

## SEMI-MARKOV PROCESS

*Prediction is difficult, particularly about the future.*

Niels Bohr

(also attributed to Mark Twain,  
but falsely attributed to Yogi Berra).

*We can chart our future clearly and wisely only when  
we know the path which has led to the present.*

Adlai Stevenson.

In many (if not most) real-world applications, the arrival of customers to a service center is not well described by renewal processes. Quite often, the times between successive arrivals are correlated, whereas renewal processes have independent interarrival times. A natural generalization is the class of **semi-Markov processes** (SMP), which when specifically applied to the arrival of customers are called **Markov Renewal Processes** (MRP) or **Markov Arrival Processes** (MAP). Of course arrivals to one station correspond to departures from some other station. So, to avoid confusion, we use the terms SMP or MRP here.

In this chapter we set up a general procedure for creating a sequence of random variables  $\{X_i\}$  which may be thought of as the interarrival times of successive customers for some arrival process. The  $\{X_i\}$  has PDFs  $\{F_i(x_i)\}$  which are generated by a sequence of representations with the same  $\mathbf{B}$  but different entrance vectors, namely  $\{\langle \boldsymbol{\varphi}_i, \mathbf{B} \rangle\}$ . Interval  $i$  we call the  *$i$ th epoch*. Recall that for a renewal process the  $X_i$  r.v.s are iid. In this chapter they are not independent, but they are asymptotically identically distributed. That is, for  $i$  large enough,  $\{F_i(x_i)\}$  approaches a limit. Hence, some researchers retain the word, renewal in describing MRPs.

The set of formulas for the joint interdeparture distributions and correlation lag- $k$  number are then set up, thereby showing that, indeed, the  $X_i$  and  $X_j$  are in most cases, correlated. We then show how they can be used to solve for various performance properties of SMP/M/1 queues. One of the first papers to try to formalize this was by Ramaswami [RAMASWAMI80]. The particular formulation presented here stems from the PhD thesis by Pierre Fiorini [FIORINI98] and other works [FIORINI LIPVDLHSIN95].

## 8.1 Introduction

Because the formulas presented here are abstract, it may be unclear how they are to be applied to specific systems. Therefore, after presenting the general formalism, we supply explicit formulas for a wide variety of processes, each having a different state-space structure from the others. The following processes are considered.

- (1) Departures from a general server that has an infinite queue (renewal process);
- (2) Markov regulated departure processes (direct-sum space);
- (3) Markov Modulated Poisson processes (MMPP);
- (4) *ON-OFF* Models, or the *N-BURST* Process;
- (5) Merging of two renewal streams (direct-product space);
- (6) Departures from overloaded generalized X/G/C queues (reduced-product space);
- (7) Departures from open G/G/1 queues (infinite direct-product space) with reduction to G/M/1 and M/G/1 queues.

The method can be applied to closed (finite number of customers) systems as easily as, or more easily than, to open ones.

Finally, we show (yet again) that departures from an open M/M/1 queue approach Poisson as the system approaches its steady state. The formulas make it very clear that this is a special property that does not carry over to other arrival or service distributions or to closed systems.

### 8.1.1 Matrix Representations of Subsystems

The equations of the previous chapters can be extended to any Markov-like subsystem with a countable state space. Let  $\boldsymbol{\wp}_0$  be the probability vector of the state of the subsystem at the time  $x = 0$  with  $\boldsymbol{\wp}_0 \boldsymbol{\epsilon}' = 1$  ( $\boldsymbol{\epsilon}'$  is the subsystem vector equivalent to  $\boldsymbol{\epsilon}$ ), and  $\mathbf{B}$  is the infinitesimal generator matrix of the process. Then, as in (3.1.7d),  $\boldsymbol{\wp}_0 \exp(-x\mathbf{B})\boldsymbol{\epsilon}'$  has the interpretation of the probability that the process has not ended by time  $x$ . Furthermore, the  $i$ th component of the vector  $\boldsymbol{\wp}_0 \exp(-x\mathbf{B})$  has the following meaning.

*Given that the subsystem was in vector state  $\boldsymbol{\wp}_0$  at time  $x = 0$ ,  $[\boldsymbol{\wp}_0 \exp(-x\mathbf{B})]_i$  is the probability that the process has not yet completed by time  $x$ , and the subsystem is in state  $i$ .*

Usually,  $\mathbf{B}$  can be constructed from the underlying Markov chain using the relation

$$\mathbf{B} = \mathbf{M}(\mathbf{I} - \mathbf{P}), \quad (8.1.1)$$

where  $\mathbf{M}$  is a diagonal matrix whose  $(ii)$ th component is the probability rate of leaving state  $i$ , and  $\mathbf{P}$  is a *substochastic matrix* whose  $ij^{th}$  component is

the probability that the subsystem will transfer to state  $j$  after leaving state  $i$ . At least one of the row sums of  $\mathcal{P}$  (i.e.,  $\mathcal{P}\mathbf{e}'$ ) is strictly less than 1. Thus, there exist state sequences that result in a departure from the subsystem (often visualized as passage to an **absorbing state**). The requirement that  $[\mathcal{I} - \mathcal{P}]$  be invertible is equivalent to there being an exit path from every state.

As we show in the next section, the following theorem about functions of matrices lets us easily calculate many integrals that are otherwise very difficult.

**Theorem 8.1.1:** Let  $\mathcal{B}$  be an invertible finite matrix with  $\mathcal{V} = \mathcal{B}^{-1}$  and eigenvalues  $\beta_1, \beta_2, \dots, \beta_m$ . Furthermore, let  $\Re(\beta_i) > 0$  for all  $i$ . Then

$$\int_0^\infty x^n \exp(-x\mathcal{B})\mathcal{B} dx = n! \mathcal{V}^n. \quad (8.1.2a)$$

(Compare with Theorem 3.1.1.) It is also true that:

$$\begin{aligned} \int_0^\infty e^{-sx} \exp(-x\mathcal{B})\mathcal{B} dx &= \int_0^\infty \exp[-x(s\mathcal{I} + \mathcal{B})]\mathcal{B} dx \\ &= [s\mathcal{I} + \mathcal{B}]^{-1} \mathcal{B} = [\mathcal{I} + s\mathcal{V}]^{-1}. \end{aligned} \quad (8.1.2b)$$

(Again compare with Theorem 3.1.1.) ■

We remind the reader that although the integration is over a scalar  $[x]$ , both sides of these equations are square matrices. With appropriate constraints, this theorem is valid even if  $\mathcal{B}$  is infinite-dimensional.

## 8.2 Markov Renewal Processes

In this section we consider the end of a process to coincide with the departure of a customer from a subsystem. As in previous chapters we refer to the periods between departures as **epochs**. The formulas given in the previous section, with generator  $\langle \mathcal{P}_0, \mathcal{B} \rangle$ , yield the distribution of the departure time of the first customer. We need the following material to describe the departure of the second, and succeeding, customers.

### 8.2.1 Interdeparture Time Distributions

Let  $\{X_n | n \geq 1\}$  be a set of random variables where  $X_n$  denotes the time for the  $n$ th epoch, or interdeparture time of the  $n$ th customer. Consider the following matrix.

#### **Definition 8.2.1**

$[\mathcal{L}]$ : Given that the subsystem is in state  $i$   $[\mathcal{L}]_{ij}\Delta$  is the probability that a departure will occur within the small time interval,  $\Delta$ , and the subsystem will be in state  $j$  immediately afterwards. In other words,  $[\mathcal{L}]_{ij}$  is the subsystem instantaneous departure rate from state  $i$  that leaves behind state  $j$ . □

From this definition, it follows that  $\sum_j \mathcal{L}_{ij} = [\mathcal{L}\boldsymbol{\varepsilon}']_i$  is the subsystem instantaneous departure rate from state  $i$ . But that is what  $[\mathcal{B}\boldsymbol{\varepsilon}']_i$  is. Therefore

$$\mathcal{L}\boldsymbol{\varepsilon}' = \mathcal{B}\boldsymbol{\varepsilon}'. \quad (8.2.1)$$

Although  $\mathcal{L}$  and  $\mathcal{B}$  are related by this relation, they describe different parts of the process of interest.  $\mathcal{B}$  generates what happens during the epoch, and  $\mathcal{L}$  tells what happens immediately after the departure. (8.2.1) states that they agree about the rate of departure.

Given that  $\mathcal{V} := \mathcal{B}^{-1}$ , we have

$$\mathcal{V}\mathcal{L}\boldsymbol{\varepsilon}' = \boldsymbol{\varepsilon}'.$$

Because of its importance, we define

$$\mathcal{Y} := \mathcal{V}\mathcal{L}, \quad \text{with the property } \mathcal{Y}\boldsymbol{\varepsilon}' = \boldsymbol{\varepsilon}' \quad (8.2.2)$$

(i.e.,  $\mathcal{Y}$  is *isometric*;  $\boldsymbol{\varepsilon}'$  is a right eigenvector of  $\mathcal{Y}$  with eigenvalue 1).

We can now observe that the  $j^{\text{th}}$  component of the vector,  $[\boldsymbol{\wp}_0 \exp(-x\mathcal{B})\mathcal{L}]$  is the instantaneous probability rate for service to end at time  $x$ , and for the subsystem to be in state  $j$  immediately after the departure. The sum over all post-departure states must yield the pdf for the process, and indeed it does. After all, from (8.2.1),

$$f_{X_1}(x) := \boldsymbol{\wp}_0 [\exp(-x\mathcal{B})\mathcal{L}]\boldsymbol{\varepsilon}' = \boldsymbol{\wp}_0 [\exp(-x\mathcal{B})\mathcal{B}]\boldsymbol{\varepsilon}' \quad (8.2.3)$$

(compare with Theorem 3.1.1.) Note that although  $f_{X_1}(x)$  is a scalar function of  $x$ , the objects in square brackets are square matrices.

The initial state for the second customer, given that  $X_1 = x$ , is

$$\boldsymbol{\wp}_1(x) = \frac{1}{f_{X_1}(x)} \boldsymbol{\wp}_0 \exp(-x\mathcal{B})\mathcal{L}. \quad (8.2.4)$$

The initial state for the second customer, averaged over all first-process times is given by

$$\begin{aligned} \boldsymbol{\wp}_1 &= \int_0^\infty f_{X_1}(x) \boldsymbol{\wp}_1(x) dx = \int_0^\infty \boldsymbol{\wp}_0 \exp(-x\mathcal{B})\mathcal{L} dx \\ &= \boldsymbol{\wp}_0 \left[ \int_0^\infty \exp(-x\mathcal{B}) dx \right] \mathcal{L} = \boldsymbol{\wp}_0 \mathcal{V}\mathcal{L} = \boldsymbol{\wp}_0 \mathcal{Y}, \end{aligned} \quad (8.2.5)$$

where (8.1.2a) for  $n = 0$  was used. One can immediately generalize that the probability state of the system immediately after the  $n$ th departure (and the starting vector for the  $(n + 1)$ st epoch) is

$$\boldsymbol{\wp}_n = \boldsymbol{\wp}_{n-1} \mathcal{Y} = \boldsymbol{\wp}_0 \mathcal{Y}^n. \quad (8.2.6)$$

Observe that the state the subsystem is in immediately after customer number  $(n-1)$  departs is the beginning state of the subsystem for generating the  $n$ th

departure. We can then say that the (unconditional) distribution function for  $X_n$  is generated by  $\langle \wp_{n-1}, \mathcal{B} \rangle$ .

The steady-state start-up vector must satisfy the equation

$$\wp := \lim_{n \rightarrow \infty} \wp_n = \lim_{n \rightarrow \infty} \wp_o \mathcal{Y}^n = \left( \lim_{n \rightarrow \infty} \wp_{n-1} \right) \mathcal{Y} = \wp \mathcal{Y}; \quad (8.2.7)$$

that is,  $\wp$  must be a left eigenvector of  $\mathcal{Y}$  with eigenvalue 1. That such a vector exists is guaranteed by the fact that  $\mathcal{Y}$  is isometric ( $\mathcal{Y}\epsilon' = \epsilon'$ ). More precisely, this limit exists if 1 is the largest eigenvalue in magnitude of  $\mathcal{Y}$ . In this case,  $\mathcal{Y}^n \rightarrow \epsilon' \wp$ . Then, as  $n$  approaches infinity, the  $X_n$ 's approach the common distribution generated by  $\langle \wp, \mathcal{B} \rangle$ . But they are almost always correlated, as shown in the next subsection.

Although  $\mathcal{L}$  and  $\mathcal{B}$  describe the same departure rate, their difference is also a useful matrix. Define the *generator of the underlying Markov process* as

$$\mathcal{Q} := \mathcal{B} - \mathcal{L}. \quad (8.2.8a)$$

Clearly,  $\mathcal{Q}\epsilon' = 0$  from (8.2.1), so there must exist a vector  $\pi$  such that  $\pi\mathcal{Q} = \mathbf{o}'$  and  $\pi\epsilon' = 1$ . In fact it can be shown by direct substitution that

$$\pi = \frac{\wp \mathcal{V}}{\wp \mathcal{V} \epsilon'}. \quad (8.2.8b)$$

If we multiply  $\pi$  by either  $\mathcal{L}$  or  $\mathcal{B}$ , we can get  $\wp$  in terms of  $\pi$ , that is,

$$(\wp \mathcal{V} \epsilon') \pi \mathcal{L} = \wp \mathcal{V} \mathcal{L} = \wp \mathcal{Y} = \wp$$

and given that  $\wp \epsilon' = 1$ , we have

$$(\wp \mathcal{V} \epsilon') (\pi \mathcal{L} \epsilon') = 1. \quad (8.2.8c)$$

This provides us with an interesting relation.  $(\wp \mathcal{V} \epsilon')$  is the mean interdeparture time, and thus its reciprocal  $(\pi \mathcal{L} \epsilon')$ , must be the long-term departure rate. This makes sense, because  $(\pi)_i$  is the fraction of time the generating system is in state  $i$ , and  $(\mathcal{L} \epsilon')_i$  is the rate of departure when the system is in state  $i$ . Their dot product averages over all states. For emphasis, we restate this now:

$$\mathbf{E}[X] = \wp \mathcal{V} \epsilon' = \text{mean interdeparture time}, \quad (8.2.8d)$$

$$\kappa := \pi \mathcal{L} \epsilon' = \text{long-term departure rate}.$$

In (8.2.8b) we expressed  $\pi$  in terms of  $\wp$ . We now reverse the relation and express  $\wp$  in terms of  $\pi$ . The equations above and the fact that  $\pi \mathcal{B} = \pi \mathcal{L}$  lead to

$$\wp = (\wp \mathcal{V} \epsilon') \pi \mathcal{L} = \frac{\pi \mathcal{L}}{\pi \mathcal{L} \epsilon'} = \frac{\pi \mathcal{B}}{\pi \mathcal{B} \epsilon'}. \quad (8.2.8e)$$

The three matrices,  $\mathcal{Q}$ ,  $\mathcal{L}$ , and  $\mathcal{B}$ , play equally important roles in SM processes, and given any two, the third follows directly. Depending on the

application it may be easier to construct  $\mathbf{Q}$  than one of the other two. We note that  $\mathbf{B}$  always has an inverse (namely,  $\mathbf{V}$ ),  $\mathbf{Q}$  never has an inverse ( $\mathbf{Q}\boldsymbol{\epsilon}' = \mathbf{o}'$  implies that  $\mathbf{Q}$  has a 0 eigenvalue), and  $\mathbf{L}$  may or may not have an inverse (but we wouldn't know what to do with it anyway).

As mentioned earlier,  $\mathbf{B}$  controls the subsystem during an epoch and  $\mathbf{L}$  connects each epoch to the next. We show in some of the applications below that  $\mathbf{Q}$  controls the subsystem irrespective of departures. This is a direct generalization of the discussion surrounding Figure 3.5.3, leading to  $\boldsymbol{\pi}_r$  the mean residual vector.  $\mathbf{Q}$  can also be thought of as the *rate matrix*, or *generator of a continuous Markov chain* as given in the discussion surrounding (1.3.2c) in Chapter 1. So if  $\boldsymbol{\wp}$  describes the state of the subsystem at the beginning of an arbitrary epoch, then  $\boldsymbol{\pi}$  describes the state of the system as seen by a random observer who has no idea when the epoch began. This is discussed further in Section 8.3.1.

Before going on, we describe how our approach differs from that of other researchers. The matrix distribution function  $Q_{ij}(x)$ , as defined in many books (e.g., [COOPER81]), in our case denotes the probability that “a departure will occur by time  $x$  and the system will find itself in state  $j$ , given that the system was in state  $i$  at time  $x = 0$ .” We, on the other hand, use the matrix density function,  $\exp(-x\mathbf{B})\mathbf{L}$ , which when integrated from  $0 \rightarrow x$  yields  $[\mathbf{I} - \exp(-x\mathbf{B})]\mathbf{V}$ , the equivalent to  $Q_{ij}(x)$ , except that, like [NEUTS81], the matrix elements themselves can be matrices. Also, our  $\mathbf{V}$  corresponds to their  $P_{ij} := \lim_{x \rightarrow \infty} Q_{ij}(x)$ . For the applications given here, because of Theorem 8.1.1, the actual values of the components of  $\exp(-x\mathbf{B})$  are not usually needed to get useful results.

### 8.2.2 Correlation of Departures

Based on the material of the previous section, we can write down the joint probability distributions for the interdeparture times. The joint density function for the departure of the first  $n + k$  customers is given by

$$\begin{aligned} & f_{X_1 X_2 \dots X_n \dots X_{n+k}}(x_1, x_2, \dots, x_n, \dots, x_{n+k}) \\ &= \boldsymbol{\wp}_o [\exp(-x_1\mathbf{B})\mathbf{L} \cdots \exp(-x_n\mathbf{B})\mathbf{L} \cdots \exp(-x_{n+k}\mathbf{B})\mathbf{L}]\boldsymbol{\epsilon}'. \end{aligned} \quad (8.2.9)$$

The joint distribution has the appearance of being separable, but the separate epochs are connected by  $\mathbf{L}$ . Only if  $\mathbf{L}$  is of rank 1 (i.e., only if  $\mathbf{L} = \mathbf{B}\boldsymbol{\epsilon}'\boldsymbol{\wp}$ ), are the interdeparture times independent of each other. For instance, let us examine the relation between two variables, say  $X_n$  and  $X_{n+k}$ . To do this, we integrate over all the other variables, and for convenience replace  $x_n$  with  $x$  and  $x_{n+k}$  with  $t$ . Then the joint density function for  $X_n$  and  $X_{n+k}$  is

$$f_{nk}(x, t) := \boldsymbol{\wp}_o [\mathbf{V}^{n-1} \exp(-x\mathbf{B})\mathbf{L} \mathbf{V}^{k-1} \exp(-t\mathbf{B})\mathbf{L}]\boldsymbol{\epsilon}'. \quad (8.2.10a)$$

We can prove the following from this.

**Theorem 8.2.1:** Let  $X_n$  and  $X_{n+k}$  ( $n, k > 0$ ) be random variables denoting the  $n$ th and  $(n + k)$ th interdeparture times. Then  $X_n$  and

$X_{n+k}$  are *independent variables* if and only if  $\mathcal{L}$  is of rank 1, or equivalently,  $\mathcal{Y} = \varepsilon' \wp$ . By *independent* we mean:

$$f_{nk}(x, t) = f_n(x) f_{n+k}(t).$$

Furthermore, except perhaps, for  $n = 1$ , they are identically distributed, and  $\{X_n\}$  is a renewal process. ■

**Proof:** By definition,  $\mathcal{L} = \mathcal{B}\mathcal{Y}$  and if  $\mathcal{L}$  is of rank 1, then  $\mathcal{Y}$  must also be of rank 1. Therefore,  $\mathcal{Y} = \varepsilon' \wp$ . Assume this is so, then  $\mathcal{Y}^n = \varepsilon' \wp$  and from (8.2.10a),

$$\begin{aligned} f_{nk}(x, t) &= \wp_o [\varepsilon' \wp \exp(-x\mathcal{B}) \mathcal{B} \varepsilon' \wp \exp(-t\mathcal{B}) \mathcal{L}] \varepsilon' \\ &= [\wp \exp(-x\mathcal{B}) \mathcal{B} \varepsilon'] [\wp \exp(-t\mathcal{B}) \mathcal{B} \varepsilon'] = f(x) f(t), \end{aligned}$$

where we have used the properties:  $\mathcal{L} \varepsilon' = \mathcal{B} \varepsilon'$  and  $\wp_o \varepsilon' = 1$ .

The converse is more complicated, but note that (8.2.10a) can be written as

$$f_{nk}(x, t) = \mathbf{a}_n(x) \cdot \mathbf{b}'_k(t),$$

where the  $\mathbf{a}_n$  and  $\mathbf{b}'_k$  are vector functions of  $x$  and  $t$ , respectively, namely:

$$\mathbf{a}_n(x) = \wp_o [\mathcal{Y}^{n-1} \exp(-x\mathcal{B}) \mathcal{L}]$$

and

$$\mathbf{b}'_k(t) = [\mathcal{Y}^{k-1} \exp(-t\mathcal{B}) \mathcal{L}] \varepsilon'.$$

$f_{nk}$  is a function of the form:  $a_1(x)b_1(t) + a_2(x)b_2(t) + \dots$ . The only way this can be separated into a single function of  $x$  times a single function of  $t$  is if the  $a_i(x)$ s are proportional to each other. Similarly for the  $b_i(t)$ s. This means that  $\mathbf{a}_n(x)$  equals a scalar function of  $x$  times a constant vector, which in turn forces  $\mathcal{L}$  to be of the form  $\mathbf{b}' \cdot \mathbf{a}$ , that is, a matrix of rank 1. It must follow that  $\mathcal{L} = \mathcal{B} \varepsilon' \wp$ . **QED**

We now move on to get expressions for covariance and autocorrelation coefficients. From its definition, again using (8.1.2a), we can evaluate the mean time for the  $n$ th epoch:

$$\begin{aligned} \mathbb{E}[X_n] &= \int_0^\infty \int_0^\infty x f_{nk}(x, t) dx dt = \wp_o [\mathcal{Y}^{n-1} \mathcal{V}^2 \mathcal{L} \mathcal{Y}^{k-1} \mathcal{V} \mathcal{L}] \varepsilon' \\ &= \wp_o [\mathcal{Y}^{n-1} \mathcal{V} \mathcal{Y}^{k+1}] \varepsilon' = \wp_o [\mathcal{Y}^{n-1} \mathcal{V}] \varepsilon' = \wp_{n-1} [\mathcal{V}] \varepsilon', \end{aligned} \quad (8.2.10b)$$

because  $\mathcal{V} \mathcal{L} = \mathcal{Y}$  and  $\mathcal{Y} \varepsilon' = \varepsilon'$ . Similarly,

$$\begin{aligned} \mathbb{E}[X_{n+k}] &= \wp_o [\mathcal{Y}^{n-1} \mathcal{V} \mathcal{L} \mathcal{Y}^{k-1} \mathcal{V}^2 \mathcal{L}] \varepsilon' \\ &= \wp_o [\mathcal{Y}^{n+k-1} \mathcal{V}] \varepsilon' = \wp_{n+k-1} [\mathcal{V}] \varepsilon'. \end{aligned} \quad (8.2.10c)$$

We see that the mean time for successive epochs is not constant. But, for  $n$  very large [because of (8.2.7)],

$$\lim_{n \rightarrow \infty} \mathbf{E}[X_n] = \lim_{n \rightarrow \infty} \mathbf{E}[X_{n+k}] = \boldsymbol{\wp} [\boldsymbol{\nu}] \boldsymbol{\varepsilon}'. \quad (8.2.10d)$$

The covariance of two random variables is given by

$$\text{Cov}(X, Y) := \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \mathbf{E}[XY] - \mathbf{E}[X] \mathbf{E}[Y]. \quad (8.2.11a)$$

The normalized **correlation coefficient** is defined by:

$$\varrho(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} \quad (8.2.11b)$$

satisfies the inequality,  $-1 \leq \varrho(X, Y) \leq 1$ .

If  $X$  and  $Y$  are two members of a sequence, as is the case here, then  $\text{Cov}(X_n, X_{n+k})$  is called the **autocovariance  $n$  lag- $k$**  of the interdeparture times, and  $\varrho$  is called the **autocorrelation coefficient  $n$  lag- $k$** . The first term on the right of (8.2.11a) evaluates to

$$\mathbf{E}[X_n X_{n+k}] = \int_0^\infty \int_0^\infty x t f_{nk}(x, t) dx dt = \boldsymbol{\wp}_o [\boldsymbol{\nu}^{n-1} \boldsymbol{\nu} \boldsymbol{\nu}^k \boldsymbol{\nu}] \boldsymbol{\varepsilon}', \quad (8.2.11c)$$

giving

$$\text{Cov}(X_n, X_{n+k}) = \boldsymbol{\wp}_{n-1} \boldsymbol{\nu} \boldsymbol{\nu}^k \boldsymbol{\nu} \boldsymbol{\varepsilon}' - (\boldsymbol{\wp}_{n-1} \boldsymbol{\nu} \boldsymbol{\varepsilon}') (\boldsymbol{\wp}_{n+k-1} \boldsymbol{\nu} \boldsymbol{\varepsilon}'). \quad (8.2.11d)$$

It is virtually impossible to measure these parameters. Instead, one must average over  $n$ , which is the same as using  $\boldsymbol{\wp}$  as the initial vector ( $\boldsymbol{\wp}_o \rightarrow \boldsymbol{\wp}$ ), making the covariance independent of  $n$  (but not of  $k$ ). That is, when the subsystem is already in its steady state (or averaged over very large  $n$ ), (8.2.11c) can be written as

$$\begin{aligned} \mathbf{E}(X, X_{+k}) &:= \lim_{n \rightarrow \infty} \mathbf{E}(X_n, X_{n+k}) = \boldsymbol{\wp} [\boldsymbol{\nu} \boldsymbol{\nu}^k \boldsymbol{\nu}] \boldsymbol{\varepsilon}' \\ &= [\boldsymbol{\wp} \boldsymbol{\nu}] \boldsymbol{\nu}^k [\boldsymbol{\nu} \boldsymbol{\varepsilon}'], \end{aligned} \quad (8.2.12a)$$

leading to the **autocovariance lag- $k$** :

$$\text{Cov}(X, X_{+k}) = [\boldsymbol{\wp} \boldsymbol{\nu}] \boldsymbol{\nu}^k [\boldsymbol{\nu} \boldsymbol{\varepsilon}'] - (\boldsymbol{\wp} \boldsymbol{\nu} \boldsymbol{\varepsilon}') (\boldsymbol{\wp} \boldsymbol{\nu} \boldsymbol{\varepsilon}'). \quad (8.2.12b)$$

Thus (8.2.11b) becomes the **autocorrelation coefficient lag- $k$** :

$$\hat{r}(k) := \lim_{n \rightarrow \infty} \varrho(X_n, X_{n+k}) = \frac{\boldsymbol{\wp} [\boldsymbol{\nu} \boldsymbol{\nu}^k - \boldsymbol{\varepsilon}' \boldsymbol{\wp} \boldsymbol{\nu}] \boldsymbol{\varepsilon}'}{2 \boldsymbol{\wp} \boldsymbol{\nu}^2 \boldsymbol{\varepsilon}' - (\boldsymbol{\wp} \boldsymbol{\nu} \boldsymbol{\varepsilon}')^2}. \quad (8.2.12c)$$

In this form, the following theorem is clearly valid.



**Theorem 8.2.2:** If 1 is larger in magnitude than all other eigenvalues of  $\mathcal{Y}$ , then

$$\lim_{k \rightarrow \infty} \mathcal{Y}^k = \boldsymbol{\varepsilon}' \boldsymbol{\varphi}$$

and

$$\lim_{k \rightarrow \infty} \text{Cov}(X, X_{+k}) = \lim_{k \rightarrow \infty} \hat{r}(k) = 0.$$

On the other hand, if  $\mathcal{Y}$  has at least one other eigenvalue of magnitude 1 (e.g., the subsystem is periodic), then the above limits are not valid. (See Example 8.3.1 below for such a case.) But even if all other eigenvalues are less than 1 in magnitude, it is possible for the **autocorrelation lag- $k$**  numbers to be significant for arbitrarily large  $k$ . This will not happen for a finite state-space, but with one caveat. The rate at which  $\hat{r}(k)$  goes to 0 depends on the difference between 1 and the next largest eigenvalue. Therefore, for some systems,  $k$  may have to be very large indeed before the covariance can be considered to be negligible. If the state-space is infinite, and 1 is an accumulation point for the set of eigenvalues (there are an infinite number of eigenvalues arbitrarily close to 1), then one must worry about this point. An important instance of this occurs in telecommunications traffic, where **long-range dependence** or **self-similar traffic** is regularly observed. We present an example of this below, where PT functions are involved. ■

Equations (8.2.12b) and (8.2.12c) can be computed as given, if  $k$  is small enough. But as  $k$  increases it can become numerically unstable. However, it can be evaluated by replacing  $\mathcal{Y}$  with its spectral decomposition over its eigenvalues and eigenvectors. Let  $\mathbf{u}_i$  and  $\mathbf{v}_i'$  be the left and right eigenvectors of  $\mathcal{Y}$  with eigenvalue  $\lambda_i$  such that  $\mathbf{u}_i \mathbf{v}_i' = 1$  (remember that  $\boldsymbol{\varphi}$  and  $\boldsymbol{\varepsilon}'$  are the left and right eigenvectors with eigenvalue 1). Then

$$\mathcal{Y} = \boldsymbol{\varepsilon}' \boldsymbol{\varphi} + \sum_{i^*} \lambda_i \mathbf{v}_i' \mathbf{u}_i,$$

where the  $*$  denotes a sum over all terms excluding eigenvalue 1. Given that  $\mathbf{u}_i \mathbf{v}_j' = \delta_{ij}$ , it follows that

$$\mathcal{Y}^k = \boldsymbol{\varepsilon}' \boldsymbol{\varphi} + \sum_{i^*} \lambda_i^k \mathbf{v}_i' \mathbf{u}_i.$$

Although this is stable, it requires knowing all the eigenvectors and eigenvalues. However, it can also be written in a form where only  $\boldsymbol{\varphi}$  has to be known. Let

$$\bar{\mathcal{Y}} := \mathcal{Y} - \boldsymbol{\varepsilon}' \boldsymbol{\varphi},$$

where  $\bar{\mathcal{Y}}$  has no unit eigenvalues and  $\bar{\mathcal{Y}} \boldsymbol{\varepsilon}' = \mathbf{o}'$ . Then

$$\mathcal{Y}^k - \boldsymbol{\varepsilon}' \boldsymbol{\varphi} = \bar{\mathcal{Y}}^k := (\mathcal{Y} - \boldsymbol{\varepsilon}' \boldsymbol{\varphi})^k$$

and (8.2.12a) can be written as

$$\text{Cov}(X, X_{+k}) = \boldsymbol{\varphi} \left[ \mathcal{Y} \bar{\mathcal{Y}}^k \mathcal{Y} \right] \boldsymbol{\varepsilon}'. \quad (8.2.13a)$$

Either of these two formulas can be used to evaluate  $\text{Cov}(X, X_{+k})$ .

An interesting parameter sometimes evaluated in studying correlations is their sum over all  $k$ . This cannot be done directly, because  $\sum_k \mathbf{Y}^k$  diverges. But when either of the two formulas above is put into (8.2.12a), one gets

$$\begin{aligned} \sum_{k=1}^{\infty} \text{Cov}(X, X_{+k}) &= \sum_{i^*} \frac{\lambda_i}{1 - \lambda_i} [\boldsymbol{\wp} \mathbf{V} \mathbf{v}_i'] [\mathbf{u}_i \mathbf{V} \boldsymbol{\epsilon}'] \\ &= \boldsymbol{\wp} \mathbf{V} \bar{\mathbf{Y}} [\mathbf{I} - \bar{\mathbf{Y}}]^{-1} \mathbf{V} \boldsymbol{\epsilon}'. \end{aligned} \quad (8.2.13b)$$

We should mention that the sum over  $k$  converges only if  $|\lambda_i| < 1$ , for all  $i^*$ .

### 8.2.3 Laplace Transforms

The formulas from the previous section allow us to find an expression for the Laplace transform of the convolution of two (correlated) variables. Let  $T = X_n + X_{n+k}$ , then, using (8.2.10a) the pdf,  $f_T(t)$ , is given by the convolution formula

$$f_T(t) = \int_0^t f_{nk}(x, t-x) dx.$$

Even for renewal processes this is not very easy to do [see Equations (3.5.2)]. However, The Laplace transform can be evaluated in a fashion identical to that used to get (8.2.11c), giving:

$$\begin{aligned} F_T^*(s) &:= \int_0^{\infty} e^{-st} f_T(t) dt \\ &= \boldsymbol{\wp}_0 \left[ \mathbf{Y}^{n-1} [\mathbf{I} + s\mathbf{V}]^{-1} \mathbf{Y}^k [\mathbf{I} + s\mathbf{V}]^{-1} \right] \boldsymbol{\epsilon}'. \end{aligned} \quad (8.2.14)$$

Because of the term  $\mathbf{Y}^k$ , it is clear that the Laplace transform of the distribution of the sum of two random variables is not usually the product of their transforms if they are correlated. Only when  $k$  becomes large enough does  $\mathbf{Y}^k \approx \boldsymbol{\epsilon}' \boldsymbol{\wp}$ , yielding:

$$\lim_{k \rightarrow \infty} F_T^*(s) = \boldsymbol{\wp}_n \left[ [\mathbf{I} + s\mathbf{V}]^{-1} \right] \boldsymbol{\epsilon}' \boldsymbol{\wp} \left[ [\mathbf{I} + s\mathbf{V}]^{-1} \right] \boldsymbol{\epsilon}' = F_{X_n}^*(s) F_X^*(s).$$

## 8.3 Some Examples

To clarify how the equations discussed in the previous sections can be applied to specific systems, we now present several examples of Markov renewal processes, each with a different matrix structure. We start with the simplest case, a renewal process.

### 8.3.1 Departures from Overloaded Server: Renewal Process

As in Chapter 3, consider a server  $S$  that can be represented by the ME pair  $\langle \mathbf{p}, \mathbf{B} \rangle$ . Furthermore, imagine that there is an infinite queue of customers waiting to use  $S$ . Then the state-space of the departure process is the same as the set of phases making up the matrix representation. (This is equivalent to Neuts' infinitesimal generator of PH-renewal processes and forms the substratum of his N-process [RAMASWAMI80]). It follows that

$$\mathbf{B} = \mathbf{B}. \quad (8.3.1a)$$

The  $\mathcal{L}$  matrix can be derived as follows.  $[\mathbf{B}\boldsymbol{\epsilon}']_i$  is the probability rate of leaving  $S$  from phase  $i$ , and  $[\mathbf{p}]_j$  is the probability that the next customer will start in phase  $j$ . Therefore from its definition, we have

$$\mathcal{L} = \mathbf{B}\boldsymbol{\epsilon}'\mathbf{p} = \mathbf{B}\mathbf{Q}, \quad (8.3.1b)$$

where  $\mathbf{Q} = \boldsymbol{\epsilon}'\mathbf{p}$ , with properties given by Lemma 3.5.1. Next,

$$\mathcal{Y} = \mathbf{V}\mathbf{B}\mathbf{Q} = \mathbf{Q}. \quad (8.3.1c)$$

Clearly, this  $\mathcal{Y}$  has the appropriate property that  $\mathcal{Y}\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}'$ , and  $\mathbf{p}$  is the steady-state start-up vector  $\boldsymbol{\rho}$ , satisfying (8.2.7).

From (8.2.8a) we can find  $\mathcal{Q}$  for this example, namely

$$\mathcal{Q} = \mathbf{B} - \mathcal{L} = \mathbf{B} - \mathbf{B}\mathbf{Q},$$

with left eigenvector

$$\boldsymbol{\pi} = \frac{1}{\Psi[\mathbf{V}]} \mathbf{p}\mathbf{V}.$$

This is the **residual vector**,  $\boldsymbol{\pi}_r$ , as given in (3.5.12a), and  $\mathbf{B} - \mathbf{B}\mathbf{Q}$  is the matrix  $\mathbf{B}_r$  as given in the discussion following (3.5.11a). The reader is referred to Section 3.5.3.1 for a full discussion of their meaning.

Following Theorem 8.2.1, we insert the values for  $\mathcal{V}$ ,  $\mathcal{B}$ , and  $\mathcal{L}$  from Equations (8.3.1) into (8.2.10a) and get (for  $n > 1$ )

$$\begin{aligned} f_{nk}(x, t) &= \mathbf{p}[\exp(-x\mathbf{B})\mathbf{B}]\mathbf{Q}[\exp(-t\mathbf{B})\mathbf{B}]\boldsymbol{\epsilon}' \\ &= [\mathbf{p}[\exp(-x\mathbf{B})\mathbf{B}]\boldsymbol{\epsilon}'] [\mathbf{p}[\exp(-t\mathbf{B})\mathbf{B}]\boldsymbol{\epsilon}']. \end{aligned}$$

Using (3.1.7d), we get

$$f_{nk}(x, t) = f(x)f(t). \quad (8.3.2a)$$

This equation is true for all  $n > 1$  and all  $k > 0$ , and is the condition that two random variables be **independent variables**. For  $n = 1$ , the initial  $\mathbf{p}$  is replaced by some initial vector  $\boldsymbol{\rho}_0$ . Then (8.3.2a) becomes

$$f_{1k}(x, t) = f_{X_1}(x)f(t). \quad (8.3.2b)$$

As in Theorem 8.2.1, we see that all  $\{X_i\}$  are mutually independent and (except perhaps for  $X_1$ ), are taken from the same distribution. Therefore,

as discussed in Section 3.5, this is a renewal process if  $\boldsymbol{\varphi}_o = \mathbf{p}$ . Otherwise it is called a **delayed renewal process**, as defined by Feller [FELLER71]. It has also been called a **generalized renewal process**. Needless to say, all autocovariances are equal to 0. Apparently, it is not generally realized (although well known in some circles) that the **counting process** (Definition 3.5.1) associated with any renewal process (with the Poisson process being the lone exception) does have a non-vanishing covariance. See Section 3.5.4.2 for an example that displays this correlation.

### 8.3.2 Markov Modulated (or Regulated) Processes

All the processes in the next few sections involve a **token** that in the course of its actions modulates, or regulates the customer departures. First we describe how the token behaves. Then we discuss several ways that the token can control traffic.

#### 8.3.2.1 The Underlying Generator, $\mathcal{Q}$

Consider a closed system with  $M$  servers,  $\{S_i \mid 1 \leq i \leq M\}$ , each with service time  $T_i$  with distribution represented by  $\langle \mathbf{p}_i, \mathbf{B}_i \rangle$  of dimension  $m_i$ . The representations are assumed to be mutually inequivalent. The token wanders from server to server, spending a time  $T_i$  at  $S_i$  and then with probability  $P_{ij}$  goes to  $S_j$ . The mean time the token spends at  $S_i$  is  $\bar{t}_i := \mathbb{E}[T_i] = \mathbf{p}_i \mathbf{V}_i \boldsymbol{\epsilon}'_i$ .  $\mathbf{P}$  is an  $M$ -dimensional Markov matrix with components  $[\mathbf{P}]_{ij} = P_{ij}$ . That is,  $\mathbf{P}\mathbf{e}' = \mathbf{e}'$ , where  $\mathbf{e}'$  is an  $M$ -dimensional column vector, all of whose components are 1. As with all Markov matrices there is a vector  $\mathbf{p}$  satisfying

$$\mathbf{p}\mathbf{P} = \mathbf{p}, \quad \text{and} \quad \mathbf{p}\mathbf{e}' = 1.$$

Only one server can be active at a time, therefore the set of states needed to describe this system is the union of the sets of states needed to describe each  $S_i$ . The vector space describing the process is the **direct sum** of the individual spaces. So if  $S_i$  is of dimension  $m_i$ , the full space is of dimension

$$M_m := \sum_{i=1}^M m_i.$$

We are dealing here with three levels of matrices. Each server  $S_i$  is described by a set of matrices (e.g.,  $\mathbf{B}_i$ ), the traffic between subsystems is governed by matrices (e.g.,  $\mathbf{P}_{ij}$ ), and the overall system has a matrix description (e.g.,  $\mathbf{P}$ ). We hope to avoid confusion by standardizing our notation with the following definition.

#### **Definition 8.3.1**

Consider an overall system  $\mathcal{S}$ , which itself is made up of subsystems,  $S_i$ . Then matrices and vectors that refer to  $\mathcal{S}$  as a whole are said to operate in **Composite-space**, or simply **C-space**, and are denoted by symbols of the form:

$$\boldsymbol{\varphi}, \boldsymbol{\pi}, \mathbf{B}, \mathbf{I}, \mathbf{L}, \mathbf{P}, \mathbf{Q}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\epsilon}' \quad (\text{bold-faced CALLIGRAPHIC}).$$

Such matrices formally have dimension  $M$ , with components  $[\mathbf{W}]_{ij}$ , where  $1 \leq i, j \leq M$ . However, each  $[\mathbf{W}]_{ij}$  is itself a matrix of dimension  $m_i \times m_j$ . Therefore,  $\mathbf{W}$  is really of dimension  $M_m \times M_m$ .

Matrices and vectors describing the individual subsystems  $S_i$  are denoted by symbols of the form:

$$\mathbf{p}_i, \mathbf{B}_i, \mathbf{I}_i, \mathbf{L}_i, \mathbf{P}_i, \mathbf{Q}_i, \mathbf{V}_i, \mathbf{Y}_i, \boldsymbol{\epsilon}'_i \text{ (bold-faced Roman).}$$

These matrices have dimension  $m_i$ , with components  $[\mathbf{W}_i]_{kl}$ , where  $1 \leq i \leq M$ , and  $1 \leq k, l \leq m_i$ .

Matrices and vectors that refer to transitions between the  $\{S_i\}$  are called *interserver operators*, and operate in  *$\mathbf{I}$ -space*. They are denoted by symbols of the form:

$$\mathbf{a}, \mathbf{p}, \mathbf{B}, \mathbf{I}, \mathbf{P}, \mathbf{V}, \mathbf{Q}, \mathbf{e}' \text{ (bold-faced Italic).}$$

These matrices are of dimension  $M$ , with components  $[\mathbf{W}]_{ij} = W_{ij}$ , where  $1 \leq i, j \leq M$ . In particular the transition matrix  $P_{ij}$ , referred to at the beginning of this section, is an element of  $\mathbf{P}$ ; that is,  $P_{ij} = [\mathbf{P}]_{ij}$ .

If  $m_i = 1$ , for all  $i$ , then  $\mathcal{S}$  reduces to an exponential network, and  $\mathcal{C}$ -space collapses to  $\mathbf{I}$ -space ( $M_m = M$ ).  $\square$

We find it useful in this section (as well as in Section 9.3) to use the following notation. Each  $S_i$  has its characteristic matrices, and often they appear as diagonal elements in the full space. We use the subscript “o” to denote such matrices. For instance:

$$\mathbf{B}_o := \begin{bmatrix} \mathbf{B}_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{B}_2 & \cdots & \mathbf{O} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{O} & \mathbf{O} & \cdots & \mathbf{B}_M \end{bmatrix}, \quad (8.3.3a)$$

with inverse

$$\mathbf{V}_o = \mathbf{B}_o^{-1} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{V}_2 & \cdots & \mathbf{O} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{O} & \mathbf{O} & \cdots & \mathbf{V}_M \end{bmatrix}. \quad (8.3.3b)$$

We also use the following notation

$$\mathcal{M}_o = \text{Diag}[\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_M]$$

to denote matrices of diagonal form. The different objects satisfy the rules of Definition 8.3.1, namely, the  $(ij)$ th element of  $\mathbf{B}_o$  is itself a matrix of dimension  $m_i \times m_j$ , where  $m_i$  is the dimension of the representation of  $S_i$ .

Given that the token wanders forever from server to server, it is governed by the same matrix described in Section 1.3.1, specifically, Equation (1.3.2c), except that there the every  $S_i$  was exponential. The token's position in time is governed by the rate matrix  $\mathbf{Q}$ , satisfying  $\mathbf{Q} = \mathcal{M}(\mathcal{I} - \mathbf{P})$ . We should

be able to construct it in a straightforward manner. Clearly,  $\mathcal{M} = \mathcal{M}_o = \text{Diag}[\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_M]$ . The identity matrix  $\mathcal{I}$  is also of this form with identity matrices  $\mathbf{I}_i$  of dimension  $m_i$  on the diagonal. The transition matrix can be seen to be

$$\mathcal{P} = \mathcal{P}_o + \begin{bmatrix} \mathbf{q}'_1 P_{11} \mathbf{p}_1 & \mathbf{q}'_1 P_{12} \mathbf{p}_2 & \cdots & \mathbf{q}'_1 P_{1M} \mathbf{p}_M \\ \mathbf{q}'_2 P_{21} \mathbf{p}_1 & \mathbf{q}'_2 P_{22} \mathbf{p}_2 & \cdots & \mathbf{q}'_2 P_{2M} \mathbf{p}_M \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{q}'_M P_{M1} \mathbf{p}_1 & \mathbf{q}'_M P_{M2} \mathbf{p}_2 & \cdots & \mathbf{q}'_M P_{MM} \mathbf{p}_M \end{bmatrix}. \quad (8.3.3c)$$

Consider a typical term  $(\mathcal{P})_{ij} = \mathbf{P}_i \delta_{ij} + \mathbf{q}'_i P_{ij} \mathbf{p}_j$ . Say the token completes service in phase  $k$  of server  $S_i$ . He then either:

- (1) Stays in  $S_i$   $[\delta_{ij}]$  and goes to phase  $l$ ,  $[(\mathbf{P}_i)_{kl}]$ , or
- (2) Leaves  $S_i$   $[(\mathbf{q}'_i)_k]$ , goes to  $S_j$   $[P_{ij}]$ , enters, and goes to phase  $l$ ,  $[(\mathbf{p}_j)_l]$ .

By thinking of  $\mathcal{E}'$  as the transpose of  $\mathcal{E} = [\epsilon_1, \epsilon_2, \dots, \epsilon_M]$ , it is easy to show that  $\mathcal{P}\mathcal{E}' = \mathcal{E}'$  when  $\mathcal{P}\mathcal{E}' = \mathcal{E}'$  even though  $\mathbf{P}_i \epsilon'_i \neq \epsilon'_i$ .

Our next task is to find  $\mathcal{Q} = \mathcal{M}(\mathcal{I} - \mathcal{P})$ . The above equations, together with the properties  $\mathbf{B}_i = \mathbf{M}_i(\mathbf{I}_i - \mathbf{P}_i)$  and  $\mathbf{M}_i \mathbf{q}'_i = \mathbf{B}_i \epsilon'_i$  yield

$$\begin{aligned} \mathcal{Q} &= \mathcal{M} - \mathcal{M}\mathcal{P}_o - \begin{bmatrix} \mathbf{M}_1 \mathbf{q}'_1 P_{11} \mathbf{p}_1 & \mathbf{M}_1 \mathbf{q}'_1 P_{12} \mathbf{p}_2 & \cdots & \mathbf{M}_1 \mathbf{q}'_1 P_{1M} \mathbf{p}_M \\ \mathbf{M}_2 \mathbf{q}'_2 P_{21} \mathbf{p}_1 & \mathbf{M}_2 \mathbf{q}'_2 P_{22} \mathbf{p}_2 & \cdots & \mathbf{M}_2 \mathbf{q}'_2 P_{2M} \mathbf{p}_M \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{M}_M \mathbf{q}'_M P_{M1} \mathbf{p}_1 & \mathbf{M}_M \mathbf{q}'_M P_{M2} \mathbf{p}_2 & \cdots & \mathbf{M}_M \mathbf{q}'_M P_{MM} \mathbf{p}_M \end{bmatrix} \\ &= \mathcal{B}_o - \begin{bmatrix} \mathbf{B}_1 \epsilon'_1 P_{11} \mathbf{p}_1 & \mathbf{B}_1 \epsilon'_1 P_{12} \mathbf{p}_2 & \cdots & \mathbf{B}_1 \epsilon'_1 P_{1M} \mathbf{p}_M \\ \mathbf{B}_2 \epsilon'_2 P_{21} \mathbf{p}_1 & \mathbf{B}_2 \epsilon'_2 P_{22} \mathbf{p}_2 & \cdots & \mathbf{B}_2 \epsilon'_2 P_{2M} \mathbf{p}_M \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{B}_M \epsilon'_M P_{M1} \mathbf{p}_1 & \mathbf{B}_M \epsilon'_M P_{M2} \mathbf{p}_2 & \cdots & \mathbf{B}_M \epsilon'_M P_{MM} \mathbf{p}_M \end{bmatrix}, \end{aligned}$$

or

$$\mathcal{Q} = \mathcal{B}_o - \mathcal{B}_o \langle \mathcal{P} \rangle = \mathcal{B}_o [\mathcal{I} - \langle \mathcal{P} \rangle], \quad (8.3.3d)$$

where we have introduced a new *embedding operation*.

### Definition 8.3.2

Let  $\mathbf{W}$  be any  $M \times M$  matrix with components  $[\mathbf{W}]_{ij} = W_{ij}$ . This can be embedded into the full  $M_n \times M_n$  space of  $\mathcal{S}$  in the following way.

$$\langle \mathbf{W} \rangle := \begin{bmatrix} W_{11} \epsilon'_1 \mathbf{p}_1 & W_{12} \epsilon'_1 \mathbf{p}_2 & \cdots & W_{1M} \epsilon'_1 \mathbf{p}_M \\ W_{21} \epsilon'_2 \mathbf{p}_1 & W_{22} \epsilon'_2 \mathbf{p}_2 & \cdots & W_{2M} \epsilon'_2 \mathbf{p}_M \\ \cdots & \cdots & \cdots & \cdots \\ W_{M1} \epsilon'_M \mathbf{p}_1 & W_{M2} \epsilon'_M \mathbf{p}_2 & \cdots & W_{MM} \epsilon'_M \mathbf{p}_M \end{bmatrix}. \quad (8.3.4a)$$

Let  $\mathbf{a}$  be any  $M$ -dimensional row vector with components  $[\mathbf{a}]_i = a_i$ ; then the  $\mathcal{C}$ -space,  $M_n$  row vector is:

$$\langle \mathbf{a} \rangle := [a_1 \mathbf{p}_1, a_2 \mathbf{p}_2, \dots, a_M \mathbf{p}_M]. \quad (8.3.4b)$$

Let  $\mathbf{b}'$  be any  $M$ -dimensional column vector with components  $[\mathbf{b}']_i = b_i$ ; then the  $\mathcal{C}$ -space  $M_n$  column vector is:

$$|\mathbf{b}'\rangle := [b_1 \epsilon'_1, b_2 \epsilon'_2, \dots, b_M \epsilon'_M]. \quad (8.3.4c)$$

These operators can be very useful when dealing with networks of non-exponential servers. For instance,

$$\epsilon' = [\epsilon'_1, \epsilon'_2, \dots, \epsilon'_M] = |\mathbf{e}'\rangle,$$

and suppose

$$\wp = \langle \mathbf{p} | := [p_1 \mathbf{p}_1, p_2 \mathbf{p}_2, \dots, p_M \mathbf{p}_M].$$

Let  $\mathbf{W}$  be any matrix in  $\mathcal{C}$ -space, then

$$\wp \mathbf{W} \epsilon' = \langle \mathbf{p} | \mathbf{W} | \mathbf{e}' \rangle = \mathbf{p} \mathbf{W} \mathbf{e}'.$$

This algebra is discussed in full in Section 9.3.  $\square$

We now examine several ways in which the token can regulate customer traffic.

### 8.3.2.2 Markov Regulated Departure Process (MRDP)

We define a *Markov Regulated Departure Process* (MRDP) as one in which a customer departs every time the token leaves a server. It follows that the time for the  $i$ th epoch is determined by where the token is after customer  $i-1$  leaves. We have already assumed that  $P_{ij}$  is a Markov matrix, but if all its rows are equal ( $P_{ij} = P_{kj} = p_j$  for all  $i, j, k$ ; i.e.,  $\mathbf{P} = \mathbf{e}' \mathbf{p}$ ) then the process reduces to the renewal process of Section 8.3.1.

An alternate but equivalent picture is of an infinite queue feeding into a network with  $M$  servers  $\{S_i | 1 \leq i \leq M\}$ , each with service time  $T_i$  from the distribution represented by  $\langle \mathbf{p}_i, \mathbf{B}_i \rangle$  of dimension  $m_i$ . The customers enter, one at a time. When a customer departs from  $S_i$  he leaves the network and the next customer goes to  $S_j$  with probability  $P_{ij}$ .

For MRDPs  $\mathcal{B}$  is easy to express, because the time between customer departures is the same as the time the token spends at  $S_i$ . Therefore,

$$\mathcal{B} = \mathcal{B}_o \quad \text{and} \quad \mathcal{V} = \mathcal{V}_o. \quad (8.3.5a)$$

Now that we have shown from (8.3.3d) that  $\mathcal{Q} = \mathcal{B}_o - \mathcal{B}_o \langle \mathbf{P} \rangle$ ,  $\mathcal{L}$  and  $\mathcal{Y}$  follow directly:

$$\mathcal{L} = \mathcal{B} - \mathcal{Q} = \mathcal{B}_o \langle \mathbf{P} \rangle \quad (8.3.5b)$$

and

$$\mathcal{Y} = \mathcal{V} \mathcal{L} = \mathcal{V}_o \mathcal{B}_o \langle \mathbf{P} \rangle = \langle \mathbf{P} \rangle. \quad (8.3.5c)$$

Because  $\sum_{j=1}^M P_{ij} = 1$  for all  $i$  ( $\mathbf{P} \mathbf{e}' = \mathbf{e}'$ ), it follows that  $\mathcal{Y} \epsilon' = \epsilon'$ . It is not hard to show that

$$\mathcal{Y}^k = \langle \mathbf{P}^k \rangle,$$

which can be useful in calculating autocorrelation lag- $k$ , or  $\mathbf{E}[X_k]$ . Note also that if  $M = 1$  (only one server) then  $\langle \mathbf{P} \rangle$  reduces to  $\mathbf{Q} = \mathbf{e}' \mathbf{p}$ .

The steady-state vector satisfying  $\wp \mathcal{V} = \wp$  is (see Definition 8.3.2):

$$\wp = \langle \mathbf{p} | := [p_1 \mathbf{p}_1, p_2 \mathbf{p}_2, \dots, p_M \mathbf{p}_M], \quad (8.3.6a)$$

where  $p_i$  is the  $i$ th component of the left eigenvector of  $\mathbf{P}$  with eigenvalue 1 (i.e.,  $\mathbf{p} \mathbf{P} = \mathbf{p}$ ). (Note that  $p_i$  is a component of the  $M$ -vector,  $\mathbf{p}$  corresponding to the steady-state probability that the token will be found at  $S_i$ , and  $\mathbf{p}_i$  is the  $m_i$ -vector whose  $k$ th component  $[\mathbf{p}_i]_k$  is the probability that the token, upon entering  $S_i$ , will go to phase  $k$ .)

In anticipation of its usefulness later, we introduce the  $M$ -dimensional,  $I$ -space matrix,

$$\mathbf{V}_o := \text{Diag} [\bar{t}_1, \bar{t}_2, \dots, \bar{t}_M],$$

where  $\bar{t}_j = \mathbf{p}_j \mathbf{V}_j \mathbf{e}'_j$  is the mean service time of  $S_j$ . It comes from

$$\wp \mathcal{V}_o \mathbf{e}' = \langle \mathbf{p} | \mathcal{V}_o | \mathbf{e}' \rangle = \mathbf{p} \mathbf{V}_o \mathbf{e}'.$$

We can now get  $\pi$  directly from (8.2.8b); that is,

$$\begin{aligned} \pi &= \frac{\wp \mathcal{V}}{\wp \mathcal{V} \mathbf{e}'} = \frac{1}{\wp \mathcal{V}_o \mathbf{e}'} [p_1 \mathbf{p}_1 \mathbf{V}_1, p_2 \mathbf{p}_2 \mathbf{V}_2, \dots, p_M \mathbf{p}_M \mathbf{V}_M] \\ &=: \frac{1}{\mathbf{p} \mathbf{V}_o \mathbf{e}'} \langle \mathbf{p} | \mathcal{V}_o. \end{aligned} \quad (8.3.6b)$$

Given that  $\pi \mathbf{e}' = 1$ , it follows that  $\mathbf{E}[X] = \mathbf{p} \mathbf{V}_o \mathbf{e}' = \sum p_i \bar{t}_i$ .

Let the initial vector be written in the form

$$\wp_o = [a_1 \mathbf{w}_1, a_2 \mathbf{w}_2, \dots, a_M \mathbf{w}_M], \quad \text{with } \wp_o \mathbf{e}' = 1, \quad (8.3.7)$$

where  $\mathbf{a}$  is an  $M$ -vector such that  $\mathbf{a} \mathbf{e}' = \sum_{i=1}^M a_i = 1$  and  $\mathbf{w}_i \mathbf{e}'_i = 1$ . That is, the first customer is initially found at  $S_i$  with probability  $a_i$ , and in vector state  $\mathbf{w}_i$ .

We defer actual derivation of these formulas to Section 9.3, but it follows from (8.2.10b) that

$$\mathbf{E}[X_n] = \mathbf{a} \mathbf{P}^{n-1} \mathbf{V}_o \mathbf{e}' = \sum_{i,j}^M a_i (\mathbf{P}^{n-1})_{ij} \bar{t}_j \quad \text{for } n > 1, \quad (8.3.8a)$$

but

$$\mathbf{E}[X_1] = \sum_{i,j}^M a_i [\mathbf{w}_i \mathbf{V}_i \mathbf{e}'_i]. \quad (8.3.8b)$$

If  $n$  is very large, or the system started in its steady state ( $a_j \rightarrow p_j$  and  $\mathbf{w}_j \rightarrow \mathbf{p}_j$ ), then the mean interdeparture time becomes

$$\mathbf{E}[X] = \lim_{n \rightarrow \infty} \mathbf{E}[X_n] = \lim_{n \rightarrow \infty} \wp_o \mathcal{V}^n \mathcal{V} \mathbf{e}' = \wp \mathcal{V} \mathbf{e}' = \mathbf{p} \mathbf{V}_o \mathbf{e}' = \sum_{i=1}^M p_i \bar{t}_i. \quad (8.3.8c)$$



Recall that  $\mathbf{E}[T_i^\ell] = \ell! \mathbf{p}_i \mathbf{V}_i^\ell \mathbf{e}' = \Psi_i[\mathbf{V}_i^\ell]$ . Then

$$\mathbf{E}[X^2] = 2 \boldsymbol{\rho} \mathbf{V}^2 \mathbf{e}' = 2 \sum p_i \Psi_i[\mathbf{V}_i^2] = 2 \mathbf{p} \mathbf{V}_0^{(2)} \mathbf{e}' = \sum p_i \mathbf{E}[T_i^2],$$

where

$$\begin{aligned} \mathbf{V}_0^{(2)} &:= \text{Diag}[\Psi_1[\mathbf{V}_1^2], \Psi_2[\mathbf{V}_2^2], \dots, \Psi_M[\mathbf{V}_M^2]] \\ &= \frac{1}{2} \text{Diag}[\mathbf{E}[T_1^2], \mathbf{E}[T_2^2], \dots, \mathbf{E}[T_M^2]]. \end{aligned}$$

Note that  $\mathbf{V}_0^{(\ell)} \neq \mathbf{V}_0^\ell$  unless all  $S_i$  are exponential.

The specific form for (8.2.12a) in this case is

$$\text{Cov}(X, X_{+k}) = \mathbf{p} \mathbf{V}_0 [\mathbf{P}^k - \mathbf{e}' \mathbf{p}] \mathbf{V}_0 \mathbf{e}' = \sum_{i,j} p_i \bar{t}_i [\mathbf{P}^k - \mathbf{e}' \mathbf{p}]_{ij} \bar{t}_j \quad (8.3.9a)$$

with interdeparture density

$$f(t) = \sum_{i=1}^M p_i f_i(t). \quad (8.3.9b)$$

Some of the properties of these equations can best be seen by examining a particular subsystem. The steady-state interdeparture density for a subsystem with two servers ( $M = 2$ ) follows.

**Example 8.3.1:** First, from  $\mathbf{p} \mathbf{P} = \mathbf{p}$ , it is seen that  $p_1 = P_{21}/(P_{12} + P_{21})$  and  $p_2 = 1 - p_1 = P_{12}/(P_{12} + P_{21})$ , so (8.3.9b) becomes

$$f(t) = \frac{P_{21} f_1(t) + P_{12} f_2(t)}{P_{12} + P_{21}},$$

and from (8.3.8c), the mean interdeparture time is

$$\mathbf{E}[X] = \frac{P_{21} \bar{t}_1 + P_{12} \bar{t}_2}{P_{12} + P_{21}}.$$

Using  $\mathbf{E}[X^\ell] = p_1 \mathbf{E}[T_1^\ell] + p_2 \mathbf{E}[T_2^\ell]$  and  $\sigma^2 = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$ , it follows that

$$\sigma^2 = p_1 \sigma_1^2 + p_2 \sigma_2^2 + p_1 p_2 (\bar{t}_1 - \bar{t}_2)^2.$$

From (8.3.9a), the steady-state covariance lag- $k$   $[(1 - P_{12} - P_{21})$  is the other eigenvalue of  $\mathbf{P}$ ] becomes

$$\text{Cov}(X, X_{+k}) = P_{12} P_{21} (1 - P_{12} - P_{21})^k \left( \frac{\bar{t}_1 - \bar{t}_2}{P_{12} + P_{21}} \right)^2,$$

and from (8.2.13b),

$$\sum_{k=1}^{\infty} \text{Cov}(X, X_{+k}) = \frac{P_{12} P_{21} (1 - P_{12} - P_{21})}{P_{12} + P_{21}} \left( \frac{\bar{t}_1 - \bar{t}_2}{P_{12} + P_{21}} \right)^2.$$

It is clear that if the  $\bar{t}_i$ s are equal, then all covariances are 0. All covariances are also 0, if  $P_{12} = P_{22}$ . For then  $P_{12} + P_{21} = 1$  and  $\mathbf{P} = \mathbf{e}'\mathbf{p}$ . That is, what happens in each epoch is independent of what happened in the previous epoch.

On the other hand, if  $P_{12} = P_{21} = 1$  ( $\mathbf{P}$  is cyclic), then

$$\text{Cov}(X, X_{+k}) = (-1)^k \left( \frac{\bar{t}_1 - \bar{t}_2}{2} \right)^2.$$

In this case, the limit as  $k \rightarrow \infty$  does not exist, and the sum over  $k$  does not converge! [See Theorem 8.2.2.]

For a last word we look at the autocorrelations,  $\hat{r}(k) = \text{Cov}(X, X_{+k})/\sigma^2$ . They depend on  $S_1$  and  $S_2$  only through their variances. The bigger  $\sigma_1^2$  and  $\sigma_2^2$  are, the smaller is  $\hat{r}(k)$ . In the other direction, if the two distributions are deterministic, their variances equal 0 and

$$\hat{r}(k) = (1 - P_{12} - P_{21})^k.$$

The dependence on the distributions is completely gone; and all that remains is a “coin-flipping” game. The ***Bernoulli process*** corresponds to  $P_{12} = P_{21}$  with  $\hat{r}(k) = 0$ . The probability of flipping a 1 is  $P_{11} = P_{21}$ . If  $P_{12}$  does not equal  $P_{21}$ , the game is biased in that the probability of a 1 depends on the result of the previous flip. ▲

What is most interesting about these processes is that their mean epoch times ( $\mathbf{E}[X_n]$ ) and correlations depend only on the means ( $\bar{t}_i$ ) of the different distributions, and not the distributions or even the higher moments. Thus, even two exponential servers regulated this way will produce a non-renewal process.

### 8.3.2.3 Markov Modulated Poisson Process (MMPP)

The most widely used SMPs are MMPPs. In particular, they have been used to model voice traffic, and recently, all telecommunications traffic (see, for instance, [MEIER-FISCHER92] and [PARK-WILL00]). In the previous section we defined the MRDP, where a “token” wanders from one server to another, spends a time  $T_j$  at  $S_j$  with distribution generated by  $\langle \mathbf{p}_j, \mathbf{B}_j \rangle$ , at which time one customer departs the system, and the token moves to another server according to the matrix  $\mathbf{P}$  (8.3.5c). In this section the token still wanders from  $S_i$  to  $S_j$ , but now customers depart continuously at a Poisson rate of  $\lambda_j$  while the token is at  $S_j$ . That is, the time between departures is exponentially distributed, with mean  $1/\lambda_j$ , and on average,  $\lambda_j \mathbf{E}[T_j]$  customers leave while the token is at  $j$ . Thus the token *modulates the rate* at which customers depart by moving from one station to another.

Most applications of MMPPs assume that each  $T_j$  is exponentially distributed. But here we assume that they are as described in Section 8.3.2.1 and have nonexponential distributions. Therefore, the token’s behavior is governed by the  $\mathbf{Q}$  of (8.3.3d). When viewed as an  $M$ -dimensional system, the

time the token spends at  $S_i$  is indeed nonexponential. But if one looks at  $\mathcal{Q}$  as an  $M_m$ -dimensional system, where each phase is thought of as a server, then the structure is again that of exponential servers. From a modeling point of view (that's what is important) we have a generalization of MMPPs. But from a purely mathematical view, this is still an MMPP (and a restricted one at that).

From its description,  $\mathcal{L}$  is easy to write down, being:

$$\mathcal{L} = \mathcal{L}_o := \begin{bmatrix} \lambda_1 \mathbf{I}_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \lambda_2 \mathbf{I}_2 & \cdots & \mathbf{O} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{O} & \mathbf{O} & \cdots & \lambda_M \mathbf{I}_M \end{bmatrix}. \quad (8.3.10a)$$

Note that adding a term of the form  $\lambda_b \mathcal{I}$  to  $\mathcal{L}$ , doesn't change its structure and doesn't change  $\mathcal{Q}$  either. But this can then be interpreted either as increasing the rate at each server, or as an MMPP with a background (or merged with a) Poisson process of rate  $\lambda_b$ . We also have occasion to use the  $M$ -dimensional matrix  $\mathcal{L}_o := \text{Diag}[\lambda_1, \lambda_2, \dots, \lambda_M]$ .

We have another notational point to make. The matrices,  $\mathcal{B}$ ,  $\mathcal{V}$ ,  $\mathcal{L}$ , and  $\mathcal{Q}$  have the same physical meaning from application to application, but they may have completely different structures. On the other hand, the matrices with subscript "o" (e.g.,  $\mathcal{L}_o$ ) are always block diagonal matrices, and may have no physical meaning in any particular application. For a summary of useful matrices, see Table 8.3.1 below.

In discussing overloaded servers in Section 8.3.1, we easily set up  $\mathcal{B}$  and  $\mathcal{L}$ , and thereby were able to get  $\mathcal{Q}$ . In Section 8.3.2.2, for MRDPs we first set up  $\mathcal{B}$  and  $\mathcal{Q}$ , from which  $\mathcal{L}$  followed. Here we have set up  $\mathcal{Q}$  and  $\mathcal{L}$  for the MMPP and now have

$$\mathcal{B} = \mathcal{L} + \mathcal{Q} = \mathcal{L}_o + \mathcal{B}_o - \mathcal{B}_o \langle P \rangle, \quad (8.3.10b)$$

where  $\mathcal{Q}$  is from (8.3.3d) and  $\mathcal{B}_o$  is from (8.3.3a).

The inverse of  $\mathcal{B}$  can be found using a technique similar to that used in Lemma 4.2.1. We present the result here, and those who wish to know more about the algebra of  $\mathcal{C}$  embeddings are referred to Section 9.3. First manipulate (8.3.10b), recalling that both  $\mathcal{V}_o$  and  $\mathcal{L}_o$  are block diagonal and commute with each other, but not with  $\langle P \rangle$ . That is,  $\mathcal{L}_o \mathcal{V}_o = \mathcal{V}_o \mathcal{L}_o$  but  $\mathcal{L}_o \langle P \rangle \neq \langle P \rangle \mathcal{L}_o$ . We get

$$\mathcal{V} = \mathcal{B}^{-1} = [\mathcal{I} - \mathcal{D}_o \langle P \rangle]^{-1} \mathcal{V}_o \mathcal{D}_o, \quad (8.3.10c)$$

where

$$\mathcal{D}_o := [\mathcal{I} + \mathcal{L}_o \mathcal{V}_o]^{-1} = \text{Diag}[\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_M]$$

with  $\mathbf{D}_i = [\mathbf{I}_i + \lambda_i \mathbf{V}_i]^{-1}$ . Furthermore,

$$\mathbf{D}_o := \text{Diag}[d_1, d_2, \dots, d_M], \quad \text{where } d_i = \mathbf{p}_i \mathbf{D}_i \mathbf{e}_i'.$$

We next make use of the special properties of  $\langle \mathbf{P} \rangle$  to take the inverse of an  $M_m \times M_m$  matrix by embedding the inverse of an  $M \times M$  matrix:

$$\left[ \mathcal{I} - \mathcal{D}_o \langle \mathbf{P} \rangle \right]^{-1} = \mathcal{I} + \mathcal{D}_o \langle (\mathcal{I} - \mathcal{P} \mathcal{D}_o)^{-1} \mathbf{P} \rangle$$

and put this into (8.3.10c) to get

$$\mathbf{v} = \left[ \mathcal{I} + \mathcal{D}_o \langle (\mathcal{I} - \mathcal{P} \mathcal{D}_o)^{-1} \mathbf{P} \rangle \right] \mathbf{v}_o \mathcal{D}_o. \quad (8.3.10d)$$

As was shown in Theorem 3.1.1,  $d_i = B^*(\lambda_i)$ , the Laplace transform of the distribution generated by  $\langle \mathbf{p}_i, \mathbf{B}_i \rangle$  and can be interpreted as the probability that the token will leave  $S_i$  before any customers depart. We take a closer look at this in the next section.

The matrix  $\mathcal{Y}$  comes easily:

$$\mathcal{Y} = \mathcal{V} \mathcal{L} = \left[ \mathcal{I} + \mathcal{D}_o \langle (\mathcal{I} - \mathcal{P} \mathcal{D}_o)^{-1} \mathbf{P} \rangle \right] \mathbf{v}_o \mathcal{D}_o \mathcal{L}_o. \quad (8.3.10e)$$

It is useful at times to use the identity  $\mathcal{D}_o \mathbf{v}_o \mathcal{L}_o = \mathcal{I} - \mathcal{D}_o$ , while noting that the three matrices commute with each other.

There are three advantages of using this notation. First, the internal structure of the matrices is more explicit (once one gets used to the notation). Second, the inverses of matrices are found by inverting matrices of smaller dimension. Third, one can go further by analytic manipulation rather than having to resort to numerical computation. The manipulations can occur without having to resort to a particular distribution representation; that is, the expressions are valid for all distributions.

We next find the steady-state mean interdeparture time. First we find  $\wp$  and  $\pi$ . We have actually done most of the work already. The  $\mathcal{Q}$  of this section is the same as that in the previous section, therefore so is  $\pi$ , as given in (8.3.6b). First multiply  $\pi$  by  $\mathcal{L}$  [remembering that for MMPP,  $\mathcal{L} = \mathcal{L}_o$  from (8.3.10a)]

$$\begin{aligned} \pi \mathcal{L} &= \frac{1}{\sum p_i \bar{t}_i} [p_1 \mathbf{p}_1 \mathbf{V}_1, p_2 \mathbf{p}_2 \mathbf{V}_2, \dots, p_M \mathbf{p}_M \mathbf{V}_M] \mathcal{L}_o \\ &= \frac{1}{\sum p_i \bar{t}_i} [p_1 \mathbf{p}_1 \mathbf{V}_1 \lambda_1, p_2 \mathbf{p}_2 \mathbf{V}_2 \lambda_2, \dots, p_M \mathbf{p}_M \mathbf{V}_M \lambda_M] = \frac{1}{p \mathbf{V}_o \mathbf{e}'} \langle \mathbf{p} | \mathbf{v}_o \mathcal{L}_o. \end{aligned}$$

Thus,

$$\pi \mathcal{L} \mathbf{e}' = \frac{\sum_{i=1}^M p_i \bar{t}_i \lambda_i}{\sum_{i=1}^M p_i \bar{t}_i} = \frac{p \mathbf{V}_o \mathcal{L}_o \mathbf{e}'}{p \mathbf{V}_o \mathbf{e}'}. \quad (8.3.11a)$$

The steady-state vector comes directly from (8.2.8e) and satisfies  $\wp \mathbf{e}' = 1$ .

$$\wp = \frac{\pi \mathcal{L}}{\pi \mathcal{L} \mathbf{e}'} = \frac{1}{p \mathbf{V}_o \mathcal{L}_o \mathbf{e}'} \langle \mathbf{p} | \mathbf{v}_o \mathcal{L}_o. \quad (8.3.11b)$$

We can now find  $\mathbb{E}[X] = \wp \mathcal{V} \mathbf{e}'$ . But wait; from (8.2.8c) and (8.3.11a) we already know what it is, namely:

$$\mathbb{E}[X] = \wp \mathcal{V} \mathbf{e}' = \frac{1}{\pi \mathcal{L}_o \mathbf{e}'} = \frac{p \mathbf{V}_o \mathbf{e}'}{p \mathbf{V}_o \mathcal{L}_o \mathbf{e}'}. \quad (8.3.11c)$$

This has a straightforward physical interpretation. The numerator is the average time spent by the token per visit to some  $S_i$ , averaged over all servers. The term  $\bar{t}_i \lambda_i = (\mathbf{V}_o \mathbf{L}_o)_{ii}$  is the average number of departures while the token is at  $S_i$ . Thus the denominator is the average number of departures per token visit, averaged over all visits.

Keep in mind that (8.3.11c) is only valid for steady-state epochs. If the system is in some vector state, say  $\wp_o$ , or  $\wp_n = \wp_o \mathcal{V}$  (the  $n$ th customer has just departed), then one must compute

$$\mathbb{E}[X_{n+1}] = \wp_n \mathcal{V} \epsilon'$$

using (8.3.10c). This takes some skill and practice to do analytically (see Section 9.3). However, all the moments, and even autocorrelation coefficients are easy to compute.

In the rest of this subsection we look at applying MMPP's to problems in telecommunications. We show how to use physical arguments to make mathematical changes to the model.

#### 8.3.2.4 Augmented MMPP's (AMMPP)

There is one problem with this model, particularly for the **ON-OFF processes** we will be discussing later. As was mentioned in the discussion following (8.3.10c),  $d_i$  is the probability that the token will leave  $S_i$  without any packets departing (customers are now called **packets**, or **cells**). That is, sometimes a token's visit to  $S_i$  will result in no sent packets. If the mean number of packets per visit is large, then  $d_i$  will be small and nothing need be done. But if that is not the case, then some modifications must be made. After all, by definition each *ON* interval must have at least one packet. After all, it represents actual transmission, not merely permission to transmit.

Thus we introduce the **Augmented MMPP** (AMMPP). The token wanders through the system as usual. While it is at  $S_i$ , customers depart at rate  $\lambda_i$ . But when the token leaves  $S_i$ , another customer leaves. Thus,  $\mathbf{Q}$  is the same as in the two previous sections, but  $\mathbf{L}$  is the sum of the MRDP and MMPP  $\mathbf{L}$ s. (note that the resulting AMMPP is not an MMPP. That is [from (8.3.3d)]:

$$\mathbf{Q} = \mathbf{B}_o - \mathbf{B}_o \langle \mathbf{P} \rangle \quad (8.3.12a)$$

and [adding (8.3.5b) to (8.3.10a)]

$$\mathbf{L} = \mathbf{L}_o + \mathbf{B}_o \langle \mathbf{P} \rangle. \quad (8.3.12b)$$

Then

$$\mathbf{B} = \mathbf{Q} + \mathbf{L} = \mathbf{B}_o - \mathbf{B}_o \langle \mathbf{P} \rangle + \mathbf{B}_o \langle \mathbf{P} \rangle + \mathbf{L}_o = \mathbf{B}_o + \mathbf{L}_o \quad (8.3.12c)$$

and

$$\mathcal{V} = [\mathbf{B}_o + \mathbf{L}_o]^{-1} = \mathcal{V}_o [\mathbf{I} + \mathbf{L}_o \mathcal{V}_o]^{-1} = \mathcal{V}_o \mathcal{D}_o. \quad (8.3.12d)$$

The various matrices for all three schemes are presented in Table 8.3.1 for comparison and reference.

Table 8.3.1. Comparison of Processes

	MRDP	MMPP	AMMPP
$\mathcal{Q}$	$\mathcal{B}_o - \mathcal{B}_o \langle \mathbf{P} \rangle$	$\mathcal{B}_o - \mathcal{B}_o \langle \mathbf{P} \rangle$	$\mathcal{B}_o - \mathcal{B}_o \langle \mathbf{P} \rangle$
$\mathcal{L}$	$\mathcal{B}_o \langle \mathbf{P} \rangle$	$\mathcal{L}_o$	$\mathcal{L}_o + \mathcal{B}_o \langle \mathbf{P} \rangle$
$\mathcal{B}$	$\mathcal{B}_o$	$\mathcal{L}_o + \mathcal{B}_o - \mathcal{B}_o \langle \mathbf{P} \rangle$	$\mathcal{B}_o + \mathcal{L}_o$
$\mathcal{V}$	$\mathcal{V}_o$	$\mathcal{X}_o \mathcal{V}_o \mathcal{D}_o$	$\mathcal{V}_o \mathcal{D}_o$
$\mathcal{Y}$	$\langle \mathbf{P} \rangle$	$\mathcal{X}_o \mathcal{V}_o \mathcal{D}_o \mathcal{L}_o$	$\mathcal{I} - \mathcal{D}_o + \mathcal{D}_o \langle \mathbf{P} \rangle$
$\pi$	$\kappa_o \langle \mathbf{p}   \mathcal{V}_o$	$\kappa_o \langle \mathbf{p}   \mathcal{V}_o$	$\kappa_o \langle \mathbf{p}   \mathcal{V}_o$
$\wp$	$\langle \mathbf{p}  $	$\kappa_1 \langle \mathbf{p}   \mathcal{V}_o \mathcal{L}_o$	$\kappa_2 \langle \mathbf{p}   (\mathcal{I} + \mathcal{V}_o \mathcal{L}_o)$
$\wp \mathcal{V}$	$\langle \mathbf{p}   \mathcal{V}_o$	$\kappa_1 \langle \mathbf{p}   \mathcal{V}_o$	$\kappa_2 \langle \mathbf{p}   \mathcal{V}_o$
$\mathbf{E}[X]$	$\mathbf{p} \mathbf{V}_o \mathbf{e}'$	$\kappa_1 / \kappa_o$	$\kappa_2 / \kappa_o$

**Notes:**

1. All three have the same  $\mathcal{Q}$ , and therefore the same  $\pi$ ;
2. Given  $\wp = c\pi\mathcal{B}$  [from (8.2.8e)], the vectors,  $\wp\mathcal{V}$ , must be proportional to  $\pi$  and to each other;
3. The number of departures per visit for the augmented process is one more than that for the MMPP, and of course, the number of departures per visit for the MRDP is 1 (see the denominators of  $\mathbf{E}[X]$ );
4.  $\mathcal{X}_o := [\mathcal{I} - \mathcal{D}_o \langle \mathbf{P} \rangle]^{-1} = \mathcal{I} + \mathcal{D}_o \langle \mathbf{P} (\mathcal{I} - \mathcal{D}_o \mathbf{P})^{-1} \rangle$ ;
5.  $\kappa_o := (\mathbf{p} \mathbf{V}_o \mathbf{e}')^{-1}$ ;
6.  $\kappa_1 := (\mathbf{p} \mathbf{V}_o \mathbf{L}_o \mathbf{e}')^{-1}$ ;
7.  $\kappa_2 := (\mathbf{p} \mathbf{V}_o \mathbf{L}_o \mathbf{e}' + 1)^{-1}$

### 8.3.2.5 ON-OFF Models (Bursty Traffic)

Researchers in telecommunications have long been aware that information traffic (e.g., voice communication or transmission of data packets) is very non-uniform. (see, e.g., Leland et al. [LELANDETAL94].) That is, the amount of traffic from time interval to time interval fluctuates enormously. This kind of behavior is called *bursty*. It is explained, at least for voice, as follows. While someone is speaking, data flows at a *peak rate*, but when that person stops speaking, no data are transmitted until someone speaks again. This can satisfactorily be modeled by a 2-state MMPP model, where, say,  $\lambda_1 = \lambda_p$  is the peak rate at which information flows when someone is talking ( $S_1$  represents the *ON time*), and  $\lambda_2 = 0$  when there is silence ( $S_2$  represents the

**OFF time**). This has been called a *one-burst process*. If the times between packets during an *ON* time are exponentially distributed (the usual assumption) the system is also called an *Interrupted Poisson Process* (IPP) (see, e.g., [LEE-LIEF-WALLACE00]). When one analyzes the superposition of several voice streams, it is difficult to tell where the *ON* and *OFF* periods are, but the burstiness remains. Still, satisfactory MMPP models were constructed where several servers, corresponding to  $1, 2, \dots, n$  simultaneous voice streams were included. In this case,  $\lambda_n = n\lambda_1$ . Reasonable  $\mathbf{P}$  matrices were constructed to reflect the probability that a new voice stream will join in, or a present one will stop. We might call these **ON-OFF MMPP's** (OOMMPP).

As data transmission became more common the MMPP models were found to be less and less useful. Further examination of data streams showed that there was *long-range autocorrelation* [CROVELLABESTAVROS96]. That is,  $\hat{r}(k)$  [see (8.2.12c)] remains measurable for very large  $k$ . This could not be modeled by the then-existing models. But further measurements of data revealed that the size of transmitted files is power-tailed for many orders of magnitude; see Hatem [HATEM97], Lipsky [LIPGARGROBBERT92], and Crovella [CROVELLABESTAVROS96]. See Section 3.3 for a full discussion, including the TPTs. When files are to be transmitted they are first broken up into packets. The packets are then sent in a smooth (Poisson?) manner, for a period of time which is PT. That is, the *ON* times must be power-tail distributed. The model presented in Section 8.3.2.3 is adequate, even reproducing the long-range autocorrelation, if a good representation of PT distributions is used. Strictly speaking, PT functions require infinite representations, but in Section 3.3.6.2 we present a truncated variety that has been shown to be more than adequate (see Schwefel [SCHWEFEL00]).

Perhaps the best way to become familiar with all the above matrices is by an example. Let us consider a simple *ON-OFF* model. It shows that previously unknown properties can be discovered without actually having to specify the PDFs of the  $S_i$ s. In fact, we come up with some interesting results.

**Example 8.3.2:** Consider a system with two servers  $S_1$  and  $S_2$ , with distributions represented by  $\langle \mathbf{p}_1, \mathbf{B}_1 \rangle$  and  $\langle \mathbf{p}_2, \mathbf{B}_2 \rangle$ , respectively. While the token is at  $S_1$  a data source sends a *burst of packets* at a *peak* rate of  $\lambda_p$  for a time  $T_1$ . When the burst is over, the token goes to  $S_2$  for a time  $T_2$ , during which time no packets are sent ( $\lambda_2 = 0$ ). The token then returns to  $S_1$ , repeating the process indefinitely. The matrices describing the system are:

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \langle \mathbf{P} \rangle = \begin{bmatrix} \mathbf{0}_1 & \epsilon'_1 \mathbf{p}_2 \\ \epsilon'_2 \mathbf{p}_1 & \mathbf{0}_2 \end{bmatrix},$$

$$\mathbf{p} = \left[ \frac{1}{2}, \frac{1}{2} \right], \quad \epsilon' = \left[ \begin{array}{c} \epsilon'_1 \\ \epsilon'_2 \end{array} \right] = |e'\rangle,$$

and (using  $\mathcal{D}_o = [\mathcal{I} + \mathcal{L}_o \mathcal{V}_o]^{-1}$  with  $\mathbf{D}_1 = [\mathbf{I} + \lambda_p \mathbf{V}_1]^{-1}$ )

$$\mathcal{L}_o = \left[ \begin{array}{cc} \lambda_p \mathbf{I}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{array} \right], \quad \mathcal{B}_o = \left[ \begin{array}{cc} \mathbf{B}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 \end{array} \right],$$

$$\mathbf{v}_o = \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 \end{bmatrix}, \quad \mathbf{D}_o = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_2 \end{bmatrix}.$$

The matrices governing the 2-server MMPP *ON-OFF process* follow.

$$\mathcal{L} = \mathcal{L}_o, \quad \mathcal{Q} = \begin{bmatrix} \mathbf{B}_1 & -\mathbf{B}_1 \boldsymbol{\epsilon}'_1 \mathbf{p}_2 \\ -\mathbf{B}_2 \boldsymbol{\epsilon}'_2 \mathbf{p}_1 & \mathbf{B}_2 \end{bmatrix},$$

$$\mathcal{B} = \begin{bmatrix} \mathbf{B}_1 + \lambda_p \mathbf{I}_1 & -\mathbf{B}_1 \boldsymbol{\epsilon}'_1 \mathbf{p}_2 \\ -\mathbf{B}_2 \boldsymbol{\epsilon}'_2 \mathbf{p}_1 & \mathbf{B}_2 \end{bmatrix},$$

(using  $d := \mathbf{p}_1 \mathbf{D}_1 \boldsymbol{\epsilon}'_1$ )

$$\begin{aligned} \mathbf{v} &= \begin{bmatrix} \mathbf{V}_1 \mathbf{D}_1 + \frac{1}{1-d} \mathbf{D}_1 \boldsymbol{\epsilon}'_1 \mathbf{p}_1 \mathbf{V}_1 \mathbf{D}_1 & \frac{1}{1-d} \mathbf{D}_1 \boldsymbol{\epsilon}'_1 \mathbf{p}_2 \mathbf{V}_2 \\ \frac{1}{1-d} \boldsymbol{\epsilon}'_2 \mathbf{p}_1 \mathbf{V}_1 \mathbf{D}_1 & \mathbf{V}_2 + \frac{d}{1-d} \boldsymbol{\epsilon}'_2 \mathbf{p}_2 \mathbf{V}_2 \end{bmatrix}, \\ \boldsymbol{\wp} &= \frac{1}{\bar{t}_1} [\mathbf{p}_1 \mathbf{V}_1, \mathbf{o}], \quad \mathcal{Y} = \lambda_p \begin{bmatrix} \mathbf{V}_1 \mathbf{D}_1 + \frac{1}{1-d} \mathbf{D}_1 \boldsymbol{\epsilon}'_1 \mathbf{p}_1 \mathbf{V}_1 \mathbf{D}_1 & \mathbf{0} \\ \frac{1}{1-d} \boldsymbol{\epsilon}'_2 \mathbf{p}_1 \mathbf{V}_1 \mathbf{D}_1 & \mathbf{0} \end{bmatrix}, \\ \pi &= \frac{1}{p \mathbf{V}_o \boldsymbol{\epsilon}'} \langle p | \mathbf{v}_o = \frac{1}{\bar{t}_1 + \bar{t}_2} [\mathbf{p}_1 \mathbf{V}_1, \mathbf{p}_2 \mathbf{V}_2], \\ \boldsymbol{\wp} \mathbf{v} &= \frac{1}{p \mathbf{V}_o \boldsymbol{\epsilon}'} \langle p | \mathbf{v}_o = \frac{1}{\lambda_p \bar{t}_1} [\mathbf{p}_1 \mathbf{V}_1, \mathbf{p}_2 \mathbf{V}_2]. \end{aligned}$$

We now find the mean and variance of the interdeparture times. The mean is simple enough to evaluate. It is

$$\mathbb{E}[X] = \boldsymbol{\wp} \mathbf{v} \boldsymbol{\epsilon}' = \frac{\bar{t}_1 + \bar{t}_2}{\bar{n}_p}, \quad (8.3.13a)$$

where  $\bar{n}_p := \lambda_p \bar{t}_1$  is the mean number of packets per cycle and the numerator is the total time for one cycle. Considered as a flow of packets, the mean flow rate  $\kappa$  [see 8.2.8d] is [the same as  $1/(\boldsymbol{\wp} \mathbf{v} \boldsymbol{\epsilon}')$ ]:

$$\kappa := \pi \mathcal{L} \boldsymbol{\epsilon}' = \left[ \frac{\bar{t}_1}{\bar{t}_1 + \bar{t}_2} \right] \lambda_p.$$

In many applications it is possible to change  $\lambda_p$ , by for instance, increasing the transmission speed of data. At that moment, the amount of data to be sent is fixed, therefore  $\bar{t}_1$  decreases in such a way that  $\bar{n}_p$  remains constant. Therefore, at least for *ON-OFF* models, it is appropriate to replace  $\lambda_p \bar{t}_1$  by  $\bar{n}_p$ . The typical picture is of data being prepared for transmission and then sent to the transmitter. In this scenario, even if the data are transmitted more rapidly, the next batch of data won't be ready for transmission until one full cycle later. In other words  $\bar{t}_1 + \bar{t}_2$ , like  $\bar{n}_p$ , is constant. In such cases the **burst parameter**,  $b$ , can be a useful variable for describing the performance of the application.

$$b := \frac{\bar{t}_2}{\bar{t}_1 + \bar{t}_2} = 1 - \frac{\kappa}{\lambda_p}. \quad (8.3.13b)$$



When  $b = 0$ ,  $\lambda_p = \kappa$ , and the *OFF* time is 0; that is, there is no burstiness and the traffic is pure Poisson. As  $\lambda_p$  increases unboundedly,  $b \rightarrow 1$ , and in the limit, all the packets are sent at the same time, that is, in *bulk*. When packets arrive at a server in this manner, it is called a *bulk arrival process*, or *batch arrival process*. (See, e.g., [GROSS-HARRIS98].)

Perhaps the easiest way to find the variance is to first evaluate  $\mathbf{V}\boldsymbol{\epsilon}'$ . We do that now, finding

$$\mathbf{V}\boldsymbol{\epsilon}' = \begin{bmatrix} \frac{1}{\lambda_p}\boldsymbol{\epsilon}'_1 + \frac{\bar{t}_2}{1-d}\mathbf{D}_1\boldsymbol{\epsilon}'_1 \\ \frac{1}{\lambda_p}\boldsymbol{\epsilon}'_2 + \frac{\bar{t}_2 d}{1-d}\boldsymbol{\epsilon}'_2 + \mathbf{V}_2\boldsymbol{\epsilon}'_2 \end{bmatrix}.$$

We know that  $\mathbb{E}[X^2] = 2\boldsymbol{\wp}\mathbf{V}^2\boldsymbol{\epsilon}'$ , so

$$\begin{aligned} \sigma_X^2 &= 2(\boldsymbol{\wp}\mathbf{V})(\mathbf{V}\boldsymbol{\epsilon}') - [\mathbb{E}[X]]^2 \\ &= \frac{1}{\lambda_p \bar{t}_1} \sigma_2^2 + \frac{1}{(\lambda_p \bar{t}_1)^2} \left[ \bar{t}_1^2 + 2\bar{t}_1 \bar{t}_2 + (\lambda_p \bar{t}_1 - 1)\bar{t}_2^2 + \frac{2\lambda_p d}{1-d} \bar{t}_1 \bar{t}_2^2 \right] \\ &= \frac{1}{\bar{n}_p} \sigma_2^2 + \left( \frac{1}{\bar{n}_p} \right)^2 \left[ (\bar{t}_1 + \bar{t}_2)^2 + \frac{2\bar{n}_p \bar{t}_2^2}{1-d} - (\bar{n}_p + 2)\bar{t}_2^2 \right]. \end{aligned} \quad (8.3.13c)$$

This somewhat unwieldy expression can be brought into simpler form by looking at the squared coefficient of variation:

$$C_X^2 := \frac{\sigma_X^2}{\mathbb{E}[X]^2} = 1 + b^2 \left[ \bar{n}_p C_2^2 + \frac{2\bar{n}_p}{1-d} - (\bar{n}_p + 2) \right]. \quad (8.3.13d)$$

If  $b = 0$  ( $\bar{t}_2 = 0$ ), then  $C_X^2 = 1$  corresponding to a Poisson process.

The dependence of  $C_X^2$  on the *OFF* time distribution is explicit in  $C_2^2$ , but the dependence on the *ON* time is implicitly contained in the behavior of  $d$ . In fact, all properties of this *ON-OFF* model depend on the *ON* time distribution through powers of  $\mathbf{D}_1 = (\mathbf{I}_1 + \lambda_p \mathbf{V}_1)^{-1}$ . This, in turn, seems to depend on the peak rate  $\lambda_p$  as well. But in our model the two are intimately connected, not merely through  $\bar{n}_p = \lambda_p \bar{t}_1$ . After all, during an *ON* time a certain number of packets are transmitted, and that should be independent of how fast they are sent. That is, the distribution of  $T_1$  depends on the distribution of the number of packets in a burst.

In Definition 3.2.1 we introduced the equivalence relation that groups together functions that have the same shape. Let  $\hat{\mathbf{V}}_1$  generate a function with the same shape as  $\mathbf{V}_1$ , but with mean  $\mathbf{p}_1 \hat{\mathbf{V}}_1 \boldsymbol{\epsilon}'_1 = 1$ . Then  $T_1$  has a distribution generated by  $\langle \mathbf{p}_1, \bar{t}_1 \hat{\mathbf{V}}_1 \rangle$ , with mean  $\mathbb{E}[T_1] = \bar{t}_1$ . We see, then, that

$$\mathbf{D}_1 = (\mathbf{I}_1 + \lambda_p \bar{t}_1 \hat{\mathbf{V}}_1)^{-1} = (\mathbf{I}_1 + \bar{n}_p \hat{\mathbf{V}}_1)^{-1}.$$

$\bar{n}_p$  is assumed to be the same irrespective of the peak rate; consequently we see that  $\mathbf{D}_1$  is also independent of  $\lambda_p$ .

We now go one step further. In Section 4.4.1 we found a relationship between the *exponential moments*  $\alpha_k(s)$  and the matrices,  $\mathbf{D}^k$ . From (4.4.1a) we have

$$\alpha_k(s) = \int_0^\infty \frac{(sx)^k}{k!} e^{-sx} b(x) dx = \Psi[(\mathbf{I} - \mathbf{D})^k \mathbf{D}] = \Psi[(s\mathbf{V}\mathbf{D})^k \mathbf{D}].$$

If we identify  $s$  with  $\lambda_p$  and  $\mathbf{D}(s)$  with  $\mathbf{D}_1(\lambda_p)$ , we can interpret  $\alpha_k$  to be the probability that exactly  $k$  packets are sent during an *ON* period. Given that  $\mathbf{D}_1$  is independent of  $\lambda_p$ , so is  $\alpha_k$ . This can also be seen directly from the integral term. From Definition 3.2.1 we have  $b(x) = \hat{b}(x/\bar{t}_1)/\bar{t}_1$ . We put this into the equation above, let  $x = \bar{t}_1 u$ , recognize that  $\lambda_p \bar{t}_1 = \bar{n}_p$ , and get:

$$\alpha_k(\bar{n}_p) = \int_0^\infty \frac{(\bar{n}_p u)^k}{k!} e^{-\bar{n}_p u} b(u) du.$$

Thus, the number of packets per *ON* time does not depend on  $\lambda_p$  or  $\bar{t}_1$  independently.

As a specific example, if  $T_1$  is exponentially distributed, we get

$$d_e = \Psi[(\mathbf{I} + \lambda_p \mathbf{V}_1)^{-1}] = \frac{1}{1 + \bar{n}_p}$$

and  $C_X^2$  simplifies to  $C_{X_e}^2 = 1 + \bar{n}_p b^2(C_2^2 + 1)$ .

If all *ON* times are the same ( $T_1 = \bar{t}_1$ ; i.e., the **d**eterministic distribution), then  $d_d = e^{-\bar{n}_p}$  and

$$C_{X_d}^2 = 1 + b^2 \left[ \bar{n}_p C_2^2 + \frac{2\bar{n}_p}{1 - e^{-\bar{n}_p}} - (\bar{n}_p + 2) \right].$$

For fixed  $b$ ,  $\bar{n}_p$ , and  $C_2^2$  this expression provides a lower bound on  $C_X^2$ , but there is no upper bound. Recall that  $d$  is the probability that an *ON* time will end without any packets. Also, distributions where the vast majority of *ON* times are very small are possible, leading to a value for  $d$  that can be very close to 1, making  $1/(1 - d)$  arbitrarily large in (8.3.13d).  $\blacktriangle$

The formula for the autocovariance leads to two interesting results, which we state as theorems. The following is really a corollary to Theorem 8.2.1.

**Theorem 8.3.1:** For any pure MMPP *ON-OFF arrival process*, if the *ON*-time distribution is exponentially distributed, then the process is a renewal process (the interarrival times are iid). This is true irrespective of the *OFF*-time distribution. The interarrival times are represented by  $\langle \boldsymbol{\varphi}, \mathcal{B} \rangle$ , where

$$\mathcal{B} = \begin{bmatrix} \mu_1 + \lambda_p & -\mu_1 \mathbf{P}_2 \\ -\mathbf{B}_2 \boldsymbol{\epsilon}'_2 & \mathbf{B}_2 \end{bmatrix}$$

and  $\wp = [1, 0, \dots, 0]$ . ■

**Proof:** The formulas of Example 8.3.2 apply here with the following substitutions:  $\mathbf{I}_1 \rightarrow 1$ ,  $\mathbf{B}_1 \rightarrow \mu_1 := 1/\bar{t}_1$ ,  $\mathbf{V}_1 \rightarrow \bar{t}_1$ ,  $\mathbf{p}_1 \rightarrow 1$ ,  $\mathbf{D}_1 \rightarrow 1/(1 + \bar{n}_p)$ , and  $\epsilon'_1 \rightarrow 1$ . Then

$$\mathcal{L} \Rightarrow \lambda_p \begin{bmatrix} 1 & \mathbf{o} \\ \mathbf{o}' & \mathbf{0} \end{bmatrix} = \lambda_p \begin{bmatrix} 1 \\ \mathbf{o}' \end{bmatrix} \wp;$$

that is,  $[\mathcal{L}]_{11} = 1$  and all other elements are 0. Also,

$$\mathcal{Y} \Rightarrow \begin{bmatrix} 1 & \mathbf{o} \\ \epsilon'_2 & \mathbf{0} \end{bmatrix} = \epsilon'_2 \wp.$$

This makes  $\mathcal{L}$  and  $\mathcal{Y}$  rank-1 matrices. Therefore by Theorem 8.2.1 the process is a renewal process. Each departure epoch (time between departures) can be described in the following way. The customer starts in  $S_1$ . Because it is exponential no  $\mathbf{p}_1$  vector is necessary. Then, after mean time  $\bar{t}_1$  he either departs [with probability  $\bar{n}_p/(1 + \bar{n}_p)$ ] or goes to  $S_2$ . After a mean time of  $\bar{t}_2 = \mathbf{p}_2 \mathbf{V}_2 \epsilon'_2$ , he returns to  $S_1$ . The cycle continues until he finally departs. Each new customer begins at the exponential server, so the interdeparture times are iid. QED

The second interesting result concerns autocovariance and autocorrelation, and is given in the following.

**Theorem 8.3.2:** The  $\text{Cov}(X, X_{+k})$ s are independent of the *OFF-time* distribution. Also,  $\hat{r}(k)$  varies inversely with  $\sigma_{OFF}^2$  but no other moments. ■

**Proof:** The autocovariance lag- $k$  is given by (8.2.12b), but first we look at (8.2.12a). From the previous example we see that  $\mathcal{Y}$  is of the form

$$\mathcal{Y} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{B} & \mathbf{0} \end{bmatrix}.$$

Direct multiplication shows that

$$\mathcal{Y}^k = \begin{bmatrix} \mathbf{A}^k & \mathbf{0} \\ \mathbf{B} \mathbf{A}^{k-1} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{A}^{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

and  $\wp \mathcal{Y} \mathcal{Y}$  is of the form  $[\mathbf{a}_1, \mathbf{o}]$ . Thus, from (8.2.12a),  $\mathbf{E}[X, X_{+k}]$  is of the form

$$[\wp \mathcal{Y} \mathcal{Y}] \mathcal{Y}^{k-1} [\mathcal{V} \epsilon'] = [\mathbf{a}_1, \mathbf{o}] \begin{bmatrix} \mathbf{A}^{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}'_1 \\ \mathbf{d}'_2 \end{bmatrix} = \mathbf{a}_1 \mathbf{A}^{k-1} \mathbf{c}'_1,$$

where

$$\begin{aligned} \mathbf{a}_1 &= \frac{1}{\lambda_p \bar{t}_1} \left[ \mathbf{p}_1 \mathbf{V}_1 + \frac{\bar{t}_2}{1-d} \mathbf{p}_1 (\mathbf{I}_1 - \mathbf{D}_1) \right], \\ \mathbf{A} &= \lambda_p \left[ \mathbf{V}_1 \mathbf{D}_1 + \frac{1}{1-d} \mathbf{D}_1 \epsilon'_1 \mathbf{p}_1 \mathbf{V}_1 \mathbf{D}_1 \right], \end{aligned}$$

and

$$\mathbf{c}'_1 = \left[ \frac{1}{\lambda_p} \boldsymbol{\epsilon}'_1 + \frac{\bar{t}_2}{1-d} \mathbf{D}_1 \boldsymbol{\epsilon}'_1 \right].$$

The calculations can be done entirely in  $S_1$  space. Next, from (8.2.12b) and (8.3.13a),

$$\text{Cov}(X, X_{+k}) = \mathbf{a}_1 \mathbf{A}^{k-1} \mathbf{c}'_1 - \left[ \frac{\bar{t}_1 + \bar{t}_2}{\lambda_p \bar{t}_1} \right]^2. \quad (8.3.14a)$$

By looking at  $\mathbf{A}$ ,  $\mathbf{a}_1$ ,  $\mathbf{c}'_1$ , and  $\mathbf{E}[X]$  it is clear that  $\text{Cov}(X, X_{+k})$  does not depend on the *OFF*-time distribution, except for  $\bar{t}_2$ . In other words, it is the same for every *OFF*-time distribution with the same mean. However, it depends very heavily on the *ON*-time distribution through the operators  $\mathbf{V}_1$  and  $\mathbf{D}_1$ , and is different for each value of  $k$  because they will appear in ever-increasing powers with increasing  $k$ .

The autocorrelation coefficient lag- $k$  is found from (8.2.12c) and (8.3.13c) by evaluating:

$$\hat{r}(k) = \frac{\text{Cov}(X, X_{+k})}{\sigma_X^2}. \quad (8.3.14b)$$

$\hat{r}(k)$  behaves in a manner similar to  $\text{Cov}(X, X_{+k})$  in that systems with different *OFF*-time distributions will show proportional behavior for all  $k$ . Interestingly, if the *OFF*-time is PT, with  $\sigma_2^2 = \infty$ , then  $\sigma_X^2 = \infty$  and  $\hat{r}(k) = 0$  even though the autocovariance is finite and measurable. Even if the PT is truncated,  $\sigma_2^2$  may be extremely large, and  $\hat{r}(k)$  may be too small to measure. **QED**

But if there is some background Poisson traffic, then all bets are off. Customers depart during the *OFF*-times, and all those components that were identically 0 are now finite. Several researchers have observed behavior such as that described here [ANTONIOSSCHWEFELLIP07]. Perhaps this analysis will explain those results. We go into this further after the following example.

**Example 8.3.3:** Here we find an explicit algebraic expression for  $\hat{r}(1)$ . From (8.3.14a),

$$\begin{aligned} & \text{Cov}(X, X_{+1}) \\ &= \frac{1}{\bar{n}_p} \left[ \mathbf{p}_1 \mathbf{V}_1 + \frac{\bar{t}_2}{1-d} \mathbf{p}_1 (\mathbf{I}_1 - \mathbf{D}_1) \right] \left[ \frac{1}{\lambda_p} \boldsymbol{\epsilon}'_1 + \frac{\bar{t}_2}{1-d} \mathbf{D}_1 \boldsymbol{\epsilon}'_1 \right] - \left[ \frac{\bar{t}_1 + \bar{t}_2}{\bar{n}_p} \right]^2. \end{aligned}$$

From Example 8.3.2,  $d = \alpha_o(\bar{n}_p)$  and the covariance can be written as

$$\text{Cov}(X, X_{+1}) = \left[ \frac{\bar{t}_2}{\bar{n}_p} \right]^2 \left[ \frac{\bar{n}_p \alpha_1(\bar{n}_p)}{[1 - \alpha_o(\bar{n}_p)]^2} - 1 \right].$$

Combining this with (8.3.13c), we get

$$\hat{r}(1) = \frac{b^2 \left[ \frac{\bar{n}_p \alpha_1}{(1-\alpha_o)^2} - 1 \right]}{1 + b^2 \left[ \bar{n}_p C_2^2 + \frac{2\bar{n}_p}{1-\alpha_o} - (\bar{n}_p + 2) \right]}.$$

Well, it could have turned out messier. It does display the properties we established previously. For instance, if  $b \rightarrow 0$  then  $\hat{r}(1) = 0$ , as should be the case for all  $\hat{r}(k)$ , because in that limit the process becomes a Poisson process. Furthermore, it depends on  $S_2$  only through  $C_2^2$  and decreases with increasing variance. In addition, it only depends on the probabilities of having 0 or 1 packet in the  $ON$  period.

Last, for exponential  $ON$  times,

$$\alpha_k(\bar{n}_p) = \left[ \frac{\bar{n}_p}{\bar{n}_p + 1} \right]^k \frac{1}{\bar{n}_p + 1}.$$

As would be expected, geometric distribution of packets per burst yields exponential  $ON$  times, and it is independent of  $\lambda_p$  and  $\bar{t}_1$ . It also follows that  $\hat{r}(1) = 0$ , consistent with Theorem 8.3.1.  $\blacktriangle$

### MRP's with Background Poisson Traffic Added

We now return to the question of what happens when there is background Poisson traffic. Here, for any MRP, whatever phase the token is in, the background source produces at the rate  $\lambda_b$ . The  $\mathbf{Q}$  matrix doesn't change, and  $\mathcal{L}$  is modified to:

$$\mathcal{L}_b = \mathcal{L} + \lambda_b \mathcal{I}$$

and  $\mathcal{B}$  is modified to

$$\mathcal{B}_b = \mathcal{Q} + \mathcal{L}_b = \mathcal{B} + \lambda_b \mathcal{I}.$$

We look at the special case of merging renewal processes in Section 8.3.3. If one of them is Poisson, then that is equivalent to what we have here. But right now we are interested in what happens to an  $ON-OFF$  process. If the term  $\lambda_b \mathcal{I}$  is added to  $\mathcal{L}$  in Example 8.3.2, then the result is a process that looks exactly like any 2-server MMPP. This leads us to the following theorem.

**Theorem 8.3.3:** Every MMPP is equivalent to some  $ON-OFF$  MMPP with background Poisson traffic.  $\blacksquare$

**Proof:** Consider any MMPP with  $\mathbf{Q}$ ,  $\mathcal{L}$ , and  $\mathcal{B}$  given. Because the process is MMPP,

$$\mathcal{L} = \text{Diag}[\lambda_1 \mathbf{I}_1, \lambda_2 \mathbf{I}_2, \dots, \lambda_M \mathbf{I}_M].$$

Define  $\lambda_b$  as the smallest  $\lambda_s$ . That is

$$\lambda_b := \text{Min}\{\lambda_i | 1 \leq i \leq M\}.$$

We now construct an  $ON-OFF$  process (using subscript “oo”) as follows. Let

$$\mathcal{L}_{oo} = \mathcal{L} - \lambda_b \mathcal{I}, \quad \mathcal{Q}_{oo} = \mathcal{Q}, \quad \text{and} \quad \mathcal{B}_{oo} = \mathcal{B} - \lambda_b \mathcal{I}.$$

Then  $\{\mathcal{L}_{oo}, \mathcal{B}_{oo}, \mathcal{Q}_{oo}\}$  is an  $ON-OFF$  process. After all, while the token is visiting the server corresponding to  $\lambda_b$  no packets are leaving; that is, it's  $OFF$ . We can now reverse the process by adding  $\lambda_b \mathcal{I}$  to the  $ON-OFF$  process and end up with the original MMPP. **QED**

One might ask if this is true in general. The answer is “No.” It is true for MMPP’s because of the particular structure of  $\mathcal{L}$ , but each process must be examined individually. We show this in what follows.

### Modified Augmented ON-OFF MMPP Model (MAOOMMPP)

We next examine the augmented MMPP that is also an *ON-OFF* process (MAOOMMPP). This process forces the *ON* server to yield at least one packet per token visit. But it cannot be a special case of the AMMPP model presented in Section 8.3.2.4. There, every token visit to every server produced at least one packet. This must not be allowed to happen at the *OFF* server. The way we augmented the MMPP model was to add  $\mathbf{B}_o \langle \mathbf{P} \rangle$  to  $\mathcal{L}$ . A particular term in that matrix is  $\mathbf{B}_i \boldsymbol{\epsilon}'_i P_{ij} \mathbf{p}_j$ , or  $\mathbf{M}_i \mathbf{q}'_i P_{ij} \mathbf{p}_j$ . Its interpretation is as follows. Start with the token finishing in phase  $k$  of server  $S_i$   $[(\mathbf{M}_i)_{kk}]$ , leaving  $S_i$   $[(\mathbf{q}_i)_k]$ , going to  $S_j$   $[(\mathbf{P})_{ij}]$ , and finally going to phase  $\ell$  in  $S_j$   $[(\mathbf{p}_j)_\ell]$ . Because this is a term in the  $\mathcal{L}$  matrix, it causes a customer to depart. If  $S_i$  is the *OFF* server, then this shouldn’t happen. Therefore we *modify* the augmented model by setting that corresponding row (now call it  $\ell$ ) to  $\mathbf{0}$ . That is, the matrix block  $\mathcal{B}_{\ell j} = \mathbf{0}$ . This can be done formally by defining  $\mathcal{I}_{oo}$  as

$$\mathcal{I}_{oo} := \text{Diag}[\mathbf{I}_1, \dots, \mathbf{I}_{\ell-1}, \mathbf{0}, \mathbf{I}_{\ell+1}, \dots, \mathbf{I}_M].$$

Then the *Modified Augmented ON-OFF MMPP* (MAOOMMPP) model is:

$$\begin{aligned} \mathcal{Q} &= \mathcal{B}_o - \mathcal{B}_o \langle \mathbf{P} \rangle; \\ \mathcal{L} &= \mathcal{L}_o + \mathcal{B}_o \mathcal{I}_{oo} \langle \mathbf{P} \rangle; \\ \mathcal{B} &= \mathcal{B}_o + \mathcal{L}_o - \mathcal{B}_o (\mathcal{I}_o - \mathcal{I}_{oo}) \langle \mathbf{P} \rangle. \end{aligned} \tag{8.3.15}$$

Compare with the comparable entries in Table 8.3.1. We now explore the two-server system in the following example.

**Example 8.3.4:** Consider a system as in Example 8.3.2, but now the *ON* time must have at least one packet, and the *OFF* time must have none. Assume that when the token leaves the *ON* server he emits a packet, giving the MAOOMMPP. Let  $M = 2$ , then (8.3.15) becomes

$$\begin{aligned} \mathcal{L} &= \begin{bmatrix} \lambda_p \mathbf{I}_1 & \mathbf{B}_1 \boldsymbol{\epsilon}'_1 \mathbf{p}_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathcal{B} = \begin{bmatrix} \mathbf{B}_1 + \lambda_p \mathbf{I}_1 & \mathbf{0} \\ -\mathbf{B}_2 \boldsymbol{\epsilon}'_2 \mathbf{p}_1 & \mathbf{B}_2 \end{bmatrix}, \\ \mathcal{Q} &= \begin{bmatrix} \mathbf{B}_1 & -\mathbf{B}_1 \boldsymbol{\epsilon}'_1 \mathbf{p}_2 \\ -\mathbf{B}_2 \boldsymbol{\epsilon}'_2 \mathbf{p}_1 & \mathbf{B}_2 \end{bmatrix}. \end{aligned}$$

It is straightforward to set up  $\mathcal{V}$  and  $\mathcal{Y}$ .

$$\mathcal{V} = \begin{bmatrix} \mathbf{V}_1 \mathbf{D}_1 & \mathbf{0} \\ \boldsymbol{\epsilon}'_2 \mathbf{p}_1 \mathbf{V}_1 \mathbf{D}_1 & \mathbf{V}_2 \end{bmatrix} \quad \text{and} \quad \mathcal{Y} = \begin{bmatrix} \mathbf{I}_1 - \mathbf{D}_1 & \mathbf{D}_1 \boldsymbol{\epsilon}'_1 \mathbf{p}_2 \\ \boldsymbol{\epsilon}'_2 \mathbf{p}_1 (\mathbf{I}_1 - \mathbf{D}_1) & d \boldsymbol{\epsilon}'_2 \mathbf{p}_2 \end{bmatrix}.$$

$\mathcal{Q}$  is the same as before. Therefore we already know what  $\boldsymbol{\pi}$  is, and get  $\boldsymbol{\wp}$  from that.

$$\boldsymbol{\pi} = \frac{1}{\bar{t}_1 + \bar{t}_2} [\mathbf{p}_1 \mathbf{V}_1, \mathbf{p}_2 \mathbf{V}_2] \quad \text{and} \quad \boldsymbol{\wp} = \frac{1}{1 + \lambda_p \bar{t}_1} [\lambda_p \mathbf{p}_1 \mathbf{V}_1, \mathbf{p}_2].$$

The flow rate is

$$\kappa = \pi \mathcal{L} \varepsilon' = \frac{1}{\wp \mathcal{V} \varepsilon'} = \frac{1 + \lambda_p \bar{t}_1}{\bar{t}_1 + \bar{t}_2}.$$

The denominator is the mean cycle time, so the mean number of packets per burst is one more than in the pure *ON-OFF* MMPP model, as we would have hoped.

Before going on, we examine  $\mathcal{Y}$ . Although it is more complex than the  $\mathcal{Y}$  for the pure MMPP *ON-OFF* process in Example 8.3.2, it has the same rank. The other  $\mathcal{Y}$  has  $m_2$  columns of all zeroes. Also its  $\mathcal{Y}_{11}$  block matrix has an inverse. Therefore it must be of rank  $m_1$ . This  $\mathcal{Y} = \mathcal{V} \mathcal{L}$ , as it must.  $\mathcal{L}$  has  $m_2$  rows of 0, and so must be of rank  $m_1$ . Therefore,  $\mathcal{Y}$  must also have rank  $m_1$ .

If in particular, the *ON* time is exponentially distributed then  $m_1 = 1$  and  $\mathcal{Y}$  is of rank 1, just as in Theorem 8.3.1. Therefore, here too, if the *ON*-time distribution is exponential, the process is a renewal process. Direct substitution shows that  $\mathcal{L}$  and  $\mathcal{Y}$  reduce to

$$\mathcal{L} = \begin{bmatrix} \lambda_p & (1/\bar{t}_1) \mathbf{p}_2 \\ \mathbf{o}' & \mathbf{0} \end{bmatrix}$$

and

$$\mathcal{Y} = \begin{bmatrix} 1-d & d \mathbf{p}_2 \\ (1-d) \boldsymbol{\epsilon}'_2 & d \boldsymbol{\epsilon}'_2 \end{bmatrix} = \begin{bmatrix} 1 \\ \boldsymbol{\epsilon}'_2 \end{bmatrix} [1-d, d \mathbf{p}_2] = \varepsilon' \wp,$$

whereas  $\mathcal{B}$  and  $\mathcal{V}$  reduce to

$$\mathcal{B} = \begin{bmatrix} 1/\bar{t}_1 d & \mathbf{o} \\ -\mathbf{B}_2 \boldsymbol{\epsilon}'_2 & \mathbf{B}_2 \end{bmatrix} \quad \text{and} \quad \mathcal{V} = \begin{bmatrix} \bar{t}_1 d & \mathbf{o} \\ \bar{t}_1 d \boldsymbol{\epsilon}'_2 & \mathbf{V}_2 \end{bmatrix},$$

where  $d = 1/(1 + \lambda_p \bar{t}_1)$ . As already stated, this is a renewal process with interdeparture times generated by  $\langle \wp, \mathcal{B} \rangle$ . The reader should show directly that

$$\begin{aligned} \mathbb{E}[X] &= \wp \mathcal{V} \varepsilon' = d(\bar{t}_1 + \bar{t}_2) \quad \text{and} \\ \sigma_X^2 &= d \sigma_2^2 + d^2 \bar{t}_1^2 + d(1-d) \bar{t}_2^2. \end{aligned}$$

We return now to nonexponential *ON* times, The formulas get quite messy as we attempt to find  $\sigma_X^2$  and other properties. We find that

$$\wp \mathcal{V} = \frac{1}{1 + \lambda_p \bar{t}_1} [\mathbf{p}_1 \mathbf{V}_1, \mathbf{p}_2 \mathbf{V}_2]$$

and

$$\mathcal{V} \varepsilon' = \begin{bmatrix} \mathbf{V}_1 \mathbf{D}_1 \boldsymbol{\epsilon}'_1 \\ \Psi[\mathbf{V}_1 \mathbf{D}_1] \boldsymbol{\epsilon}'_2 + \mathbf{V}_2 \boldsymbol{\epsilon}'_2 \end{bmatrix}.$$

It is easy enough to evaluate  $\wp \mathcal{V} \mathcal{V} \varepsilon'$  numerically for any specific values of the various parameters, but the analytic expression is somewhat long

and not too informative. However, it can be seen that it, and therefore  $\sigma_X^2$ , depends on the *OFF* time explicitly through  $\mathbf{p}_2 \mathbf{V}_2^2 \boldsymbol{\epsilon}'_2$  and  $\bar{t}_2$  only, similar to (8.3.13c). For autocorrelation  $\mathcal{Y}$  is not as simple as that in Example 8.3.2, but it does not contain any information about the *OFF*-time. It pays for us to look at

$$\wp \mathcal{Y} = \frac{1}{1 + \lambda_p \bar{t}_1} [\mathbf{p}_1 [(\mathbf{V}_1 + \bar{t}_2 \mathbf{I}_1)(\mathbf{I}_1 - \mathbf{D}_1)], (\Psi[\mathbf{V}_1 \mathbf{D}_1] + d \bar{t}_2) \mathbf{p}_2].$$

This also has no dependence on the *OFF*-time distribution except for its mean.  $\mathcal{Y} \boldsymbol{\epsilon}'$  does contain  $\mathbf{V}_2$ , but  $\mathcal{Y} \mathcal{Y} \boldsymbol{\epsilon}'$  does not, as displayed below.

$$\mathcal{Y} \mathcal{Y} \boldsymbol{\epsilon}' = \begin{bmatrix} [(\mathbf{I}_1 - \mathbf{D}_1) \mathbf{V}_1 \mathbf{D}_1 + \Psi[\mathbf{V}_1 \mathbf{D}_1] + \bar{t}_2 \mathbf{D}_1] \boldsymbol{\epsilon}'_1 \\ [\Psi[(\mathbf{I}_1 - \mathbf{D}_1) \mathbf{V}_1 \mathbf{D}_1] + d \Psi[\mathbf{V}_1 \mathbf{D}_1] + \bar{t}_2 d] \boldsymbol{\epsilon}'_2 \end{bmatrix}.$$

the autocorrelation lag- $k$  depends on  $(\wp \mathcal{Y}) \mathcal{Y}^{k-2} (\mathcal{Y} \mathcal{Y} \boldsymbol{\epsilon}')$ , and given that none of the bracketed terms depends on  $\mathbf{V}_2$ , it therefore follows that the MAOOMMPP does not depend on the *OFF*-time distribution, similar to the unmodified process in Theorem 8.3.2.

We have seen that although the OOMMPP and MAOOMMPP processes have similar behavior, they are not equivalent. After all, they do have different mean interdeparture times. Furthermore, the OOMMPP is an MMPP, but the MAOOMMPP is not. In the next subsection we quickly look at another variation.

### Alternative Modified Augmented *ON-OFF* Model

This process abbreviates to AMAOOMMPP. (YES, we have been carried away with acronyms, but they do carry some meaning.) We finally present our last variation of Markov modulated processes. Our purpose (besides generating long strings) is to show how physical ideas can be implemented into mathematical models, and also to see that systems that are differently described can still produce physically identical results even though they appear to be different mathematically. In Section 3.4 we showed, through *isometric transformations*, that a single ME distribution can have an infinite number of distinct representations. The same could be true for more general systems.

In the MAOOMMPP model each *ON*-time ends with a packet transmission. An outside observer cannot tell for sure when that *ON*-time began, because the motion of the token is not observable, but the final packet could tell her when it ended if marking of packets were allowed. In this section we look at a different scenario, namely, each *ON*-time begins with a packet transmission. We do this by allowing the token to send a packet whenever it leaves the *OFF* server. Then that packet can be considered to be the first of the next *ON* period for whichever server the token moves to next.

Let  $S_\ell$  be the *OFF* server. In the previous section by setting the matrix block  $\mathcal{B}_{\ell j}$  to 0, we were able to have the token not emit a packet as he left  $S_\ell$ . Now, instead, we want the token not to emit a packet as he enters  $S_\ell$ .



Then  $\mathbf{B}_{j\ell} = \mathbf{0}$  for all  $j$ . This can be done in the following way. In analogy with Equations (8.3.15) we have

$$\begin{aligned}\mathcal{Q} &= \mathbf{B}_o - \mathbf{B}_o \langle P \rangle \\ \mathcal{L} &= \mathcal{L}_o + \mathbf{B}_o \langle P \rangle \mathcal{I}_{oo} \\ \mathbf{B} &= \mathbf{B}_o + \mathcal{L}_o - \mathbf{B}_o \langle P \rangle (\mathcal{I}_o - \mathcal{I}_{oo}).\end{aligned}\tag{8.3.16}$$

We now consider the two-server system where the token emits a packet upon entering the *ON* server (or leaving the *OFF* server).

**Example 8.3.5:** Consider a system as in Example 8.3.4, but now a packet is emitted as the token enters the *ON* state. Then

$$\begin{aligned}\mathcal{L} &= \begin{bmatrix} \lambda_p \mathbf{I}_1 & \mathbf{0} \\ \mathbf{B}_2 \epsilon'_2 \mathbf{p}_1 & \mathbf{0} \end{bmatrix}, \quad \mathcal{Q} = \begin{bmatrix} \mathbf{B}_1 & -\mathbf{B}_1 \epsilon'_1 \mathbf{p}_2 \\ -\mathbf{B}_2 \epsilon'_2 \mathbf{p}_1 & \mathbf{B}_2 \end{bmatrix}, \\ \mathbf{B} &= \begin{bmatrix} \mathbf{B}_1 + \lambda_p \mathbf{I}_1 & -\mathbf{B}_1 \epsilon'_1 \mathbf{p}_2 \\ \mathbf{0} & \mathbf{B}_2 \end{bmatrix}.\end{aligned}$$

It is straightforward to set up  $\mathcal{V}$  and  $\mathcal{Y}$ :

$$\mathcal{V} = \begin{bmatrix} \mathbf{V}_1 \mathbf{D}_1 & \mathbf{D}_1 \epsilon'_1 \mathbf{p}_2 \mathbf{V}_2 \\ \mathbf{0} & \mathbf{V}_2 \end{bmatrix} \quad \text{and} \quad \mathcal{Y} = \begin{bmatrix} \mathbf{I}_1 - \mathbf{D}_1 + \mathbf{D}_1 \epsilon'_1 \mathbf{p}_1 & \mathbf{0} \\ \epsilon'_2 \mathbf{p}_1 & \mathbf{0} \end{bmatrix}.$$

Once again  $\mathcal{Q}$  is the same as before, so we get

$$\boldsymbol{\pi} = \frac{1}{\bar{t}_1 + \bar{t}_2} [\mathbf{p}_1 \mathbf{V}_1, \mathbf{p}_2 \mathbf{V}_2] \quad \text{and} \quad \boldsymbol{\wp} = \frac{1}{1 + \lambda_p \bar{t}_1} [p_1 (\mathbf{I}_1 + \lambda_p \mathbf{V}_1), \mathbf{o}].$$

The flow rate is the same as that for the MAOOMMPP, namely

$$\kappa = \boldsymbol{\pi} \mathcal{L} \boldsymbol{\varepsilon}' = \frac{1}{\boldsymbol{\wp} \mathcal{V} \boldsymbol{\varepsilon}'} = \frac{1 + \lambda_p \bar{t}_1}{\bar{t}_1 + \bar{t}_2}.$$

We can also show that the two processes have the same variance.

So, what have we here? Observe that  $\mathcal{Y}$  and  $\mathcal{L}$  have zeroes in their second columns. It is that property we used to prove Theorems 8.3.1 and 8.3.2 for the pure *ON-OFF* process. Therefore, those theorems must apply here as well.  $\blacktriangle$

We summarize this section with a theorem and a conjecture which is very likely true.

**Theorem 8.3.4:** Markov modulated *ON-OFF* processes, with and without the modifications discussed above, share the following properties.

1. The autocovariance lag- $k$  is independent of the *OFF*-time distribution;

2. The variance for the interdeparture times  $\sigma_X^2$  depends on the *OFF*-time distribution only through its variance  $\sigma_{OFF}^2$ . Therefore, the  $\hat{r}(k)$ s are proportional for all *OFF*-time distributions with the same mean;
3. If the *ON*-time is exponentially distributed, then the process is a renewal process.

**Conjecture:** Although the MAOOMMPP and AMOOAMMPP processes have different representations (Examples 8.3.4 and 8.3.5), they are equivalent.

**Evidence:** They have the same mean and variance; They both satisfy 1. to 3. above, their  $\mathbf{B}$  matrices have the same set of eigenvalues, and therefore there exists an isometric transformation that connects them, their  $\mathbf{Y}$ s have the same rank ( $m_1$ ), they have the same  $\hat{r}(1)$ , previous calculations ([SCHWEFEL00]) of specific models have yielded results that are the same to within calculation error. ■

### 8.3.3 Merging Renewal Processes

It would seem at first thought that the merging of two independent renewal streams would produce a composite stream with zero covariance. Except for Poisson processes, this is not the case. In fact, this is more complicated to describe than the processes of the previous sections. In what follows, we restrict ourselves to two processes, although a generalization to more streams is straightforward.

Visualize an infinite queue feeding into two general servers  $S_1$  and  $S_2$  represented, respectively, by  $\langle \mathbf{p}_j, \mathbf{B}_j \rangle$ ,  $j = 1, 2$ . Then two customers are being served simultaneously and independently. Thus the state space needed must be a *direct product* of the spaces needed to describe each. We use the standard *Kronecker product* representation here (see Chapter 7 and, e.g., [GRAHAM81]). The following are square matrices of dimension  $m_1 m_2$ .

$$\begin{aligned}\hat{\mathbf{B}}_1 &:= \mathbf{B}_1 \otimes \mathbf{I}_2 \\ \hat{\mathbf{B}}_2 &:= \mathbf{I}_1 \otimes \mathbf{B}_2.\end{aligned}\tag{8.3.17}$$

Then the generator matrix for the interdeparture times is

$$\mathbf{B} = \hat{\mathbf{B}}_1 + \hat{\mathbf{B}}_2, \quad \text{and} \tag{8.3.18a}$$

$$\mathbf{V} = [\hat{\mathbf{B}}_1 + \hat{\mathbf{B}}_2]^{-1} = \hat{\mathbf{V}}_1 \hat{\mathbf{V}}_2 [\hat{\mathbf{V}}_1 + \hat{\mathbf{V}}_2]^{-1}.\tag{8.3.18b}$$

From its definition, the  $\mathcal{L}$  matrix is (where  $\mathbf{Q}_j = \mathbf{e}'_j \mathbf{p}_j$ )

$$\mathcal{L} = \hat{\mathbf{B}}_1 \hat{\mathbf{Q}}_1 + \hat{\mathbf{B}}_2 \hat{\mathbf{Q}}_2 \tag{8.3.18c}$$

which certainly satisfies (8.2.1). Then from (8.2.2),

$$\mathbf{Y} = [\hat{\mathbf{B}}_1 + \hat{\mathbf{B}}_2]^{-1} [\hat{\mathbf{B}}_1 \hat{\mathbf{Q}}_1 + \hat{\mathbf{B}}_2 \hat{\mathbf{Q}}_2] = [\hat{\mathbf{V}}_1 + \hat{\mathbf{V}}_2]^{-1} [\hat{\mathbf{V}}_2 \hat{\mathbf{Q}}_1 + \hat{\mathbf{V}}_1 \hat{\mathbf{Q}}_2], \tag{8.3.19a}$$

and by (8.2.8a)

$$\mathbf{Q} = \mathbf{B} - \mathbf{L} = \hat{\mathbf{B}}_1(\mathbf{I} - \hat{\mathbf{Q}}_1) + \hat{\mathbf{B}}_2(\mathbf{I} - \hat{\mathbf{Q}}_2). \quad (8.3.19b)$$

By direct substitution, it can be shown that

$$\boldsymbol{\wp} = \frac{1}{\bar{x}_1 + \bar{x}_2}(\mathbf{p}_1 \otimes \mathbf{p}_2)(\hat{\mathbf{V}}_1 + \hat{\mathbf{V}}_2) \quad (8.3.20a)$$

and satisfies  $\boldsymbol{\wp}\mathcal{Y} = \boldsymbol{\wp}$ . Also, from (8.2.8b)

$$\boldsymbol{\pi} = c\boldsymbol{\wp}\mathcal{Y} = \frac{1}{\bar{x}_1 \bar{x}_2}(\mathbf{p}_1 \otimes \mathbf{p}_2)(\hat{\mathbf{V}}_1 \hat{\mathbf{V}}_2), \quad (8.3.20b)$$

$$\begin{aligned} \mathbb{E}[X] &= \boldsymbol{\wp}[\hat{\mathbf{B}}_1 + \hat{\mathbf{B}}_2]^{-1}\boldsymbol{\epsilon}' = \boldsymbol{\wp}[\hat{\mathbf{V}}_1 + \hat{\mathbf{V}}_2]^{-1}\hat{\mathbf{V}}_1 \hat{\mathbf{V}}_2 \boldsymbol{\epsilon}' \\ &= \frac{\bar{x}_1 \bar{x}_2}{\bar{x}_1 + \bar{x}_2} = \frac{1}{(1/\bar{x}_1 + 1/\bar{x}_2)}. \end{aligned} \quad (8.3.20c)$$

Clearly, the mean arrival rate  $(1/\bar{x})$  is equal to the sum of the arrival rates  $[(1/\bar{x}_1 + 1/\bar{x}_2)]$  of the two streams. This is only true for the steady state, or when many customers have already departed.

These equations are perfectly amenable to numerical computation, but we can get analytical results if  $S_1$  or  $S_2$  is an exponential server (or equivalently, if one of the processes is Poisson). Let  $m_2 = 1$ . Then  $\mathbf{B}_2$  is a scalar, say  $\lambda$ , and we can drop the subscript for  $S_1$ . The product space is now the same as the state-space for  $S_1$ . The above equations become

$$\mathbf{B} = \mathbf{B} + \lambda\mathbf{I}, \quad (8.3.21a)$$

$$\mathbf{L} = \mathbf{B}\mathbf{Q} + \lambda\mathbf{I}, \quad \mathbf{Q} = \mathbf{B} - \mathbf{B}\mathbf{Q}, \quad (8.3.21b)$$

and

$$\mathcal{Y} = [\lambda\mathbf{I} + \mathbf{B}]^{-1}[\lambda\mathbf{I} + \mathbf{B}\mathbf{Q}] = [\mathbf{I} + \lambda\mathbf{V}]^{-1}[\lambda\mathbf{V} + \mathbf{Q}]. \quad (8.3.21c)$$

Then (8.3.20a) becomes

$$\boldsymbol{\wp} = \frac{1}{1 + \lambda\bar{x}}\mathbf{p}(\mathbf{I} + \lambda\mathbf{V}), \quad (8.3.21d)$$

satisfying  $\boldsymbol{\wp}\mathcal{Y} = \boldsymbol{\wp}$  and  $\boldsymbol{\wp}\boldsymbol{\epsilon}' = 1$ . Also, from (8.3.20c), it follows that

$$\mathbb{E}[X] = \frac{\bar{x}}{1 + \lambda\bar{x}}. \quad (8.3.22a)$$

This process can be considered to be an AMMPP with one server where an additional customer departs when the token leaves the server and then returns. (See Section 8.3.2.4 and let  $M = 1$ . See also Theorem 8.3.3.) Define  $\mathbf{D} := [\mathbf{I} + \lambda\mathbf{V}]^{-1}$ . Then the autocovariance lag-1 turns out to be

$$\lambda^2 \text{Cov}(X, X_{+1})$$

$$= \frac{1}{1 + \lambda \bar{x}} [(\Psi[\lambda \mathbf{V} \mathbf{D}])^2 + \Psi[\lambda^3 \mathbf{V}^3 \mathbf{D}^2]] - \left[ \frac{\lambda \bar{x}}{1 + \lambda \bar{x}} \right]^2. \quad (8.3.22b)$$

Recall that  $\lambda \mathbf{V} \mathbf{D} = \mathbf{I} - \mathbf{D}$ , so from (4.4.1c)

$$\alpha_k(\lambda) := \Psi[(\lambda \mathbf{V} \mathbf{D})^k \mathbf{D}] = \int_0^\infty \frac{(\lambda x)^k}{k!} e^{-\lambda x} f(x) dx. \quad (8.3.22c)$$

The integral clearly shows that  $\alpha_k(\lambda)$  is the probability that there will be  $k$  departures from  $S_2$  between departures from  $S_1$ . For future reference, it is not hard to see (in at least two different ways) that  $\sum_{k=0}^\infty \alpha_k(\lambda) = 1$ .

Finally it can be shown that

$$\lambda^2 \text{Cov}(X, X_{+1}) = \frac{\alpha_0^2 + \alpha_1}{1 + \lambda \bar{x}} - \frac{1}{[1 + \lambda \bar{x}]^2}. \quad (8.3.22d)$$

If  $f(x)$  is exponential (two Poisson processes), then  $\alpha_0 = 1/(1 + \lambda \bar{x})$ ,  $\alpha_1 = \lambda \bar{x}/(1 + \lambda \bar{x})^2$ , and the covariance is 0. Of course in this case, (8.3.21c) clearly shows that  $\mathcal{Y}$  reduces to 1, and all correlations are 0. Thus we reprove the well-known theorem that the merging of Poisson processes is a Poisson process with mean arrival rate equal to the sum of the arrival rates of the individual processes.

Similar expressions can be derived for lag-2 or more, but with increasing difficulty. Note, however, that the last equation does not depend on any ME representation, so it is true for all distributions.

### 8.3.4 Departures from Overloaded Multiprocessor Systems

In Chapter 6 we discussed “generalized X/G/C-type systems”. Such systems can be  $C$  identical servers, or even an arbitrary *Jackson networklike* collection of load-dependent exponential servers, for which only  $C$  customers can be active at once. The other customers must queue up. The matrices needed here are already defined in that chapter. The correspondence is as follows, where all matrices with subscript  $c$  are the *reduced-product space* operators explicitly defined in Chapter 6.

$$\begin{aligned} \mathcal{B} &= \mathbf{B}_c \\ \mathcal{L} &= \mathbf{M}_c \mathbf{Q}_c \mathbf{R}_c \\ \mathcal{Y} &= \mathbf{V}_c \mathbf{M}_c \mathbf{Q}_c \mathbf{R}_c \\ \wp_o &= \mathbf{p} \mathbf{R}_2 \mathbf{R}_3 \cdots \mathbf{R}_c \\ \wp &= \boldsymbol{\pi}_c. \end{aligned} \quad (8.3.23)$$

Imagine a large number of customers waiting to be served with  $C$  of them entering service simultaneously at the start. Then  $\wp_o$  is the initial vector, and many properties, including the mean time to drain the queue, as well as the interdeparture distributions and correlations can be calculated according to the formulas in this chapter.

### 8.3.5 Departures from ME/ME/1 Queues

Consider two servers  $S_1$  and  $S_2$  represented by vector-matrix pairs,  $\langle \mathbf{p}_i, \mathbf{B}_i \rangle$ , with dimension  $m_i$ . Further suppose that there is an infinite queue of customers waiting to be served one at a time by  $S_2$ . As was shown in Section 8.3.1, departures from  $S_2$  constitute a renewal process. After being served, a customer moves “downstream” to  $S_1$ . The behavior of the queue at  $S_1$  constitutes a G/G/1 queue. Assuming that  $\mathbf{E}[T_2] > \mathbf{E}[T_1]$ ,  $S_1$  will sometimes be empty, so the departure process from  $S_1$  (almost always) is not a renewal process. It is this process that we analyze here. The subsystem includes everything upstream from the departure point of  $S_1$ . We must now deal with an infinite state-space, because not only must the states of the customers in service be tracked, but also the length of the queue at  $S_1$ .

#### 8.3.5.1 If $S_2$ Is Exponential (M/ME/1 Queues)

Let  $S_2$  be an exponential server with service rate  $\lambda$ . Then it generates a Poisson arrival process to  $S_1$ , now represented without subscripts by an  $m$ -dimensional vector-matrix pair,  $\langle \mathbf{p}, \mathbf{B} \rangle$ . The full-system matrices in this section (e.g.,  $\mathcal{L}$ ,  $\mathcal{Y}$ ,  $\mathcal{V}$ ,  $\mathcal{B}$ , and  $\wp$ ) have elements that are of different size, because an empty queue is represented by a single state. Thus the  $\{1, 1\}$  element of the  $\mathcal{B}$  matrix below is a scalar, whereas the other elements in row 1 are  $m$ -dimensional row vectors. Similarly, the other elements in the first column are column  $m$ -vectors. All other elements are  $m \times m$  matrices. Analogous conditions hold for row- and column-vectors (e.g.,  $\wp$  and  $\varepsilon'$ ).

$$\mathcal{B} = \begin{bmatrix} \lambda & -\lambda\mathbf{p} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{o}' & \mathbf{B} + \lambda\mathbf{I} & -\lambda\mathbf{I} & \mathbf{0} & \cdots \\ \mathbf{o}' & \mathbf{0} & \mathbf{B} + \lambda\mathbf{I} & -\lambda\mathbf{I} & \cdots \\ \mathbf{o}' & \mathbf{0} & \mathbf{0} & \mathbf{B} + \lambda\mathbf{I} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \quad (8.3.24a)$$

and thus

$$\mathcal{B}\varepsilon' = \begin{bmatrix} 0 \\ \mathbf{B}\varepsilon' \\ \mathbf{B}\varepsilon' \\ \mathbf{B}\varepsilon' \\ \cdots \end{bmatrix}.$$

As in the previous subsections, let  $\mathbf{D} = [\mathbf{I} + \lambda\mathbf{V}]^{-1}$ , then it can be shown that  $\mathcal{B}^{-1}$  is

$$\mathcal{V} = \frac{1}{\lambda} \begin{bmatrix} 1 & \mathbf{p}\lambda\mathbf{V}\mathbf{D} & \mathbf{p}(\lambda\mathbf{V}\mathbf{D})^2 & \mathbf{p}(\lambda\mathbf{V}\mathbf{D})^3 & \cdots \\ \mathbf{o}' & \lambda\mathbf{V}\mathbf{D} & (\lambda\mathbf{V}\mathbf{D})^2 & (\lambda\mathbf{V}\mathbf{D})^3 & \cdots \\ \mathbf{o}' & \mathbf{0} & \lambda\mathbf{V}\mathbf{D} & (\lambda\mathbf{V}\mathbf{D})^2 & \cdots \\ \mathbf{o}' & \mathbf{0} & \mathbf{0} & \lambda\mathbf{V}\mathbf{D} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}. \quad (8.3.24b)$$

It is not hard to see that the departure matrix is

$$\mathcal{L} = \begin{bmatrix} 0 & \mathbf{o} & \mathbf{o} & \mathbf{o} & \cdots \\ \mathbf{B}\boldsymbol{\epsilon}' & \mathbf{O} & \mathbf{O} & \mathbf{O} & \cdots \\ \mathbf{o}' & \mathbf{BQ} & \mathbf{O} & \mathbf{O} & \cdots \\ \mathbf{o}' & \mathbf{O} & \mathbf{BQ} & \mathbf{O} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \quad \text{and} \quad \mathcal{L}\boldsymbol{\epsilon}' = \begin{bmatrix} 0 \\ \mathbf{B}\boldsymbol{\epsilon}' \\ \mathbf{B}\boldsymbol{\epsilon}' \\ \mathbf{B}\boldsymbol{\epsilon}' \\ \cdots \end{bmatrix} = \mathbf{B}\boldsymbol{\epsilon}'. \quad (8.3.25a)$$

The matrix  $\mathcal{Y}$  can now be written down:

$$\mathcal{Y} = \mathcal{V}\mathcal{L} = \begin{bmatrix} \Psi[\mathbf{D}] & \Psi[\lambda\mathbf{VD}^2]\mathbf{p} & \Psi[(\lambda\mathbf{VD})^2\mathbf{D}]\mathbf{p} & \Psi[(\lambda\mathbf{VD})^3\mathbf{D}]\mathbf{p} & \cdots \\ \mathbf{D}\boldsymbol{\epsilon}' & (\lambda\mathbf{VD})\mathbf{DQ} & (\lambda\mathbf{VD})^2\mathbf{DQ} & (\lambda\mathbf{VD})^3\mathbf{DQ} & \cdots \\ \mathbf{o}' & \mathbf{DQ} & (\lambda\mathbf{VD})\mathbf{DQ} & (\lambda\mathbf{VD}^2)\mathbf{DQ} & \cdots \\ \mathbf{o}' & \mathbf{O} & \mathbf{DQ} & (\lambda\mathbf{VD})\mathbf{DQ} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}. \quad (8.3.25b)$$

Using  $\lambda\mathbf{VD} = \mathbf{I} - \mathbf{D}$ , one can show that  $\mathcal{Y}\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}'$  by recognizing that the matrix geometric series sum over  $n$  of  $(\lambda\mathbf{VD})^n$  converges to

$$\begin{aligned} \mathbf{I} + \lambda\mathbf{VD} + (\lambda\mathbf{VD})^2 + \cdots &= [\mathbf{I} - \lambda\mathbf{VD}]^{-1} \\ &= [\mathbf{I} - (\mathbf{I} - \mathbf{D})]^{-1} = \mathbf{D}^{-1} = \mathbf{I} + \lambda\mathbf{V}. \end{aligned}$$

The vector  $\mathcal{V}\boldsymbol{\epsilon}'$  can also be evaluated from the above. We use the utilization parameter  $\rho = \lambda\bar{x}$  and get

$$\lambda\mathcal{V}\boldsymbol{\epsilon}' = \begin{bmatrix} 1 \\ \mathbf{o}' \\ \mathbf{o}' \\ \mathbf{o}' \\ \cdots \end{bmatrix} + \begin{bmatrix} \rho \\ \lambda\mathbf{V}\boldsymbol{\epsilon}' \\ \lambda\mathbf{V}\boldsymbol{\epsilon}' \\ \lambda\mathbf{V}\boldsymbol{\epsilon}' \\ \cdots \end{bmatrix}. \quad (8.3.26)$$

From Theorems 4.2.3 and 4.2.4, the steady-state departure vectors for the M/ME/1 queue are given by

$$\mathbf{d}(n) = (1 - \rho)\Psi[\mathbf{U}^n]\mathbf{p},$$

where from Equations (4.1.4)

$$\mathbf{U}^{-1} := \mathbf{A} := \mathbf{I} + \frac{1}{\lambda}\mathbf{B} - \mathbf{Q}. \quad (8.3.27)$$

Given that  $\sum_{n=0}^{\infty} \mathbf{d}(n) = \mathbf{p}$ , it must be true that  $\sum_{n=0}^{\infty} \Psi[\mathbf{U}^n] = \Psi[(\mathbf{I} - \mathbf{U})^{-1}] = (1 - \rho)^{-1}$ , which it is [see (4.2.3)]. We also mention that  $\Psi[\mathbf{U}] = (1 - \alpha_o)/\alpha_o$ .

The steady-state vector over all states is given by

$$\boldsymbol{\varphi} = (1 - \rho) [1, \Psi[\mathbf{U}]\mathbf{p}, \Psi[\mathbf{U}^2]\mathbf{p}, \Psi[\mathbf{U}^3]\mathbf{p}, \dots]. \quad (8.3.28)$$

where  $\rho\epsilon' = 1$ . With some algebraic manipulation it can be shown that  $\rho\mathcal{Y} = \rho$ .

Note, it is the departure vectors of the M/G/1 queue,  $\mathbf{d}(n)$  (see theorem 4.2.4), not the steady-state vectors  $\pi(n)$ , that make up  $\rho$ , the left eigenvector of  $\mathcal{Y}$ . Also note that if the elements of  $\mathcal{Y}$  are reduced to scalars by pre- and postmultiplying them by appropriately dimensioned  $\mathbf{p}$  (or 1) and  $\epsilon'$  (or 1), the following matrix results [see (8.3.22c)].

$$\bar{\mathcal{Y}} := \begin{bmatrix} \alpha_o & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \cdots \\ \alpha_o & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \cdots \\ 0 & \alpha_o & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\ 0 & 0 & \alpha_o & \alpha_1 & \alpha_2 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}. \quad (8.3.29)$$

Finding the left-eigenvector of this matrix is the standard way one finds the scalar steady-state probabilities, as given in [KLEINROCK75] (see also Section 4.4.3). But its derivation depends on the knowledge that a random observer sees the same probabilities as a departing customer for the M/G/1 queue. This, in turn, is only true because the arrival process to  $S_1$  is Poisson.

Returning to evaluation of the various covariances, we need expressions for  $\mathcal{Y}\mathcal{V}\epsilon'$  and  $\mathcal{Y}\mathcal{Y}\mathcal{V}\epsilon'$ . It does not take too much effort to get them. They are:

$$\lambda\mathcal{Y}\mathcal{V}\epsilon' = \begin{bmatrix} \alpha_o \\ \mathbf{D}\epsilon' \\ \mathbf{o}' \\ \mathbf{o}' \\ \cdots \end{bmatrix} + \rho\epsilon' \quad (8.3.30a)$$

and

$$\lambda\mathcal{Y}\mathcal{Y}\mathcal{V}\epsilon' = \begin{bmatrix} \alpha_o + \alpha_1 \\ (\lambda\mathbf{V}\mathbf{D})\mathbf{D}\epsilon' \\ \mathbf{o}' \\ \mathbf{o}' \\ \cdots \end{bmatrix} + \rho \begin{bmatrix} 1 \\ \mathbf{o}' \\ \mathbf{o}' \\ \mathbf{o}' \\ \cdots \end{bmatrix} + \rho \begin{bmatrix} \rho \\ \lambda\mathbf{V}\epsilon' \\ \lambda\mathbf{V}\epsilon' \\ \lambda\mathbf{V}\epsilon' \\ \cdots \end{bmatrix}. \quad (8.3.30b)$$

Three different initial conditions are presented here for  $\rho_o$ . They are:

1. The process starts with an empty queue (designated by subscript “a”),
2. The process starts with the arrival of a customer to an empty queue (subscript “b”),
3. The process starts in its steady state (8.3.28) (no subscript).

The first two are:

$$\rho_a = [1, \mathbf{o}, \mathbf{o}, \mathbf{o}, \dots] \quad (8.3.31a)$$

and

$$\rho_b = [0, \mathbf{p}, \mathbf{o}, \mathbf{o}, \dots]. \quad (8.3.31b)$$





and

$$\mathbf{v}\boldsymbol{\varepsilon}' = \frac{1}{\lambda}\boldsymbol{\varepsilon}' + \begin{bmatrix} \mathbf{V}\boldsymbol{\varepsilon}' \\ \mathbf{o}' \\ \mathbf{o}' \\ \mathbf{o}' \\ \dots \end{bmatrix}, \quad (8.3.32d)$$

where  $\mathbf{D} := [\mathbf{I} + \lambda\mathbf{V}]^{-1}$  and  $d := \alpha_o = \Psi[\mathbf{D}]$  [see (8.3.22c)]. We also have occasion once again to use  $\lambda\mathbf{V}\mathbf{D} = \mathbf{I} - \mathbf{D}$ . It is not hard to see that the departure matrix is

$$\mathcal{L} = \begin{bmatrix} \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{O} & \dots \\ \lambda\mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} & \dots \\ \mathbf{O} & \lambda\mathbf{I} & \mathbf{O} & \mathbf{O} & \dots \\ \mathbf{O} & \mathbf{O} & \lambda\mathbf{I} & \mathbf{O} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix} \quad \text{and} \quad \mathcal{L}\boldsymbol{\varepsilon}' = \lambda \begin{bmatrix} \mathbf{o}' \\ \boldsymbol{\varepsilon}' \\ \boldsymbol{\varepsilon}' \\ \boldsymbol{\varepsilon}' \\ \dots \end{bmatrix} = \mathcal{B}\boldsymbol{\varepsilon}'. \quad (8.3.33a)$$

We see, then, that (8.2.1) is satisfied. We can calculate  $\mathcal{Y}$  and, with some effort, we can also show that it is isometric,

$$\mathcal{Y} = \mathcal{V}\mathcal{L} = \lambda \begin{bmatrix} \mathbf{QVD} & d\mathbf{QVD} & d^2\mathbf{QVD} & d^3\mathbf{QVD} & \dots \\ \mathbf{VD} & d\mathbf{QVD} & d^2\mathbf{QVD} & d^3\mathbf{QVD} & \dots \\ \mathbf{O} & \mathbf{VD} & d\mathbf{QVD} & d^2\mathbf{QVD} & \dots \\ \mathbf{O} & \mathbf{O} & \mathbf{VD} & d\mathbf{QVD} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix} \quad (8.3.33b)$$

and  $\mathcal{Y}\boldsymbol{\varepsilon}' = \boldsymbol{\varepsilon}'$ . To fully define the process, the initial vector  $\boldsymbol{\wp}_o$  must be specified. We present three interesting options here. First, we can imagine the process beginning immediately after an arrival to an empty queue (the beginning of a busy period). Then

$$\boldsymbol{\wp}_a := [\mathbf{o}, \mathbf{p}, \mathbf{o}, \mathbf{o}, \dots]. \quad (8.3.34)$$

Each element is itself a vector of dimension  $m$ .

A second interesting case occurs at the end of a busy period, that is, when a customer leaves an empty queue behind. Consider the matrix  $\mathbf{A}$ , defined in (8.3.27), and let  $s$  be its smallest eigenvalue between 0 and 1, with left eigenvector  $\hat{\mathbf{u}}$  (i.e.,  $\hat{\mathbf{u}}\mathbf{A} = s\hat{\mathbf{u}}$ ). From Corollary 5.1.2 we know that the arrival process (at  $S_1$ ) is in vector state

$$\hat{\mathbf{u}} := \lambda\mathbf{pV}[\mathbf{I} + \lambda(1-s)\mathbf{V}]^{-1} \quad (8.3.35)$$

at that moment, so

$$\boldsymbol{\wp}_b := [\hat{\mathbf{u}}, \mathbf{o}, \mathbf{o}, \mathbf{o}, \dots]. \quad (8.3.36)$$

Pictorially,  $S_2$  is in state  $i$  with probability  $[\hat{\mathbf{u}}]_i$  at the moment a customer leaves an empty queue behind at  $S_1$ . The requirement that  $\hat{\mathbf{u}}\boldsymbol{\varepsilon}' = 1$  is equivalent to requiring that  $s$  satisfy the equation

$$s = F^*(\lambda(1 - s)).$$

The geometric parameter for the steady-state G/M/1 queue  $s$  is the smallest root between 0 and 1 that satisfies the above.

The most important example is the steady-state vector. Again from Theorem 5.1.3 we know that the steady-state vector probability of having  $k$  customers at  $S_1$  at the time of a departure is

$$\mathbf{d}(k) = (1 - s)s^k \hat{\mathbf{u}}.$$

Therefore, the infinite steady-state vector over all queue lengths is

$$\boldsymbol{\wp} = (1 - s)[\hat{\mathbf{u}}, s\hat{\mathbf{u}}, s^2\hat{\mathbf{u}}, s^3\hat{\mathbf{u}}, \dots]. \quad (8.3.37)$$

One can show by direct calculation that this  $\boldsymbol{\wp}$  satisfies (8.2.7) ( $\boldsymbol{\wp}\mathbf{V} = \boldsymbol{\wp}$ ), as it must. Of course, all three vectors have “length” 1; that is,

$$\boldsymbol{\wp}_a \boldsymbol{\epsilon}' = \boldsymbol{\wp}_b \boldsymbol{\epsilon}' = \boldsymbol{\wp} \boldsymbol{\epsilon}' = 1.$$

In order to calculate the covariance for each of the three cases, we need:

$$\boldsymbol{\wp}_x [\mathbf{V}] \boldsymbol{\epsilon}', \quad \boldsymbol{\wp}_x [\mathbf{V}\mathbf{V}] \boldsymbol{\epsilon}', \quad \text{and} \quad \boldsymbol{\wp}_x [\mathbf{V}\mathbf{V}\mathbf{V}] \boldsymbol{\epsilon}',$$

where  $\mathbf{x} = \mathbf{a}$ ,  $\mathbf{b}$ , and *blank*. This is easiest done by first setting up  $[\mathbf{V}\boldsymbol{\epsilon}']$ ,  $[\mathbf{V}\mathbf{V}\boldsymbol{\epsilon}']$ , and  $[\mathbf{V}\mathbf{V}\mathbf{V}\boldsymbol{\epsilon}']$ . We already know  $[\mathbf{V}\boldsymbol{\epsilon}']$  from (8.3.32d). The second term is

$$\mathbf{V}\mathbf{V}\boldsymbol{\epsilon}' = \mathbf{V}[\mathbf{V}\boldsymbol{\epsilon}'] = \frac{1}{\lambda}\mathbf{V}\boldsymbol{\epsilon}' + \mathbf{V} \begin{bmatrix} \mathbf{V}\boldsymbol{\epsilon}' \\ \mathbf{o}' \\ \mathbf{o}' \\ \mathbf{o}' \\ \dots \end{bmatrix} = \frac{1}{\lambda}\boldsymbol{\epsilon}' + \lambda \begin{bmatrix} \Psi[\mathbf{V}^2 \mathbf{D}] \boldsymbol{\epsilon}' \\ \mathbf{V}^2 \mathbf{D}\boldsymbol{\epsilon}' \\ \mathbf{o}' \\ \mathbf{o}' \\ \dots \end{bmatrix}. \quad (8.3.38a)$$

The third term can be evaluated in a similar fashion

$$\begin{aligned} \mathbf{V}\mathbf{V}\mathbf{V}\boldsymbol{\epsilon}' &= \mathbf{V}[\mathbf{V}\mathbf{V}\boldsymbol{\epsilon}'] = \frac{1}{\lambda}\mathbf{V}\boldsymbol{\epsilon}' + \lambda \mathbf{V} \begin{bmatrix} \Psi[\mathbf{V}^2 \mathbf{D}] \boldsymbol{\epsilon}' \\ \mathbf{V}^2 \mathbf{D}\boldsymbol{\epsilon}' \\ \mathbf{o}' \\ \mathbf{o}' \\ \dots \end{bmatrix} \\ &= \frac{1}{\lambda^2}\boldsymbol{\epsilon}' + \frac{1}{\lambda} \begin{bmatrix} \mathbf{V}\boldsymbol{\epsilon}' \\ \mathbf{o}' \\ \mathbf{o}' \\ \mathbf{o}' \\ \dots \end{bmatrix} + \lambda \begin{bmatrix} \Psi[\mathbf{V}^2 \mathbf{D}] \mathbf{V}\boldsymbol{\epsilon}' + \Psi[\mathbf{V}^3 \mathbf{D}^2] \boldsymbol{\epsilon}' \\ \mathbf{V}^3 \mathbf{D}^2 \boldsymbol{\epsilon}' \\ \mathbf{o}' \\ \mathbf{o}' \\ \dots \end{bmatrix}. \end{aligned} \quad (8.3.38b)$$

Next, define the random variables  $X_{an}$ ,  $X_{bn}$ , and  $X_n$ , where  $n = 1, 2, \dots$ . Then combining (8.3.34), (8.3.36), and (8.3.37) with (8.3.32d), we get for the departure time of the first customer:

$$\begin{aligned}\lambda \mathbb{E}[X_{a1}] &= 1 \\ \lambda \mathbb{E}[X_{b1}] &= 1 + \lambda (\hat{\mathbf{u}} \mathbf{V} \boldsymbol{\epsilon}') \\ \lambda \mathbb{E}[X_1] &= 1 + \lambda (1 - s)(\hat{\mathbf{u}} \mathbf{V} \boldsymbol{\epsilon}').\end{aligned}\tag{8.3.39a}$$

Next, combining (8.3.34) and (8.3.36) with (8.3.38a), and recalling that  $\boldsymbol{\wp} \mathbf{Y} = \boldsymbol{\wp}$  [from (8.2.7)], the mean interdeparture time for the second customer is

$$\begin{aligned}\lambda \mathbb{E}[X_{a2}] &= 1 + \lambda^2 \Psi[\mathbf{V}^2 \mathbf{D}] \\ \lambda \mathbb{E}[X_{b2}] &= \lambda \mathbb{E}[X_{a2}] \\ \lambda \mathbb{E}[X_2] &= \lambda \mathbb{E}[X_1].\end{aligned}\tag{8.3.39b}$$

Finally, the same three equations are combined with (8.3.38b) to get the double expectations

$$\begin{aligned}\lambda^2 \mathbb{E}[X_{a1} X_{a2}] &= 1 + \lambda^3 \Psi[\mathbf{V}^3 \mathbf{D}^2] \\ \lambda^2 \mathbb{E}[X_{b1} X_{b2}] &= \lambda^2 \mathbb{E}[X_{a1} X_{a2}] + \lambda^2 (\hat{\mathbf{u}} \mathbf{V} \boldsymbol{\epsilon}') \mathbb{E}[X_{a2}] \\ \lambda^2 \mathbb{E}[X_1 X_2] &= s + (1-s)\lambda^2 \mathbb{E}[X_{b1} X_{b2}] + \lambda^3 s(1-s) (\hat{\mathbf{u}} \mathbf{V}^3 \mathbf{D}^2 \boldsymbol{\epsilon}').\end{aligned}\tag{8.3.39c}$$

(See also Section 4.4.1). The matrix terms  $\Psi[\mathbf{V}^2 \mathbf{D}]$ ,  $\Psi[\mathbf{V}^3 \mathbf{D}^2]$ ,  $(\hat{\mathbf{u}} \mathbf{V} \boldsymbol{\epsilon}')$  and  $(\hat{\mathbf{u}} \mathbf{V}^3 \mathbf{D}^2 \boldsymbol{\epsilon}')$  can be written as nonmatrix expressions by algebraic manipulation of  $\lambda \mathbf{V} \mathbf{D} = \mathbf{I} - \mathbf{D}$ ,

$$\lambda \Psi[\mathbf{V}^2 \mathbf{D}] = \lambda \Psi[\mathbf{V}(\mathbf{I} - \mathbf{D})] = \lambda \bar{x} - \Psi[\mathbf{I} - \mathbf{D}] = \lambda \bar{x} + \alpha_o - 1.$$

Similarly,

$$\lambda^3 \Psi[\mathbf{V}^3 \mathbf{D}^2] = \lambda \bar{x} + 2\alpha_o + \alpha_1 - 2.$$

The  $\hat{\mathbf{u}}$  terms are more difficult, but straightforward when using  $\hat{\mathbf{u}}[\mathbf{I} + \lambda(1-s)\mathbf{V}] = \lambda \mathbf{pV}$ , from (8.3.35). Algebraic manipulation yields

$$\begin{aligned}s \hat{\mathbf{u}} \mathbf{D} &= \lambda \mathbf{pV} \mathbf{D} - (1-s)\hat{\mathbf{u}} \\ \lambda(1-s)\hat{\mathbf{u}} \mathbf{V} &= \lambda \mathbf{pV} - \hat{\mathbf{u}} \\ \lambda s \hat{\mathbf{u}} \mathbf{V} \mathbf{D} &= \hat{\mathbf{u}} - \lambda \mathbf{pV} \mathbf{D}.\end{aligned}$$

From these, on multiplying from the right with any powers of  $\mathbf{V}$  and  $\mathbf{D}$ , and then with  $\boldsymbol{\epsilon}'$ , all terms of the form  $[\hat{\mathbf{u}} \mathbf{V}^k \mathbf{D}^j \boldsymbol{\epsilon}']$  can be expressed. In particular,

$$\lambda(1-s)[\hat{\mathbf{u}} \mathbf{V} \boldsymbol{\epsilon}'] = \lambda \bar{x} - 1$$

and

$$\lambda^3 s^2(1-s)[\hat{\mathbf{u}} \mathbf{V}^3 \mathbf{D}^2 \boldsymbol{\epsilon}'] = \lambda \bar{x} s^2 - 1 + (1 - \alpha_o)(1-s)(1+2s) - s(1-s)\alpha_1.$$

After all of the above are put into (8.2.11a) with  $n = k = 1$ , we get for the covariance lag-1



Finally, we have

$$\mathbf{Y} = \mathbf{V}\mathcal{L}$$

$$= \begin{bmatrix} \hat{\mathbf{p}}_1 \hat{\mathbf{B}}_1 \hat{\mathbf{Q}}_2 \mathbf{D} \hat{\boldsymbol{\epsilon}}_1' & \hat{\mathbf{p}}_1 \hat{\mathbf{B}}_1 \hat{\mathbf{Q}}_2 \mathbf{D} \mathbf{X} \hat{\mathbf{Q}}_1 & \hat{\mathbf{p}}_1 \hat{\mathbf{B}}_1 \hat{\mathbf{Q}}_2 \mathbf{D} \mathbf{X}^2 \hat{\mathbf{Q}}_1 & \cdots \\ \hat{\mathbf{B}}_1 \mathbf{D} \hat{\boldsymbol{\epsilon}}_1' & \hat{\mathbf{B}}_1 \mathbf{D} \mathbf{X} \hat{\mathbf{Q}}_1 & \hat{\mathbf{B}}_1 \mathbf{D} \mathbf{X}^2 \hat{\mathbf{Q}}_1 & \cdots \\ \mathbf{O} & \hat{\mathbf{B}}_1 \mathbf{D} \hat{\mathbf{Q}}_1 & \hat{\mathbf{B}}_1 \mathbf{D} \mathbf{X} \hat{\mathbf{Q}}_1 & \cdots \\ \mathbf{O} & \mathbf{O} & \hat{\mathbf{B}}_1 \mathbf{D} \hat{\mathbf{Q}}_1 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \quad (8.3.40d)$$

and  $\mathbf{Y}\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}'$ . We have made use of the fact that all subscripted matrices commute with matrices with different subscripts (i.e.,  $\hat{\mathbf{B}}_1 \hat{\mathbf{Q}}_2 = \hat{\mathbf{Q}}_2 \hat{\mathbf{B}}_1$  but  $\hat{\mathbf{B}}_1 \hat{\mathbf{Q}}_1 \neq \hat{\mathbf{Q}}_1 \hat{\mathbf{B}}_1$ ). Also,  $\hat{\mathbf{B}}_1$  and  $\hat{\mathbf{B}}_2$  commute with  $\mathbf{D}$ , but  $\hat{\mathbf{Q}}_1$  and  $\hat{\mathbf{Q}}_2$  do not.

Note that  $\mathbf{Y}$  in (8.3.40d) is a stochastic matrix (if its elements are all nonnegative) which satisfies the canonical form described as “M/G/1-type” by M. Neuts [NEUTS81]. Thus there exist standard numerical procedures for solving the equation  $\boldsymbol{\rho}\mathbf{Y} = \boldsymbol{\rho}$ . The  $n$ th component of  $\boldsymbol{\rho}$  is itself a vector, the sum of whose elements is the probability  $d(n)$  that a departing customer will leave  $n$  other customers behind at  $S_1$ . Each component must be of the form  $d(n)\mathbf{p}_1 \times \mathbf{v}_2(n)$ , where  $\mathbf{v}_2(n)$  is the vector state of  $S_2$  at the moment of the departure.

### 8.3.5.4 M/M/1//N Queues

It has long been well known that the departure process from a steady-state open M/M/1 queue is itself a Poisson process. We give here a simple demonstration of why this is so. The expressions also show that this is an exceptional property, and that in general, except for the steady-state M/M/1 queue (see [DISNEYKIESSLER87] for minor exceptions), there *is* correlation. This includes departures from servers in closed systems and also finite buffered queues (i.e., departures from queued servers are not generally renewal processes).

Let  $\lambda$  be the arrival rate of the Poisson process to an exponential server whose rate is  $\mu = 1/\bar{x}$ . The formulas of the previous sections simplify when the following substitutions are made.  $\mathbf{Q} \rightarrow 1$ ,  $\mathbf{V} \rightarrow 1/\mu$ ,  $\mathbf{B} \rightarrow \mu$ , and  $\mathbf{D} \rightarrow 1/(1 + \lambda/\mu)$ . If the formulas from the M/G/1 section are used, then the utilization factor is  $\rho = \lambda/\mu$ . But if the formulas from the G/M/1 section are used, the roles of  $\lambda$  and  $\mu$  must be interchanged; then  $\rho = \mu/\lambda$ . In either case, the subsystem  $\mathbf{V}$  and  $\mathbf{Y}$  matrices become:

$$\mathbf{V} = \frac{\bar{x}}{1 + \rho} \begin{bmatrix} 1/\alpha & 1 & \alpha & \alpha^2 & \alpha^3 & \cdots \\ 0 & 1 & \alpha & \alpha^2 & \alpha^3 & \cdots \\ 0 & 0 & 1 & \alpha & \alpha^2 & \cdots \\ 0 & 0 & 0 & 1 & \alpha & \cdots \\ 0 & 0 & 0 & 0 & 1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \quad (8.3.41a)$$

and

$$\mathbf{Y} = \frac{1}{1 + \rho} \begin{bmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 & \dots \\ 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 & \dots \\ 0 & 1 & \alpha & \alpha^2 & \alpha^3 & \dots \\ 0 & 0 & 1 & \alpha & \alpha^2 & \dots \\ 0 & 0 & 0 & 1 & \alpha & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix}, \quad (8.3.41b)$$

where  $\alpha := \rho/(1 + \rho)$ . Equations (8.3.25b), (8.3.29) and (8.3.33b) all reduce to this one for the M/M/1 queue. The steady-state probabilities are  $p(n) = (1 - \rho)\rho^n$ , so

$$\boldsymbol{\wp} = (1 - \rho)[1, \rho, \rho^2, \rho^3, \rho^4, \dots]. \quad (8.3.42a)$$

The extraordinary property is that  $\boldsymbol{\wp}$  is a left eigenvector of both  $\mathbf{V}$  and  $\mathbf{Y}$ . That is,

$$\boldsymbol{\wp}\mathbf{Y} = \boldsymbol{\wp}, \quad (8.3.42b)$$

as it should. But it is also true that

$$\boldsymbol{\wp}\mathbf{V} = \frac{1}{\lambda}\boldsymbol{\wp}. \quad (8.3.42c)$$

Clearly, then,  $\boldsymbol{\wp}\mathbf{V}\boldsymbol{\varepsilon}' = 1/\lambda$ . Equation (8.2.11c), with  $\boldsymbol{\wp}$  replacing  $\boldsymbol{\wp}_0$ , simplifies to

$$\mathbf{E}[X_n X_{n+k}] = \boldsymbol{\wp}[\mathbf{V} Y^k \mathbf{V}]\boldsymbol{\varepsilon}' = \frac{1}{\lambda}\boldsymbol{\wp}[\mathbf{Y}^k \mathbf{V}]\boldsymbol{\varepsilon}' = \frac{1}{\lambda}\boldsymbol{\wp}[\mathbf{V}]\boldsymbol{\varepsilon}' = \frac{1}{\lambda^2} \quad (8.3.43)$$

for all  $n$  and all  $k$ . This, together with (8.2.10d) and (8.2.11a) shows that the autocovariance is 0 for all  $n$  and all  $k$ . Be reminded, though, that this assumes the subsystem to be in its steady state initially. If the initial vector is not  $\boldsymbol{\wp}$ , then all bets are off. For the finite customer (M/M/1//N) and finite buffer (M/M/1/N) queues, the last column of  $\mathbf{V}$  does not fit the pattern for the other elements, so (8.3.42c) is not satisfied. Thus only the steady-state, open M/M/1 queue yields a Poisson departure process.

## 8.4 MRP/M/1 Queues

All the examples given can be used as arrival processes to a queueing system. We discuss how to do this here, where the queue feeds to an exponential server. The general method is also applicable to M/G/1 (Chapter 4), G/M/1 queues (Chapter 5), and with some extension, generalized M/G/C and G/G/1 queues (Chapters 6 and 7). Whereas previously we were able to find explicit solutions, now we must find the correct solution by iteration. The method depends on a very powerful theorem by Wallace [WALLACE69] on **QBD processes** of which all of these are special cases. Recall that **Birth-Death** Processes are those for which the population grows and contracts by single steps (arrivals

and departures). For QBD processes the steps are multistate sets, exactly as we have been dealing with here.

Let  $\pi(n)$  be the steady-state vector probability that the system is in vector state  $\{i, n\}$  and  $r(n) = \pi(n)\epsilon'$  is the associated scalar probability. The theorem states that if the matrices that govern the transitions are independent of the population  $n$ , then

$$\pi(n) = c\mathbf{u}\mathbf{R}^n \quad \text{and} \quad r(n) = \pi(n)\epsilon',$$

where  $\mathbf{R}$  is a matrix satisfying some *matrix quadratic equation*,  $\mathbf{u}$  is a special vector with  $\mathbf{u}\epsilon' = 1$ , and  $c$  is determined by the normalization condition,  $\sum_{n=0}^{\infty} r(n) = 1$ .

We next consider queueing systems where the arrivals to an exponential server are generated by some MRP satisfying the rules defined in this chapter. By “system” we mean the combination of the arrival process, the exponential server, and the customers in the queue.

#### 8.4.1 Balance Equations

Let  $n$  be the number of customers at an exponential server (called  $S_\nu$ ) with service rate  $\nu$ . The arrival process is described by the matrices  $\mathbf{B}$ ,  $\mathbf{Q}$ , and  $\mathbf{L}$ , as defined previously. The  $i$ th component of the ss vector,  $\pi_i(n)$ , refers to the state the MRP is in when there are  $n$  customers at  $S_\nu$ . This is a straightforward generalization of the description we gave in Chapter 5 from the G/M/1 queue. The system can leave state  $\{i; n\}$  by either a change at the MRP  $[\pi_i(n)(\mathbf{M})_{ii}]$  or a customer completion at  $S_\nu$   $[\pi_i(n)\nu]$ . The system can enter this state by one of three ways:

1. A change of state from some  $j$  to  $i$  in the arrival process  $[\pi_j(n)(\mathbf{M})_{jj}(\mathbf{P})_{ji}]$ ,
2. A customer completion at  $S_\nu$  when there are  $n+1$  customers there  $[\pi_i(n+1)\nu]$ ,
3. The MRP has a departure when there are  $n-1$  customers at  $S_\nu$   $[\pi_j(n-1)(\mathbf{L})_{ji}]$ .

By summing over all intermediate subscripts we get the vector balance equations:

$$\pi(n)(\mathbf{M} + \mu\mathbf{I}) = \pi(n)\mathbf{M}\mathbf{P} + \pi(n+1)\nu + \pi(n-1)\mathbf{L}.$$

Making use of the relation,  $\mathbf{B} = \mathbf{M} - \mathbf{M}\mathbf{P}$  we get for  $n \geq 1$ ,

$$\pi(n+1)\nu - \pi(n)(\mathbf{B} + \nu\mathbf{I}) + \pi(n-1)\mathbf{L} = 0. \quad (8.4.1a)$$

[Compare with (4.1.3d).] For  $n = 0$  there is no possibility for a customer to complete service, so instead we have

$$\pi(1)\nu = \pi(0)\mathbf{B}. \quad (8.4.1b)$$

We now substitute  $\pi(n) = \pi\mathbf{R}^n$  into (8.4.1a), but  $\pi$  and  $\mathbf{R}$  are yet to be determined. For  $n > 1$ ,

$$\pi(1) [\nu\mathbf{R}^{n+1} - \mathbf{R}^n(\mathbf{B} + \nu\mathbf{I}) + \mathbf{R}^{n-1}\mathbf{L}] = 0.$$

Because this must be true for all  $n > 1$  and  $\pi(1)$  cannot be 0, the expression in square brackets must be 0. Therefore

$$\mathcal{R}^{n-1} [\nu \mathcal{R}^2 - \mathcal{R}(\mathcal{B} + \nu \mathcal{I}) + \mathcal{L}] = 0.$$

Again, if  $\mathcal{R}$  has an inverse (something that is not always true, as we show below) then the expression in square brackets must be 0. Thus

$$\nu \mathcal{R}^2 - \mathcal{R}(\mathcal{B} + \nu \mathcal{I}) + \mathcal{L} = 0. \quad (8.4.2a)$$

This equation doesn't hold for  $n = 1$ , so we must go back to (8.4.1a), using  $\pi(2) = \pi(1)\mathcal{R}$ ,  $\mathcal{Y} = \nu \mathcal{L}$ , and (8.4.1b) to get

$$\pi(1) [\nu \mathcal{R} - \mathcal{B} - \nu \mathcal{I} + \nu \mathcal{Y}] = 0. \quad (8.4.2b)$$

Ah, if only we could argue that the expression in square brackets is zero, we would have an explicit expression for  $\mathcal{R}$ . But it is, instead, an eigenvector equation for  $\pi(1)$  (once we know what  $\mathcal{R}$  is). A necessary and sufficient condition that  $\mathcal{R} = \mathcal{I} + \mathcal{B}/\nu - \mathcal{Y}$  satisfy (8.4.2a) is that  $\mathcal{Y}^2 = \mathcal{Y}$ . From (8.2.12c) this condition leads to  $\text{Cov}(X, X_{+k}) = \text{constant}$ , independent of  $k$ . Furthermore, all the eigenvalues of  $\mathcal{Y}$  must be either 0 or 1, the number of unit eigenvalues being equal to the rank of  $\mathcal{Y}$ . The only processes of interest to us that have these properties are the renewal processes, where the covariance equals 0 for all  $k$  and  $\mathcal{Y} = \epsilon' \rho$  has one unit eigenvalue. In fact this is exactly what we used in Chapters 4 and 5. But that does not work here, for we are now interested in the more general MRPs. We discuss this in Section 8.4.3.

Equation (8.4.2a) is the defining equation for  $\mathcal{R}$ , but it is not that easy to solve. First we search for some other properties. We multiply this equation from the right by  $\epsilon'$  and note that  $\mathcal{L}\epsilon' = \mathcal{B}\epsilon'$  to get:

$$\nu(\mathcal{R} - \mathcal{I})\mathcal{R}\epsilon' = (\mathcal{R} - \mathcal{I})\mathcal{B}\epsilon'.$$

But  $(\mathcal{R} - \mathcal{I})$  must have an inverse unless at least one of the eigenvalues of  $\mathcal{R}$  equals 1. This happens when the arrival rate ( $\kappa = \pi\mathcal{L}\epsilon'$ ) equals  $\nu$ , in which case the system is unstable and there is no steady-state solution. Otherwise, a unit eigenvalue implies decomposability, a property which we assume has been removed *a priori*. So, assuming that  $(\mathcal{I} - \mathcal{R})^{-1}$  exists, we get

$$\nu \mathcal{R}\epsilon' = \mathcal{B}\epsilon'. \quad (8.4.3)$$

This relation also satisfies (8.4.2b).

The  $\pi$  vectors must still satisfy the normalization property  $\sum r(n) = 1$ . But more than that, we can assume that if the MRP is observed without reference to queue length it must be found in state  $i$  with the same probability as the residual vector. That is,

$$\sum_{n=0}^{\infty} \pi(n) = \pi,$$



where  $\boldsymbol{\pi}$  (no argument) is defined by (8.2.8b), namely,  $\boldsymbol{\pi}\mathbf{Q} = \mathbf{o}$ , and  $\boldsymbol{\pi}\boldsymbol{\varepsilon}' = 1$ . Thus [note that  $\boldsymbol{\pi}(0)$  may not be of the same form as the other vector probabilities]

$$\boldsymbol{\pi} = \sum_{n=1}^{\infty} \boldsymbol{\pi}(1)\mathbf{R}^{n-1} + \boldsymbol{\pi}(0) = \boldsymbol{\pi}(1) [(\mathbf{I} - \mathbf{R})^{-1} + \nu\mathbf{V}]$$

or

$$\boldsymbol{\pi}(1)[\mathbf{I} + \nu\mathbf{V}(\mathbf{I} - \mathbf{R})] = \boldsymbol{\pi}(\mathbf{I} - \mathbf{R}). \quad (8.4.4)$$

But equations (8.4.2a), (8.4.2b), (8.4.3), and (8.4.4) together are not sufficient to uniquely determine the vector-matrix pair  $\langle \mathbf{u}, \mathbf{R} \rangle$ . In fact, there may be multiple distinct solutions, all of which produce the same queue-length probabilities  $\boldsymbol{\pi}(n)$ .

At this point, following [MEIER-FISCHER92], we assume that

$$\mathbf{c}\mathbf{u} = \boldsymbol{\pi}(\mathbf{I} - \mathbf{R}).$$

From this we have

$$\boldsymbol{\pi}(n) = \boldsymbol{\pi}(\mathbf{I} - \mathbf{R})\mathbf{R}^n \quad \text{and} \quad r(n) = \boldsymbol{\pi}(\mathbf{I} - \mathbf{R})\mathbf{R}^n\boldsymbol{\varepsilon}'. \quad (8.4.5)$$

This equation clearly satisfies  $\sum \boldsymbol{\pi}(n) = \boldsymbol{\pi}$ , which then implies that  $\sum r(n) = 1$ . But, we still don't know how to solve for  $\mathbf{R}$ .

A standard procedure for finding  $\mathbf{R}$  follows.

**Algorithm 8.4.1:** First rewrite (8.4.2a) as

$$\mathbf{R} = \nu\mathbf{R}^2\mathbf{D} + \mathbf{L}\mathbf{D},$$

where  $\mathbf{D} := (\nu\mathbf{I} + \mathbf{B})^{-1}$ . Consider this to be a formula for *fixed point iteration*. That is, let  $\mathbf{R}_0 = \mathbf{0}$  and

$$\mathbf{R}_{\ell+1} = \nu(\mathbf{R}_{\ell})^2\mathbf{D} + \mathbf{L}\mathbf{D}, \quad \text{for } \ell \geq 0.$$

Iterate on  $\ell$  until  $(\mathbf{R}_{\ell+1} - \mathbf{R}_{\ell})$  is “sufficiently small” by some pre-established criterion.

This procedure is guaranteed to converge if the MRP was constructed from PH representations, but may not converge otherwise. Nonconvergence does not mean there is no solution, just that another method, or a different  $\mathbf{R}_0$ , must be chosen. Furthermore, this is not the unique solution to (8.4.2a). Given that this is a quadratic equation, one might expect to find two independent solutions. But this is a *matrix quadratic* equation, for which the number of independent solutions is given by

$$\binom{2M}{M}, \quad \text{where } M = \text{Dim}(\mathbf{R}).$$

We can say that the algorithm produces an  $\mathbf{R}$  whose eigenvalues are all less than 1 in magnitude, otherwise the algorithm would not converge. For more

information, see [LATOUCHE-RAM99] or [NEUTS89]. We give an example of this ambiguity when we look at the G/M/1 queue from the point-of-view of this chapter.

Before showing how to calculate various performance measures we prove the following.

**Theorem 8.4.1:** The matrix  $\mathcal{R}$ , as found by the iterative method described above, must be of the form

$$\mathcal{R} = \mathcal{L}\mathcal{X}, \quad (8.4.6)$$

and has at most the same rank as  $\mathcal{L}$ . Therefore if  $\mathcal{L}$  has no inverse, then  $\mathcal{R}$  has no inverse. But there may be another solution of (8.4.2a) that is invertible. ■

**Proof:** Observe that  $\mathcal{R}_1 = \mathcal{L}\mathcal{D}$ . Next assume that  $\mathcal{R}_k = \mathcal{L}\mathcal{X}_k$  for  $k \leq \ell$  where  $\mathcal{X}_\ell$  follows from the recursive formula. Then

$$\mathcal{R}_{\ell+1} = \nu \mathcal{L}\mathcal{X}_\ell \mathcal{L}\mathcal{X}_\ell \mathcal{D} + \mathcal{L}\mathcal{D} = \mathcal{L}[\nu \mathcal{X}_\ell \mathcal{L}\mathcal{X}_\ell \mathcal{D} + \mathcal{D}] = \mathcal{L}\mathcal{X}_{\ell+1}.$$

Therefore,  $\mathcal{R}_\ell = \mathcal{L}\mathcal{X}_\ell$  for all  $\ell$ . Furthermore, the limit (if it exists) is

$$\mathcal{R} = \lim_{\ell \rightarrow \infty} \mathcal{L}\mathcal{X}_\ell = \mathcal{L}\mathcal{X},$$

where

$$\mathcal{X} := \lim_{\ell \rightarrow \infty} \mathcal{X}_\ell.$$

Therefore,  $\text{Rank}(\mathcal{R}) \leq \text{Rank}(\mathcal{L})$ .

The method described here may not apply to some more general systems, but it does apply to all MRP/M/1 queues.

### 8.4.2 Some Performance Measures

As in previous chapters, given  $\pi(n)$  one can compute the mean queue length, the mean system time, and probability of overflow. We do that now.

Let  $N$  be the r.v. denoting the number of customers queued at  $S_\nu$ ; then

$$\bar{q} := \mathbb{E}[N] = \sum_{n=1}^{\infty} n r(n) = \pi(\mathcal{I} - \mathcal{R}) \left[ \sum_{n=1}^{\infty} n \mathcal{R}^n \right] \epsilon' = \pi \mathcal{R} [\mathcal{I} - \mathcal{R}]^{-1} \epsilon'. \quad (8.4.7a)$$

The *mean system time* is given by Little's formula (1.1.2). This is also called *Mean Cell Delay* (MCD) or *Mean Packet Delay* (MPD) when studying telecommunications traffic. We need  $\kappa$ , the arrival rate of cells to  $S_\nu$ , to use Little's formula. This is given by (8.2.8d).

$$MCD = \frac{\mathbb{E}[N]}{\kappa} = (\wp \mathcal{V} \epsilon') \pi \mathcal{R} [\mathcal{I} - \mathcal{R}]^{-1} \epsilon'. \quad (8.4.7b)$$

Recall from Section 4.2.4 what is meant by *buffer overflow probability* (BOP), namely  $\mathbf{Pr}(N \geq B_s)$ , where  $B_s$  is the size of the primary buffer. The

probability that an arriving customer will see  $n$  customers at  $S_\nu$  is needed to find this.  $[\pi(n)]_i$  is the steady-state probability that there are  $n$  cells already at  $S_\nu$  and the MRP is in state  $i$ . Multiplying by  $\mathcal{L}\epsilon'$  gives the probability rate that a new cell will arrive under these conditions. Upon dividing by the overall arrival rate, we get the arrival probability.

$$a(n) = (\wp \mathcal{V} \epsilon') \pi [(\mathcal{I} - \mathcal{R}) \mathcal{R}^n \mathcal{L}] \epsilon'. \quad (8.4.8a)$$

Then the BOP is

$$\begin{aligned} \Pr(N \geq B_s) &= \sum_{n=B_s}^{\infty} a(n) = (\wp \mathcal{V} \epsilon') \pi (\mathcal{I} - \mathcal{R}) \left[ \sum_{n=B_s}^{\infty} \mathcal{R}^n \right] \mathcal{L} \epsilon' \\ &= (\wp \mathcal{V} \epsilon') \pi [\mathcal{R}^{B_s} \mathcal{L}] \epsilon'. \end{aligned} \quad (8.4.8b)$$

For further information of the utility of these formulas see, for instance, [SCHWEFEL00], [SCHWEFEL-LIP01], and [PARK-WILL00].

### 8.4.3 The G/M/1 Queue as an Example

Recall that in Chapters 4 and 5 we solved the M/G/1 and G/M/1 queues by finding the special matrices (replacing  $\lambda$  with  $\nu$ )  $\mathbf{A} = \mathbf{I} + \mathbf{B}/\nu - \mathbf{Q}$  and  $\mathbf{U} = \mathbf{A}^{-1}$  in (4.1.4a), yielding  $r(n) = (1 - \rho) \mathbf{p} \mathbf{U}^n \epsilon'$  for the M/G/1 queue. The G/M/1 queue was more difficult, and the limit in going from the G/M/1// $N$  to the open G/M/1 queue for  $N \rightarrow \infty$  has to be taken very carefully. But we found the solution in Theorem 5.1.3 to be

$$\begin{aligned} \pi(0) &= (1 - \varrho) \frac{\hat{\mathbf{u}} \mathbf{V}}{\hat{\mathbf{u}} \mathbf{V} \epsilon'} \quad \text{and} \\ \pi(k) &= (1 - s) \varrho s^{k-1} \hat{\mathbf{u}}, \end{aligned}$$

where  $\varrho = 1/(\nu \Psi[\mathbf{V}])$  is the utilization parameter of Chapter 5, and

$$\hat{\mathbf{u}} = \nu \mathbf{p} [\nu(1 - s) \mathbf{I} + \mathbf{B}]^{-1}$$

with normalization,  $\hat{\mathbf{u}} \epsilon' = 1$ .  $s$  is the smallest positive root of the equation

$$s = B^*[\nu(1 - s)] = \mathbf{p} \mathbf{B} (\mathbf{B} + \nu(1 - s) \mathbf{I})^{-1} \epsilon'.$$

Of more relevance to us here, we also showed that  $\hat{\mathbf{u}} \mathbf{A} = s \hat{\mathbf{u}}$ , that is,  $s$  is the smallest positive eigenvalue of  $\mathbf{A}$ , with left eigenvector  $\hat{\mathbf{u}}$ . Because of this, the solution could be written in matrix geometric form as

$$\pi(k) = (1 - s) \varrho \hat{\mathbf{u}} \mathbf{A}^{k-1}.$$

Interestingly enough,  $\mathbf{A}$  has eigenvalues that are greater than 1 in magnitude, so  $\mathbf{A}^k$  grows unboundedly large with  $k$ . But  $\hat{\mathbf{u}}$  is orthogonal to all the corresponding eigenvectors, so  $\hat{\mathbf{u}} \mathbf{A}^k = \hat{\mathbf{u}} s^k \rightarrow 0$ . Its relevance here is that  $\mathbf{A}$  satisfies (8.4.2a) with  $\mathcal{L} = \mathbf{B} \mathbf{Q}$ , yet it does not satisfy Theorem 8.4.1. That is,

$$\nu \mathbf{A}^2 - \mathbf{A}(\mathbf{B} + \nu \mathbf{I}) + \mathbf{B} \mathbf{Q} = 0.$$

(Recall that here,  $\mathbf{Q} = \boldsymbol{\epsilon}'\mathbf{p}$ .) But although  $\mathbf{BQ}$  is of rank 1,  $\text{Rank}(\mathbf{A}) = \text{Dim}(\mathbf{A}) > 1$  if the renewal process is not Poisson.

We now use Algorithm 8.4.1 (well, not quite) to find a solution of (8.4.2a). We know from Theorem 8.4.1 that  $\mathbf{R}$  must be of rank 1, because  $\mathbf{BQ}$  is of rank 1. Therefore,  $\mathbf{R}$  must be of the form  $s\mathbf{v}'\mathbf{u}$  where  $\mathbf{u}\mathbf{v}' = 1$ . Its one non-zero eigenvalue is  $s$ . Given that we know its form so precisely, we can substitute it into (8.4.2a) and find what  $s$ ,  $\mathbf{u}$ , and  $\mathbf{v}'$  are. Note that  $(s\mathbf{v}'\mathbf{u})^2 = s^2\mathbf{v}'\mathbf{u}$ , therefore  $\mathbf{R}^2 = s\mathbf{R}$ . (In fact, for any  $n > 0$ ,  $\mathbf{R}^n = s^{n-1}\mathbf{R}$ .) Then

$$\nu\mathbf{R}^2 - \mathbf{R}(\mathbf{B} + \nu\mathbf{I}) + \mathbf{BQ} = \nu s\mathbf{R} - \mathbf{R}(\mathbf{B} + \nu\mathbf{I}) + \mathbf{BQ} = \mathbf{BQ} - \mathbf{R}(\mathbf{B} + \nu(1-s)\mathbf{I}) = 0.$$

Rearranging the terms and multiplying from the right by  $(\mathbf{B} + \nu(1-s)\mathbf{I})^{-1}$  yields

$$\mathbf{R} = \mathbf{B}\boldsymbol{\epsilon}'[\mathbf{p}(\mathbf{B} + \nu(1-s)\mathbf{I})].$$

But if  $s$  is right, the expression in square brackets is precisely what was defined above as  $\hat{\mathbf{u}}/\nu$ . In order to have  $\mathbf{R} = s\mathbf{v}'\hat{\mathbf{u}}$  it must follow that  $\hat{\mathbf{u}}\mathbf{B}\boldsymbol{\epsilon}' = s\nu$ . But this reduces to the expression  $s = \mathbf{pB}(\mathbf{B} + \nu(1-s)\mathbf{I})^{-1}\boldsymbol{\epsilon}'$ , the equation that defined  $s$ . Therefore  $s\nu\mathbf{v}' = \mathbf{B}\boldsymbol{\epsilon}'$ , and because  $\mathbf{B}\boldsymbol{\epsilon}' = \mathcal{L}\boldsymbol{\epsilon}'$ ,

$$\mathbf{R} = \frac{1}{\nu}\mathbf{B}\boldsymbol{\epsilon}'\hat{\mathbf{u}} = \frac{1}{\nu}\mathcal{L}\boldsymbol{\epsilon}'\hat{\mathbf{u}} = \mathbf{A}\boldsymbol{\epsilon}'\hat{\mathbf{u}},$$

thereby explicitly satisfying Theorem 8.4.1. The rightmost expression for  $\mathbf{R}$  explicitly yields the idempotent property for  $\mathbf{R}$ . That is,

$$\mathbf{R}^2 = (\mathbf{A}\boldsymbol{\epsilon}'\hat{\mathbf{u}})(\mathbf{A}\boldsymbol{\epsilon}'\hat{\mathbf{u}}) = \mathbf{A}\boldsymbol{\epsilon}'(\hat{\mathbf{u}}\mathbf{A}\boldsymbol{\epsilon}')\hat{\mathbf{u}} = s\mathbf{A}\boldsymbol{\epsilon}'\hat{\mathbf{u}} = s\mathbf{R},$$

and thus,

$$\mathbf{R}^n = s^{n-1}\mathbf{R}.$$

Using what we have found so far we can say that, for  $n > 0$ ,

$$\boldsymbol{\pi}(n) = \boldsymbol{\pi}(\mathbf{I} - \mathbf{R})\mathbf{R}^n = s^{n-1}\boldsymbol{\pi}(\mathbf{R} - \mathbf{R}^2) = s^{n-1}(1-s)\boldsymbol{\pi}\mathbf{R}.$$

Next we look at  $\boldsymbol{\pi}\mathbf{R}$ ,

$$\boldsymbol{\pi}\mathbf{R} = \left(\frac{1}{\mathbf{pV}\boldsymbol{\epsilon}'}\mathbf{pV}\right)\left(\frac{1}{\nu}\mathbf{B}\boldsymbol{\epsilon}'\hat{\mathbf{u}}\right) = \varrho\hat{\mathbf{u}}.$$

This yields

$$\boldsymbol{\pi}(n) = (1-s)s^{n-1}\varrho\hat{\mathbf{u}} \quad \text{for } n \geq 1,$$

exactly the same as Chapter 5. With some contortions (identical with those we did in Chapter 5) we can reproduce the expression for  $\boldsymbol{\pi}(0)$  given above.

In conclusion, we have found two completely distinct solutions for the simplest nontrivial MRP/M/1 queue, and have shown that they produce identical results. For more complicated systems it may be impossible to show that two different solutions yield the same results except by direct computation, but then we can only be sure to within numerical accuracy, and even then only for the particular parameters chosen.



<http://www.springer.com/978-0-387-49704-4>

Queueing Theory

A Linear Algebraic Approach

Lipsky, L.

2009, XIV, 548 p., Hardcover

ISBN: 978-0-387-49704-4