Journal of the Operations Research
Society of Japan
Vol. 30, No. 3, September 1987

ROW-CONTINUOUS FINITE MARKOV CHAINS: STRUCTURE AND ALGORITHMS

Julian Keilson
University of Rochester

Ushio Sumita
University of Rochester

M. Zachmann
University of Rochester

(Received January 21, 1986; Revised January 12, 1987)

Abstract A variety of bivariate finite Markov chain models are employed in the performance analysis of computer and production systems. Many such chains have a skip-free structure in their row index and can be described as row-continuous. For such a row-continuous chain, systematic treatment of the states in a row as a probabilistic entity permits development of a rank reducing algorithmic procedure for calculating the ergodic probabilities of the states. The algorithmic procedure produces, as byproducts, dynamic information such as first passage times, ergodic exit times and ergodic sojourn times, shedding additional light in understanding structural properties of row-continuous Markov chains. Many motivating examples are provided and some numerical results are exhibited.

§0. Introduction

A variety of finite Markov chain models are employed in the performance analysis of computer and production systems. Often such chains have so many states in their state space that algorithms for the study of the ergodic and transient behavior based on powers of the underlying Q-matrix fail. Some success with such problems has been obtained through the aggregation-dissagregation methods of Takahashi[18], Takahashi and Takami[19] and Schweitzer[17], and the matrix geometric result of Neuts[14,15]. The utilization of rows as probabilistic entities, pioneered by Neuts, is extended in this paper to address the ergodic and dynamic behavior sought both algorithmically and analytically.

The work described, largely completed in 1980, and widely quoted since, has experienced unfortunate delays in publication. Many of the basic ideas have since been published by other authors, see e.g.[3] and [4]. The topic, however, is of fundamental importance, the presentation is unique, and a variety of applications and structural insights are exhibited. In particular the ideas in Section 5, the matrix polynomial representation theorems in Section 6 and the applications of Sections 1 and 7 are of interest. The authors are grateful therefore, for the opportunity to have the paper appear formally in a journal of quality which is accessible.

A variety of Markov chains in continuous time have for their state space a rectangular lattice of states $B = \{(j,n) : 0 \leq j \leq J, \ \ 0 \leq n \leq N\}$. When the number of states $(J+1)(N+1)$ is large, say larger than 100, evaluation numerically of the ergodic distribution, and moments of exit times from and entrance times to subsets of interest is costly and simulation is often resorted to.

For many such chains, changes in column index j or row index n at transition epochs have values 0, ± 1. The chains may then be described as column-continuous and row-continuous respectively. When such row-continuity is present, for example, systematic treatment of the row subsets of states as probabilistic entities provides a theoretical basis for the discussion of the chain. Algorithms for the description of the chain involving matrices of order $J+1$, rather than $(J+1)(N+1)$, can then be developed better suited to the capacity constraints of computers. The procedure may therefore be described as rank reducing.

In the treatment of Neuts[14,15], the number of rows is infinite and the transition rates for the chain are independent of row index n, except near the boundary $n = 0$. His methods are oriented largely toward the ergodic behavior of such chains. The present treatment is primarily directed toward finite Markov chains with transition rates depending on both j and n. The methodology permits the number of states in a row n to vary with n as the reader will see. It should be emphasized that the row or column orientation, natural for some systems, may be an effective tool for chain description even when no natural row or column meaning is present.

In the first section, the basic bivariate process is described and notation is developed. Several motivating examples are given. Subsequent sections develop the methodology and algorithmic procedures, and discuss computer storage requirements and execution times. In a concluding section a tandem queue with Poisson arrivals, exponential service of different rates, multiple servers, and finite waiting rooms is presented.

§1. The Bivariate Markov Process $\underline{B}(t) = [J(t), N(t)]$

Consider a bivariate Markov process $\underline{B}(t) = [J(t), N(t)]$ on $\mathcal{B} = \mathcal{J} \times \mathcal{N}$ where $\mathcal{J} = \{j : 0 \leq j \leq J\}$, $\mathcal{N} = \{n : 0 \leq n \leq N\}$. Suppose that $\underline{B}(t)$ is governed by the set of hazard rates $\{\nu_{(j,m)(k,n)} \ : \ j, \ k \in \mathcal{J} \ ; m, \ n \in \mathcal{N}\}$. For one important case, one might have $\sum_n \nu_{(j,m)(k,n)}$ independent of m. The marginal process $J(t)$ is then a finite Markov chain in continuous time, but the marginal process $N(t)$, "defined on $J(t)$" in the sense of [12], need not be Markov. The formalism developed, however, is more general and does not require that either $J(t)$ or $N(t)$ be Markov.

Of interest in this paper are irreducible finite Markov chains $\underline{B}(t)$ for which $N(t)$ is skip-free in both directions so that

$$\nu_{(j,m)(k,n)} = 0 \quad if \ |n-m| > 1. \tag{1.1}$$

It will be convenient to work with the set of states $\{(j,n)\}$ with common row index n as an entity, and to introduce the corresponding notation $\underline{\underline{\nu}}_n^+$, $\underline{\underline{\nu}}_n^0$, $\underline{\underline{\nu}}_n^-$ to designate the transition rate matrices of order $J+1$ by

$$\begin{aligned} &(a) \quad \underline{\underline{\nu}}_n^0 = [\nu_{(j,n),(k,n)}] \\ &(b) \quad \underline{\underline{\nu}}_n^+ = [\nu_{(j,n),(k,n+1)}] \\ &(c) \quad \underline{\underline{\nu}}_n^- = [\nu_{(j,n),(k,n-1)}]. \end{aligned} \tag{1.2}$$

We will also work with the matrix

$$\underline{\underline{\nu}}_n = \underline{\underline{\nu}}_n^0 + \underline{\underline{\nu}}_n^- + \underline{\underline{\nu}}_n^+ \tag{1.3}$$

and the diagonal matrix

$$\underline{\underline{\nu}}_{Dn} = diag[\nu_{nj}] \quad ; \nu_{nj} = \sum_k \sum_m \nu_{(j,n),(k,m)}. \tag{1.4}$$

In the same spirit we will employ the transition probability substochastic matrices of order $J+1$ defined by

$$\underline{\underline{p}}_{mn}(t) = [p_{(j,m)(k,n)}(t)] \ ; \ p_{(j,m)(k,n)}(t) = P[\underline{B}(t) = (k,n) | \underline{B}(0) = (j,m)] \tag{1.5}$$

and state probability row vectors

$$\underline{p}_n^T(t) = [p_{(j,n)}(t)] \ ; \ p_{(j,n)}(t) = P[\underline{B}(t) = (j,n)]. \tag{1.6}$$

The ergodic row vectors will be designated by

$$\underline{e}_n^T = [e_{(j,n)}] \tag{1.7}$$

where $e_{(j,n)} = \lim_{t\to\infty} p_{(j,n)}(t)$. Laplace transforms will often be used, with the notation typical of subsequent usage,

$$\underline{\pi}_n^T(s) = L[\underline{p}_n^T(t)] = \int_0^\infty e^{-st} \underline{p}_n^T(t) dt. \tag{1.8}$$

We will also employ the notation of a matrix probability density function (p.d.f).

Definition 1.1 A matrix function $\underline{\underline{f}}(\tau) = [f_{mn}(\tau)]$ is said to be a matrix p.d.f. iff

a) $f_{mn}(\tau) \geq 0 \quad for \ \ all \ \ m,n,\tau$ and

b) $\sum_k \int_{-\infty}^{\infty} f_{jk}(\tau) d\tau = 1 \quad 0 \leq j \leq J.$

Such matrices play an important role in processes defined on a Markov chain [11,12] and in Markov renewal processes[1,9,16]. In our setting $f_{mn}(\tau) = 0$ for $\tau < 0$.

Examples To motivate the analysis that follows and indicate the prevalence of such row-continuous processes, some examples are appropriate. Transition diagrams for these examples are given in Figure 1.1.

A. Contiguous Processes

Any chain $\underline{B}(t) = [J(t),\ N(t)]$ on $\mathcal{B} = \{(j,n) |\ 0 \leq j \leq J,\ 0 \leq n \leq N\}$ for which $\nu_{(j,n),(k,m)} = 0$ if $|j-k| > 1$ or $|m-n| > 1$ will be called contiguous. The row-continuous chains are more general in that the marginal column process $J(t)$ need not be skip-free.

For any contiguous process the transition rate matrices of (1.2) are tri-diagonal. All such processes that are irreducible are amenable to the methods we describe.

B. Contiguous Horizontal-Vertical Processes

A subset of the contiguous processes is of interest for which either $J(t)$ or $N(t)$ can change in a lattice continuous manner, but not both simultaneously. If, for example, $J(t)$ and $N(t)$ were independent truncated birth-death processes, $\underline{B}(t)$ would be horizontal-vertical, since the probability of simultaneous change would be zero. In some cases, the independence between $J(t)$ and $N(t)$ may not be present, e.g. where $J(t)$ is a truncated birth-death process and $N(t)$ is a dependent birth-death-like process for which the upward and downward transition rates change when $J(t)$ changes. One example of this type is a communication link carrying both voice and data[2]. We note that for such processes $\underline{\underline{\nu}}_n^+$ and $\underline{\underline{\nu}}_n^-$ are diagonal.

C. Tandem Queues with Blocking

A contiguous process of interest is the tandem two station queue where each station has finite waiting room, arrivals are Poisson with rate λ, and service times are exponential with service rates η_1, η_2. For this process $\underline{\underline{\nu}}_n^+$ is diagonal, $\underline{\underline{\nu}}_n^-$ is upper diagonal, and $\underline{\underline{\nu}}_n^0$ is lower diagonal. (A matrix $\underline{\underline{a}} = [a_{ij}]$ is upper diagonal if $a_{ij} \neq 0$ implies $j = i+1$.) The methodology treated here allows each station to consist of a finite number of servers. Various types of blocking and feedback could as easily be analyzed. Extensive numerical analysis and further discussion of this tandem queue model can be found in Section 7.

D. Assembly Line

One interesting non-contiguous process is an assembly line with two machines in tandem and a buffer with finite storage in between. The marginal process $J(t)$ describes fluctuating working and non-working states of the two machines, M_1 and M_2 which are governed by failure rates μ_1, μ_2 and repair rates λ_1, λ_2 respectively. The two machines process work at the same rate of speed, i.e. have a common hazard rate ρ for completing their operation. When the second machine is down, items accumulate in the buffer. When the first machine is down, the flow of incoming items is cut off and the second machine goes idle after the buffer is emptied. The first machine stops when the buffer is full. $N(t)$ is the number of half-finished items in the buffer. For this model, both $\underline{\underline{\nu}}_n^+$ and $\underline{\underline{\nu}}_n^-$ are diagonal with two non-zero elements. The transition rate matrix $\underline{\underline{\nu}}_n^0$ has a zero diagonal.

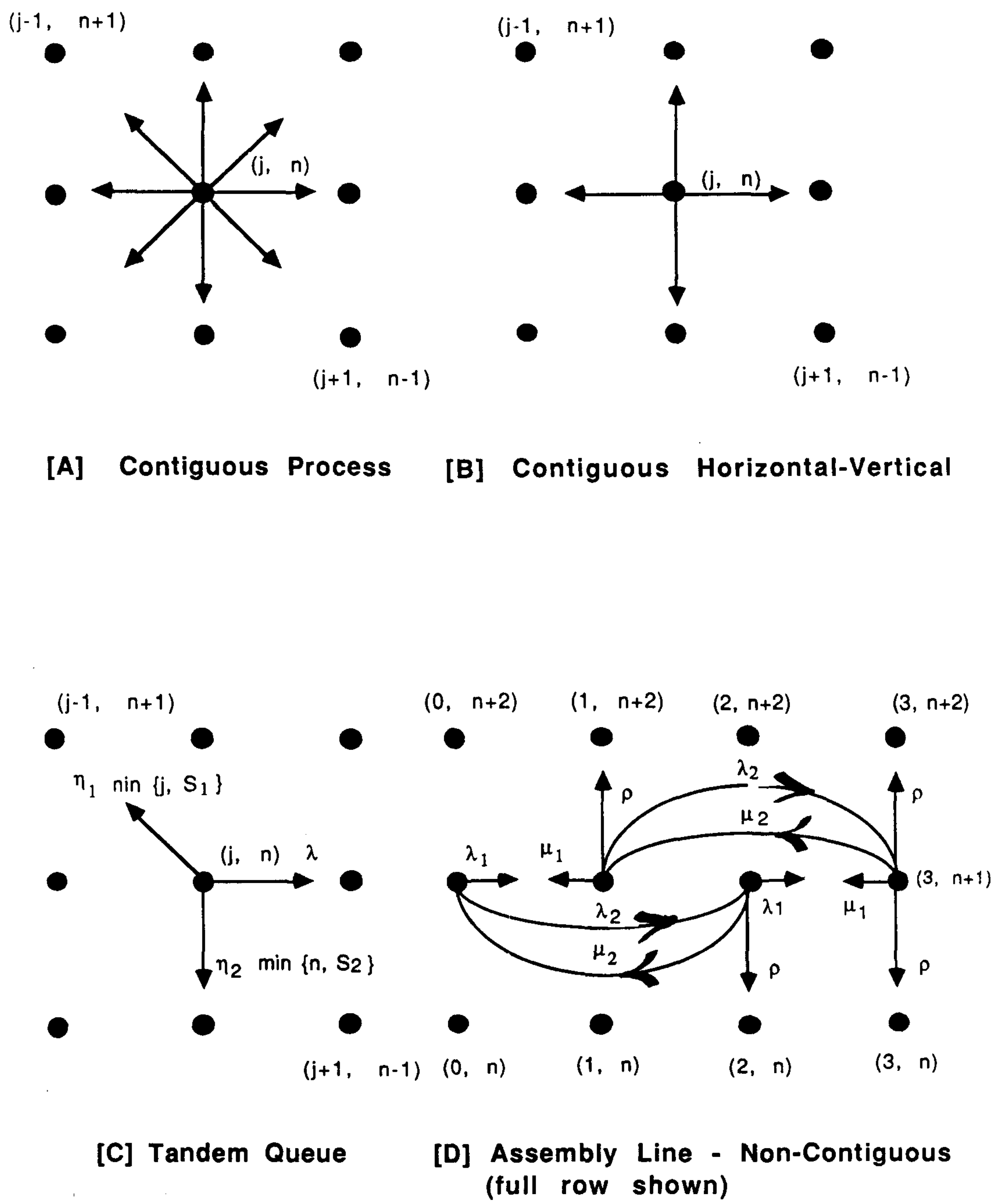

Fig. 1.1 Possible Transitions From An Interior Point

§2. Passage Time Densities

The description of the ergodic and dynamic behavior we will develop is based upon the passage time densities for the bivariate process. To exploit the skip-free feature of row-continuous processes, the passage times between states of adjacent rows are treated as building blocks. The focus on a row of states as an entity then gives rise to a matrix probability density function describing the joint distribution between the time of arrival at the adjacent row and the state reached, as a function of state of origin. From the passage time p.d.f.'s for adjacent rows one can then describe regeneration times for the states of a row and hence the ergodic probabilities. This will be clearer from the details of the development.

To discuss the passage time densities it will be useful to work with the lossy process $\tilde{J}_n(t)$ for row n on the set $\{(j,n)\}$ for which other states of $\mathcal{B}$ (i.e., other rows) are absorbing. Let $\tilde{p}_{(j,n),(k,n)}(t) = P[J(t) = k;\ N(t') = n,\ 0 \le t' \le t \mid J(0) = j,\ N(0) = n]$ and as for (1.5) let $\underline{\underline{\tilde{p}}}_{nn}(t) = [\tilde{p}_{(j,n),(k,n)}(t)]$. Since $\frac{d}{dt}\underline{\underline{\tilde{p}}}_{nn}(t) = -\underline{\underline{\tilde{p}}}_{nn}(t)\underline{\underline{\nu}}_{Dn} + \underline{\underline{\tilde{p}}}_{nn}(t)\underline{\underline{\nu}}_n^0$, we may rewrite $\underline{\underline{\tilde{p}}}_{nn}(t)$ in the from of

$$\underline{\underline{\tilde{p}}}_{nn}(t) = \exp\{\underline{\underline{\tilde{Q}}}_n t\}, \tag{2.1}$$

where $\underline{\underline{\tilde{Q}}}_n$ is the $\underline{\underline{Q}}$ matrix for the lossy process, i.e.,

$$\underline{\underline{\tilde{Q}}}_n = -\underline{\underline{\nu}}_{Dn} + \underline{\underline{\nu}}_n^0. \tag{2.2}$$

Note that $\underline{\underline{\tilde{p}}}_{nn}(t)$ is strictly substochastic. Form (2.1) one then has for $\underline{\underline{\tilde{\pi}}}_{nn}(s) = L[\underline{\underline{\tilde{p}}}_{nn}(t)]$,

$$\underline{\underline{\tilde{\pi}}}_{nn}(s) = \int_0^\infty e^{-st}\underline{\underline{\tilde{p}}}_{nn}(t)dt = [s\underline{\underline{I}} - \underline{\underline{\tilde{Q}}}_n]^{-1}, \quad s > 0. \tag{2.3}$$

The chain $B(t)$ is uniformizable [10] in the sense that there exists a positive constant ν such that $\infty > \nu > \max_{(j,m)} \sum_{(k,n)} \nu_{(j,m),(k,n)}$ since J and N are finite. If we let

$$\underline{\underline{a}}_n^+ = \frac{1}{\nu}\underline{\underline{\nu}}_n^+ \ ; \ \underline{\underline{a}}_n^- = \frac{1}{\nu}\underline{\underline{\nu}}_n^- \ ; \ \underline{\underline{a}}_n^0 = \underline{\underline{I}} - \frac{1}{\nu}[\underline{\underline{\nu}}_{Dn} - \underline{\underline{\nu}}_n^0] \tag{2.4}$$

we have from (2.3)

$$\underline{\underline{\tilde{\pi}}}_{nn}(s) = [(s+\nu)\underline{\underline{I}} - \nu\underline{\underline{a}}_n^0]^{-1}. \tag{2.5}$$

To proceed further, we require the passage time densities and some associated notation. Let $s_{n:jk}^+(\tau)$ be the joint probability density that, given one starts at (j,n), a) the row $(n+1)$ is visited for the first time at time τ and b) the row is visited at $(k, n+1)$. One then has $\sum_k \int_0^\infty s_{n:jk}^+(\tau)d\tau = 1$. Similarly, let $s_{n:jk}^-(\tau)$ be defined with respect to visiting the row $(n-1)$ for $n \ge 1$. The irreducibility of $\underline{B}(t)$ implies that $\underline{\underline{s}}_n^+(\tau) = [s_{n:jk}^+(\tau)]$ and $\underline{\underline{s}}_n^-(\tau) = [s_{n:jk}^-(\tau)]$ are matrix p.d.f's, see Definition 1.1. Correspondingly, $\underline{\underline{s}}_n^+ = \int_0^\infty \underline{\underline{s}}_n^+(\tau)d\tau$ and $\underline{\underline{s}}_n^- = \int_0^\infty \underline{\underline{s}}_n^-(\tau)d\tau$ are stochastic matrices. In terms of the Laplace transforms $\underline{\underline{\sigma}}_n^+(s) =$

$L[\underline{\underline{s}}_n^+(\tau)]$ and $\underline{\underline{\sigma}}_n^-(s) = L[\underline{\underline{s}}_n^-(\tau)]$, and moment matrices $\underline{\underline{\mu}}_n^+ = \int_0^\infty \tau \underline{\underline{s}}_n^+(\tau)d\tau$ and $\underline{\underline{\mu}}_n^- = \int_0^\infty \tau \underline{\underline{s}}_n^-(\tau)d\tau$, one has

$$\underline{\underline{s}}_n^+ = \underline{\underline{\sigma}}_n^+(0), \quad \underline{\underline{\mu}}_n^+ = -\frac{d}{ds}\underline{\underline{\sigma}}_n^+(s)|_{s=0} \ , \tag{2.6a}$$

$$\underline{\underline{s}}_n^- = \underline{\underline{\sigma}}_n^-(0), \quad \underline{\underline{\mu}}_n^- = -\frac{d}{ds}\underline{\underline{\sigma}}_n^-(s)|_{s=0} \ . \tag{2.6b}$$

The skip-free feature of the row-continuous processes provides a recursion between the upward passage time p.d.f.'s $\underline{\underline{s}}_n^+(\tau)$, and a recursion between the downward passage time p.d.f.'s $\underline{\underline{s}}_n^-(\tau)$. There are direct matrix counterparts to those present for birth-death processes with $J = \{0\}$. The recursion is based on an argument similar to that employed in [7,8,10]. For $N(t)$ to go from n to $n+1$, it must either do so directly after motion on row n, or there must be a first downward transition to row $n-1$, a first subsequent return to row n, and then a first subsequent arrival at row $n+1$. A probabilistic argument based on this then gives our basic recursive equation

$$\underline{\underline{s}}_n^+(t) = \tilde{\underline{\underline{p}}}_{nn}(t)\underline{\underline{\nu}}_n^+ + \tilde{\underline{\underline{p}}}_{nn}(t)\underline{\underline{\nu}}_n^- * \underline{\underline{s}}_{n-1}^+(t) * \underline{\underline{s}}_n^+(t) \ , \quad n \geq 1 \tag{2.7}$$

where the asterisk denotes convolution in time. That is, $\underline{a}(t) * \underline{b}(t) = \int_0^t \underline{a}(t-\tau)\underline{b}(\tau)d\tau$. Consequently, from (2.7) and (2.4), we have

$$\underline{\underline{\sigma}}_n^+(s) = \nu\tilde{\underline{\underline{\pi}}}_{nn}(s)\underline{\underline{a}}_n^+ + \nu\tilde{\underline{\underline{\pi}}}_{nn}(s)\underline{\underline{a}}_n^-\underline{\underline{\sigma}}_{n-1}^+(s)\underline{\underline{\sigma}}_n^+(s) \ , \quad n \geq 1. \tag{2.8}$$

If we solve (2.8) for $\underline{\underline{\sigma}}_n^+(s)$ and use (2.5), we obtain

$$[\underline{\underline{I}} - (\frac{\nu}{s+\nu})\{\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^-\underline{\underline{\sigma}}_{n-1}^+(s)\}]\underline{\underline{\sigma}}_n^+(s) = (\frac{\nu}{s+\nu})\underline{\underline{a}}_n^+. \tag{2.9}$$

When $n = 0$ one has from $\underline{\underline{s}}_0^+(t) = \tilde{\underline{\underline{p}}}_{00}(t)\underline{\underline{\nu}}_0^+$ and (2.5) that

$$[\underline{\underline{I}} - (\frac{\nu}{s+\nu})\underline{\underline{a}}_0^0]\underline{\underline{\sigma}}_0^+(s) = (\frac{\nu}{s+\nu})\underline{\underline{a}}_0^+. \tag{2.10}$$

Equations (2.9) and (2.10) can be used to generate $\underline{\underline{\sigma}}_n^+(s)$ recursively. Similar equations generate $\underline{\underline{\sigma}}_n^-(s)$ recursively from $\underline{\underline{\sigma}}_N^-(s)$, see(4.2). By letting $s = 0$ we obtain:

$$[\underline{\underline{I}} - \{\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^-\underline{\underline{s}}_{n-1}^+\}]\underline{\underline{s}}_n^+ = \underline{\underline{a}}_n^+ \quad n \geq 1 \tag{2.11}$$

and

$$[\underline{\underline{I}} - \underline{\underline{a}}_0^0]\underline{\underline{s}}_0^+ = \underline{\underline{a}}_0^+. \tag{2.12}$$

Similar relations for $\underline{\underline{s}}_n^-$ are given in (4.4). Differentiation of (2.9) and (2.10) with respect to s at $s = 0$ leads to the following relations for the mean first passage times:

$$[\underline{\underline{I}} - \underline{\underline{a}}_n^0 - \underline{\underline{a}}_n^-\underline{\underline{s}}_{n-1}^+]\underline{\underline{\mu}}_n^+ = \frac{1}{\nu}[\underline{\underline{I}} + \nu\underline{\underline{a}}_n^-\underline{\underline{\mu}}_{n-1}^+]\underline{\underline{s}}_n^+ \tag{2.13}$$

and

$$[\underline{\underline{I}} - \underline{\underline{a}}_0^0]\underline{\underline{\mu}}_0^+ = \frac{1}{\nu}\underline{\underline{s}}_0^+. \tag{2.14}$$

We note that both $\underline{\underline{s}}_n^+$ and $(\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^- + \underline{\underline{a}}_n^+)$ are stochastic. Since $\underline{\underline{a}}_n^+$ is not a zero matrix because of irreducibility, the matrix $(\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^- \underline{\underline{s}}_{n-1}^+)$ is strictly substochastic, its spectral radius is less than one, and $[\underline{\underline{I}} - \underline{\underline{a}}_n^0 - \underline{\underline{a}}_n^- \underline{\underline{s}}_{n-1}^+]$ is invertible. One may therefore obtain $\underline{\underline{s}}_n^+$ from (2.11) and $\underline{\underline{\mu}}_n^+$ from (2.13). We also see that $\underline{\underline{0}} \le \underline{\underline{\sigma}}_n^+(s) \le \underline{\underline{s}}_n^+$ and $\nu/(s+\nu) \le 1$ for $s > 0$ real. Hence $[\underline{\underline{I}} - (\frac{\nu}{s+\nu})\{\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^- \underline{\underline{\sigma}}_{n-1}^+(s)\}]$ is invertible for $s \ge 0$ real. Therefore (2.9) may be used to obtain $\underline{\underline{\sigma}}_n{}^+(s)$.

Of frequent interest in applications is the matrix passage time p.d.f. $\underline{\underline{s}}_{0n}(\tau)$ describing the distribution of the time at which row n is first reached given a start at row 0. Specifically $\underline{\underline{s}}_{0n}(\tau)$ is $[s_{0n:ij}(\tau)]$ where $s_{0n:ij}(\tau)$ is the joint probability density that the set $\{(j,n) \ : \ j \in J\}$ is visited for the first time at (j,n) and the time of first arrival is τ, given a start at $(0,i)$. A simple probabilistic argument shows that $\underline{\underline{s}}_{0n}(\tau) = \underline{\underline{s}}_0^+(\tau) * \underline{\underline{s}}_1^+(\tau) * \ldots * \underline{\underline{s}}_{n-1}^+(\tau)$. Correspondingly, $\underline{\underline{\sigma}}_{0n}(s) = \underline{\underline{\sigma}}_0^+(s)\underline{\underline{\sigma}}_1^+(s)\ldots\underline{\underline{\sigma}}_{n-1}^+(s)$. This will be discussed further in Section 6.

§3. Ergodic Probabilities

We can now use the passage time densities of Section 2 to find the ergodic probabilities. The basic probabilistic argument for finding the transition probabilities from the passage times goes as follows. If one is in row n at time 0, to be in row n at time t, either

a) one never left row n;

b) one went to row $n+1$ before row $n-1$, returned to row n for the first time subsequently, and was in row n at time t, possibly after further wanderings; or

c) one went to row $n-1$ before row $n+1$, etc. as in b).

Consequently, one has, for the cases $0 < n < N$, $n = 0$, and $n = N$ respectively,

$$\underline{\underline{p}}_{nn}(t) = \tilde{\underline{\underline{p}}}_{nn}(t) + \nu\tilde{\underline{\underline{p}}}_{nn}(t)\underline{\underline{a}}_n^+ * \underline{\underline{s}}_{n+1}^-(t) * \underline{\underline{p}}_{nn}(t) + \nu\tilde{\underline{\underline{p}}}_{nn}(t)\underline{\underline{a}}_n^- * \underline{\underline{s}}_{n-1}^+(t) * \underline{\underline{p}}_{nn}(t), \tag{3.1}$$

$$\underline{\underline{p}}_{00}(t) = \tilde{\underline{\underline{p}}}_{00}(t) + \nu\tilde{\underline{\underline{p}}}_{00}(t)\underline{\underline{a}}_0^+ * \underline{\underline{s}}_1^-(t) * \underline{\underline{p}}_{00}(t), \tag{3.2}$$

and

$$\underline{\underline{p}}_{NN}(t) = \tilde{\underline{\underline{p}}}_{NN}(t) + \nu\tilde{\underline{\underline{p}}}_{NN}(t)\underline{\underline{a}}_N^- * \underline{\underline{s}}_{N-1}^+(t) * \underline{\underline{p}}_{NN}(t). \tag{3.3}$$

From (3.1) it follows that

$$\underline{\underline{\pi}}_{nn}(s) = \tilde{\underline{\underline{\pi}}}_{nn}(s) + \nu\tilde{\underline{\underline{\pi}}}_{nn}(s)\underline{\underline{a}}_n^+\underline{\underline{\sigma}}_{n+1}^-(s)\underline{\underline{\pi}}_{nn}(s) + \nu\tilde{\underline{\underline{\pi}}}_{nn}(s)\underline{\underline{a}}_n^-\underline{\underline{\sigma}}_{n-1}^+(s)\underline{\underline{\pi}}_{nn}(s), \quad 0 < n < N$$

which in turn leads to

$$[\underline{\underline{I}} - \nu\tilde{\underline{\underline{\pi}}}_{nn}(s)\underline{\underline{a}}_n^+\underline{\underline{\sigma}}_{n+1}^-(s) - \nu\tilde{\underline{\underline{\pi}}}_{nn}(s)\underline{\underline{a}}_n^-\underline{\underline{\sigma}}_{n-1}^+(s)]\underline{\underline{\pi}}_{nn}(s) = \tilde{\underline{\underline{\pi}}}_{nn}(s). \tag{3.4}$$

One then finds, from (2.5), that

$$\underline{\underline{\pi}}_{nn}(s) = (\frac{1}{s+\nu})[\underline{\underline{I}} - (\frac{\nu}{s+\nu})\{\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^+ \underline{\underline{\sigma}}_{n+1}^-(s) + \underline{\underline{a}}_n^- \underline{\underline{\sigma}}_{n-1}^+(s)\}]^{-1}. \tag{3.5}$$

To show that the matrix expressed within brackets is in fact invertible, we let

$$\underline{\underline{\beta}}_n(s) = (\frac{\nu}{s+\nu})\{\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^+ \underline{\underline{\sigma}}_{n+1}^-(s) + \underline{\underline{a}}_n^- \underline{\underline{\sigma}}_{n-1}^+(s)\} \quad n = 1, \ldots, N-1. \tag{3.6}$$

Note that $(\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^+ + \underline{\underline{a}}_n^-)$ is stochastic, and further

$$L^{-1}[\underline{\underline{\beta}}_n(s)] = (\nu e^{-\nu t}\underline{\underline{I}}) * (\underline{\underline{a}}_n^0 \delta(t) + \underline{\underline{a}}_n^+ \underline{\underline{s}}_{n+1}^-(t) + \underline{\underline{a}}_n^- \underline{\underline{s}}_{n-1}^+(t)). \tag{3.7}$$

Therefore each row of $\underline{\underline{\beta}}_n(s)$ is a convex combination of rows of matrix p.d.f.'s. Hence $[\underline{\underline{I}} - \underline{\underline{\beta}}_n(s)]$ is invertible for $s > 0$ real. With this identification, Equation(3.5) becomes

$$\underline{\underline{\pi}}_{nn}(s) = (\frac{1}{s+\nu})[\underline{\underline{I}} - \underline{\underline{\beta}}_n(s)]^{-1}. \tag{3.8}$$

In fact, with the definition

$$\underline{\underline{\beta}}_0(s) = (\frac{\nu}{s+\nu})[\underline{\underline{a}}_0^0 + \underline{\underline{a}}_0^+ \underline{\underline{\sigma}}_1^-(s)] \tag{3.9a}$$

and

$$\underline{\underline{\beta}}_N(s) = (\frac{\nu}{s+\nu})[\underline{\underline{a}}_N^0 + \underline{\underline{a}}_N^- \underline{\underline{\sigma}}_{N-1}^-(s)] \tag{3.9b}$$

we immediately obtain (3.8) for all $n \in N$.

To find $\underline{e}_n^T$ we must evaluate $\lim_{s\to 0+} s\underline{\underline{\pi}}_{nn}(s) = \underline{1}\underline{e}_n^T$. In [9] it is shown that

$$\lim_{s\to 0+} s\underline{\underline{\pi}}_{nn}(s) = \frac{1}{\nu} \cdot \frac{1}{\underline{e}_{bn}^T \underline{\underline{\mu}}_{bn} \underline{1}} \cdot \underline{1}\, \underline{e}_{bn}^T \tag{3.10}$$

where $\underline{e}_{bn}^T$ is the stationary left eigenvector for $\underline{\underline{\beta}}_n(0)$ and $\underline{\underline{\mu}}_{bn}$ is the mean of $\underline{\underline{b}}_n(t)$, that is

$$\underline{e}_{bn}^T \underline{\underline{\beta}}_n(0) = \underline{e}_{bn}^T \tag{3.11}$$

and

$$\underline{\underline{\mu}}_{bn} = \int_0^\infty t\underline{\underline{b}}_n(t)dt. \tag{3.12}$$

We can easily find $\underline{\underline{\beta}}_n(0)$ by setting $s = 0$ in (3.6), (3.9a) and (3.9b), while $\underline{\underline{\mu}}_{bn}$ is accessible by differentiation via

$$\underline{\underline{\mu}}_{bn} = -\underline{\underline{\beta}}_n'(0) = \frac{1}{\nu}\underline{\underline{\beta}}_n(0) + \underline{\underline{a}}_n^+ \underline{\underline{\mu}}_{n+1}^- + \underline{\underline{a}}_n^- \underline{\underline{\mu}}_{n-1}^+ \qquad n = 1, \ldots, N-1, \tag{3.13}$$

and

$$\underline{\underline{\mu}}_{b0} = \frac{1}{\nu}\underline{\underline{\beta}}_0(0) + \underline{\underline{a}}_0^+ \underline{\underline{\mu}}_1^- \quad ; \quad \underline{\underline{\mu}}_{bN} = \frac{1}{\nu}\underline{\underline{\beta}}_N(0) + \underline{\underline{a}}_N^- \underline{\underline{\mu}}_{N-1}^+. \tag{3.14}$$

As we will see in Section 5, row balance

$$\underline{e}_{n+1}^T \underline{\underline{a}}_{n+1}^- \underline{1} = \underline{e}_n^T \underline{\underline{a}}_n^+ \underline{1} \quad , \quad 0 \le n \le N-1 \tag{3.15}$$

is always present. This enables one to evaluate

$$m_{bn} = \nu \underline{e}_{bn}^T \underline{\underline{\mu}}_{bn} \underline{1} \tag{3.16}$$

recursively without computing $\underline{\underline{\mu}}_{bn}$, and thus $\underline{\underline{\mu}}_n^+$ and $\underline{\underline{\mu}}_n^-$, in the following manner. From (3.10) one has $\underline{e}_n^T = \frac{1}{m_{bn}} \underline{e}_{bn}^T$ so that

$$\underline{e}_{n+1}^T \underline{\underline{a}}_{n+1}^- \underline{1} = \frac{1}{m_{b,n+1}} \underline{e}_{b,n+1}^T \underline{\underline{a}}_{n+1}^- \underline{1} \quad ; \quad \underline{e}_n^T \underline{\underline{a}}_n^+ \underline{1} = \frac{1}{m_{b,n}} \underline{e}_{bn}^T \underline{\underline{a}}_n^+ \underline{1}. \tag{3.17}$$

Hence, from (3.16), one sees that

$$m_{b,n+1} = m_{b,n} \frac{\underline{e}_{b,n+1}^T \underline{\underline{a}}_{n+1}^- \underline{1}}{\underline{e}_{b,n}^T \underline{\underline{a}}_n^+ \underline{1}} \qquad 0 \le n \le N-1. \tag{3.18}$$

One may start with an arbitrary positive m_{b0} and then normalize at the end. More specifically,

$$\underline{e}_n^T = \frac{K}{m_{bn}} \underline{e}_{bn}^T \quad , \quad m_{b0} > 0 \quad , \quad \text{arbitrary} \tag{3.19}$$

and

$$K = [\sum_{n \in \mathcal{N}} \frac{1}{m_{bn}} \underline{e}_{bn}^T \underline{1}]^{-1}. \tag{3.20}$$

§4. Summary of Computational Procedures

A tabulation of the key results obtained in the previous sections is given to provide an overview of the formalism, and ease of access to the formulae needed for implementation.

$$\underline{\underline{\sigma}}_0^+(s) = [\underline{\underline{I}} - (\frac{\nu}{s+\nu}) \underline{\underline{a}}_0^0]^{-1} (\frac{\nu}{s+\nu}) \underline{\underline{a}}_0^+ \tag{4.1}$$

$$\underline{\underline{\sigma}}_n^+(s) = [\underline{\underline{I}} - (\frac{\nu}{s+\nu}) \{\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^- \underline{\underline{\sigma}}_{n-1}^+(s)\}]^{-1} (\frac{\nu}{s+\nu}) \underline{\underline{a}}_n^+$$

$$\underline{\underline{\sigma}}_N^-(s) = [\underline{\underline{I}} - (\frac{\nu}{s+\nu}) \underline{\underline{a}}_N^0]^{-1} (\frac{\nu}{s+\nu}) \underline{\underline{a}}_N^- \tag{4.2}$$

$$\underline{\underline{\sigma}}_n^-(s) = [\underline{\underline{I}} - (\frac{\nu}{s+\nu}) \{\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^+ \underline{\underline{\sigma}}_{n+1}^-(s)\}]^{-1} (\frac{\nu}{s+\nu}) \underline{\underline{a}}_n^-$$

$$\underline{\underline{s}}_0^+ = [\underline{\underline{I}} - \underline{\underline{a}}_0^0]^{-1} \underline{\underline{a}}_0^+ \tag{4.3}$$

$$\underline{\underline{s}}_n^+ = [\underline{\underline{I}} - \{\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^- \underline{\underline{s}}_{n-1}^+\}]^{-1} \underline{\underline{a}}_n^+$$

$$\underline{\underline{s}}_N^- = [\underline{\underline{I}} - \underline{\underline{a}}_N^0]^{-1} \underline{\underline{a}}_N^- \tag{4.4}$$

$$\underline{\underline{s}}_n^- = [\underline{\underline{I}} - \{\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^+ \underline{\underline{s}}_{n+1}^-\}]^{-1} \underline{\underline{a}}_n^-$$

$$\underline{\underline{\mu}}_0^+ = \frac{1}{\nu} [\underline{\underline{I}} - \underline{\underline{a}}_0^0]^{-1} \underline{\underline{s}}_0^+ \tag{4.5}$$

$$\underline{\underline{\mu}}_n^+ = \frac{1}{\nu} [\underline{\underline{I}} - \{\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^- \underline{\underline{s}}_{n-1}^+\}]^{-1} [\underline{\underline{I}} + \nu \underline{\underline{a}}_n^- \underline{\underline{\mu}}_{n-1}^+] \underline{\underline{s}}_n^+$$

$$\underline{\underline{\mu}}_N^- = \frac{1}{\nu}[\underline{\underline{I}} - \underline{\underline{a}}_N^0]^{-1}\underline{\underline{s}}_N^- \tag{4.6}$$

$$\underline{\underline{\mu}}_n^- = \frac{1}{\nu}[\underline{\underline{I}} - \{\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^+\underline{\underline{s}}_{n+1}^-)\}]^{-1}[\underline{\underline{I}} + \nu\underline{\underline{a}}_n^+\underline{\underline{\mu}}_{n+1}^-\}]\underline{\underline{s}}_n^-$$

$$\underline{\underline{\beta}}_n(s) = (\frac{\nu}{s+\nu})[\underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^+\underline{\underline{\sigma}}_{n+1}^-(s) + \underline{\underline{a}}_n^-\underline{\underline{\sigma}}_{n-1}^+(s)] \tag{4.7}$$

$$\underline{\underline{\beta}}_n(0) = \underline{\underline{a}}_n^0 + \underline{\underline{a}}_n^+\underline{\underline{s}}_{n+1}^- + \underline{\underline{a}}_n^-\underline{\underline{s}}_{n-1}^+$$

$$\pi_{nn}(s) = (\frac{1}{s+\nu})[\underline{\underline{I}} - \underline{\underline{\beta}}_n(s)]^{-1} \tag{4.8}$$

$$\underline{e}_{bn}^T\underline{\underline{\beta}}_n(0) = \underline{e}_{bn}^T \tag{4.9}$$

$$m_{bn} = m_{b,\ n-1}\frac{(\underline{e}_{bn}^T\underline{\underline{a}}_n^-\underline{1})}{(\underline{e}_{b,n-1}^T\underline{\underline{a}}_{n-1}^+\underline{1})}, \quad m_{b0} > 0 \quad \text{arbitrary} \tag{4.10}$$

$$\underline{e}_n^T = \frac{K}{m_{bn}}\underline{e}_{bn}^T \quad ; \quad K = [\sum_{n\in\mathcal{N}} \frac{1}{m_{bn}}\underline{e}_{bn}^T\underline{1}]^{-1} \tag{4.11}$$

§5. Row Balance and Row Generation

The computational procedure outlined in Section 3 and 4 is a rank-reducing procedure. The procedure can further be improved by using set balance on the state space, when mean passage times are not needed. A still greater improvement can be realized when the $\underline{\underline{a}}_n^+$ (or $\underline{\underline{a}}_n^-$) are invertible for all n. We have seen in Section 1 that the required invertibility is often present.

To exhibit row balance, some preliminary tools are needed.

Lemma 5.1

$$\underline{e}_{n+1}^T\underline{\underline{a}}_{n+1}^- = \underline{e}_n^T[\underline{\underline{I}} - \underline{\underline{a}}_n^0] - \underline{e}_{n-1}^T\underline{\underline{a}}_{n-1}^+ \ , \quad 1 \le n \le N-1 \tag{5.1}$$

and

$$\underline{e}_1^T\underline{\underline{a}}_1^- = \underline{e}_0^T[\underline{\underline{I}} - \underline{\underline{a}}_0^0]. \tag{5.2}$$

Proof: The forward equations are:

$$\frac{d}{dt}\underline{p}_n^T(t) = -\nu\underline{p}_n^T(t) + \nu\underline{p}_n^T(t)\underline{\underline{a}}_n^0 + \nu\underline{p}_{n-1}^T(t)\underline{\underline{a}}_{n-1}^+ + \nu\underline{p}_{n+1}^T(t)\underline{\underline{a}}_{n+1}^- \ , \quad 1 \le n \le N-1 \tag{5.3}$$

and

$$\frac{d}{dt}\underline{p}_0^T(t) = -\nu\underline{p}_0^T(t) + \nu\underline{p}_0^T(t)\underline{\underline{a}}_0^0 + \nu\underline{p}_1^T(t)\underline{\underline{a}}_1^-. \tag{5.4}$$

Since $\underline{p}_n^T(t) \to \underline{e}_n^T$, and $\frac{d}{dt}\underline{p}_0^T(t) \to \underline{0}^T$ as $t \to \infty$, the result is immediate.

Theorem 5.2 If $\underline{\underline{a}}_n^-$ is invertible for all $1 \le n \le N$ then

$$\underline{e}_{n+1}^T = \{\underline{e}_n^T[\underline{\underline{I}} - \underline{\underline{a}}_n^0] - \underline{e}_{n-1}^T \underline{\underline{a}}_{n+1}^+\}(\underline{\underline{a}}_{n+1}^-)^{-1} \quad , \quad 1 \le n \le N-1 \tag{5.5}$$

and

$$\underline{e}_1^T = \underline{e}_0^T[\underline{\underline{I}} - \underline{\underline{a}}_0^0](\underline{\underline{a}}_1^-)^{-1}. \tag{5.6}$$

Proof: The theorem follows directly from (5.1) and (5.2).

We see that when the ergodic distribution is desired, and this invertibility present, the $\underline{e}_n^T$ are available recursively via (5.5). In such a case, one needs to calculate only one eigenvector ($\underline{e}_{b0}^T$) in order to obtain the other $\underline{e}_n^T$ recursively.

A similar result holds when $\underline{\underline{a}}_n^+$ for $0 \le n \le N$ are invertible. The recursion then begins with $\underline{e}_N^T$.

Theorem 5.3 If $\underline{\underline{a}}_n$ is invertible for all $n > 0$ then

$$\underline{e}_{n-1}^T = \{\underline{e}_n^T[\underline{\underline{I}} - \underline{\underline{a}}_n^0] - \underline{e}_{n+1}^T \underline{\underline{a}}_{n+1}^-\}(\underline{\underline{a}}_{n-1}^+)^{-1} \quad , \quad n = 1, \ldots, N-1 \tag{5.7}$$

and

$$\underline{e}_{N-1}^T = \underline{e}_N^T[\underline{\underline{I}} - \underline{\underline{a}}_N^0](\underline{\underline{a}}_{N-1}^+)^{-1}. \tag{5.8}$$

Proof: Equation (5.7) follows by rearranging (5.1). The forward equation (5.4) has its counterpart on the top row

$$\frac{d}{dt}\underline{p}_N^T(t) = -\nu \underline{p}_N^T(t) + \nu \underline{p}_N^T(t)\underline{\underline{a}}_N^0 + \nu \underline{p}_{N-1}^T(t)\underline{\underline{a}}_{N-1}^+. \tag{5.9}$$

By letting $t \to \infty$, (5.8) follows from (5.9).

Even when neither set is invertible, set balance can be used to normalize the $\underline{e}_{bn}^T$, as described at the end of Section 3. We are now ready to prove our basic result regarding the set balance.

Theorem 5.4 (Row Balance)

$$\underline{e}_{n+1}^T \underline{\underline{a}}_{n+1} \underline{1} = \underline{e}_n^T \underline{\underline{a}}_n^+ \underline{1} \quad for \quad 0 \le n \le N-1 \,. \tag{5.10}$$

Proof: We recall that $\underline{\underline{a}}_n^0 \underline{1} = \underline{1} - (\underline{\underline{a}}_n^+ + \underline{\underline{a}}_n^-)\underline{1}$ and $\underline{\underline{a}}_0^0 \underline{1} = \underline{1} - \underline{\underline{a}}_0^+ \underline{1}$. Combining this with Lemma 5.1 we obtain

$$\underline{e}_n^T \underline{\underline{a}}_n^+ \underline{1} - \underline{e}_n^T \underline{\underline{a}}_{n+1}^- \underline{1} = \underline{e}_{n-1}^T \underline{\underline{a}}_{n-1}^+ \underline{1} - \underline{e}_n^T \underline{\underline{a}}_n^- \underline{1} \tag{5.11}$$

and

$$\underline{e}_0^T \underline{\underline{a}}_0^+ \underline{1} = \underline{e}_1^T \underline{\underline{a}}_1^- \underline{1}. \tag{5.12}$$

The theorem now follows by induction.

§6. The Matrix Polynomial Representation of $\underline{\underline{\sigma}}_{0n}(s)$

Important information on the dynamical behavior of a Markov chain is contained in the spectral structure of its first passage time densities. Knowledge of this structure, and that of the related relaxation time, is essential when one wishes to use the ergodic distribution [10]. In this section, we discuss a representation of $\underline{\underline{\sigma}}_{mn}(s)$ in terms of matrix polynomials, where the set of poles of $\underline{\underline{\sigma}}_{mn}(s)$ corresponds to spectral lines.

Our approach closely parallels the work done on one-dimensional birth-death process in [5,6] and [7,8,10]. This representation will be used to discuss the structure of $\underline{\underline{s}}_{0n}(\tau)$ and simple related results. We also indicate how this representation may be used, in principle, to obtain the relaxation times.

Theorem 6.1 If $\underline{\underline{a}}_n^+$ is non-singular for all n, let

$$\underline{\underline{Q}}_{n+1}(s) = (\underline{\underline{a}}_n^+)^{-1}[\{(\frac{s+\nu}{\nu})\underline{\underline{I}} - \underline{\underline{a}}_n^0\}\underline{\underline{Q}}_n(s) - \underline{\underline{a}}_n^- \underline{\underline{Q}}_{n-1}(s)] \tag{6.1}$$

with

$$\underline{\underline{Q}}_0(s) = \underline{\underline{I}} \ ; \ \underline{\underline{Q}}_1(s) = (\underline{\underline{a}}_n^+)^{-1}[(\frac{s+\nu}{\nu})\underline{\underline{I}} - \underline{\underline{a}}_0^0]. \tag{6.2}$$

Then the matrix polynomial $\underline{\underline{Q}}_n(s)$ is invertible for $s \geq 0$, and

$$\sigma_n^+(s) = \underline{\underline{Q}}_n(s)\underline{\underline{Q}}_{n+1}^{-1}(s). \tag{6.3}$$

Proof: We prove the theorem by induction. Clearly $\underline{\underline{Q}}_1(s) = (\frac{\nu}{s+\nu})(\underline{\underline{a}}_0^+)^{-1}[\underline{\underline{I}} - (\frac{s+\nu}{\nu})\underline{\underline{a}}_0^0]$ where $\underline{\underline{a}}_0^0$ is strictly substochastic. Hence $\underline{\underline{Q}}_1(s)$ is invertible for all $s \geq 0$. Thus (6.3) holds true for $n = 0$. Now assume that $\underline{\underline{Q}}_n(s)$ is non-singular for $0 \leq n \leq M$ and $\sigma_n^+(s) = \underline{\underline{Q}}_n(s)\underline{\underline{Q}}_{n+1}(s)^{-1}$ for $0 \leq n \leq M-1$. Then

$$\begin{aligned}\underline{\underline{Q}}_{M+1}(s) &= (\underline{\underline{a}}_M^+)^{-1}[\{(\frac{s+\nu}{\nu})\underline{\underline{I}} - \underline{\underline{a}}_M^0\}\underline{\underline{Q}}_M(s) - \underline{\underline{a}}_M^- \underline{\underline{Q}}_{M-1}(s)] \\ &= (\frac{s+\nu}{\nu})(\underline{\underline{a}}_M^+)^{-1}[\underline{\underline{I}} - (\frac{\nu}{s+\nu})\{\underline{\underline{a}}_M^0 - \underline{\underline{a}}_M^- \underline{\underline{\sigma}}_{M-1}^+(s)\}]\underline{\underline{Q}}_M(s).\end{aligned} \tag{6.4}$$

Since $\underline{\underline{\sigma}}_{M-1}^+(s)$ is the transform of a matrix p.d.f., we see that $\underline{\underline{Q}}_{M+1}(s)$ is invertible for $s \geq 0$. Postmultiplying (6.4) by $\underline{\underline{Q}}_M^{-1}(s)$ and inverting one sees from (2.8), that $\underline{\underline{\sigma}}_M^+(s) = \underline{\underline{Q}}_M(s)\underline{\underline{Q}}_{M+1}(s)^{-1}$, completing the proof.

We note that the recursion relation (6.1) implies that $\underline{\underline{Q}}_n(s)_{ij}$ is a polynomial in s of degree n. The decomposition (6.3) will be seen to be useful in a variety of ways.

The $(\underline{\underline{Q}}_n(s))_{ij}$ arrays allow one to evaluate first passage times upwards and downwards over a number of rows rather easily. The following theorems illustrate this point.

Theorem 6.2 In the context of Theorem 6.1, one has :

a) $\underline{\underline{\sigma}}_{0n}(s) = \underline{\underline{Q}}_n(s)^{-1} \qquad for \quad n \geq 1$; (6.5)

b) $\quad Q'_{n+1}(0) = (\underline{\underline{a}}_n^+)^{-1}\{(\frac{1}{\nu})\underline{\underline{Q}}_n(0) + (\underline{\underline{I}} - \underline{\underline{a}}_n^0)Q'_n(0) - \underline{\underline{a}}_n^-\underline{\underline{Q}}'_{n-1}(0)\}$
with $\underline{\underline{Q}}'_0(0) = \underline{\underline{0}}$ *and* $\underline{\underline{Q}}'_1(0) = \frac{1}{\nu}\underline{\underline{a}}_0^{+-1}$; (6.6)
c) $\quad \underline{\underline{\mu}}_{0n} = \underline{\underline{Q}}_n^{-1}(0)\underline{\underline{Q}}'_n(0)\underline{\underline{Q}}_n^{-1}(0)$.

Proof: We observe from Theorem 6.1 that

$$\underline{\underline{\sigma}}_{on}(s) = \prod_{m=0}^{n-1} \underline{\underline{\sigma}}_m^+(s) = \underline{\underline{Q}}_0(s)(\prod_{m=0}^{n-1} \underline{\underline{Q}}_m^{-1}(s)\underline{\underline{Q}}_m(s))\underline{\underline{Q}}_n^{-1}(s) = \underline{\underline{Q}}_n^{-1}(s),$$

proving (a). For (b), we see from (6.2) that for $n \geq 2$,

$$\underline{\underline{Q}}'_{n+1}(s) = (\underline{\underline{a}}_n^+)^{-1}[\frac{1}{\nu}\underline{\underline{Q}}_n(s) + \{(\frac{s+\nu}{\nu})\underline{\underline{I}} - \underline{\underline{a}}_n^0\}\underline{\underline{Q}}'_n(s) - \underline{\underline{a}}_n^-\underline{\underline{Q}}'_{n-1}(s)]. \tag{6.7}$$

Part (b) then follows for $n \geq 2$ by setting $s = 0$. The cases for $n = 0, 1$ are trivial. We note from (a) that $\underline{\underline{Q}}_n(s)\underline{\underline{\sigma}}_{0n}(s) = \underline{\underline{I}}$. It then follows that

$$\underline{\underline{Q}}'_n(s)\underline{\underline{\sigma}}_{0n}(s) + \underline{\underline{Q}}_n(s)\underline{\underline{\sigma}}'_{0n}(s) = \underline{\underline{0}}. \tag{6.8}$$

Therefore, one has

$$\underline{\underline{\mu}}_{0n} = -\underline{\underline{\sigma}}'_{0n}(s) = \underline{\underline{Q}}_n^{-1}(0)\underline{\underline{Q}}'_n(0)\underline{\underline{Q}}_n^{-1}(0). \tag{6.9}$$

proving (c).

Theorem 6.2 provides a computationally easy way to calculate the mean upward passage times when $\underline{\underline{a}}_n^+$ are invertible. A dual argument eases the computation of downward passage times when $\underline{\underline{a}}_n^-$ are invertible. The explicit formulation for arbitrary upward passage times is given next.

Theorem 6.3

a) $\quad \underline{\underline{\sigma}}_{mn}(s) = \underline{\underline{Q}}_m(s)Q_n^{-1}(s) \qquad$ *for* $0 \leq m \leq n \leq N$ (6.10)

b) $\quad \underline{\underline{\mu}}_{mn} = \underline{\underline{Q}}_m(0)\underline{\underline{Q}}_n^{-1}(0)\underline{\underline{Q}}'_n(0)\underline{\underline{Q}}_n^{-1}(0) - \underline{\underline{Q}}'_m(0)\underline{\underline{Q}}_n^{-1}(0) \qquad$ *for* $0 \leq m \leq n \leq N$ (6.11)

Proof: Since $\underline{\underline{\sigma}}_{0m}(s)\underline{\underline{\sigma}}_{mn}(s) = \underline{\underline{\sigma}}_{0n}(s)$, part (a) follows from (6.5). From (a), we see that $\underline{\underline{\sigma}}_{mn}(s)\underline{\underline{Q}}_n(s) = \underline{\underline{Q}}_m(s)$. By differentiating this at $s = 0$, one finds that

$$\underline{\underline{\sigma}}'_{mn}(0)\underline{\underline{Q}}_n(0) + \underline{\underline{\sigma}}_{mn}(0)\underline{\underline{Q}}'_n(0) = \underline{\underline{Q}}'_m(0). \tag{6.12}$$

From (a) and the identity $\underline{\underline{\mu}}_{mn} = -\underline{\underline{\sigma}}'_{mn}(0)$, statement (b) now follows.

We note that $\underline{\underline{\sigma}}_n^+(s) = \underline{\underline{Q}}_n(s)\underline{\underline{Q}}_{n+1}^{-1}(s)$ implies that

$$\underline{\underline{s}}_n^+ = \underline{\underline{Q}}_n(0)\underline{\underline{Q}}_{n+1}^{-1}(0). \tag{6.13}$$

Thus, calculation of $\underline{\underline{Q}}_n(0)$, and $\underline{\underline{Q}}'_n(0)$ is sufficient to give $\underline{\underline{s}}_{mn}$ and $\underline{\underline{\mu}}_{mn}$ directly.

The matrices $\underline{\underline{Q}}_n(0)$ and $\underline{\underline{Q}}'_n(0)$ can be calculated recursively using (6.1) and (6.6) at $s = 0$. Knowledge of the matrices allows us to evaluate the mean ergodic exit time and mean stationary sojourn time, two useful measures of the dynamical behavior of the queue [10]. Both are defined with respect to a partition of the state space into two disjoint connected sets, called the good set and the bad set.

The ergodic exit time is the time required to leave the good set given that the system has reached its stationary state and is in the good set. Let a partition of the state space be defined by,

$$G_m = \{(j,\ n)\ :\ n < m\}\quad ,\quad B_m = \{(j,\ n)\ :\ n \geq m\}.$$

Since entry into B_m is at row m, the mean ergodic exit time from G_m, denoted by μ_{Em}, is given by

$$\mu_{Em} = \frac{\sum_{n<m} \underline{e}_n^T \underline{\underline{\mu}}_{mn} \underline{1}}{\sum_{n<m} \underline{e}_n^T \underline{1}}. \tag{6.14}$$

Using Theorem 6.3, this then leads to

$$\mu_{Em} = \frac{\sum_{n<m} \underline{e}_n^T \{\underline{\underline{Q}}_m(0)\underline{\underline{Q}}_n^{-1}(0)\underline{\underline{Q}}'_n(0)\underline{\underline{Q}}_n^{-1}(0) - \underline{\underline{Q}}'_m(0)\underline{\underline{Q}}_n^{-1}(0)\}\underline{1}}{\sum_{n<m} \underline{e}_n^T \underline{1}}. \tag{6.15}$$

The stationary sojourn time from G_m is the time required to first leave the good set given that the system was stationary and a transition into the good set from the bad set just occurred. The mean stationary sojourn time μ_{Vm} is given by

$$\mu_{Vm} = \frac{\underline{e}_m^T \underline{\underline{a}}_m^- \underline{\underline{\mu}}_{m-1}^+ \underline{1}}{\underline{e}_m^T \underline{\underline{a}}_m^- \underline{1}}. \tag{6.16}$$

This can be rewritten in a simpler form using a well known result [10] as $\mu_{Vm} = \frac{P_\infty(G_m)}{i_{B_m G_m}}$ where $P_\infty(G_m)$ is the total ergodic probability on the set G_m, and $i_{B_m G_m}$ is the ergodic probability flow from B_m to G_m. It then follows that

$$\mu_{Vm} = \frac{\sum_{n<m} \underline{e}_n^T \underline{1}}{\nu \underline{e}_m^T \underline{\underline{a}}_m^- \underline{1}}. \tag{6.17}$$

§7. A Tandem Queue Example

A tandem queue is a queueing system with distinct service facilities connected in series, i.e. the customer output stream of the first facility is the input stream of the second [13]. To illustrate the algorithmic procedure we have developed, one such tandem queue will be analyzed. The example selected has been chosen primarily for didactic reasons. A more complex and realistic example could have been analyzed with little increase in computational time.

The tandem queue evaluated is a two server series queue with blocking and finite buffers, see Section 1, Example C. The first service facility has 8 buffer slots and 4 servers. The second facility has 4 buffer slots and 5 severs. A flow diagram is given in Figure 7.1.

The corresponding rate matrices are given in Figure 7.2. When the first queue is full, customers balk and go elsewhere. When the second queue is full, the first queue stops serving until a space at the second facility becomes available. The model can easily be modified to allow different types of blocking, and features such as feedforward, feedback, etc.

In Figure 7.1 the occupancy level at the first facility is given by the coordinate j which corresponds to the number of people in service or waiting there. The occupancy level at the second facility is given by the coordinate n. The blocking may be seen in the absence of the transition rates associated with μ_1 on the top row.

Figure 7.2 displays the three matrices $\underline{\underline{\nu}}_n^0$, $\underline{\underline{\nu}}_n^+$, and $\underline{\underline{\nu}}_n^-$ when $n = 3$. The matrix $\underline{\underline{\nu}}_n^0$ is upper diagonal for all n. The only transitions which have no impact on the second queue are the arrivals (at rate λ) to the first queue, which increase that occupation level by one. The matrix $\underline{\underline{\nu}}_n^-$ is diagonal, with diagonal elements $\mu_2 = min\{n,\ k_2\}$ where k_2 is the number of servers in the second service facility. Finally, $\underline{\underline{\nu}}_n^+$ is lower diagonal. Increases in queue length at the second service facility are caused by departures from the first facility.

The results are shown in Figures 7.3 (a), (b), (c), 7.4 and 7.5 for a traffic intensity of 0.4 at the first facility and 0.6 at the second. Figures 7.3 (a), (b), (c) describe the ergodic probabilities as a function of $(j,\ n)$. An examination of these figures immediately reveals two features. First, the ergodic probabilities are significant for modest values of j and n, arising from the moderate traffic intensities. The small ridge visible along the upper edge of Figure 7.3 (a) indicates some blocking.

The ergodic probabilities may be useless when system parameters change before ergodicity is reached. The information contained in the first passage times may then be helpful. The ergodic distribution, for example, might cause concern over large probabilities of saturation, blocking, long waiting times and so on. From the last table in Figure 7.5 describing the mean first passage times from idleness, we see that the mean time to saturation is on the order of 4 days provided that our time scale is in hours. If this queue modeled a system that starts anew each day, one would be less inclined to worry.

The tables in Figure 7.4 give, for differing levels, the mean sojourn time, i.e., the mean time spent above a certain level after the level is reached. This gives a feeling, for example, for the persistence time in a congested state. The first two tables in Figure 7.5 present mean first passage times one step upward (downward). The $(j,\ n)$ entry in the table is the mean time to go from $(j,\ n)$ to a state in the row $n+1$ $(n-1)$. Again, the effect of blocking is noticeable in the dropoff (for the upward table) of the passage time with increasing values of j.

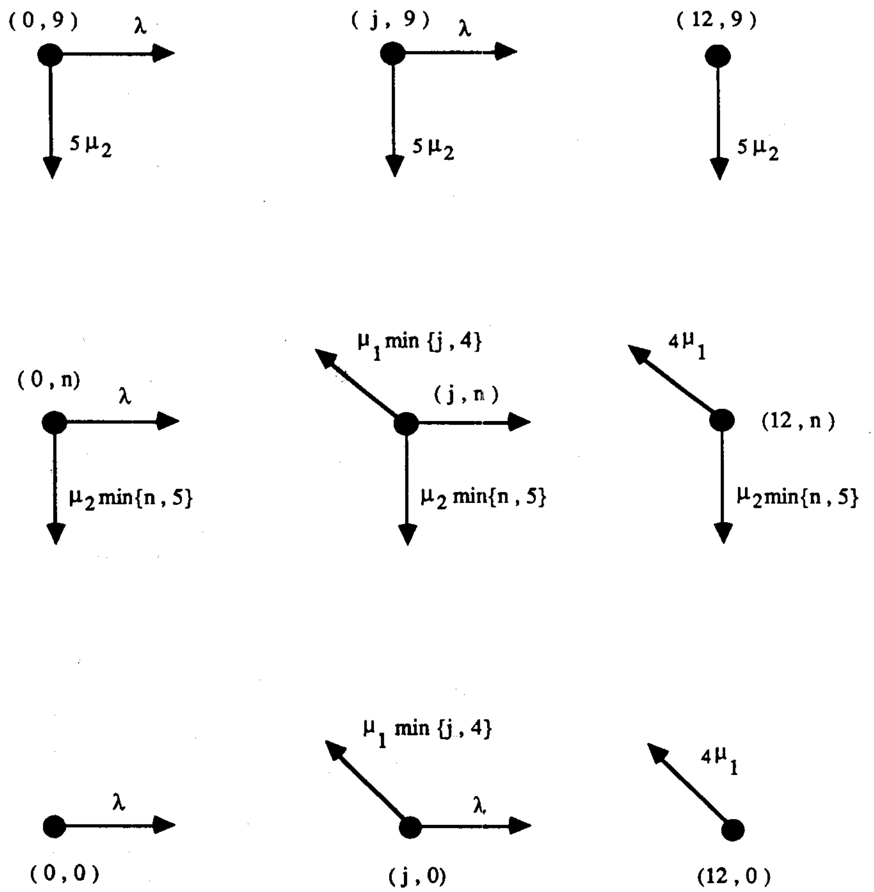

Fig. 7.1 Transition Rate Structure For The Tandem Queue

Transition rates are shown for representative edge, coner, and interior states.

13×13

$$\underline{\underline{\nu}}_n^0 = \begin{pmatrix} 0 & \lambda & 0 & \cdots & \cdots & 0 & 0 \\ \vdots & 0 & \lambda & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 0 & \lambda & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 & 0 & \lambda \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 \end{pmatrix}$$

13×13

$$\underline{\underline{\nu}}_3^- = \begin{pmatrix} 3\mu_2 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 3\mu_2 & 0 & \vdots & \vdots & \vdots & \vdots \\ \vdots & 0 & 3\mu_2 & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 3\mu_2 \end{pmatrix}$$

13×13

$$\underline{\underline{\nu}}_3^+ = \begin{pmatrix} 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \mu_1 & 0 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 2\mu_1 & 0 & \vdots & \vdots & \vdots & \vdots \\ \vdots & 0 & 3\mu_1 & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & 0 & 4\mu_1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 4\mu_1 & 0 \end{pmatrix}$$

TANDEM QUEUE EXAMPLE

SIZES: BUFFER (1) = 08 BUFFER (2) = 04 SERVES (1) = 04 SERVES (2) = 05

ARRIVAL RATE = 4.000 SERVICE RATE (1) = 2.500 SERVICE RATE (2) = 1.300

ERGODIC PROBABILITIES

I ↓ J →	0	1	2	3	4	5	6	7	8	9	10	11	12
9	0.0017	0.0038	0.0047	0.0042	0.0031	0.0020	0.0013	0.0008	0.0005	0.0003	0.0002	0.0001	0.0001
8	0.0038	0.0069	0.0067	0.0046	0.0025	0.0015	0.0009	0.0006	0.0003	0.0002	0.0001	0.0001	0.0000
7	0.0069	0.0117	0.0101	0.0060	0.0028	0.0014	0.0008	0.0004	0.0003	0.0001	0.0001	0.0001	0.0000
6	0.0118	0.0192	0.0158	0.0088	0.0038	0.0017	0.0008	0.0004	0.0002	0.0001	0.0001	0.0000	0.0000
5	0.0194	0.0313	0.0254	0.0138	0.0056	0.0024	0.0010	0.0005	0.0002	0.0001	0.0001	0.0000	0.0000
4	0.0317	0.0509	0.0410	0.0220	0.0089	0.0036	0.0015	0.0006	0.0003	0.0001	0.0001	0.0000	0.0000
3	0.0413	0.0662	0.0531	0.0284	0.0114	0.0046	0.0019	0.0008	0.0003	0.0001	0.0001	0.0000	0.0000
2	0.0403	0.0645	0.0516	0.0276	0.0111	0.0044	0.0018	0.0007	0.0003	0.0001	0.0001	0.0000	0.0000
1	0.0262	0.0419	0.0335	0.0179	0.0072	0.0029	0.0012	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000
0	0.0085	0.0136	0.0109	0.0058	0.0023	0.0009	0.0004	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000

Fig. 7.3(a) The Ergodic Probabilities For The Tandem Queue Model

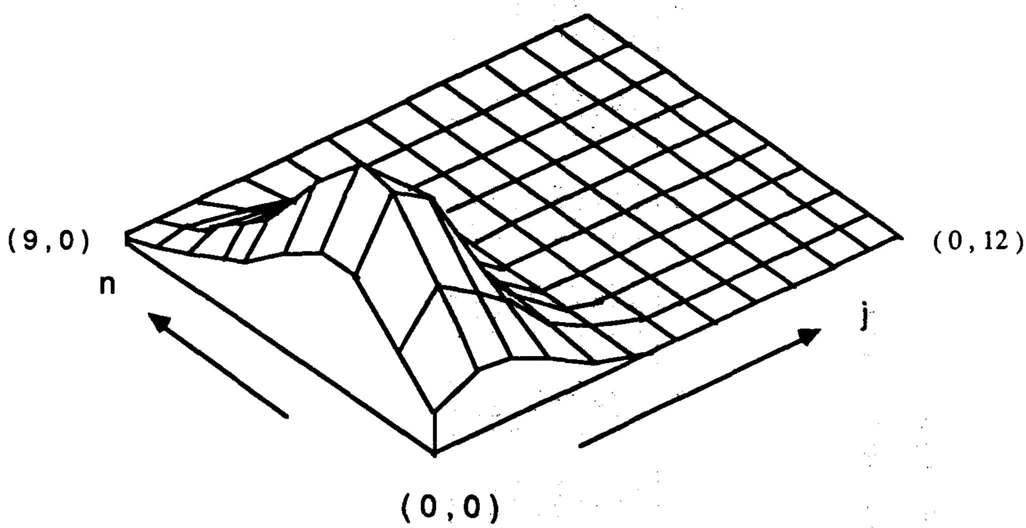

Fig. 7.3(b) The ergodic probabilities for the tandem queue model

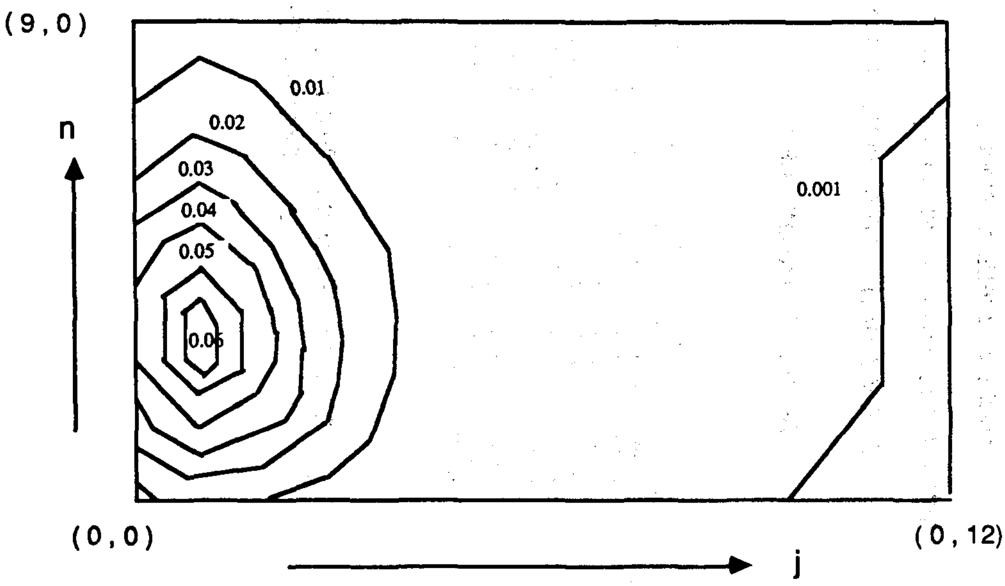

Fig. 7.3(c) The contours of the ergodic probabilities

MEAN SOJOURN TIMES ON THE SET { (J,N) : N ≥ X }

X =	1	2	3	4	5	6	7	8	9
	5.600746	1.568603	0.768045	0.496777	0.392214	0.379189	0.347591	0.277949	0.153846

MEAN SOJOURN TIMES ON THE SET { (J,N) : J ≥ X }

X =	1	2	3	4	5	6	7	8	9	10	11	12
	0.039110	0.014885	0.009001	0.007106	0.007581	0.008174	0.008832	0.009430	0.009751	0.009483	0.008210	0.005347

Fig. 7.4 The Mean Sojourn Time For The Tandem Queue Model

MEAN FIRST PASSAGE TIMES ONE STEP UPWARDS (↑N)

N ↓ J →	0	1	2	3	4	5	6	7	8	9	10	11	12
8	18.8602	12.3717	7.7164	4.6200	2.7606	1.9648	1.4850	1.1628	0.9357	0.7732	0.6594	0.5870	0.5534
7	11.3294	7.4297	4.6352	2.7777	1.6617	1.1814	0.8929	0.7024	0.5721	0.4824	0.4219	0.3847	0.3678
6	6.7328	4.3919	2.7236	1.6223	0.9665	0.6860	0.5235	0.4225	0.3583	0.3174	0.2919	0.2771	0.2707
5	3.9476	2.5233	1.5257	0.8898	0.5278	0.3828	0.3078	0.2670	0.2444	0.2319	0.2251	0.2216	0.2202
4	2.2708	1.3542	0.7750	0.4433	0.2735	0.2186	0.1958	0.1859	0.1816	0.1797	0.1789	0.1786	0.1785
3	1.4271	0.7981	0.4451	0.2636	0.1786	0.1590	0.1531	0.1513	0.1507	0.1505	0.1505	0.1505	0.1505
2	0.9615	0.5103	0.2888	0.1846	0.1374	0.1308	0.1296	0.1294	0.1294	0.1294	0.1294	0.1294	0.1294
1	0.6831	0.3502	0.2092	0.1451	0.1150	0.1131	0.1130	0.1130	0.1130	0.1130	0.1130	0.1130	0.1130
0	0.5055	0.2555	0.1652	0.1217	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000

MEAN FIRST PASSAGE TIMES ONE STEP DOWNWARDS (↓N)

N ↓ J →	0	1	2	3	4	5	6	7	8	9	10	11	12
9	0.1538	0.1538	0.1538	0.1538	0.1538	0.1538	0.1538	0.1538	0.1538	0.1538	0.1538	0.1538	0.1538
8	0.1848	0.2351	0.2839	0.3295	0.3672	0.3816	0.3871	0.3892	0.3900	0.3903	0.3905	0.3905	0.3905
7	0.1975	0.2685	0.3576	0.4581	0.5582	0.6226	0.6659	0.6950	0.7145	0.7275	0.7362	0.7417	0.7448
6	0.2033	0.2836	0.3920	0.5267	0.6748	0.7894	0.8820	0.9579	1.0202	1.0711	1.1121	1.1439	1.1648
5	0.2057	0.2900	0.4068	0.5573	0.7307	0.8766	1.0043	1.1180	1.2198	1.3107	1.3907	1.4583	1.5065
4	0.2728	0.3773	0.5185	0.6971	0.9027	1.0854	1.2535	1.4103	1.5574	1.6951	1.8224	1.9348	2.0175
3	0.4333	0.6057	0.8166	1.0622	1.3279	1.5609	1.7811	1.9925	2.1963	2.3924	2.5782	2.7457	2.8699
2	0.9427	1.3055	1.6983	2.1009	2.4898	2.8046	3.0961	3.3773	3.6516	3.9186	4.1742	4.4057	4.5761
1	4.0319	5.0923	6.0317	6.8311	7.4829	7.9339	8.3233	8.6893	9.0436	9.3862	9.7177	10.0143	10.2292

MEAN FIRST PASSAGE TIME FROM IDLE TO GIVEN N

1	2	3	4	5	6	7	8	9
0.5055	0.9382	1.4672	2.2167	3.4175	5.5906	9.3470	15.6818	26.2116

Fig. 7.5 The Mean First Passage Times For The Tandem Queue Model

Acknowledgment

The authors wish to thank the referees for helpful comments, S. Graves for useful discussions and Y. Ishikawa for his extensive editorial assistance.

References

[1] Cinlar, E., "Markov Renewal Theory : A Survey," **Management Science**, Vol.21, pp.727-752, 1975.

[2] Friedman, D., "Queueing Analysis of A Dynamically Shared Voice-Data Link", Ph.D. Thesis, Department of Electrical Engineering and Computer Science, M.I.T, 1981.

[3] Gaver, D. P., Jacobs P. A. and Latouche, G., "Finite Birth-and-Death Models in Randomly Changing Environments," **Advances in Applied Probability**, Vol.16, pp.715-731, 1984.

[4] Hajek, B.,"Birth-and-Death Processes on Integers with Phases and General Boundaries," **Journal of Applied Probability**, Vol.19, pp.448-499, 1982.

[5] Karlin, S. and McGregor, J. L., "The Differential Equations of Birth-and-Death Processes and the Stieltjes Moment Problem", **Transactions of American Mathematical Society**, Vol.85, pp.489-546, 1957.

[6] Karlin, S. and McGregor, J. L., "The Classification of Birth and Death Processes", **Transactions of American Mathematical Society**, Vol.86, 1957.

[7] Keilson, J., "A Review of Transient Behavior in Regular Diffusion and Birth-Death Processes", **Journal of Applied Probability**, Vol.1, pp.247-266, 1964.

[8] Keilson, J., "A Review of Transient Behavior in Regular Diffusion and Birth-Death Processes – Part II", **Journal of Applied Probability**, Vol.2, pp.405-428, 1965.

[9] Keilson, J., "On the Matrix Renewal Function for Markov Renewal Processes", **The Annals of Mathematical Statistics**, Vol.40, pp.1901-1907, 1969.

[10] Keilson, J., **Markov Chain Models – Rarity and Exponentiality**, Springer-Verlag, New York, 1979.

[11] Keilson, J. and Wishart, D. M. G., "A Central Limit Theorem for Processes Defined on a Finite Markov Chain," **Proceeding of the Cambridge Philosophical Society**, Vol.60, pp.547-567, 1964.

[12] Keilson, J. and Wishart, D. M. G., "Boundary Problems for Additive Processes Defined on a Finite Markov Chain," **Proceeding of the Cambridge Philosophical Society**, Vol.61, pp.173-190, 1965.

[13] Kleinrock, L., **Queueing Systems, Vol. 2: Computer Applications**, John Wiley & Sons, New York, 1976.

[14] Neuts, M. F., "Markov Chains with Applications in Queueing Theory Which Have a Matrix-geometric Invariant Vector", **Advances in Applied Probability**, Vol.10, pp.185-212, 1978.

[15] Neuts, M. F., **Matrix Geometric Solutions in Stochastic Models – An Algorithmic Approach**, The Johns Hopkins University Press, 1981.

[16] Pyke, R., "Markov Renewal Processes : Definition and Preliminary Properties," **Annals of Mathematical Statistics**, Vol.35, pp.1746-1764, 1964.

[17] Schweitzer, P., "Aggregation Methods for Large Markov Chains," Invited Paper in the **Proceedings of Mathematical Computer Performance and Reliability**, P. J. Courtois and A. Hordijk (editors), pp.275-286, 1984.

[18] Takahashi,Y., "A Lumping Method for Numerical Calculations of Stationary Distibutions of Markov Chains," Research Report B-18, Department of Information Sciences, Tokyo Institute of Techinology, Tokyo, Japan, 1975.

[19] Takahashi, Y. and Takami, Y., "A Numerical Method for the Steady-State Probabilities of a GI/G/C Queueing System in a General Class," **Journal of Operations Research Society of Japan**, Vol.19, No.2, pp.147-157, 1976.

Ushio SUMITA: William E. Simon
Graduate School of Business Administration,
University of Rochester,
Rochester, NY 14627, U.S.A.