Difference and differential equations in Stochastic Operations Research

Ivo Adan and Jan van der Wal

Department of Mathematics and Computing Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Email: iadan@win.tue.nl, wscojvdw@win.tue.nl

November 12, 1998

Contents

Preface

This text is the result of continuing reseach on difference and differential equations for Operations Research Models in the Stochastic Operations Research Group in the Department of Mathematics and Computing Science at the Eindhoven University of Technology.

This first draft is prepared especially for the one week summercourse held in the week August 19 - 23, 1996, in Jyväskylä in Finland. We thank the organizer Prof. Timo Tiihonen for providing us with the pressure we needed to organize our thoughts and to prepare this text.

Chapter 1

Introduction

In this course we want to show that many problems in Stochastic Operations Research can be modelled and analysed by difference or differential equations. Most of the problems studied in this course are inspired by problems in the field of production, communication and inventory systems.

Our aim is not to give an extensive treatment of the analysis of difference and differential equations. We focus on the translation from problem to model and will try to show how the solution of the model can be used to answer questions posed in the original problem.

The models involve discrete Markovian models and fluid flow models (or a mixture of both). The first models usually give rise to recurrence relations of the form

$$x_n = p_{-l}x_{n-l} + \ldots + p_{-1}x_{n-1} + p_0x_n + p_1x_{n+1} + \ldots + p_kx_{n+k} \tag{1.1}$$

and for the other models we get a finite set of differential equations of the form

$$f_i'(x) = \sum_{j=1}^{N} q_{ij}f_j(x), \qquad i = 1, \ldots, N. \tag{1.2}$$

Exact as well as approximate techniques to solve these equations will be presented.

In the next chapter we will present some simple examples in which we encounter equations of the form (1.1) and (1.2). In chapter 3 we will refresh some well known results for analysis of the equations (1.1) and (1.2). Chapter 4 discusses a number of fairly basic concepts and results from probability that we will use. Chapters 5 and 6 discuss the $M/M/1$ queue, a number of variants and some applications. Chapter 7 deals with queueing problems with non-exponential interarrival or service times. Chapter 8 considers multi server systems. Chapter 9 treats a model in which production to order and production to stock are combined. In chapter 10 the performances of several rules to assign jobs to parallel machines are evaluated and compared. Chapter 10 is devoted to a comparison of the performance of several rules the analysis of the shortest queue problem. Finally chapter 11 considers the fluid models for production lines.

The text contains a lot of exercises.

Chapter 2

Examples

In this chapter we present some simple problems and show how they can be modelled by recurrence relations.

2.1 The Fibonacci series

One of the most famous and oldest examples of a recurrence relation was formulated by Leonardo Pisano in his book 'Liber Abaci'. He formulates a model to describe the offspring of one pair of rabbits. In this model he assumes that each pair of rabbits, from the second month they were born, produces each month one pair of young rabbits. Let us denote the number of pairs in month n by x_n. We start with one pair, so $x_0 = x_1 = 1$. Assuming that no rabbits die, the number of pairs in month $n + 2$ is the sum of the number of pairs in month $n+1$ and the number of pairs born in that month, which is equal to the number of pairs in month n. So we obtain the recurrence relation

$$x_{n+2} = x_{n+1} + x_n, \qquad n = 0, 1, 2, \ldots \tag{2.1}$$

With $x_0 = x_1 = 1$ we get for $x_0, x_1, x_2, \ldots$ the following, so-called Fibonacci sequence

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, \ldots$$

The quotients x_{n+1}/x_n seem to converge to $1.618\ldots$ (this number is also known as the golden mean, which is the desired quotient between the width and height of e.g. paintings). Does it indeed converge? To what? How can we prove that?

2.2 Markov chain

Let us consider a discrete time Markov chain with N states and $N \times N$ transition probability matrix P. The distribution at time 0 is given by x_0. Then the distribution x_n at time n is given by the recurrence relation

$$x_{n+1} = x_n P, \qquad n = 0, 1, 2, \ldots$$

Hence

$$x_n = x_0 P^n, \qquad n = 0, 1, 2, \ldots \tag{2.2}$$

If the Markov chain is irreducible and aperiodic, then x_n converges to the limiting and equilibrium distribution π.

Let us assume that the probability transition matrix P has N distinct, and hence single eigenvalues, denoted by $\lambda_1, \lambda_2, \ldots, \lambda_N$ with corresponding row eigenvectors $v_1, v_2, \ldots, v_N$. It is well known and easy to verify that all eigenvalues of a stochastic matrix are less than or equal to 1 in absolute value. If the matrix is irreducible and aperiodic, then there is one eigenvalue equal to 1 and all other eigenvalues are less than 1 in absolute value. Let us assume that $\lambda_1 = 1$.

We can write x_0 as a linear combination of the eigenvectors v_i as follows (remark that the eigenvectors $v_1, \ldots, v_N$ are linearly independent, and thus form a basis)

$$x_0 = \sum_{i=1}^{N} c_i v_i. \tag{2.3}$$

Then we get from (2.2) and (2.3) that

$$x_n = \sum_{i=1}^{N} c_i v_i P^n = \sum_{i=1}^{N} c_i \lambda_i^n v_i. \tag{2.4}$$

Recall that $|\lambda_i| < 1$ for all $i > 1$, so x_n converges to $c_1 v_1$ which apparently equals the equilibrium distribution π. Also the difference between x_n and π decreases exponentially fast.

Example 2.2.1 Consider the 2×2 matrix P,

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}.$$

The matrix P has eigenvalues 1 and $1-a-b$ with corresponding row eigenvectors (b, a) and $(1, -1)$, resp. Set $x_0 = (1, 0)$. We can write x_0 as linear combination of the two eigenvectors,

$$x = \frac{1}{a+b}(b, a) + \frac{a}{a+b}(1, -1).$$

From (2.4) we obtain that

$$x_n = \frac{1}{a+b}(b, a) + (1-a-b)^n \frac{a}{a+b}(1, -1).$$

As we see, there are two special cases, $a = b = 1$ and $a = b = 0$. In all other cases the second eigenvalue $1-a-b$ is less than 1 in absolute value. The case $a = b = 1$ corresponds to a periodic chain (switching between state 1 and state 2). The case $a = b = 0$ gives a chain with two absorbing states, and hence, which is not irreducible.

We may conclude that in all irreducible and aperiodic Markov chains the distribution converges exponentially fast to the equilibrium distribution. The rate of convergence is the so-called subdominant eigenvalue, i.e., the second largest eigenvalue in absolute value.

Remark 2.2.2 A complication arises is there are multiple eigenvalues. Then we may have to take into account generalized eigenvectors as well. But the results are very similar.

2.3 Roulette

In the previous examples we had to find x_n for infinitely many values of n. Sometimes we are only interested in finitely many x_n. An example in which only finitely many x_n are relevant will be presented below.

Let us consider someone playing roulette in a casino. The player is very cautious and only plays pair and impair. He plays a dollar at a time. The probability p that he wins is equal to 18/37. The probability q that he looses is 19/37. So with probability p he wins a dollar, with probability q he looses a dollar. The player starts with N dollar and stops as soon as he has $2N$ dollar, or more likely, he has lost his N dollar.

Let us denote the probability that he can go home with $2N$ dollar if now he has n dollar by x_n. Then we have the following relation between the probabilities x_n

$$x_n = px_{n+1} + qx_{n-1}, \tag{2.5}$$

with boundary conditions $x_0 = 0$ and $x_{2N} = 1$. Similar relations can be derived in case the player plays with more than one dollar at a time.

In table 2.1 we list the probability of leaving with $2N$ dollar for $N = 64$ dollar for a player playing 1, 2, 4, 8, 16, 32 and 64 dollar at a time. Apparently the cautious player has little chance of winning.

Stake	Chance of winning
1	0.0305
2	0.1506
4	0.2963
8	0.3935
16	0.4461
32	0.4730
64	0.4865

Table 2.1: Chance of winning for a player starting with $N = 64$ dollar

2.4 Loan

You are going to buy a car, but you do not have the money. So you need a loan for $K = 10000$ dollars. You pay $a = 350$ dollar a month during $N = 36$ months. What is the interest rate?

Let us denote the interest per month by r. Further define x_n to be the remaining debt after n months. Then we get the following recurrence relation

$$x_{n+1} = (1 + r)x_n - a, \qquad n = 0, 1, ..., N - 1,$$

with boundary values $x_0 = K$ and $x_N = 0$. So given the values of K, a and N we should be able to compute the interest rate r.

2.5 Transmission channel

Consider a channel for the transmission of data packets. The transfers are governed by a discrete time clock. At the beginning of each clock cycle, or time slot, a batch of packets arrives. The number of packets in a batch is k with probability p_k, $k = 0, 1, \ldots, K$. Packets that arrived at the beginning of a clock cycle are immediately candidates for transmission. In each time slot exactly one packet can be transmitted (if there is any). Packets which cannot be transferred immediately are put in a buffer where they wait for transmission. What is the distribution of the number of packets in the buffer at the beginning of a time slot, and what is the delay of the packets at this channel?

This system can be described by a Markov chain with states n, $n = 0, 1, 2, \ldots$ where n denotes the number of packets in the buffer at the beginning of a time slot. The transition probabilities $p_{n,m}$ satisfy $p_{n,m} = p_{m-n+1}$, the only exception here is $p_{0,0}$ which is given by $p_{0,0} = p_0 + p_1$. Let $x_n(t)$ denote the probability that there are n packets in the buffer at the beginning of time slot t, $t = 0, 1, 2, \ldots$ The probabilities $x_n(t)$ satisfy the recurrence relation

$$x_n(t + 1) = \sum_{k=0}^{K} x_{n+1-k}(t)p_k,$$

where by convention $x_n(t) = 0$ for $n < 0$. If we let t tend to infinity then we get for the limiting probabilities x_n the relation

$$x_n = \sum_{k=0}^{K} x_{n+1-k}p_k. \tag{2.6}$$

Above we assumed that the size of the buffer is infinite. In fact, it is finite. Suppose that the buffer offers room for at most B data packets. Packets arriving when the buffer is full are lost. Hence $x_n = 0$ for $n > B$. What should be the size of the buffer such that no more than a given fraction of the packets will be lost?

2.6 Unreliable machine

Consider a machine which is alternatingly working and in repair. The time that the machine works before breaking down is exponentially distributed with mean $1/\lambda$. The density $f(\cdot)$ and distribution function $F(\cdot)$ of the exponential distribution with mean $1/\lambda$ are given by

$$f(x) = \lambda e^{-\lambda x}, \qquad F(x) = 1 - e^{-\lambda x}, \qquad x \geq 0.$$

The repair time is also exponentially distributed with mean $1/\mu$. What is the probability that at time t the machine is working?

Let $p_0(t)$ (resp. $p_1(t)$) be the probability that the machine at time t is in repair (resp. working). Suppose that at time 0 the machine is working, so $p_0(0) = 1 - p_1(0) = 0$. We want to derive differential equations for $p_0(t)$ and $p_1(t)$. To do so we need the following property of the exponential distribution. This property is called the *memoryless property* of exponential distributions.

Property 2.6.1 *Let X be an exponentially distributed random variable with mean $1/\lambda$. Then the following holds for all $x \geq 0$ and $t \geq 0$,*

$$P(X > x + t | X > t) = P(X > x) = e^{-\lambda x}.$$

We often use this property in the form

$$P(X < t + \Delta t | X > t) = 1 - e^{\lambda \Delta t} = \lambda \Delta t + o(\Delta t), \qquad (\Delta t \to 0).$$

Based on this property we get

$$\begin{aligned} p_0(t + \Delta t) &= p_0(t)(1 - \mu \Delta t) + p_1(t)\lambda \Delta t + o(\Delta t), \qquad (\Delta t \to 0), \\ p_1(t + \Delta t) &= p_0(t)\mu \Delta t + p_1(t)(1 - \lambda \Delta t) + o(\Delta t), \qquad (\Delta t \to 0). \end{aligned}$$

Hence, by letting $\Delta t \to 0$,

$$\begin{aligned} p_0'(t) &= -\mu p_0(t) + \lambda p_1(t), \\ p_1'(t) &= \mu p_0(t) - \lambda p_1(t). \end{aligned}$$

or in matrix notation

$$p'(t) = p(t)Q, \qquad t > 0, \tag{2.7}$$

where

$$p(t) = (p_0(t), p_1(t)), \qquad Q = \begin{pmatrix} -\mu & \mu \\ \lambda & -\lambda \end{pmatrix}.$$

The matrix Q is called the *generator of the Markov process*. To solve the system of differential equations (2.7) we try a solution of the form

$$p(t) = v e^{at},$$

where v is a row vector and a some constant. Substituting this form into (2.7) we get

$$av = vQ.$$

So a must be an eigenvalue of Q and v a corresponding row eigen vector. The eigenvalues of Q are 0 and $-(\lambda+\mu)$ with corresponding eigenvectors (λ, μ) and $(1, -1)$, resp. Hence, for any c_1 and c_2,

$$p(t) = c_1(\lambda, \mu) + c_2(1, -1)e^{-(\lambda+\mu)t}, \qquad t \geq 0, \tag{2.8}$$

is a solution of (2.7). The constants c_1 and c_2 follow from the initial condition $p(0) = (0, 1)$ yielding

$$c_1 = \frac{1}{\lambda+\mu}, \qquad c_2 = \frac{-\lambda}{\lambda+\mu}.$$

This completes the determination of the probabilities $p_0(t)$ and $p_1(t)$. By letting t tend to infinity in (2.8) we obtain for the limiting probabilities p_0 and p_1 that

$$p_0 = 1 - p_1 = \frac{\lambda}{\lambda+\mu}. \tag{2.9}$$

This could also have been obtained by letting t tend to infinity in (2.7) and assuming that $p'(t)$ tends to zero as t tends to infinity. This yields a set of two equations for p_0 and p_1, which, together with the normalization condition $p_0 + p_1 = 1$, has (2.9) as solution.

Chapter 3

Refresher

The aim of this section is to refresh some well-known results from the theory of linear recurrence relations and linear differential equations with constant coefficients. The results can be found in standard textbooks, such as e.g. [13]. We first start to review some results for recurrence relations and then we treat differential equations. It will appear that the results for these two types of equations are very similar.

3.1 Linear recurrence relations of order k

A recurrence relation of the form

$$x_{n+k} + a_1 x_{n+k-1} + a_2 x_{n+k-2} + \cdots + a_k x_n = b_n, \qquad n = 0, 1, 2, \ldots \tag{3.1}$$

is called a *linear recurrence relation with constant coefficients of order* k. If $b_n = 0$ for all n, then (3.1) is called *homogeneous*, and *inhomogeneous* otherwise. So, for example, the Fibonacci series in section (2.1) is determined by a homogeneous second order linear recurrence relation with coefficients $a_1 = a_2 = -1$.

Solutions of recurrence relations are not unique. To make it unique one has to impose extra conditions. One possibility is to prescribe the starting values $x_0, \ldots, x_{k-1}$. In this case we speak of an *initial-value problem.* The solution of an initial-value problem is unique. It is also possible to impose other conditions. For example, we may want to solve a second order recurrence relation with prescribed values at two end points, say x_0 and x_N. Then we have to determine $x_1, \ldots, x_{N-1}$ given x_0 and x_N. Such a problem is usually referred to as a *boundary-value problem.* The solution of a boundary-value problem may not exist, and if it does exist, it is not necessarily unique. The Fibonacci series is an example of a initial value problem, the roulette problem in section 2.3 and the loan in section 2.4 are examples of boundary value problems.

Let us first consider the homogeneous recurrence relation

$$x_{n+k} + a_1 x_{n+k-1} + a_2 x_{n+k-2} + \cdots + a_k x_n = 0, \qquad n = 0, 1, 2, \ldots \tag{3.2}$$

If two sequences $\{x_n^(1)\}$ and $\{x_n^{(2)}\}$ are solutions of (3.2) and c_1 and c_2 are two complex numbers, then the sequence $\{c_1 x_n^{(1)} + c_2 x_n^{(2)}\}$ is again a solution of (3.2). This shows that the solutions form a vector space.

To show that l solutions $\{x_n^{(1)}\}, \ldots, \{x_n^{(l)}\}$ are linearly independent, one has to check that the relation

$$c_1 x_n^{(1)} + \cdots + c_l x_n^{(l)} = 0, \qquad n = 0, 1, 2, \ldots$$

is only possible if all coefficients $c_1, \ldots, c_l$ are zero. The following result states that it suffices to check this for $n = 0, \ldots, k-1$. It is a direct consequence of the fact that when $\{x_n\}$ is a solution of (3.2) with $x_0 = \cdots = x_{k-1} = 0$, then $x_n = 0$ for all n.

Lemma 3.1.1 *Let the sequences $\{x_n^{(1)}\}, \ldots, \{x_n^{(l)}\}$ be solutions of (3.2). These l solutions are linearly independent if and only if the vectors*

$$\begin{pmatrix} x_0^{(1)} \\ x_1^{(1)} \\ \vdots \\ x_{k-1}^{(1)} \end{pmatrix}, \begin{pmatrix} x_{2,0}^{(2)} \\ x_{2,1}^{(2)} \\ \vdots \\ x_{2,k-1}^{(2)} \end{pmatrix}, \ldots, \begin{pmatrix} x_{l,0}^{(l)} \\ x_{l,1}^{(l)} \\ \vdots \\ x_{l,k-1}^{(l)} \end{pmatrix}$$

are linearly independent.

From this lemma we immediately obtain the following result.

Theorem 3.1.2 *The set of all solutions of (3.2) forms a k-dimensional vector space over the complex field.*

This result is useful to find the general solution of (3.2). We first try to find k linearly independent solutions $\{x_n^{(1)}\}, \ldots, \{x_n^{(k)}\}$ of (3.2). Then the general solution is given by

$$z_n = c_1 x_n^{(1)} + \cdots + c_k x_n^{(k)}, \qquad n = 0, 1, 2, \ldots,$$

where $c_1, \ldots, c_k$ are arbitrary complex numbers. To obtain k independent solutions we try solutions of the form

$$x_n = y^n.$$

Substituting this form into (3.2) and then dividing by common factors we get

$$y^k + a_1 y^{k-1} + a_2 y^{k-2} + \cdots + a_k = 0. \tag{3.3}$$

This equation is called the *characteristic equation* of (3.2). Suppose that this equation has k distinct roots $y_1, \ldots, y_k$. Then the solutions $\{y_1^n\}, \ldots, \{y_k^n\}$ are linearly independent, which follows lemma 3.1.1 and the fact that the *VanderMonde matrix*

$$\begin{pmatrix} 1 & 1 & \ldots & 1 \\ y_1 & y_2 & \cdots & y_k \\ \vdots & \vdots & & \vdots \\ y_1^{k-1} & y_2^{k-1} & \ldots & y_k^{k-1} \end{pmatrix}$$

is non-singular. Namely, suppose that we have a linear combination of rows which vanishes, i.e. there are coefficients $c_0, \ldots, c_{k-1}$ such that

$$c_0 + c_1 y_i + \cdots + c_{k-1} y_i^{k-1} = 0, \qquad i = 1, \ldots, k.$$

This means that the polynomial

$$p(y) := c_0 + c_1 y + \cdots + c_{k-1} y^{k-1}$$

has $y_1, \ldots, y_k$ as zeros. Since the degree of $p(\cdot)$ is k this implies that $p(\cdot)$ is identical to zero, so $c_0 = \ldots = c_{k-1} = 0$. Hence the general solution of (3.2) is given by

$$z_n = c_1 y_1^n + \cdots + c_k y_k^n, \qquad n = 0, 1, 2, \ldots$$

In case the characteristic equation (3.3) has multiple roots, we have to find for each root y with multiplicity m also m independent solutions. It is easy to verify that when we take, if y is a root of (3.3) with multiplicity m, the m sequences

$$y^n, ny^{n-1}, \binom{n}{2} y^{n-2}, \ldots, \binom{n}{m-1} y^{n-m+1}, \qquad n = 0, 1, 2, \ldots$$

as solutions, and we do so for each root, then we get k independent solutions of (3.2). Summarizing, we find the following result.

Theorem 3.1.3 *Let $y_1, \ldots, y_l$ be the roots of the characteristic equation (3.3) and suppose that y_i has multiplicity m_i, $i = 1, \ldots, l$. Then the sequences*

$$y_1^n, ny_1^{n-1}, \binom{n}{2} y_1^{n-2}, \ldots, \binom{n}{m_1-1} y_1^{n-m_1+1},$$

$$\ldots$$

$$y_l^n, ny_l^{n-1}, \binom{n}{2} y_l^{n-2}, \ldots, \binom{n}{m_l-1} y_l^{n-m_l+1}$$

for $n = 0, 1, 2, \ldots$ form a solution basis for recurrence relation (3.2).

Now consider the inhomogeneous relation (3.1). It is readily verified that the general solution of the inhomogeneous relation is the sum of a *particular solution* of the inhomogeneous relation and the general solution of the homogeneous relation. To find a particular solution one usually tries simple, suitable solutions, possibly with some parameters which have to be determined yet.

3.2 Systems of recurrence relations

The general form of a system of linear recurrence relations with constant coefficients of the first order is, in vector-matrix notation,

$$x_{n+1} = Ax_n + b_n, \qquad n = 0, 1, 2, \ldots \tag{3.4}$$

where

$$x_n = \begin{pmatrix} x_{1,n} \\ x_{2,n} \\ \vdots \\ x_{k,n} \end{pmatrix}, \qquad b_n = \begin{pmatrix} b_{1,n} \\ b_{2,n} \\ \vdots \\ b_{k,n} \end{pmatrix}, \qquad A = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,k} \\ a_{2,1} & a_{2,2} & \ldots & a_{1,k} \\ \vdots & & & \vdots \\ a_{k,1} & a_{k,2} & \ldots & a_{k,k} \end{pmatrix}.$$

The linear recurrence relations of order k treated in the previous section may also be written as a system of recurrence relations of the first order. For instance, consider the second order relation

$$y_{n+2} + a_1 y_{n+1} + a_2 y_n = c_n, \qquad n = 0, 1, 2, \ldots$$

Define the vector x_n by

$$x_n = \begin{pmatrix} y_n \\ y_{n+1} \end{pmatrix}.$$

For the vectors x_n we get the first order relation

$$x_{n+1} = \begin{pmatrix} y_{n+1} \\ y_{n+2} \end{pmatrix} = \begin{pmatrix} y_{n+1} \\ -a_2 y_n - a_1 y_{n+1} \end{pmatrix} + \begin{pmatrix} 0 \\ c_n \end{pmatrix} = Ax_n + b_n$$

where

$$A = \begin{pmatrix} 0 & 1 \\ -a_2 & -a_1 \end{pmatrix}, \qquad b_n = \begin{pmatrix} 0 \\ c_n \end{pmatrix}.$$

In the same way higher order recurrence relations can be translated to a system of recurrence relations of the first order.

Let us consider the homogeneous system

$$x_{n+1} = Ax_n, \qquad n = 0, 1, 2, \ldots \tag{3.5}$$

The solutions of the homogeneous system form a vector space and the analogue of lemma 3.1.1 holds.

Lemma 3.2.1 *Let the sequences $\{x_n^{(1)}\}, \ldots, \{x_n^{(l)}\}$ be solutions of (3.5). These l solutions are linearly independent if and only if the vectors $x_0^{(1)}, \ldots, x_0^{(l)}$ are linearly independent.*

This implies the following result.

Theorem 3.2.2 *The set of all solutions of (3.5) forms a k-dimensional vector space over the complex field.*

By iterating relation (3.5) it follows that the general solution of (3.5) is given by

$$x_n = A^n x_0, \qquad n = 0, 1, 2, \ldots$$

So, in fact, solving (3.5) amounts to determining the powers of the matrix A. Below we describe an approach to find these powers. We try to find a solution of (3.5) of the form

$$x_n = \lambda^n v$$

where λ is some number and v a vector. Substitution of this form into (3.5) yields

$$\lambda v = Av.$$

Hence λ is an eigenvalue and v a corresponding eigenvector. Suppose that A has k distinct eigenvalues $\lambda_1, \ldots, \lambda_k$. Since the corresponding eigenvectors $v_1, \ldots, v_k$ are linearly independent (see e.g. [28]) we can conclude that

$$x_n = c_1\lambda_1^n v_1 + c_2\lambda_2^n v_2 + \cdots + c_k\lambda_k^n v_k, \qquad n = 0, 1, 2, \ldots \tag{3.6}$$

is the general solution, where $c_1, \ldots, c_k$ are arbitray complex numbers. In case λ is an eigenvalue of A with multiplicity $m > 1$, but the dimension of the eigenspace of λ is also m (so we can find m independent eigenvectors), then the general solution is again of the form (3.6). A difficulty arises when the number of independent eigenvectors is less than the multiplicity of λ, since then we do not have sufficient solutions. In that case we call λ *defective*. For a defective eigenvalue λ we try to find extra solutions of the form $\lambda^n(nv + w)$ or even $\lambda^n(n^2v + nw + u)$ (and so on). For instance, suppose that λ is a double eigenvalue which has exactly one independent eigenvector v. Then we try to find a solution of the form

$$x_n = \lambda^n(nw + u),$$

where w and u are vectors to be determined yet. Substituting this form into (3.5) and dividing by common powers of λ leads to

$$\lambda(nw + w + u) = nAw + Au.$$

This equation is satisfied if

$$Aw = \lambda w, \qquad Au = \lambda w + \lambda u.$$

So we can choose $w = v$ and it can be show that the equation for u has a solution. To formulate more precisely how the general solution looks like we need the following result stating that each matrix is similar to a block-diagonal matrix, the blocks of which are so-called Jordan boxes (see e.g. [20, 28]).

Theorem 3.2.3 *For every complex $k \times k$ matrix A there exists a non-singular $k \times k$ matrix V and a $k \times k$ matrix J such that*

$$A = VJV^{-1}.$$

The matrix J is of the form

$$J = \begin{pmatrix} J_0 & 0 & 0 & \ldots & 0 \\ 0 & J_1 & 0 & \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \ldots & J_s \end{pmatrix}$$

where J_0 is a diagonal matrix with diagonal $\lambda_1, \lambda_2, \ldots, \lambda_q$, and J_i is an $r_i \times r_i$ matrix given by

$$J_i = \begin{pmatrix} \lambda_{q+i} & 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & \lambda_{q+i} & 1 & 0 & \ldots & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & 0 & 0 & \ldots & \lambda_{q+i} & 1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & \lambda_{q+i} \end{pmatrix}$$

for $i = 1, \ldots, s$ (so $k = q + r_1 + \ldots + r_s$). The λ_j, $j = 1, \ldots, q+s$, are the eigenvalues of A, which need not all be distinct. If λ_j is a simple eigenvalue, then it occurs in J_0, and therefore, if A has k distinct eigenvalues, then J is the diagonal matrix

$$J = \begin{pmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \ldots & \lambda_k \end{pmatrix}.$$

The matrix J in the theorem above is called the *Jordan form* of A, and V is a corresponding *Jordan basis*. Suppose that V has as columns the vectors $v_1, \ldots, v_k$. Since $AV = VJ$ the vectors $v_1, \ldots, v_k$ satisfy the relations

$$\begin{aligned} Av_1 &= \lambda_1 v_1 \\ &\ldots \\ Av_q &= \lambda_q v_q \\ Av_{q+1} &= \lambda_{q+1} v_{q+1} \\ Av_{q+2} &= v_{q+1} \lambda_{q+2} v_{q+2} \\ &\ldots \\ Av_{q+r_1} &= v_{q+r_1-1} \lambda_{q+r_1} v_{q+r_1} \\ &\ldots \\ Av_{k-r_s+1} &= \lambda_{q+s} v_{k-r_s+1} \\ &\ldots \\ Av_k &= v_{k-1} + \lambda_{q+s} v_k \end{aligned}$$

We can now construct k independent solutions of (3.5) by using the vectors $v_1, \ldots, v_k$. By exploiting the structure of J it is not difficult to show that the sequences

$$\lambda_1^n v_1, \lambda_2^n v_2, \ldots, \lambda_q^n v_q,$$

$$\lambda_{q+1}^n v_{q+1},$$

$$\lambda_{q+1}^n (n v_{q+1} + v_{q+2}),$$

$$\lambda_{q+1}^n (\binom{n}{2} v_{q+1} + n v_{q+2} + v_{q+3}),$$

$$\ldots$$

$$\lambda_{q+1}^n (\binom{n}{r_1 - 1} v_{q+1} + \binom{n}{r_1 - 2} n v_{q+2} + \cdots + v_{q+r_1}),$$

$$\ldots$$

$$\lambda_{q+s}^n v_{k-r_s+1},$$

$$\ldots$$

$$\lambda_{q+s}^n (\binom{n}{r_s - 1} v_{k-r_s+1} + \binom{n}{r_s - 2} n v_{k-r_s+2} + \cdots + v_k)$$

for $n = 0, 1, 2, \ldots$ make up a set of k independent solutions of (3.5). So we have found a basis of solutions of the form $\lambda^n (n^l w_l + n^{l-1} w_{l-1} + \ldots + n w_1 + w_0)$ where λ is an eigenvalue of A, the degree l is less or equal to the multiplicity of λ and the vectors w_i can be expressed in terms of the corresponding (generalized) eigenvectors. We remark from a practical point of view that in general the computation of the Jordan form is difficult.

Hence the general solution of (3.5) can be expresses as a linear combination of the preceding k independent solutions. By adding to this solution a particular solution of the inhomogeneous system (3.4) we finally obtain the general solution of the inhomogeneous system.

3.3 Systems of linear differential equations

The theory of systems of linear differential equations is very similar to the theory of systems of linear recurrence relations. Therefore it will suffice to present a brief description of the results for the differential equations.

The general form of a system of linear differential equations with constant coefficients of the first order is, in vector-matrix notation,

$$x'(t) = Ax(t) + b(t), \qquad t \in I, \tag{3.7}$$

where I is some real interval and

$$x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_k(t) \end{pmatrix}, \qquad b(t) = \begin{pmatrix} b_1(t) \\ b_2(t) \\ \vdots \\ b_k(t) \end{pmatrix}, \qquad A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,k} \\ a_{2,1} & a_{2,2} & \dots & a_{1,k} \\ \vdots & & & \vdots \\ a_{k,1} & a_{k,2} & \dots & a_{k,k} \end{pmatrix}.$$

We assume that $b(t)$ is a continuous vector on the interval I.

Let us first consider the homogeneous system

$$x'(t) = Ax(t), \qquad t \in I. \tag{3.8}$$

Theorem 3.3.1 *The set of all solutions of (3.8) on I forms a k-dimensional vector space over the complex field.*

If $x^{(1)}(t), \dots, x^{(k)}(t)$ are a set of k independent solutions of (3.8), they are said to form a basis of solutions of (3.8). The general solution can be expressed as a linear combination of the basis solutions,

$$x(t) = c_1 x^{(1)}(t) + c_2 x^{(2)}(t) + \cdots + c_k x^{(k)}(t), \qquad t \in I,$$

where $c_1, \dots, c_k$ are arbitrary complex numbers.

To obtain k independent solutions of (3.8) we try to find solutions of the form

$$x(t) = e^{\lambda t} v$$

where λ is some number and v a vector. Substitution of this form into (3.8) yields, after dividing by $e^{\lambda t}$,

$$\lambda v = Av.$$

Hence λ is an eigenvalue and v a corresponding eigenvector. If A has k distinct eigenvalues $\lambda_1, \dots, \lambda_k$, with associated eigenvectors $v_1, \dots, v_k$, the solutions

$$e^{\lambda_1 t} v_1, e^{\lambda_2 t} v_2, \dots, e^{\lambda_k t} v_k$$

for $t \in I$ form a solution basis and so we can conclude that

$$x(t) = c_1 e^{\lambda_1 t} v_1 + c_2 e^{\lambda_2 t} v_2 + \cdots + c_k e^{\lambda_k t} v_k, \qquad t \in I,$$

is the general solution, where $c_1, \dots, c_k$ are arbitray complex numbers. If λ is an eigenvalue of A with multiplicity m and the number of eigenvectors for λ is less than m, we try to find extra solutions of the form $e^{\lambda t}(t^l w_l + t^{l-1} w_{l-1} + \ldots + t w_1 + w_0)$. To formulate precisely how the general solution looks like we need again the Jordan form of A.

Let J be the Jordan form of A, as given by theorem 3.2.3, and suppose V is a nonsingular matrix such that

$$AV = VJ.$$

Denote the columns of V by $v_1, \ldots, v_k$. Then the solutions

$$e^{\lambda_1 t}v_1, e^{\lambda_2 t}v_2, \ldots, e^{\lambda_q t}v_q,$$

$$e^{\lambda_{q+1} t}v_{q+1},$$

$$e^{\lambda_{q+1} t}(tv_{q+1} + v_{q+2}),$$

$$\ldots$$

$$e^{\lambda_{q+1} t}(\frac{t^{r_1-1}}{(r_1-1)!}v_{q+1} + \frac{t^{r_1-2}}{(r_1-2)!}v_{q+2} + \cdots + tv_{q+r_1-1}) + v_{q+r_1}),$$

$$\ldots$$

$$e^{\lambda_{q+s} t}v_{k-r_s+1},$$

$$\ldots$$

$$e^{\lambda_{q+s} t}(\frac{t^{r_s-1}}{(r_s-1)!}v_{k-r_s+1} + \frac{t^{r_s-2}}{(r_s-2)!}v_{k-r_s+2} + \cdots + tv_{k-1}) + v_k),$$

for $t \in I$ make up a set of k independent solutions of (3.8). So the general solution of (3.5) can be expresses as a linear combination of these solutions. We finally remark that by adding to this solution a particular solution of the inhomogeneous system (3.7) we obtain the general solution of the inhomogeneous system.

3.4 Exercises

Exercise 3.4.1 *(Fibonacci series)*
The Fibonacci series is determined by the recurrence relation (2.1).

(i) Solve this recurrence relation.

(ii) Determine the limit of x_{n+1}/x_n as $n \to \infty$ (this ratio is called the golden mean).

Exercise 3.4.2 *(Roulette game)*
In section 2.3 a recurrence relation is derived for the winning chances of a roulette player who is playing one dollar at a time.

(i) Solve this recurrence relation.

Suppose that the player plays with 2 dollars at a time, unless he has one dollar left.

(ii) Formulate and solve the recurrence relation for the winning chances when the player uses this strategy.

Exercise 3.4.3 *(Transmission channel)*
Consider the problem in section 2.5 and suppose that the buffer size B is infinite.

(i) Give the general solution of the recurrence relation (2.6) for the probabilities x_n.

(ii) What are the proper conditions to uniquely determine the unknown coefficients in this solution?

Now suppose that the buffer size B is finite.

(iii) Formulate the recurrence relation for the probabilities x_n and show how the probabilities x_n can be determined.

(iv) Express the fraction of packets that is lost in terms of the probabilities x_n.

Exercise 3.4.4
Prove lemma 3.1.1 and the theorems 3.1.2 and 3.1.3.

Exercise 3.4.5 *(Machine with limited buffer)*
Let us consider a machine where jobs are processed one at a time and in order of arrival. The interarrival times of the jobs are exponentially distributed with mean $1/\lambda$. The processing times are exponentially distributed with mean $1/\mu$. Jobs arriving while the machine is busy wait for their turn in a buffer, if there is room, and otherwise, if the buffer is full, they are processed somewhere else. The buffer has space for one job only.
We want to find the distribution of the number of jobs at the machine (the one being processed and the number of jobs waiting in the buffer) at time t. Due to the memoryless property of the exponential distribution (see property 2.6.1) the state of the system at time t can be described by the number of jobs in the system. Let $p_i(t)$ be the probability that at time t there are i jobs in the system, $i = 0, 1, 2$.

(i) Show that, when there is a job being processed, the time that elapses till something happens (an arrival or job completion) starting from time t is exponentially distributed with mean $1/(\lambda+\mu)$.

(ii) Derive the following differential equations for $p_i(t)$, $i = 0, 1, 2$,

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t) + \mu p_1(t), \\ p_1'(t) &= \lambda p_0(t) - (\lambda+\mu) p_1(t) + \mu p_2(t), \\ p_2'(t) &= \lambda p_1(t) - \mu p_2(t). \end{aligned}$$

(iii) Solve the differential equations in (ii) with initial condition

$$(p_0(0), p_1(0), p_2(0)) = (1, 0, 0),$$

i.e. at time 0 the system is empty.

(iv) Determine the limiting probabilities $p_i = \lim_{t\to\infty} p_i(t)$, $i = 0, 1, 2$.

(v) What is in the long run the occupation rate of the machine and the fraction of jobs that is processed somewhere else?

Chapter 4

Basic concepts

This chapter is devoted to some basic concepts from probability theory and queueing theory. These concepts can be found is every standard textbook on these topics, see e.g. [27, 44].

4.1 Random variable

In this report random variables are denoted by capitals, X, Y, etc. The expected value or mean of X is denoted by $E(X)$ and its variance by $\mathrm{var}(X)$, or by $\sigma^2(X)$ where $\sigma(X)$ is the standard deviation of X.

An important quantity is the *coefficient of variation* of the positive random variable X defined as

$$c_X = \frac{\sigma(X)}{E(X)}.$$

The coefficient of variation is a measure of the variability of the random variable X. For instance, the deterministic distribution has $c_X = 0$ and the exponential distribution has $c_X = 1$.

4.2 Useful probability distributions

This section discusses a number of important distributions which have been found useful for describing random variables in many applications.

4.2.1 Exponential distribution

The density of an exponential distribution with parameter μ is given by

$$f(t) = \mu e^{-\mu t}, \qquad t > 0.$$

The distribution function equals

$$F(t) = 1 - e^{-\mu t}, \qquad t \geq 0.$$

For this distribution we have

$$E(X) = \frac{1}{\mu}, \qquad \mathrm{var}(X) = \frac{1}{\mu^2}, \qquad c_X = 1.$$

As we already mentioned in section 2.6 an important property of an exponential random variable X with parameter μ is the memoryless property

$$P(X > x + t | X > t) = P(X > x) = e^{-\mu x}.$$

4.2.2 Erlang distribution

A random variable X has an *Erlang-k* ($k = 1, 2, \ldots$) distribution with mean k/μ if X is the sum of k independent random variables $X_1, \ldots, X_k$ having a common exponential distribution with mean $1/\mu$. The common notation is $E_k(\mu)$ or briefly E_k. The density of an $E_k(\mu)$ distribution is given by

$$f(t) = \mu \frac{(\mu t)^{k-1}}{(k-1)!} e^{-\mu t}, \qquad t > 0.$$

The distribution function equals

$$F(t) = 1 - \sum_{j=0}^{k-1} \frac{(\mu t)^j}{j!} e^{-\mu t}, \qquad t \geq 0.$$

The parameter μ is called the scale parameter, k is the shape parameter. A phase diagram of the E_k distribution is shown in figure 4.1.

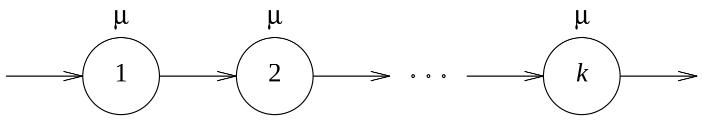

Figure 4.1: Phase diagram for the Erlang-k distribution with scale parameter μ

The mean, variance and squared coefficient of variation are equal to

$$E(X) = \frac{k}{\mu}, \qquad \mathrm{var}(X) = \frac{k}{\mu^2}, \qquad c_X^2 = \frac{1}{k}.$$

A convenient distribution arises when we mix an E_{k-1} and E_k distribution with the same scale parameters. The notation used is $E_{k-1,k}$. A random variable X has an $E_{k-1,k}(\mu)$

distribution, if X is with probability p (resp. $1-p$) the sum of $k-1$ (resp. k) independent exponentials with common mean $1/\mu$. The density of this distribution has the form

$$f(t) = p\mu\frac{(\mu t)^{k-2}}{(k-2)!}e^{-\mu t} + (1-p)\mu\frac{(\mu t)^{k-1}}{(k-1)!}e^{-\mu t}, \qquad t > 0,$$

where $0 \le p \le 1$. As p runs from 0 to 1, the squared coefficient of variation of the mixed Erlang distribution varies from $1/(k-1)$ to $1/k$. It will appear (later on) that this property is useful for fitting distributions on the first two moments.

4.2.3 Hyperexponential distribution

A random variable X is hyperexponentially distributed if X is with probability p_i, $i = 1,\dots,k$ an exponential X_i with mean $1/\mu_i$. The notation used is $H_k(p_1,\dots,p_k;\mu_1,\dots,\mu_k)$, or simply H_k. The density is given by

$$f(t) = \sum_{i=1}^{k} p_i\mu_i e^{-\mu_i t}, \qquad t > 0,$$

and the mean is equal to

$$E(X) = \sum_{i=1}^{k} \frac{p_i}{\mu_i}.$$

The coefficient of variation c_X of this distribution is greater or equal to 1. A phase diagram of the H_k distribution is shown in figure 4.2.

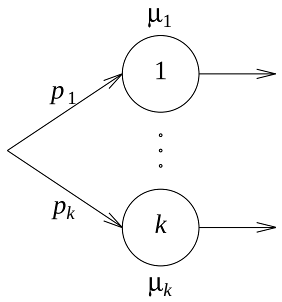

Figure 4.2: Phase diagram for the hyperexponential distribution

4.2.4 Phase-type distribution

The preceding distributions are all special cases of the *phase-type distribution*. The notation is PH. This distribution is characterized by a Markov chain with states $1,\dots,k$ (the so-called phases) and a transition probability matrix P which is *transient*. This means that P^n tends to zero as n tends to infinity. In words, eventually you will always leave the

Markov chain. The residence time in state i is exponentially distributed with mean $1/\mu_i$, and the Markov chain is entered with probability p_i in state i, $i = 1, \ldots, k$. Then X has a phase-type distribution if X is the total residence time in the preceding Markov chain, i.e. X is the total time elapsing from start in the Markov chain till departure from the Markov chain.

We mention two important classes of phase-type distributions which are *dense in the class of all non-negative distribution functions.* This is ment in the sense that for any non-negative distribution function $F(\cdot)$ a sequence of phase-type distributions can be found which pointwise converges at the points of continuity of $F(\cdot)$. The denseness of the two classes makes them very useful as a practical modelling tool. A proof of the denseness can be found in [37, 38]. The first class is the class of *Coxian distributions*, notation C_k, and the other class consists of *mixtures of Erlang distributions with the same scale parameters.* The phase representations of these two classes are shown in the figures 4.3 and 4.4

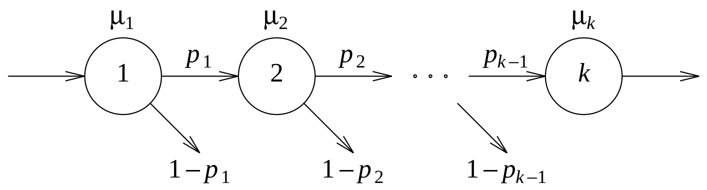

Figure 4.3: Phase diagram for the Coxian distribution

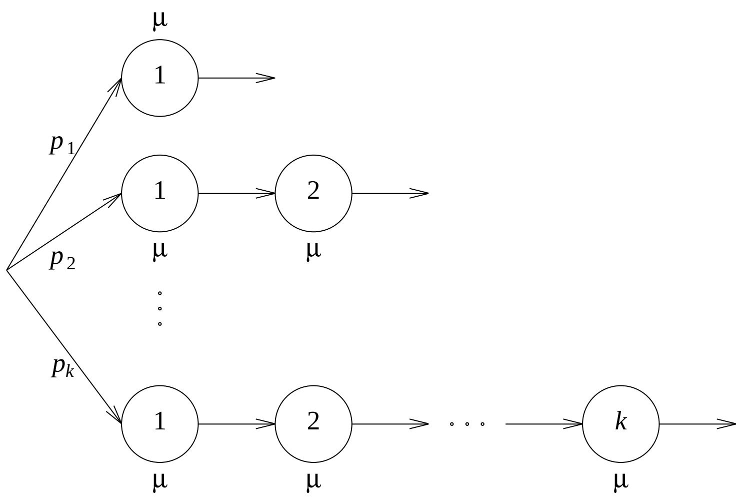

Figure 4.4: Phase diagram for the mixed Erlang distribution

A random variable X has a Coxian distribution of order k if it has to go through up to at most k exponential phases. The mean length of phase n is μ_n, $n = 1, \ldots, k$. It starts in phase 1. After phase n it comes to an end with probability $1 - p_n$ and it enters the next phase with probability p_n. Obviously $p_k = 0$. For the Coxian-2 distribution it holds that the squared coefficient of variation is greater or equal to 0.5 (see exercise 4.10.6).

A random variable X has a mixed Erlang distribution of order k if it is with probability p_n the sum of n exponentials with the same mean $1/\mu$, $n = 1, \dots, k$.

4.3 Fitting distributions

In practice it often occurs that the only information of random variables that is available is their mean and standard deviation, or if one is lucky, some real data. To obtain an approximating distribution it is common to fit a phase-type distribution on the mean, E(X), and the coefficient of variation, c_X, of a given positive random variable X, by using the following simple approach.

In case $0 < c_X < 1$ one fits an $E_{k-1,k}$ distribution (see subsection 4.2.2). More specifically, if

$$\frac{1}{k} \leq c_X^2 \leq \frac{1}{k-1},$$

for certain $k = 2, 3, \dots$, then the approximating distribution is with probability p (resp. $1-p$) the sum of $k-1$ (resp. k) independent exponentials with common mean $1/\mu$. By choosing (see e.g. [44])

$$p = \frac{1}{1+c_X^2}[kc_X^2 - \{k(1+c_X^2) - k^2c_X^2\}^{1/2}], \qquad \mu = \frac{k-p}{E(X)},$$

the $E_{k-1,k}$ distribution matches $E(X)$ and c_X.

In case $c_X \geq 1$ one fits a $H_2(p_1, p_2; \mu_1, \mu_2)$ distribution. The hyperexponential distribution however is not uniquely determined by its first two moments. In applications, the H_2 distribution with *balanced means* is often used. This means that the normalization

$$\frac{p_1}{\mu_1} = \frac{p_2}{\mu_2}$$

is used. The parameters of the H_2 distribution with balanced means and fitting $E(X)$ and c_X (≥ 1) are given by

$$p_1 = \frac{1}{2}\left(1 + \sqrt{\frac{c_X^2 - 1}{c_X^2 + 1}}\right), \qquad p_2 = 1 - p_1,$$

$$\mu_1 = \frac{2p_1}{E(X)}, \qquad \mu_1 = \frac{2p_2}{E(X)}.$$

It also possible to make a more sophisticated use of phase-type distributions by, e.g., trying to match the first three (or even more) moments of X or to approximate the shape of X. For the case $c_X^2 > 1$ it is shown in [48] that a three moment fit to the random variable X by a Coxian-2 distribution exists only if the third moment of X satisfies

$$E(X^3) \geq \frac{3}{2}(1 + c_X^2)^2 E(X)^3. \tag{4.1}$$

For the case $0.5 \le c_X^2 \le 1$ in [21] lower and upper bounds on the third moment of X are given such that a three moment fit by a C_2 distribution exists. If it exists, then the parameters μ_1, μ_2 and p_1 of the three moment fit are given by

$$\mu_1 = \frac{1}{2}\left[a_1 + \sqrt{a_1^2 - 4a_2}\right], \quad \mu_2 = \frac{1}{2}\left[a_1 - \sqrt{a_1^2 - 4a_2}\right], \quad p_1 = \frac{\mu_2(m_1\mu_1 - 1)}{\mu_1},$$

where $m_i = E(X^i)$, $i = 1, 2, 3$, and

$$a_2 = \frac{6m_1^2 - 3m_2}{6m_2^2/4 - m_1 m_3} \qquad a_1 = \frac{1}{m_1} + \frac{m_2 a_2}{2m_1}.$$

The Coxian-2 distribution can of course also be used for a two-moment fit. The following set is suggested by [31],

$$\mu_1 = 2/E(X), \quad p_1 = 0.5/c_X^2, \quad \mu_2 = \mu_1 p_1.$$

In [25] a queueing model is used to explore the quality of two- and three-moment approximations, and in some cases, approximations considering the distribution shape.

Phase-type distributions may of course also naturally arise in practical applications. For example, if the processing of a job involves performing several tasks, where each task takes an exponential amount of time, then the processing time can be described by an Erlang distribution.

4.4 Poisson process

Let $N(t)$ be the number of arrivals in $[0, t]$ for a *Poisson process* with intensity λ, i.e. the time between successive arrivals is exponentially distributed with parameter λ and independent of the past. Then $N(t)$ has a *Poisson distribution* with parameter λt, so

$$P(N(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \qquad k = 0, 1, 2, \ldots$$

The mean, variance and coefficient of variation of $N(t)$ are equal to

$$E(N(t) = \lambda t, \qquad \mathrm{var}(X) = \lambda t, \qquad c_{N(t)}^2 = \frac{1}{\lambda t}.$$

The Poisson process is an extremely useful process for modelling purposes in many practical applications, such as, e.g. to model arrival processes for queueing models or demand processes for inventory systems. It is empirically found that in many circumstances the arising stochastic processes can be well approximated by a Poisson process.

Next we mention two important properties of a Poisson process (see e.g. [34]).

(i) Suppose that $N_1(t)$ and $N_2(t)$ are two independent Poisson processes with respective intensities λ_1 and λ_2. Then the sum $N_1(t) + N_2(t)$ of the two processes is again a Poisson process with intensity $\lambda_1 + \lambda_2$.

(ii) Suppose that $N(t)$ is a Poisson process with intensity λ and that each arrival is marked with probability p independent of all other arrivals. Let $N_1(t)$ and $N_2(t)$ denote respectively the number of marked and unmarked arrivals in $[0, t]$. Then $N_1(t)$ and $N_2(t)$ are both Poisson processes with respective intensities λp and $\lambda(1-p)$. And these two processes are independent.

So Poisson processes remain Poisson under merging and splitting.

4.5 Queueing models and Kendall's notation

A queueing model is, among others, characterized by the arrival process of customers, the service times and the numbers of servers. Usually we assume that the interarrival times are independent and have a common distribution. The type of distribution is specified by a letter. For example, for a general distribution the letter G is used, M for the exponential distribution (M stands for Memoryless) and D for deterministic times. The same letters are used for the service times, which are also assumed to be independent and identically distributed. The basic queueing model is shown in figure 4.5.

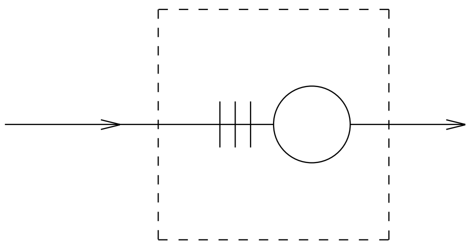

Figure 4.5: Basic queueing model

The basic model can be described by a three-part code $a/b/c$. The first letter specifies the interarrival time distribution, the second one the service time distribution and the last one the number of servers. Some examples are $M/M/1$, $M/M/c$, $M/G/1$, $G/M/1$ and $M/D/1$. The notation can be extended with an extra letter to cover other queueing models. For example, a system with exponential interarrival and service times, one server and having waiting room only for N customers (including the one in service) is abbreviated by the four letter code $M/M/1/N$.

Queueing models have been proved to very useful in many practical applications in areas such as, e.g., production systems, inventory systems and communication systems. These applications concern in particular *design problems*, where answers are required to questions like: Is the capacity sufficient?, What should be the layout? or How do we have to divide work among several capacities? An explanation why queueing models are important is that in many applications the variability in the arrival and service process are essential to the behavior of the system.

4.6 Little's law

Little's law (see [30]) gives a very important relation between the mean sojourn time in the system $E(S)$ on one hand and the mean number of customers in the system $E(L)$ and the average number of customers entering the system per unit time, denoted by λ. Little's law states that

$$E(L) = \lambda E(S).$$

Intuitively this result can be understood as follows. Suppose that all customers pay 1 dollar per unit time while in the system. This money can be earned in two ways. The first possibility is to let pay all customers 'continuously' in time. Then the average reward earned by the system equals $E(L)$ dollar per unit time. The second possibility is to let customers pay 1 dollar per unit time for their residence in the system when they leave. In equilibrium, the average number of customers leaving the system per unit time is equal to the average number of customers entering the system. So the system earns an average reward of $\lambda E(S)$ dollar per unit time. Obviously, the system earns the same in both cases. For a rigorous proof, see [41].

4.7 PASTA property

For queueing systems with Poisson arrivals, so for $M/\cdot/\cdot$ systems, the very special property holds that arriving customers find on the average the same situation in the queueing system as someone who is passing by (and not entering) at a random moment. In general this is not true (in fact, it is only true for Poisson arrivals). For instance, in a $D/D/1$ system which is empty at time 0, and with arrivals at $1, 3, 5, \ldots$ and service times 1, every arriving customer finds an empty system, whereas someone who is passing by at random finds with probability $1/2$ one customer in the system and with probability $1/2$ an empty system.

This property of Poisson arrivals is called PASTA property, which is the acrynom for Poisson Arrivals See Time Averages. Clearly Time Average is what is seen by someone passing by at random.

A rigorous proof of this property can be found in [52, 53].

4.8 Residual service time

In an $M/G/1$ system an arriving customer may find several customers waiting for service and one in service. What is the residual service time of the one in service?

If customers arrive according to a Poisson proces, the density of the service time of the job that is being served for a customer, who arrives when the server is busy, is equal to

$$f(x) = \frac{x f_X(x)}{\int_0^\infty y f_X(y) dy}, \qquad x > 0,$$

where $f_X(x)$ is the density of the service time distribution. So the mean service time for the job in service is equal to $E(X^2)/E(X)$. Since a customer arrives on the average in the middle of a service time, we find for the mean residual service time of the one in service

$$E(R) = \frac{E(X^2)}{2E(X)} = \frac{1}{2}(1 + c_X^2)E(X).$$

4.9 Occupation rate

In an $G/G/1$ system with arrival rate λ and mean service time $E(X)$ the amount of work arriving per unit time equals $\lambda E(X)$. The server can handle 1 unit work per unit time. So $\lambda E(X)$ must be less than 1, since otherwise the amount of work in the system will grow to infinity. It is common to use the notation

$$\rho = \lambda E(X)$$

where ρ is called the *occupation rate* or server utilization. In a multi server system $G/G/c$ we have to require $\lambda E(X) < c$. The occupation rate per server is $\rho = \lambda E(X)/c$ and the mean number of busy servers equals $\lambda E(X)$.

4.10 Exercises

Exercise 4.10.1
Show that the coefficient of variation of a hyperexponential distribution is greater or equal to 1.

Exercise 4.10.2
Let $X_1, \ldots, X_n$ be independent exponential random variables with mean $E(X_i) = 1/\mu_i$, $i = 1, \ldots, n$. Define

$$Y_n = \min(X_1, \ldots, X_n), \qquad Z_n = \max(X_1, \ldots, X_n).$$

Determine the distributions of Y_n and Z_n.

Exercise 4.10.3
Let $X_1, X_2, \ldots$ be independent exponential random variables with mean $1/\mu$ and let N be geometrically distributed:

$$P(N = k) = (1 - p)p^{k-1}$$

with $k = 1, 2, \ldots$ and $0 \leq p < 1$. Show that S defined as

$$S = \sum_{n=1}^{N} X_n$$

is again exponentially distributed with parameter $(1 - p)\mu$.

Exercise 4.10.4 *(Poisson process)*
Suppose that arrivals occur at $T_1, T_2, \ldots$. The interarrival times $A_n = T_n - T_{n-1}$ are independent and have common exponential distribution with mean $1/\lambda$, where $T_0 = 0$ by convention. Let $N(t)$ denote the number of arrivals is $[0, t]$ and define for $n = 0, 1, 2, \ldots$

$$p_n(t) = P(N(t) = n), \qquad t > 0.$$

(i) Determine $p_0(t)$.

(ii) Show that for $n = 1, 2, \ldots$

$$p_n'(t) = -\lambda p_n(t) + \lambda p_{n-1}(t), \qquad t > 0,$$

with initial condition $p_n(0) = 0$.

(iii) Solve the preceding initial value problem for $n = 1, 2, \ldots$

Exercise 4.10.5 *(Fitting a distribution)*
Suppose that processing a job on a certain machine takes on the average 4 minutes with a standard deviation of 3 minutes. Show that if we model the processing time as a mixture of an exponential distribution and an Erlang-2 distribution with density

$$f(t) = p\mu e^{-\mu t} + (1-p)\mu^2 t e^{-\mu t},$$

the parameters p and μ can be chosen in such a way that this distribution matches the mean and standard deviation of the processing times on the machine.

Exercise 4.10.6
Consider a random variable X with a Coxian-2 distribution with parameters μ_1 and μ_2 and branching probability p_1.

(i) Show that we can assume without loss of generality that $\mu_1 \geq \mu_2$.

(ii) Show that $c_X^2 \geq 0.5$.

Exercise 4.10.7
Let X and Y be exponentials with parameters μ and λ, respectively. Suppose that $\lambda < \mu$. Let Z be equal to X with probability λ/μ and equal to $X + Y$ with probability $1 - \lambda/\mu$. Show that Z is an exponential with parameter λ.

Exercise 4.10.8
Consider a H_2 distribution with parameters $\mu_1 > \mu_2$ and branching probabilities q_1 and q_2, respectively. Show that the C_2 distribution with parameters μ_1 and μ_2 and branching probability p_1 given by

$$p_1 = 1 - (q_1\mu_1 + q_2\mu_2)/\mu_1\,,$$

is equivalent to the H_2 distribution.

Chapter 5

Elementary queueing problems

In this chapter we will treat some elementary queueing models. In these models we assume that both the interarrival and service times are exponentially distributed. This assumption allows for a very simple state description, namely the number of jobs in the systems. We do not have to remember when the previous job arrived nor do we have to register when the jobs went into service. Since the exponential distribution is memoryless (see 2.6.1), this information does not yield a better prediction of the future.

5.1 $M/M/1/N$ queue

We first consider the $M/M/1$ queue with $N-1$ waiting positions in its queue. Let us think of a machine processing jobs one at a time and in order of arrival. In front of the machine there is a buffer offering space to at most $N-1$ jobs. The processing times are exponential with mean $1/\mu$ and jobs arrive according to a Poisson stream with rate λ. Jobs for which there is no space in the buffer at the moment they arrive are processed somewhere else.

In the following subsection we will first study the time-dependent behavior of the system. Next we consider the limiting behavior.

5.1.1 Time-dependent behavior

The state of the system at time t can be described by the number of jobs in the system (i.e. the jobs in the buffer and the one being processed). Let $p_n(t)$ be the probability that at time t there are n jobs in the system, $n = 0, 1, \ldots, N$. In the same way as in section 2.6 we obtain the following set of differential equations for the probabilities $p_n(t)$.

$$
\begin{aligned}
p_0'(t) &= -\lambda p_0(t) + \mu p_1(t), \\
p_n'(t) &= \lambda p_{n-1}(t) - (\lambda+\mu) p_n(t) + \mu p_{n+1}(t), \qquad n = 1, \ldots, N-1, \\
p_N'(t) &= \lambda p_{N-1}(t) - \mu p_N(t),
\end{aligned}
\tag{5.1}
$$

or in vector-matrix notation,

$$
p'(t) = p(t)Q, \tag{5.2}
$$

where

$$p(t) = (p_0(t), \ldots, p_N(t)),$$

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots & 0 & 0 & 0 \\ \mu & -(\lambda+\mu) & \lambda & 0 & \cdots & 0 & 0 & 0 \\ 0 & \mu & -(\lambda+\mu) & \lambda & \cdots & 0 & 0 & 0 \\ \vdots & & & & & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \mu & -(\lambda+\mu) & \lambda \\ 0 & 0 & 0 & 0 & \cdots & 0 & \mu & -\mu \end{pmatrix}. \tag{5.3}$$

At time 0 we assume that the system is empty. So the initial condition for (5.2) is

$$p(0) = (1, 0, \ldots, 0). \tag{5.4}$$

Equation (5.2) is a system of linear differential equations with initial-value condition (5.4). We have seen in section (3.3) that the general solution of (5.2) can be expressed in terms of the eigenvalues and eigenvectors of Q. The following lemma gives us information on the eigenvalues of Q.

Lemma 5.1.1 *The eigenvalues of Q are all real and simple; one eigenvalue is 0 and all other eigenvalues are negative and greater than or equal to $-2(\lambda+\mu)$.*

Denote these eigenvalues by $\xi_1, \ldots, \xi_N$ with $\xi_1 = 0$. We remark from a computational point of view that the eigenvalues can be easily found. They are characterized as the N roots in the interval $[-2(\lambda+\mu), 0]$ of characteristic equation

$$\mathrm{Det}(Q - \xi I) = 0,$$

where I is the $N \times N$ identity matrix. So the roots may be found by, e.g., simple bisection. Denote the associated row eigenvectors by $y_1, \ldots, y_N$. For given eigenvalue ξ_j the eigenvector y_j is a non-zero solution of the second order recurrence relation

$$\begin{aligned} -\lambda y_{j,0} + \mu y_{j,1} &= \xi_j y_{j,0}, \\ \lambda y_{j,n-1} - (\lambda+\mu) y_{j,n} + \mu y_{j,n+1} &= \xi_j y_{j,n}, \qquad n = 1, \ldots, N-1, \\ \lambda y_{j,N-1} - \mu y_{j,N} &= \xi_j y_{j,N}. \end{aligned}$$

If we set $y_{j,0} = 1$, then the eigenvector y_j is given by

$$y_{j,n} = a_j x_{j,1}^n + (1 - a_j) x_{j,2}^n, \qquad n = 0, 1, \ldots, N,$$

where $x_{j,1}$ and $x_{j,2}$ are the two roots of the quadratic equation

$$(\lambda + \mu + \xi_j) x = \lambda + \mu x^2,$$

and

$$a_j = \frac{x_{j,2} - (\lambda + \xi_j)/\mu}{x_{j,2} - x_{j,1}}.$$

Note that the eigenvector for $\xi_1 = 0$ simplifies to

$$y_{1,n} = \left(\frac{\lambda}{\mu}\right)^n, \qquad n = 1, 2, \ldots, N.$$

Once the eigenvalues and eigenvectors are known, the general solution of (5.2) can be written as

$$p(t) = c_1 y_1 + c_2 y_2 e^{\xi_2 t} + \ldots c_N y_N e^{\xi_N t}, \qquad t > 0. \tag{5.5}$$

To complete the solution the coefficients $c_1, \ldots, c_N$ have to be determined from the initial condition (5.4).

5.1.2 Limiting behavior

We now investigate what happens as t tends to infinity. From (5.5) and lemma (5.1.1) we directly obtain, by letting t tend to infinity,

$$p_n = \lim_{t\to\infty} p_n(t) = c_1 y_1 = \frac{(1-\rho)\rho^n}{1-\rho^{N+1}}, \qquad n = 1, 2, \ldots, N,$$

where $\rho = \lambda/\mu$. Remark that ρ is not the machine utilization, since part of the work is done somewhere else. From this distribution we find for the mean number of jobs in the system

$$E(L) = \sum_{n=0}^{N} n p_n = \frac{1 + N\rho^{N+1} - (N+1)\rho^N}{(1-\rho)(1-\rho^{N+1})}.$$

By using Little's law we obtain for the mean time spent in the system by an order that is processed by the machine (and not somewhere else)

$$E(S) = \frac{E(L)}{(1-p_0)\mu}.$$

5.2 $M/M/1$ queue

In this section we consider the case of an infinite buffer. Let, as in the previous section, λ denote the arrival rate and μ the service rate. We have to require

$$\rho = \frac{\lambda}{\mu} < 1,$$

since, otherwise, the amount of work in the system will explode. The quotient ρ is the utilization of the machine. In the next subsection we start to look at the time-dependent behavior of the system.

5.2.1 Time-dependent behavior

Let $p_n(t)$ denote again the probability that at time t there are n jobs in the system, $n = 0, 1, 2, \ldots$ They satisfy the following infinite set of differential equations.

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t) + \mu p_1(t), \\ p_n'(t) &= \lambda p_{n-1}(t) - (\lambda+\mu)p_n(t) + \mu p_{n+1}(t), \qquad n = 1, 2, \ldots, \end{aligned} \tag{5.6}$$

or in vector-matrix notation,

$$p'(t) = p(t)Q,$$

where

$$p(t) = (p_0(t), p_1(t), p_2(t), \ldots),$$

and Q is a countably infinite matrix given by

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & & & & \\ \mu & -(\lambda+\mu) & \lambda & 0 & \cdots & & & \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \cdots & & \\ \vdots & 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \cdots & \\ & \vdots & & \vdots & & \vdots & & \end{pmatrix}.$$

We now have an *infinite system of differential equations*, and therefore we cannot apply the theory of chapter 3. An attempt to analyze this infinite system is to let N tend to infinity in solution (5.5) for the finite buffer model. However, this limit is very tricky and complicated. We may end up with an infinite set of values ξ for which $Q - \xi I$ is not invertible. This set can be countably infinite, or even continuous, and the resulting solution may be expressed as an infinite sum, or an integral. An explicit solution for the probabilities $p_n(t)$ can be found in [27]. The expression presented there (see p. 77) is an infinite sum of modified Bessel functions. So already one of the simplest interesting queueing models leads to an ugly expression for the time-dependent behavior of its state probabilities. For more general systems we can only expect more complexity. Therefore, in the remainder we will focus on the *limiting or equilibrium behavior* of queueing systems.

5.2.2 Limiting behavior

One may show that as $t \to \infty$, then $p_n'(t) \to 0$ and $p_n(t) \to p_n$. So there exists a limiting distribution. Hence, from (5.6) it follows that the limiting (or equilibrium) probabilities p_n satisfy the recurrence relations

$$0 = -\lambda p_0 + \mu p_1, \tag{5.7}$$

$$0 = \lambda p_{n-1} - (\lambda+\mu)p_n + \mu p_{n+1}, \qquad n = 1, 2, \ldots \tag{5.8}$$

It is also possible to derive these equations directly from a so-called *flow diagram*, as shown in figure 5.1.

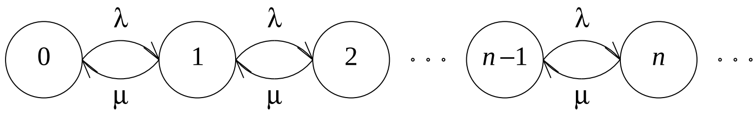

Figure 5.1: Flow diagram for the $M/M/1$ model

The arrows indicate possible transitions. The rate at which a transition occurs is λ for a transition from n to $n+1$ (an arrival) and μ for a transition from $n+1$ to n (a departure). The number of transitions per unit time from n to $n+1$, which is also called the *flow* from n to $n+1$, is equal to p_n, the fraction of time the system is in state n, times λ, the rate at arrivals occur while the system is in state n. The equilibrium equations (5.7) and (5.8) follow by equating the flow out of state n and the flow into state n.

The equations (5.8) are a second order recurrence relation with constant coefficients. Its general solution is of the form

$$p_n = c_1 x_1^n + c_2 x_2^n, \qquad n = 0, 1, 2, \ldots \tag{5.9}$$

where x_1 and x_2 are roots of the quadratic equation

$$\lambda - (\lambda + \mu)x + \mu x^2 = 0.$$

This equation has two zeros, namely $x = 1$ and $x = \lambda/\mu = \rho$. So all solutions to (5.8) are of the form

$$p_n = c_1 + c_2 \rho^n, \qquad n = 0, 1, 2, \ldots$$

There is in fact, besides (5.7), a second boundary condition, namely the condition that the sum of all probabilities is equal to 1. This condition of course directly implies that c_1 must be equal to 0. That c_1 must be equal to 0 also follows from (5.7) by substituting the solution (5.9) into (5.7).

The limiting or equilibrium distribution finally follows from the normalization equation, stating that the sum of all probabilities is equal to 1. So

$$p_n = (1 - \rho)\rho^n, \qquad n = 0, 1, 2, \ldots \tag{5.10}$$

Remark 5.2.1 *(alternative solution approaches)*
For this simple model there are several ways to determine the solution of the equations (5.8) and (5.7). We mention the following ones.

(i) *Recursion.* One can use (5.7) to express p_1 in p_0 and (5.8) to recursively express all other probabilities in p_0. Normalization yields (5.10).

(ii) *Generating function.* Define the probability generating function

$$P(z) = \sum_{n=0}^{\infty} p_n z^n,$$

which is properly defined for z with $|z| \leq 1$. By multiplying the nth equilibrium equation with z^n and then summing the equations over all n, the equilibrium equations for p_n can be transformed to a single equation for $P(z)$ which can be easily solved. The transformation back gives the solution (5.10).

(iii) *Global balance principle.* This principle states that for each set of states A, the flow out of set A is equal to the flow into that set. In fact, the equilibrium equations follow by applying this principle to a single state. However if we apply the balance principle to the set $A = \{0, 1, \dots, n-1\}$ we get

$$\lambda p_{n-1} = \mu p_n, \qquad n = 1, 2, \dots$$

This relation can be trivially solved yielding the solution (5.10).

From the equilibrium probabilities we can derive expressions for the mean number of jobs in the system and the mean time spent in the system. For the first one we get

$$E(L) = \sum_{n=0}^{\infty} n p_n = \frac{\rho}{1-\rho},$$

and by applying Little's law,

$$E(S) = \frac{1/\mu}{1-\rho}. \tag{5.11}$$

The mean waiting time follows from $E(S)$ by subtracting the mean processing time, so

$$E(W) = E(S) - 1/\mu = \frac{\rho/\mu}{1-\rho}.$$

It is also possible to derive the distribution of the waiting time. Denote by N_a the number of jobs in the system just before the arrival of a job and let X_m be the service time of the mth job in the buffer. Then we have

$$P(W > t) = P(\sum_{m=1}^{N_a} X_m > t).$$

By conditioning on N_a and using that N_a and X_m are independent it follows that

$$P(W > t) = \sum_{n=0}^{\infty} P(\sum_{m=1}^{n} X_m > t) P(N_a = n). \tag{5.12}$$

The problem is to find the probability that an arriving job finds n jobs in the system. The average number of jobs per unit time finding on arrival n jobs in the system is equal to λp_n. Dividing this number by the average number of jobs arriving per unit time gives the desired fraction. So

$$P(N_a = n) = \frac{\lambda p_n}{\lambda} = p_n. \tag{5.13}$$

In words, the fraction of jobs finding on arrival n jobs in the system is equal to the fraction of time there are n jobs in the system. This is known as the PASTA property (see section 4.7). Substituting (5.13) in (5.12) and using that $\sum_{m=1}^{n} X_m$ is Erlang distributed leads after some calculations to (cf. exercise 4.10.3)

$$P(W > t) = \rho e^{-\mu(1-\rho)t}, \qquad t > 0.$$

The probability that one does not have to wait is equal to $p_0 = 1 - \rho$. So the *conditional waiting time* which is the waiting time given that it is greater than zero is exponentially distributed with parameter $\mu(1-\rho)$.

In fact, the means $E(L)$ and $E(S)$ can also be directly determined, i.e. without knowing the probabilities $p(n)$, by combining Little's law and the PASTA property. Based on PASTA we know that the average number of jobs in the system seen an arriving job equals $E(L)$ and each of them (also the one being processed) has a (residual) processing time with mean $1/\mu$. The job further has to wait for its own processing time. Hence

$$E(S) = E(L)\frac{1}{\mu} + \frac{1}{\mu}$$

and together with

$$E(L) = \lambda E(S)$$

we find expression (5.11).

If we look at the expressions for $E(L)$ and $E(S)$ we see that both quantities grow to infinity as ρ approaches unity. The dramatic behavior is caused by the variation in the arrival and service process. This type of behavior with respect to ρ is characteristic for almost every queueing system. In table 5.1 we list for increasing values of ρ the mean mean waiting time and some waiting time probabilities.

We can conclude that the server utilization and the variation in the arrival and service process are *essential* to the behavior of the system.

ρ	$E(W)$		$P(W > t)$		
		t	5	10	20
0.5	1		0.04	0.00	0.00
0.8	4		0.29	0.11	0.02
0.9	9		0.55	0.33	0.12
0.95	19		0.74	0.58	0.35

Table 5.1: Performance characteristics for the $M/M/1$ with mean processing time 1

5.3 $M/M/c$ queue

In this section we consider a production system with c parallel identical machines and a *central buffer*. So jobs are not assigned to machines immediately on arrival, but there is a buffer for all machines together. Jobs are processed in order of arrival. As before we assume that jobs arrive according to a Poisson stream with rate λ. The processing times are exponentially distributed with mean $1/\mu$. We now need

$$\rho = \frac{\lambda}{c\mu} < 1$$

to guarantee that the system can handle the amount of work offered per unit time. The quantity ρ is the occupation rate of the machines.

The state of the system is completely characterized by the number of jobs in the system. Let p_n denote the equilibrium probability that there are n jobs in the system. Similar as for the $M/M/1$ we can derive the equilibrium equations for the probabilities p_n from the flow diagram shown in figure 5.2.

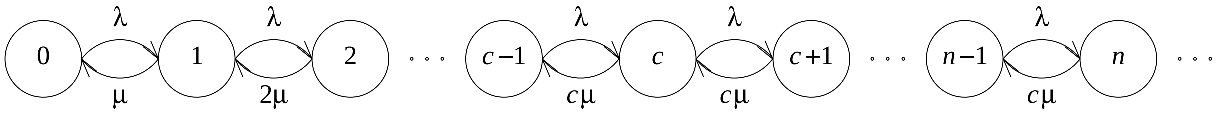

Figure 5.2: Flow diagram for the $M/M/c$ model

Instead of equating the flow into and out of a single state n, we get simpler equations by equating the flow out of and into the set of states $\{0, 1, \ldots, n-1\}$. This in fact amounts to equating the flow between the two neighboring states $n-1$ and n yielding

$$\lambda p_{n-1} = \min(n, c)\mu p_n, \qquad n = 1, 2, \ldots$$

Iterating gives

$$p_n = \frac{(c\rho)^n}{n!} p_0, \qquad n = 0, \ldots, c$$

and

$$p_{c+n} = \frac{(c\rho)^{c+n}}{c!c^n} p_0, \qquad n = 0, 1, 2, \ldots$$

The probability p_0 follows from normalization. This gives

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{(c\rho)^n}{c!} + \frac{(c\rho)^c}{c!(1-\rho)} \right)^{-1}.$$

An important quantity is the probability that a job has to wait. Denote this probability by Π_W. It is usually referred to as the *delay probability* and it is given by

$$\Pi_W = \frac{(c\rho)^c}{c!(1-\rho)} \left(\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!(1-\rho)} \right)^{-1}.$$

The conditional waiting time can be easily determined. The probability that an arriving job finds n jobs waiting in the buffer, given that all machines are busy on arrival, is equal to $(1-\rho)\rho^n$. Namely, by PASTA, this probability is equal to p_{c+n} divided by the sum of $p_c, p_{c+1}, \ldots$, which yields the preceding result. Since the buffer is cleared with rate $c\mu$, it follows that

$$E(W|W>0) = \sum_{n=0}^{\infty} \frac{n+1}{c\mu}(1-\rho)\rho^n = \frac{1}{1-\rho}\frac{1}{c\mu}.$$

Hence the mean (unconditional) waiting time equals

$$E(W) = \Pi_W E(W|W>0) = \Pi_W \frac{1}{1-\rho}\frac{1}{c\mu}.$$

Table 5.2 lists the delay probability and the mean waiting time in an $M/M/c$ with mean processing time 1 for $\rho = 0.9$.

c	Π_W	$E(W)$
1	0.90	9.00
2	0.85	4.26
5	0.76	1.53
10	0.67	0.67
20	0.55	0.28

Table 5.2: Performance characteristics for the $M/M/c$ with $\mu = 1$ and $\rho = 0.9$

We see that the delay probability slowly decreases as c increases. The mean waiting time however decreases fast (a little faster than $1/c$). One can also look somewhat differently at the performance of the system. We do not look at the occupation rate of a machine, but at the average number of idle machines. Let us call this the surplus capacity. Table 5.3 shows for fixed surplus capacity (instead of for fixed occupation rate as in the previous table) and c varying from 1 to 20 the mean waiting time and the mean number of jobs in the system.

c	ρ	$E(W)$	$E(L)$
1	0.90	9.00	10
2	0.95	9.26	19
5	0.98	9.50	51
10	0.99	9.64	105
20	0.995	9.74	214

Table 5.3: Performance characteristics for fixed surplus capacity of 0.1 machine

Although the mean number of jobs in the system sharply increases, the mean waiting time remains nearly constant.

5.4 Exercises

Exercise 5.4.1
Prove lemma (5.1.1).

Exercise 5.4.2 *(bulk arrivals)*
In a work station orders arrive according to a Poisson arrival process with arrival rate λ. An order consists of a number of independent jobs. The number N of jobs in an order is geometrically distributed:

$$P(N = k) = (1 - p)p^{k-1}$$

with $k = 1, 2, \ldots$ and $0 \le p < 1$. Each job requires an exponentially distributed amount of processing time with mean $1/\mu$.

(i) Derive the distribution of the workload of a job.

(ii) Determine the distribution of the number of orders in the system.

Exercise 5.4.3 *(variable production rate)*
Consider a work station where jobs arrive according to a Poisson process with arrival rate λ. The jobs have an exponentially distributed service time with mean $1/\mu$. So the service completion rate (the rate at which jobs depart from the system) is equal to μ.
If the queue length drops below the threshold Q_L the service completion rate is lowered to μ_L. If the queue length reaches Q_H, where $Q_H \ge Q_L$, the service rate is increased to μ_H. (L stands for low, H for high.) Determine the queue length distribution and the mean time spent in the system.

Exercise 5.4.4 *(a fast and a slow machine)*
Consider two parallel machines with a common (infinite) buffer where jobs arrive according to a Poisson stream with rate λ. The processing times are exponentially distributed with mean $1/\mu_1$ on machine 1 and $1/\mu_2$ on machine 2 ($\mu_1 > \mu_2$). Jobs are processed in order of arrival. A job who arrives when both machines are idle is assigned to the fast machine. To guarantee that the system can handle the amount of work offered per unit time we assume $\rho < 1$ where

$$\rho = \frac{\lambda}{\mu_1 + \mu_2}\,.$$

(i) Determine the distribution of the number of jobs in the system.

(ii) Use this distribution to derive the mean number of jobs in the system.

(iii) When is it better to not use the slower machine at all?

(iv) Calculate for the following two cases the mean number of jobs in the system with and without the slow machine.

(a) $\lambda = 2$, $\mu_1 = 5$, $\mu_2 = 1$;

(b) $\lambda = 3$, $\mu_1 = 5$, $\mu_2 = 1$.

For more details concerning this problem the interested reader is referred to [35]. There it is shown that one should not remove the slow machine if $r > 0.5$ where $r = \mu_2/\mu_1$. When $0 \leq r < 0.5$ the slow machine should be removed (and the resulting system is stable) whenever $\rho \leq \rho_c$, where

$$\rho_c = \frac{2 + r^2 - \sqrt{(2+r^2)^2 + 4(1+r^2)(2r-1)(1+r)}}{2(1+r^2)}.$$

Exercise 5.4.5 *(base-stock policy)*
For a given product one has a stock of N items in order to directly satisfy customer demand. The demand (for one item at a time) is a Poisson process with rate λ per week. When demand for an item occurs, one directly orders a new one at the supplier. This ordering policy is called *base-stock policy*. The lead time is exponentially distributed with mean $1/\mu$ weeks. Each demand which cannot be satisfied directly from stock on hand is lost.
This inventory system is completely characterized by the number of outstanding orders. Let p_n, $n = 0, \ldots, N$ denote the probability that there are n outstanding orders.

(i) Determine the probabilities p_n.

(ii) What is the fraction of customer demand which is lost?

(iii) Denote the fraction of customer which is lost when the (maximal) stock level is N by B_N. Derive a recurrence relation to determine $B_0, B_1, B_2, \ldots$

Suppose that $\lambda = \mu = 2$.

(iv) Determine the minimum stock level such that at least 90% of customer demand can be satisfied from stock on hand.

Exercise 5.4.6 *(mean clearance time)*
Consider a work station where jobs arrive according to a Poisson process with rate λ. The jobs have an exponentially distributed service time with mean $1/\mu$. Assume that $\lambda < \mu$. Let C_n be the time till the work station is empty again if there are now n jobs present in the system, $n = 0, 1, 2, \ldots$ (so $C_0 = 0$).

(i) Show that the means $E(C_n)$ satisfy the following recurrence relation,

$$E(C_n) = \frac{1}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}E(C_{n-1}) + \frac{\lambda}{\lambda+\mu}E(C_{n+1}), \qquad n = 1, 2, \ldots$$

(ii) Show that

$$E(C_1) = \frac{1}{\mu - \lambda}.$$

(iii) Solve the recursion relation for $E(C_n)$.

Chapter 6

Variations on the $M/M/1$ theme

In this chapter we will treat some variations of $M/M/1$ model. In the first section we consider a production system with setup times. Next we will analyze an unreliable production system. In the final section we revisit the problem with a fast and a slow machine which is introduced in exercise 5.4.4.

6.1 Machine with setup times

Consider a single machine where jobs are being processed in order of arrival. Since it is expensive to keep the machine in operation while there are no jobs, the machine is turned off as soon as the system is empty. When a new job arrives the machine is turned on again, but it takes some time till the machine is ready for processing. So turning off the machine leads to longer production leadtimes. The question arises whether it is better to keep the machine continuously in operation and taking the extra costs for granted. A simple model for answering this question will be formulated and studied below.

Suppose that the jobs arrive according to a Poisson stream with rate λ and that the processing times are exponential with mean $1/\mu$. For stability we require that $\rho = \lambda/\mu < 1$. The setup time of the machine is also exponential with mean $1/\theta$. Due to the exponential assumptions the state of the system can be characterized by the pair (n, m) where n is the number of jobs in the system and m denotes the state of the machine. If $m = 1$ the machine is in operation, if $m = 0$ the machine is off or in its setup phase. The flow diagram of this system is shown in figure 6.1.

By equating in each state the flow out of and into that state we obtain the following set of balance (or equilibrium) equations.

$$
\begin{aligned}
p(0,0)\lambda &= p(1,1)\mu, && (6.1)\\
p(n,0)(\lambda+\theta) &= p(n-1,0)\lambda, \qquad n = 1,2,\ldots && (6.2)\\
p(1,1)(\lambda+\mu) &= p(1,0)\theta + p(2,1)\mu, && (6.3)\\
p(n,1)(\lambda+\mu) &= p(n,0)\theta + p(n+1,1)\mu + p(n-1,1)\lambda, \qquad n = 2,3,\ldots && (6.4)
\end{aligned}
$$

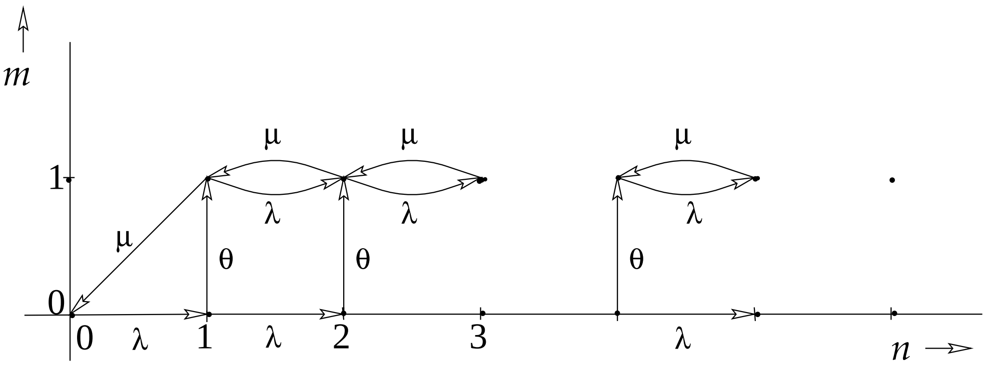

Figure 6.1: Flow diagram for the $M/M/1$ model with setup times

These balance equations can be solved as follows. Equation (6.2) is a first order difference equation which has as solution

$$p(n,0) = p(0,0)x^n, \qquad n = 0, 1, \ldots \tag{6.5}$$

with

$$x = \frac{\lambda}{\lambda + \theta}.$$

Note that ρ is the fraction of time the machine is processing jobs. Hence the fraction of time the machine is not processing jobs is given by

$$\sum_{n=0}^{\infty} p(n,0) = 1 - \rho,$$

whence with (6.5),

$$p(0,0) = (1-\rho)(1-x).$$

To find the probabilities $p(n,1)$ we substitute (6.5) into (6.4) yielding

$$p(n,1)(\lambda + \mu) = (1-\rho)(1-x)x^n\theta + p(n+1,1)\mu + p(n-1,1)\lambda. \tag{6.6}$$

This is a second order *inhomogeneous* recurrence relation. Its general solution is given by

$$p(n,1) = A(1-\rho)(1-x)x^n + B_1 y_1^n + B_2 y_2^n, \qquad n = 1, 2, \ldots \tag{6.7}$$

The first term is a particular solution to the inhomogeneous relation, where

$$A = \frac{\theta x}{(\lambda+\mu)x - \mu x^2 - \lambda} = \frac{\lambda + \theta}{\mu - \lambda - \theta}.$$

For convenience we assume $\theta \neq \mu - \lambda$, so that the denominator above does not vanish. In case this assumption does not hold the particular solution is of the form nx^n, see exercise

6.4.1. The second term is the general solution of the homogeneous equation in which the coefficients B_1 and B_2 are free and y_1 and y_2 are the roots of

$$(\lambda + \mu)y = \mu y^2 + \lambda.$$

Hence $y_1 = \lambda/\mu = \rho$ and $y_2 = 1$. Since the probabilities $p(n, m)$ have to add up to one, we have to set $B_2 = 0$. Finally substitution of (6.5) and (6.7) into equation (6.3) yields $B_1 = -A$. Remark that since the equilibrium equations are dependent the remaining equation in (0,0) is also satisfied. This completes the solution of the balance equations. Summarizing we have found

$$\begin{aligned} p(n,0) &= (1-\rho)(1-x)x^n, \qquad n = 0, 1, \ldots \\ p(n,1) &= A(1-\rho)(1-x)(x^n - \rho^n), \qquad n = 1, 2, \ldots \end{aligned}$$

From the equilibrium distribution we can derive the mean number of jobs in the system, yielding

$$\begin{aligned} E(L) &= \sum_{n=1}^{\infty} n(p(n,0) + p(n,1)) \\ &= (1-\rho)(1-x)\left[\frac{x}{(1-x)^2} + \frac{Ax}{(1-x)^2} - \frac{A\rho}{(1-\rho)^2}\right], \end{aligned}$$

and by using Little's law we obtain the mean production leadtime $E(S)$. It is also possible to derive the distribution of the production leadtime. By conditioning on the state seen on arrival and using PASTA we get

$$\begin{aligned} P(S > t) &= \sum_{n=0}^{\infty} p(n,0)P(Y + X_1 + \ldots + X_{n+1} > t) \\ &+ \sum_{n=1}^{\infty} p(n,1)P(X_1 + \ldots + X_{n+1} > t), \end{aligned}$$

where Y is an exponential (residual) setup time with mean $1/\theta$ and $X_1, X_2, \ldots$ are exponential processing times with mean $1/\mu$. Substituting the expressions for $p(n, 0)$ and $p(n, 1)$ and using that a geometric sum of exponential random variables is again exponential (see exercise 4.10.3), we get

$$P(S > t) = \frac{1}{\mu - \lambda - \theta}\,[\mu(1-\rho)e^{-\theta t} - \theta e^{-\mu(1-\rho)t}]. \tag{6.8}$$

Now we have seen how to calculate, for instance, the mean and distribution of the production lead time, one should be able to decide what to do: keep the machine continuously in operation or to turn it off when there is no work.

The means $E(L)$ and $E(S)$ can also be determined by combining the PASTA property and Little's law, in the same way as we have seen for the $M/M/1$ system. Based on PASTA we know that the average number of jobs in the system seen an arriving job equals $E(L)$ and each of them (also the one being processed) has a (residual) processing time with mean

$1/\mu$. With probability $1-\rho$ the machine is not in operation on arrival, so that the job also has to wait for the setup phase with mean $1/\theta$. Further the job has to wait for its own processing time. Hence

$$E(S) = (1-\rho)\frac{1}{\theta} + E(L)\frac{1}{\mu} + \frac{1}{\mu}$$

and together with

$$E(L) = \lambda E(S)$$

we find

$$E(S) = \frac{1/\mu}{1-\rho} + \frac{1}{\theta}\,.$$

So the mean setup time is the extra mean delay caused by turning off the machine when there is no work. In fact, it can be shown (by using, e.g., a sample path argument) that the extra delay is an exponential time with mean $1/\theta$.

6.2 Unreliable machine

In this section we consider an unreliable machine processing jobs. The machine breaks down at random instants, whether it is processing or not. What is the impact of these breakdowns on the production leadtimes? To obtain some insight in the effects of the breakdowns we formulate and study the following simple model.

At the machine jobs arrive according to a Poisson stream with rate λ. The processing times are exponential with mean $1/\mu$. The time between two breakdowns is exponentially distributed with mean $1/\eta$. The repair time is also exponentially distributed with mean $1/\theta$.

The state of the system can be characterized by the pair (n, m) where n is the number of jobs in the system and m denotes the state of the machine. If $m = 1$ the machine is up, if $m = 0$ the machine is down and in repair. The flow diagram of this system is shown in figure 6.2.

What is the condition under which the machine can handle the amount of work offered per unit time? Let ρ_U be the fraction of time the machine is up, so

$$\rho_U = \frac{1/\eta}{1/\eta + 1/\theta}\,.$$

Then we have to require that

$$\frac{\lambda}{\mu} < \rho_U. \tag{6.9}$$

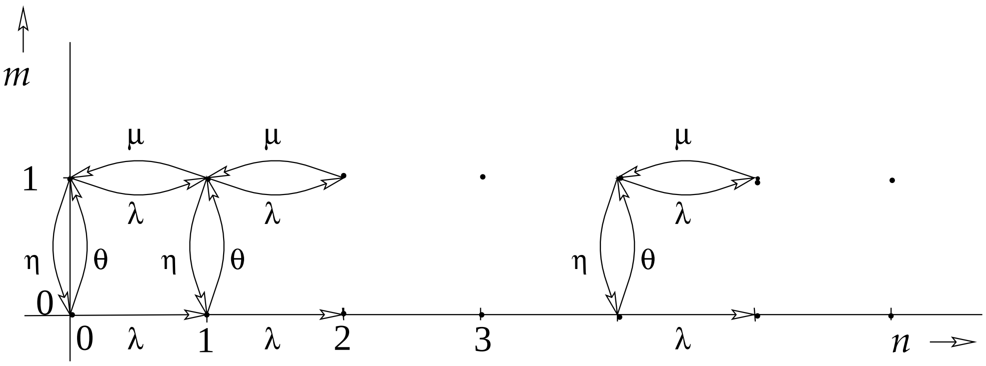

Figure 6.2: Flow diagram for the $M/M/1$ model with random breakdowns

From the flow diagram we obtain the following set of balance equations.

$$\begin{aligned}
p(0,0)(\lambda+\theta) &= p(0,1)\eta, && (6.10)\\
p(n,0)(\lambda+\theta) &= p(n-1,0)\lambda + p(n,1)\eta, \qquad n=1,2,\ldots && (6.11)\\
p(0,1)(\lambda+\eta) &= p(0,0)\theta + p(1,1)\mu, && (6.12)\\
p(n,1)(\lambda+\eta+\mu) &= p(n,0)\theta + p(n+1,1)\mu + p(n-1,1)\lambda, \quad n=1,2,\ldots && (6.13)
\end{aligned}$$

These balance equations can be solved as follows. We first try to find solutions of (6.11) and (6.13) of the form

$$\begin{aligned}
p(n,0) &= x^n u,\\
p(n,1) &= x^n v.
\end{aligned}$$

Substituting this form into (6.11) and (6.13) yields

$$\begin{aligned}
xu(\lambda+\theta) &= u\lambda + xv\eta,\\
xv(\lambda+\eta+\theta) &= ux\theta + x^2 v\mu + v\lambda.
\end{aligned}$$

Set $u = 1$. Then we get from the first equation

$$v = \frac{(\lambda+\theta)x-\lambda}{\eta x}. \qquad (6.14)$$

Substitution of this relation into the second equation leads to

$$x((\lambda+\theta)x-\lambda)(\lambda+\eta+\theta) = \eta\theta x^2 + ((\lambda+\theta)x-\lambda)\mu x^2 + ((\lambda+\theta)x-\lambda)\lambda$$

which can be rewritten as

$$(x-1)(\mu(\lambda+\theta)x^2 - \lambda(\lambda+\theta+\eta+\mu)x + \lambda^2) = 0.$$

The root $x = 1$ is not useful, since we must be able to normalize the solution. It can be easily shown that condition (6.9) implies that the quadratic equation above has two roots x_1 and x_2 satisfying $0 < x_1 < x_2 < 1$. Hence the solution of (6.11) and (6.13) is

$$\begin{aligned}
p(n,0) &= c_1 x_1^n + c_2 x_2^n, && (6.15)\\
p(n,1) &= c_1 v_1 x_1^n + c_2 v_2 x_2^n,
\end{aligned}$$

where v_i is given by (6.14) with $x = x_i$, $i = 1, 2$ and the coefficients c_1 and c_2 are to be determined yet. These coefficients follow from the boundary conditions (6.10) and (6.12) together with the normalization condition. Inserting (6.15) into (6.10) yields

$$\frac{c_1}{x_1} + \frac{c_2}{x_2} = 0.$$

When c_1 and c_2 satisfy this relation, it can be verified that equation (6.12) is also satisfied (in fact, this is due to the dependence of the balance equations). From the condition that the probabilities have to add up to one, we obtain

$$\frac{c_1(1+v_1)}{1-x_1} + \frac{c_2(1+v_2)}{1-x_2} = 1$$

The two equations for c_1 and c_2 have as solution

$$c_1 = Cx_1, \qquad c_2 = -Cx_2,$$

where

$$C = \left\{ \frac{x_1(1+v_1)}{1-x_1} - \frac{x_2(1+v_2)}{1-x_2} \right\}^{-1}.$$

This completes the determination of the probabilities $p(n, m)$. Based on the solution for $p(n, m)$ it is easy to find the mean number of jobs in the system,

$$E(L) = C\left(\frac{(1+v_1)x_1^2}{(1-x_1)^2} - \frac{(1+v_2)x_2^2}{(1-x_2)^2}\right),$$

and $E(S)$ follows by using Little's law.

In table 6.1 we investigate the impact of the variability in the availability of the machine on the mean production leadtime. The mean production time is 1 hour ($\mu = 1$). The average number of jobs that arrives in a week (40 hours) is 32, so $\lambda = 0.8$ jobs per hour. The fraction of time the machine is available is kept constant at 90%, so $\rho_U = 0.9$. The rate η at which the machine breaks down is varied from (on the average) every 10 minutes till once a week. In the former case the mean repair time is 1.1 minute, in the latter case it is more dramatic, namely nearly half a day (4.4 hours). The results show that the variation in the availability of the machine is essential to the behavior of the system.

6.3 A fast and a slow machine

We now revisit the problem studied in exercise 5.4.4. Consider two parallel machines with a common (infinite) buffer where jobs arrive according to a Poisson stream with rate λ. The processing times are exponentially distributed with mean $1/\mu_1$ on machine 1 and $1/\mu_2$ on machine 2 ($\mu_1 > \mu_2$). Jobs are processed in order of arrival. For stability we assume

$$\rho = \frac{\lambda}{\mu_1 + \mu_2} < 1$$

η	θ	$E(S)$
6	54	10.02
3	27	10.03
1	9	10.1
0.125	1.125	10.8
0.0625	0.5625	11.6
0.025	0.225	14

Table 6.1: The mean production leadtime in hours as a function of the break-down rate η per hour for fixed availability of 90%

In exercise 5.4.4 we looked at two extreme options: make always use of the slow machine, or do not use it at all. The choice between the two options depends on the workload and the ratio of the processing speeds of the machines. In this section we consider the same model except that jobs are allowed to wait for the busy fast machine at times when the slow one is idle. The problem is to find the best operating policy, i.e. a rule which tells you when to use the slow machine, so as to minimize the mean time spent in the system.

From the point of view of an individual job the decision problem is simple. If an arriving job finds an idle fast machine, he will join it. If he finds a busy fast machine with n jobs and an idle slow machine, then he will go to the slow machine if $(n+1)/\mu_1 > 1/\mu_2$ and otherwise, he will wait for the fast machine. This rule will minimize his expected time in the system. So the *individually optimal policy* is characterized by the integer part of μ_1/μ_2, say N. Jobs wait for the fast machine as long as the number of jobs in the system is less or equal to N.

However we are interested the *overall or socially optimal policy*, i.e., in assigning jobs to the machines so as to minimize the mean time in the system over all jobs. In [29, 46] a proof is given that the overall optimal policy is of the same type as the individual optimal policy, but presumable with a different threshold N. The issue of individually versus socially optimal policies is also studied in [42] for the problem of controlling the arrivals to a single-server queue.

Below we will study the model in which jobs are processed by the slow machine only when the number of jobs in the system is more than N. It will be shown that the equilibrium probabilities for this model, and also quantities such as the mean time spent in the system, can be easily computed. Based on these results the optimal threshold level N can readily be computed.

The state of the system can be described by the pair (n, m) where n is the number of jobs in the system and m denotes the state of machine 2 which is the slower one. If $m = 1$ machine 2 is busy, if $m = 0$ it is idle. The possible states are the pairs $(n, 0)$ with n running from 0 to N and the pairs $(n, 1)$ with $n = 1, 2, \ldots$ The flow diagram of this system is shown in figure 6.3.

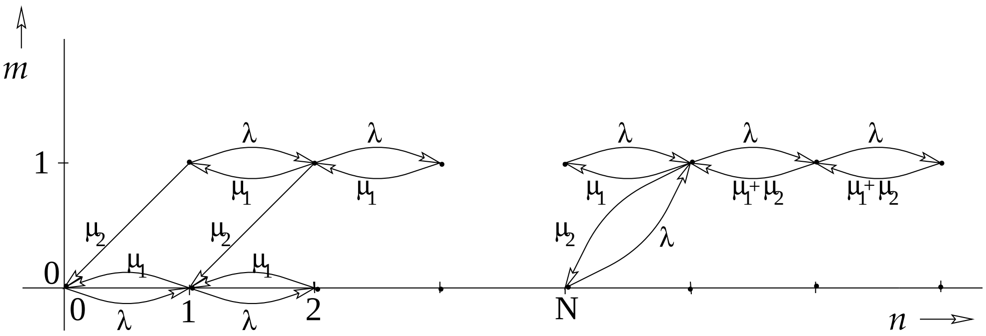

Figure 6.3: Flow diagram for the system with a fast and a slow machine

The equilibrium probabilities $p(n,m)$ satisfy the following set of equations.

$$
\begin{aligned}
p(0,0)\lambda &= p(1,0)\mu_1 + p(1,1)\mu_2, && (6.16)\\
p(n,0)(\lambda+\mu_1) &= p(n-1,0)\lambda + p(n+1,0)\mu_1 + p(n+1,1)\mu_2, && (6.17)\\
& \qquad n = 1,\ldots,N-1, &&\\
p(N,0)(\lambda+\mu_1) &= p(N-1,0)\lambda + p(N+1,1)\mu_2, && (6.18)\\
p(1,1)(\lambda+\mu_2) &= p(2,1)\mu_1, && (6.19)\\
p(n,1)(\lambda+\mu_1+\mu_2) &= p(n-1,1)\lambda + p(n+1,1)\mu_1, && (6.20)\\
& \qquad n = 2,\ldots,N, &&\\
p(N+1,1)(\lambda+\mu_1+\mu_2) &= p(N,1)\lambda + p(N+2,1)(\mu_1+\mu_2) + p(N,0)\lambda, && (6.21)\\
p(n,1)(\lambda+\mu_1+\mu_2) &= p(n-1,1)\lambda + p(n+1,1)(\mu_1+\mu_2), && (6.22)\\
& \qquad n = N+2, N+3, \ldots &&
\end{aligned}
$$

To solve these balance equations we first consider the equations (6.22). These equations form a second order recurrence relation the general solution of which is given by

$$p(n,1) = A\rho^n + B, \qquad n = N+1, N+2, \ldots \tag{6.23}$$

Since the probabilities $p(n,m)$ have to add up to one, we must set $B = 0$. The constant A is still free and it will be determined later on from the normalization condition. It now remains to solve the (finitely many) balance equations in the states (n,m) with $n \leq N$. Let us first look at the equations (6.20). The general solution of these equations is

$$p(n,1) = C_1 x_1^n + C_2 x_2^n, \qquad n = 1,2,\ldots,N+1, \tag{6.24}$$

where x_1 and x_2 are the two roots of

$$x(\lambda+\mu_1+\mu_2) = \lambda + x^2\mu_1.$$

To determine C_1, C_2 note that the forms (6.23) and (6.24) have to agree for $n = N+1$, so

$$C_1 x_1^{N+1} + C_2 x_2^{N+1} = A\rho^{N+1}. \tag{6.25}$$

Further we have to satisfy (6.19) yielding

$$(\lambda - \mu_1 x_1)C_1 + (\lambda - \mu_1 x_2)C_2 = 0. \tag{6.26}$$

Hence, from (6.25) and (6.26) we find

$$\begin{aligned} C_1 &= \frac{A\rho^{N+1}(\lambda - \mu_1 x_2)}{(\lambda - \mu_1 x_2)x_1^{N+1} - (\lambda - \mu_1 x_1)x_2^{N+1}}, \\ C_2 &= \frac{A\rho^{N+1}(\lambda - \mu_1 x_1)}{(\lambda - \mu_1 x_1)x_2^{N+1} - (\lambda - \mu_1 x_2)x_1^{N+1}}. \end{aligned} \tag{6.27}$$

The equations (6.17) form an second order recurrence relation in $p(n,0)$ with the probabilities $p(n+1,1)$ as inhomogeneous terms. The general solution is

$$p(n,0) = D_1\rho_1^n + D_2 - C_1 x_1^{n+1} - C_2 x_2^{n+1}, \qquad n = 0, 1, \ldots, N, \tag{6.28}$$

where $\rho_1 = \lambda/\mu_1$. The first two terms are the general solution of the homogeneous relation, the other ones are a particular solution of the inhomogeneous relation. The constants D_1 and D_1 follow form the boundary conditions (6.16) and (6.18). The first condition yields $D_2 = 0$ and the other one

$$D_1\rho_1^{N+1} = C_1 x_1^{N+2} + C_2 x_2^{N+2},$$

whence, by (6.27),

$$D_1 = A\,\frac{(\lambda - \mu_1 x_2)x_1^{N+2} - (\lambda - \mu_1 x_1)x_2^{N+2}}{(\lambda - \mu_1 x_2)x_1^{N+1} - (\lambda - \mu_1 x_1)x_2^{N+1}}\left(\frac{\rho}{\rho_1}\right)^{N+1}. \tag{6.29}$$

To complete the solution we have to solve A from the normalization condition, stating that all probabilities have to add up to one. Substituting the forms (6.23), (6.24) and (6.28) we get

$$A\,\frac{\rho^{N+2}}{1-\rho} + D_1\frac{1-\rho_1^{N+1}}{1-\rho_1} = 1.$$

Inserting the expression for D_1 it follows that

$$A = \left\{\frac{\rho^{N+2}}{1-\rho} + \frac{1-\rho_1^{N+1}}{1-\rho_1}\,\frac{(\lambda - \mu_1 x_2)x_1^{N+2} - (\lambda - \mu_1 x_1)x_2^{N+2}}{(\lambda - \mu_1 x_2)x_1^{N+1} - (\lambda - \mu_1 x_1)x_2^{N+1}}\left(\frac{\rho}{\rho_1}\right)^{N+1}\right\}^{-1}. \tag{6.30}$$

Summarizing we can conclude the equilibrium probabilities are given by

$$\begin{aligned} p(n,0) &= D_1\rho_1^n - C_1 x_1^{n+1} - C_2 x_2^{n+1}, \qquad n = 0, 1, \ldots, N, \\ p(n,1) &= C_1 x_1^n + C_2 x_2^n, \qquad n = 1, 2, \ldots, N, \\ p(n,1) &= A\rho^n, \qquad n = N+1, N+2, \ldots \end{aligned}$$

where the constants A, C_1, C_2, D_1 satisfy (6.30), (6.27) and (6.29), respectively. From these expressions for the equilibrium probabilities it is straightforward to derive an expression for, e.g., the mean number of jobs in the system and the mean time spent in the system.

In table 6.2 we list for fixed $\lambda = 1$ and varying ρ and μ_1/μ_2 the value of the threshold N minimizing the mean time spent in the system. The optimal threshold is found by a simple search procedure. Denote the mean time spent in the system with threshold N by $E(S(N))$. Then $E(S(N))$ is evaluated for increasing values of N until, for the first time, $E(S(N+1)) > E(S(N))$. When this happens the associated N is the optimal one and it will be denoted by N_0. The corresponding mean time spent in the system is denoted by $E(S(N_0))$. Having obtained the optimal system it is interesting to compare its performance with the two systems studied in exercise 5.4.4, i.e., the system which never uses the slow machine and the one which always uses both machines. The performance of the first system is indicated by $E(S_1)$, that of the second one by $E(S_2)$. Note that when $\rho_1 > 1$ the system with the fast machine only is unstable, and in table 6.2 this is indicated by $E(S_1) = \infty$.

ρ	μ_1/μ_2	N_0	$E(S(N_0))$	$E(S_1)$	$E(S_2)$
0.3	20	14	0.46	0.46	0.96
	10	7	0.49	0.49	0.80
	5	3	0.55	0.56	0.68
	3	2	0.60	0.67	0.63
0.5	20	10	1.10	1.11	1.72
	10	5	1.19	1.22	1.56
	5	3	1.28	1.50	1.41
	3	2	1.32	2.00	1.33
0.8	20	6	4.45	5.25	4.83
	10	3	4.49	7.33	4.72
	5	2	4.48	24.00	4.57
	3	1	4.48	∞	4.48
0.9	20	5	9.56	17.18	9.85
	10	3	9.56	99.00	9.75
	5	2	9.54	∞	9.61
	3	1	9.52	∞	9.52

Table 6.2: Optimal thresholds and system performance where in each example $\lambda = 1$

We see in table 6.2 that the performance of the optimal system does not change much as μ_1/μ_2 changes while ρ is fixed. Even when μ_1/μ_2 is large, 20 say, substantial savings can be achieved by using the slow machine properly. The practical lesson may be that one should not throw away an old machine when the new one is much faster. Also observe that for low values of ρ the system with the fast machine only is just as good as the optimal system, while for high values of ρ the system using both machines is just as good as the optimal one. Finally remark that since the optimal threshold N_0 is (usually) smaller than (the integer part of) μ_1/μ_2 the overall optimal system makes more use of the slow machine than the individually optimal system.

6.4 Exercises

Exercise 6.4.1
Consider the $M/M/1$ model with setup times in section 6.1. Suppose that $\theta = \lambda - \mu$. Show that

$$p(n, 1) = (1 - \rho)^2 n\rho^n, \qquad n = 1, 2, \ldots$$

is a particular solution of (6.6).

Exercise 6.4.2
Consider the $M/M/1$ model with setup times in section 6.1. Derive expression (6.8) for the probability that the production leadtime is longer than t.

Exercise 6.4.3
Consider a single machine which is turned off when there is no work. It is turned on and restarts work when enough orders, say N, arrived to the machine. The setup times are negligible. The processing times are exponentially distributed with mean $1/\mu$ and the average number of orders arriving per unit time is λ.

(i) How can you describe the state of the system?

(ii) Formulate the balance equations for the (equilibrium) state probabilities.

(iii) Solve the balance equations.

Suppose that $\lambda = 20$ orders per hour, $1/\mu = 2$ minutes and that the setup cost is 54 dollar. In operation the machine costs 12 dollar per minute. The waiting cost is 1 dollar per minute per order.

(iv) Show how for given threshold N the average cost per unit time can be determined.

(v) Compute the threshold N minimizing the average cost per unit time.

Exercise 6.4.4
Consider a channel for the transmission of data messages. The length of the messages are exponentially distributed. Hence also the transfer times are exponentially distributed, with mean $1/\mu$ say. The packets are generated by a source (for instance a computer system) which is alternatingly on and off. The length of an on (resp. off) period is exponentially distributed with parameter θ_1 (resp. θ_0). While the source is on messages are generated with rate λ.
The state of this system can be described by the pair (n, m) where n is the number of messages in the system (waiting for transfer including the one being transferred) and m is the state of the source. If $m = 0$ then the source is off, if $m = 1$ it is on and generating messages.

(i) Describe the flow diagram and give the set of equations for the equilibrium state probabilities.

(ii) Solve these equations and give an expression for the mean delay at the transmission channel.

Exercise 6.4.5 *(Maintenance of trains)*
Consider a maintenance facility for trains. In this facility there is one track for the maintenance of small trains which are used for local distances only. The track offers space to exactly two trains. It can be entered on one side only. This implies that when the train at the front is ready, then it has to wait until the one at the back is also ready before it can leave the track.
Assume that the trains arrive according to a Poisson stream with rate λ and that maintenance of a train takes an exponentially distributed time with mean $1/\mu$.

(i) How can the state of the maintenance facility be described?

(ii) Formulate the balance equations for the state probabilities.

(iii) Under which condition is the facility capable of handling the amount of work offered per unit time?

(iv) Solve the balance equations.

(v) Give an expression for the mean number of trains in the system.

Chapter 7

Non-exponential systems

In the previous chapters we considered queueing models in which both interarrival and processing times are exponential. In this chapter we will treat systems in which the processing times or the interarrival times (or both) are not exponential, but phase-type distributed. The use of phase-type distributions still allows a simple description of the system and the class is sufficiently rich for the modelling of random variables in many practical applications.

7.1 $M/E_r/1$ queue

We consider a single machine. Orders arrive according to a Poisson stream with rate λ and are processed in order of arrival. The processing times are Erlang-r distributed with mean r/μ (see section 4.2.2). So the processing times can be thought of as consisting of r independent exponential phases with mean $1/\mu$. The Erlang distribution can be used to model processing times with a low coefficient of variation (less than one), but it can also arise naturally. For instance, if an order has to pass, stage by stage, through a series of r independent production stages, where each stage takes a exponentially distributed time. For stability we have to require

$$\rho = \frac{\lambda r}{\mu} < 1. \tag{7.1}$$

There are two ways in which one can describe the state of the system: a natural way, giving a two-dimensional description, and another way that is found in most textbooks and leads to a one-dimensional state.

7.1.1 Natural description

The natural way to describe the state of a nonempty system is by the pair (n, k) where n denotes the number of jobs in the system and k the remaining number of phases to be processed for the job in service. The empty state can be denoted by 0. This leads to the flow diagram in figure 7.1.

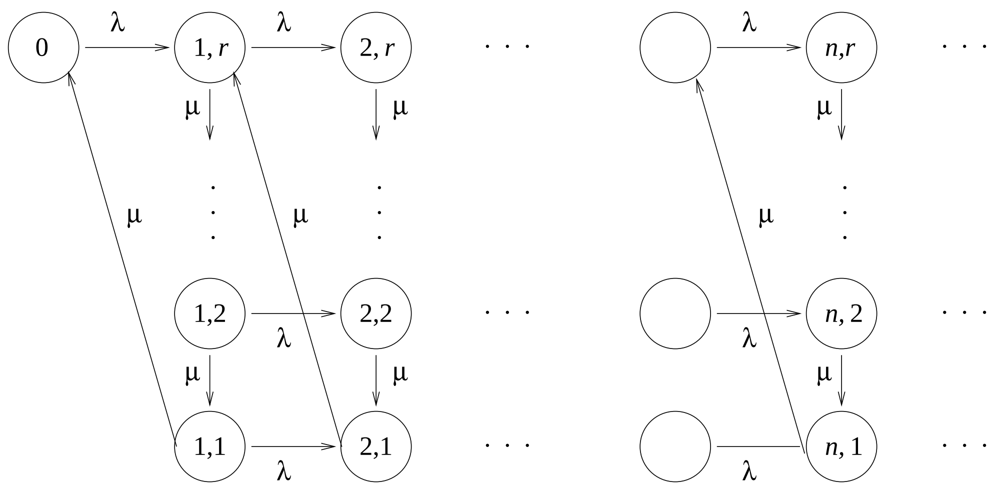

Figure 7.1: Two dimensional flow diagram for the $M/E_r/1$ model

Let $p(n,k)$ denote the probability that the system is in state (n,k). Then we can write down the following balance equations.

$$\begin{aligned}
p(n,k)(\lambda+\mu) &= p(n-1,k)\lambda + p(n,k+1)\mu, \qquad n>1, k<r, & (7.2)\\
p(n,r)(\lambda+\mu) &= p(n-1,r)\lambda + p(n+1,1)\mu, \qquad n>1, & (7.3)\\
p(1,k)(\lambda+\mu) &= p(1,k+1)\mu, \qquad k<r, & (7.4)\\
p(1,r)(\lambda+\mu) &= p(0)\lambda + p(2,1)\mu, & (7.5)\\
p(0)\lambda &= p(1,1)\mu. & (7.6)
\end{aligned}$$

If we now substitute the solution $p(n,k) = y^n v_k$ into equation (7.2) then we get

$$(\lambda+\mu)yv_k = \lambda v_k + \mu y v_{k+1},$$

from which follows

$$v_{k+1} = \frac{(\lambda+\mu)y-\lambda}{\mu y} v_k, \qquad k = 1,2,\ldots,r-1.$$

From equation (7.3) we get

$$(\lambda+\mu)yv_r = \lambda v_r + \mu y^2 v_1,$$

with $v_r = x^{r-1}v_1$ and $x = ((\lambda+\mu)y-\lambda)/(\mu y)$. So

$$(\lambda+\mu)yx^{r-1} = \lambda x^{r-1} + \mu y^2,$$

or

$$[(\lambda+\mu)y-\lambda]x^{r-1} = \mu y^2.$$

From this we see that $x^r = y$. Hence

$$(\lambda+\mu)x^r = \lambda + \mu x^{r+1}. \qquad (7.7)$$

Clearly one of the solutions for x is $x = 1$ which implies $y = 1$ and thus cannot be a solution. Using Rouché's theorem and stability condition (7.1) one may show that the other r roots $x_1, \ldots, x_r$ say are all distinct and inside the unit circle (see exercise 7.3.1). Of course then the corresponding values for $y_1, ..., y_r$ are also inside the unit circle. From this we can conclude that the overall solution for the probabilities $p(n, k)$ satisfies

$$p(n,k) = \sum_{i=1}^{r} A_i x_i^{nr+k}, \tag{7.8}$$

where the unknown coefficients A_i and $p(0)$ can be determined from the r boundary conditions (7.4)–(7.6) together with the normalization equation.

7.1.2 Phase description

The second way to describe the state is by counting the number of phases of work in the system. The number $nr + k$ that we met in the preceding section is just that number minus n (for the job in service we have k phases of work instead of n). There is a one to one correspondence between the number of phases in the system and the pair (n, k). The number of phases in the system is equal to the number of jobs waiting in the queue multiplied by r plus the number of not yet completed phases of the job in service.

The analysis for the one-dimensional system is of course closely related to the one in the previous section. For the one-dimensional description we get the flow diagram of figure 7.2.

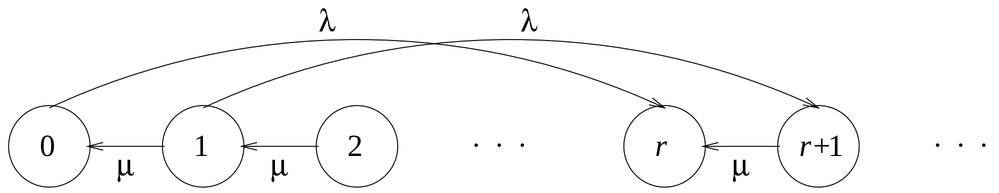

Figure 7.2: One-dimensional flow diagram for the $M/E_r/1$ model

Let p_n be the equilibrium probability of n phases work in the system. From the flow diagram we obtain the following set of balance equations for p_n.

$$\begin{aligned} p_0\lambda &= p_1\mu, & (7.9)\\ p_n(\lambda+\mu) &= p_{n-r}\lambda + p_{n+1}\mu, \qquad n = 1, 2, \ldots & (7.10) \end{aligned}$$

where by convention

$$p_{-1} = p_{-2} = \cdots = p_{-r+1} = 0. \tag{7.11}$$

The balance equations form a recurrence relation of order $r + 1$. So its general solution is

$$p_n = \sum_{k=1}^{r+1} B_k x_k^n, \qquad n = 0, 1, 2, \ldots$$

where the factors x_k are the roots of the equation

$$(\lambda + \mu)x^r = \lambda + \mu x^{r+1},$$

provided this equation has $r+1$ distinct roots. Note that this equation is the same as (7.7). So there are indeed $r+1$ distinct roots. One of them equals 1, say $x_{r+1} = 1$, and the others are all inside the unit circle.

Since the probabilities p_n have to add up to one, we have to set $B_{r+1} = 0$. It is convenient to rewrite p_n in the form of a sum of r geometric distributions, yielding

$$p_n = \sum_{k=1}^{r} C_k(1 - x_k)x_k^n\,, \qquad n = 0, 1, 2, \ldots \tag{7.12}$$

The coefficients C_k can be determined from convention (7.11) and the normalization condition. Note that since the balance equations are dependent, equation (7.9) may be omitted. Substitution of (7.12) into (7.11) directly leads to the following set of equations for the coefficients C_k.

$$C_1(1 - x_1)y_1^n + \cdots + C_r(1 - x_r)y_r^n = 0\,, \qquad n = 1, \ldots, r-1\,,$$

where $y_k = 1/x_k$. These equations are of *VanderMonde-type* and therefore can be solved explicitly using Cramer's rule (see exercise 7.3.2). Then we get

$$C_k = \frac{C}{\prod_{j=1}^{r}(1 - y_j)} \frac{\prod_{j\neq k}(1 - y_j)}{\prod_{j\neq k}(y_k - y_j)}\,, \qquad k = 1, \ldots, r\,,$$

for some constant C. This constant can be determined from the normalization condition, which leads to (see exercise 7.3.3)

$$C = \prod_{j=1}^{r}(1 - y_j). \tag{7.13}$$

Hence the coefficients C_k are given by

$$C_k = \frac{\prod_{j\neq k}(1 - y_j)}{\prod_{j\neq k}(y_k - y_j)}\,, \qquad k = 1, \ldots, r.$$

This completes the determination of the equilibrium probabilities p_n. Of course, it is easily verified that expression (7.8) found for the probabilities $p(n, k)$ is equivalent to the one for the probabilities p_n.

Apparently the problem of finding the probabilities boils down to finding the roots of an equation. How these roots can be found will be briefly discussed in the next section.

Solution (7.12) is a generalization of the one for $M/M/1$ model. Based on this observation, results for the mean and distribution of the waiting time in the $M/M/1$ can be readily generalized to the case of Erlang processing times, yielding

$$P(W > t) = \sum_{k=1}^{r} C_k x_k e^{-\mu(1-x_k)t}\,, \qquad t > 0, \tag{7.14}$$

and

$$E(W) = \sum_{k=1}^{r} C_k \frac{x_k}{\mu(1 - x_k)} \,. \tag{7.15}$$

By using the PASTA property and Little's law it is also possible to directly derive an expression for $E(W)$ (see exercise 7.3.4).

In table 7.1 we list for varying values of ρ and r the mean waiting time and some waiting time probabilities. The squared coefficient of variation of the processing time is denoted by c_X^2. We see that the variation in the processing times is important to the behavior of the system.

ρ	r	c_X^2	$E(W)$		$P(W > t)$		
				t	5	10	20
0.8	1	1	4		0.29	0.11	0.02
	2	0.5	3		0.21	0.05	0.00
	4	0.25	2.5		0.16	0.03	0.00
	10	0.1	2.2		0.12	0.02	0.00
0.9	1	1	9		0.55	0.33	0.12
	2	0.5	6.75		0.46	0.24	0.06
	4	0.25	5.625		0.41	0.18	0.04
	10	0.1	4.95		0.36	0.14	0.02

Table 7.1: Performance characteristics for the $M/E_r/1$ with mean processing time 1

We finally remark that the results in this section can be easily extended to the case that the processing times are mixtures of Erlang distributions with the same scale parameters.

7.1.3 Root finding

In the preceding sections we have shown that the problem of solving the equilibrium probabilities reduces to that of finding the roots of the polynomial equation (7.7). Standard methods to find the zeros of a polynomial of degree $r + 1$ may suffer from numerical instability if r is large. Note that a coefficient of variation of 0.1 already gives rise to $r = 100$. Below we will show that these roots can be efficiently determined by transforming equation (7.7) to r fixed point equations which can be solved by successive substitutions.

Equation (7.7) can be rewritten as

$$x^r = \frac{\lambda}{\lambda + \mu(1 - x)} \,.$$

Raising both sides to the power $1/r$ yields

$$x = \phi F(x), \tag{7.16}$$

where ϕ satisfies $\phi^r = 1$ and

$$F(x) = \sqrt[r]{\lambda/(\lambda + \mu(1 - x))}\,.$$

By applying Rouché's Theorem it can be proved for each feasible ϕ that equation (7.16) has exactly one root with $|x| < 1$. For given ϕ we can try to find the root of (7.16) with $|x| < 1$ by using the iteration scheme

$$x^{(k+1)} = \phi F(x^{(k)})\,, \qquad k = 0, 1, \ldots \tag{7.17}$$

starting with $x^{(0)} = 0$. Let us investigate whether the sequence $x^{(0)}, x^{(1)}, \ldots$ indeed converges to the desired root. It is easily verified that $F(x)$ is convex for $0 \leq x \leq 1$. This implies, together with $F(0) > 0$, $F(1) = 1$ and $F'(1) = 1/\rho > 1$, that there is a unique χ with $0 < \chi < 1$ satisfying

$$\chi = F(\chi)\,.$$

We will now show that $\phi F(x)$ is a *contraction* on $|x| \leq \chi$. For $|x| \leq \chi$ it follows that

$$|\phi F(x)| \leq F(|x|) \leq F(\chi) = \chi\,,$$

so, ϕF maps $|x| \leq \chi$ into itself. Further, for $|\tilde{x}|, |\hat{x}| < \chi$ we have

$$\begin{aligned} |\phi F(\tilde{x}) - \phi F(\hat{x})| &= \left| \int_0^1 F'(\hat{x} + t(\tilde{x} - \hat{x}))(\tilde{x} - \hat{x})dt \right| \\ &\leq |\tilde{x} - \hat{x}| \max_{0 \leq t \leq 1} |F'(\hat{x} + t(\tilde{x} - \hat{x}))|\,. \end{aligned} \tag{7.18}$$

Since for all $|x| \leq \chi$,

$$|F'(x)| \leq F'(|x|) \leq F'(\chi) < 1\,,$$

we get from (7.18) that

$$|\phi F(\tilde{x}) - \phi F(\hat{x})| \leq |\tilde{x} - \hat{x}| F'(\chi)\,.$$

Hence, ϕF is a contraction, so that we can conclude that for given ϕ satisfying $\phi^r = 1$ the root of (7.16) with $|x| < 1$ can indeed be found by using the iteration scheme (7.17).

Let us investigate some properties of the roots $x_1, \ldots, x_r$. Equation (7.16) and thus also equation (7.7) has a unique positive root, less than 1 (namely χ). Denote this root by x_1. Remark that the computation of x_1 can also be done by simple bisection on the interval $(0, 1)$ (instead of iterating). It is easily seen that the other (possible complex) roots are in absolute value strictly less than x_1. Hence as $n \to \infty$,

$$p_n \approx C_1(1 - x_1)x_1^n.$$

This geometric tail behavior arises in many queueing and inventory applications (see e.g. [43]). The property is used in [45] to derive a simple approximation for solving the balance equations of an infinite Markov chain. The idea of the approach is to reduce the infinite system of balance equations to a finite one using the geometric tail behavior (if it exists) of the state probabilities. Many applications of this approximation can be found in [15]. The approach will be explained for the present model. Replace p_n by $p_N x_1^{n-N}$ in (7.9)–(7.10) and in the normalization equation, where N is a sufficiently large integer. Then we get the following finite set of linear equations.

$$\begin{aligned} p_0\lambda &= p_1\mu, && (7.19)\\ p_n(\lambda+\mu) &= p_{n-r}\lambda + p_{n+1}\mu, \qquad n = 1,2,\ldots,N-2, && (7.20)\\ p_{N-1}(\lambda+\mu) &= p_{N-1-r}\lambda + p_N\mu, && (7.21) \end{aligned}$$

and

$$\sum_{n=0}^{N-1} p_n + \frac{p_N}{1-x_1} = 1.$$

This finite system can routinely be solved using a standard Gaussian elimination procedure. For large N the quality of the resulting approximation for the state probabilities will be excellent, but it turns out that already for moderate values of N the approximations are reasonable. This is due to the fact that the geometric behavior, although being an asymptotic property, dominates very fast. Denote the resulting approximation for the mean waiting time by $E(S(N))$. In table 7.2 we list for the same examples as in table 7.1 (except for $r = 1$ in which case the approximation is exact) the approximate mean waiting times $E(S(N))$ for several values of the truncation parameter N.

ρ	r	c_X^2		$E(S(N))$					
			N	5	10	15	20	25	∞
0.8	2	0.5		3.09	2.99	3.00	3.00	3.00	3
	4	0.25		2.76	2.43	2.50	2.50	2.50	2.5
	10	0.1		1.84	2.40	2.13	2.26	2.20	2.2
0.9	2	0.5		6.86	6.74	6.75	6.75	6.75	6.75
	4	0.25		5.92	5.53	5.628	5.627	5.625	5.625
	10	0.1		4.51	5.18	4.86	5.02	4.95	4.95

Table 7.2: Approximations for the mean waiting time for the $M/E_r/1$ with mean processing time 1

7.2 $G/E_r/1$ queue

In practice Poisson arrivals often provide a good description of the arrival process, but there are also many cases of interest in which Poisson arrivals are not appropriate. Think for instance of appointment systems or communication systems where the input may have a strong bursty character (during long periods no traffic and for short periods a lot of traffic, see also exercise 6.4.4). In this section we deal with systems with generally distributed interarrival times and we show in this case how the waiting time distribution can be found.

Consider a single machine with Erlang-r distributed processing times with mean r/μ and an arbitrary interarrival time distribution $A(t)$ with mean $1/\lambda$. For stability we assume

$$\frac{\lambda r}{\mu} < 1.$$

At an arbitrary point t in time the state of the system can be described by the pair (x, n) where x is the time elapsed till the last arrival before t and n is the number of uncompleted phases work in the system. Since the first variable is continuous this description leads to a complicated analysis. Therefore we will look at special points in time, namely at arrival instants. On arrival the state description simplifies to n, since $x = 0$.

So we consider the system at arrival instants, or in other words, we look at the *imbedded Markov chain* just prior to an arrival. The states are the number of uncompleted phases work in the system just prior to an arrival and the transition probabilities are given by

$$p_{nm} = \begin{cases} \sum_{k=n+r}^{\infty} \beta_k\,, & n \geq 0\,,\ \ m = 0\,, \\ \beta_{n+r-m}\,, & n \geq 0\,,\ \ 1 \leq m \leq n+r\,, \\ 0\,, & n \geq 0\,,\ \ m \geq n+r+1\,, \end{cases}$$

where β_k is defined as the probability that k service stages are completed during an interarrival time, so

$$\beta_k = \int_0^\infty \frac{(\mu t)^k}{k!} e^{-\mu t} dA(t)\,, \qquad k \geq 0\,.$$

Note that the Markov chain is irreducible and aperiodic. The equilibrium probabilities $p_n, n = 0, 1, \ldots$ satisfy the following balance equations.

$$p_0 = \sum_{k=r}^{\infty} \beta_k p_0 + \sum_{k=r+1}^{\infty} \beta_k p_1 + \sum_{k=r+2}^{\infty} \beta_k p_2 + \ldots\,, \tag{7.22}$$

$$p_n = \beta_0 p_{n-r} + \beta_1 p_{n-r+1} + \beta_2 p_{n-r+2} + \ldots\,, \qquad n = 1, 2, \ldots\,, \tag{7.23}$$

where by convention

$$p_{-1} = p_{-2} = \cdots = p_{-r+1} = 0. \tag{7.24}$$

The probability p_n can be interpreted as the probability that an arbitrary job finds on arrival n phases work in the system. So once we know these probabilities we can calculate the waiting time of a job.

Remark that the balance equations form a recurrence relation the order of which is *not finite*. But to solve these equations we will use the same idea as for finite order recurrence relations. We seek solutions of equation (7.23) of the form $p_n = x^n$ and then linearly combine these solutions to also satisfy the convention (7.24) and the normalization condition. Note that, since the balance equations are dependent, equation (7.22) will be satisfied automatically. Inserting the form $p_n = x^n$ into (7.23) and then dividing by x^{n-r} yield

$$x^r = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots = \tilde{A}(\mu(1-x)), \tag{7.25}$$

where $\tilde{A}(s)$ is the Laplace-Stieltjes transform of $A(t)$, i.e.,

$$\tilde{A}(s) = E(e^{-As}) = \int_0^\infty e^{-\mu t} dA(t), \qquad Re\, s > 0.$$

Clearly, because of the normalization condition, only solutions with $|x| < 1$ are useful. Applying Rouché's theorem it can be shown that equation (7.25) has exactly r roots inside the unit circle. For convenience, we assume that the roots of (7.25) with $|x| < 1$ are simple. Hence, we find r solutions of the form $p_n = x^n$ with $|x| < 1$ satisfying equation (7.23). These solutions are labeled x_k^n, $k = 1, \ldots, r$. It now remains to determine coefficients C_k such that the linear combination

$$p_n = \sum_{k=1}^{r+1} C_k (1 - x_k) x_k^n, \qquad n = 0, 1, 2, \ldots$$

satisfies the convention (7.24) and the normalization condition. This proceeds along exactly the same lines as in section 7.1. We can conclude that the solution to the $G/E_r/1$ model is of the same form as the one for the $M/E_r/1$ model, and thus also the waiting time distribution is given by expression (7.14). Under certain conditions for the interarrival time distribution it can be shown that the roots x_k can be found by successive substitutions similar as in section 7.1.3 (see [4]).

Finally we remark that the analysis in this section can easily be extended to the case of processing times with a mixed Erlang distribution (see figure 4.4).

7.3 Exercises

Exercise 7.3.1 *(Rouché's theorem)*
Rouché's theorem reads as follows.
Let the bounded region D have as its boundary a contour C. Let the functions $f(z)$ and $g(z)$ be analytic both in D and on C, and assume that $|f(z)| < |g(z)|$ on C. Then $f(z) + g(z)$ has in D the same number of zeros as $g(z)$, all zeros counted according to their multiplicity.

(i) Use Rouché's theorem to prove that the polynomial equation (7.7) has exactly r (possibly complex) roots x with $|x| < 1$.
(Hint: take the left-hand side of (7.7) as $f(z)$, its right-hand side as $g(z)$ and take as contour C the circle with centre 0 and radius $1 - \epsilon$ with ϵ small and positive.)

(ii) Show that all roots are simple.

Exercise 7.3.2 *(VanderMonde determinant)*
Let $\alpha_1, \alpha_2, \ldots$ be a sequence of numbers and define the determinant Δ_n for $n = 1, 2, \ldots$ by

$$\Delta_n = \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \vdots & & & \vdots \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \cdots & \alpha_n^{n-1} \end{vmatrix}.$$

The determinant Δ_n is called a *VanderMonde determinant* of order n.

(i) Derive the following recurrence relation for the determinants Δ_n,

$$\Delta_n = \Delta_{n-1} \prod_{m=1}^{n-1} (\alpha_m - \alpha_n), \qquad n = 1, 2, \ldots$$

(ii) Show that the solution of this recurrence relation is given by

$$\Delta_n = \prod_{1 \le i < j \le n} (\alpha_j - \alpha_i).$$

Exercise 7.3.3
Prove expression (7.13) for the constant C.
(Hint: consider the polynomial

$$p(x) = \sum_{k=1}^{r} \frac{\prod_{j \ne k}(x - y_j)}{\prod_{j \ne k}(y_k - y_j)},$$

and show that it is identical to 1.)

Exercise 7.3.4
Consider the $M/E_r/1$ queue. Derive, by using the PASTA property and Little's law, that

$$E(W) = \frac{\rho E(R)}{1-\rho},$$

where $E(R)$ is the mean residual service time given by

$$E(R) = \frac{1+r}{2\mu}.$$

Exercise 7.3.5 *(Hyper-exponential processing times)*
Consider a single machine where orders arrive according to a Poisson stream with rate λ. The processing times are hyperexponentially distributed. With probability q_1 the processing time is exponential with mean $1/\mu_1$ and with probability q_2 it is exponential with mean $1/\mu_2$ ($q_1 + q_2 = 1$). The hyperexponential distribution can be used to model processing times with a coefficient of variation greater than 1 (see section 4.3), but it can also arise naturally (when there are two types of jobs).
The state of this system can be described by a pair (n, m) where n is the number of jobs in the system and m indicates the type of job being processed. If the job being processed is exponential with parameter μ_1 (resp. μ_2) then $m = 1$ (resp. $m = 2$).

(i) Describe the flow diagram and give the set of balance equations for the equilibrium state probabilities.

(ii) Solve these equations.

(iii) Derive an expression for the mean and distribution of the time spent in the system.

Exercise 7.3.6 *(Coxian processing times)*
Consider the same system as in the previous exercise, except that now the the processing times are Coxian distributed of order 2 (see figure 4.3). The processing time consists of an exponential phase with mean $1/\mu_1$ which is followed with probability q by a second exponential phase with mean $1\mu_2$. The Coxian-2 distribution can be used to model processing times with a coefficient of variation greater than 0.5 (see section 4.3).

(i) How can the state of this system be characterized?

(ii) Describe the flow diagram and give the set of balance equations for the state probabilities.

(iii) Find the solution to these equations.

(iv) Derive an expression for the mean and distribution of the time spent in the system.

Exercise 7.3.7 *(Two phase production)*
A machine produces products in two phases. The first phase is standard and the same for all products. The second phase is customer specific (the finishing touch). The first (resp. second) phase takes an exponential time with mean $1/\mu_1$ (resp. $1/\mu_2$) hour. Orders for the production of one product arrive according to a Poisson stream with rate λ orders per hour. Orders are processed in order of arrival.
The state of this system can be described by the pair (n, m) where n is the number of orders in the system and m indicates the production phase (so m is 1 or 2).

(i) Formulate and solve the balance equations for the state probabilities.

Suppose that $\mu_1 = 6, \mu_2 = 30$ and $\lambda = 3$.

(ii) Determine the mean production lead time of an order.

To reduce the production lead time one decides to start already with the production of phase 1 when the system is empty. If upon completion of phase 1 no order arrived yet, the machine stops the production and waits for the first order to arrive. When the first order arrives the machine can directly start with phase 2.

(iii) Formulate and solve the balance equations for this slightly modified system.

(iv) Determine the reduction in the mean production lead time. Try to explain this reduction.

Exercise 7.3.8
Consider a machine where jobs are being processed. The mean production time is 4 minutes and the standard deviation is 3 minutes. We have seen in exercise 4.10.5 that the production times can be modelled as a mixture of an exponential distribution and an Erlang-2 distribution with the same scale parameters.
The mean number of jobs arriving per hour is 10. The interarrival times are uniformly distributed between 4 and 8 minutes.
The state of this system just prior to the arrival of an order can be described by n, where n is the number of uncompleted production phases in the system.

(i) Formulate the balance equations for the (equilibrium) arrival probabilities.

(ii) Determine the arrival probabilities.

(iii) What is the fraction of the orders which has to wait longer than 10 minutes?

Exercise 7.3.9
Orders arrive according to a Poisson process, so the interarrival times are independent and exponentially distributed. Each order has to pass, stage by stage, through a series of R independent production stages. Each stage takes a exponentially distributed time. The total production time, the workload, of the order is the sum of R independent, identically distributed random variables. So the distribution of the workload is the R-stage Erlangian distribution.

(i) The state of this system (i.e., the work station together with the queue in front of it) can be characterized by the number of orders in the system and by the number of not yet completed stages for the order in the work station.
Describe the flow diagram and give the set of equations for the equilibrium state probabilities.

(ii) The state of the system at a certain time can also be described by the total number of production stages that have to be completed before all the orders in the system are ready.
Give the flow diagram and formulate the equations for the equilibrium state probabilities.

(iii) Define

$$\begin{aligned} p_n &= P(\text{total number of orders in the system} = n), \\ q_k &= P(\text{total number of production stages in the system } = k). \end{aligned}$$

Give the relation between these two probabilities.

Exercise 7.3.10
Orders arrive in so called bulks. Each bulk consists of R independent orders. The bulks themselves arrive according to a Poisson process, with independent exponentially distributed interarrival times. The sequence in which orders of one bulk are processed in the work station is unimportant. All orders of a bulk have to wait until all orders of bulks that have arrived earlier are completed. The workload of each order is exponentially distributed. The state of the system is simply characterized by the number of orders in the system.
Describe the flwo diagram for this situation. Compare the result with that of the previous exercise, part (ii).

Exercise 7.3.11
Two types of orders arrive according to independent Poisson processes with arrival rates λ_1 and λ_2 respectively. Type i orders require an exponentially distributed time with mean $1/\mu_i$, $i = 1, 2$.
Describe the state of the system, give the flow diagram and the balance equations for the equilibrium state probabilities.

Chapter 8

Multi-server systems

The topic of this chapter is the analysis of queueing models with multiple parallel servers. One of the simplest examples is the system with Poisson arrivals and identical parallel exponential servers. This system has been studied in section 5.3. The analysis of non-exponential systems however is much more complicated. In this chapter we will analyze in detail a system with 2 servers and Coxian service times. It will appear that the analysis of this simple special case is already much more involving than the analysis of the $M/M/c$.

In section 6.3 we considered a production system with two parallel machines, a fast and a slow one. We investigated what to do with the slow machine: when to use it and when not. In this chapter we will look at a system with two identical machines, but with two different types of jobs, small and big ones. Then the problem we address is how to process the jobs. Should we use both machines for both types of jobs, or should we use each machine for one job type only? For answering this question a simple model will be formulated and studied in the next section.

8.1 A production system with two types of jobs

Consider two parallel machines where two types of jobs, labeled 1 and 2, arrive according to Poisson streams with rate λ_1, respectively λ_2. The processing times are exponentially distributed with mean $1/\mu_1$ (resp. $1/\mu_2$) for type 1 (resp. type 2) jobs. We assume $\mu_1 > \mu_2$, so the type 1 jobs are the small ones, the type 2 jobs the big ones. The question is whether it is better dedicate each machine to one job type or to pool the machines and use them for both types. In the first case we have two $M/M/1$ models. This option is only sensible if each of the machines is capable of handling one type job, so assume

$$\rho_1 = \frac{\lambda_1}{\mu_1} < 1, \qquad \rho_2 = \frac{\lambda_2}{\mu_2} < 1.$$

Denoting the mean time spent in the system of type 1 (resp. type 2) jobs by $E(S(1))$ (resp. $E(S(2))$ we get

$$E(S(1)) = \frac{1}{\mu_1 - \lambda_1}, \qquad E(S(2)) = \frac{1}{\mu_2 - \lambda_2}.$$

The overall mean time spent in the system equals

$$E(S) = \frac{\lambda_1}{\lambda_1 + \lambda_2} E(S(1)) + \frac{\lambda_2}{\lambda_1 + \lambda_2} E(S(2)). \tag{8.1}$$

Note that it can happen that one machine is idle while there are jobs waiting at the other machine. So the capacity is not optimally used.

We now want to compare the dedicated situation with the situation where the two machines are pooled, i.e. they are both used for the processing of the small and big jobs. Clearly in this case it will not happen that a machine is idle while there are jobs waiting. The problem with pooled machines can be modelled by a system with two parallel identical machines where one big stream of jobs arrives with rate $\lambda = \lambda_1 + \lambda_2$ (we merge the type 1 and type 2 streams). When a machine starts processing a job, it is with probability p_1 a type 1 job with mean processing time $1/\mu_1$ and with probability p_2 a type 2 job with mean processing time $1/\mu_2$, where

$$p_1 = \frac{\lambda_1}{\lambda}, \qquad p_2 = \frac{\lambda_2}{\lambda}.$$

So the processing times can be modelled by a hyperexponential distribution. To compare this system with the dedicated one it is now our task to find the mean time spent in the system in a $M/H_2/2$ system. This is the subject of the next section. For the analysis it appears to be easier to deal with Coxian-2 processing times than hyperexponential processing times. In exercise 4.10.8 we have seen that a hyperexponential distribution with parameters $\mu_1 > \mu_2$ and branching probabilities p_1 and p_2, respectively, is equivalent to a Coxian-2 distribution with parameters μ_1 and μ_2 and branching probability q given by

$$q = 1 - (p_1\mu_1 + p_2\mu_2)/\mu_1 .$$

The phase representation of this Coxian-2 distribution is shown in figure 8.1.

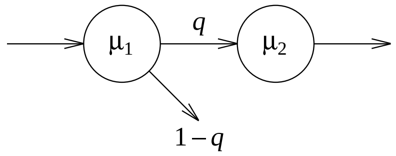

Figure 8.1: Phase diagram for the Coxian-2 distribution

8.2 $M/C_2/2$ queue

We consider a system with two parallel identical machines. Jobs arrive according to a Poisson stream with rate λ and are processed in order of arrival. The processing times are

Coxian-2 distributed with parameters $\mu_1 > \mu_2$ and branching probability q. For stability we assume

$$\rho = \frac{\lambda}{2}\left(\frac{1}{\mu_1} + \frac{q}{\mu_2}\right) < 1. \tag{8.2}$$

The quantity ρ is the machine utilization. The state of the system can be described by the pair (n, m) where n is the number of jobs in the system (waiting plus the ones being processed) and m is the number of jobs in phase 2 of the production. So n ranges from 0 to ∞ and m from 0 to 2, unless $n = 0$ or $n = 1$ in which case $m = 0$, respectively $m = 0, 1$. Below we will analyze the equilibrium behavior of the system. The flow diagram for this system is depicted in figure 8.2.

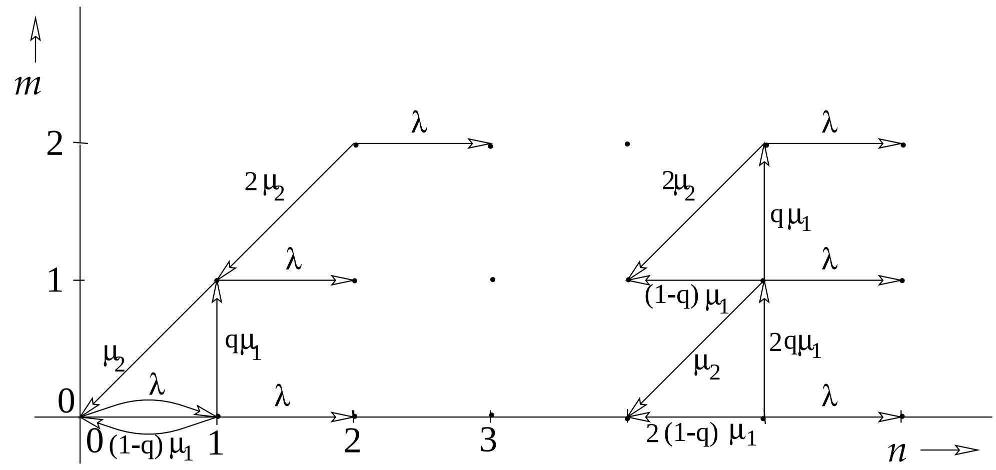

Figure 8.2: Flow diagram for the $M/C_2/2$ model

Let $p(n, m)$ be the equilibrium probability of state (n, m). From the flow diagram we obtain the following set of balance equations for $p(n, m)$.

$$\begin{aligned}
p(0,0)\lambda &= p(1,0)(1-q)\mu_1 + p(1,1)\mu_2, & (8.3)\\
p(1,0)(\lambda+\mu_1) &= p(0,0)\lambda + p(2,1)\mu_2 + p(2,0)2(1-q)\mu_1, & (8.4)\\
p(1,1)(\lambda+\mu_2) &= p(2,2)2\mu_2 + p(2,1)(1-q)\mu_1 + p(1,0)q\mu_1, & (8.5)\\
p(n,0)(\lambda+2\mu_1) &= p(n-1,0)\lambda + p(n+1,1)\mu_2 & \\
&\quad +p(n+1,0)2(1-q)\mu_1, \qquad n = 2,3,\ldots & (8.6)\\
p(n,1)(\lambda+\mu_1+\mu_2) &= p(n-1,1)\lambda + p(n+1,2)2\mu_2 & \\
&\quad +p(n+1,1)(1-q)\mu_1 + p(n,0)2q\mu_1, \quad n = 2,3,\ldots & (8.7)\\
p(2,2)(\lambda+2\mu_2) &= p(2,1)q\mu_1, & (8.8)\\
p(n,2)(\lambda+2\mu_2) &= p(n-1,2)\lambda + p(n,1)q\mu_1, \qquad n = 3,4,\ldots & (8.9)
\end{aligned}$$

A compact formulation of the equations (8.6)–(8.9) is given below.

$$p(n,m)(\lambda + (2-m)\mu_1 + m\mu_2) = p(n-1,m)\lambda + p(n+1,m+1)(m+1)\mu_2$$

$$
\begin{aligned}
&+p(n+1,m)(2-m)(1-q)\mu_1 \\
&+p(n,m-1)(2-(m-1))q\mu_1, \qquad (8.10)
\end{aligned}
$$

valid for all $n = 2, 3, \ldots$ and $m = 0, 1, 2$, where by convention $p(2, 2) = 0$ and $p(n, m) = 0$ for $m < 0$ and $m > 2$. To solve the balance equations we first consider equation (8.10). The boundary conditions (8.3)–(8.5) will be satisfied at the end of the analysis.

We seek solutions of (8.10) of the form

$$
p(n,m) = y(m)x^n, \qquad n = 1, 2, 3, \ldots, \ m = 0, 1, 2.
$$

Substitution of this form into (8.10) and then dividing both sides by x^{n-1} yields

$$
\begin{aligned}
&y(m-1)(2-(m-1))q\mu_1 x \\
&\quad +y(m)(\lambda+(2-m)(1-q)\mu_1 x^2-(\lambda+(2-m)\mu_1+m\mu_2)x) \\
&\quad +y(m+1)(m+1)\mu_2 x^2 = 0, \qquad m = 0, 1, 2, \qquad (8.11)
\end{aligned}
$$

where by convention $y(m) = 0$ for $m < 0$ and $m > 2$. Now we have to find the values of x for which the linear system above has a non-null solution. Since we must be able to normalize the solution afterwards, we need values of x satisfying $|x| < 1$. To determine the desired values of x we employ an idea of [39, 9] to transform the difference equations (8.11) to a single differential equation for the *generating function* of the sequence $y(m)$.

Let us introduce the generating function

$$
Y(z) = \sum_{m=0}^{2} y(m)z^m.
$$

Multiplying (8.11) by z^m and adding with respect to m we obtain the differential equation

$$
\begin{aligned}
&Y(z)z2q\mu_1 x - Y'(z)z^2 q\mu_1 x \\
&\quad +Y(z)(\lambda+2(1-q)\mu_1 x^2-(\lambda+2\mu_1)x) - Y'(z)z((1-q)\mu_1 x^2-\mu_1 x+\mu_2 x) \\
&\quad +Y'(z)\mu_2 x^2 = 0,
\end{aligned}
$$

which may be written in the form

$$
\begin{aligned}
\frac{Y'(z)}{Y(z)} &= \frac{2q\mu_1 xz+\lambda+2(1-q)\mu_1 x^2-(\lambda+2\mu_1)x}{q\mu_1 xz^2+((1-q)\mu_1 x^2-\mu_1 x+\mu_2 x)z-\mu_2 x^2} \\
&= \frac{2z+n_0(x)}{z^2+d_1(x)z+d_0(x)}, \qquad (8.12)
\end{aligned}
$$

where we introduced the abbreviations

$$
\begin{aligned}
n_0(x) &= (\lambda+2(1-q)\mu_1 x^2-(\lambda+2\mu_1)x)/(q\mu_1 x), \\
d_1(x) &= ((1-q)\mu_1 x^2-\mu_1 x+\mu_2 x)/(q\mu_1 x), \\
d_0(x) &= -\mu_2 x^2/(q\mu_1 x).
\end{aligned}
$$

Partial fraction decomposition of the right-hand side of (8.12) leads to

$$\frac{Y'(z)}{Y(z)} = \frac{A(x)}{z - z_1(x)} + \frac{2 - A(x)}{z - z_2(x)}. \tag{8.13}$$

In this equation $z_1(x)$ and $z_2(x)$ are the roots of

$$z^2 + d_1(x)z + d_0(x) = 0,$$

so

$$z_1(x) = \frac{-d_1(x) + \sqrt{d_1^2(x) - 4d_0(x)}}{2}, \qquad z_2(x) = \frac{-d_1(x) - \sqrt{d_1^2(x) - 4d_0(x)}}{2}.$$

The coefficient $A(x)$ satisfies

$$(2 - A(x))z_1(x) + A(x)z_2(x) + n_0(x) = 0.$$

By substituting $z_2(x) = -d_1(x) - z_1(x)$ and the expression for $z_1(x)$ the equation $A(x)$ becomes

$$(1 - A(x))\sqrt{d_1^2(x) - 4d_0(x)} = d_1(x) - n_0(x). \tag{8.14}$$

The solution of the differential equation (8.13) is

$$Y(z) = K(z - z_1(x))^{A(x)}(z - z_2(x))^{2-A(x)},$$

with K an arbitrary constant. Now the key idea to proceed is that since $Y(z)$ is a polynomial of z, the exponent $A(x)$ should equal an integer, i.e. $A(x) = j$, $j = 0, 1, 2$. Hence from (8.14) we can conclude that x must satisfy

$$(1 - j)\sqrt{((1-q)\mu_1 x^2 - \mu_1 x + \mu_2 x)^2 + 4q\mu_1\mu_2 x^2} = -\lambda + (\lambda + \mu_1 + \mu_2)x - (1-q)\mu_1 x^2,$$

for some $j = 0, 1, 2$. So we have three different equations for x. For $j = 1$ we have

$$(1-q)\mu_1 x^2 - (\lambda + \mu_1 + \mu_2)x + \lambda = 0, \tag{8.15}$$

and for $j = 0, 2$ we obtain, by squaring both sides of the equation,

$$(x - 1)(2\mu_1((1-q)\lambda + 2\mu_2)x^2 - \lambda(\lambda + 2\mu_1 + 2\mu_2)x + \lambda^2) = 0. \tag{8.16}$$

The root $x = 1$ is not useful. Provided $\rho < 1$ it can be shown (see exercise 8.4.1) that the quadratic equation (8.15) has one positive root, less than unity and equation (8.16) has two positive roots, less than unity. For convenience we assume that the roots are different (otherwise we can slightly modify one the parameters of the Coxian distribution). Let us denote these roots by x_0, x_1 and x_2. For given $x = x_j$ the corresponding $y_j(m)$ can easily be solved from (8.11) yielding

$$y_j(2) = 1, \quad y_j(1) = \frac{(\lambda + 2\mu_2)x_j - \lambda}{q\mu_1 x_j}, \quad y_j(0) = \frac{y_j(1)\mu_2 x^2}{\lambda + 2(1-q)\mu_1 x^2 - (\lambda + 2\mu_1)x}.$$

We have now succeeded in finding three different solutions to (8.10). These solutions are linearly combined to also satisfy the boundary conditions. Set

$$p(n,m) = \sum_{j=0}^{2} C_j y_j(m) x_j^n \,, n = 1,2,\ldots,\ m = 0,1,2,$$

where C_j are constants to be determined yet. The probability $p(0,0)$ directly follows from equation (8.3) yielding

$$p(0,0) = \frac{1}{\lambda}\sum_{j=0}^{2} C_j[y_j(0)x_j(1-q)\mu_1 + y_j(1)x_j\mu_2]\,.$$

The coefficients C_j follow from the convention $p(1,2) = 0$, so

$$\sum_{j=0}^{2} C_j y_j(2) x_j = 0,$$

and equation (8.5) which leads to

$$\sum_{j=0}^{2} C_j[y_j(1)(\lambda - \mu_1 x_j) + y_j(0) q \mu_1 x_j] = 0.$$

Due to the dependence of the balance equations the remaining equation in $(1,0)$ will be automatically satisfied. Finally the solution must be normalized which gives the desired third equation needed to determine the three coefficients C_j.

$$\begin{aligned}\sum_{n,m} p(m,n) &= \frac{1}{\lambda}\left[\sum_{j=0}^{2} C_j y_j(0) x_j (1-q)\mu_1 + \sum_{j=0}^{2} C_j y_j(1) x_j \mu_2\right] \\ &\quad + \sum_{j=0}^{2} C_j \left[(y_j(0)+y_j(1)+y_j(2))\frac{x_j}{1-x_j} - y_j(2)x_j\right] \\ &= 1.\end{aligned}$$

From the three linear equations above the coefficients C_j can readily be solved. This completes the determination of the state probabilities $p(n,m)$. Based on these probabilities we obtain for the mean number in the system

$$E(L) = \sum_{j=0}^{2} C_j \left[(y_j(0)+y_j(1)+y_j(2))\frac{x_j^2}{1-x_j} - y_j(2)x_j\right]$$

from which by Little's law the mean time spent in the system follows.

Knowing how to find the mean time spent in the system in a $M/C_2/2$ system, and thus in a $M/H_2/2$ system, we can return to the original question: do we pool or not?

Remark 8.2.1 *(extensions)*
The approach demonstrated in this section also works for the $E_k/C_2/s$ system, see [9]. In [7, 8] it is shown that even the more general $C_k/C_r/s$ system can analyzed along these lines. The state description of the latter system is a vector $(n, l, i_1, \ldots, i_r)$ where n is the number of jobs in the system, l is the arrival phase currently occupied by the arriving job and i_j is the number of jobs in phase j of the service time. In [7] it is proved that the state probabilities $p(n, l, i_1, \ldots, i_r)$ can be expressed as a finite linear combination of geometric terms,

$$p(n, l, i_1, \ldots, i_r) = \sum_k C_k B_k(l, i_1, \ldots, i_r) x_k^n ,$$

for all states with $n \geq s$ (i.e. states with all servers occupied). This representation is in [8] exploited to obtain a closed form expression for the waiting time distribution. In case of Erlang distributed interarrival and service times a refinement of the representation above can be obtained, i.e., the coefficients $B_k(l, i_1, \ldots, i_r)$ can be explicitly determined (see [5]).

8.3 Pooling or not?

In this section we will compute and compare the mean production lead time over all jobs for the dedicated system and the system in which the machines are pooled. In the figures 8.3 and 8.4 we show the mean production lead times for increasing values of μ_1/μ_2 while keeping ρ_1 and ρ_2 and the overall mean production time constant. The overall mean production time is given by

$$\frac{\lambda_1}{\lambda} \frac{1}{\mu_1} + \frac{\lambda_2}{\lambda} \frac{1}{\mu_2} ,$$

which in each example is set to 1. Since ρ_1 and ρ_2 are constant, the overall mean production lead time in the dedicated system is also constant and equal to (see (8.1))

$$E(S) = \frac{1}{\rho_1 + \rho_2} \left[\frac{\rho_1}{1 - \rho_1} + \frac{\rho_2}{1 - \rho_2} \right] .$$

In the figures 8.3 and 8.4 we see that pooling leads to smaller lead times than in the dedicated system, except for large values of μ_1/μ_2. This might have been expected because pooling makes better use of the capacity. When μ_1/μ_2 is large there will be many small jobs and a few very big ones. So if the machines are pooled it can occur that both machines are occupied by very big jobs, while there are a lot of small ones waiting. Then it may be better to dedicate the machines to one job type. Indeed this is shown in the figures below. Note that the squared coefficient of variation of the production time of an arbitrary job equals

$$1 + \frac{2\rho_1\rho_2}{(\rho_1 + \rho_2)^2} \left(\frac{\mu_1}{\mu_2} + \frac{\mu_2}{\mu_1} - 2 \right) ,$$

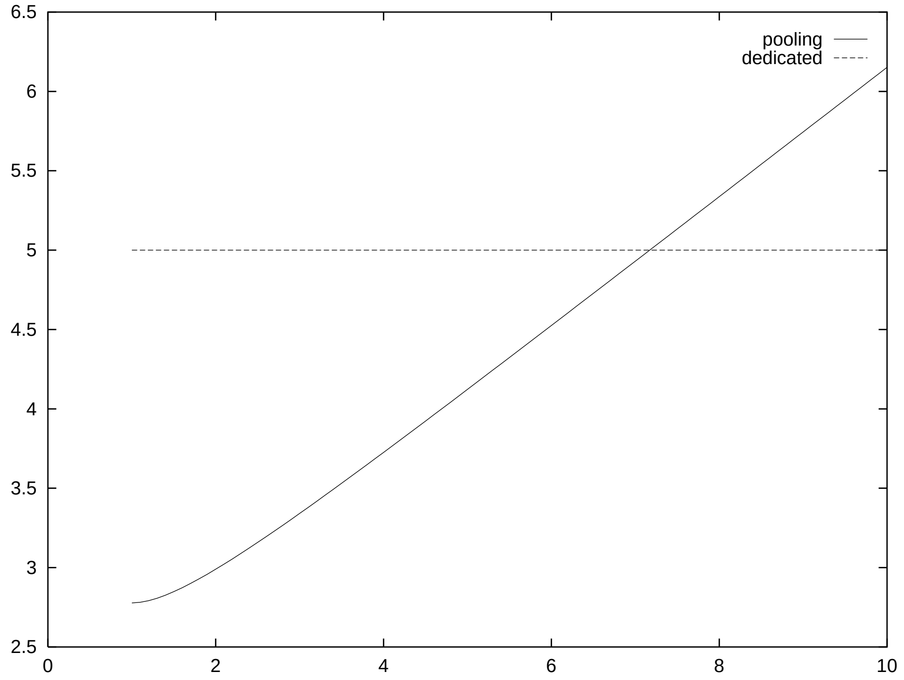

Figure 8.3: Mean production lead times as a function of μ_1/μ_2 for $\rho_1 = \rho_2 = 0.8$

so it linearly increases in μ_1/μ_2 provided μ_1/μ_2 is not close to 1.

The conclusion may be that pooling leads to smaller production lead times than the dedicated system, unless the production times of the two job types extremely differ.

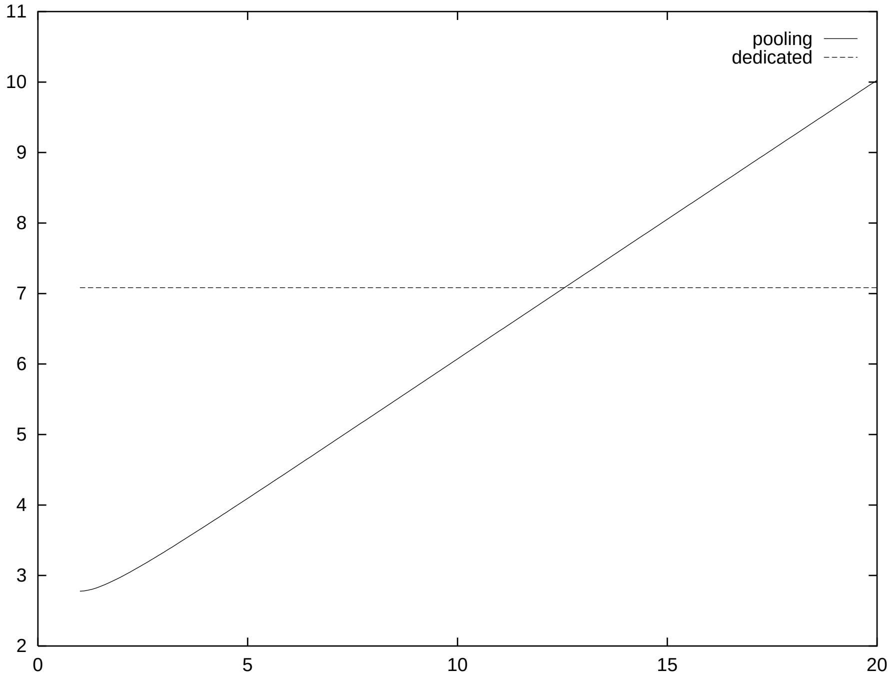

Figure 8.4: Mean production lead times as a function of μ_1/μ_2 for $\rho_1 = 0.9$ and $\rho_2 = 0.7$

8.4 Exercises

Exercise 8.4.1
Show that, provided $\rho < 1$, the quadratic equation (8.15) has a unique positive root, less than unity and (8.16) has two positive roots, less than unity.

Exercise 8.4.2 *($M/E_2/2$ system)*
In a post office financial transactions are treated by two servers. Customers arrive according to a Poisson stream with rate λ and line up in one queue. Since the squared coefficient of the service times is approximately $1/2$, the service times can be modelled by an Erlang-2 distribution with mean $2/\mu$.
The state of this system can be described by the pairs (n, m) where n is the number of customers (waiting or in service) and m is the number of customers with 1 remaining service phase.

(i) For which values of λ is the system capable of handling the amount of work offered per unit time?

(ii) Show that for $n = 2, 3, \ldots$ and $m = 0, 1, 2$ the state probabilities $p(n, m)$ can be expressed as a linear combination of exactly three terms of the form

$$p(n, m) = x^n y(m).$$

(iii) Derive expressions for the mean number of customers in the system and the mean waiting time.

Chapter 9

Combining make to order and make to stock

Lesser and lesser production systems are fully Make to Stock (MTS) organized. In MTS systems the clients receive their products from a warehouse near the client. A lot of research concerns the performance and control of these systems (multi-echelon inventory control). Some other systems are completely Make to order (MTO). No product is made without a client. The analysis of these systems calls for queueing models. An important problem in MTO systems is the relation between the lead time and the utiization of the capacity. It is well-known that the lead time grows with the utilization as $1/(1-\rho)$. Utilizations of 80% are often considered to be far to low. On the other hand a utilization of 95% may give lead times which are not acceptable any more.

If one wants high utilizations and short lead times then one has to be able to deal with the natural day to day fluctuations in the demand (we are not thinking of seasonal variations). One possibility is flexible capacity. Another option studied in this chapter is the combination of MTO with some MTS.

We will consider two models in which this mixture is studied. In the first one we assume that there is one product which is standard and can be made to stock. If the machine is idle it will produce a limited number of these standard products to stock.

In the second model we think of products being produced in two phases. The first phase is standard and common for all products. The second phase is customer specific. If the machine is idle it can continue to produce a limited number of 'first phases' as a capacity inventory.

Assuming exponential production and interarrival times the state of each of the two systems can be described by a pair (n, m), where n indicates the number of outstanding orders and m the number of products on stock. We show that the equilibrium state probabilities of the two systems can be determined analytically. From these probabilities it is straightforward to compute the distribution and moments of the production lead time. Numerical results show that combining MTO and MTS effectively reduces production lead times.

The literature on these hybrid systems is growing fast. The papers [50, 12, 40, 18] study models where several items are produced by a common facility with limited capacity. Some of these items ar produced to stock, others are produced to order. The MTO items have priority over MTS items. For these systems a variety of questions is addressed, such as: 'Which items should be stocked or made to order?' and 'How must production capacity allocated among MTO and MTS items?'

In section 9.1 we first consider some simple queueing models for inventory systems related to the produce-to-stock systems in [11], and investigate the relation between inventory and the production lead time. In particular tails of the production lead time distribution tend to be rather thick. Section 9.2 deals with the two type system, standard products to stock, specific products to order. In section 9.3 we consider the production in two phases, the first one being standard, the second one customer specific.

9.1 Inventories and production lead times

Let us first consider the following model. Jobs arrive according to a Poisson process with rate $\rho(< 1)$, the production times are exponential with mean 1. In this system the production lead time is exponential with mean $1/(1-\rho)$. We now assume that some of the idle time is used to produce items on stock, say at most M items. As state space for this system we take the set $\{-M, -M+1, \ldots, 0, 1, 2, \ldots\}$ where the negative numbers refer to a number of items on stock and the positive numbers to jobs waiting to be processed. The equilibrium probability p_n for the system to be in state n clearly equals

$$p_n = (1-\rho)\rho^{n+M}, \qquad n = -M, -M+1, \ldots$$

So the production lead time $S(M)$ is equal to 0 with probability $1-\rho^M$ and exponential with mean $1/(1-\rho)$ with probability ρ^M. Hence, for ρ close to 1, the maximum stock M has to be fairly large to get a significant effect. Table 9.1 shows the mean production lead time and the probability of a production lead time greater than 10 times the mean production time as a function of ρ and M.

In case the production times have an Erlang distribution with r phases, each phase with mean $1/r$ (so the mean production time is 1), and the maximum stock is M items the analysis proceeds in an almost identical way. Now we take as state space the set $\{-Mr, -Mr+1, \ldots, 0, 1, 2, \ldots\}$ where the negative numbers refer to a number of phases completed in advance (so in state $-r$ there is one item on stock) and the positive numbers to the number of uncompleted phases in the system. The equilibrium probabilities p_n are for $n = -Mr, -Mr+1, \ldots$ given by (see e.g. [27])

$$p_n = \sum_{i=1}^{r} A_i x_i^{n+Mr},$$

where the x_i's are the roots of the equation

$$x^r - \rho(x^{r-1} + x^{r-2} + \cdots + 1)/r = 0,$$

		$E(S(M))$					
M		0	8	16	24	32	40
	0.8	5	0.839	0.141	0.024	0.004	0.001
ρ	0.9	10	4.30	1.85	0.798	0.343	0.148
	0.95	20	13.3	8.80	5.84	3.87	2.57

		$P(S(M) > 10)$					
M		0	8	16	24	32	40
	0.8	0.135	0.023	0.004	0.001	0.000	0.000
ρ	0.9	0.368	0.158	0.068	0.029	0.013	0.005
	0.95	0.607	0.402	0.267	0.177	0.118	0.078

Table 9.1: Performance characteristics for exponential production times.

and the coefficients A_i satisfy

$$A_i = (1-\rho) / \prod_{j \neq i} (1 - x_j/x_i).$$

The production lead time $S(M)$ is equal to 0 with probability $\sum_{n=-Mr}^{-r} p_n$ and it is the sum of $n+r$ exponentials with mean $1/r$ with probability p_n for $n = -r+1, -r+2, \ldots$ So for $t \geq 0$ we obtain that

$$\begin{aligned} P(S(M) > t) &= \sum_{n=-r+1}^{\infty} p_n \sum_{l=0}^{n+r-1} e^{-rt} \frac{(rt)^l}{l!} \\ &= \sum_{i=1}^{r} \frac{A_i}{1-x_i} x_i^{-r+1+Mr} e^{-r(1-x_i)t}, \end{aligned}$$

and thus also

$$E(S(M)) = \sum_{i=1}^{r} \frac{A_i}{r(1-x_i)^2} x_i^{-r+1+Mr}.$$

Table 9.2 shows how the production lead times are reduced by the production to stock of the standard items.

By comparing the results in table 9.1 and table 9.2 we see that in case of Erlang production times the reductions in the production lead time are much larger than in case of exponential production times. This is also illustrated in figure 9.1.

We finally note that for general phase-type production time distributions the analysis is equally simple.

		$E(S(M))$					
M		0	8	16	24	32	40
	0.8	3.2	0.140	0.006	0.000	0.000	0.000
ρ	0.9	5.95	1.32	0.291	0.064	0.014	0.003
	0.95	11.5	5.47	2.61	1.24	0.592	0.282

		$P(S(M) > 10)$					
M		0	8	16	24	32	40
	0.8	0.027	0.001	0.000	0.000	0.000	0.000
ρ	0.9	0.172	0.038	0.008	0.002	0.000	0.000
	0.95	0.420	0.200	0.096	0.046	0.022	0.010

Table 9.2: Performance characteristics for Erlang-10 production times.

9.2 Two types of jobs, standard (MTS) and customer specific (MTO)

We now turn to a model which combines make to order and make to stock. There are two types of products, standard and non-standard products. When the machine is idle (i.e. there are no orders) it is used to produce standard products until the stock reaches a a certain maximum level, M say. A customer asking for a standard product receives it directly (if possible) from stock. If the stock is empty the customer order joins the queue. Non-standard, customer specific products are never delivered from stock, but always produced to order. Demand for standard products arrives according to a Poisson process with rate λ_1, the customer specific demand is Poisson as well with rate λ_2. Further we set $\lambda = \lambda_1 + \lambda_2$. The production times for both types of products are exponential with the same mean $1/\mu$. Production to stock is preempted by production to order. For stability we assume that $\rho = \lambda/\mu < 1$. Below we will first present an exact analysis of this system and then investigate the effect of producing to stock on the production lead times of standard and non-standard orders.

The system can be described by the states (n, m) where n is the number of orders in the production queue and m is the number of items on stock. If $m > 0$ all n jobs in the queue are type 2, but if $m = 0$ the n jobs are a mixture of type 1 and type 2 jobs. The flow diagram for this model is depicted in figure 9.2.

By balancing in each state (n, m) the flow out of and into that state we get the following set of equations:

$$p(n, M)(\lambda + \mu) = p(n-1, M)\lambda_2 + p(n+1, M)\mu, \quad n > 0, \tag{9.1}$$

$$p(n, m)(\lambda + \mu) = p(n, m+1)\lambda_1 + p(n-1, m)\lambda_2 + p(n+1, m)\mu, \tag{9.2}$$
$$n > 0, 0 < m < M,$$

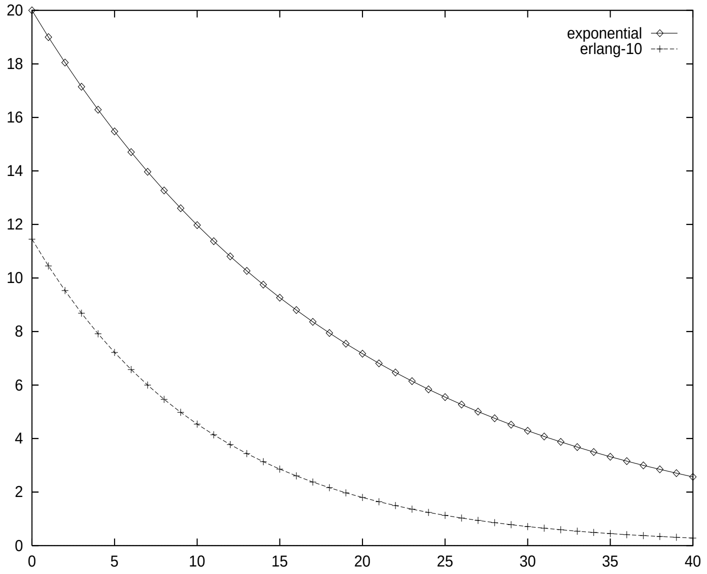

Figure 9.1: The mean production lead time as a function of the maximum stock M for $\rho = 0.95$ in case of exponential and Erlang-10 production times, resp.

$$
\begin{aligned}
p(n,0)(\lambda+\mu) &= p(n,1)\lambda_1 + p(n-1,0)\lambda + p(n+1,0)\mu, \quad n>0, && (9.3)\\
p(0,M)\lambda &= p(0,M-1)\mu + p(1,M)\mu, && (9.4)\\
p(0,m)(\lambda+\mu) &= p(0,m+1)\lambda_1 + p(0,m-1)\mu + p(1,m)\mu, && (9.5)\\
& \quad 0<m<M, &&\\
p(0,0)(\lambda+\mu) &= p(0,1)\lambda_1 + p(1,0)\mu. && (9.6)
\end{aligned}
$$

Note that in the equations (9.2) there are no probabilities $p(l,k)$ with $k<m$, since it is only possible to consume and not produce stock items in states (l,k) with $l>0$. Hence the balance equations can be solved as follows. We start to solve the equations for $m=M$ and then work our way down from $m=M-1$ to $m=0$. Equation (9.1) is a difference equation in n only. Its solution is given by

$$p(n,M) = A_{0,0}x^n,$$

where $A_{0,0}$ is (for the time being) free and x is a root of

$$x(\lambda+\mu) = \lambda_2 + x^2\mu. \tag{9.7}$$

One of the two roots of (9.7) is larger than 1, and thus not useful. The other one equals

$$x = \frac{\lambda+\mu-\sqrt{(\lambda+\mu)^2-4\lambda_2\mu}}{2\mu}.$$

This root is positive and less than 1. Substituting $p(n,M) = A_{0,0}x^n$ into equation (9.2) for $m=M-1$ we obtain an inhomogeneous recurrence relation for the probabilities $p(n,M-1)$,

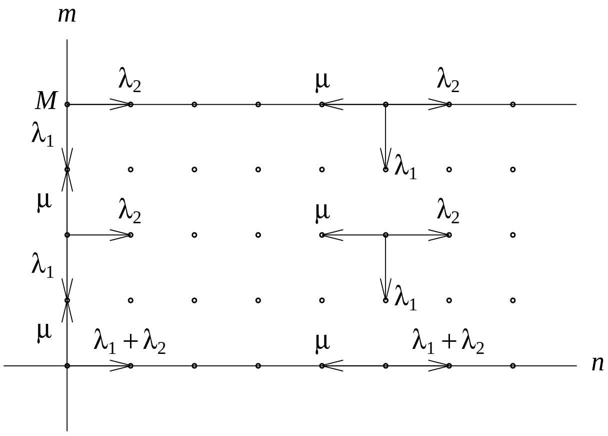

Figure 9.2: Flow diagram for the model with standard and customer specific jobs

with the $p(n, M)$'s as the inhomogeneous terms. The solution of this recurrence relation is of the form

$$p(n, M-1) = A_1 x^n + A_2 n x^n \,.$$

The first term is the solution of the homogeneous equation, the second one is a particular solution of the inhomogeneous equation. It is convenient to rewrite the expression for $p(n, M-1)$ in the form

$$p(n, M-1) = \sum_{j=0}^{1} A_{1,j} \binom{1+n}{j} x^n \,. \tag{9.8}$$

Next we substitute this expression into (9.2) for $m = M-2$ and solve the probabilities $p(n, M-2)$. Repeating this procedure we can work our way down from $M-1$ to 1. This leads to the solution

$$p(n, M-k) = \sum_{j=0}^{k} A_{k,j} \binom{k+n}{j} x^n, \quad k = 0, \ldots, M-1, \; n = 0, 1, \ldots, \tag{9.9}$$

where for $k = 1, \ldots, M-1$ the constants $A_{k,j}$ satisfy

$$A_{k,k} = \frac{\lambda_1 A_{k-1,k-1}}{\lambda + \mu - 2\mu x},$$

$$A_{k,j} = \frac{\mu x A_{k,j+1} + \lambda_1 A_{k-1,j-1}}{\lambda + \mu - 2\mu x}, \quad j = 1, \ldots, k-1.$$

The constants $A_{0,0}, A_{1,0}, \ldots, A_{M-1,0}$ are still free. Equation (9.3) is a difference equation in $p(n, 0)$ with the $p(n, 1)$'s given by (9.9) as inhomogeneous terms. The solution of this equation is given by

$$p(n, 0) = B \left(\frac{\lambda}{\mu}\right)^n + \sum_{j=0}^{M-1} A_{M,j} \binom{M+n}{j} x^n. \tag{9.10}$$

The first term with B free is the solution of the homogeneous equation. The sum is a particular solution of the inhomogeneous equation. The constants $A_{M,j}$ in this sum satisfy

$$A_{M,M-1} = -A_{M-1,M-1}x,$$

and

$$A_{M,j} = \lambda_1^{-1}\left[\{(\lambda+\mu)x - 2\mu x^2\}A_{M,j+1} - \mu x^2 A_{M,j+2} - \lambda_1 x A_{M-1,j}\right],$$

for $j = 0, \ldots, M-2$, where $A_{M,M} = 0$ by definition. We can conclude that the solution of the equations (9.1)-(9.3) is given by (9.9) and (9.10) in which the constants $A_{0,0}, \ldots, A_{M-1,0}$ and B are free. These constants can finally be determined from the boundary equations (9.4)-(9.5) together with the normalization condition.

Once the probabilities $p(n, m)$ are known, performance characteristics, such as e.g. the mean number of orders in the production queue and the mean production lead time for the two types of orders can be easily obtained. For instance, let $L(M)$ be the number of orders in the production queue. Then, by inserting the expressions for the probabilities $p(n, m)$ we find that

$$E(L(M)) = \sum_{n,m} np(n,m) = B\frac{\lambda/\mu}{1-\lambda/\mu} + \sum_{k=0}^{M}\sum_{j=0}^{k} A_{k,j}\sum_{n=0}^{\infty}\binom{k+n}{j}nx^n,$$

where the infinite sum can be easily calculated exactly (cf. (9.24). The mean overall production lead time $E(S(M))$ (i.e. for type 1 and 2 orders together) can be obtained from $E(L(M))$ by application of Little's law, i.e.,

$$E(S(M)) = E(L(M))/\lambda\,.$$

Denote by $S_1(M)$ and $S_2(M)$ the production lead time for a standard and non-standard order, respectively. From the PASTA property (see [52]) we obtain

$$E(S_2(M)) = (E(L(M)) + 1)/\mu\,.$$

Further, it holds that

$$E(S(M)) = \frac{\lambda_1}{\lambda}E(S_1(M)) + \frac{\lambda_2}{\lambda}E(S_2(M)),$$

from which $E(S_1(M))$ directly follows.

The figures 9.3 and 9.4 illustrate for $\rho = 0.9$ and $\mu = 1$ how the mean production lead times for standard and non-standard orders decrease with M for the ratios λ_1/λ_2 equal to 3, 1 and 1/3, respectively. Note that for $M = 0$ and $M = \infty$ we have

$$E(S_1(0)) = E(S_2(0)) = \frac{1/\mu}{1-\rho_1-\rho_2},$$

$$E(S_1(\infty)) = 0, \quad E(S_2(\infty)) = \frac{1/\mu}{1-\rho_2},$$

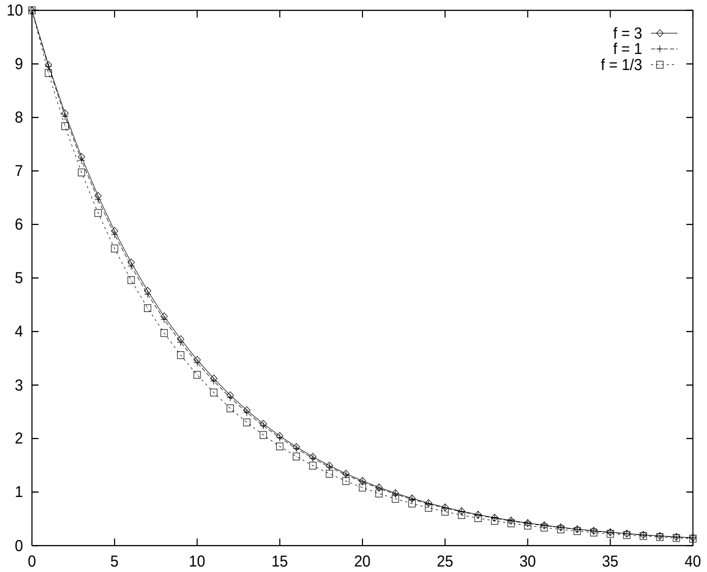

Figure 9.3: The mean production lead time for standard orders as a function of M for $\mu = 1$ and $\rho = 0.9$ where $f = \lambda_1/\lambda_2$ is varied as 3, 1 and 1/3, resp.

where $\rho_1 = \lambda_1/\mu$ and $\rho_2 = \lambda_2/\mu$.

We see that significant reductions in the lead time are possible, and that, apparently, the reductions for the standard orders are quite insensitive to the ratio λ_1/λ_2. Most of the reduction, 80% say, is already achieved for moderate M. This is also shown in table 9.3 where we list the minimal M for which α% of the possible reduction in the lead time is realized, i.e.

$$E(S_1(0)) - E(S_1(M)) \geq \frac{\alpha}{100} E(S_1(0)) ,$$

$$E(S_2(0)) - E(S_2(M)) \geq \frac{\alpha}{100} (E(S_1(0)) - E(S_1(\infty))) .$$

Again observe that the minimal M is quite insensitive to the ratio λ_1/λ_2.

The figures 9.3 and 9.4 and table 9.3 illustrate the reductions in the lead time achieved for given *maximum* stock M. The maximum stock size is important for the allocation of space needed for storage, the average stock is relevant for the calculation of holding costs. The average stock can be determined as follows. There is a simple relation between the probability $p(0, m-1)$ and the probability $q(m)$ that there are m items on stock. Namely, balancing the flow between the sets $\{(0, m-1), (1, m-1), \ldots\}$ and $\{(0, m), (1, m), \ldots\}$ yields

$$p(0, m-1)\mu = q(m)\lambda_1 , \quad m = 1, \ldots, M .$$

From this relation we can compute the probabilities $q(m)$ and thus also the average stock. In figure 9.5 we show for $\rho = 0.9$ and $\mu = 1$ the average stock as a function of M for the

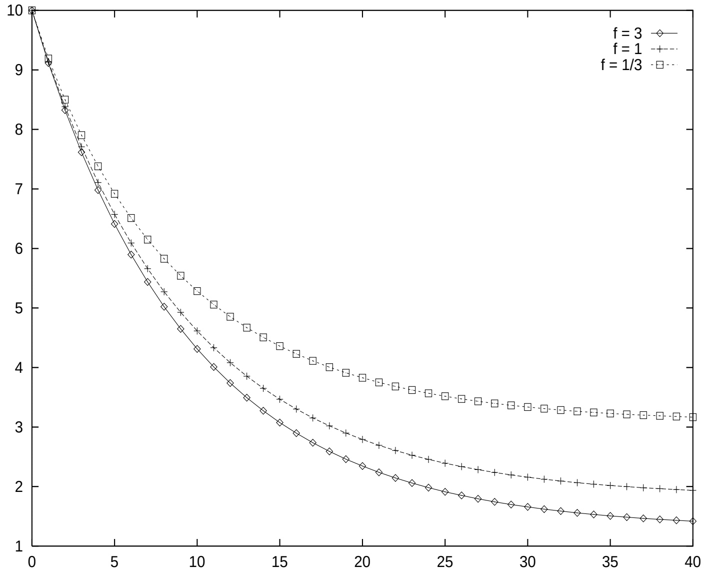

Figure 9.4: The mean production lead time for non-standard orders as a function of M for $\mu = 1$ and $\rho = 0.9$ where $f = \lambda_1/\lambda_2$ is varied as 3, 1 and 1/3, resp.

ratios λ_1/λ_2 equal to 3, 1 and 1/3, respectively. Observe that from $M = 20$ say, the lines are nearly straight with slope 1.

It is also possible to compute the distribution of the production lead time from the equilibrium probabilities $p(n,m)$. The production lead time $S_1(M)$ for a standard order is the sum of $n+1$ exponentials with mean $1/\mu$ with probability $p(n,0)$ and it is equal to 0 otherwise. For a non-standard order the production lead time $S_2(M)$ is equal to the sum of $n+1$ exponentials with probability $p(n,m)$. Hence, for $t \geq 0$ we obtain that

$$P(S_1(M) > t) = \sum_{n=0}^{\infty} p(n,0) \sum_{l=0}^{n} e^{-\mu t} \frac{(\mu t)^l}{l!},$$

$$P(S_2(M) > t) = \sum_{n=0}^{\infty} \sum_{m=0}^{M} p(n,m) \sum_{l=0}^{n} e^{-\mu t} \frac{(\mu t)^l}{l!}.$$

Table 9.4 lists the percentages of the standard and non-standard orders that will get a production lead time more than 10 times the mean production time.

9.3 Half-finished products to stock

In this section we consider products that are produced in two phases. The first phase is standard and identical for all products. The second phase is customer specific. We will analyze how the production lead time decreases if we store some, at most M, half-finished products, that is products with phase 1 completed. The first phase of a product takes an exponential time with mean $1/\mu_1$, the second phase is exponential with mean $1/\mu_2$. Orders

		M			
$\lambda_1/\lambda_2 = 3$	α	20	50	80	90
	0.8	2	5	9	12
ρ	0.9	4	8	17	23
	0.95	6	15	33	46
$\lambda_1/\lambda_2 = 1$					
	0.8	2	4	8	11
ρ	0.9	4	8	17	23
	0.95	6	15	33	46
$\lambda_1/\lambda_2 = 1/3$					
	0.8	2	4	7	10
ρ	0.9	3	7	16	22
	0.95	6	15	32	46

Table 9.3: Minimal M to achieve $\alpha\%$ of the possible reduction in the lead time

arrive for one item at a time according to a Poisson process with rate λ. For stability we assume that

$$\frac{\lambda}{\mu_1} + \frac{\lambda}{\mu_2} < 1. \tag{9.11}$$

The system can be described by the states (n, m) with n the number of orders in the system and m the number of half-finished products in stock or in use. So state $(1, 2)$ denotes the situation with 1 job in the system for which the machine is processing phase 2, and 1 half-finished product on the shelf. If phase 2 is completed the state changes to $(0, 1)$ and the machine continues with the production of phase 1 products until the limit M is reached, or a new customer order enters. So, a newly arriving order preempts a stock production. In state $(n, 0)$ with n positive the machine is working on phase 1. If phase 1 is completed the state changes to $(n, 1)$ and the machine continues with phase 2. The flow diagram of this Markov process is depicted in figure 9.6.

By equating in each state (n, m) the flow out of and into that state we get:

$$p(n, M)(\lambda + \mu_2) = p(n-1, M)\lambda, \quad n > 0, \tag{9.12}$$

$$p(n, m)(\lambda + \mu_2) = p(n-1, m)\lambda + p(n+1, m+1)\mu_2, \quad n > 0, 1 < m < M, \tag{9.13}$$

$$p(n, 1)(\lambda + \mu_2) = p(n-1, 1)\lambda + p(n, 0)\mu_1 + p(n+1, 2)\mu_2, \quad n > 0, \tag{9.14}$$

$$p(n, 0)(\lambda + \mu_1) = p(n-1, 0)\lambda + p(n+1, 1)\mu_2, \quad n > 0, \tag{9.15}$$

$$p(0, M)\lambda = p(0, M-1)\mu_1, \tag{9.16}$$

$$p(0, m)(\lambda + \mu_1) = p(0, m-1)\mu_1 + p(1, m+1)\mu_2, \quad 2 < m < M, \tag{9.17}$$

$$p(0, 2)(\lambda + \mu_1) = p(0, 1)\mu_1 + p(1, 3)\mu_2, \tag{9.18}$$

$\lambda_1/\lambda_2 = 3$		M	$P(S_1(M) > 10) \times 100\%$				
			0	8	16	24	32
	0.8		14	2.3	0.38	0.06	0.01
ρ	0.9		37	16	6.8	2.9	1.3
	0.95		61	40	27	18	12
$\lambda_1/\lambda_2 = 1$							
	0.8		14	2.3	0.38	0.06	0.01
ρ	0.9		37	16	6.8	2.9	1.3
	0.95		61	40	27	18	12
$\lambda_1/\lambda_2 = 1/3$							
	0.8		14	2.1	0.32	0.05	0.01
ρ	0.9		37	15	6.5	2.8	1.2
	0.95		61	40	26	17	12

$\lambda_1/\lambda_2 = 3$		M	$P(S_2(M) > 10) \times 100\%$					
			0	8	16	24	32	∞
	0.8		14	2.3	0.41	0.10	0.04	0.03
ρ	0.9		37	16	6.8	3.0	1.3	0.04
	0.95		61	40	27	18	12	0.05
$\lambda_1/\lambda_2 = 1$								
	0.8		14	2.4	0.61	0.31	0.26	0.25
ρ	0.9		36	16	7.1	3.3	1.6	0.41
	0.95		61	40	27	18	12	0.52
$\lambda_1/\lambda_2 = 1/3$								
	0.8		14	3.4	2.1	1.9	1.8	1.8
ρ	0.9		37	17	9.4	6.2	4.9	3.9
	0.95		61	41	29	21	16	5.6

Table 9.4: Percentages of the standard and non-standard orders with a lead time longer than 10 times the mean production time

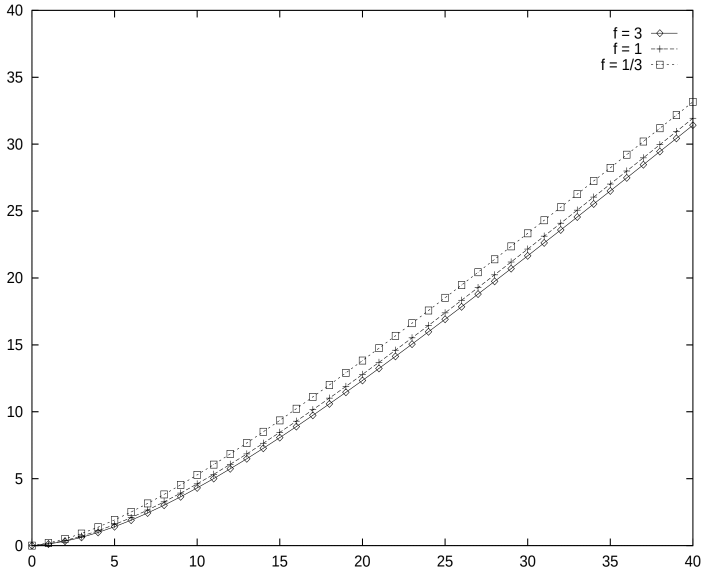

Figure 9.5: The average stock as a function of M for $\mu = 1$ and $\rho = 0.9$ where $f = \lambda_1/\lambda_2$ is varied as 3, 1 and 1/3, resp.

$$p(0,1)(\lambda+\mu_1) \;=\; p(0,0)\mu_1 + p(1,2)\mu_2, \tag{9.19}$$
$$p(0,0)(\lambda+\mu_1) \;=\; p(1,1)\mu_2. \tag{9.20}$$

The approach to solve the equations (9.12)-(9.20) is similar to the one in the previous section. We start with the equations for $m = M$ and then subsequently solve the equations for $m = M-1$, $m = M-2$ and so on. Equation (9.12) is a difference equation in n only. Its solution is given by

$$p(n, M) = A_{0,0} x^n\,,$$

where $A_{0,0}$ is free and

$$x = \frac{\lambda}{\lambda+\mu_2}\,.$$

By substituting $p(n,M) = A_{0,0}x^n$ into (9.13) for $m = M-1$ we get an inhomogeneous recurrence relation for the probabilities $p(n, M-1)$, with the $p(n,M)$ as inhomogeneous terms. The solution of this recurrence relation can be written as (cf. (9.8))

$$p(n, M-1) = \sum_{j=0}^{1} A_{1,j} \binom{1+n}{j} x^n\,.$$

Next we insert this expression into (9.13) for $m = M-2$. This yields an inhomogeneous recurrence relation for the probabilities $p(n, M-2)$, which can readily be solved. By repeating this procedure for $m = M-3$ down to $m = 2$ we obtain

$$p(n, M-k) = \sum_{j=0}^{k} A_{k,j} \binom{n+j}{j} x^n\,, \quad k = 0,\ldots,M-2,\; n = 0,1,\ldots, \tag{9.21}$$

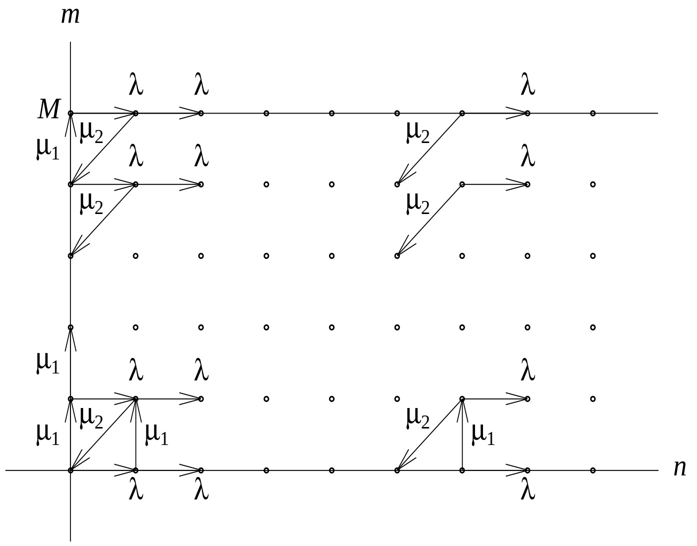

Figure 9.6: Flow diagram for the model with half-finished products

where for $k = 1, \ldots, M-2$ the constants $A_{k,j}$ satisfy

$$A_{k,k} = \frac{x\mu_2}{\lambda + \mu_2} A_{k-1,k-1} \,,$$

$$A_{k,j} = A_{k,j+1} + \frac{x\mu_2}{\lambda + \mu_2} A_{k-1j-1} \,, \quad j = k-1, \ldots, 1 \,.$$

This completes the solution of the equations (9.12)-(9.13). The constants $A_{0,0}, \ldots, A_{M-2,0}$ in this solution are still free. We now consider the equations (9.14)-(9.15). Note that these equations are difference equations in $p(n, 0)$ and $p(n, 1)$ with the $p(n, 2)$'s given by (9.21) as inhomogeneous terms. The solution is given by

$$p(n,1) = B_1 y_1^n + B_2 y_2^n + \sum_{j=0}^{M-2} A_{M-1,j} \binom{n+j}{j} x^n \,, \tag{9.22}$$

$$p(n,0) = B_1 \frac{\lambda + \mu_2 - \lambda y_1^{-1}}{\mu_1} y_1^n + B_2 \frac{\lambda + \mu_2 - \lambda y_2^{-1}}{\mu_1} y_2^n + \sum_{j=0}^{M-2} A_{M,j} \binom{n+j}{j} x^n \,. \tag{9.23}$$

The first two terms in (9.22)-(9.23) are the solution of the homogeneous equations, where B_1 and B_2 are free and y_1 and y_2 are the two roots of

$$\mu_1 \mu_2 y^2 - [\lambda^2 + \lambda(\mu_1 + \mu_2)] y + \lambda^2 = 0 \,.$$

Condition (9.11) guarantees that the roots y_1 and y_2 are both positive and less than 1. The two sums in (9.22)-(9.23) are a particular solution of the inhomogeneous equations. Note that the particular solution has the same form as the inhomogeneous term $p(n,2)$ given by (9.21). For the constants $A_{M,j}$ and $A_{M-1,j}$ in the two sums the following relations can be derived. The constants $A_{M,j}$ in (9.23) satisfy

$$A_{M,M-2} = -\frac{\mu_2 x}{\mu_1} A_{M-2,M-2} \,,$$

and

$$\begin{aligned} A_{M,j} &= \frac{\lambda x^{-1}}{\mu_1 \mu_2 x} \left[(\mu_1 - \mu_2) A_{M,j+1} + \lambda x^{-1} A_{M,j+2} \right] \\ &\quad - \sum_{l=j+1}^{M-2} A_{M,l} - \frac{\mu_2 y}{\mu_1} \sum_{l=j}^{M-2} (l+1-j) A_{M-2,l} \,, \end{aligned}$$

for $j = M-3, \ldots, 0$, where $A_{M,M-1} = 0$ by definition. The constants $A_{M-1,j}$ in expression (9.22) follow from the relations

$$A_{M-1,j} = \frac{1}{\lambda x^{-1}} \left[\mu_1 A_{M,j-1} + \mu_2 x \sum_{m=j-1}^{M-2} A_{M-2,m} \right]$$

for $j = 1, \ldots, M-2$ and

$$A_{M-1,0} = \frac{1}{\mu_2 x^{-1}} \left[(\lambda + \mu_1 - \lambda x^{-1}) A_{M,0} + \lambda x^{-1} A_{M,1} - \mu_2 x \sum_{l=1}^{M-2} A_{M-1,l} \right] .$$

We can now conclude that the solution of the equations (9.12)-(9.17) is given by (9.21)-(9.23) in which the constants $A_{0,0}, \ldots, A_{M-2,0}$ together with B_1 and B_2 are free. These constants can finally be solved from the boundary equations (9.16)-(9.20) and the normalization condition.

It is straightforward to determine performance characteristics, such as the mean number of orders in the system and the mean lead time from the equilibrium probabilities $p(n,m)$. For instance, for the mean number of orders in the system we obtain

$$\begin{aligned} E(L(M)) &= \left(1 + \frac{\lambda + \mu_2 - \lambda y_1^{-1}}{\mu_1}\right) \frac{B_1 y_1}{(1-y_1)^2} + \left(1 + \frac{\lambda + \mu_2 - \lambda y_2^{-1}}{\mu_1}\right) \frac{B_2 y_2}{(1-y_2)^2} \\ &\quad + \sum_{k=0}^{M-2} \sum_{j=0}^{k} A_{k,j} \frac{(1+j)x}{(1-x)^{2+j}} \\ &\quad + \sum_{j=0}^{M-2} A_{M-1,j} \frac{(1+j)x}{(1-x)^{2+j}} + \sum_{j=0}^{M-2} A_{M,j} \frac{(1+j)x}{(1-x)^{2+j}} \,, \end{aligned}$$

where we used that

$$\sum_{n=0}^{\infty} \binom{n+j}{j} n x^n = \frac{(1+j)x}{(1-x)^{2+j}} \,. \tag{9.24}$$

The mean production lead time follows from $E(L(M))$ and Little's law, i.e.,

$$E(S(M)) = E(L(M))/\lambda \,.$$

Note that for $M = 0$ and $M = \infty$ it holds that

$$E(S(0)) = \frac{1/\mu_1 + 1/\mu_2 - \rho_2/\mu_1}{1 - \rho_1 - \rho_2} \,,$$

$$E(S(\infty)) = \frac{1/\mu_2}{1 - \rho_2} \,,$$

where $\rho_1 = \lambda/\mu_1$ and $\rho_2 = \lambda/\mu_2$. Figure 9.7 illustrates how the mean production lead time decreases with M for the ratios μ_1/μ_2 equal to 3, 1 and 1/3, respectively. In each case the mean production time (i.e., $1/\mu_1 + 1/\mu_2$) is equal to 1.

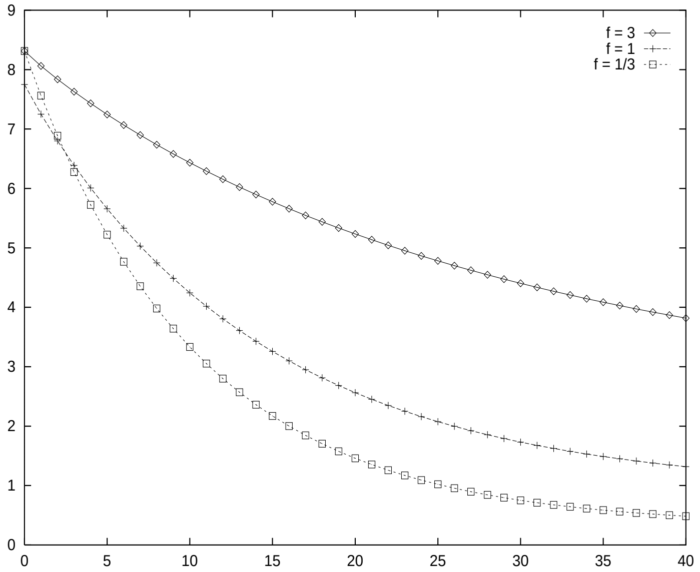

Figure 9.7: The mean production lead time as a function of M for mean production time 1 and $\rho = 0.9$ where $f = \mu_1/\mu_2$ is varied as 3, 1 and 1/3, resp.

We see that significant reductions in the lead time are possible. The reduction is large if phase 1 constitutes a large part of the total production time, thus if μ_1/μ_2 is small. In table 9.5 we list the minimal M for which α% of the possible reduction in the lead time is realized, i.e.

$$E(S(0)) - E(S(M)) \geq \frac{\alpha}{100}(E(S(0)) - E(S(\infty)) \,.$$

Figure 9.7 and table 9.5 show the reductions in the lead time for given maximum stock level M. It is interesting to know the average stock in these cases. To compute the average stock we can use the following relation between the probability $q(m)$ that there are m

$\mu_1/\mu_2 = 3$	α	M 20	50	80	90
	0.8	3	8	20	29
ρ	0.9	6	20	48	69
	0.95	14	43	102	146
$\mu_1/\mu_2 = 1$					
	0.8	2	5	11	16
ρ	0.9	4	10	23	33
	0.95	7	21	48	68
$\mu_1/\mu_2 = 1/3$					
	0.8	2	4	8	12
ρ	0.9	3	8	17	24
	0.95	5	15	34	49

Table 9.5: Minimal M to achieve α% of the possible reduction in the lead time

half-finished products in stock and $p(0, m-1)$. By balancing the flow between the sets $\{(0, m-1), (1, m-1), \ldots\}$ and $\{(0, m), (1, m), \ldots\}$ it follows that

$$p(0, m-1)\mu_1 = (q(m) - p(0, m))\mu_2 , \quad m = 2, \ldots, M .$$

Also note that $q(1) = p(0, 1)$. From this relation we can calculate the probabilities $q(m)$, and thus also the average stock. Figure 9.8 illustrates the average stock for the examples in figure 9.7. We see that for $\mu_1/\mu_2 = 1, 1/3$ the lines are straight with slope 1 from $M = 20$ say, so that from there on an extra half-finished product on stock will not leave the shelf. For $\mu_1/\mu_2 = 3$, however, it will help a little bit to produce extra half-finished products, even when there are already 40 items on stock.

It is also possible to compute the distribution of the production lead time from the equilibrium probabilities $p(n, m)$. Let $S(n, m)$ be the lead time if on arrival the system is in state (n, m). So, if $n + 1 \leq m$, then $S(n, m)$ is the sum of $n + 1$ phase 2 productions (and no phase 1 productions), and otherwise, $S(n, m)$ is the sum of $n + 1 - m$ phase 1 and $n + 1$ phase 2 productions. Since an arriving order finds the system with probability $p(n, m)$ in state (n, m) we obtain for $t \geq 0$ that

$$P(S(M) > t) = \sum_{n,m} p(n, m)P(S(n, m) > t) .$$

From this relation it is straightforward to calculate the probability $P(S(M) > t)$ once the probabilities $p(n, m)$ are known. This is done for the examples in table 9.6 where the percentages of the orders that will get a production lead time more than 10 times the mean production time are listed.

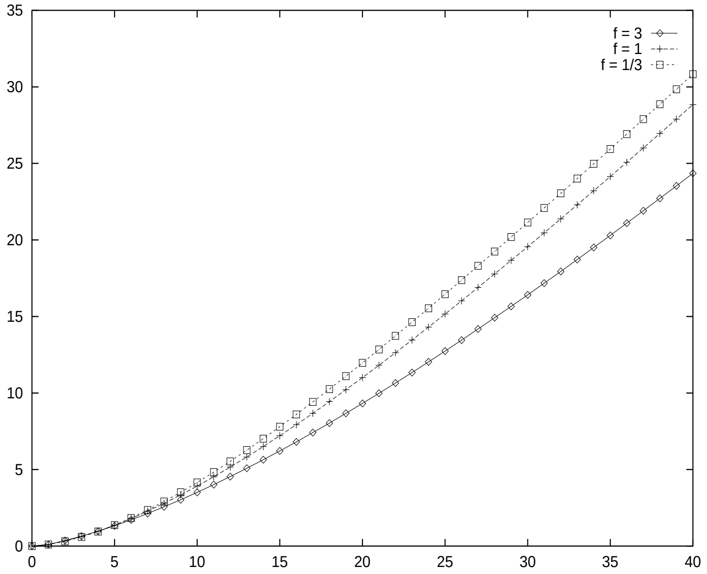

Figure 9.8: The average stock as a function of M for mean production time 1 and $\rho = 0.9$ where $f = \mu_1/\mu_2$ is varied as 3, 1 and 1/3, resp.

			$P(S(M) > 10) \times 100\%$				
$\mu_1/\mu_2 = 3$		M	0	8	16	24	32	∞
	0.8		8.8	5.3	3.2	2.0	1.4	0.48
ρ	0.9		30	23	18	14	11	1.3
	0.95		55	48	43	38	34	2.2
$\mu_1/\mu_2 = 1$								
	0.8		7.2	2.2	0.68	0.21	0.07	0.00
ρ	0.9		27	16	8.9	5.1	2.9	0.00
	0.95		52	40	30	23	18	0.00
$\mu_1/\mu_2 = 1/3$								
	0.8		8.8	1.7	.32	0.06	0.01	0.00
ρ	0.9		30	14	6.3	2.9	1.3	0.00
	0.95		55	37	26	18	12	0.00

Table 9.6: Percentages of the orders with a lead time longer than 10 times the mean production time

Chapter 10

Assigning jobs to machines

In this chapter we consider a production system consisting of several parallel and identical machines, each with its own buffer with infinite capacity. Jobs which have to be processed by this system have to be assigned to one of machines immediately upon arrival, thereafter to be processed on a first-come-first-served basis. We assume that on arrival the only information available is the number of jobs waiting and being processed at each machine. So the problem is to decide, based on this information, which queue to join. Then the obvious candidate is to join the shortest queue, but is it optimal? The shortest queue rule may be a good rule to use, but it need *not* minimize the individual expected delay of each job or the expected delay over all jobs. It is known to be optimal when the processing times are exponentially distributed or have a nondecreasing failure rate which means that the longer the job has been in process, the more likely it is completed shortly (see [51, 47, 22]). But in [49] it is shown that there also distributions of the processing times for which it is not optimal to join the shortest queue.

The aim of this chapter is not to identify an optimal rule, but we want to compare for a simple model the expected delays when *dynamic rules* like joining the shortest queue are used with the delays under *static rules* like random or round robin assignment. And also, how good are these rules compared to the situation in which there is a central buffer?

The model we consider throughout this chapter consists of two parallel and identical machines, each with its own buffer (see figure 10.1). The processing times are exponentially distributed with mean 1. Jobs arrive according to a Poisson stream with rate 2ρ where we assume that $0 < \rho < 1$. Immediately upon arrival jobs must be assigned to one of the machines. In the following two sections we first consider two static rules. Then we will analyze the situation in which the shortest queue rule is used.

10.1 Random assignment

Suppose that an arriving job joins either queue with probability 1/2. So the Poisson stream is split up into two streams, both Poisson again and rate ρ (see section 4.4). Hence, if we

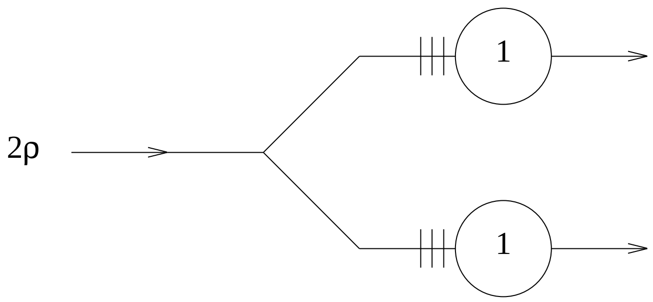

Figure 10.1: The model with two parallel machines, each with its own buffer

denote the mean production lead time by $E(S)$ we obtain

$$E(S) = \frac{1}{1-\rho}.$$

10.2 Round robin assignment

Now consider the situation in which arriving jobs are alternatingly sent to machine 1 and 2. So the first one is sent to machine 1, the second one to machine 2, the third again to 1, the fourth to 2 and so on. Let us look at the arrival process in queue 1 say (the situation in queue 2 is the same). The interarrival times are composed of two independent exponential times, each with mean $1/2\rho$, and thus are Erlang-2 distributed. Hence queue 1 (and 2) can be described by an $E_2/M/1$ queue, so the mean production lead time is given by (see section 7.2)

$$E(S) = \frac{1}{1-x},$$

where x is the unique positive root, less than unity, of

$$x = \left(\frac{2\rho}{2\rho + 1 - x}\right)^2.$$

10.3 Shortest queue routing

In this section we analyze the model where jobs join the shortest queue, and if the queues have equal length, join either queue with probability 1/2. It will appear that the analysis of this model is much more complicated than its simple formulation suggests.

The state of the system can be described by the pair (i, j) where i and j are the lengths of the two queues. But it is somewhat simpler to use instead of i and j the state variables m and n where $m = \min(i, j)$ and $n = j - i$. So m is the length of the shortest queue and n is the difference between the queue lengths. Note that the present model has *two real dimensions*, i.e. both variables m and n are unbounded. So far we have only seen models which are essentially one-dimensional. The treatment of two-dimensional models is in general much more complicated.

Let $p(m,n)$ be the equilibrium probability for state (m,n). The flow diagram is depicted in figure 10.2. The rates in the region $n \leq 0$ can be obtained by reflection in the m-axis. By symmetry $p(m,n) = p(m,-n)$, so the analysis can be restricted to the probabilities $p(m,n)$ in the region $n \geq 0$.

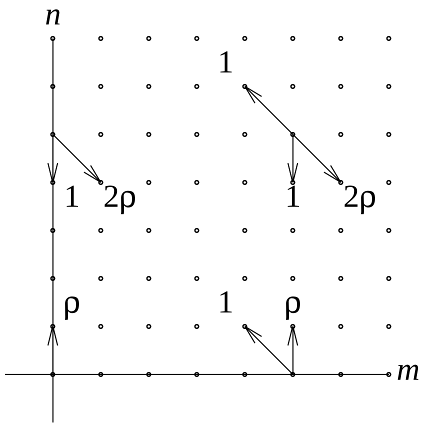

Figure 10.2: Flow diagram for the shortest queue model

The balance equations for the probabilities $p(m,n)$ can be found by equating for each state the rate into and the rate out of that state. In the equations below we have eliminated the probabilities $p(m,0)$ from (10.2) and (10.4) by substituting (10.5) and (10.6). This is done to simplify the presentation. The analysis can now be restricted to the probabilities $p(m,n)$ with $n > 0$. These probabilities satisfy equations (10.1)–(10.4). The equations (10.5)–(10.6) are further treated as definition for $p(m,0)$.

$$\begin{aligned}
p(m,n)2(\rho+1) &= p(m-1,n+1)2\rho + p(m,n+1) + p(m+1,n-1), \\
&\qquad m>0, n>1, && (10.1)\\
p(m,1)2(\rho+1) &= p(m-1,2)2\rho + p(m,2) \\
&\quad +(p(m,1)2\rho + p(m+1,1))\frac{1}{\rho+1} \\
&\quad +(p(m-1,1)2\rho + p(m,1))\frac{\rho}{\rho+1}, \qquad m>0, && (10.2)\\
p(0,n)(2\rho+1) &= p(0,n+1) + p(1,n-1), \qquad n>1, && (10.3)\\
p(0,1)(2\rho+1) &= p(0,2) + (p(0,1)2\rho + p(1,1))\frac{1}{\rho+1} + p(0,1), && (10.4)\\
p(m,0)(\rho+1) &= p(m-1,1)2\rho + p(m,1), \qquad m>0, && (10.5)\\
p(0,0)\rho &= p(0,1). && (10.6)
\end{aligned}$$

So the problem is to find the solution to find solutions to this set of *partial difference* *equations*. The usual approaches are based on generating functions, see [26, 19, 14, 16, 17, 23]. These approaches require much work and do not lead to very explicit results.

In this section we show how the empirical finding that the probabilities $p(m,n)$ approximately behave as a product of powers already at a short distance from the boundaries, can be exploited to develop a so-called compensation approach by which the probabilities can be found efficiently. This approach leads to an explicit characterization of the probabilities in the form of an infinite sum of products, i.e., there are α_i, β_i and c_i such that

$$p(m,n) = \sum_{i=0}^{\infty} c_i \alpha_i^m \beta_i^n.$$

Based on this expression for the probabilities $p(m,n)$, it can be shown that also performance criteria, such as the mean waiting time, can be expressed as infinite sums of products. Moreover, the results can be easily exploited for numerical calculations.

Numerical experiments show that the probabilities $p(m,n)$ behave as a product $K\alpha^m\beta^n$ already at a short distance from the boundaries. This feature is illustrated in table 10.1 for the case $\rho = 0.5$ by displaying the ratios $p(m+1,n)/p(m,n)$ and $p(m,n+1)/p(m,n)$.

In table 10.1 we see that for almost all m and n the ratios $p(m+1,n)/p(m,n)$ and $p(m,n+1)/p(m,n)$ are constant, which suggests that for some constant K,

$$p(m,n) \approx K\alpha^m\beta^n, \qquad n > 0, m \to \infty, \tag{10.7}$$

where $\alpha = 0.25$ and $\beta = 0.1$ for the case $\rho = 0.5$. The first question that arises is: what are α and β in general? To obtain α, consider the process on the aggregate states k where k is the total number of jobs in the system. An approximation of the flow diagram is depicted in figure 10.3.

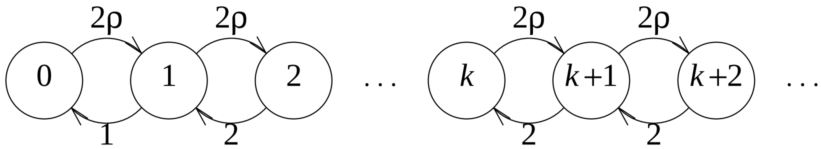

Figure 10.3: Approximation of the flow diagram for aggregated states k where k is the total number of jobs in the system

In fact, this is the flow diagram for the $M/M/2$ system with arrival rate 2ρ and service rate 1 for both servers. It is an approximation for the shortest queue system, since the average service rate in state $k > 1$ is less than 2 due to the fact that one of the servers can be idle. Intuitively, it will be obvious that the average service rate tends to 2 as k tends to infinity, so the approximation is better for large k. From the flow diagram we obtain for the equilibrium probabilities $P(k)$ that for large k

$$P(k) \approx C\rho^k, \tag{10.8}$$

↑	n	0	1	2	3	4	5
↑	6	0.19	0.24	0.25	0.25	0.25	0.25
n	5	0.19	0.24	0.25	0.25	0.25	0.25
	4	0.19	0.24	0.25	0.25	0.25	0.25
	3	0.19	0.24	0.25	0.25	0.25	0.25
	2	0.20	0.24	0.25	0.25	0.25	0.25
	1	0.28	0.25	0.25	0.25	0.25	0.25
	0						
		0	1	2	3	4	5
						$m \rightarrow$	

The ratios $p(m+1,n)/p(m,n)$

↑	6	0.10	0.10	0.10	0.10	0.10	0.10
n	5	0.10	0.10	0.10	0.10	0.10	0.10
	4	0.10	0.10	0.10	0.10	0.10	0.10
	3	0.10	0.10	0.10	0.10	0.10	0.10
	2	0.11	0.10	0.10	0.10	0.10	0.10
	1	0.15	0.11	0.10	0.10	0.10	0.10
	0						
		0	1	2	3	4	5
						$m \rightarrow$	

The ratios $p(m,n+1)/p(m,n)$

Table 10.1: The ratios $p(m+1,n)/p(m,n)$ and $p(m,n+1)/p(m,n)$ for $\rho = 0.5$

for some constant C. On the other hand, from (10.7) and using that empirically $\beta < \alpha$, for large k we get

$$P(2k-1) = 2\sum_{l=1}^{k} p(k-l, 2l-1) \approx 2K\alpha^k \sum_{l=1}^{k} \frac{\beta^{2l-1}}{\alpha^l} \approx 2K\alpha^k \sum_{l=1}^{\infty} \frac{\beta^{2l-1}}{\alpha^l}. \tag{10.9}$$

Relations (10.8) and (10.9) suggest that

$$\alpha = \rho^2,$$

which agrees with the empirical value $\alpha = 0.25$ for $\rho = 0.5$. The parameter β can be found by observing that if the product 10.7 describes the asymptotic behavior as $m \to \infty$, then it has to satisfy equations (10.1) and (10.2) for $m > 0$. Inserting $K\alpha^m\beta^n$ into (10.1) and then dividing by the common factor $K\alpha^{m-1}\beta^{n-1}$ leads to a *quadratic equation* for α and β.

Lemma 10.3.1 *The product $K\alpha^m\beta^n$ satisfies (10.1) if*

$$\alpha\beta 2(\rho+1) = \beta^2 2\rho + \alpha\beta^2 + \alpha^2. \tag{10.10}$$

Substituting $\alpha = \rho^2$ in (10.10) leads to a quadratic equation in β with roots $\beta = \rho$ and $\beta = \rho^2/(2+\rho)$. The first root yields the asymptotic solution $p(m,n) \approx K\rho^{2m}\rho^n$ for some K, corresponding to the equilibrium distribution of two independent $M/M/1$ queues, each with workload ρ. The queues of the shortest queue model, however, are strongly dependent. Therefore, the only sensible choice is

$$\beta = \frac{\rho^2}{2+\rho},$$

which agrees with the empirical value $\beta = 0.1$ for $\rho = 0.5$. It can easily be verified that for these values of α and β equation (10.2) is also satisfied. Hence, we find that for some K

$$p(m,n) \approx K\alpha_0^m\beta_0^n, \qquad n > 0, m \to \infty \tag{10.11}$$

with

$$\alpha_0 = \rho^2, \qquad \beta_0 = \frac{\rho^2}{2+\rho}. \tag{10.12}$$

Actually, the asymptotic result is rigorously proved in [26, 19]. We now come to the important question of how to exploit this asymptotic result to obtain better approximations. The product $\alpha_0^m\beta_0^n$ does not describe the behavior near the vertical axis $m = 0$, as can be seen in table 10.1 for $\rho = 0.5$. Indeed $\alpha_0^m\beta_0^n$ violates equation (10.3) for $m = 0$. The idea to improve the initial approximation $\alpha_0^m\beta_0^n$ is:

> Try to find c_1, α, β with α, β satisfying (10.10) such that $\alpha_0^m\beta_0^n + c_1\alpha^m\beta^n$ satisfies (10.3).

Inserting this linear combination into (10.3) yields the condition:

$$A(\alpha_0, \beta_0)\beta_0^{n-1} + c_1 A(\alpha, \beta)\beta^{n-1} = 0, \qquad n > 1 \tag{10.13}$$

with

$$A(x, y) = y(2\rho + 1) - y^2 - x.$$

Since $A(\alpha_0, \beta_0) \neq 0$ and condition (10.13) must hold for *all* $n > 1$, we have to take

$$\beta = \beta_0$$

and thus

$$\alpha = \alpha_1,$$

where α_1 is the second, smaller root of (10.10) with $\beta = \beta_0$ (the other root being α_0). The coefficient c_1 can now be solved from (10.13) with $\beta = \beta_0$, yielding

$$c_1 = -\frac{A(\alpha_0, \beta_0)}{A(\alpha_1, \beta_0)} = -\frac{\beta_0(2\rho + 1) - \beta_0^2 - \alpha_0}{\beta_0(2\rho + 1) - \beta_0^2 - \alpha_1}. \tag{10.14}$$

Since α_0 and α_1 are roots of (10.10) for $\beta = \beta_0$, we have

$$\alpha_0 + \alpha_1 = \beta_0 2(\rho + 1) - \beta_0^2,$$

so (10.14) can be simplified to

$$c_1 = -\frac{\alpha_1 - \beta_0}{\alpha_0 - \beta_0}.$$

For this choice of c_1 the sum $\alpha_0^m \beta_0^n + c_1 \alpha^m \beta^n$ satisfies (10.3) and, of course, also (10.1). This procedure can be generalized as follows.

Lemma 10.3.2 *Let x_1 and x_2 be the roots of the quadratic equation (10.10) for fixed β. Then the linear combination $x_1^m \beta^n + c x_2^m \beta^n$ satisfies equations (10.1) and (10.3) if c is given by*

$$c = -\frac{x_2 - \beta}{x_1 - \beta}. \tag{10.15}$$

For the case $\rho = 0.5$ we display in table 10.2 the same ratios as in table 10.1 for the approximation $\alpha_0^m \beta_0^n + c_1 \alpha_1^m \beta_0^n$.

Comparing tables 10.1 and 10.2 we see that $\alpha_0^m \beta_0^n + c_1 \alpha_1^m \beta_0^n$ also describes the behavior of the probabilities near the boundary $m = 0$. Hence, we find

$$p(m, n) \approx K(\alpha_0^m \beta_0^n + c_1 \alpha_1^m \beta_0^n), \qquad n > 0, m + n \to \infty,$$

for some K. In fact, this statement, which is stronger than (10.11), has been proved in [19]. We added $c_1 \alpha_1^m \beta_0^n$ to compensate for the error on the vertical boundary $m = 0$ and by doing so introduced a new error on the horizontal boundary $n = 1$, since this term violates condition (10.2). Since $\alpha_1 < \alpha_0$, the term $c_1 \alpha_1^m \beta_0^n$ is very small compared to $\alpha_0^m \beta_0^n$ even for small m. Therefore its disturbing effect near the horizontal boundary is negligible. However, we can compensate for the error of $c_1 \alpha_1^m \beta_0^n$ on the horizontal boundary by again adding a term:

$\uparrow$ n						
6	0.19	0.24	0.25	0.25	0.25	0.25
5	0.19	0.24	0.25	0.25	0.25	0.25
4	0.19	0.24	0.25	0.25	0.25	0.25
3	0.19	0.24	0.25	0.25	0.25	0.25
2	0.19	0.24	0.25	0.25	0.25	0.25
1	0.19	0.24	0.25	0.25	0.25	0.25
0						
	0	1	2	3	4	5
					$m \rightarrow$	

The ratios $p(m+1,n)/p(m,n)$

$\uparrow$ n						
6	0.10	0.10	0.10	0.10	0.10	0.10
5	0.10	0.10	0.10	0.10	0.10	0.10
4	0.10	0.10	0.10	0.10	0.10	0.10
3	0.10	0.10	0.10	0.10	0.10	0.10
2	0.10	0.10	0.10	0.10	0.10	0.10
1	0.10	0.10	0.10	0.10	0.10	0.10
0						
	0	1	2	3	4	5
					$m \rightarrow$	

The ratios $p(m,n+1)/p(m,n)$

Table 10.2: The ratios $p(m+1,n)/p(m,n)$ and $p(m,n+1)/p(m,n)$ for the approximation $p(m,n) = \alpha_0^m \beta_0^n + c_1 \alpha_1^m \beta_0^n$ and the case $\rho = 0.5$

Try to find α, β, d_1 with α, β satisfying (10.10) such that
$c_1\alpha_1^m\beta_0^n + d_1c_1\alpha^m\beta^n$ satisfies (10.2).

If we succeed, then the total sum $\alpha_0^m\beta_0^n+c_1\alpha_1^m\beta_0^n+d_1c_1\alpha_1^m\beta^n$ satisfies (10.2) by linearity. The procedure to find α, β, d_1 is analogous to the one used for the vertical boundary. To satisfy (10.2) for *all* $m > 0$ we are forced to take

$$\alpha = \alpha_1$$

and thus

$$\beta = \beta_1,$$

where β_1 is the second, smaller root of (10.10) with $\alpha = \alpha_1$ (the other root being β_0). Inserting $c_1\alpha_1^m\beta_0^n + c_1d_1\alpha_1^m\beta_1^n$ into (10.2) and dividing by $c_1\alpha_1^{m-1}$ leads to an equation for d_1 which is solved by

$$d_1 = -\frac{\beta_0(2\rho+\alpha_1)(\alpha_1+\rho)/(\rho+1) - \alpha_1^2}{\beta_1(2\rho+\alpha_1)(\alpha_1+\rho)/(\rho+1) - \alpha_1^2}\,. \tag{10.16}$$

Since β_0 and β_1 are the roots of the quadratic equation (10.10) for fixed $\alpha = \alpha_1$ we have

$$\beta_0\beta_1(2\rho+\alpha_1) = \alpha_1^2.$$

This equality reduces (10.16) to

$$d_1 = -\frac{(\alpha_1+\rho)/\beta_1 - (\rho+1)}{(\alpha_1+\rho)/\beta_0 - (\rho+1)}\,.$$

This procedure can be generalized as follows.

Lemma 10.3.3 *Let y_1 and y_2 be the roots of the quadratic equation (10.10) for fixed α. Then the linear combination $\alpha^m y_1^n + d\alpha^m y_2^n$ satisfies (10.1) and (10.2) if d is given by*

$$d = -\frac{(\alpha+\rho)/y_2 - (\rho+1)}{(\alpha+\rho)/y_1 - (\rho+1)}\,. \tag{10.17}$$

We added $d_1c_1\alpha_1^m\beta_1^n$ to compensate for $c_1\alpha_1^m\beta_0^n$ on the horizontal boundary and in doing so introduced a new error, since $d_1c_1\alpha_1^m\beta_1^n$ violates the vertical boundary conditions (10.2), so we have to add again a term, and so on. It is clear how to continue: the compensation procedure consists of adding on terms so as to compensate alternately for the error on the vertical boundary, according to lemma 10.3.2, and for the error on the horizontal boundary, according to lemma 10.3.3. This results in the infinite sum depicted in figure 10.4.

Each term in the sum in figure 10.4 satisfies (10.1), each sum of two terms with the same β-factor satisfies (10.2) and each sum of two terms with the same α-factor satisfies (10.3). Since the equilibrium equations are linear, we can conclude that the sum in figure

$$\overbrace{d_0c_0\alpha_0^m\beta_0^n}^{H} + \overbrace{d_0c_1\alpha_1^m\beta_0^n + d_1c_1\alpha_1^m\beta_1^n}^{H} + \overbrace{d_1c_2\alpha_2^m\beta_1^n + d_2c_2\alpha_2^m\beta_2^n}^{H} + \cdots$$

$$V \qquad\qquad V$$

Figure 10.4: The final sum of compensation terms. By definition $c_0 = d_0 = 1$. Sums of two terms with the same β-factor satisfy the vertical boundary conditions (V) and sums of two terms with the same α-factor satisfy the horizontal boundary conditions (H)

10.4 formally satisfies the equations (10.1)–(10.3). Let us define $x(m,n)$ as the infinite sum of compensation terms, so for $m \geq 0$, $n > 0$,

$$x(m,n) = \sum_{i=0}^{\infty} d_i(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n, \tag{10.18}$$

$$= c_0d_0\beta_0^n\alpha_0^m + \sum_{i=0}^{\infty} c_{i+1}(d_i\beta_i^n + d_{i+1}\beta_{i+1}^n)\alpha_{i+1}^m. \tag{10.19}$$

In the first sum we formed pairs with the same β-factor, in the second sum we formed pairs with the same α-factor. Below we formulate the recurrence relations for α_i, β_i, c_i and d_i.

For the initial values $\alpha_0 = \rho^2$ and $\beta_0 = \rho^2/(2+\rho)$, the sequence

$$\alpha_0 \nearrow \beta_0 \searrow \alpha_1 \nearrow \beta_1 \searrow \alpha_2 \nearrow \beta_2 \searrow \dots$$

is generated such that for all $i \geq 0$ the numbers α_i and α_{i+1} are the roots of (10.10) for fixed $\beta = \beta_i$ and the numbers β_i and β_{i+1} are the roots of (10.10) for fixed $\alpha = \alpha_{i+1}$. The generation of α_i and β_i is graphically illustrated in figure 10.5.

The coefficients c_i are generated such that for all i the term $(c_i\alpha_i^m + c_{i+1}\alpha_{i+1}^m)\beta_i^n$ satisfies (10.3). Application of lemma 10.3.2 yields that c_{i+1} can be obtained from c_i by

$$c_{i+1} = -\frac{\alpha_{i+1} - \beta_i}{\alpha_i - \beta_i}\, c_i, \qquad i \geq 0,$$

where initially

$$c_0 = 1.$$

The coefficients d_i are generated such that for all i the term $(d_i\beta_i^n + d_{i+1}\beta_{i+1}^n)\alpha_{i+1}^m$ satisfies (10.2). Application of lemma 10.3.3 yields that d_{i+1} can be obtained from d_i by

$$d_{i+1} = -\frac{(\alpha_{i+1}+\rho)/\beta_{i+1} - (\rho+1)}{(\alpha_{i+1}+\rho)/\beta_i - (\rho+1)}\, d_i, \qquad i \geq 0,$$

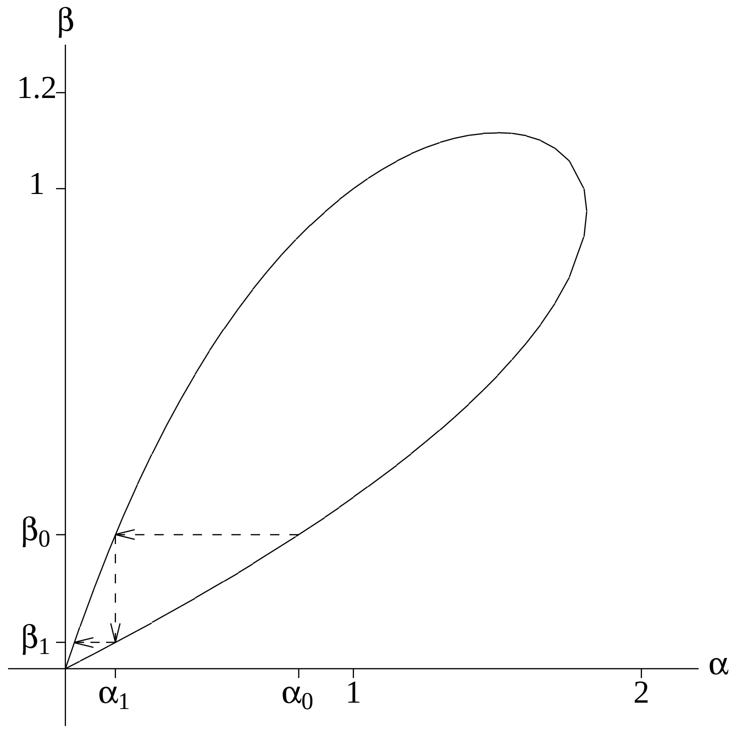

Figure 10.5: The curve: $2\rho\beta^2 + \alpha\beta^2 + \alpha^2 - 2(\rho+1)\alpha\beta = 0$ in the positive quadrant for the case $\rho = 0.9$. This curve generates α_i and β_i

where initially

$$d_0 = 1.$$

The sequence $x(m, n)$ is a *formal solution* of (10.1)–(10.1), including (10.4) due to the dependence of the equilibrium equations. The remaining problem is to prove the convergence of $x(m, n)$. It can be shown that $x(m, n)$ converges absolutely for fixed m and n (absolute convergence guarantees equality of (10.18) and (10.19), that $x(m, n) > 0$ for all m and n and that the sum of $x(m, n)$ over all m and n converges (so normalization is possible). Rigorous proofs of these statements can be found in [1, 2].

Hence we may conclude that $x(m, n)$ is a convergent solution of the equilibrium equations and normalization of $x(m, n)$ produces the desired equilibrium distribution. Based on these expressions for the equilibrium probabilities and the PASTA property of Poisson arrivals it is straightforward to derive a similar expression for the mean waiting time.

Remark 10.3.4 *(scope of the compensation approach)*
Above we developed an approach to solve the shortest queue problem. Now the question arises: what is the scope of this approach? In [3] this approach has been applied to a class of two dimensional processes on the lattice in the positive quadrant of R^2 and conditions

have been derived under which the approach works. [3] considers processes for which the transition rates are constant in the interior points and also constant on each of the axes. To simplify the analysis, it is assumed that the transitions are restricted to neighboring states. The flow diagram of such processes is depicted in figure 10.6 (compare figure 10.2).

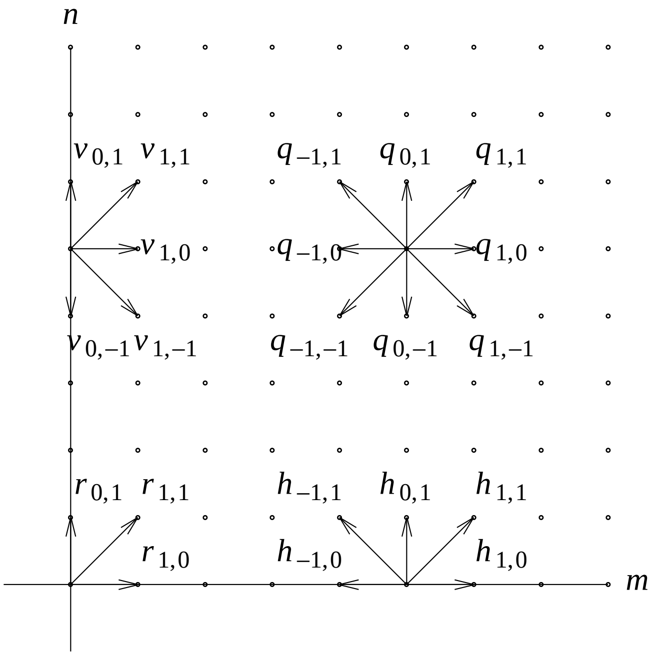

Figure 10.6: Flow diagram for a process with constant rates and transitions restricted to neighboring states. $q_{i,j}$ is the transition rate from (m, n) to $(m + i, n + j)$ with m, $n > 0$ and a similar notation is used for the transition rates on each of the axes

It appears that the compensation approach works for a subset of these models only, namely for those models satisfying

$$q_{0,1} = q_{1,1} = q_{1,0} = 0.$$

In words, the approach is only successful for models for which there are no transitions in the interior points to the North, North-East and East. Provided this condition is satisfied, it can be shown by using the same technique as for the shortest queue problem that the equilibrium probabilities $p(m, n)$ can be expressed as an infinite sum of products, i.e. there are α_i, β_i and c_i such that

$$p(m, n) = \sum_{i=0}^{\infty} c_i \alpha_i^m \beta_i^n.$$

10.4 Comparison of the assignment rules

In this section we present for each of the assignment rules considered in the previous section some numerical results for the mean production lead time.

In figure 10.7 we list the mean production lead time as a function of the occupation rate ρ for random assignment, round robin assignment and shortest queue routing. We also list the mean production lead time in case there is a central buffer.

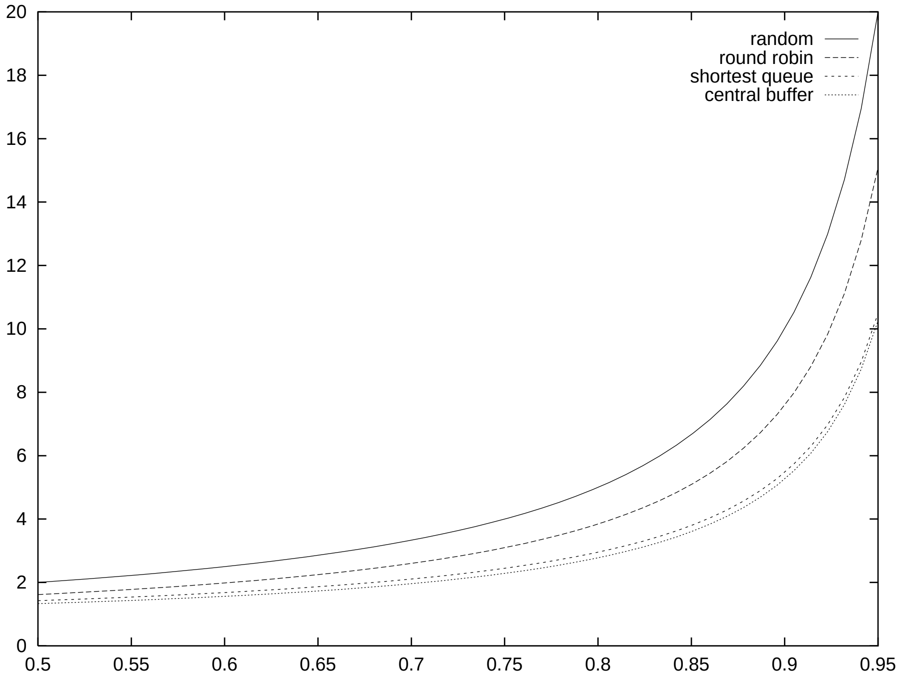

Figure 10.7: Mean production lead times as a function of the occupation rate ρ for random, round robin and shortest queue routing and the central buffer model

Several conclusions can be drawn from the data in figure 10.7. First of all shortest queue routing works nearly as good as the central buffer system, even when ρ is close to one. In fact, it can be shown that as ρ tends to one, the difference between the mean production lead time for shortest queue routing and the central buffer system tends to 1/4. Comparing the performance of round robin with the performance of random routing shows that when the occupation rate is high round robin approximately yields a reduction of 25% in the mean production lead time. Shortest queue routing leads to reduction of approximately 50%.

10.5 Exercises

Exercise 10.5.1 *(Shortest queue routing with threshold jockeying)*
Consider the model in section 10.3. Now we assume that one job switches from the longest to the shortest queue if the difference between the lengths of both queues exceeds some threshold value T.
The state of the system can again be described by the pair (m, n) where m is the length of the shortest queue and n is the difference between the queue lengths. The variable n is not unbounded, but less or equal to T.

(i) Describe the flow diagram and give the set of equations for the equilibrium state probabilities $p(m, n)$.

(ii) Solve these equations for $T = 1$ and $T = 2$.

(iii) Show for general T that for all $m + n > T$ and for $m = T$ and $n = 0$ the probabilities $p(m, n)$ are given by

$$p(m,n) = \begin{cases} K\alpha_0^m \beta_0^n, & 0 < n < T, \\ K f_0 \alpha_0^m, & n = 0, \\ K g_0 \alpha_0^m \beta_0^n, & n = T, \end{cases}$$

where K is a normalizing constant, α_0 and β_0 are given by (10.12) and

$$f_0 = \frac{\alpha_0}{\alpha_0 + \rho}, \qquad g_0 = \frac{\alpha_0}{2(\rho + 1)\beta_0}.$$

Exercise 10.5.2 *(2×2 switch)*
In parallel data processing networks, switching systems are being used to route messages and to resolve conflicts. The simplest basic unit in a switching system is the 2×2 clocked buffered switch, i.e. a switch with two input and two output ports. Let us consider the 2×2 clocked buffered switch under the simplest possible assumptions (see figure 10.8):

(i) all transfers are governed by a single discrete time clock;

(ii) at input port i, $i = 1, 2$, arrives at the beginning of each clock cycle one message with probability r_i and zero with probability $1 - r_i$;

(iii) input port i, $i = 1, 2$, routes these messages to output port j with probability $t_{i,j}$;

(iv) Messages that arrived at the beginning of a clock cycle are immediately candidates for transmission;

(v) each output port can transmit one message per clock cycle.

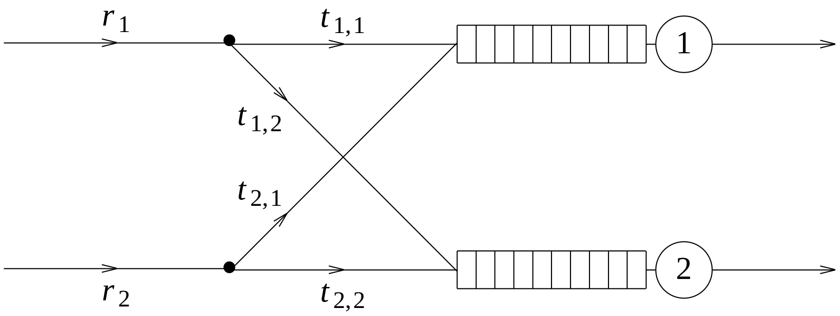

Figure 10.8: 2×2 clocked buffered switch

To guarantee that the switch can handle the messages we have to assume

$$r_1 t_{1,j} + r_2 t_{2,j} < 1, \qquad j = 1, 2.$$

The behavior of the switch can be described by a Markov chain with states (m, n) where m and n denote the number of waiting messages at output port 1 and 2, respectively, at the beginning of a clock cycle.

(i) What are the transition probabilities of this Markov chain?

(ii) Describe the flow diagram.

(iii) Verify whether the state probabilities $p(m, n)$ can determined by using the compensation approach.

For a detailed analysis of this model the reader is referred to [24, 10].

Chapter 11

Fluid flow models

A situation often encountered in the Operations Research practise is the production line. In a production line the production is organized in stages. Each stage is performed by a different machine. And the products move from one one machine to the next with as little waiting as possible. Between the machines there is little or no waiting space. Although the products are usually discrete items and not a fluid, the process is often analyzed using fluid-flow models. In this chapter we will present the basic fluid-flow model which leads to a set of differential equations that can be solved and from which we can obtain the relevant performance measures.

Before discussing some of the important Operations Research applications that can be described and analyzed using fluid-flow models we first introduce the model.

The presentation is based on the paper [54], who used a somewhat different analysis and the paper [6].

For a very extensive list of references we refer [32].

11.1 The mathematical model

Mathematically, the model we consider is the following. There are two machines, Machine 1 and Machine 2. Machine 1 produces a material and puts it into a buffer and Machine 2 takes the material out of the buffer. The material is a fluid flowing in and out the buffer, hence we talk of fluid-flow models.

The speeds at which Machine 1 produces and Machine 2 consumes however are not constant but change over time. The speed change processes of the two machine are independent. For the speed changes we use the following Markov model. The speed of Machine 1 is equal to $v_i^{(1)}$ if the Machine 1 is in state i. When Machine 1 is in state i it takes an exponential amount of time with mean $1/\lambda_i^{(1)}$ for the state to change. The next state will be state j with probability $p_{i,j}^{(1)}$. For Machine 2 we have a similar model.

So the state of the machines is a pair (i_1, i_2) and the net speed of the flow into the buffer is $v_{i_1}^{(1)} - v_{i_2}^{(2)}$.

Later, in the analysis, we will simplify the description of the model. Instead of dealing with the states and the speeds of each machine separate, we only consider the combined state, i say, instead of the pair (i_1, i_2) and the speed by which the buffer is filled or emptied, which we will denote α_i instead of $v_{i_1}^{(1)} - v_{i_2}^{(2)}$. When the combined state is equal to i it changes into state j at the rate λ_{ij}. The total number of combined states is N. So the state and net speed changes are described by a Markov generator.

11.2 Applications

In this section we first describe some applications.

Chemical processes
In a chemical process Machine 1 produces a standard substance that is used by Machine 2 in the production of a range of products. Whenever Machine 2 changes from one product to another it has to be cleaned. Switching off Machine 1 is costly, so the buffer allows Machine 1 to continue the production. In the simplest model Machine 1 has only 1 speed and Machine 2 has two speeds, one of them being 0 (cleaning). How large should the buffer be?

The model not only works very well for real flows, it also gives a very useful analysis for processes with discrete products.

Data communication
In an ATM data network standard packages called cells are sent from one switch to the next. In a switch there are several incoming lines that are 'multiplexed' on one outgoing line. If the outgoing line can process only 1 cell per time unit there will be a problem if in a certain time unit more than 1 cell arrives. If the average number of incoming cells is less than 1 buffering is helpful. If temporarily the number of incoming cells exceeds the capacity of the outgoing line, the excess demand is buffered. Once the buffer is full an incoming cell will be lost.

It is important to know the relation between arrival process, capacity, buffer size and loss probability. In ATM networks one thinks of loss probabilities in the order of 10^{-6}!

Truck cabin production
The production of a truck cabin is an assembly process that mainly involves spotwelding. Some of the spotwelding is done by robots. Unfortunately the robots are not fully reliable. Occasionally they break down and then it may take upto the time to process 5 to 10 cabins before it is repaired again.

Now consider two consecutive workstations the first one manned and somewhat slower than the second which is a robot. A buffer enables the man to continue while the robot is being repaired. If there is no buffer some of the capacity of the first station and hence of the total production line would be lost. On the other hand a truck cabin requires a lot of space so the buffer has to be small. What is a good size?

11.3 The Analysis

In the fluid flow model the state of the system is described by on the one hand the state i corresponding to speed α_i, $i = 1, 2, \ldots, N$, and on the other hand the buffer content x say, $0 \le x \le K$, with K the capacity of the buffer. In order to analyze the system we introduce the distribution functions $P_i(x,t)$, defined as the probability that at time t the system is state i and the buffer content does not exceed x.

As we will see it is possible to derive a set of differential equations for the way in which the functions P_i change over time.

Let us first relate $P_i(x, t + \Delta)$ to the functions P at time t. For convenience we define

$$\lambda_i = \sum_{j \neq i} \lambda_{ij}.$$

Then we have

$$P_i(x, t+\Delta) = P_i(x - \alpha_i \Delta, t)(1 - \lambda_i \Delta) + \sum_{j \neq i} P_j(x + O(\Delta), t)\lambda_{ji}\Delta + O(\Delta^2).$$

Dividing by Δ and letting Δ tend to 0 gives

$$\frac{\partial P_i(x,t)}{\partial t} + \alpha_i \frac{\partial P_i(x,t)}{\partial x} = -\lambda_i P_i(x,t) + \sum_{j \neq i} P_j(x,t)\lambda_{ji}.$$

We are interested in the equilibrium situation (or limiting situation) so we let t tend to ∞. Defining

$$F_i(x) = \lim_{t \to \infty} P_i(x,t),$$

we get

$$\alpha_i \frac{dF_i(x)}{dx} = -\lambda_i F_i(x) + \sum_{j \neq i} F_j(x)\lambda_{ji}.$$

Now let $F(x)$ be the N-column vector with i-th component $F_i(x)$. Further define the diagonal matrix D, with elements α_i and the Markov generator M, with diagonal elements $-\lambda_i$ and non-diagonal elements λ_{ij}. And define the matrix A as $D^{-1}M^T$, then we can rewrite this in the form

$$\frac{dF(x)}{dx} = AF(x),$$

which is just the form of the homogeneous equation (3.8).

There also are a number of boundary conditions. It is clear that if in a state α_i is positive then $F_i(0) = 0$ because the system cannot stay in the state $(i, 0)$ because in state i the buffer immediately starts to fill. On the other hand the system can stay in state (i, K) for quite a while, by the assumption that excess production overflows and is lost. So $F_i(x)$ has a jump in $x = K$. If α_i is negative the state $(i, 0)$ has positive probability but $F_i(x)$ is continuous in $x = K$. Further note that $F_i(K)$ is just the equilibrium probability π_i of the system being in state i. These probabilities can be computed from the set of equilibrium equations $\pi M = 0$.

Let H be the set of states for which α_i is positive and L the set for which α_i is negative, then we have the boundary conditions

$$F_i(0) = 0 \qquad \text{for } i \in H$$

and

$$F_i(K) = \pi_i \qquad \text{for } i \in L.$$

For the time being we assume that the set of states for which $\alpha_i = 0$ is empty. Of course we also assume that neither H nor L is empty.

So in order to solve for the functions F_i we first compute all eigenvalues σ_k and corresponding eigenvectors u_k of the matrix A. Let us assume that all eigenvalues are different, then the general solution is of the form

$$F(x) = \sum_{k=1}^{N} C_k u_k e^{\sigma_k x}.$$

The constants C_k have to be determined from the boundary conditions.

11.4 A simple example

Consider the following situation. There are two machines, the first of which is faster but breaks down every now and then where as the second one is perfect. Between the two machines a buffer will be installed. The size of the buffer has not been decided yet. The capacity of Machine 1 is 5 tons per hour, the capacity of Machine 2 only 4 ton per hour. If Machine 1 fails the speed drops to 0. So the total system has two possible speeds $\alpha_1 = 1$ and $\alpha_2 = -4$. The repair time for Machine 1 is exponentially distributed with a mean of $1/\lambda_{21} = 1$ hour. Once it is repaired it will work properly again for an exponential time with mean $1/\lambda_{12} = 9$ hours.

Following the analysis we first compute the matrix A:

$$A = \begin{pmatrix} -\lambda_{12}/\alpha_1 & \lambda_{21}/\alpha_1 \\ \lambda_{12}/\alpha_2 & -\lambda_{21}/\alpha_2 \end{pmatrix}.$$

Next we compute the eigenvalues and eigenvectors:

$$\sigma_1 = 0, \qquad u_1^T = (\lambda_{21}, \lambda_{12}),$$

$$\sigma_2 = -\lambda_{12}/\alpha_1 - \lambda_{21}/\alpha_2, \qquad u_2^T = (1/\alpha_1, -1/\alpha_2).$$

And we get the boundary conditions

$$F_1(0) = 0, \qquad F_2(K) = \pi_2 = \lambda_{12}/(\lambda_{12} + \lambda_{21}).$$

From this we can obtain the general solution

$$F(x) = \pi_2 \frac{u_1/\alpha_1 - \lambda_{21} e^{\sigma_2 x} u_2}{\lambda_{12}/\alpha_1 + \lambda_{21} e^{\sigma_2 K}/\alpha_2}.$$

The most important performance measure is the average production rate. Note that Machine 2 always produces at the rate 4, unless the buffer is empty. In which case the production rate of Machine 2 drops to 0. So the average production rate is equal to $(1 - F_2(0))4$. Table 11.1 shows how the average production rate varies with K.

K	$F_2(0)$	$(1 - F_2(0))4$
0	0.10000	3.60000
1	0.07885	3.68459
2	0.06344	3.74625
3	0.05180	3.79279
4	0.04279	3.82886
5	0.03565	3.85738
6	0.02992	3.88030
7	0.02526	3.89896
8	0.02142	3.91431
9	0.01824	3.92704
10	0.01558	3.93768

Table 11.1: The average production as a function of K

11.5 Exercises

Exercise 11.5.1
Consider the example in section 11.4 and now suppose that $\lambda_{12} = 1/3$ and $\lambda_{21} = 1$. Find the average production rate depending on K.

Exercise 11.5.2
Consider the example in section 11.4 and now suppose that the repair times are Erlang-2 distributed with a mean of 1 hour. All other data are the same as in the example in section 11.4. Find the average production rate depending on K.

Bibliography

[1] I.J.B.F. Adan, *A compensation approach for queueing problems*, CWI Tract 104, Stichting Mathematisch Centrum, Amsterdam, 1994.

[2] I.J.B.F. Adan, J. Wessels, W.H.M. Zijm, *Analysis of the symmetric shortest queue problem*, Stochastic Models, 6 (1990), pp. 691–713.

[3] I.J.B.F. Adan, J. Wessels, W.H.M. Zijm, *A compensation approach for two-dimensional Markov processes*, Adv. Appl. Prob., 25 (1993), pp. 783–817.

[4] I.Adan and Y.Zhao, *Analyzing $GI/E_r/1$ queues*, Opns. Res. Lett., 1996 (to appear).

[5] I.J.B.F. Adan, W.A. van de Waarsenburg, J. Wessels, *Analyzing $E_k|E_r|c$ queues*, EJOR, 92 (1996), pp. 112–124.

[6] D. Anick, D. Mitra, M.M. Sondhi, *Stochastic theory of a data handling system*, Bell System Tech. J., 61 (1982), pp. 1871–1894.

[7] D. Bertsimas, *An analytic approach to a general class of $G/G/s$ queueing systems*, Opns. Res., 38 (1990), pp. 139–155.

[8] D. Bertsimas, *An exact FCFS waiting time analysis for a class of $G/G/s$ queueing systems*, QUESTA, 3 (1988), pp. 305–320.

[9] D. Bertsimas, X.A. Papaconstantinou *Analysis of the stationary $E_k/C_2/s$ queueing system*, EJOR, 37 (1988), pp. 272–287.

[10] O.J. Boxma, G.J. van Houtum, *the compensation approach applied to a 2×2 switch*, Prob. Engineer. Inform. Sci., 7 (1993), pp. 471–493.

[11] J.A. Buzacott, J.G. Shanthikumar, *Stochastic models of manufacturing systems*, Prentice Hall, 1993.

[12] A.S. Carr, A.R. Güllü, P.L. Jackson, J.A. Muckstadt, *An exact analysis of a production-inventory strategy for industrial suppliers*, Working paper, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1993.

[13] E.A. Coddington, N. Levinson, *Theory of ordinary differential equations*, McGraw-Hill, London, 1955.

[14] J.W. Cohen, O.J. Boxma, *Boundary value problems in queueing system analysis*, North-Holland, Amsterdam, 1983.

[15] M.J.A. van Eenige, *Queueing systems with periodic service*, Thesis, Eindhoven University of Technology, Eindhoven, 1996.

[16] G. Fayolle, *Méthodes analytiques pour les files d'attente coupleés*, Thesis, Univ. de Paris VI, Paris ,1979.

[17] G. Fayolle, R. Iasnogorodski, *Two coupled processors: the reduction to a Riemann–Hilbert problem*, Z. Wahrscheinlichkeitsth. Verw. Gebiete, 47 (1979), pp. 325–351.

[18] A. Federgruen, Z. Katalan, *Make-to-stock or make-to-order: that is the question; novel answers to an ancient debate*, Working paper, Graduate School of Business, Columbia University, New York, 1994.

[19] L. Flatto, H.P. McKean, *Two queues in parallel*, Comm. Pure Appl. Math., 30 (1977), pp. 255–263.

[20] P.R. Halmos, *Finite-dimensional vector spaces*, Van Nostrand, London, 1958.

[21] M.C. van der Heijden, *Performance analysis for reliability and inventory models*, Thesis, Vrije Universiteit, Amsterdam, 1993.

[22] A. Hordijk, G. Koole, *On the optimality of the generalized shortest queue policy*, Prob. Engineer. Inform. Sci., 4 (1990), pp. 477–487.

[23] R. Iasnogorodski, *Problèmes-frontières dans les files d'attente*, Thesis, Univ. de Paris IV, Paris, 1979.

[24] S. Jaffe, *The equilibrium distribution for a clocked buffered switch*, Prob. Engineer. Inform. Sci., 6 (1992), pp. 425–438.

[25] M.A. Johnson, *An emperical study of queueing approximations based on phase-type approximations*, Stochastic Models, 9 (1993), pp. 531–561.

[26] J.F.C. Kingman, *Two similar queues in parallel*, Ann. Math. Statist., 32 (1961), pp. 1314–1323.

[27] L. Kleinrock, *Queueing Systems, Vol. I: Theory.* Wiley, New York, 1975.

[28] S. Lang, *Linear algebra*, Addison-Wesley, new York, 1978.

[29] W. Lin, P.R. Kumar, *Optimal control of a queueing system with two heterogeneous servers*, Mathematics Research Report No. 82–83, Dept. of Math., University of maryland, Baltimore County, 1982.

[30] J.D. LITTLE, *A proof of the queueing formula $L = \lambda W$*, Opns. Res., 9 (1961), pp. 383–387.

[31] R.A. MARIE, *Calculating equilibrium probabilities for $\lambda(n)/C_k/1/N$ queue*, in: Proceedings Performance' 80, Toronto, (May 28–30, 1980), pp. 117–125.

[32] D. MITRA, *Stochastic theory of a fluid model of producers and consumers coupled by a buffer*, Adv. Appl. Prob., 20 (1988), pp. 646–676.

[33] M.F. NEUTS, *Matrix-geometric solutions in stochastic models*, Johns Hopkins University Press, Baltimore, 1981.

[34] S.M. ROSS, *Introduction to probability models*, Academic Press, New York, 1985.

[35] M. RUBINOVITCH, *The slow server problem*, J. Appl. Prob., 22 (1985), pp. 205–213.

[36] M. RUBINOVITCH, *The slow server problem: a queue with stalling*, J. Appl. Prob., 22 (1985), pp. 879–892.

[37] R.S. SCHASSBERGER, *On the waiting time in the queueing system $GI/G/1$*, Ann. Math. Statist., 41 (1970), pp. 182–187.

[38] R.S. SCHASSBERGER, *Warteschlangen,* Springer-Verlag, Berlin, 1973.

[39] S. SHAPIRO, *The M-server queue with Poisson input and Gamma-distributed service of order two*, Opns. Res., 14 (1966), pp. 685–694.

[40] C.R. SOX, L.J. THOMAS, J.O. MCCLAIN, *Jellybeans: Sorting capacity to improve service*, Working paper, Department of Industrial Engineering, Auburn University, Alabama, 1994.

[41] S. STIDHAM, *A last word on $L = \lambda W$*, Opns. Res., 22 (1974), pp. 417–421.

[42] S. STIDHAM, *Socially and individually optimal control of arrivals to a $GI/M/1$ queue*, Management Sci., 24 (1978), pp. 1598–1610.

[43] Y.TAKAHASHI, *Asymptotic exponentiality of the tail of the waiting time distribution in a $Ph/Ph/1$ queue*, Adv. Appl. Prob., 13 (1981), pp. 619–630.

[44] H.C. TIJMS, *Stochastic modelling and analysis: a computational approach*, John Wiley & Sons, Chichester, 1986.

[45] H.C. TIJMS, M.C.T. VAN DE COEVERING, *A simple numerical approach for infinite-state Markov chains*, Prob. Engineer. Inform. Sci., 5 (1991), pp. 285–295.

[46] J. WALRAND, *A note on 'Optimal control of a queueing system with two heterogeneous servers'*, Technical Report, Dept. of Electrical Enigeering and Computer Sciences and Electronic Research Laboratory, University of California, Berkeley, 1983.

[47] R.W. Weber, *On the optimal assignment of customers to parallel servers*, J. Appl. Prob., 15 (1978), pp. 406–413.

[48] W. Whitt *Approximating a point process by a renewal process I: two basic methods*, Opns. Res., 30 (1986), pp. 125–147.

[49] W. Whitt *Deciding which queue to join: some counterexamples*, Opns. Res., 34 (1986), pp. 55–62.

[50] T.M. Williams, *Special products and uncertainty in production/inventory systems*, EJOR, 15 (1984), pp. 46–54.

[51] W. Winston, *Optimality of the shortest line discipline*, J. Appl. Prob., 14 (1977), pp. 181–189.

[52] R.W. Wolff, *Poisson arrivals see time averages*, Opns. Res., 30 (1982), pp. 223–231.

[53] R.W. Wolff, *Stochastic modeling and the theory of queues*, Prentice-Hall, London, 1989.

[54] J. Wijngaard, *The effect of interstage buffer storage on the output of two unreliable production units in series, with different production rates*, AIIE Trans., 11 (1979), pp. 42–47.

Index