a queueing system is to attempt to model the real world. In many situations, it is not appropriate to use a stationary Poisson process. In an example of a queue in a shop, it is expected that the rate of customers arriving throughout the day is not constant. For example, the lunchtime rush at a cafe. Thus the model of a queue can be expanded to use a Poisson process with rate λ_t. This still has it's limitations, and alternatively a Cox process can be used (Boxma et al., 2019), which is a non-stationary Poisson process where the rate λ_t is itself a non-stationary stochastic process.

4 Server Networks

A queueing network describes a system of nodes, where a node is a server or queue. This is combined with some routing method which dictates whether, after a customer is served, what node the customer is then transferred to or if the customer leaves the system. Arrivals into the system and service times are still modelled randomly. This model can mimic situations such as telecommunications networks.

Consider a series of J connected single server queues. Let the first queue have a stationary Poisson arrival process with rate λ. The service rate is exponential with rate μ. Restrict $\lambda < \mu$ so that a steady state exists. Call $N_j(t)$ the number of customers in the j^{th} queue at time t. On it's own, the first queue is simply $M/M/1$.

Since the number of customers in an $M/M/1$ queue is a reversible Markov process (Kelly, 1979), the rate at which customers leave the system is also a Poisson process. Furthermore, the joint distribution of the departure process up to a point in time t_0 and the number of customers in the queue at t_0 has the same distribution as the arrival process after time $-t_0$ and the number of customers in the queue at $-t_0$. It then follows that the departure process prior to t_0 is independant of the number of customers in the queue at t_0.

Due to the departure process of an $M/M/1$ queue being Poisson, this ensures that the arrival distribution for the second queue is also Poisson, and so on. Perpetuating this means that each

queue viewed on it's own is just an $M/M/1$ queue. Therefore the steady state distribution for the number of customers in queue j is $P(N_j = n_j) = (1-\rho)\rho^{n_j}$.

As stated above, $N_1(t_0)$ is independant of the departure process of the first queue prior to t_0. But the joint distribution of $(N_2(t_0), N_3(t_0), \ldots, N_J(t_0))$ has two deciding factors; the departure process of the first queue prior to t_0, and the service times of queues 2 through J. Therefore, $N_1(t_0)$ is independant of $(N_2(t_0), N_3(t_0), \ldots, N_J(t_0))$. Moreover, the same argument can be made for the j^{th} queue; $N_j(t_0)$ is independant of $(N_{j+1}(t_0), N_{j+2}(t_0), \ldots, N_J(t_0))$. As a result, all the $N_j(t_0)$'s must be independant of each other. So the joint distribution of $(N_1(t_0), N_2(t_0), N_3(t_0), \ldots, N_J(t_0))$ can be calculated using

$$\pi(N_1, N_2, N_3, \ldots, N_J) = \prod_{j=1}^{J} \left(1 - \frac{\lambda}{\mu_j}\right) \left(\frac{\lambda}{\mu_j}\right)^{n_j}$$

where π denotes the joint stationary distribution (Kelly, 1979).

4.1 Routes through Open Networks

Consider a network of J nodes, and that the routing method is some kind of deterministic function based upon some customer "type". So for J nodes there are I customer types. Thus a customer can arrive at a rate λ_i and follows a route $r(i,s)$ where s is the stage of the route $1, 2, \ldots, S(i)$. This set up allows for modelling the behaviour of a system where there is dependence between a customers route and the customers past. The system is called open as customers can enter and then leave the system once their route is complete.

To find the stationary distribution of a system like this requires some set up. First let $t_j(l)$ describe the type of the customer in position l of queue j (with $l = 1, 2, \ldots, n_j$ for n_j being the length of the queue). Then define $c_j(l) = (t_j(l), s_j(l))$ with s being defined in the previous paragraph. This $c_j(l)$ is the *class* of the customer. This sets up a vector $\mathbf{c}_j = (c_j(1), c_j(2), \ldots, c_j(n_j))$ which describes the state of queue j. Therefore overall $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_J)$ is a Markov process describing the system state.

Theorem 4.1 *The stationary distribution of the system described above is*

$$\pi(\mathbf{C}) = \prod_{j=1}^{J} \pi_j(\mathbf{c}_j)$$

where π_j is the stationary distribution of queue j.

To fully define π_j first define

$$a_j = \sum_{i=1}^{I} \sum_{s=1}^{S(i)} \lambda_i \mathbb{1}(r(i,s) = j)$$

which is the average number of customers arriving to the j^{th} queue per unit time when the system is in steady state. The indicator function restricts the sum to only including customers at the relevant queue j and the λ_i is the corresponding arrival rate. This is summed over customers at any point

in their route, as it doesn't matter *when* in their route a customer has arrived at queue j, only that they have arrived. It is then also summed over customer type to get the average arrival per unit time for the whole queue.

Now define

$$\frac{1}{b_j} = \sum_{n_j=1}^{\infty} \frac{a_j^{n_j}}{\prod_{l=1}^{n_j} \mu_j(l)}$$

with $\mu_j(l)$ being the service rate. This ensures that a steady state exists, as it guarantees that $\pi(\mathbf{C})$ sums to 1. Altogether this results in

$$\pi_j(\mathbf{c}_j) = b_j \prod_{l=1}^{n_j} \frac{\lambda_{t_j(l)} \mathbb{1}(r(t_j(l), s_j(l)) = j)}{\mu_j(l)}.$$

For a full proof see Kelly (1979, pg. 61). This set up can then adapted for the situation of a *closed* network in which a customer restarts their route when they are finished, so there is no entering or leaving the system. This kind of model is good for situations like machine repairs in a factory; where the service time is how long a machine takes to be repaired and the time spent in a queue is how long until the machine breaks (plus how long it has to wait for an engineer to be free). This topic of server networks is then furthered by Massey and Whitt (1993) which details a model of a network of infinite server queues with a non-stationary Poisson input. Further to this is the work of Fiems et al. (2018) which continues the infinite server network model where sudden changes within the network population can be analysed.

References

Boxma, O., Kella, O., and Mandjes, M. (2019). Infinite-server systems with coxian arrivals. *Queueing Systems*, 92:233–255.

Fiems, D., Mandjes, M., and Patch, B. (2018). Networks of infinite-server queues with multiplicative transitions. *Performance Evaluation*, 123:35–49.

Gross, D., Shortle, J. F., Thompson, J. M., and Harris, C. M. (2008). *Fundamentals of Queueing Theory*. John Wiley & Sons, Hoboken, New Jersey, 4th edition.

Kelly, F. P. (1979). *Reversibility and Stochastic Networks*. Wiley series in probability and mathematical statistics. Wiley, Chichester, New York.

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics*, 24(3):338 – 354.

Little, J. D. C. (1961). A proof for the queuing formula: $L = \lambda W$. *Operations Research*, 9(3):383–387.

Massey, W. A. and Whitt, W. (1993). Networks of infinite-server queues with nonstationary poisson input. *Queueing Systems*, 13:183–250.

Stewart, W. J. (2009). *Probability, Markov Chains, Queues, and Simulation: the Mathematical Basis of Performance Modeling*. Princeton University Press, Princeton, New Jersey.