CAMBRIDGE

Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Excerpt
More Information

Part I

The Single Queue

In the first part of this book we introduce the single queue, a system in which arriving customers require a single service from the system, and this service is provided by one or more servers. We study properties of some special queueing systems that are amenable to exact analysis.

In Chapter 1 we define the single queue, introduce notations and some relations and properties, and present the most tractable examples of queues, so-called birth and death queues . We also discuss simulation of queues.

In Chapter 2 we study a queueing system with memoryless Poisson arrivals and generally distributed processing times, the so-called M/G/1 system. Performance measures of this system can be derived exactly, using the principle of work conservation, and the property of PASTA (Poisson arrivals see time averages).

Chapter 3 considers the scheduling of batches of jobs, and of stationary streams of jobs. We discuss priority queues and other service policies.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Excerpt
More Information

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Excerpt
More Information

1

Queues and Their Simulations, Birth and Death Queues

In this chapter we start our exploration of queues. We describe a single queue consisting of a single stream of customers where each of them requires a single service operation and this service is provided by a single service station, manned by one or more servers, operating under some service policy. We introduce notation to describe such queues and derive two basic relationships: the first presents the queue length as arrivals minus departures, the second, Lindley's equation, provides a recursive calculation of waiting times of successive customers. We discuss simulation of queues using the Lindley's equation recursion. Next we study examples of single queues with Poisson arrivals and exponential service times, which are modeled by Markovian birth and death processes, and derive the stationary distribution of the queue length, using detailed balance equations.

1.1 The Single Queue

A queueing system consists of two parts: on the demand side there are the streams of customers, each with its service requests; on the service side there are one or more service stations, with one or more servers in each. We start our exploration of queueing theory by considering a single stream of customers, each requiring a single service operation, and a single service station that provides the service. We refer to the sequence of arrivals and services as the primitives of the system. We model the customer arrivals by a stochastic point process $\mathcal{A}(t)$, $t > 0$, which counts the number of arrivals in the time interval $(0, t]$. We let A_n, $n = 1, 2, \ldots$ be the arrival times of the customers, and $T_1 = A_1$, $T_n = A_n - A_{n-1}$, $n = 2, 3, \ldots$ be the interarrival times. We have that $\mathcal{A}(t) = \max\{n : A_n \leq t\}$. We will frequently assume that T_n are independent identically distributed random variables with distribution F and finite expectation $\mathbb{E}(T_1) = 1/\lambda$, so that $\mathcal{A}(t)$ is a renewal process, with arrival rate λ. In particular, if interarrivals

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Excerpt
More Information

are exponentially distributed, then $\mathcal{A}(t)$ is a Poisson process and we say that arrivals are Poisson with rate λ.

Customer n requires service for a duration X_n, $n = 1, 2, \ldots$. We will always assume that the sequence of service durations is independent of the arrival times, and service durations are independent identically distributed with distribution G and finite expectation $m = 1/\mu$.

The customers are served by a single service station, which may have one or more servers that provide the service. Typically, an arriving customer will join a queue and wait, and will then move to a server and be served.

A common notation introduced by D.G. Kendall describes the single queue by a three-field mnemonic: the first describes the arrival process, the second the service distribution, and the third the service station. Thus M/M/1 denotes a queue with Poisson arrivals, exponential service times, and a single server, where M stands for memoryless. D/G/s denotes a queue with deterministic arrivals, generally distributed independent service times, and s servers. G/G/∞ is a queue with independent, generally distributed interarrival times, general independent service times, and an infinite number of servers, which means that arriving customers start service immediately and there is no waiting. G/$\cdot$/$\cdot$ will denote a queue that has a general stationary sequence of interarrival times. If the system can only contain a limited number of customers at any time, this limit is sometimes added as a fourth field. Thus, M/M/K/K is a queueing system with Poisson arrivals, exponential service times, K servers, and a total space for K customers. This system is the famous Erlang loss system, which Erlang has used to model a telephone exchange with K lines. In this system, when all the lines are busy, arriving customers are lost.

The interaction between the customers described by $\mathcal{A}(t)$, $t > 0$ and X_n, $n = 1, 2, \ldots$ on the one hand and the service station on the other hand creates waiting and queues. To describe this interaction we need to specify also the service policy. We list a few commonly used service policies: FCFS – first come first served (also known as FIFO – first in first out) in which customers enter service in order of arrival; LCFS – last come first served (sometimes called LIFO – last in first out) in which whenever a customer arrives it enters service immediately, sometimes preempting the service of an earlier customer; PS – processor sharing, the station divides its service capacity equally between all the customers in the system.

The following two very simple relationships are the basis of much of queueing theory. The first describes the dynamics of customers. We denote the queue length by $Q(t)$, which is the number of customers in the system at time t, including both those waiting for service and those being served. We

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Excerpt
More Information

also denote by $\mathcal{D}(t)$ the number of customers that have left the system in the time interval $(0, t]$, which we call the departure process. Then we have the obvious relation

$$\mathcal{Q}(t) = \mathcal{Q}(0) + \mathcal{A}(t) - \mathcal{D}(t), \tag{1.1}$$

i.e. what is in the system at time t is what was there initially at time 0, plus all arrivals, minus all departures. The queueing dynamics are illustrated in Figure 1.1.

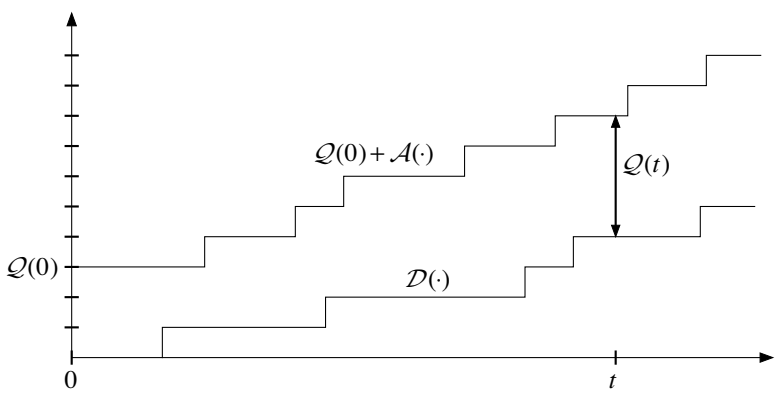

Figure 1.1 Queue length is arrivals minus departures.

The second relation calculates the waiting time of a customer under FCFS, and is known as Lindley's equation. We denote by V_n the waiting time of the nth arriving customer, from his arrival time to the start of service. For a single server operating under FCFS, we then have:

$$V_{n+1} = (V_n + X_n - T_{n+1})^+, \tag{1.2}$$

where $(x)^+ = \max(0, x)$ is the positive part of x. We explain this relation: Customer n departs the system $V_n + X_n$ time units after his arrival, while customer $n + 1$ arrives T_{n+1} time units after the arrival of customer n. If T_{n+1} exceeds $V_n + X_n$, then customer $n + 1$ will enter service immediately and not wait. If T_{n+1} is less than $V_n + X_n$, then customer $n + 1$ will wait $V_n + X_n - T_{n+1}$. This proves (1.2). Lindley's equation is illustrated in Figure 1.2.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Excerpt
More Information

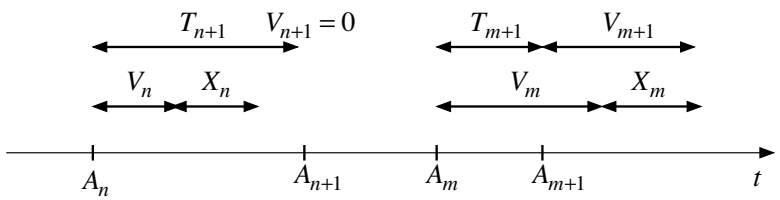

Figure 1.2 Waiting time calculation using Lindley's equation.

1.2 Simulation of a Queue

Simulation is a powerful tool for studying queueing systems. The detailed analysis of many queueing systems is intractable, but various performance measures of the queues can be estimated by simulation. There is a rich theory of how to use simulation, which we will not cover in this text; however, we suggest some books in the sources in Section 1.4. Here we indicate only the elementary method that can be used for simple explorations by the reader. Thinking of simulation is also a way of getting a different view from what can be obtained from theorems and equations.

Recursive relations such as Lindley's equation can be used to simulate the queues. In the case of a single queue with a single server, operating under FCFS, the simulation will work as follows: Initialize the system with the time at which the server will be available after serving all the customers present at time zero. Thereafter, generate successive interarrival and service times for successive customers, and use these to obtain arrival time, service start time, and departure time of successive customers. From this, obtain the waiting time and sojourn time (we define sojourn time as waiting plus service, or more generally, time from arrival to departure) of each customer. Furthermore, by counting arrivals minus departures, the queue lengths at any time can be obtained.

Example 1.1 The following tables illustrate part of a simulation of a queue. In the first table we start from the 17th customer who arrives at time 39.1, and the server is available to start serving this customer at time 42.8. We denote by S_n the start of service of customer n, by D_n his departure, by W_n his sojourn time, and by V_n his waiting time before being served. The simulation then proceeds as follows. Successive interarrivals T_n and service requirements X_n are generated pseudorandomly from the interarrival and service time distributions, and we then calculate recursively: $A_n = A_{n-1} + T_n$, $S_n = \max(A_n, D_{n-1})$, $V_n = S_n - A_n$, $D_n = S_n + X_n$, $W_n = V_n + X_n$. The round-off numbers are:

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Excerpt
More Information

Customer	T_n	A_n	X_n	Start	Depart	Sojourn	Wait
17		39.1	2.2	42.8	45.1	6.0	3.8
18	2.8	41.9	2.7	45.1	47.7	5.9	3.2
19	4.3	46.1	1.0	47.7	48.7	2.6	1.6
20	2.5	48.6	0.3	48.7	49.0	0.4	0.1
21	4.1	52.8	1.5	52.8	54.2	1.5	0.0
22	4.8	57.6	2.0	57.6	59.5	2.0	0.0
23	1.3	58.9	3.9	59.5	63.5	4.6	0.6

The second table calculates the queue length. Order all the arrival and departure times, attaching 1 to each arrival and -1 to each departure. The queue length is then obtained, by adding the initial queue length and all the $+1$'s and -1's up to each time t. The table lists times of arrival and departure, the identity of the customer that arrives or departs (with positive sign for arrival and negative sign for departure), and the queue length at the time of this event. The queue lengths are plotted in Figure 1.3.

Time	39.1	41.9	42.1	42.8	45.1	46.1	47.7	48.6	48.7
Customer	17	18	-15	-16	-17	19	-18	20	-19
Queue	3	4	3	2	1	2	1	2	1

Time	49.0	52.8	54.2	57.6	58.9	59.5	59.8	61.6	63.5
Customerr	-20	21	-21	22	23	-22	24	25	-23
Queue	0	1	0	1	2	1	2	3	2

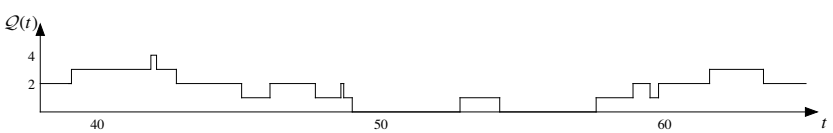

Figure 1.3 Simulation of queue length.

1.3 Birth and Death Queues

We now consider queues with Poisson arrivals and exponential service times. For such a queue, at any time t, the remaining time to the next arrival is exponentially distributed, as are also the remaining processing times of all the customers present in the system, either waiting or currently in process. All these times are independent of anything that happened prior to t. As a

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Excerpt
More Information

result, the queue length process $Q(t)$ is a continuous time Markov chain, with states $m = 0, 1, 2, \ldots$. Furthermore, when there are m customers in the system, $Q(t) = m$, the next event to happen will, with probability 1, be either a single arrival or a single service completion and departure, so the state will change by ± 1. Such queues are called birth and death queues. They are fully described by the transition rates

$$q(m, m+1) = \lambda_m, \qquad q(m, m-1) = \mu_m. \tag{1.3}$$

Here λ_m, the birth rate, is the rate at which arrivals occur when there are m customers in the system, and μ_m, the death rate, is the rate at which departures occur when there are m customers in the system. Note, for Poisson arrivals $\lambda_n = \lambda$, but letting the birth rate depend on the state allows for more general models, some of which we will encounter soon.

A major descriptor of the queueing process $Q(t)$ is its stationary distribution,

$$\pi(m) = \lim_{t\to\infty} \mathbb{P}(Q(t) = m), \quad m \geq 0, \tag{1.4}$$

provided these limits exist. This is sometimes called the limiting or long-run distribution, since it describes the state of the process after a time at which it no longer depends on the initial state. We will define stability of the queueing system if such a stationary distribution exists.

Continuous time birth and death processes are time reversible, and their stationary probabilities, $\pi(m) = \lim_{t\to\infty} \mathbb{P}(Q(t) = m)$, satisfy the detailed balance equations:

$$\pi(m)q(m, m+1) = \pi(m+1)q(m+1, m), \text{ i.e. } \pi(m+1) = \pi(m)\frac{\lambda_m}{\mu_{m+1}}. \tag{1.5}$$

We will discuss reversibility and balance equations in greater detail in Section 8.3. To interpret (1.5), note that it equates the rate at which transitions from m to $m+1$ occur, to the rate at which transitions back from $m+1$ to m occur. These rates are sometimes referred to as flux, borrowing a term from electricity networks. The detailed balance equations say that at stationarity these must be equal. From the detailed balance equations (1.5) we obtain the stationary distribution of a general birth and death queue:

$$\pi(m) = \pi(0)\frac{\lambda_0\lambda_1\cdots\lambda_{m-1}}{\mu_1\mu_2\cdots\mu_m}, \tag{1.6}$$

where $\pi(0)$ is obtained as the normalizing constant. The necessary and sufficient condition for ergodicity (irreducibility and positive recurrence of the Markov chain, see later Definition 2.4) is that the normalizing constant

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Excerpt
More Information

is > 0, i.e. that the sum of the terms on the r.h.s. of (1.6) converges. In that case, we say that the queue is stable.

We now describe several important models of birth and death queues. We use (1.6), to derive their stationary distributions. It is customary to denote by ρ the offered load of the system, which is the average amount of work that arrives at the service station per unit of time: if arrivals are Poisson with rate λ and service is exponential with rate μ then average service time is $m = 1/\mu$, and the average amount of work arriving per unit time is $\rho = \lambda/\mu$.

Example 1.2 (The M/M/1 queue) The M/M/1 queue is the simplest queueing model, for which almost every property or performance measure can be expressed by a closed form formula. Arrivals are Poisson at rate λ, service is exponential at rate μ, and there is a single server. Figure 1.4 illustrates the states and transition rates of $Q(t)$.

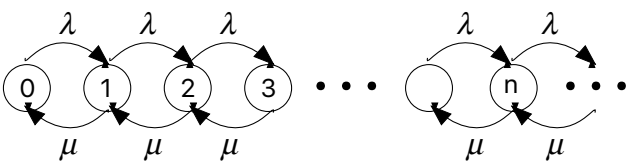

Figure 1.4 The M/M/1 queue, states and transition rates.

From (1.6) we have immediately:

Stationary distribution of the M/M/1 queue:
$$\pi(n) = (1-\rho)\rho^n, \quad n = 0, 1, 2, \ldots, \quad \rho < 1. \tag{1.7}$$

The queue is stable if and only if $\rho < 1$, and the stationary distribution of the queue length is geometric with parameter $1-\rho$ (denoted $\sim \text{Geom}_0(1-\rho)$), and mean $\frac{\rho}{1-\rho}$.

Theorem 1.3 *The sojourn time of a customer in the stationary M/M/1 queue under FCFS is exponentially distributed with rate $\mu - \lambda$, $W_n \sim Exp(\mu - \lambda)$, with mean $\frac{1}{\mu-\lambda}$.*

Proof If a customer arrives and there are j customers in the queue then his waiting time will be the sum of j i.i.d. exponential rate μ random variables, and his sojourn time will be the sum of $j+1$; this has an Erlang distribution. The probability that there are j customers in the queue is $(1-\rho)\rho^j$. So the pdf of his sojourn time $f_W(t)$ is

$$f_W(t) = \sum_{j=0}^{\infty}(1-\rho)\rho^j \frac{\mu^{j+1}t^j}{j!}e^{-\mu t} = \mu(1-\rho)e^{-\mu(1-\rho)t},$$

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Excerpt
More Information

where $\frac{\mu^{j+1}t^j}{j!}e^{-\mu t}$ is the density of the Erlang distribution with parameters $j+1$ and μ, denoted ~Erlang$(j+1,\mu)$. Hence, W_n is distributed exponentially with parameter $\mu(1-\rho) = \mu - \lambda$. Here we assume that the number of customers in the system at the time of an arrival is distributed as the stationary distribution of $Q(t)$. We justify this assumption, by proving that Poisson arrivals see time averages (PASTA) in Section 2.6. □

Remark (Resource pooling) There is an important lesson to be learned here: We note that the queue length, described by (1.7), depends only on $\rho = \lambda/\mu$, and does not depend directly on λ or μ. If we speed up the server, and speed up the arrival rate, say by a factor s, the number of customers in the system will remain the same. However, the expected sojourn time will decrease by a factor of s: $\frac{1}{s\mu - s\lambda}$. In other words, suppose we had s single-server M/M/1 queues to process s streams of customers, and were able to use instead an s time faster service rate and pool them all into one queue. In that case we would see the same length of queue at the pooled single queue as we saw in each of the s queues, but customers would move at a speed increased by a factor of s. This is the phenomena of *resource pooling*. We will encounter it later throughout the text.

Example 1.4 (The M/M/∞ queue) Arrivals are Poisson at rate λ, service time is exponential with rate μ, and there is an unlimited number of servers, so that customers enter service immediately on arrival and there is no waiting. The queue length process $Q(t)$ is now the number of customers in service, which is also the number of busy servers. Figure 1.5 illustrates the states and transition rates of $Q(t)$.

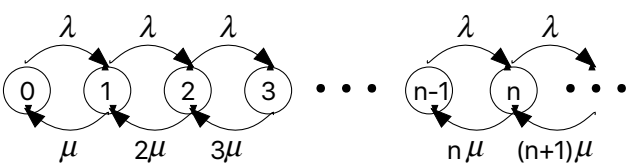

Figure 1.5 The M/M/∞ queue, states and transition rates.

The M/M/∞ queue is always stable, and the stationary distribution of the queue length is, by (1.6), Poisson with parameter ρ, with mean and variance ρ:

Stationary distribution of the M/M/∞ queue:

$$\pi(n) = \frac{\rho^n}{n!}e^{-\rho}, \quad n = 0, 1, 2, \ldots . \tag{1.8}$$

© in this web service Cambridge University Press www.cambridge.org