

Chapter 1

Basic Tools of Probabilistic Modelling

1.1. General background

On a certain level of abstraction, computer systems belong to the same family as, for example, job-shops, supermarkets, hairdressing salons and airport terminals; all these are sometimes described as “mass service systems” and more often as “queueing systems”. Customers (or tasks, or jobs, or machine parts) arrive according to some random pattern; they require a variety of services (execution of arithmetic and logical operations, transfer of information, seat reservations) of random durations. Services are provided by one or more servers, perhaps at different speeds. The order of service is determined by a set of rules which constitutes the “scheduling strategy”, or “service discipline”.

The mathematical analysis of such systems is the subject of queueing theory. Since A. K. Erlang’s studies of telephone switching systems, in 1917–1918, that theory has progressed considerably; today it boasts an impressive collection of results, methods and techniques. Interest in queueing theory has always been stimulated by problems with practical applications. In particular, most of the theoretical advances of the last decade are directly attributable to developments in the area of computer systems performance evaluation.

Because customer interarrival times and the demands placed on the various servers are random, the state $S(t)$ of a queueing system at time t of its operation is a random variable. The set of these random variables $\{S(t), t \geq 0\}$ is a stochastic process. A particular realisation of the random variables — that is, a particular realisation of all arrival events, service demands, etc. — is a “sample path” of the stochastic process. For example, in a single-server queueing system where all customers are of the same type, one might be interested in the stochastic process $\{N(t), t \geq 0\}$, where $N(t)$ is the number of customers waiting and/or being served at time t . A portion

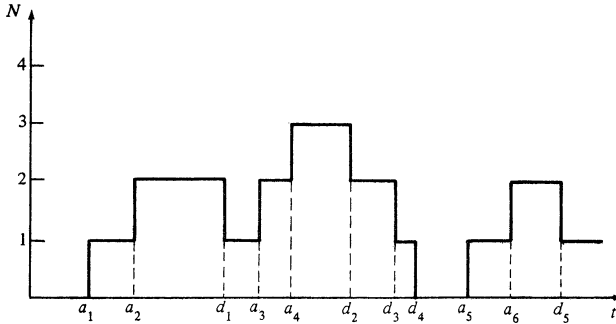


Fig. 1.1.

of a possible sample path for this process is shown in Fig. 1.1: customers arrive at moments a_1, a_2, \dots and depart at moments d_1, d_2, \dots .

An examination of the sample paths of a queueing process can disclose some general relations between different quantities associated with a given path. For instance, in the single-server system, if $N(t_1) = N(t_2)$ for some $t_1 < t_2$, and there are k arrivals in the interval (t_1, t_2) , then there are k departures in that interval. Since a sample path represents a system in operation, relations of the above type are sometimes called “operational laws” or “operational identities” (Buzen [1]). We shall derive some operational identities in section 1.7. Because they apply to individual sample paths, these identities are independent of any probabilistic assumptions governing the underlying stochastic process. Thus, the operational approach to performance evaluation is free from the necessity to make such assumptions. It is, however, tied to specific sample paths and hence to specific runs of an existing system where measurements can be taken.

The probabilistic approach involves studying the stochastic process which represents the system. The results of such a study necessarily depend on the probabilistic assumptions governing the process. These results are themselves probabilistic in nature and concern the population of all possible sample paths. They are not associated with a particular run of an existing system, or with any existing system at all. It is often desirable to evaluate not only the expected performance of a system, but also the likely deviations from that expected performance. Dealing with probability distributions makes this possible, at least in principle.

We shall be concerned mainly with steady-state system behaviour — that is, with the characteristics of a process which has been running for a long time and has settled down into a “statistical equilibrium regime”. Long-run performance measures are important because they are stable;

state i and of anything that happened before that time. This very important property will be referred to as the “memoryless property”.

The probability $p_{i,j}(y)$, regarded as a function of y , is called the “transition probability function”. The memoryless property immediately implies the following set of functional equations:

$$p_{i,j}(x+y) = \sum_{k=0}^{\infty} p_{i,j}(x)p_{k,j}(y), \quad x, y \geq 0, \quad i, j = 0, 1, \dots \quad (1.3)$$

These equations express simply the fact that, in order to move from state i to state j in time $x+y$, the process has to be in some state k after time x and then move to state j in time y (and the second transition does not depend on i and x). They are the Chapman–Kolmogorov equations of the Markov process. Introducing the infinite matrix $\mathbf{P}(y)$ of transition functions $p_{i,j}(y)$, we can rewrite (1.3) as

$$\mathbf{P}(x+y) = \mathbf{P}(x)\mathbf{P}(y), \quad x, y \geq 0. \quad (1.4)$$

We shall assume that the functions $p_{i,j}(y)$ are continuous at $y=0$:

$$\lim_{y \rightarrow 0} p_{i,j}(y) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \quad (1.5)$$

That assumption, together with (1.3), ensures that $p_{i,j}(y)$ is continuous, and has a continuous derivative, for all $y \geq 0$; $i, j = 0, 1, \dots$ (we state this without proof).

A special role is played by the derivatives $a_{i,j}$ of the transition functions at $t=0$. By definition,

$$a_{i,i} = \lim_{y \rightarrow 0} \frac{p_{i,i}(y) - 1}{y}, \quad i = 0, 1, \dots \quad (1.6)$$

$$a_{i,j} = \lim_{y \rightarrow 0} \frac{p_{i,j}(y)}{y}, \quad i \neq j = 0, 1, \dots$$

Hence, if h is small,

$$p_{i,j}(h) = a_{i,j}h + o(h), \quad i \neq j = 0, 1, \dots, \quad (1.7)$$

where $o(x)$ is a function such that $\lim_{x \rightarrow 0} [o(x)/x] = 0$.

In other words, if the Markov process is in state i at some moment t , then the probability that at time $t+h$ it is in state j is nearly proportional to h , with coefficient of proportionality $a_{i,j}$. That is why $a_{i,j}$ is called

the “instantaneous transition rate from state i to state j ”, $i \neq j$. The probability that the process leaves state i by $t+h$ is approximately equal to

$$1 - p_{i,i}(h) = -a_{i,i}h + o(h), \quad i = 0, 1, \dots, \quad (1.8)$$

so $-a_{i,i}$ is the instantaneous rate of transition out of state i . Of course, we must have

$$-a_{i,i} = \sum_{\substack{j=0 \\ j \neq i}}^{\infty} a_{i,j}. \quad (1.9)$$

In fact, since $\mathbf{P}(y)$ is a stochastic matrix (its rows sum up to 1), the rows of $\mathbf{P}'(y)$ must sum up to 0 for all $y \geq 0$.

Let $\mathbf{A} = [a_{i,j}]$, $i, j = 0, 1, \dots$ be the matrix of instantaneous transition rates. Differentiating (1.4) with respect to x and then letting $x \rightarrow 0$ yields a system of equations known as the Chapman–Kolmogorov backward differential equations:

$$\mathbf{P}'(y) = \mathbf{A}\mathbf{P}(y). \quad (1.10)$$

Similarly, differentiating (1.4) with respect to y and letting $y \rightarrow 0$ yields the Chapman–Kolmogorov forward differential equations

$$\mathbf{P}'(x) = \mathbf{P}(x)\mathbf{A}. \quad (1.11)$$

Either (1.10) or (1.11) can be solved for the transition probability functions, subject to the initial conditions $\mathbf{P}(0) = \mathbf{I}$ (the identity matrix) and $\mathbf{P}'(0) = \mathbf{A}$. In a purely formal way, treating $\mathbf{P}(y)$ as a numerically valued function and \mathbf{A} as a constant, (1.10) and (1.11) are satisfied by

$$\mathbf{P}(y) = e^{\mathbf{A}y}. \quad (1.12)$$

This turns out, indeed, to be the solution, provided that (1.12) is interpreted as

$$\mathbf{P}(y) = \sum_{n=0}^{\infty} \frac{y^n}{n!} \mathbf{A}^n, \quad y \geq 0. \quad (1.13)$$

Thus, the transition probability functions are completely determined by their derivatives at $y = 0$. It should be clear, however, that to find them in practice is by no means a trivial operation. The matrix $\mathbf{P}(y)$, for finite values of y , is referred to as the “transient solution” of the Markov process. As far as closed-form expressions are concerned, transient solutions are unobtainable for all but a few very simple Markov processes.

Let $\{S(t), t \geq 0\}$ be a Markov process with instantaneous transition rate matrix \mathbf{A} . Suppose that at time t the process is in state i . What is the distribution of the interval η_i until the first exit from state i (that interval is called the “holding time”)? And what is the probability $q_{i,j}$ that the next state to be entered will be state j ? According to the memoryless property, the answers to both these questions are independent of t and of the process history prior to t . In particular, they are independent of how long the process has already spent in state i . Consider first the holding time; denote by $\hat{H}_i(x)$ the complementary distribution function of η_i : $\hat{H}_i(x) = P(\eta_i > x)$. From the memoryless property, if the process stays in state i for time x , the probability that it will remain there for at least another interval y is independent of x . Therefore,

$$\hat{H}_i(x + y) = \hat{H}_i(x)\hat{H}_i(y), \quad x, y \geq 0. \quad (1.14)$$

Any distribution function which satisfies (1.14) must fall into one of the following three categories:

- (i) $\hat{H}_i(x) = 1$ for all $x \geq 0$. If this is the case, once the process enters state i it remains there forever (properly speaking, the holding time does not have a distribution function then). States of this type are called “absorbing”.
- (ii) $\hat{H}_i(x) = 0$ for all $x \geq 0$. In this case the process bounces out of state i as soon as it enters it. Such states are called “instantaneous”.
- (iii) $\hat{H}_i(x)$ is monotone decreasing from 1 to 0 on the interval $[0, \infty)$ and is differentiable. States in this category are called “stable”.

From now on, we shall assume that all states are stable. Differentiating Eq. (1.4) with respect to y and letting $y \rightarrow 0$ we obtain $\hat{H}'_i(x) = -\lambda_i \hat{H}_i(x)$, where $\lambda_i = -\hat{H}'_i(0)$. Hence

$$\hat{H}_i(x) = e^{-\lambda_i x}, \quad x \geq 0,$$

and the distribution function $H_i(x) = P(\eta_i \leq x)$ is given by

$$H_i(x) = 1 - e^{-\lambda_i x}, \quad x \geq 0. \quad (1.15)$$

To determine the parameter λ_i in terms of the matrix \mathbf{A} , note that according to (1.15) the probability of leaving state i in a small interval h is equal to $H_i(h) = \lambda_i h + o(h)$. Comparing this with (1.8) shows that λ_i is exactly the instantaneous transition rate out of state i :

$$\lambda_i = -a_{i,i}, \quad i = 0, 1, \dots \quad (1.16)$$

From (1.15), (1.7) and the memoryless property it follows that the probability that the process remains in state i for time x and then moves to state j in the infinitesimal interval $(x, x + dx)$ is equal to

$$e^{-\lambda_i x} a_{i,j} dx, \quad x \geq 0, \quad j \neq i.$$

Integrating this expression over all $x \geq 0$ gives us the probability that the next state to be entered will be state j :

$$q_{i,j} = \int_0^\infty e^{-\lambda_i x} a_{i,j} dx = \frac{a_{i,j}}{\lambda_i} = -\frac{a_{i,j}}{a_{i,i}}, \quad i \neq j = 0, 1, \dots \quad (1.17)$$

We derived (1.15) and (1.17) under the assumption that the Markov process was observed at some arbitrary, but fixed, moment t . These results continue to hold if, for example, the process is observed just after it enters state i . Moreover, a stronger assertion can be made (we state it without proof): given that the process has just entered state i , the time it spends there and the state it enters next are mutually independent.

The behaviour of a Markov process can thus be described as follows: at time $t = 0$ the process starts in some state, say i ; it remains there for an interval of time distributed exponentially with parameter λ_i (average length $1/\lambda_i$); the process then enters state j with probability $q_{i,j}$, remains there for an exponentially distributed interval with mean $1/\lambda_j$, enters state k with probability $q_{j,k}$, etc. The successive states visited by the process form a “Markov chain” — that is, the next state depends on the one immediately before it, but not on all the previous ones and not on the number of moves made so far. This Markov chain is said to be “embedded” in the Markov process.

We shall conclude this section by examining a little more closely the exponential distribution defined in (1.15). That distribution plays a central role in most probabilistic models that are analytically tractable. It owes its preminent position to the memoryless property. If the duration η of a certain activity is distributed exponentially with parameter λ , and if that activity is observed at time x after its beginning, then the remaining duration of the activity is independent of x and is also distributed exponentially with parameter λ :

$$P(\eta > x + y \mid \eta > x) = \frac{P(\eta > x + y)}{P(\eta > x)} = \frac{e^{-\lambda(x+y)}}{e^{-\lambda x}} = e^{-\lambda y} = P(\eta > y). \quad (1.18)$$

On the other hand, we have seen in the derivation of (1.15) that (excluding the degenerate cases) the memoryless property implies the exponential distribution. There are, therefore, no other distributions with that property.

Let η_1 and η_2 be two independent random variables with distribution functions

$$F_1(x) = 1 - e^{-\lambda_1 x}; \quad F_2(x) = 1 - e^{-\lambda_2 x},$$

and density functions

$$f_1(x) = \lambda_1 e^{-\lambda_1 x}; \quad f_2(x) = \lambda_2 e^{-\lambda_2 x},$$

respectively. Think of η_1 and η_2 as the durations of two activities which are in progress simultaneously. The two activities are observed at a given moment; neither of them has completed. It is then of interest to know the distribution of the interval, η , until the first completion of an activity and the probability, q_i , that the i -th activity will complete first ($i = 1, 2$). Denote the distribution function and the density function of η by $F(x)$ and $f(x)$, respectively. Using the conventional notation $P(\eta = x)/dx$ in place of $\lim_{\Delta x \rightarrow 0} [P(x \leq \eta < x + \Delta x)/\Delta x]$, and the memoryless property, we can write

$$\begin{aligned} f(x)dx &= P(\eta = x) = P(\min(\eta_1, \eta_2) = x) \\ &= P(\eta_1 = x)P(\eta_2 \geq x) + P(\eta_1 \geq x)P(\eta_2 = x) \\ &= f_1(x)dx[1 - F_2(x)] + f_2(x)dx[1 - F_1(x)] \\ &= \lambda_1 e^{-\lambda_1 x} e^{-\lambda_2 x} dx + \lambda_2 e^{-\lambda_2 x} e^{-\lambda_1 x} dx \\ &= (\lambda_1 + \lambda_2)e^{-(\lambda_1 + \lambda_2)x} dx. \end{aligned} \tag{1.19}$$

The time until the first completion is thus distributed exponentially with parameter $(\lambda_1 + \lambda_2)$. The probability that activity 1 will complete first is given by

$$q_1 = P(\eta_1 < \eta_2) = \int_0^\infty f_1(x)[1 - F_2(x)]dx = \lambda_1/(\lambda_1 + \lambda_2). \tag{1.20}$$

Similarly, $q_2 = P(\eta_2 > \eta_1) = \lambda_2/(\lambda_1 + \lambda_2)$. Moreover, it is easily seen that the time until the nearest completion does not depend on which activity completes first. For instance,

$$P(\eta = x \mid \eta_1 < \eta_2) = (\lambda_1 + \lambda_2)e^{-(\lambda_1 + \lambda_2)x} dx = P(\eta = x). \tag{1.21}$$

These results, which can be generalised in an obvious way to any (even infinite) number of activities, give an intuitive meaning to expressions (1.15) and (1.17) concerning the holding times and transition probabilities of a Markov process.

When the process enters state i , we can imagine exponentially distributed activities, representing the transitions from state i to state j ($j = 0, 1, \dots$), being started all at once. The parameter of the j -th distribution is $a_{i,j}$. The holding time in state i is then the time until the first completion of an activity; the next state entered is the index of that first activity.

1.3. Poisson arrival streams. Important properties

The telephone calls received at a switchboard, the impacts by molecules to which a small particle immersed in liquid is subjected, the breakdown of machines in a large factory — all these, and many other physical phenomena, give rise to Poisson processes. In general, a Poisson process is used to model a sequence of events — we shall refer to them as “arrivals” — whose moments of occurrence satisfy certain probabilistic conditions. In textbooks on stochastic processes, the definition and treatment of the Poisson process usually precede those of general Markov processes. Here, however, we wish to be as economical as possible; having developed some Markov process theory, we shall apply it to this very special case.

The Poisson process, $\{N(t), t \geq 0\}$, is a Markov process which satisfies the following restrictions:

- (i) $N(0) = 0$ with probability 1,
- (ii) from state i ($i = 0, 1, \dots$) the process moves to state $i + 1$ with probability 1; the instantaneous transition rate $a_{i,i+1}$ does not depend on i ($a_{i,i+1} = \lambda, i = 0, 1, \dots$).

We have thus defined a counting process: the value of $N(t)$ is equal to the number of moves, or the number of arrivals, in the interval $(0, t]$. The distribution of that number, $p_k(t) = P(N(t) = k \mid N(0) = 0)$, $k = 0, 1, \dots$, constitutes the first row of the transition probability matrix $\mathbf{P}(t)$ defined in the last section. We are now in the happy position of being able to use the general result (1.12) to find the desired distribution; the Poisson process is just simple enough to permit such an approach.

Restriction (ii), together with (1.9) and (1.17), imply that the instantaneous transition matrix of the Poisson process has the form

$$\mathbf{A} = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & \cdots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix} = \lambda(-\mathbf{I} + \mathbf{U}). \tag{1.22}$$

Here, \mathbf{I} is the (infinite) identity matrix and \mathbf{U} is the matrix which has ones on the first upper diagonal and zeros everywhere else:

$$\mathbf{U} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}.$$

Substituting (1.22) into (1.12) yields

$$\mathbf{P}(t) = e^{\lambda(\mathbf{U}-\mathbf{I})t} = e^{-\lambda t} e^{\lambda \mathbf{U}t} = e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} \mathbf{U}^n. \tag{1.23}$$

Now, the matrix \mathbf{U}^n has ones on the n -th upper diagonal and zeros everywhere else. Therefore, the first row of the matrix defined by the series on the right-hand side of (1.23) is $(1, \lambda t, (\lambda t)^2/2!, \dots)$. The probability of k arrivals in the interval $(0, t]$ is equal to

$$p_k(t) = \frac{e^{-\lambda t} (\lambda t)^k}{k!}, \quad k = 0, 1, \dots \tag{1.24}$$

Because of the memoryless property, the probability of k arrivals in any interval of length t is also given by (1.24). In a small interval of length h , there is one arrival with probability $p_1(h) = \lambda h + o(h)$. The probability that there are two or more arrivals in an interval of length h is $P_{>1}(h) = o(h)$. These last properties (plus the memoryless one) are sometimes given as defining axioms for the Poisson process.

Since the Poisson process is a Markov process, the holding times, i.e. the intervals between consecutive arrivals, are independent and distributed exponentially with parameter λ . This property too, can be taken as a definition of the Poisson process; it implies the Markov property and everything else. The expected length of the interarrival intervals is $1/\lambda$. Therefore, the average number of arrivals per unit time is λ . For that reason, the parameter λ is called the “rate” of the Poisson process. The average

number of arrivals in an interval of length t is

$$E[N(t)] = \lambda t, \quad (1.25)$$

as can also be seen directly from (1.24).

Often in practice, arrival streams from two or more different sources merge before reaching a single destination. We shall see this happening, for example, in queueing networks (Chapter 3). Now, if the component processes are Poisson, then the result of this merging, or superposition operation is also Poisson. Indeed, let $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ be two independent Poisson processes with rates λ_1 and λ_2 , respectively, and let $\{N(t) = N_1(t) + N_2(t), t \geq 0\}$ be their superposition. Consider the interval η between an arbitrary moment t_0 and the next arrival instant of $\{N(t)\}$. Clearly, $\eta = \min(\eta_1, \eta_2)$, where η_i is the interval between t_0 and the next arrival instant of $\{N_i(t)\}$, $i = 1, 2$. Since the component processes are Poisson, η_1 and η_2 are exponentially distributed with parameters λ_1 and λ_2 , respectively; also they are mutually independent. By (1.19), η is exponentially distributed with parameter $\lambda = \lambda_1 + \lambda_2$. This, in turn, implies that $\{N(t), t \geq 0\}$ is Poisson with rate λ .

The above argument generalises easily. The superposition of an arbitrary number of independent Poisson processes is Poisson, with rate equal to the sum of the component rates. Moreover, the superposition is approximately Poisson even if the individual components are not, as long as they are independent and there is a large number of them. This explains why Poisson arrival processes are frequently observed in practice. For example, if each user of a computing facility submits jobs independently of the others, and there are many users, the total stream of jobs will be approximately Poisson.

Consider now the splitting, or “decomposition”, of a Poisson process $\{N(t), t \geq 0\}$ into two components $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$. The decomposition is performed by a sequence of independent Bernoulli trials: every arrival of the process $\{N\}$ is assigned to the process $\{N_i\}$ with probability α_i ($i = 1, 2; \alpha_1 + \alpha_2 = 1$). The joint distribution of $N_1(t)$ and $N_2(t)$ can be obtained as follows:

$$\begin{aligned} P(N_1(t) = n_1, N_2(t) = n_2) &= P(N_1(t) = n_1, N_2(t) = n_2) \\ &= P(N_2(t) = n_2 \mid N(t) = n_1 + n_2) P(N(t) = n_1 + n_2) \\ &= \frac{(n_1 + n_2)!}{n_1! n_2!} \alpha_1^{n_1} \alpha_2^{n_2} \frac{e^{-\lambda t} (\lambda t)^{n_1 + n_2}}{(n_1 + n_2)!} \\ &= \frac{e^{-\alpha_1 \lambda t} (\alpha_1 \lambda t)^{n_1}}{n_1!} \frac{e^{-\alpha_2 \lambda t} (\alpha_2 \lambda t)^{n_2}}{n_2!}, \end{aligned} \quad (1.26)$$

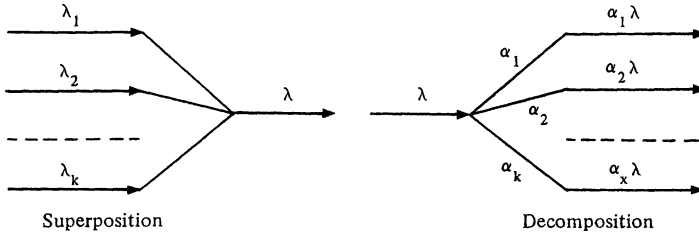


Fig. 1.2.

where we have used (1.24). We see that the processes resulting from the decomposition are both Poisson (with rates $\alpha_1 \lambda$ and $\alpha_2 \lambda$, respectively). Not only that, these processes are independent of each other. This result, too, generalises to arbitrary number of components.

The superposition and decomposition of Poisson processes are illustrated in Fig. 1.2.

In analysing system performance, we frequently employ the technique of “tagging” an incoming customer and following his progress through the system. It is therefore important to know something about the system state distribution that customers see when they arrive. In this respect, Poisson arrivals have a very useful, and apparently unique property: they behave like random observers. More precisely, let $\{S(t), t \geq 0\}$ be a stochastic process representing the state of a queueing system. That system is fed with customers by one or more arrival streams. Consider an arbitrary moment t_0 ; let $S(t_0^-)$ be the system state just prior to t_0 . Then, if the arrival streams are Poisson, the random variable $S(t_0^-)$ is independent of whether there is an arrival at t_0 or not (Strauch [8]). This is because $S(t_0^-)$ is influenced only by the past history of the arrival processes, and that is independent of whether there is an arrival at t_0 (looking backwards in time, the interarrival intervals are still exponentially distributed and hence memoryless).

Thus, an arrival from a Poisson stream sees the same system state distribution as someone who just happens to look at the system, having otherwise nothing to do with it (a random observer).

To appreciate this remarkable property better, let us take a contrasting example where the arrival stream is decidedly not Poisson. Imagine a conveyor belt bringing machine parts to an operator at intervals ranging between 20 and 30 minutes; the operation performed on each part lasts between 10 and 18 minutes. Two hours after starting the belt, a random observer (the shop floor supervisor?) may well see the operator diligently

at work. But if a machine part arrives at that time, it is guaranteed to find him idle!

Before leaving the topic of Poisson processes, let us derive the distribution of the time T_n until the n -th arrival instant. That random variable — the sum of n independent exponentially distributed intervals with the same mean — plays an important role in modelling. Denote its distribution function by $G_n(x)$. From the definition of T_n , and from (1.24), we have

$$\begin{aligned} G_n(x) &= P(T_n \leq x) = P(N(x) \geq n) = \sum_{k=n}^{\infty} e^{-\lambda x} (\lambda x)^k / k! \\ &= 1 - \sum_{k=0}^{n-1} e^{-\lambda x} (\lambda x)^k / k!. \end{aligned}$$

That function is called “the n -stage Erlang distribution function”. Its derivative

$$g_n(x) = G'_n(x) = \lambda e^{-\lambda x} (\lambda x)^{n-1} / (n-1)!,$$

is “the n -stage Erlang density function”. The mean and variance of T_n are, respectively, n/λ and n/λ^2 .

1.4. Steady-state. Balance diagrams. The “Birth and Death” process

So far, we have been concerned with time-dependent properties of stochastic processes. The chief objects of interest in a Markov process were the transition probability functions $p_{i,j}(y)$ relating the state of the process at a given moment to its state at time y later. Now, although the process state at time t depends on the initial state (at time 0), we feel intuitively that in a “well-behaved” system that dependence should weaken as t increases. In the long run, the probability of finding the process in a given state should be independent of where the process started and should cease to vary with time.

Let us give these intuitive ideas a more precise meaning. Consider a Markov process $\{S(t), t \geq 0\}$ with state space $\{0, 1, \dots\}$ and instantaneous transition rate matrix $\mathbf{A} = [a_{i,j}]$, $i, j = 0, 1, \dots$. The time-dependent behaviour of the process is described by the matrix of transition probability functions $\mathbf{P}(t) = [p_{i,j}(t)]$, $i, j = 0, 1, \dots$. We say that steady-state (or equilibrium, or long-run) regime exists for that process if (i),

the limits

$$\pi_j = \lim_{t \rightarrow \infty} p_{i,j}(t) = \lim_{t \rightarrow \infty} P(S(t) = j \mid S(0) = i), \quad j = 0, 1, \dots, \quad (1.27)$$

exist and are independent of the initial state, and (ii), these limits constitute a probability distribution:

$$\sum_{j=0}^{\infty} \pi_j = 1. \quad (1.28)$$

To justify the term “steady-state”, suppose that the distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \dots)$ exists and let $x \rightarrow \infty$ in the Chapman–Kolmogorov equations (1.4). Since every row of $\mathbf{P}(x)$ tends to $\boldsymbol{\pi}$, and every row of $\mathbf{P}(x + y)$ tends to $\boldsymbol{\pi}$, this yields

$$\boldsymbol{\pi} \mathbf{P}(y) = \boldsymbol{\pi}, \quad y \geq 0. \quad (1.29)$$

In other words, if at any moment the process state has the steady-state distribution, then it has the steady-state distribution at time y later, no matter how large or small y is. The state distribution becomes invariant with respect to time.

There are two important questions which arise in this connection. First, under what conditions does a steady-state regime exist for a Markov process? Second, how does one determine the steady-state distribution of the process? We shall leave the question of existence until the end of this section and concentrate now on the determination of the vector $\boldsymbol{\pi}$, assuming that it exists.

Differentiating (1.29) at $y = 0$, and remembering that $\mathbf{P}'(0) = \mathbf{A}$, we obtain a system of linear equations for $\boldsymbol{\pi}$:

$$\boldsymbol{\pi} \mathbf{A} = \mathbf{0}. \quad (1.30)$$

This is known as the system of “balance equations”, for reasons which will become apparent shortly. Being homogeneous, that system determines the vector $\boldsymbol{\pi}$ up to a multiplicative constant; the normalising equation (1.28) then completes the determination.

The balance equations have a strong intuitive appeal. To see this, let us write the i -th equation in the form

$$-a_{i,i} \pi_i = \sum_{\substack{j=0 \\ j \neq i}}^{\infty} a_{j,i} \pi_j. \quad (1.31)$$

Now, we can think of π_j as the proportion of time (in the steady-state) that the process spends in state j . While the process is in state j , it moves to state i at rate $a_{j,i}$ (since $a_{j,i}$ is the instantaneous transition rate from state j to state i ; see (1.7)). Therefore, the product $\pi_j a_{j,i}$ is equal to the average number of transitions from state j to state i per unit time. The right-hand side of (1.31) thus represents the average number of times that the process enters state i per unit time. Similarly, the left-hand side of (1.31) represents the average number of times that the process leaves state i per unit time (since $-a_{i,i}$ is the instantaneous transition rate out of state i ; see (1.8)). If the process is in equilibrium, these two averages must be equal. More generally, if $I = (i_1, i_2, \dots)$ is any group of states, finite or infinite, then the average number of times that the process enters group I per unit time is equal, in the steady-state, to the average number of times that the process leaves group I per unit time. The balance equations obtained by considering groups of states are not, of course, independent of the system (1.30); however, they are sometimes simpler and easier to deal with.

It is very convenient to describe a Markov process in equilibrium by means of a marked directed graph. This representation, called a “balance diagram”, makes it easier to visualise the process structure and often helps to select the set of balance equations best suited for determining the steady-state distribution. The nodes of the balance diagram correspond to the process states. With node i is associated the steady-state probability π_i ($i = 0, 1, \dots$). There is an arc from node i to node j ($i \neq j$) if the instantaneous transition rate $a_{i,j}$ is non-zero; that arc is labelled $a_{i,j}$. To obtain a balance equation from the diagram, cut off a group of nodes from the rest of the diagram by an imaginary closed curve. If an arc from node i to node j crosses the curve we say that there is a flow $\pi_i a_{i,j}$ across the cut. The total flow out of the cut (from nodes inside to nodes outside) is then equal to the total flow into the cut (from nodes outside to nodes inside). For instance, making a cut around node i alone, we obtain the balance equation (1.31). Note that the term “flow” used here is simply an abbreviation for “average number of transitions per unit time”.

Consider, as an example, the celebrated “Birth and Death” Markov process. As well as illustrating the methods of analysis, this example is of interest in its own right since a number of queueing system models turn out to be special cases of it. We think of the Birth and Death process $\{N(t), t \geq 0\}$ as representing the size of a certain population at time t . The only possible transitions out of state i are to states $i + 1$ and $i - 1$, with instantaneous transition rates λ_i , and μ_i , respectively (these are the

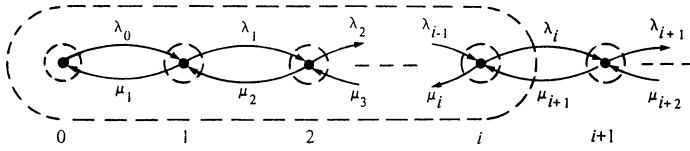


Fig. 1.3.

rates of “Birth” and “Death” when the population size is i), $i = 1, 2, \dots$. From state 0 the process moves to state 1, with instantaneous rate λ_0 . The balance diagram for the Birth and Death process is shown in Fig. 1.3.

Making a cut around each node in succession we obtain the system of balance equations (1.30);

$$\begin{aligned}
 \lambda_0\pi_0 &= \mu_1\pi_1 \\
 (\lambda_1 + \mu_1)\pi_1 &= \lambda_0\pi_0 + \mu_2\pi_2 \\
 &\dots\dots\dots \\
 (\lambda_i + \mu_i)\pi_i &= \lambda_{i-1}\pi_{i-1} + \mu_{i+1}\pi_{i+1} \\
 &\dots\dots\dots
 \end{aligned}
 \tag{1.32}$$

(there are two arcs going out and two arcs coming into each cut, except for node 0). Alternatively, cutting off the group of states $(0, 1, \dots, i)$, for $i = 0, 1, \dots$, we obtain an equivalent system of balance equations:

$$\begin{aligned}
 \lambda_0\pi_0 &= \mu_1\pi_1 \\
 \lambda_1\pi_1 &= \mu_2\pi_2 \\
 &\dots\dots\dots \\
 \lambda_i\pi_i &= \mu_{i+1}\pi_{i+1} \\
 &\dots\dots\dots
 \end{aligned}
 \tag{1.33}$$

(one arc going out and one arc coming into each cut). The general solution of (1.33) is easily obtained by successive elimination:

$$\pi_i = \frac{\lambda_0\lambda_1 \dots \lambda_{i-1}}{\mu_1\mu_2 \dots \mu_i}\pi_0, \quad i = 1, 2, \dots
 \tag{1.34}$$

This leaves one unknown constant, π_0 , which is determined from the normalising condition (1.28):

$$\pi_0 = \left(1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0\lambda_1}{\mu_1\mu_2} + \dots \right)^{-1}.
 \tag{1.35}$$

Note that we have here a necessary condition for equilibrium of the Birth and Death process in order for the solution given by (1.34) and (1.35) to be a probability distribution, the infinite series on the right-hand side of (1.35) must converge. We shall see that the inverse implication also holds: if the series converges, the Birth and Death process has a steady-state. However, in order to state the general result, some preliminaries are needed.

The state j of a Markov process is said to be “reachable” from state i if there is a non-zero probability of finding the process in state j at time t , given that it started in state i : $p_{i,j}(t) > 0$ for all $t > 0$. Since transition probability functions are zero either everywhere or nowhere on the open half-line, the state j is not reachable from i if $p_{i,j}(t) = 0$ for all $t > 0$. A subset σ of process states is said to be “closed” if no state outside σ is reachable from a state in σ . Thus, if the process once enters a closed subset of states, it remains in that subset for ever afterwards. A set of states is said to be “irreducible” if no proper and non-empty subset of it is closed. As far as the long-run behaviour of the process is concerned, an irreducible set of states can be treated in isolation, so we can assume that the set of all states, i.e. the Markov process, is irreducible.

Every state of an irreducible Markov process is reachable from every other state. Indeed, suppose that this is not so, and let i and j be two states such that j is not reachable from i . Consider the set σ of all states reachable from i . That set is closed, since any state k reachable from a state in σ is also reachable from i (this follows from (1.3)) and hence $k \in \sigma$. But σ does not contain j , which contradicts the irreducibility of the process.

The states of a Markov process $\{S(t), t \geq 0\}$ can be classified according to the time the process spends in them. Let $R_{i,j}(t)$ be the average amount of time spent in state j during the interval $[0, t)$, given that $S(0) = i$. Introducing the indicator function of a Boolean B

$$I_B = \begin{cases} 1 & \text{if } B \text{ is true} \\ 0 & \text{if } B \text{ is false} \end{cases}$$

we can write

$$\begin{aligned} R_{i,j}(t) &= E \left[\int_0^t I_{(S(u)=j)} du \mid S(0) = i \right] \\ &= \int_0^t E[I_{(S(u)=j)} \mid S(0) = i] du \\ &= \int_0^t P(S(u) = j \mid S(0) = i) du = \int_0^t p_{i,j}(u) du. \end{aligned} \quad (1.36)$$

Further, let $R_{i,j}$ be the total average amount of time spent in state j , given that $S(0) = i$:

$$R_{i,j} = \lim_{t \rightarrow \infty} R_{i,j}(t) = \int_0^{\infty} p_{i,j}(u) du. \quad (1.37)$$

A state j is said to be “transient” if $R_{i,j}$ is finite; otherwise j is “recurrent”. Since the average time the process remains in state j on every visit is finite (it is equal to $-1/a_{j,j}$), the average number of visits to state j is finite if j is transient and it is infinite if j is recurrent. Denote by $f_{i,j}$ the probability that, starting in state i , the process will ever be in state j . From the remarks above it follows that if state j is recurrent, then $f_{j,j} = 1$ and if state j is transient, then $f_{j,j} < 1$. The inverse implications also hold.

If the Markov process is irreducible, and if $R_{i,j} = \infty$ for some pair of states i and j , then $R_{r,k} = \infty$ for any pair of states r, k . Indeed, taking two arbitrary positive constants v and w , we can write

$$\begin{aligned} R_{r,k} &= \int_0^{\infty} p_{r,k}(u) du \geq \int_0^{\infty} p_{r,k}(v + u + w) du \\ &\geq \int_0^{\infty} p_{r,i}(v) p_{i,j}(u) p_{j,k}(w) du = p_{r,i}(v) p_{j,k}(w) R_{i,j} = \infty. \end{aligned}$$

(The first inequality is obvious; the second follows from the Chapman–Kolmogorov equations (1.3); the irreducibility of the process ensures that $p_{r,i}(v) > 0$ and $p_{j,k}(w) > 0$.) Hence, either all states are transient, or all states are recurrent.

The case of all transient states can be disposed of quickly: if $R_{i,j}$ is finite for all i, j then, according to (1.37),

$$\lim_{t \rightarrow \infty} p_{i,j}(t) = 0, \quad i, j = 0, 1, \dots$$

In that case, steady-state does not exist.

Suppose now that the Markov process is recurrent, as well as irreducible. Every state is guaranteed to be visited, no matter what the initial state is (if the probability of eventually moving from state i to state j were not 1, there would be a non-zero probability of moving from j to i and not returning to j ; state j would not be recurrent). Having once visited a state, the process keeps returning to it *ad infinitum*. Let m_j be the average length of the intervals between consecutive returns to state j , $j = 0, 1, \dots$. That average length may be finite, in which case state j is said to be “recurrent non-null”, or it may be infinite, and then j is “recurrent null”.

The moments t_1, t_2, \dots of successive visits to state j are “regeneration points” for the Markov process: the process behaviour in the interval $[t_n, t_{n+1})$ is a probabilistic replica of that in the interval $[t_{n-1}, t_n)$. The average time spent in state j during each of these intervals is $-1/a_{j,j}$. Therefore, in the long run, the fraction of time that the process spends in state j is independent of the initial state and is given by

$$\lim_{t \rightarrow \infty} \frac{R_{i,j}(t)}{t} = \frac{-(1/a_{j,j})}{m_j}, \quad i, j = 0, 1, \dots \quad (1.38)$$

On the other hand, that fraction of time is equal to the long-run probability of finding the process in state j :

$$\lim_{t \rightarrow \infty} p_{i,j}(t) = \frac{-(1/a_{j,j})}{m_j}, \quad i, j = 0, 1, \dots \quad (1.39)$$

Equations (1.38) and (1.39) seem intuitively clear, yet to prove them rigorously is not easy. Some fundamental results from renewal theory are involved (see, for example, Cinlar [3]).

It follows from (1.39) that the limiting probability of state j is zero if $m_j = \infty$, i.e. if j is recurrent null, and vice versa. Moreover, if one state, j , is recurrent null, then all other states are also recurrent null. Choose an arbitrary state k ($k \neq j$) and two positive constants v and w . The following inequality follows from the Chapman–Kolmogorov equations (1.3):

$$p_{j,j}(v+t+w) \geq p_{j,k}(v)p_{k,k}(t)p_{k,j}(w).$$

Since $p_{j,j}(t)$ tends to 0 as $t \rightarrow \infty$, so must $p_{k,k}(t)$; hence, state k is recurrent null.

Let us recapitulate the results obtained so far. In an irreducible Markov process, either all states are transient, or all states are recurrent null, or all states are recurrent non-null. In the first two cases, all limiting probabilities are equal to 0; steady-state does not exist. In the last case, all limiting probabilities are non-zero; steady-state exists.

We have seen already that if a steady-state distribution vector $\boldsymbol{\pi}$ exists, it satisfies the system of balance equations (1.30) and the normalising equation (1.28). Now we shall demonstrate that if equations (1.30) and (1.28) have a solution, $\boldsymbol{\pi}$, then steady-state exists.

First, taking the known expression (1.13) for the transition probability matrix

$$\mathbf{P}(t) = \sum_{n=0}^{\infty} \mathbf{A}^n t^n / n!$$

and multiplying both sides by $\boldsymbol{\pi}$ on the left, we see that if $\boldsymbol{\pi}\mathbf{A} = \mathbf{0}$, then

$$\boldsymbol{\pi}\mathbf{P}(t) = \boldsymbol{\pi} \quad \text{for all } t \geq 0. \quad (1.40)$$

Let $t \rightarrow \infty$ in this section. If the Markov process were transient or recurrent null, then every column of $\mathbf{P}(t)$ would tend to $\mathbf{0}$ and we would have $\boldsymbol{\pi} = \mathbf{0}$. That, however, is impossible since $\boldsymbol{\pi}$ satisfies (1.28). Therefore, the process must be recurrent non-null and hence steady-state exists. In the latter case, all elements of the j -th column of $\mathbf{P}(t)$ tend to the same constant, γ_j (given by (1.39)). The j -th equation in (1.40) becomes, in the limit,

$$\pi_j = \sum_{i=0}^{\infty} \pi_i \gamma_j = \gamma_j, \quad j = 0, 1, \dots$$

In other words, if a solution of (1.30) and (1.28) exists, then it is unique and is precisely the steady-state distribution of the process.

So, an irreducible Markov process $\{S(t), t \geq 0\}$ has a steady-state regime if, and only if, the balance equations (1.30) have a solution $\boldsymbol{\pi} = (\pi_0, \pi_1, \dots)$ whose elements sum up to 1; that solution is then unique and represents the steady-state distribution of the process:

$$\pi_j = \lim_{t \rightarrow \infty} P(S(t) = j), \quad j = 0, 1, \dots$$

This important result is the point of departure for most analytic and numerical studies of systems modelled by Markov processes.

Returning to the Birth and Death process considered earlier, we can assert now that the necessary and sufficient condition for existence of steady-state is the convergence of the series appearing on the right-hand side of (1.35); when it exists, the steady-state distribution is given by (1.34) and (1.35). That assertion follows from the result above and from the fact that the Birth and Death process is irreducible; the probability $p_{i,j}(t)$ of moving from state i to state j in time t is obviously non-zero, for all i, j and all $t > 0$.

1.5. The $M/M/1$, $M/M/c$ and related queueing systems

We shall examine here several models which fit easily into the framework of the theory developed in the last section. Although these models are rather simple, they manage to capture and display some essential features of mass-service systems. In particular, they illustrate very clearly the way in which

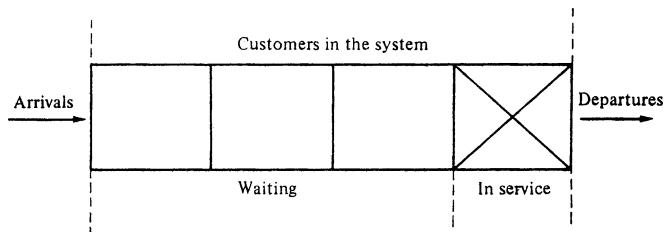


Fig. 1.4.

system performance is influenced by the level of user demand and by the capacity and availability of servers.

Consider a single-server queueing system where all customers are of the same type and are served in order of arrival (that service discipline is usually referred to as FIFO, first-in-first-out, or FCFS first-come-first-served). There is no restriction on the size of the queue that may develop and no customer leaves the queue before completing service (Fig. 1.4). Such a system can be used to model a counter at a bank, a car-washing station, a uniprogrammed computer, etc. Under a suitable set of assumptions the model becomes a Markov process which lends itself to analysis. The simplest way to ensure that the memoryless property holds is to assume that consecutive interarrival times are independent and distributed exponentially with mean $1/\lambda$ (i.e. the arrival stream is Poisson with rate λ), and consecutive service times are independent and distributed exponentially with mean $1/\mu$; also, the arrival and service processes are mutually independent. We thus obtain the $M/M/1$ queueing model. Let $N(t)$ be the number of customers in the system (waiting and in service) at time t . From the memoryless property of the exponential distribution it follows that

$$P(N(t+y) = j \mid N(t) = i) = P_{i,j}(y), \quad i, j = 0, 1, \dots$$

independently of t and of the past history $\{N(u), u < t\}$. Therefore, $\{N(t), t \geq 0\}$ is a Markov process. The only possible transitions out of state i ($i = 1, 2, \dots$) are to states $i+1$ (if an arrival occurs before a service completion) and $i-1$ (if a service completion occurs before an arrival). The instantaneous transition rates are $a_{i,i+1} = \lambda$ and $a_{i,i-1} = \mu$. From state 0 the process always moves to state 1, with instantaneous rate $a_{0,1} = \lambda$.

We recognise here a special case of the Birth and Death process introduced in the last section, with $\lambda_i = \lambda$ ($i = 0, 1, \dots$) and $\mu_i = \mu$ ($i = 1, 2, \dots$).

Denoting $(\lambda/\mu) = \rho$, the general solution (1.34) of the balance equations becomes

$$\pi_i = \rho^i \pi_0, \quad i = 0, 1, \dots \quad (1.41)$$

The necessary and sufficient condition for the existence of a solution whose elements sum up to 1, and hence for the existence of steady-state, is $\rho < 1$. When the system is in equilibrium, the number of customers in it is distributed geometrically:

$$P(N = i) = \pi_i = \rho^i (1 - \rho), \quad i = 0, 1, \dots \quad (1.42)$$

The expectation, $E[N]$, and the variance, $\text{Var}[N]$, of that number are given by

$$E[N] = \sum_{i=1}^{\infty} i \pi_i = \rho / (1 - \rho), \quad (1.43)$$

and

$$\text{Var}[N] = E[N^2] - E^2[N] = \sum_{i=1}^{\infty} i^2 \pi_i - E^2[N] = \rho / (1 - \rho)^2. \quad (1.44)$$

In order to give physical meaning to these results, it is helpful to distinguish the amount of service required by a customer, or the “job length”, from the speed of the server. Job lengths are measured in “units of work” (in computer systems the unit of work is usually a machine instruction), while the speed of the server is measured in “units of work per unit time”. The time unit can always be chosen so that the server speed is 1; then the service time of a customer is simply the amount of work that he requires.

The average number of customers arriving into the system per unit time is λ . The average amount of work required by a customer is $1/\mu$. Hence, the quantity ρ represents the average amount of work brought into the system per unit time; for that reason, it is referred to as “traffic intensity”. The condition for existence of steady-state now reads: the average amount of work brought into the system per unit time must be less than the speed of the server (the amount of work that it can do per unit time). This is a very natural requirement; we shall come across it many times, under much more general assumptions.

When the traffic intensity is less than 1, the process $\{N(t), t \geq 0\}$ is recurrent non-null. Every state, and in particular the state $N = 0$, occurs infinitely many times, at intervals whose expectations are finite. The system goes through alternating “busy” and “idle” periods. We shall see at the end

of this section that the steady-state distribution can, in fact, be determined directly from these regeneration cycles.

As $\rho \rightarrow 1$, the steady-state average number of customers in the system tends to infinity. The state $N = 0$ occurs less and less often. The variance of N also tends to infinity, which means that a randomly observed queue size is likely to be very far from the expected one.

When $\rho = 1$, the process is recurrent null (we state this without proof). Every state is still visited infinitely many times but the intervals between visits are infinitely long on the average. The long-run mean and variance of N are infinite, and the probability of observing any given N is zero.

When $\rho > 1$, the process is transient (again we give no proof). The number of jobs in the system grows eventually above any finite number, never to drop below it again. Not only is the fraction of time that the system spends in any given state zero in the long run, but the total time it spends in any state is finite.

We shall sometimes use the terms “non-saturated system” and “saturated system” to describe the cases $\rho < 1$ and $\rho \geq 1$, respectively.

A random variable of central importance in a queueing system is the steady-state response time, w (the time a customer spends in the system). The average response time is often taken as a measure of system performance.

We now proceed to find the probability density function $f_w(x)$ of the response time in an $M/M/1$ system in equilibrium. First, from the random observer property of the Poisson stream (see section 1.3), it follows that an arriving customer sees the steady-state distribution (1.42) of the number of customers in the system. Next, from the memoryless property of the exponential distribution, if the new arrival finds a customer in service, the remaining service time of that customer is distributed exponentially with mean $1/\mu$. The response time of a customer who finds n customers in the system is therefore the sum of $n + 1$ independent exponentially distributed random variables. Such a sum has the $n + 1$ stage Erlang density function $g_{n+1}(x)$ defined in section 1.3:

$$g_{n+1}(x) = \mu(\mu x)^n e^{-\mu x} / n!. \quad (1.45)$$

Combining (1.42) and (1.45), and remembering that $\rho = \lambda/\mu$, we obtain

$$f_w(x) = \sum_{n=0}^{\infty} \pi_n g_{n+1}(x) = (1 - \rho) \mu e^{-\mu x} \sum_{n=0}^{\infty} (\rho \mu x)^n / n! = (\mu - \lambda) e^{-(\mu - \lambda)x}. \quad (1.46)$$

The response time is thus distributed exponentially. No matter how long a customer has already spent in the system, his remaining time there still has the same distribution. The average response time $W = E[w]$ is equal to

$$W = \frac{1}{\mu - \lambda} = \frac{1}{\mu(1 - \rho)}. \quad (1.47)$$

Note that this performance measure differs from those in (1.43) and (1.44) in that it depends on λ and μ not just through their ratio ρ . It is possible for a system to be nearly saturated, with large queue sizes, and yet to have a very short expected response time.

Let us now generalise the model by allowing c parallel servers (each of unit speed), keeping the other assumptions as before. This is the $M/M/c$ queueing system. If at a given moment there are i customers in the system, the number of customers in service is $\min(i, c)$. Since each service time is distributed exponentially with parameter μ , the interval until the nearest service completion is distributed exponentially with parameter $\mu \min(i, c)$. The process representing the number of customers in the system, $\{N(t), t \geq 0\}$, is therefore a Birth and Death process with constant birth rate, $\lambda_i = \lambda$ ($i = 0, 1, \dots$), and state-dependent death rate, $\mu_i = \mu \min(i, c)$. The general solution (1.34) of the balance equations is

$$\pi_i = \begin{cases} (\rho^i / i!) \pi_0, & i = 0, 1, \dots, c \\ [\rho^i / (c! c^{i-c})] \pi_0 = (\rho/c)^{i-c} \pi_c, & i > c. \end{cases} \quad (1.48)$$

Steady-state exists if, and only if, $\rho < c$. As before, this is a requirement that the average amount of work brought into the system per unit time should be less than the amount of work that can be done per unit time. When $\rho \geq c$ the system is saturated (recurrent null if $\rho = c$ and transient if $\rho > c$).

To determine the steady-state distribution we need the probability of the idle state:

$$\pi_0 = \left[\sum_{i=0}^{c-1} (\rho^i / i!) + (\rho^c / c!) c / (c - \rho) \right]^{-1}. \quad (1.49)$$

Various performance measures can now be obtained, although the expressions tend to be complicated. In general, an $M/M/c$ system is less efficient than an $M/M/1$ system with an equivalent service capacity. Let us carry out the comparison between an $M/M/2$ system with parameters λ and μ , and $M/M/1$ system with parameters λ and 2μ . The non-saturation

condition is, in both cases, $\lambda < 2\mu$. We shall use the expected number of customers in the system, $E[N]$, as a measure of performance. In the $M/M/1$ system we have, from (1.43),

$$E[N]_{M/M/1} = \frac{\lambda}{2\mu - \lambda}.$$

For the $M/M/2$ system, we find first

$$\pi_0 = [1 + \rho + \rho^2/(2 - \rho)]^{-1} = (2 - \rho)/(2 + \rho).$$

The expression for $E[N]$ now becomes

$$E[N]_{M/M/2} = \sum_{i=1}^{\infty} i\pi_i = \frac{4\lambda\mu}{(2\mu - \lambda)(2\mu + \lambda)}.$$

The non-saturation condition implies that

$$\frac{4\mu}{2\mu + \lambda} > 1.$$

Therefore

$$E[N]_{M/M/2} > E[N]_{M/M/1}.$$

A similar inequality holds for any number of servers. The reason for the worse performance of the $M/M/c$ system is that its full service capacity is not always utilised: when there are less than c customers in the system, some servers are idle. The $M/M/c$ system is, in its turn, more efficient than c independent servers with separate queues (i.e. c $M/M/1$ systems), where each new arrival joins any of the queues with equal probability. We leave that comparison as an exercise to the reader. The lesson that emerges from all this is that, other things being equal, a pooling of resources leads to improved performance.

A limiting case of the $M/M/c$ system is the system with infinitely many servers, $M/M/\infty$. Clearly, there can be no queue of waiting customers here. The solution of the balance equations is as in (1.48), top case, for all i :

$$\pi_i = (\rho^i/i!)\pi_0, \quad i = 0, 1, \dots \quad (1.50)$$

That solution can always be normalised:

$$\pi_0 = \left[\sum_{i=0}^{\infty} (\rho^i/i!) \right]^{-1} = e^{-\rho}.$$

Hence, steady-state always exists. This, of course, is hardly surprising since the service capacity is infinite. The expected number of customers in the system is $E[N] = \rho$.

Other members of the Birth and Death family are models with limited waiting room: there is a maximum number K of customers that can be allowed into the system at any one time. All new arrivals who find K customers in the system are turned away and are lost. Steady-state always exists in these systems because the number of states is finite. When a limit on the waiting room is imposed, it is included in the Kendall notation as another descriptor after the number of servers, e.g. $M/G/1/K$. We shall mention here two systems of this type. The first is the $M/M/1/K$ system, where there can be one customer in service and at most $K - 1$ waiting.

For us the interest of this model lies in the fact that it is equivalent to the following closed cyclic system: K customers circulate endlessly between two servers, 1 and 2, whose service times are distributed exponentially with means $1/\mu$ and $1/\lambda$, respectively. The order of service is FIFO at both servers (Fig. 1.5). The cyclic model can be applied, for example, to a computer system consisting of one CPU and one Input/Output device, with K jobs sharing the main memory.

To see the equivalence between the $M/M/1/K$ and the cyclic system note that as long as the number of customers at server 1 is less than K , customers arrive there at intervals distributed exponentially with mean $1/\lambda$; when all K customers are at server 1, the arrivals stop. This is the same as having a Poisson arrival stream which is turned off in state K .

The steady-state distribution of the $M/M/1/K$ system state is given by

$$\pi_i = \rho^i \pi_0, \quad i = 0, 1, \dots, K, \quad (1.51)$$

where $\rho = \lambda/\mu$ and $\pi_0 = (1 - \rho)/(1 - \rho^{K+1})$; when $\rho = 1$, $\pi_i = 1/(K + 1)$, $i = 0, 1, \dots, K$.

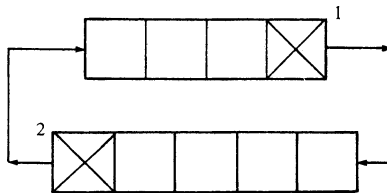


Fig. 1.5.

Our other example is the $M/M/c/c$ system, where only customers who find idle servers are admitted. The classic application for such a model is a telephone exchange with c lines. The steady-state distribution of the number of busy servers is

$$\pi_i = (\rho^i/i!)\pi_0, \quad i = 0, 1, \dots, c, \quad (1.52)$$

where

$$\pi_0 = \sum_{i=0}^c (\rho^i/i!)^{-1}.$$

A measure of performance for this system is the fraction of customers that is lost. Since Poisson arrivals behave like random observers, that fraction is equal to

$$\pi_c = (\rho^c/c!) \left/ \left[\sum_{i=0}^c (\rho^i/i!) \right] \right. . \quad (1.53)$$

Expression (1.53) is known as “Erlang’s loss formula”.

Let us now return to the $M/M/1$ queueing system and analyse it in the steady-state by applying a renewal theory argument. We have mentioned that when $\rho < 1$, every state is entered infinitely many times at intervals whose expectation is finite. Let t_1, t_2, \dots be the consecutive moments when the queueing process $\{N(t), t \geq 0\}$ enters state 0. These moments are regeneration points for the process: the behaviour of $N(t)$ on the interval $[t_j, t_{j+1})$ is an independent probabilistic replica of its behaviour on the interval $[t_{j-1}, t_j)$, $j = 2, 3, \dots$. In particular, the interval lengths $(t_{j+1} - t_j)$, $j = 1, 2, \dots$, are independent and identically distributed. Denote their expectation by T :

$$T = E[t_{j+1} - t_j].$$

Let T_i be the total expected amount of time that the process spends in state i during a regeneration period ($i = 0, 1, \dots$). By the same argument that led to equations (1.38) and (1.39) it can be shown that the long-run fraction of time that the process spends in state i , and hence the steady-state probability of state i , is given by

$$\pi_i = T_i/T, \quad i = 0, 1, \dots \quad (1.54)$$

Now we proceed to find the expectations T_i . Denote by M_i the average number of visits to state i during a regeneration period. Then, since the

average time the process remains in state 0 on each visit is $1/\lambda$, and the average time it remains in state i ($i > 0$) on each visit is $1/(\lambda + \mu)$, we have

$$\begin{aligned} T_0 &= M_0/\lambda \\ T_i &= M_i/(\lambda + \mu), \quad i = 1, 2, \dots \end{aligned} \quad (1.55)$$

But a visit to state i is either a result of a transition from state $i - 1$, with probability $\lambda/(\lambda + \mu)$, or a result of a transition from state $i + 1$, with probability $\mu/(\mu + \lambda)$, $i = 1, 2, \dots$. State 0 can be entered only from state 1, with probability $\mu/(\mu + \lambda)$ (for the transition probabilities, see (1.17)). Hence,

$$\begin{aligned} M_0 &= \frac{\mu}{\lambda + \mu} M_1 \\ M_i &= \frac{\lambda}{\lambda + \mu} M_{i-1} + \frac{\mu}{\lambda + \mu} M_{i+1}, \quad i = 1, 2, \dots \end{aligned} \quad (1.56)$$

Substituting (1.55) into (1.56) we obtain

$$\begin{aligned} \lambda T_0 &= \mu T_1 \\ (\lambda + \mu) T_i &= \lambda T_{i-1} + \mu T_{i+1}, \quad i = 1, 2, \dots \end{aligned} \quad (1.57)$$

These last equations, together with $T_0 = 1/\lambda$ (there is only one visit to state 0 during a regeneration interval) can be solved by successive elimination:

$$T_i = \rho^i/\lambda, \quad i = 0, 1, \dots \quad (1.58)$$

The average length of a regeneration interval is, of course, equal to

$$T = \sum_{i=0}^{\infty} T_i = 1/[\lambda(1 - \rho)]. \quad (1.59)$$

Substituting (1.58) and (1.59) into (1.54) we finally obtain the desired distribution

$$\pi_i = \rho^i(1 - \rho), \quad i = 0, 1, \dots$$

Note that the above approach can be applied to the general Birth and Death process as well, with obvious minor modifications.

1.6. Little's result. Applications. The $M/G/1$ system

We shall derive here a simple relation between the average response time and the average number of customers in a queueing system in equilibrium.

The first rigorous proof of that relation was given by Little [6]; hence, it is known as “Little’s result”, or “Little’s theorem”. However, the validity of the result had been realised earlier and there were also proofs for some special cases.

Consider an arbitrary queueing system in equilibrium, and let N , W and λ be the average number of customers in the system, the average time customers spend in the system and the average number of arrivals per unit time, respectively. Little’s theorem states that

$$N = \lambda W, \tag{1.60}$$

regardless of the interarrival and service time distributions, the service discipline and any dependencies within the system. Note that we have not even specified what constitutes “the system”, nor what customers do there. It is just a place where customers arrive, remain for some time and then depart. The only requirement is that the processes involved should be stationary (independent of time).

Let us first give an intuitive justification for (1.60). Suppose that the system receives a reward (or penalty) of 1 for every unit of time that a customer spends in it. Then the total expected reward per unit time is equal to the average number of customers in the system, N . On the other hand, the average number of customers coming into the system per unit time is λ ; the expected reward contributed by each customer is equal to his average residence time, W . Since it does not matter whether the reward is collected on arrival or continuously, we must have $N = \lambda W$. (This, and the following argument and proof, are due to Foster [5].)

A different interpretation of relation (1.60) is obtained by rewriting it in the form $\lambda = (N/W)$. Since a customer in the system remains there for an average time of W , his average rate of departure is $1/W$. The total average departure rate is, therefore, N/W . Thus, the relation holds if the average arrival rate is equal to the average departure rate. But the latter is clearly the case since the system is in equilibrium.

The above arguments should suffice to convince us that Little’s result holds in its full generality. To prove it formally (admittedly in a slightly less general case: arrivals in batches will be excluded), denote by $F_w(x)$ the probability distribution function of the response time. The average response time is given by

$$W = \int_0^{\infty} (1 - F_w(x)) dx.$$

Fix an arbitrary moment t in the steady-state. The customers who are in the system at that moment are those who arrived before t and will depart after t . Since the arrival process is stationary with rate λ and customers arrive one at a time, the probability that there was an arrival at time $t - u$ is λdu . Such an arrival is still in the system at time t with probability $1 - F_w(u)$. Therefore, point $t - u$ contributes an average of $\lambda(1 - F_w(u))du$ customers to the ones present at time t . Integrating over all values of u yields

$$N = \int_0^\infty \lambda(1 - F_w(u))du = \lambda W,$$

thus establishing the result.

Little's original proof of the relation contained another, more basic assertion: if the space averages N , λ and W are replaced by time averages — that is, averages over an individual realisation of the queueing process — then in every such realisation (1.60) holds with probability 1. This is an instance of an operational identity.

Let us now turn to some applications. Consider first a queueing system where customers are served by a number (finite or infinite) of identical servers of unit speed. Denote, as before, the arrival rate by λ and the average service time by $1/\mu$. The relevant distributions can be general, as can be the scheduling discipline. Assume further that customers do not leave before receiving service. Define the set of servers, σ , as “the system”, for the purpose of Little's theorem. Since every incoming customer enters a server eventually, the rate of arrivals into σ is also λ . The average time a customer spends in σ is equal to $1/\mu$. According to the theorem, the average number of customers in σ is λ/μ .

Thus, in any $G/G/c$ or $G/G/\infty$ system in equilibrium, the average number of busy servers is equal to the traffic intensity, ρ . One consequence of this is that the condition $\rho < c$ is necessary for the existence of equilibrium in the general case (we have already seen that it is necessary and sufficient in the case $M/M/c$). When $c = 1$, the average number of busy servers is equal to the probability that the server is busy. Therefore, in any single-server system in the steady-state we have

$$P(\text{there are customers in the system}) = \rho, \tag{1.61}$$

$$P(\text{idle system}) = 1 - \rho.$$

Suppose that the customer population is split into classes, numbered $1, 2, \dots$, with different characteristics. Let the arrival rate and the average

service time of class i customers be λ_i and $1/\mu_i$, respectively ($i = 1, 2, \dots$). Then, applying Little's theorem to the class i customers only, we find (in exactly the same way as above) that the average number of class i customers in service is $\rho_i = (\lambda_i/\mu_i)$, and it is necessary that $\sum_i \rho_i < c$.

In single-server systems we have

$$P(\text{a customer of class } i \text{ is in service}) = \rho_i, \quad (1.62)$$

$$P(\text{idle system}) = 1 - \sum_i \rho_i.$$

As our next example, we shall find the steady-state average number of customers, N , and the average response time, W , in a single-server system with Poisson arrivals and generally distributed service times ($M/G/1$). The scheduling discipline is FIFO; interarrival and service times are assumed to be mutually independent; there is a single-customer class. The techniques of the previous sections cannot be applied to this system (at least not directly), because the process $\{N(t), t \geq 0\}$ representing the number of customers in the system is not Markov in general. When the distribution of service times is not exponential, the process behaviour after a given moment depends on its history prior to that moment. However, here we are only interested in the averages N and W , which can be obtained by a rather simple argument. A more detailed study of the $M/G/1$ system will be presented in Chapter 2.

By the random observer property of the Poisson stream, a new arrival into the system finds an average of N customers there. Of these, we saw that an average of ρ are being served and $N - \rho$ are waiting in the queue. Each of the waiting customers will take an average of $1/\mu$ to serve, as will the new arrival himself. Denote by W_0 the expected remaining service time of a customer found in service by a random observer. We can then write, for the expected residence time of the new arrival,

$$W = \rho W_0 + (N - \rho)(1/\mu) + (1/\mu).$$

Substituting Little's result, $W = N/\lambda$, in this equation, and solving for N , we obtain

$$N = \rho + \lambda W_0 \frac{\rho}{1 - \rho}. \quad (1.63)$$

It remains to determine the quantity W_0 . To do this, imagine the consecutive service intervals laid end-to-end on the time axis, thus eliminating any idle periods. The resulting sequence of independent and identically

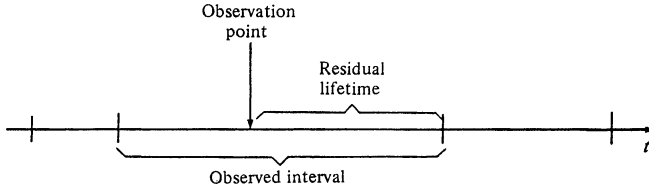


Fig. 1.6.

distributed intervals forms a “renewal process”. The end-points of the renewal intervals are called “renewal epochs”. We are interested in the random variable representing the time between a random observation point and the next renewal epoch (Fig. 1.6); this is the “residual lifetime” of the renewal interval (also sometimes called “random modification” or “forward recurrence time”).

Let $f(x)$, m and M_2 be the probability density function, the mean and the second moment of the renewal interval, respectively (in our case, $m = 1/\mu$). Consider the renewal process over a very long period of time, T . Since, on average, there are T/m renewal intervals during T , and since a renewal interval is of length x with probability $f(x)dx$, the average number of renewal intervals of length x during the period T is equal to $[Tf(x)dx]/m$. Hence, the average portion of T covered by renewal intervals of length x is equal to $[Tx f(x)dx]/m$. The random observation point is, by definition, equally likely to fall anywhere in T ; therefore, the probability $\tilde{f}(x)dx$ that the observed renewal interval is of length x is given by

$$\tilde{f}(x)dx = (xf(x)dx)/m. \quad (1.64)$$

From (1.64) we obtain the average length, \tilde{m} , of the observed renewal interval:

$$\tilde{m} = \int_0^{\infty} x\tilde{f}(x)dx = M_2/m. \quad (1.65)$$

Note that \tilde{m} is always greater than or equal to m , with equality only when $M_2 = m^2$, i.e. when the variance of the renewal interval is zero. This is because a renewal interval which receives the observation point is more likely to be long than one which does not. Since the observation point is equally likely to fall anywhere in the observed interval, the expected residual lifetime is equal to

$$W_0 = \tilde{m}/2 = M_2/(2m). \quad (1.66)$$

It is interesting that, in some cases, the expected residual lifetime can be greater than the expected lifetime!

We can now substitute (1.66) with $m = 1/\mu$, into (1.63). This yields

$$N = \rho + \frac{\lambda^2 M_2}{2(1 - \rho)}. \quad (1.67)$$

The above expression is known as Pollaczek-Khintchine's formula. It is usually written in the form

$$N = \rho + \frac{\rho^2(1 + C^2)}{2(1 - \rho)}. \quad (1.68)$$

where $C^2 = \text{Var}[s]/(E[s])^2 = (\mu^2 M_2) - 1$ is the squared coefficient of variation of the service time s . Here, as in the $M/M/1$ system, we note the appearance of $(1 - \rho)$ in the denominator; the expected number of customers in the system approaches infinity as $\rho \rightarrow 1$. For fixed λ and μ , the value of N is determined by the coefficient of variation of the service times. When $C^2 > 1$, the $M/G/1$ system performance is worse than that of the $M/M/1$ system ($C^2 = 1$ for the exponential distribution); when $C^2 < 1$ it is better. The average response time W in the $M/G/1$ system can, of course, be determined easily from Little's theorem: $W = N/\lambda$.

For the last example we return to a topic covered twice already: the steady-state distribution $\{\pi_0, \pi_1, \dots\}$ of the number of customers in the $M/M/1$ system. An ingenious derivation, using Little's theorem, was proposed by Foster [5]. Its interest lies in the conjuring trick whereby a distribution is pulled out of a hat containing only averages.

Identify individual queue positions by numbering them $1, 2, \dots$: 1 is the service position, 2 is the first waiting position, etc. The steady-state probability q_j that the j -th position is occupied is equal to the probability that there are j or more customers in the system:

$$q_j = \pi_j + \pi_{j+1} + \dots; \quad j = 1, 2, \dots$$

This is also the average number of customers in the j -th position.

After a service completion, every customer in the system moves by one queue position to the next lower index. Every customer who finds, on arrival, $j - 1$ or more customers in the system, passes eventually through position j . Therefore, the rate of arrivals into position j is λq_{j-1} ($j = 1, 2, \dots$; $q_0 = 1$ by definition). The average time that customers remain in position j is equal to $1/\mu$, regardless of whether they arrive there directly or from position $j + 1$ (this is because of the memoryless property

of the exponential distribution). Applying Little's theorem to the "system" consisting of the j -th queue position gives

$$q_j = \lambda q_{j-1} / \mu = \rho q_{j-1}; \quad j = 1, 2, \dots$$

This, together with $q_0 = 1$, yields

$$q_j = \rho^j \quad j = 1, 2, \dots, \quad \text{or} \quad \pi_j = q_j - q_{j+1} = \rho^j (1 - \rho), \quad j = 0, 1, \dots;$$

the same expression as (1.42).

1.7. Operational identities

It will be instructive, at this point, to examine more closely the sample path behaviour of a general queueing process, $\{N(t), t \geq 0\}$, representing the number of customers in a system. Any such sample path is a step function of the type illustrated in Fig. 1.1: the function jumps up by one at arrival instants and it jumps down by one at service completion instants (bulk arrivals and departures are excluded). In this section only, $N(t)$ will denote a sample path function; it should be remembered that this is not now a random variable, but an ordinary function of t describing a particular realisation of the queueing process.

Consider a sample path $N(t)$ over a time interval $[a, b]$ such that $N(a) = N(b)$ (Buzen [2]). Let m and M be, respectively, the minimum and the maximum values reached by $N(t)$ on $[a, b]$. Since all jumps are of unit magnitude, every value n in the range $m \leq n \leq M$ is attained at least once during that interval. For each such n , denote:

$T(n)$, the total amount of time the sample path remains at level n during $[a, b]$;

$A(n)$, the number of jumps from n to $n + 1$ during $[a, b]$ (i.e. the number of arrivals who find n customers in the system);

$D(n)$, the number of jumps from n to $n - 1$ during $[a, b]$ (i.e. the number of departures who leave $n - 1$ customers behind). Clearly, $A(n) > 0$ for $n = m, m + 1, \dots, M - 1$ and $D(n) > 0$ for $n = m + 1, m + 2, \dots, M$.

From the "operational equilibrium" condition $N(a) = N(b)$ it follows that

$$A(n) = D(n + 1), \quad n = m, m + 1, \dots, M - 1. \quad (1.69)$$

This equation yields, in a straightforward manner,

$$\frac{A(n)}{T(n)} \frac{T(n)}{T} = \frac{D(n+1)}{T(n+1)} \frac{T(n+1)}{T}, \quad n = m, m+1, \dots, M-1, \quad (1.70)$$

where $T = b - a$ is the length of the observation interval. Now, $A(n)/T(n)$ is the observed average number of arrivals per unit time in state n ; denote it by $\lambda(n)$. Similarly, $\mu(n) = D(n)/T(n)$ is the observed average number of service completions per unit time in state n . The ratio $T(n)/T$, which we shall denote by $p(n)$, represents the observed proportion of time that the system remains in state n . In this notation, (1.70) becomes

$$\lambda(n)p(n) = \mu(n+1)p(n+1), \quad n = m, m+1, \dots, M-1. \quad (1.71)$$

Note the similarity of form between (1.71) and the balance equations (1.33) of the Birth and Death process. It should be realised, however, that the content is very different. The relations (1.33) were between the parameters λ_i, μ_i of a certain stochastic process, and the probabilities π_i , taken over the set of all sample paths of that process. Those relations could be used to determine the probabilities. Here, on the other hand, we have identities valid for any sample path of any queueing process. The equations (1.71) can also be solved for $p(n)$:

$$p(n) = p(m) \prod_{k=m}^{n-1} \frac{\lambda(k)}{\mu(k+1)}, \quad n = m+1, \dots, M, \quad (1.72)$$

where $p(m)$ is obtained from the normalising equation

$$\sum_{n=m}^M p(n) = 1.$$

The fractions $p(n)$ can thus be determined in terms of the fractions $\lambda(n)$ and $\mu(n)$. The latter are not, however, parameters of the process; they are characteristics of the same sample path for which the former are sought. Knowing the values of $\lambda(n)$ and $\mu(n)$ for one sample path does not help to find the values of $p(n)$ for another sample path.

Suppose now that the sample path $N(t)$ is observed over longer and longer periods of time, and that during those periods it attains wider and

wider ranges of values. In other words, let $T \rightarrow \infty$, $m \rightarrow 0$, $M \rightarrow \infty$. Suppose further, that the limits

$$\lambda_n = \lim_{T \rightarrow \infty} \frac{A(n)}{T(n)}; \quad \mu_n = \lim_{T \rightarrow \infty} \frac{D(n)}{T(n)}; \quad p_n = \lim_{T \rightarrow \infty} \frac{T(n)}{T} \quad (1.73)$$

exist and are non-zero for all $n = 0, 1, \dots$ (except for μ_0). Continuing the analogy with the Birth and Death process, one would naturally expect the fractions p_n to be the unique solution of the infinite system of equations

$$\sum_{n=0}^{\infty} p_n = 1 \quad (1.74)$$

$$\lambda_n p_n = \mu_{n+1} p_{n+1}, \quad n = 0, 1, \dots$$

This is not necessarily the case, as can be seen from the following example.

Consider the sample path illustrated in Fig. 1.7. $N(t)$ goes through alternating busy and idle periods of unit length. During the i -th busy period ($i = 2, 3, \dots$), it spends time $\varepsilon/2^{n-1}$ at level n , $n = 1, 2, \dots, i-1$, and the rest of the time at level i ($0 < \varepsilon < \frac{1}{2}$). It is easily seen that the limits (1.73) for this sample path are

$$\lambda_0 = 1, \quad \lambda_n = \mu_n = 2^{n-1}/\varepsilon, \quad n = 1, 2, \dots$$

$$p_0 = 1/2, \quad p_n = \varepsilon/2^n, \quad n = 1, 2, \dots$$

Equations (1.74), on the other hand, yield

$$p_n = (\varepsilon/2^{n-1})p_0, \quad p_0 = 1/(1 + 2\varepsilon).$$

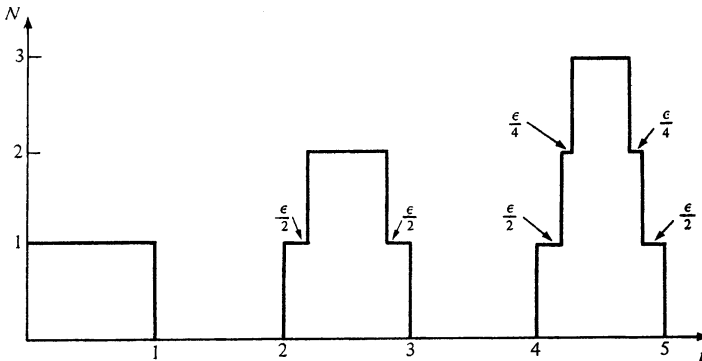


Fig. 1.7.

If we were dealing with a Birth and Death process with the above parameters, then a sample path should spend, in the long run, a fraction $1/(1 + 2\varepsilon)$ of its time in state 0, with probability 1. A pathological sample path like the one in Fig. 1.7 may occur, but the probability of such an event is zero.

1.8. Priority queueing

Let us now move away from the First-In-First-Out scheduling discipline and study some queueing models where the order of service is determined by externally assigned priorities. The customer population is split into a set R of distinct classes, numbered $1, 2, \dots$. That set may be finite or infinite. The class indices are used as priority levels: customers of class i have priority over those of class j if $i < j$.

The models that we shall consider have several common features. In all cases, customers of different classes are assumed to arrive into the system according to independent Poisson streams, with rate λ_i for class i ($i = 1, 2, \dots$). Service is given by a single server of unit speed and within each class customers are served in FIFO order. The server cannot be idle when there are customers in the system. If customers of different classes are waiting for service, the ones with higher priority (lower class index) will be served first.

There are several possibilities concerning the action to be taken when a higher-priority customer arrives to find a lower-priority one in service. In our first model, the new arrival waits until the current service is completed before beginning his own. This is the “non-pre-emptive” or “head-of-the-line” priority discipline (Cobham [4]): after each service completion, the customer with the highest priority among those waiting is selected and served to completion. The service times for class i customers may be generally distributed, with mean $1/\mu_i$ and second moment M_{2i} ($i = 1, 2, \dots$). We shall denote, as usual, the traffic intensity for class i by $\rho_i = \lambda_i/\mu_i$; this is the expected amount of work of class i brought into the system per unit time.

The condition for non-saturation is that the server should be able to cope with the work brought in:

$$\sum_{i \in R} \rho_i < 1.$$

Under that condition, we shall be interested in the steady-state average number of class i customers in the system, N_i , and the average response time for class i , W_i .

It was shown in section 1.6 that the expected number of class i customers in service is ρ_i (this is also the probability that a new arrival finds a class i customer being served). If a class i customer is found in service, his expected remaining service time W_{0i} is given by equation (1.66):

$$W_{0i} = \frac{1}{2}\mu_i M_{2i}; \quad i \in R.$$

Therefore, the overall expected delay W_0 caused by any customer that might be found in service is equal to

$$W_0 = \sum_{i \in R} \frac{1}{2}(\rho_i \mu_i M_{2i}) = \frac{1}{2} \sum_{i \in R} \lambda_i M_{2i}. \quad (1.75)$$

Consider the expected total delay, W_1 , to which a top-priority customer is subjected. Apart from W_0 , that delay comprises the service times of all class 1 customers that our customer finds in the queue (their average number is $N_1 - \rho_1$), plus his own service time. Hence,

$$W_1 = W_0 + (N_1 - \rho_1)/\mu_1 + 1/\mu_1.$$

Substituting, from Little's theorem, $N_1 = \lambda_1 W_1$, and solving for W_1 , we obtain

$$W_1 = 1/\mu_1 + W_0/(1 - \rho_1). \quad (1.76)$$

Let us examine now the total average delay, W_2 , suffered by a class 2 customer. First we make the following remark: suppose that a class 2 customer has to wait for time T (no matter for what reason). All class 1 customers who arrive during T will be served before him. Since class 1 work is brought into the system at rate ρ_1 per unit time, this causes an additional delay of $\rho_1 T$. But all class 1 customers who arrive during that additional delay will be served before our customer, causing a further delay $\rho_1^2 T$, etc. Thus any delay T inflicted on a class 2 customer is stretched to

$$T(1 + \rho_1 + \rho_1^2 + \dots) = T/(1 - \rho_1)$$

due to the continuing arrival of class 1 customers.

On arrival, a class 2 customer is subjected to delays by the customer in service (average of W_0), the class 1 customers in the queue (average of $(N_1 - \rho_1)/\mu_1$) and the class 2 customers in the queue (average of $(N_2 - \rho_2)/\mu_2$). Each of these delays is stretched by a factor of $1/(1 - \rho_1)$ because

of subsequent class 1 arrivals. On top of all that, there is the customer's own service time. The expression for W_2 takes the form

$$W_2 = [W_0 + (N_1 - \rho_1)/\mu_1 + (N_2 - \rho_2)/\mu_2]/(1 - \rho_1) + 1/\mu_2.$$

Substituting $N_1 = \lambda_1 W_1$ (where W_1 is given by (1.76)) and $N_2 = \lambda_2 W_2$, and solving for W_2 yields

$$W_2 = \frac{1}{\mu_2} + \frac{W_0}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}. \quad (1.77)$$

We can now write a similar formula for an arbitrary customer class, j . Note that if customer classes $1, 2, \dots, j - 1$ are lumped together into a single class, H , and are served in FIFO order, this will not affect in any way the customers of class j . Class H would then be the top-priority class and class j the second-priority class. The value of W_0 will remain the same. The traffic intensity for class H , ρ_H is equal to

$$\rho_H = \rho_1 + \rho_2 + \dots + \rho_{j-1}.$$

Applying formula (1.77) to class j gives

$$\begin{aligned} W_j &= \frac{1}{\mu_j} + \frac{W_0}{(1 - \rho_H)(1 - \rho_H - \rho_j)} \\ &= \frac{1}{\mu_j} + \left[\sum_{i \in R} \lambda_i M_{i2} \right] / \left[2 \left(1 - \sum_{i=1}^{j-1} \rho_i \right) \left(1 - \sum_{i=1}^j \rho_i \right) \right], \end{aligned} \quad (1.78)$$

where we have used (1.75). The average number of class j customers in the system is obtained, of course, from Little's theorem: $N_j = \lambda_j W_j$.

It is intuitively clear that, with priority scheduling, higher-priority customers receive better treatment at the expense of lower-priority ones. The above expressions make that intuition quantitative. They also allow one to address various optimisation problems. For instance, given the arrival and service characteristics, and a cost function of the form

$$C = \sum_{i \in R} c_i W_i,$$

how should one assign priorities to classes in order to minimise C ? We shall solve this problem in Chapter 6.

As an application of formulae (1.78), consider the $M/G/1$ system under the Shortest-Processing-Time-first (SPT) scheduling discipline. Service times are assumed to be known in advance and, after each service

completion, the customer with the shortest service time of those waiting is selected and served to completion. Customers arrive in a Poisson stream with rate λ ; the probability distribution function of their service times is $F(x)$.

This model can be reduced to the one with head-of-the-line priorities by introducing an infinity of “artificial” customer classes, using the service time x as a class index (for a rigorous derivation, service times should be first assumed discrete and then a limit taken). Customers of class x arrive at rate $\lambda_x = \lambda dF(x)$; the first and second moments of their service times are, of course, x and x^2 , respectively. The traffic intensity for class x is $\rho_x = \lambda x dF(x)$. Substituting these parameters into (1.78) and replacing the sums by integrals we obtain the conditional expected response time W_x of a customer whose service time is x (Phipps [7]):

$$\begin{aligned} W_x &= x \\ &+ \left[\int_0^\infty \lambda u^2 dF(u) \right] / \left[2 \left(1 - \int_0^{x^-} \lambda u dF(u) \right) \left(1 - \int_0^{x^+} \lambda u dF(u) \right) \right] \\ &= x + (\lambda M_2 / 2) / \left[\left(1 - \int_0^{x^-} \lambda u dF(u) \right) \left(1 - \int_0^{x^+} \lambda u dF(u) \right) \right], \end{aligned} \quad (1.79)$$

where M_2 is the second moment of $F(x)$ and x^- and x^+ denote limits from the left and from the right (if $F(u)$ is continuous at point x , the two are identical). The unconditional expected response time W is given by

$$W = \int_0^\infty W_x dF(x). \quad (1.80)$$

We shall see in Chapter 6 that, of all non-pre-emptive scheduling disciplines, SPT yields the least average response time W .

Let us now return to the priority model with classes $1, 2, \dots$. Suppose that when a higher-priority customer finds a lower-priority one in service, he interrupts the service in progress and starts his own immediately. This is a pre-emptive priority discipline: customers of class j are served only when there are no customers of classes $1, 2, \dots, j - 1$ in the system. To define the discipline completely, one should specify what happens to a pre-empted customer. Does he later continue his service from the point of interruption (pre-emptive-resume discipline), or does he restart the same service from the beginning (pre-emptive-repeat without resampling), or does he request a new independent service (pre-emptive-repeat with

resampling)? To avoid these complications, and to make the analysis easier, we shall assume that class i service times are distributed exponentially with mean $1/\mu_i$ ($i = 1, 2, \dots$). Now, it does not matter which of the above policies is chosen, because of the memoryless property.

Again, we are interested in the expected response time W_j for customers of class j ($j = 1, 2, \dots$). Because priorities are pre-emptive, class j customers are not affected in any way by the existence of classes $j + 1, j + 2, \dots$. In particular, class 1 customers behave as they would in a single-class $M/M/1$ system with parameters λ_1 and μ_1 . Their expected response time is given by expression (1.47):

$$W_1 = \frac{1}{\mu_1(1 - \rho_1)}.$$

Following a similar argument as before, we note that every delay to which a class 2 customer is subjected is stretched by a factor of $1/(1 - \rho_1)$ because of subsequent class 1 arrivals. The delays that should be included in this calculation are due to the class 1 customers he finds in the system (average number N_1 each taking an average of $1/\mu_1$ to serve), the class 2 customers he finds in the system (average N_2/μ_2) and his own service time (average $1/\mu_2$). Hence,

$$W_2 = \frac{(N_1/\mu_1) + (N_2/\mu_2) + (1/\mu_2)}{(1 - \rho_1)}.$$

Substituting $N_1 = \lambda_1 W_1$, $N_2 = \lambda_2 W_2$, using the known expression for W_1 and then solving for W_2 yields

$$W_2 = \frac{1}{\mu_2(1 - \rho_1)} + \frac{(\rho_1/\mu_1) + (\rho_2/\mu_2)}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}.$$

This expression generalises easily to an arbitrary class j :

$$W_j = 1 \left/ \left[\mu_j \left(1 - \sum_{i=1}^{j-1} \rho_i \right) \right] \right. \\ \left. + \left[\sum_{i=1}^j (\rho_i/\mu_i) \right] \left/ \left[\left(1 - \sum_{i=1}^{j-1} \rho_i \right) \left(1 - \sum_{i=1}^j \rho_i \right) \right] \right. . \quad (1.81)$$

Note the similarity between (1.81) and (1.78). The numerator in the second term of (1.81) also represents expected residual service, this time averaged over classes $1, 2, \dots, j$ only.

References

1. Buzen, J. P. (1976). Fundamental operational laws of computer system performance. *Acta Informatica*, **7**, 167–182.
2. Buzen, J. P. (1977). “Operational Analysis: An Alternative to Stochastic Modelling.” Research Report, Harvard University.
3. Cinlar, E. (1954). “Introduction to Stochastic Processes.” Prentice-Hall, Englewood Cliffs, New Jersey.
4. Cobham, A. (1954). Priority assignment in waiting-line problems. *Operations Research*, **9**, 383–387.
5. Foster, F. G. (1972). “Stochastic Processes” Proc. IFORS Conference, Dublin.
6. Little, J. D. C. (1961). A proof for the queueing formula $L = \lambda W$. *Operations Research*, **9**, 383–387.
7. Phipps, T. E. (1961). Machine repair as a priority waiting-line problem. *Operations Research*, **9**, 732–742.
8. Strauch, R. E. (1970). When a queue looks the same to an arriving customer as to an observer. *Man. Sci.*, **17**, 140–141.