**THESIS REPORT**
*Master's Degree*

SYSTEMS
RESEARCH
CENTER

*Supported by the*
*National Science Foundation*
*Engineering Research Center*
*Program (NSFD CD 8803012),*
*Industry and the University*

# Tandem Queueing Systems Subject to Blocking with Phase Type Servers: Analytic Solutions and Approximations

*by: L. Gün*
*Advisor: A.M. Makowski*

M.S. 87-1
*Formerly TR 87-2*

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*

| | | |
|---|---|---|
| 1. REPORT DATE<br>**1986** | 2. REPORT TYPE | 3. DATES COVERED<br>**00-00-1986 to 00-00-1986** |

4. TITLE AND SUBTITLE
**Tandem Queueing Systems Subject to Blocking with Phase Type Servers: Analytical Solutions and Approximations**

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
**University of Maryland,The Graduate School,2123 Lee Building,College Park,MD,20742**

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT
**see report**

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT<br>**unclassified** | b. ABSTRACT<br>**unclassified** | c. THIS PAGE<br>**unclassified** | | **160** | |

# TANDEM QUEUEING SYSTEMS SUBJECT TO BLOCKING WITH PHASE TYPE SERVERS : ANALYTIC SOLUTIONS AND APPROXIMATIONS

*by*

Levent Gün

Thesis submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Master of Science
1986

# ACKNOWLEDGMENTS

I wish to express my deep-felt gratitude to my advisor Prof. Armand M. Makowski for his guidance and generous assistance during this research. I also thank Dr. Harry Heffes for many useful comments and suggestions for this thesis and for his excellent supervision during my stay at Bell Laboratories. Also, my thanks are due to the financial supporters of this research who made this reasearch possible: National Science Foundation, through NSFD Grant NSFD CDR-85-00108, Office of Naval Research, through Grant N00014-84-K-0614 and AT&T Bell Laboratories. Finally, I would like to pay tribute to my wife, Hediye, who have stood by me and supported me over the past several years.

# TABLE OF CONTENTS

**Chapter VII: An approximation method for general tandem queueing systems subject to blocking**

**Appendix**

# LIST OF TABLES

## LIST OF FIGURES

**Figure** **Page**

# CHAPTER I

# INTRODUCTION

Consider a production system composed of many processing stages (servers) through which the material (jobs) must pass in some prespecified order. The performance of such a system is impaired by variations in servers behavior due either to failures or to fluctuations in service times. The effects of these variations can be mitigated by using intermediate storage spaces (buffers) between the servers. However, because of physical limitations on buffer spaces and variations in the service times, the flow of jobs through the system may get *blocked.* Queueing systems with blocking have a wider applicability in that they can also be used to model computer systems, telecommunication networks and distributed systems, to name a few applications.

This thesis is devoted to various analytical and algorithmic issues for *tandem* queueing systems with blocking. The blocking systems considered here are composed of a series arrangement of service stations with *finite* intermediate buffers between these stations. A typical tandem configuration is shown in Figure 1.1.1. At each station some work is performed on a job which is then passed on to the next station and finally ejected from the last station. It is assumed here that a *single* server which operates according to the FCFS (first-come-first -served) queueing discipline is in attendance at each station.

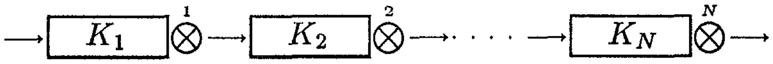

**Figure 1.1.1**

The literature considers two distinct blocking policies for transfer lines with finite buffers. Let S be one of the servers in Figure 1.1.1. Under the first policy, called *immediate blocking*, at a time of service completion, the server S is blocked if the downstream buffer becomes full due to this service completion; the server S remains blocked until the congestion is reduced downstream, at which time it resumes service and *begins* to process its next job (if any). Under the second policy, called *non-immediate* blocking, the server S is blocked at a service completion time if the job that has just completed service cannot proceed to the next buffer due to congestion. When the congestion is reduced downstream, this job proceeds to the next buffer without receiving any further service at server S, and the server S resumes service and begins processing its next job (if any). As noted in Altıok and Stidham [4], these two blocking policies are *not* equivalent in the sense that the solution of the system under one policy cannot be obtained from the solution of the system (with possibly different parameter values) under the other policy, and therefore cannot be reduced to each other, except when N=2.

Queueing networks with blocking are typically very difficult to analyze, and closed form results at steady state (or otherwise) are usually not available, except in tandem models with two exponential servers and a finite intermediate buffer. This model is probably the simplest configuration of an open queueing network with blocking. A rigorous Markovian analysis of this system was performed in the pioneering work of Hunt [36], who upon enumerating the different possible states in which the system can exist, did set up and solve the equations for the state probabilities in statistical equilibrium. Since then, queueing networks with blocking have been studied by researchers from different research communities. To fix the terminology, a brief classification of the blocking systems discussed in the literature is provided in the Appendix, where an annotated bibliography of some of these papers has been compiled. A basic assumption that was made in all the models studied in the surveyed literature is that *the last stage is never blocked*, i. e., there is always space available for a job whose service has been completed at the last server. Three classes of systems have been studied:

**(i) Systems without failures :** In this class, the service stations in Figure 1.1.1 are attended by *reliable* servers. As long as there is a part to process and the server is not blocked, it can always give service according to some known service distribution. Authors like Altıok [2], Clarke [19,20], Foster and Perros [25], Hildebrand [32,33], Hunt [36], Konheim and Reiser [40], Latouche and Neuts [45], Lavenberg [46], Neuts

[54,57], Perros and Altıok [63] and Pinedo and Wolf [64] allow an infinite capacity input buffer upstream of the first server and assume jobs to arrive according to some statistical pattern. Others, like Hatcher [31], Hillier and Boling [34], Knott [40], Muth [51,52] and Rao [66], assume that there is an inexhaustible supply of jobs to the first server, and so the first server is never starved. In most of the models studied in the literature, servers are assumed to have exponentially distributed service times. Exceptions are Hiller and Boling [34] which allow Erlang distributions, and Muth [51,52] and Rao [66] who postulate more general service time distributions.

**(ii) Systems subject to failures :** The service stations are attended by *unreliable* servers, i. e., servers subject to failures which are non-deterministic in both occurence and duration. This class of models can be further divided into two subclasses according to the assumptions imposed on the service times.

**(ii.a) Deterministic processing times :** The service times are assumed to be *deterministic* while both the failure and repair times of the servers are allowed to be *random.* All the models discussed in the literature postulate that the *first* server is never starved, and that the duration of up and down times are geometrically distributed. Relevant papers include the work of Artamonov [6], Buzacott [15,16], Buzacott and Hanifin [17], Freeman [24], Gershwin and Ammar [27], Gershwin and Schick [28], Ignall and Silver [37], Masso and Smith [48], Muth and Yeralan [53], Ohmi [60], Okamura and Yamashina [61], Sheskin [68], Soyster, Schmidt and Rohrer [69] and Wijngaard [76].

**(ii.b) Random processing times :** Service times as well as the failure and repair times are *random,* as the case in the work of Buzacott [16], and Gershwin and Berman [28]. In both papers, the failure and repair times and the service times were assumed to be exponentially distributed. Only two stages were considered with the assumption that the first server is never starved.

**(iii) Flow models :** In this class of models, jobs are not treated as discrete items but rather as a *continuous* fluid. This model was considered in papers by Buzacott and Hanifin [17] and Wijngaard [76], where the servers were subject to failures with deterministic service times as in class ii.a.

To the author's knowledge, the *exact* closed-form solution of the *joint* queue length distribution at steady-state has not yet been reported in the literature, and this even in the simplest of models under exponential assumptions. One of the contributions of this thesis is to identify a class of two node tandem systems with blocking for which the *joint* queue length distribution at steady-state can be obtained in closed-form.

In this thesis attention is given to the analytical aspects of various models of two node tandem systems. Although two node blocking systems may not constitute very realistic models for many real life systems, these simple systems could serve as *building* blocks for approximating general tandem queueing systems with blocking. Indeed, many approximation schemes developed for general tandem queueing systems, e.g., Altıok [1], Brandwajn and Jow [13], Hillier and Boling [34] and Sheskin [67], to name a few, are based on *decomposition* and *isolation* algorithms that reduce the problem to one of simultaneously solving several simple systems which are special cases of the models studied here. Such ideas clearly motivate a careful study of the *general* two node tandem systems with blocking, with a view towards providing exact analytical results so as to enhance existing and future approximation schemes.

The focus of the work reported here is on models which are in classes (i) and (ii.a). As a rule, the models presented here adopt *immediate blocking* and include servers with *phase type* (PH-type) service distributions. For the benefit of the unfamiliar reader, Chapter II includes a brief introduction to PH-type distributions. With the exception of the model discussed in Chapter V where an an arrival process to the first buffer in Figure 1.1.1 is considered, the system is assumed *saturated* in that the input queue in Figure 1.1.1 contains, at all times, an infinite supply of jobs, so that the first node server is *never starved.* In all cases, it is assumed that the server in the last node is *never blocked.*

In Chapter III, a two node tandem system with an intermediate buffer of finite capacity is analyzed when *both* servers have PH-type service distributions. Although the discussion focuses on the discrete-time formulation, results for the continuous-time model are also presented. A system state process is introduced and its probabilistic structure is carefully described. In contrast with the continuous-time model, the discrete-time formulation may give rise to a *non-irreducible* Markov chain. Necessary and sufficient conditions for its irreducibility are given, and its ergodic and transient states are identified in terms of the parameters defining the service distributions. Moreover, the steady state probabilities of the system states are shown to have a *matrix geometric* form and *closed form* expressions are obtained for the quantities of interest. The effect of reversing the order of the servers is discussed and it is observed that in statistical equilibrium the probability of finding $i$ customers in the buffer in one case coincides with finding $K-i$ customers in the case when the order of the servers is changed, $K$ being the buffer capacity. The results are illustrated through numerical examples on some well-known PH-type distributions.

In Chapter IV, the general model of Chapter III is specialized to the case where one of the servers is *geometric.* The discussion focuses on the situation where the *first* and *second* node servers have *geometric* and *PH-type* service distributions, respectively. The case when the order of the servers is reversed is only briefly discussed, as the results follow from the discussion in Chapter III. A system state process is introduced and its probabilistic structure is carefully described. For this special case, it is shown that the steady state probability vector is *always unique* and a *complete characterization* of the system states is given by identifying the irreducible and transient states of the underlying Markov chain. Explicit expressions for the steady state probabilities are obtained and attempts are made to give an expression for the average system throughput, in order to study its behavior as a function of the intermediate buffer capacity. The continuous-time formulation of the model is also briefly discussed and similar results are obtained. The discussion on this particular model concludes by applying the results to some specific PH-type distributions.

In Chapter V, a two node system with PH-type servers at *both* nodes (as in Chapter III) is considered by relaxing the assumption that the first node server is never starved to capture the situation of a Bernoulli arrival stream to a *finite* capacity queue in front of the first node server. A system state is defined and the probabilistic structure of the corresponding process is described. Neccessary and sufficient conditions for the irreducibility of the Markov chain are obtained in terms of the system parameters. It is shown that the steady state probability vector is *always* unique, and can be obtained in *closed* form by grouping the states in pairs. The *joint* queue length distribution at steady state is then easily obtained as entries of the steady state probability vector of these auxilary states. Although the discussion focuses on the discrete-time formulation, the continuous-time model is also discussed and the irreducibility of the corresponding Markov process is shown.

In Chapter VI, *unreliable* servers with *PH-type* up and down time distributions are considered. The *effective* service time distributions of such servers are represented by PH-distributions of higher order. Although the results of Chapter III do not directly apply the exact same methodology can be used. Specifically, when the idling servers are subject to failure, the transition probabilities among the boundary states are explicitly written and a *matrix geometric* solution is derived for the steady state probabilities through calculations similar to the ones given in previous chapters. The case when only operational servers can fail is obtained as a special case of this discussion. The results are illustrated by several numerical examples.

Finally, in Chapter VII an iterative approximation scheme that uses the two node system as a building block is presented for solving general tandem queueing systems with *finite* capacity intermediate buffers and *PH-type* servers. Since the algorithm is based on a two node system it is applicable for both *immediate* and *nonimmediate* blocking since these two blocking policies can be reduced to each other for two node systems as mentioned earlier. However only immediate blocking is considered here. The approximation algorithm gives results in the form of marginal probability distribution of the number of jobs in each queue of the tandem configuration so that the line throughput can be computed. In view of the results derived in Chapter VI, this approximation scheme also applies to failure type servers with PH-type service and repair distributions. The algorithm is presented and its accuracy tested through numerical examples both for continuous and discrete-time systems, under both types of blocking. Comparison of the results against simulations indicate reasonable accuracy even in the presence of significant blocking. Program listings are available from the author upon request.

# CHAPTER II

# MOTIVATION AND BACKGROUND

## II.1. Motivation :

The work reported in this thesis aims at understanding the effect of blocking on system performance through a detailed study of two node tandem models. Although blocking systems with only two or three nodes may not be realistic models for most applications, they can be thought out as *building* blocks for approximating general queueing networks subject to blocking. In this section, a few examples are given as to how several authors have used such simple models to approximate more general tandem queueing systems.

The first example is borrowed from Brandwajn and Jow [13], where the solution of a two node system is used as a building block for an approximate analysis of systems of tandem queues with blocking. *Reliable* and *exponential* servers with load dependent service rates were considered. The method uses the notion of *equivalence* to produce approximations to the joint queue length probability distributions for pairs of neighboring servers and to various performance measures for individual servers. The proposed method is applicable to both types of blocking since they are equivalent for two node systems.

The second example is from Hillier and Boling [34], who consider a general tandem manufacturing system with a single *exponential* server at each node and *finite* intermediate buffers between the nodes. They develop an efficient algorithm to compute the steady state output rate via single node analysis. The effective mean arrival rate and the effective mean service rate at a particular node are obtained by considering the remainder of the system as a *black box* which generates input to this particular node and accepts the output of this particular node. Each node can therefore be viewed as a single server queueing model with Poisson input, exponential service times and a finite buffer. This already suggests that, provided an interconnection methodology were available, tandem configurations of PH-type servers could be approximated by several isolated two node models of the type studied in Chapter IV.

The third example comes from Sheskin [68], where the servers are subject to failures. He studied a general tandem line where service times are *constant* whereas failure and repair times are random, i. e., a model in class (ii.a) described earlier. The transitions between the up and down states of the servers are assumed to occur according to a *time-homogenous* Markov chain. For larger systems, Sheskin proposes to approximate the system by analyzing each server separately by ignoring the dependence between the arrivals and the departures to a node. Hence, this *decomposition* algorithm reduces the problem to several simple problems. Due to the memoryless nature of the up and down time distributions, the service distribution of such a failure server can be represented by a PH-type distribution with only two phases, the so-called *up* and *down* phases of the service. This decomposition procedure yields simple systems which are special cases of the models studied here. It is noteworthy that, this decomposition algorithm is reported to approximate general systems within 5% of relative error, without requiring much computer time and memory.

The last example was considered by Altıok [1], who studied a model in class (ii.b) with independent *Erlang* type service distributions, *exponential* up times and *arbitrary* down time distributions. Although expressions for the cumulative distribution of service completion times can be obtained, they are very complicated and not readily implemented. When the down times have also exponential distributions, Altıok proposes to assume that during the processing time of a job at most one failure may occur and uses this fact to incorporate failures into the service completion times. These service completion times are then approximated by specific PH-type distributions through empirical observations, and Then, by using the results of Perros and Altıok [63], the effective service time distribution of the $i^{th}$ server is represented with a phase structure involving $2 \times (M - i + 1)$ phases, where $M$ is the total number of nodes in the system. Again, the model is approximated by simple systems (of the type studied here) in isolation.

The last two examples suggest that even models with unreliable servers can be approximated by simultaneously solving several simple models. Since these models are of the type discussed in Chapter III, it is particularly important that explicit solutions for such models be obtained and that their algorithmic properties be discussed in some detail.

In order to capture a fairly general class of service time distributrions PH-type servers are considered. To fix the notation and terminology for the unfamiliar reader it is convenient at this point to briefly describe properties of the class of PH-type distributions. For additional information on this topic the reader is invited to consult the monograph by Neuts [58].

## II.2. Phase Type Distributions :

Probability distributions of the *phase-type*, denoted by *PH-type,* were introduced both in discrete and continuous-time by M.F. Neuts [58]. Loosely speaking, a ***discrete-time*** PH-type distribution is any probability distribution on the nonnegative integers obtainable as the distribution of the time until ***absorption*** in a discrete-time finite state space Markov chain with a *single* absorbing state. The class of PH-type distributions includes well known distributions such as the Generalized Negative Binomial and Hypergeometric distributions, as well as any distribution with *finite* support on the nonnegative integers. The usefulness of these distributions to the algorithmic solution of many queueing models has been forcefully advocated by Neuts for quite some time now, and is illustrated in some detail in [58]. The reader is also refered to [11] and [55] for a discussion of the many interesting properties enjoyed by PH-type distributions.

Specifically, consider a discrete-time Markov chain on the state-space $S = \{1,,2,\ldots,m+1\}$, where the states $\{1,\ldots,m\}$ are *transient* and the state $m+1$ is *absorbing.* The chain starts in state $i \in S$ with probability $\alpha_i$ and evolves in accordance with the one-step transition matrix $P$. It is customary to write the one-step transition matrix $P$ and the initialization probabilities in the form

$$P = \begin{pmatrix} Q & p \\ 0_m & 1 \end{pmatrix} \quad , \ (\alpha, \alpha_{m+1}) \qquad \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m) \tag{2.2.1}$$

where

$Q$ is a $m \times m$ substochastic matrix,

$p$ is a $m \times 1$ column vector of absorption probabilities to the absorbing state $m+1$ from the transient states $\{1,\ldots,m\}$,

$\alpha$ is a $1 \times m$ row vector of initialization probabilities,

$0_m$ is a $1 \times m$ row vector with zero entries.

The following relationships

$$p = (I_m - Q)\, e_m \quad , \qquad\qquad \alpha_{m+1} = 1 - \alpha\, e_m$$

obviously hold, where $I_m$ is the $m \times m$ identity matrix and $e_m$ is the $m \times 1$ column vector of ones.

It is assumed that the matrix $(I_m - Q)$ is *nonsingular* or equivalently, that the states $\{1, 2, \ldots, m\}$ are all *transient*, so that the absorption into the state $m+1$, from any initial state, is certain [58, p. 45].

The *PH-type* probability distribution associated with the pair $(\alpha, Q)$ is the distribution function $F(\cdot)$ on the non-negative integers with probability mass function $\{q_k\}_1^\infty$ given by

$$q_k = \begin{cases} \alpha_{m+1} \ , & k = 0 \ , \\ \alpha Q^{k-1} p \ , & k \geq 1 \ . \end{cases} \tag{2.2.2}$$

The pair $(\alpha, Q)$ is called the *representation* of the distribution $F$, which it uniquely determines. The converse is not true in that a given distribution function $F(\cdot)$ admits infinitely many PH-representations. The probability generating function $P(\cdot)$ of the PH-distribution with representation $(\alpha, Q)$ is given for all $0 < z \leq 1$ by

$$\begin{aligned} P(z) &= \sum_{k=0}^{\infty} q_k \ z^k \\ &= \alpha_{m+1} + z\alpha(I_m - zQ)^{-1} p \qquad (2.2.3) \\ &= \alpha_{m+1} + 1 - \left(\frac{1-z}{z}\right) \alpha \left(\frac{1}{z} I_m - Q\right)^{-1} e_m \ . \end{aligned}$$

The following probabilistic construction is given in Neuts [58, p. 48]. Upon absorption into state $m+1$, *independent* multinomial trials with probabilities

$(\alpha, \alpha_{m+1})$ are *instantaneously* and *repeatedly* performed until one of the alternatives $1, 2, \ldots, m$ occurs. The process is restarted in the corresponding state and the same procedure is repeated at the next absorption. It is easy to show that upon indefinitly continuing this procedure, a new Markov process is constructed on $\{1, 2, \ldots, m\}$ in which the state $m+1$ is an *instantaneous* state. The corresponding one-step probability matrix is readily seen to be

$$Q^* = Q + \frac{1}{1 - \alpha_{m+1}} \, p \, \alpha \tag{2.2.4}$$

The successive visits to the instantaneous state form a *renewal* process with the underlying distribution $F$, called a *PH-renewal process.* The representation $(\alpha, Q)$ is said *irreducible* if and only if the Markov chain on $\{1, \ldots, m\}$ with one-step probability matrix $Q^*$ is *irreducible.*

As shown in [58, pp. 49-50], every PH-type distribution function $F$ admits an irreducible representation, whence only irreducible representations will be considered hereafter. Moreover, the condition $\alpha_{m+1} = 0$ will be assumed in order to avoid complications of limited interest in applications. Under this assumption, it is plain that $\alpha \neq 0_m$, and that $p \neq 0_m^T$, where $^T$ denotes transposition, for otherwise the matrix $Q$ would be stochastic, thus contradicting the nonsingularity assumption of the matrix $(I_m - Q)$.

To fix ideas, the representations of some well-known PH-type distributions are displayed below; they will be used in the numerical examples of later chapters. In these examples, $0 < \mu_i < 1$ for $1 \leq i \leq n$.

$(i)$: A *Negative Binomial* distribution of order $n$ is represented by the pair $(\alpha, Q)$, where

$$\alpha = (1, 0, 0, \ldots, 0) = (1, 0_{n-1})$$

and

$$Q = \begin{pmatrix} (1-\mu_1) & \mu_1 & & & & \\ & (1-\mu_2) & \mu_2 & & & \\ & & \cdot & \cdot & & \\ & & & \cdot & \cdot & \\ & & & & \cdot & \cdot \\ & & & & (1-\mu_{n-1}) & \mu_{n-1} \\ & & & & & (1-\mu_n) \end{pmatrix} . \tag{2.2.5}$$

$(ii)$: A *Hypergeometric* distribution of order $n$ is represented by the pair $(\alpha, Q)$, where

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n) \quad \text{and} \quad Q = diag(1-\mu_1, 1-\mu_2, \ldots, 1-\mu_n) \ . \tag{2.2.6}$$

$(iii)$: A *Deterministic* service time of $n$ units is represented by the pair $(\alpha, Q)$, where

$$\alpha = (1, 0, 0, \ldots, 0) = (1, 0_{n-1})$$

and

$$Q = \begin{pmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ & & \cdot & \cdot & & \\ & & & \cdot & \cdot & \\ & & & & 0 & 1 \\ & & & & & 0 \end{pmatrix} . \tag{2.2.7}$$

PH-type representations can also be introduced in a similar way for distributions on $[0, \infty)$. In this case, a Markov process on the states $\{1, \ldots, m+1\}$ with infinitesimal generator $\begin{pmatrix} Q & p \\ 0_m & 0 \end{pmatrix}$ is considered, where the $m \times m$ matrix $Q$ satisfies the conditions $Q_{ii} < 0$ and $p_i \geq 0$ for $1 \leq i \leq m$, and $Q_{ij} \geq 0$, $1 \leq i \neq j \leq m$. Moreover, $Qe_m = -p$ and the initial probability vector of $Q$ is given by $(\alpha, \alpha_{m+1})$, where $\alpha$ is defined as before. It is again assumed that the states $\{1, \ldots, m\}$ are all *transient,* or equivalently, that the matrix $Q$ is *nonsingular* [58, p. 45].

A probability distribution $F(\cdot)$ on $[0,\infty)$, is a *distribution of PH-type* if and only if it is the distribution of the time until absorption in a finite Markov process of the type defined above. In that case, the distribution function $F(\cdot)$ can be expressed as

$$F(x) = 1 - \alpha\, exp(Qx) e_m \quad , \qquad x \geq 0 \; , \tag{2.2.8}$$

with a jump at $x = 0$ of height $\alpha_{m+1}$ and its density $F'(x)$ on $(0,\infty)$ given by

$$F'(x) = \alpha\, exp(Q\, x)\, p \;, \quad x > 0 \; . \tag{2.2.9}$$

The pair $(\alpha, Q)$ is again called the *representation* of the distribution $F$.

## II.3. Preliminaries :

In this section, several definitions and results of interest from the theory of nonnegative matrices are presented, and the nonsingularity of several matrices of interest for subsequent developments is proved. The reader is referred to Berman and Plemmons [10] for a general reference on the theory of nonnegative matrices. Throughout the discussion, all the matrices have real entries unless otherwise mentioned. The notation $A \geq 0$ is used whenever each entry of the vector or matrix $A$ is nonnegative whereas $A > 0$ is used if $A \geq 0$ and at least one entry is positive, and $A >> 0$ is used if *all* entries of $A$ are positive.

**Definition 2.3.1.** : *An $n \times n$ nonnegative matrix $A$ is* **cogredient** *to an $n \times n$ matrix $E$ if for some permutation matrix $P$, $PAP^T = E$. The matrix $A$ is said to be* **reducible** *if it is cogredient to*

$$E = \begin{pmatrix} B & 0 \\ C & D \end{pmatrix},$$

*where $B$ and $D$ are square matrices, otherwise, $A$ is said to be* **irreducible.**

**Definition 2.3.2.** : *The* **directed graph** *$G(A)$ associated with an $n \times n$ nonnegative matrix $A$ is the graph made up of $n$ vertices, say $P_1, P_2, \ldots P_n$, with an edge leading from $P_i$ to $P_j$ if and only if $A_{ij} > 0$, $1 \leq i, j \leq n$.*

It is well known [10, p. 30] that a matrix $A$ is *irreducible* if and only if $G(A)$ is *strongly connected*, that is, for every ordered pair $(P_i, P_j)$ of vertices of $G(A)$, there exists a path, i. e., a sequence of edges which leads from $P_i$ to $P_j$. The following Lemma is a simple application of the Definition 2.3.2 and will be useful in the proofs of Theorems 4.3.2. and 5.3.1.

**Lemma 2.3.3.** : *For any two nonnegative square matrices $A$ and $B$, the matrix $(A+B)$ is irreducible if and only if the matrix $(c_1 A + c_2 B)$ is irreducible for all scalars $c_1, c_2 > 0$.*

**Proof :** Owing to the nonnegativity of the matrices $A$ and $B$, $(A+B)_{ij} > 0$ if and only if $(c_1\,A + c_2\,B)_{ij} > 0$, whenever $c_1, c_2 > 0$, while $(A + B)_{ij} = 0$ if and only if $A_{ij} = B_{ij} = c_1\,A_{ij} + c_2\,B_{ij} = (c_1\,A + c_2\,B)_{ij} = 0$ . The topology of the directed graphs $G(A + B)$ and $G(c_1\,A + c_2\,B)$, $c_1$, $c_2 > 0$, are thus exactly same, and the result follows by Definition 2.3.1.

△

When a nonnegative matrix is a *stochastic* one, the notion of irreducibility in the Definition 2.3.1 is related to the probabilistic one in that every stochastic matrix can be viewed as the one-step transition matrix of a Markov chain. In order to clarify the terminology used hereafter, it is imperative to classify the states of a Markov chain.

Consider a finite state Markov chain with the state space $S = \{s_1, \ldots, s_n\}$.

**Definition 2.3.4. :** *If in the state transition diagram of the Markov chain $S$ there exists a path from state $s_i$ to state $s_j$, $s_i, s_j \in S$, then the state $s_i$ is said to have* **access** *to state $s_j$, written $s_i \longrightarrow s_j$. If $s_i$ has access to $s_j$ and $s_j$ has access to $s_i$, then $s_i$ and $s_j$ are said to* **communicate**, *written as $s_i \leftrightarrow s_j$.*

Implicit in Definition 2.3.4 is that every state in $S$ communicates with itself, and with this convention, the communication relation is an *equivalence* relation on the set of states and thus partitions $S$ into equivalence classes. With this in mind, the following definition is given.

**Definition 2.3.5. :** *A state $s_i \in S$ is called* **transient** *if there exists some $s_j \neq s_i$ in $S$ with the property that $s_i \longrightarrow s_j$ but $s_j \not\rightarrow s_i$, that is, $s_i$ has access to some other state which does not have access to $s_i$. Otherwise, the state $s_i$ is called* **ergodic.**

Thus, $s_i$ is ergodic if and only if $s_i \longrightarrow s_j$ implies $s_j \longrightarrow s_i$ for some $s_j \neq s_i$ in $S$. It follows that if one state in an equilance class of states associated with a Markov chain is transient (resp. ergodic), then each state in that class is transient (resp.

ergodic). This leads to the next definition.

**Definition 2.3.6. :** *A class induced by the communication relation on the set $S$ is called* **transient** *if it contains a transient state and* **ergodic** *otherwise.*

**Definition 2.3.7. :** *A Markov chain is called* **irreducible** *if it consists of a single ergodic class.*

In view of the above definitions, the following Lemma can easily be proved.

**Lemma 2.3.8. :** *A finite state Markov chain is irreducible if and only if the corresponding one-step transition matrix is irreducible.*

Therefore, these two concepts of irreducibility can and will be used interchangeably hereafter. The following notion of *reachability* will be used repeateadly in later chapters.

**Definition 2.3.9 :** *Let $U$ and $V$ be subsets of the state space of a Markov chain with corresponding one-step probability transition matrix $T$. The set $U$ is* **reachable** *from the set $V$ if there is a path from some state in $V$ to a state in $U$ in the directed graph $G(T)$ of the matrix $T$.*

The following class of matrices is very useful in many applications.

**Definition 2.3.10 :** *An $n \times n$ matrix $A$ is called an* **M-matrix** *if it is of the form $A = sI_n - B$ for some $s > 0$ and $B \geq 0$, with $s \geq \rho(B)$, where $\rho(B)$ denotes the spectral radius of $B$.*

For many useful properties of M-matrices, the reader is referred to [10, Chapter 6]. The following Lemma will be used in the proof of Lemma 2.3.17.

**Lemma 2.3.11 :** *If $(\alpha, Q)$ is an irreducible discrete (resp. continuous) PH-representation of order $m$, then the matrix $(I_m - Q)$ (resp. $-Q$) is a nonsingular M-matrix and its eigenvalues all have positive real parts.*

**Proof :** It is well known, [10], that for any $n \times n$ nonnegative matrix $A$, the following bounds

$$\min_i\{\sum_{j=1}^{n} A_{ij}\} \leq \rho(A) \leq \max_i\{\sum_{j=1}^{n} A_{ij}\}$$

hold true for its spectral radius. When the matrix $A$ is stochastic, each one of its rows sum to 1, whence $\rho(A) = 1$.

Since the nonnegative matrix $Q$ is substochastic, $\rho(Q) \leq 1$, and the matrix $I_m - Q$ is an M-matrix by Definition 2.3.10 (with $s = 1$ and $B = Q$). Furthermore, the matrix $(I_m - Q)$ being invertible implies $\rho(Q) < 1$ and the second part of Lemma 2.3.11 is now immediate: Indeed, $(\lambda, x)$ is a rigth (left) eigenpair for $Q$ if and only if $(1 - \lambda, x)$ is a right (left) eigenpair for $I_m - Q$, whence $\Re e(1 - \lambda) > 0$ since $-1 < \Re e(\lambda) < 1$, where $\Re e(x)$ denotes the real part of the complex number $x$.

The Lemma can be proved in a similar way for an irreducible continuous PH-representation with $s = -q^*$ and $B = (-q^* I_m + Q)$, where $-q^* := \max_{1 \leq i \leq m}\{-Q_{ii}\}$.

△

The following theorem is a basic result in the Perron-Frobenius theory of nonnegative matrices.

**Theorem 3.2.12.** : *If $A$ is an $n \times n$ nonnegative matrix, then*

*(i) The spectral radius $\rho(A)$ of $A$ is an eigenvalue of $A$, and*

*(ii) There always exists left and right eigenvectors with nonnegative components which corresponds to $\rho(A)$.*

The *invariant* probability vector $\pi$ of an m-state Markov chain with one-step probability transition matrix $A$ is defined as the $1 \times m$ vector $\pi$ that satisfies

$$\pi A = \pi \quad , \quad \pi e_m = 1 \quad .$$

The following results investigate the existence and uniqueness of the invariant probability vector of a finite state Markov chain.

**Theorem 2.3.13.** : *Every finite state Markov chain has an invariant probability vector.*

**Proof :** If the $m \times m$ matrix $A$ is the state transition matrix associated with the chain, then $\rho(A) = 1$ by Lemma 2.3.11 and by Theorem 2.3.12 there exists a row vector $x > 0$ with $xA = x$. Normalizing $x$ gives $\pi = (xe_m)^{-1}x$ with $\pi A = \pi$ and $\pi e_m = 1$, i.e., $\pi$ is an invariant probability vector of the chain.

△

Although Theorem 2.3.13 guarantees the existence of an invariant probability vector, it is *not* unique in general. The next result investigates the general form of an invariant probability vector.

**Theorem 2.3.14.** : *Let $S_i$, $1 \leq i \leq r$, be the ergodic classes of a finite state Markov chain. For each $S_i$ there is a unique invariant probability vector $\pi(i)$ with the property that the entries of $\pi(i)$ corresponding to the states of $S_i$ are positive whereas all other entries are zero. Moreover, any invariant probability vector $\pi$ of the chain can be expressed as a linear convex combination of the vectors $\pi(i)$, $1 \leq i \leq r$, i. e.,*

$$\pi = \sum_{i=1}^{r} \lambda_i \pi(i) \ , \quad \lambda_i \geq 0 \ , \quad \sum_{i=1}^{r} \lambda_i = 1 \ .$$

**Proof :** See Berman and Plemnons [10, pp. 224-225].

In view of Theorem 2.3.14, even when the Markov chain is not irreducible, if it has a single ergodic class, then the invariant probability vector is unique with positive entries for positions corresponding to the ergodic class and zero entries for positions corresponding to the transient states. Note that the ergodic classes are defined as equivalence classes induced by the communication relation as defined above, with no assumptions made on the *(a)periodicity* of the Markov chain. Therefore, in cases where there is a single ergodic class, the unique invariant probability vector will coincide with the *long-run* average probability vector of the Markov chain ( defined in the *Cesaro* sense).

Next, conditions for the nonsingularity of several matrices of interest for subsequent developments are established. To that end, the *Kronecker product* of matrices is defined and some of its properties are introduced.

**Definition 2.3.15 :** *If $B$ and $A$ are $p \times q$ and $m \times n$, rectangular matrices, respectively, then their Kronecker product, denoted $B \otimes A$, is the $pm \times qn$ matrix defined in block-partitioned form by*

$$B \otimes A := \begin{bmatrix} B_{11}A & B_{12}A & \cdots & B_{1q}A \\ B_{21}A & B_{22}A & \cdots & B_{2q}A \\ \cdot & & & \\ \cdot & & & \\ B_{p1}A & B_{p2}A & \cdots & B_{pq}A \end{bmatrix} . \tag{2.3.1}$$

Some properties of the Kronecker product are collected in the following Lemma for future use. The proof of these properties and of other useful properties of the Kronecker product can be found in [14].

**Lemma 2.3.16. :** *(i) : If A,B,C,D are rectangular matrices such that the ordinary matrix products below are defined, then*

$$(A \otimes B) \otimes C = A \otimes (B \otimes C) \ , \tag{2.3.2a}$$

$$(A \otimes B)\,(C \otimes D) = AC \otimes BD \ , \tag{2.3.2b}$$

$$(A \otimes B)^T = A^T \otimes B^T \ . \tag{2.3.2c}$$

*(ii) : If the row vectors $\alpha$ and $\beta$ are the left eigenvectors of the $m \times m$ and $n \times n$ matrices $A$ and $B$ corresponding to eigenvalues $\lambda$ and $\mu$, respectively, then $\alpha \otimes \beta$ is an eigenvector of $A \otimes B$ with eigenvalue $\lambda\,\mu$ and all eigenpairs of $A \otimes B$ have this*

*form, whence the relation*

$$det(A \otimes B) = det(A)^n \; det(B)^m \tag{2.3.2d}$$

*holds, where det(X) is the determinant of the matrix X.*

Let the pairs $(\alpha, A)$ and $(\beta, B)$ be two irreducible discrete-time PH-representations. The row vectors $\alpha$ and $\beta$ and the matrices $A$ and $B$ have dimensions $1 \times l$, $1 \times m$, $l \times l$ and $m \times m$, respectively, and the corresponding $l \times 1$ and $m \times 1$ column vectors of absorption probabilities are denoted by $a$ and $b$, respectively. The $(l\; m \times l\; m)$ matrices $M$ and $N$ are now defined by

$$M := I_m \otimes (I_l - e_l\, \alpha) + B \otimes (e_l\, \alpha - A) \tag{2.3.3a}$$

$$N := (I_m - e_m\, \beta) \otimes I_l + (e_m\, \beta - B) \otimes A, \tag{2.3.3b}$$

respectively. Conditions for the nonsingularity of the matrices $M$ and $N$ are now given. To that end, let $C$ be the open disc centered at (0.5,0) in the complex plane with radius 0.5 and let $Sp(X)$ denote the spectrum of the matrix $X$. The following result gives a sufficient but not necessary condition.

**Lemma 2.3.17. :** *If $Sp(B) \subseteq C$ (resp. $Sp(A) \subseteq C$) then the matrix M (resp. N) defined by (2.3.3) is nonsingular.*

**Proof :** Let $\Re e(x)$ again denote the real part of the complex number $x$. The lemma is proved only for $M$ as nonsingularity of $N$ follows along the same lines. Since the matrices $(I_l - A)$ and $(I_m - B)$ are nonsingular, Lemma 2.3.16 allows a rewriting of the matrix $M$ in the form

$$M = I_m \otimes (I_l - A) + (I_m - B) \otimes (A - e_l\, \alpha)$$

$$= [I_m \otimes (I_l - A)] \;\; D \;\; [(I_m - B) \otimes I_l]$$

where

$$D = (I_m - B)^{-1} \otimes I_l + I_m \otimes (I_l - A)^{-1} \; (A - e_l\; \alpha) \; .$$

This clearly shows that $M$ is nonsingular if and only if the matrix $D$ is nonsingular. By Lemma 2.3.16 again, each eigenvalue of the matrix $D$ is the sum of the eigenvalues of the matrices $(I_m - B)^{-1}$ and $(I_l - A)^{-1}\ (A - e_l\ \alpha)$. By assumption, it is an easy exercise to check that the eigenvalues of $(I_m - B)^{-1}$ have real parts strictly greater than one, and the matrix $D$ is thus invertible if the real parts of the eigenvalues of the matrix $(I_l - A)^{-1}\ (A - e_l\ \alpha)$ are greater than or equal to -1.

The proof will be completed by showing that if $\gamma \in Sp\left[(I_l - A)^{-1}\ (A - e_l\,\alpha)\right]$, then $\Re e(\gamma) \geq -1$. The argument proceeds by contradiction and assumes $\Re e(\gamma) < -1$. If $y$ denotes the corresponding right eigenvector, then

$$(I_l - A)^{-1}\,A\,y - (\alpha\,y)\,(I_l - A)^{-1}\,e_l = \gamma\,y \quad , \tag{2.3.4}$$

and therefore $\alpha\,y \neq 0$, for otherwise the second term on the left hand side of (2.3.4) drops and adding $y$ to both sides then yields

$$\left(I_l + (I_l - A)^{-1}\,A\right)\ y = (I_l - A)^{-1}\,y = \gamma\,y + y$$

and therefore $(\gamma + 1) \in Sp\left[(I_l - A)^{-1}\right]$ and $\Re e(\gamma + 1) \geq 0$ by Lemma 2.3.11, thus contradicting the assumption that $\Re e(\gamma) < -1$.

Since $\alpha\,y \neq 0$, assume without loss of generality that $\alpha\,y = 1$ and note from (2.3.4) that $-\gamma\left[I_l - (\frac{1+\gamma}{\gamma})\,A\right]\ y = e_l$. The assumption $\Re e(\gamma) < -1$ implies $\left|\frac{1+\gamma}{\gamma}\right| < 1$, whence $\left(I_l - (\frac{1+\gamma}{\gamma})\,A\right)$ is invertible. The relation

$$y = -\left[\gamma\,I_l - (1+\gamma)\,A\right]^{-1}\,e_l$$

thus obtains, and therefore upon premultiplying by $\alpha$,

$$1 = -\frac{1}{1+\gamma}\ \alpha\ \left[\frac{\gamma}{1+\gamma}\ I_l - A\right]^{-1}\ e_l\ . \tag{2.3.5}$$

Equations (2.3.5) and (2.2.3) combine to yield

$$P\left(\frac{1+\gamma}{\gamma}\right) = 1 - (\frac{-1}{1+\gamma})\,\alpha\,\left[\frac{\gamma}{1+\gamma}\,I_l - A\right]^{-1}\,e_l = 0$$

where $P(\cdot)$ is the probability generating function of the PH-representation $(\alpha, A)$. The condition $\Re e(\gamma) < -1$ means $0 < \Re e\left(\frac{1+\gamma}{\gamma}\right) \leq 1$, and the probability generating function of a non-negative random variable being *strictly* positive in the right half-plane, it thus follows by contradiction that $\Re e(\gamma) \geq -1$.

$\triangle$

It should be noted that for many interesting discrete PH-type distributions such as Erlang or hypergeometric distributions the eigenvalues of the associated $Q$ matrix as given by (2.2.5) and (2.2.6), respectively, satisfy the conditions of Lemma 2.3.17.

The following corollary is an immediate consequence of this last proof.

**Corollary 2.3.18. :** *The matrix $M$ (resp. $N$) is singular if the matrix $B$ (resp. $A$) is singular.*

**Proof :** The matrix $B$ being singular, $\lambda = 1$ is an eigenvalue of $(I_m - B)^{-1}$ that corresponds to eigenvalue $\lambda = 0$ for $B$, while $(-1, e_l)$ is a right eigenpair for the matrix $(I_l - A)^{-1}\,(A - e_l\,\alpha)$. Therefore, zero is an eigenvalue of the matrix $D$ defined above, whence the matrix $D$ is singular and so is the matrix $M$.

$\triangle$

Similarly, let the pairs $(\alpha, A)$ and $(\beta, B)$ be two irreducible continuous-time PH-representations. The row vectors $\alpha$ and $\beta$ and the matrices $A$ and $B$ have dimensions $1 \times l$, $1 \times m$, $l \times l$ and $m \times m$, respectively, and let the corresponding $l \times 1$ and $m \times 1$ column vectors of absorption probabilities be again denoted by $a$ and $b$, respectively. The $(l\ m \times l\ m)$ matrices $\widetilde{M}$ and $\widetilde{N}$ are now defined by

$$\widetilde{M} := B \otimes (I_l - e_l\,\alpha) + I_m \otimes A \qquad (2.3.6a)$$

$$\widetilde{N} := (I_m - e_m\,\beta) \otimes A + B \otimes I_l \ , \tag{2.3.6b}$$

respectively.

**Lemma 2.3.19. :** *The matrices $\widetilde{M}$ and $\widetilde{N}$ defined by (2.3.6) are nonsingular.*

**Proof :** The Lemma is proved only for $\widetilde{N}$ as the nonsingularity of $\widetilde{M}$ follows along the same lines. With a similar line of thoughts as in the proof of Lemma 2.3.17, the matrix $\widetilde{N}$ can be rewitten as

$$\widetilde{N} = (B \otimes I_l)\ \left[B^{-1}\ (I_m - e_m\,\beta) \otimes I_l + I_m \otimes A^{-1}\right]\ (I_m \otimes A)$$

Since the matrices $A$ and $B$ are invertible, it follows from the second part of Lemma 2.3.6 that the matrices $(B \otimes I_l)$ and $(I_m \otimes A)$ are both invertible, whence $\widetilde{N}$ is invertible if and only if the matrix $\left[B^{-1}\ (I_m - e_m\,\beta) \otimes I_l + I_m \otimes A^{-1}\right]$ is invertible. Note from Lemma 2.3.11 that $A$ has eigenvalues with strictly negative real parts, therefore to show the nonsingularity of $\widetilde{N}$, it suffices to show that if $\gamma \in Sp[B^{-1}\,(I_m - e_m\,\beta)]$, then $\Re e(\gamma) \leq 0$. That this fact holds true follows by arguments similar to the ones given in the proof of Lemma 2.3.17, and the matrix $\widetilde{N}$ is thus invertible.

△

Finally, the well-known *Gerschgorin Circle Theorem* is stated here for easy reference in later chapters.

**Theorem 2.3.20. :** *Let $X$ be an $n \times n$ complex matrix. Define*

$$r_i(X) := \sum_{\substack{j=1 \\ j \neq i}}^{n} |X_{ij}| \qquad\qquad 1 \leq i \leq n \ ,$$

*and let $Z_i$, $1 \leq i \leq n$, denote the closed disc in the complex plane $C$, with center $X_{ii}$ and radius $r_i(X)$,i. e.,*

$$Z_i = \{z \in C\ : |z - x_{ii}| \leq r_i(X)\}\,.$$

*Any eigenvalue $\lambda$ of $X$ belongs to at least one of the discs $Z_i$, $1 \leq i \leq n$. Moreover if $m$ of the discs form a connected set $S$, disjoint from the $n-m$ remaining ones, then $S$ contains exactly $m$ of the eigenvalues of $X$, counted according to their multiplicity as roots of the characteristic polynomial of $X$. The numbers $r_i(X)$, $1 \leq i \leq n$, are called the* **Gershgorin radii** *of the matrix $X$.*

**Proof :** See Atkinson [7, p. 500].

△

# CHAPTER III

# TWO NODE SYSTEM WITH PHASE TYPE SERVERS

## III.1. Introduction :

This chapter is devoted to the analysis of a two node system with PH-type servers in *both* nodes. Under the discrete-time formulation discussed in Sections 2 and 3, neccesary and sufficient conditions for the irreducibility of the underlying Markov chain are given and its ergodic and transient states are identified. Explicit analytical expressions for the invariant probabilities of the system are shown to have a *matrix geometric* form and *closed form* expressions are obtained.

In Section 4, the effect of reversing the order of the servers is studied. The results for the discrete-time formulation are illustrated in Section 5 through numerical examples, for some specific PH-type service distributions. In Section 6, a continuous-time formulation of the same model is discussed; the underlying Markov chain is shown to *always* be *irreducible*, and closed form expressions for the invariant probabilities are obtained in matrix geometric form.

## III.2. The discrete-time model :

The model to be discussed consists of two nodes separated by a *finite* intermediate buffer of capacity $K$, i. e., there are exactly $K$ positions in the buffer, inclusive of the one taken by the job in service at the second node server. The service times at both nodes have *PH-type* distributions with irreducible representations $(\alpha, A)$ and $(\beta, B)$, respectively. The row vectors $\alpha$ and $\beta$ and the matrices $A$ and $B$ have dimensions $1 \times l$, $1 \times m$, $l \times l$ and $m \times m$, respectively, and the corresponding $l \times 1$ and $m \times 1$ column vectors of absorption probabilities for the first and the second node server are denoted by $a$ and $b$, respectively. There is an *infinite* supply of jobs available in front of the first node server and the second node server *never* gets blocked. *Immediate* blocking is assumed in that blocking of the first node server occurs as soon as the intermediate buffer becomes full. The first node server resumes service as soon as there is a space available in the buffer, i. e., whenever a service completion takes place at the second node server.

A natural state space $E$ for this system is the one that contains $r := (K-1)lm + l + m$ states with

$$E = \begin{cases} (i,0)\ , & 1 \leq i \leq l\ ,\ \ k = 0, \\ (i,k,j)\ , & 1 \leq i \leq l, 0 < k < K \text{ and } 1 \leq j \leq m, \\ (K,j)\ , & 1 \leq j \leq m\ ,\ \ k = K, \end{cases}$$

where $k$ indicates the buffer size, while $i$ and $j$ represent the service phase in the first and the second node server, respectively. The phase of the second node server is not defined when it has no jobs to process and the phase of the first server is not defined when the buffer is full since blocked.

The invariant probability vector of these states is denoted by the $1 \times r$ row vector $\pi$. This vector is partioned into $K+1$ blocks of components, say $\pi = (\pi_0, \pi_1, \ldots, \pi_K)$, with $\pi_0$, $\pi_k$, $0 < k < K$, and $\pi_K$ being row vectors of dimension $1 \times l$, $1 \times l\,m$ and $1 \times m$, respectively.

By lexicographically ordering the states, i.e., $(1,0), \ldots, (l,0), (1,1,1), (2,1,1), \ldots, (l,1,1), (1,1,2), \ldots$, the one-step state transition matrix $T$ of the underlying Markov chain can be obtained in block tridiagonal form. If $T_{k_1,k_2}$ denotes the $(k_1,k_2)$-th block of the matrix $T$, $0 \leq k_1, k_2 \leq K$, then the reader will easily check that

$$\begin{aligned} T_{k,k-1} &= b\,\beta \otimes A && 1 < k < K, \\ T_{k,k} &= B \otimes A + b\,\beta \otimes a\,\alpha && 1 \leq k < K, \\ T_{k,k+1} &= B \otimes a\,\alpha && 1 \leq k < K-1, \end{aligned}$$

where each block is an $l\,m \times l\,m$ matrix. The following interpretation can be given for the above expressions: For the equation $T_{k,k-1} = b\beta \otimes A$, the buffer size will decrease by one when a service completion take place in the second node server, according to the absorption vector $b$, and the new phase of service is initialized according to the initialization vector $\alpha$ (if there is a job availible in the buffer), while no service completion occur in the first node server and the phase of service change according to the transition matrix $A$. A parallel reasonning can also be given for the term $T_{k,k+1}$. On the other hand, at the end of the time epoch the buffer size may remain unchanged for two reasons: i) A service completion may occur at *both* servers, $b\beta \otimes a\alpha$, or ii) no service completion occur in *both* servers, $B \otimes A$, whence the interpretation for the term $T_{k,k}$. For the *boundary* states, the entries are given in block form by

$$\begin{aligned} T_{00} &= A && l \times l \text{ matrix}, \\ T_{10} &= b \otimes A && m\,l \times l \text{ matrix}, \\ T_{01} &= \beta \otimes a\,\alpha && l \times m\,l \text{ matrix}, \end{aligned}$$

$$
\begin{aligned}
T_{K-1,K} &= B \otimes a && m\,l \times m \text{ matrix,}\\
T_{K,K-1} &= b\,\beta \otimes \alpha && m \times m\,l \text{ matrix,}\\
T_{K,K} &= B && m \times m \text{ matrix.}
\end{aligned}
$$

The resulting transition matrix then takes the form

$$
T = \left( \begin{array}{cccccc}
A & \beta \otimes a\,\alpha & & & & \\
b \otimes A & \begin{array}{c} B \otimes A + \\ b\,\beta \otimes a\,\alpha \end{array} & B \otimes a\,\alpha & & & \\
 & b\,\beta \otimes A & \begin{array}{c} B \otimes A + \\ b\,\beta \otimes a\,\alpha \end{array} & B \otimes a\,\alpha & & \\
 & & \ddots & \ddots & \ddots & \\
 & & & b\,\beta \otimes A & \begin{array}{c} B \otimes A + \\ b\,\beta \otimes a\,\alpha \end{array} & B \otimes a \\
 & & & & b\,\beta \otimes \alpha & B
\end{array} \right) . \qquad (3.2.1)
$$

## III.3. Analysis of the discrete-time model :

In this section, necessary and sufficient conditions are obtained for the *irreducibility* of the Markov chain with one-step transition matrix $T$ given by (3.2.1) and explicit solutions for the invariant probability vector are obtained. For notational convenience, when the intermediate buffer capacity is $K$, the one-step transition matrix (3.2.1) is denoted by $T_K$.

The case $K = 1$ is discussed first since the results for the general discussion do not cover this case. The results are summarized in the following theorem.

**Theorem 3.3.1. :** *For the model described above with $K = 1$, i. e., there is no intermediate buffer, the underlying Markov chain is* **always irreducible** *and the invariant probability vector is given explicitly by*

$$\pi_0 = c\alpha\,(I_l - A)^{-1} \ , \qquad \pi_1 = c\,\beta\,(I_m - B)^{-1} \ , \tag{3.3.1a}$$

where

$$c = \left(\alpha\,(I_l - A)^{-1}e_l + \beta\,(I_m - B)^{-1}e_m\right)^{-1} \ . \tag{3.3.1b}$$

**Proof :** Let $E_0$ and $E_1$ be subsets of the state set $E$ given by

$$E_0 = \{x \in E : x = (i,0),\ 1 \le i \le l\}$$

$$E_1 = \{x \in E : x = (1,j),\ 1 \le j \le m\}$$

Since no intermediate buffer space is available other than the one taken by the job in service at the second node and immediate blocking is assumed, the servers cannot be active simultaneously and the decomposition $E = E_0 \cup E_1$ thus holds. Note that the one-step transition matrix $T_1$ takes the form

$$T_1 = \begin{bmatrix} A & a\beta \\ b\,\alpha & B \end{bmatrix} \ .$$

Since the PH-representation $(\alpha, A)$ is irreducible, by viewing the transitions from the states in the set $E_0$ to the states in the set $E_1$ as visits to the *instantaneous* phase $l+1$ of the first server, it readily follows that the states in $E_0$ belong to the same irreducible class of $E$. By a similar argument, the states in $E_1$ are also in an irreducible class. The irreducibility of the Markov chain thus follows since these two irreducible classes of $E$ communicate.

The equation $\pi T_1 = \pi$ can be rewritten as

$$\pi_0 A + \pi_1 (b\, \alpha) = \pi_0 \tag{3.3.2a}$$

$$\pi_0 (a\, \beta) + \pi_1 B = \pi_1 \ . \tag{3.3.2b}$$

Solving $\pi_0$ from (3.3.2a) and using (3.3.2b) with the relation $Be_m + b = e_m$ gives

$$\pi_0 = \pi_0\, a\, \alpha\, (I_l - A)^{-1} \ . \tag{3.3.3}$$

or eqivalently,

$$\pi_0 = c\, \alpha\, (I_l - A)^{-1} \ , \tag{3.3.4}$$

with the identifation $c = \pi_0\, a$. Substitution of (3.3.4) into (3.3.2b) readily leads to

$$\pi_1 = \pi_0 (a\, \beta)\, (I_m - B)^{-1} = c\, \beta\, (I_m - B)^{-1} \ . \tag{3.3.5}$$

Use of the normalization condition $\pi_0\, e_l + \pi_1\, e_m = 1$ gives the desired result.

△

Next, the transitions among the states of the matrix $T_2$ are discussed in detail since they will be used to characterize the irreducibility of the matrices $T_K$, $K > 2$.

Let $E$ denote the set of states for the Markov chain with one-step transition matrix

$$T_2 = \begin{bmatrix} A & \beta \otimes a\alpha & 0_{l\times m} \\ b \otimes A & B \otimes A + b\beta \otimes a\alpha & B \otimes a \\ 0_{m\times l} & b\beta \otimes \alpha & B \end{bmatrix} .$$

If the sets $E_k,\ 0 \le k \le 2$, are defined by

$$E_0 := \{e \in E : e = (i,0),\ 1 \le i \le l\},$$
$$E_1 := \{e \in E : e = (i,1,j),\ 1 \le i \le l,\ 1 \le j \le m\},$$
$$E_2 := \{e \in E : e = (2,j),\ 1 \le j \le m\}\ ,$$

then $E_0, E_1$ and $E_2$ form a partition of the set $E$ and the transitions among the states *within* these sets are induced by the directed graphs $G(A), G(B \otimes A + b\beta \otimes a\alpha)$ and $G(B)$, respectively. In general, these directed graphs may induce several irreducible classes, say $I_k^r,\ 1 \le r \le R_k,\ 0 \le k \le 2$, and a set of transient states, say $T_k,\ 0 \le k \le 2$, respectively. In order to more completely understand the transition mechanism of the underlying Markov chain among the states in $E$, and characterize the ergodic and transient states, the following sets are introduced:

$$\begin{aligned}
&E_1^* := \{(i,1,j) \in E_1\ : \alpha_i \beta_j > 0\},\\
&T_1^* := T_1 \cap E_1^*\ ,\\
&E_1^r := I_1^r \cup \{e \in T_1^*\ : I_1^r \text{ is reachable from } e\} \qquad\qquad 1 \le r \le R_1,\\
&\qquad \cup\ \{e \in T_1 : e \text{ is reachable from } T_1^* \text{ and } I_1^r \text{ is reachable from } e\}\ ,\\
&\Omega_1 := \{r,\ 1 \le r \le R_1 :\ E_1^r \cap E_1^* \ne \emptyset\}\\
&E_k^r := I_k^r \cup \{(k,i) \in T_k\ :\ (k,i) \text{ is reachable from } \cup_{s\in\Omega_1} E_1^s\\
&\qquad\qquad \text{and } I_k^r \text{ is reachable from } (k,i)\} \qquad k = 0,2,\quad 1 \le r \le R_k,\\
&\Omega_k := \{r,\ 1 \le r \le R_k : E_k^r \text{ is reachable from the set } \cup_{s\in\Omega_1} E_1^s\},\quad k = 0,2,\\
&\Gamma_k := \{1,2,\ldots,R_k\} \setminus \Omega_k\ ,\quad 0 \le k \le 2,
\end{aligned} \tag{3.3.6}$$

where $1 \le i \le l$ and $1 \le j \le m$. Note that each set $E_k^r,\ 0 \le k \le 2,\ 1 \le r \le R_k$, has two components, one in the $r^{th}$ irreducible class $I_k^r$ of $E_k$ and the one formed by states which are transient for the directed graph governing the law of motion within the set $E_k$, but for which there exists a path that connects the irreducible classes in the sets $E_{k'}$, $0 \leq k' \neq k \leq 2$, to the irreducible class $I_k^r$. The index sets $\Omega_k$, $0 \leq k \leq 2$, identify the sets $E_k^r$, $0 \leq k \leq 2$, $1 \leq r \leq R_k$, that communicate with each other through a service completion process. The transitions from any state in $E_0 \cup E_2$ to any state in $E_1$ is due to a service completion either in the first or the second server. In either case, the phases in *both* PH-distributions are ***reinitialized.*** Therefore, the Markov chain can enter the set $E_1$ only from the states in the set $E_1^*$. Furthermore, *every* state in the set $E_1^*$ is reachable from *every* state in the set $E_0 \cup E_2$. Therefore, in view of these definitions the following property (P) holds, where

> (P): The sets $E_1^r$, $r \in \Omega_1$, are reachable from *every* set $E_0^r$ and $E_2^{r'}$, $1 \leq r \leq R_0$ and $1 \leq r' \leq R_2$, and the sets $E_1^r$, $r \in \Gamma_1$ and the set $T_1 \setminus E_1^*$ are *not* reachable from *any* state in $E_0 \cup E_2$.

If $\Omega_0 \cup \Omega_2 = \emptyset$ then, once the process reaches a state in the set $\cup_{r\in\Omega_1}$, the service completions in the PH-type servers occur at the same time so that there will always be one job in the buffer. Since all the sets $E_1^r$, $r \in \Omega_1$, are reachable through a service completion in both servers, the states in the sets $E_1^r$, $r \in \Omega_1$ communicate, whence the set $\bigcup_{r\in\Omega_1} E_1^r$ forms an irreducible class for the matrix $T_2$ and all the other states will be transient. On the other hand, if $\Omega_0 \cup \Omega_2 \neq \emptyset$, then from the above argument *none* of the sets $E_1^r$, $r \in \Omega_1$ are ***absorbing***, and owing to property (P), the set $T_2^E := \bigcup_{k=0}^2 \bigcup_{r\in\Omega_k} E_k^r$ form an irreducible class for $T_2$. The sets $E_1^r$, $r \in \Gamma_1$, can not be absorbing sets since this would mean that the service completions occur at the same time in both servers, and this would contradict the fact that they are not reachable from $E_0 \cup E_2$. Therefore, the states which are not in the set $T_2^E$ are the *transient* states of $T_2$.

The above discussion can be summarized in the following theorem.

**Theorem 3.3.2.** : *The Markov chain with one-step probability transition matrix $T_2$ is either irreducible or it has a single irreducible class and the remaining states are transient. Furthermore, the irreducible class of the chain, denoted by $T_2^E$, is given by*

$$T_2^E = \bigcup_{k=0}^{2} \bigcup_{r \in \Omega_k} E_k^r \ . \tag{3.3.7}$$

The following corollary follows immediately from Theorems 3.3.2 and 2.3.13.

**Corollary 3.3.3.** : *For the model described above, when $K = 2$, the invariant probability distribution vector is always* unique.

The analytical expression for the invariant distribution when $K = 2$ is included in the statement of Theorem 3.3.7.

The directed graph, $G(T_2)$, of the matrix $T_2$ can be used to obtain information about the classification of the states of the Markov chain induced by the matrix $T_K$, $K > 2$. In order to avoid confusion in the discussion, the state space for the model with intermediate buffer capacity $K > 2$ is denoted by $\widetilde{E}$. As in the case $K = 2$ the state space $\widetilde{E}$ is partitioned into sets $\widetilde{E}_k$, $1 \leq k \leq K$, where $\widetilde{E}_k$ are defined by

$$\begin{aligned} \widetilde{E}_0 &:= \{e \in \widetilde{E} : \ e = (i,0), \ 1 \leq i \leq l\} \ , \\ \widetilde{E}_k &:= \{e \in \widetilde{E} : \ e = (i,k,j), \ 1 \leq i \leq l, \ 1 \leq j \leq m\} \ , \\ \widetilde{E}_K &:= \{e \in \widetilde{E} : \ e = (K,j), \ 1 \leq j \leq m\} \ , \end{aligned} \tag{3.3.8a}$$

and the subsets $\widetilde{E}_k^r$, $1 \leq r \leq R_k$ of $\widetilde{E}_k$ are defined as

$$\begin{aligned}
&\widetilde{E}_0^r := E_0^r \ , \\
&\widetilde{E}_k^r := \{(i,k,j) : \ (i,1,j) \in E_1^r, \ \ 1 \leq i \leq m, \ 1 \leq j \leq l\} \ , \ 0 < k < K \qquad (3.8.8b) \\
&\widetilde{E}_K^r := \{(K,j) : \ (2,j) \in E_2^r, \ \ 1 \leq j \leq l\} \ .
\end{aligned}$$

Comparison of the matrices $T_K$ and $T_2$ quickly reveals that the equailities

$$\begin{aligned}
&G(E_0) \equiv G(\widetilde{E}_0) \\
&G(E_1) \equiv G(\widetilde{E}_k) \ , \ \ 0 < k < K \qquad (3.3.9) \\
&G(E_2) \equiv G(\widetilde{E}_K)
\end{aligned}$$

holds, where the notation $\equiv$ indicates that the associated directed graphs have same topological structure.

By using the notation developed for $T_2$ the following cases are studied:

(i) $\Omega_0 = \emptyset$ and $\Omega_2 = \emptyset$; As discussed for $K = 2$, this means that when both servers are operational the service completions occur at the same time. From the discussion given for $K = 2$ and relations (3.3.9) each set $\cup_{r \in \widetilde{\Omega}_1} \widetilde{E}_k^r$, $0 < k < K$, $1 \leq r \leq R_k$, form an ergodic class for $T_K$. Note that these are not the only ergodic classes of $T_K$ and there may be other ergodic classes formed by the communication of the transient states of $\widetilde{E}_k$ that are not in any of the sets $\widetilde{E}_k^r$, $1 \leq k < K$, $1 \leq r \leq R_k$.

(ii) $\Omega_0 \neq \emptyset$ and $\Omega_2 = \emptyset$; In this case, again owing to (3.3.6) and (3.3.9) the sets $\widetilde{E}_k$ are *not* reachable from the sets $\widetilde{E}_{k-1}$, while the sets $\widetilde{E}_{k-1}$ are reachable from the sets $\widetilde{E}_k$, for $1 < k \leq K$. Therefore all the states in the set $\left[\bigcup_{k=2}^{K} \widetilde{E}_k\right] \bigcup \left[\bigcup_{k=0}^{1} \bigcup_{r \in \Gamma_k} \widetilde{E}_k^r\right]$ are transient states for $T_K$ and the states in the set $\bigcup_{k=0}^{1} \bigcup_{r \in \Omega_k} \widetilde{E}_k^r$ form a single irreducible class, where $\Gamma_k$ and $\Omega_k$, $0 \leq k \leq 2$ are as defined in (3.3.6).

(iii) $\Omega_0 = \emptyset$ and $\Omega_2 \neq \emptyset$; A similar argument as in case (ii) shows that the states in the set $\left[\bigcup_{k=0}^{K-2} \widetilde{E}_k\right] \bigcup \left[\left(\cup_{r\in\widetilde{\Gamma}_1} \widetilde{E}^r_{K-1}\right) \bigcup \left(\cup_{r\in\widetilde{\Gamma}_2} \widetilde{E}^r_K\right)\right]$ are transient states, while the states in the set $\left(\cup_{r\in\Omega_1} \widetilde{E}^r_{K-1}\right) \bigcup \left(\cup_{r\in\Omega_2} \widetilde{E}^r_K\right)$ form a single ergodic class.

(iv) $\Omega_0 \neq \emptyset$, $\Omega_2 \neq \emptyset$ and $\widetilde{E}_0$ and $\widetilde{E}_K$ are reachable from *every* state in the set $\bigcup_{k=1}^{K-1} \widetilde{E}_k$; A sufficient but not necessary condition for this to happen is that $E_0$ and $E_2$ are reachable from *every* $E^r_1$ in $G(T_2)$. Let the set $X$ be defined as

$$X := \{e \in \widetilde{E} : \ e \text{ is reachable from the set } \widetilde{E}_0 \cup \widetilde{E}_K\} \ .$$

Note that $\widetilde{E}_0 \cup \widetilde{E}_K \subseteq X$. In this case, every state in $X$ communicates with each other, possibly through the states in the set $\widetilde{E}_0 \cup \widetilde{E}_K$. Therefore, the set $X$ form an ergodic class for $T_K$ and the states in the set $\widetilde{E} \setminus X$ are the transient states as they have access to the set $\widetilde{E}_0 \cup \widetilde{E}_K$ but not reachable from $\widetilde{E}_0 \cup \widetilde{E}_K$.

The following results are immediate consequences of this discussion.

**Lemma 3.3.4.** : *If cases (ii)-(iv) above holds, then the matrices $T_K$, $K > 2$, have a single ergodic class.*

**Corollary 3.3.5** : *If cases (ii)-(iv) above holds, then the invariant probability distribution vector $\pi$ for the Markov chain with one-step transition matrix $T_K$, $K > 2$, is* unique.

**Lemma 3.3.6.** : *If case (i) above holds, then the matrices $T_K$, $K > 2$, have several ergodic classes.*

In order to obtain a closed form expression for the vector $\pi$, proceed as follows: The equation

$$\pi\, T_K = \pi \quad , \quad \pi e_r = 1$$

satisfied by the invariant vector can be rewritten as

$$\pi_0\, A + \pi_1\, (b \otimes A) = \pi_0 \tag{3.3.10$a$}$$

$$\pi_0\, (\beta \otimes a\, \alpha) + \pi_1\, (B \otimes A + b\,\beta \otimes a\, \alpha) + \pi_2\, (b\,\beta \otimes A) = \pi_1 \tag{3.3.10$b$}$$

$$\pi_{k-1}\, (B \otimes a\, \alpha) + \pi_k\, (B \otimes A + b\,\beta \otimes a\, \alpha) + \pi_{k+1}\, (b\,\beta \otimes A) = \pi_k \tag{3.3.10$c$}$$

$$1 < k < K-1$$

$$\pi_{K-2}\, (B \otimes a\, \alpha) + \pi_{K-1}\, (B \otimes A + b\,\beta \otimes a\, \alpha) + \pi_K\, (b\,\beta \otimes \alpha) = \pi_{K-1} \tag{3.3.10$d$}$$

$$\pi_{K-1}\, (B \otimes a) + \pi_K\, B = \pi_K \quad . \tag{3.3.10$e$}$$

Postmultiplication of (3.3.10a) by $e_l$, of (3.3.10b)-(3.3.10d) by $(e_m \otimes e_l)$ and of (3.3.10e) by $e_m$ leads, after simplifications, to the relations

$$\pi_k\, [b \otimes (e_l - a)] - \pi_{k-1}\, [(e_m - b) \otimes a] = 0\ , \qquad 1 < k < K \quad , \tag{3.3.11}$$

whence upon postmultiplication by $(\beta \otimes \alpha)$,

$$\pi_k\, [b\,\beta \otimes (e_l - a)\, \alpha] = \pi_{k-1}\, [\,(e_m - b)\,\beta \otimes a\, \alpha] \ , \quad 1 < k < K \quad . \tag{3.3.12}$$

Note that the left hand side of equation (3.3.11) is the transition probability from the set $\widetilde{E}_k$ to the set $\widetilde{E}_{k-1}$ in one step while the right hand side of (3.3.11) is the transition probability from the set $\widetilde{E}_{k-1}$ to the set $\widetilde{E}_k$. Now, postmultiplication of (3.3.10b)-(3.3.10d) with $(e_m\,\beta \otimes I_l)$ and use of (3.3.10b)-(3.3.10d)readily yield

$$\pi_1\, [B\,(e_m\,\beta - I_m) \otimes A + (I_m - e_m\,\beta) \otimes I_l] = 0_{ml} \tag{3.3.13$a$}$$

$$\pi_k\, [B\,(e_m\,\beta - I_m) \otimes A + (I_m - e_m\,\beta) \otimes I_l]$$
$$+\pi_{k-1}\, [B\,(e_m\,\beta - I_m) \otimes a\, \alpha] = 0_{ml} \quad 1 < k < K\ , \tag{3.3.13$b$}$$

whereas postmultiplication of (3.3.10b) and (3.3.10c) by $(I_m \otimes (e_l\, \alpha - I_l))$ gives

$$\begin{aligned} &\pi_{k+1}\ [b\,\beta \otimes A\,(e_l\,\alpha - I_l)] + \\ &\pi_k\ [B \otimes A\,(e_l\,\alpha - I_l) - I_m \otimes (e_m\,\alpha - I_l)] = 0_{ml}\ , \quad 1 \le k < K-1\ . \end{aligned} \tag{3.3.14}$$

Although both (3.3.13) and (3.3.14) provide a relation between the vectors $\pi_k$ and $\pi_{k-1}$ for $1 < k < K$, neither of these relations yields a recursive solution for the vectors $\pi_k$, $1 \le k < K$, since all the matrices in square brackets in these equations turn out to be *singular*. Indeed, $(0, e_m \otimes e_l)$ is eaily seen to be a right eigenpair for both matrices upon postmultipliying them by $e_m \otimes e_l$. Now, recall the definition (2.3.3) of the matrices $M$ and $N$. Equation (3.3.13b) gives the relation

$$\begin{aligned} \pi_k\ N &= \pi_k\ [\,(I_m - e_m\,\beta) \otimes I_l + (e_m\,\beta - B) \otimes A] \\ &= \pi_k\ [e_m\,\beta \otimes A - B\,e_m\,\beta \otimes A] + \pi_{k-1}\,[B\,(I_m - e_m\,\beta) \otimes a\,\alpha] \\ &= \pi_k\,[b\,\beta \otimes A] + \pi_{k-1}\,[B\,(I_m - e_m\,\beta) \otimes a\,\alpha]\,, \quad 1 < k < K\ \ , \end{aligned}$$

which combines to (3.3.12) to yield

$$\pi_k\ N = \pi_k\ [b\,\beta \otimes (A - (e_l - a)\,\alpha)] + \pi_{k-1}[B \otimes a\,\alpha]\ , \quad 1 < k < K\ . \tag{3.3.15}$$

Similarly, use of (3.3.13a) and (3.3.10a) readily gives

$$\begin{aligned} \pi_1\ N &= \pi_1\ [b\,\beta \otimes A] \\ &= \pi_0\ [\beta \otimes (I_l - A)]\ \ . \end{aligned} \tag{3.3.16}$$

On the other hand, (3.3.14) can be rewritten in the form

$$\pi_k\ M = \pi_k\ [B \otimes a\,\alpha] + \pi_{k+1}\ [b\,\beta \otimes (A - (e_l - a)\,\alpha)\,]\ \ , \ \ 1 \le k < K-1\ \ . \tag{3.3.17}$$

Upon combining (3.3.15) and (3.3.17), the reader will check that

$$\pi_k\ N = \pi_{k-1}\ M\ , \qquad\qquad 1 < k < K\ \ . \tag{3.3.18}$$

The system of equations (3.3.18) can be solved recursively if either the matrix $N$ or the matrix $M$ is invertible. The explicit solution is presented when the matrix $N$ is invertible. Necessary and sufficient conditions for the invertibility of $N$ are given in Lemma 2.3.17 and Corollary 2.3.18, respectively. From (3.3.16), (3.3.10e) and (3.3.18) it follows that

$$\begin{aligned} \pi_1 &= \pi_0 \; (\beta \otimes (I_l - A)\,) \; N^{-1} \\ \pi_k &= \pi_{k-1} \; M\,N^{-1} \qquad\qquad 1 < k < K, \qquad\qquad (3.3.19) \\ \pi_K &= \pi_{K-1}\,(B \otimes a)\,(I_m - B)^{-1} \;. \end{aligned}$$

Postmultiplication of (3.3.10b)–(3.3.10d) by $(e_m \otimes I_l)$ and of (3.3.10e) by $e_m \otimes \alpha$ gives

$$\pi_0 \; a\,\alpha + \pi_1 \;\; (B\,e_m \otimes A + b \otimes a\,\alpha) + \pi_2(b \otimes A) = \pi_1\,(e_m \otimes I_l) \qquad (3.3.20a)$$

$$\pi_{k-1} \; (B\,e_m \otimes a\,\alpha\,) + \pi_k \;\; (B\,e_m \otimes A + b \otimes a\,\alpha) \qquad (3.3.20b)$$
$$\pi_{k+1} \; (b \otimes A) = \pi_k \;\; (e_m \otimes I_l) \;\; , \;\; 1 < k < K-1$$

$$\pi_{K-2} \;\; (B\,e_m \otimes a\,\alpha) + \pi_{K-1} \;\; (B\,e_m \otimes A + b \otimes a\,\alpha) \qquad (3.3.20c)$$
$$+\pi_K \;\; (b \otimes \alpha) = \pi_{K-1} \;\; (e_m \otimes I_l)$$

$$\pi_{K-1} \;\; (B\,e_m \otimes a\,\alpha) + \pi_K \;\; (B\,e_m \otimes \alpha) = \pi_K \;\; (e_m \otimes I_l) \qquad (3.3.20d)$$

Summation of (3.3.20b) over $k$, $1 < k < K-1$ and use of (3.3.20a), (3.3.20c), (3.3.20d) and (3.3.10a) lead after some simplification to the relation

$$\pi_0 + \sum_{i=1}^{K-1} \; \pi_k \;\; (e_m \otimes I_l) = \left( \pi_0 + \sum_{i=1}^{K-1} \; \pi_k(e_m \otimes I_l) \right) \;\; (A + a\,\alpha) \;\; . \qquad (3.3.21)$$

Equations (3.3.19) and (3.3.21) combine to give the main result, which is summarized in the following theorem.

**Theorem 3.3.7.** : *Let the* $1 \times l$ *row vector* $x$ *be the unique solution to the equation*

$$x\,(A + a\alpha) = x \quad , \quad x e_l = 1 \quad . \tag{3.3.22}$$

*If the matrix* $N$ *is invertible, then any invariant probabilty vector* $\pi = (\pi_0, \ldots, \pi_K)$ *has the following* matrix geometric *form*

$$\pi_k = \pi_0 \; S \; R^{k-1} \qquad\qquad 1 \le k < K \tag{3.3.23a}$$

$$\pi_K = \pi_0 \; S \; R^{K-2} \; U \; , \tag{3.3.23b}$$

*where the vector* $\pi_0$ *satisfies the equation*

$$\pi_0 \; Z = x \quad . \tag{3.3.24}$$

*The matrices R,S,U and Z are defined through the equations*

| | |
|---|---|
| $M = I_m \otimes (I_l - e_l\,\alpha) + B \otimes (e_l\,\alpha - A)$ | $m\,l \times m\,l$ matrix, |
| $N = (I_m - e_m\,\beta) \otimes I_l + (e_m\,\beta - B) \otimes A$ | $m\,l \times m\,l$ matrix, |
| $R = M \; N^{-1}$ | $m\,l \times m\,l$ matrix, |
| $S = [\beta \otimes (I_l - A)] \; N^{-1}$ | $l \times m\,l$ matrix, |
| $U = (B \otimes a)\,(I_m - B)^{-1}$ | $m\,l \times m$ matrix, |
| $W = I_l + \left[\sum_{i=1}^{K-1} \; S \; R^{i-1}\right] \; (e_m \otimes I_l)$ | $l \times l$ matrix, |
| $Z = W + S\,R^{K-2} \; U\,e_m\,x$ | $l \times l$ matrix. |

**Proof :** The irreducibility of the matrix $(A + a\alpha)$ implies that the system (3.3.22) has a unique solution $x$. Equation (3.3.23) follows from (3.3.19) and combines with (3.3.21) to yield

$$\pi_0\,W\,(A + a\alpha) = \pi_0\,W \;,$$

whence, $\pi_0\,W = c\,x$ for some constant $c$. The normalization condition $\pi\,e_r = 1$ and (3.3.23) lead to $c = 1 - \pi_K\,e_m$. Therefore,

$$\pi_0 W = (1 - \pi_K e_m) x = x - \pi_0 S R^{K-2}\, U e_m\, x \quad ,$$

and the relation

$$\pi_0 \left[W + SR^{K-2}\, Ue_m\, x\right] = \pi_0\, Z = x$$

follows upon grouping terms appropriately. This completes the proof.

△

## III.4 The effect of reversing the order of the servers :

In this section, the model of Section 3.2 is studied when the order of the servers is reversed, i. e., the first and the second node servers have PH-representations $(\beta, B)$ and $(\alpha, A)$, respectively. The states of the system are now

$$\begin{cases} (0,j) \ , & 1 \le j \le m \ , \quad k = 0, \\ (i,k,j) \ , & 1 \le i \le l, \, 0 < k < K, \text{ and } 1 \le j \le m, \\ (i,K) \ , & 1 \le i \le l \ , \quad k = K. \end{cases}$$

Note that now the *first* index denote the phase of the *second* node server. The invariant probability vector is denoted by $\widetilde{\pi}$ and is partitioned into $K+1$ block components $\widetilde{\pi}_k$, $0 \le k \le K$, as in Section 2. By lexicographically ordering the states, the one-step state transition matrix $\widetilde{T}$ of the underlying Markov chain can be obtained in the block tridiagonal form

$$\widetilde{T} = \begin{pmatrix} B & b\beta \otimes \alpha & & & & \\ B \otimes a & \begin{array}{c} B \otimes A + \\ b\beta \otimes a\alpha \end{array} & b\beta \otimes A & & & \\ & B \otimes a\alpha & \begin{array}{c} B \otimes A + \\ b\beta \otimes a\alpha \end{array} & b\beta \otimes A & & \\ & & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot & \\ & & & B \otimes a\alpha & \begin{array}{c} B \otimes A + \\ b\beta \otimes a\alpha \end{array} & b \otimes A \\ & & & & \beta \otimes a\alpha & A \end{pmatrix} \tag{3.4.1}$$

while the equation $\widetilde{\pi}\,\widetilde{T} = \widetilde{\pi}$ fot the invariant probability vector $\widetilde{\pi}$ can be rewritten as

$$\widetilde{\pi}_K \; A + \widetilde{\pi}_{K-1} \, (b \otimes A) = \widetilde{\pi}_K \qquad (3.4.2a)$$

$$\widetilde{\pi}_K \; (\beta \otimes a \, \alpha) + \widetilde{\pi}_{K-1} \, (B \otimes A + b \, \beta \otimes a \, \alpha) + \widetilde{\pi}_{K-2} \, (b \, \beta \otimes A) = \widetilde{\pi}_{K-1} \, (3.4.2b)$$

$$\widetilde{\pi}_{k+1} \; (B \otimes a \, \alpha) + \widetilde{\pi}_k \, (B \otimes A + b \, \beta \otimes a \, \alpha) + \widetilde{\pi}_{k-1} \, (b \, \beta \otimes A) = \widetilde{\pi}_k \qquad (3.4.2c)$$

$$1 < k < K - 1$$

$$\widetilde{\pi}_2 \, (B \otimes a \, \alpha) + \widetilde{\pi}_1 \, (B \otimes A + b \, \beta \otimes a \, \alpha) + \widetilde{\pi}_0 \, (b \, \beta \otimes \alpha) = \widetilde{\pi}_1 \qquad (3.4.2d)$$

$$\widetilde{\pi}_1 \, (B \otimes a) + \widetilde{\pi}_0 \; B = \widetilde{\pi}_0 \qquad (3.4.2e)$$

A direct comparision of (3.4.2) and (3.3.8) quickly reveals that

$$\pi_k = \widetilde{\pi}_{K-k} \quad , \qquad 0 \leq k \leq K \quad . \qquad (3.4.3)$$

Denoting the service phase in the first and the second node servers, and the buffer size at steady state by the random variables $PH1$, $PH2$ and $C$, respectively, (3.4.3) now reads as

$$P\{PH1 = i, \; C = 0\} = P_r\{PH2 = i, \; C = K\} \; ,$$

$$P\{PH1 = i, C = k, PH2 = j\} = P_r\{PH1 = j, C = K - k, PH2 = i\}, \;\; 0 < k < K,$$

$$P\{PH2 = j, \; C = K\} = P_r\{PH1 = j, \; C = 0\} \; ,$$

$$(3.4.4)$$

for $1 \leq i \leq l$, $1 \leq j \leq m$, where $P_r$ corresponds to the probability measure when the order of the servers is reversed. In particular the marginal probabilities satisfy the relations

$$P\{C = k\} = P_r\{C = K - k\} \; , \;\; 0 \leq k \leq K \; ,$$

$$P\{PH1 = i\} = P_r\{PH2 = i\} \; , \;\; 1 \leq i \leq l \; ,$$

$$P\{PH2 = j\} = P_r\{PH1 = j\} \; , \;\; 1 \leq j \leq m \; .$$

A probabilistic interpretation for this kind of *reversibility* is provided in Melamed [49] by embedding both the original and the reversed system in the same probability space and by viewing the vacant buffer locations (holes) of the original system as "occupied" by fictitious *dual jobs.* As *regular* jobs march through the buffer in one direction, the holes march in the opposite direction and they both receive *identical* service times.

## III.5. Numerical examples :

In this section, Theorem 3.3.7 is illustrated with specific PH-type distributions. Typical examples of PH-type distributions, such as Negative Binomial and Hypergeometric distributions, are used. First, three different examples are considered by using PH-type distributions with the following numerical values:

1. Negative Binomial of order 3 ($NB_3$) with $\mu_1 = 0.75$, $\mu_2 = 0.5$ and $\mu_3 = 0.6$,
2. Hypergeometric of order 3 ($HG_3$) with $\alpha = (0.5, 0.3, 0.2)$ and $Q = \text{diag}(0.75, 0.85, 0.8)$,
3. Hypergeometric of order 4 ($HG_4$) with $\alpha = (0.3, 0.1, 0.2, 0.4)$ and $Q = \text{diag}(0.4, 0.5, 0.2, 0.5)$.

The expected service time $E\,S(\cdot)$ can be found by using $P'(1) = \alpha(I_m - Q)^{-1}\, e_m$. The numerical values given above lead to $E\,S(NB_3) = 5$, $E\,S(HG_3) = 5$ and $E\,S(HG_4) = 1.75$.

The following three examples are considered.

| | 1st Node | 2nd Node |
|---|---|---|
| (i) | $NB_3$ | $HG_3$ |
| (ii) | $HG_3$ | $HB_3$ |
| (iii) | $HG_4$ | $HG_3$ |

In Table (3.5.1), the values of $p_k = \pi_k\, e$, the invariant probability of having $k$ customers in the buffer, are displayed for buffer size K=6. The calculations are carried out in double precision complex arithmetic up to 15 digits after the decimal point. Only the first four digits after the decimal point are displayed in Table (3.5.1). The last row checks the normalizing condition and it is accurate up to $10^{-15}$. Many of the matrices defined in Theorem 3.3.7 are of the form $X = B\,A^{-1}$ and are computed as the solution of equation $X\,A = B$, by using Gaussian Elimination, instead of directly computing the inverses.

$K = 6$

| buffer size | $NB_3 - HG_3$ | $HG_3 - NB_3$ | $HG_4 - HG_3$ |
|---|---|---|---|
| 0 | 0.0875 | 0.0875 | $1.0836 \times 10^{-4}$ |
| 1 | 0.1773 | 0.1447 | $5.9041 \times 10^{-4}$ |
| 2 | 0.1705 | 0.1645 | 0.0027 |
| 3 | 0.1680 | 0.1680 | 0.0120 |
| 4 | 0.1645 | 0.1705 | 0.0560 |
| 5 | 0.1447 | 0.1773 | 0.2786 |
| 6 | 0.0875 | 0.0875 | 0.6500 |
| sum | 1.0 | 1.0 | 1.0 |

Table 3.5.1.

Cases (i) and (ii) show the effect of reversing the order of the servers and the results confirm the conclusion of Section 4. In the third case, the first node server is being faster than the second node server, the buffer is usually full and the first node server is mostly in blocked position.

As a second example, a system that contains *unreliable* servers with *constant* production times of one time slot is considered. The up and down times of the servers are assumed *geometric* with parameters $p$ and $q$, respectively, and the servers are assumed to fail only when processing a job. A two node blocking system that involves such unreliable servers has been studied by many authors (class ii.a). The service distribution of such an unreliable server has a PH-type distribution which can be represented by the pair $(\alpha, Q)$, where

$$\alpha = (1,0) \qquad \text{and} \qquad Q = \begin{pmatrix} 0 & \overline{p} \\ q & \overline{q} \end{pmatrix}.$$

The following numerical values are used for the servers,

$$p_1 = 0.95 \qquad p_2 = 0.6$$

$$q_1 = 0.8 \qquad q_2 = 0.2$$

For these values, the expected service times for server one and server two are 1.12 and 5, respectively. The vectors $\pi_k$ and the probability of finding $k$ customers at steady state, $P\{C = k\}$, $0 \leq k \leq 6$ , are shown in Table 3.5.2. As expected the buffer is mostly full since the first server is faster than the second one.

$K = 6$

| buffer size ($k$) | $\pi_k$ | $P\{C = k\}$ |
|---|---|---|
| 0 | $10^{-4} \times (0.5356,\ 0.2965)$ | $0.8321 \times 10^{-4}$ |
| 1 | $10^{-3} \times (0.4528, 0.0622, 0.0315, 0.0182)$ | $0.5647 \times 10^{-3}$ |
| 2 | $10^{-3} \times (0.8628, 0.2936, 0.3311, 0.0643)$ | 0.0016 |
| 3 | $(0.0066, 0.0007, 0.0010, 0.0003)$ | 0.0086 |
| 4 | $(0.0011, 0.0057, 0.0056, 0.0008)$ | 0.0133 |
| 5 | $(0.1866, 0.0003, 0.0079, 0.0048)$ | 0.1996 |
| 6 | $(0.1306, 0.6457)$ | 0.7763 |

sum = 1.0 exact up to 15 digits.

**Table 3.5.2.**

A word of caution: After computing the matrices $M$ and $N$ to form $R$, the use of the equations $\pi_2 = \pi_1 R$ and (3.3.10a) in (3.3.10b) gives

$$\pi_1 = L\,\pi_1 \tag{3.5.1}$$

with the matrix $L$ being given by

$$L = (b \otimes A)\,(I_l - A)^{-1}\,(\beta \otimes a\,\alpha) + B \otimes A + b\,\beta \otimes a\,\alpha + R\,(b\,\beta \otimes A) \qquad (3.5.2)$$

To get the invariant probability vector $\pi$, a direct approach might be to solve (3.5.1) for $\pi_1$, as suggested in Neuts [58, p. 126] for the continuous-time version of the problem. The difficulty with such a procedure is that the solution $\pi_1$ to equation (3.5.1) may not be unique. For instance, in the last example the eigenvalues of $L$ are 1, 1, 0.1294 and -0.0494, and the vectors $x_1 = (0.9995, -0.0220, 0.0071, 0.0220)$ and $x_2 = (1.3539, 1.2397, 0.5084, 0.1745)$ are both eigenvectors of $L$ with eigenvalue $\lambda = 1$. Hence, the eigenvectors of $L$ that corresponds to $\lambda = 1$ lie in the subspace spanned by $x_1$ and $x_2$. Therefore, the eigenvalue problem (3.5.1) does not yield to a unique solution $\pi_1$, and (3.3.24) needs to be solved for $\pi_0$ in order to get the invariant probability vector $\pi$.

## III.6. The continuous-time formulation :

In this section, a *complete* solution to the same problem is presented in continuous-time to show that the two formulations do not subsume each other. The continuous-time formulation has two major advantages: The underlying continuous-time Markov chain is *always* irreducible, and the computations to obtain the explicit solution are simpler.

The service distributions of the first and the second node servers again admit representations $(\alpha, A)$ and $(\beta, B)$, respectively, which are assumed irreducible. The row vectors $\alpha$ and $\beta$ and the infinitesimal generator matrices $A$ and $B$ have dimensions $1 \times l$, $1 \times m$, $l \times l$ and $m \times m$, respectively. The corresponding $l \times 1$ and $m \times 1$ column vectors of absorption rates for the first and the second node server are denoted by $a$ and $b$, respectively.

The state-space of the continuous-time Markov chain is defined as in Section 2, while the corresponding generator matrix $T$ of the underlying Markov process can be obtained by ordering the states in the lexicographical order as before. The resulting infinitesimal generator matrix then takes the form

$$T = \begin{pmatrix} A & \beta \otimes a\,\alpha & & & & \\ b \otimes I_l & \begin{array}{c} B \otimes I_l + \\ I_m \otimes A \end{array} & I_m \otimes a\,\alpha & & & \\ & b\,\beta \otimes I_l & \begin{array}{c} B \otimes I_l + \\ I_m \otimes A \end{array} & I_m \otimes a\,\alpha & & \\ & & \ddots & \ddots & \ddots & \\ & & & b\,\beta \otimes I_l & \begin{array}{c} B \otimes I_l + \\ I_m \otimes A \end{array} & I_m \otimes a \\ & & & & b\,\beta \otimes \alpha & B \end{pmatrix} . \qquad (3.6.1)$$

The irreducibility of this continuous-time Markov chain is first established.

**Theorem 3.6.1. :** *For the contiuous-time version of the model described in Section 2, the underlying continuous-time Markov chain is always* **irreducible.**

**Proof :** In the continuous-time formulation, two events cannot occur simultaneously (with a positive probability) unlike for the discrete-time case. This fact together with the independence and irreducibility of the service representations $(\alpha, A)$ and $(\beta, B)$, yields the following access relations:

$$\begin{aligned}
&(i,0) \rightarrow (i',1,j') \\
&(i,k,j) \rightarrow \begin{cases} (i',k+1,j) & 0 < k < K-1, \\ (K,j) & k = K-1 \end{cases} \\
&(i,k,j) \rightarrow \begin{cases} (i,k-1,j') & 1 < k < K, \\ (i,0) & k = 1, \end{cases} \\
&(K,j) \rightarrow (i',K-1,j')
\end{aligned} \tag{3.6.2}$$

for every $1 \leq i, i' \leq l$ and $1 \leq j, j' \leq m$.

It is now easy to see that starting from an arbitrary state $(i,k,j)$, any other arbitrary state $(i', k+n, j')$ can be reached, for $1 < k,\ k+n < K-1$. For the case when $n \geq 0$, based on (3.6.2), there is a path from the state $(i,k,j)$ to $(i', k+n+1, j)$, and then from this state to the state $(i', k+n, j')$. A similar argument can be made when $n < 0$, first by fixing $i$ and then $j$. Since these two states are arbitrary all the states of the Markov chain in the sets $E_k$, $1 < k < K-1$, communicate, where the sets $E_k$, $0 \leq k \leq K$, are defined as in Section 3. Similar arguments reveal that the states in the sets $E_0$ and $E_1$, and $E_{K-1}$ and $E_K$, respectively, also communicate. Therefore, all the states in $E$ communicate with each other since the sets $E_{K-2}$ and $E_{K-1}$, and $E_1$ and $E_2$, respectively, are reachable from each other.

$\triangle$

In order to obtain a closed form solution for the invariant vector $\pi$, the equation

$\pi\ T = 0_r$ it satisfies can be rewritten as

$$\pi_0\ A + \pi_1\,(b \otimes I_l) = 0_l \qquad (3.6.3a)$$

$$\pi_0\ (\beta \otimes a\,\alpha) + \pi_1\,(B \otimes I_l + I_m \otimes A) + \pi_2\,(b\,\beta \otimes I_l) = 0_{lm} \qquad (3.6.3b)$$

$$\pi_{k-1}\ (I_m \otimes a\,\alpha) + \pi_k\,(B \otimes I_l + I_m \otimes A) + \pi_{k+1}\,(b\,\beta \otimes I_l) = 0_{lm} \qquad (3.6.3c)$$

$$1 < k < K - 1$$

$$\pi_{K-2}\,(I_m \otimes a\,\alpha) + \pi_{K-1}\,(B \otimes I_l + I_m \otimes A) + \pi_K\,(b\,\beta \otimes \alpha) = 0_{lm} \qquad (3.6.3d)$$

$$\pi_{K-1}\,(I_m \otimes a) + \pi_K\ B = 0_m \qquad (3.6.3e)$$

Postmultiplication of (3.6.3a) by $e_l$, of (3.6.3b)-(3.6.3d) by $(e_m \otimes e_l)$ and of (3.6.3) by $e_m$, after simplifications, gives

$$\pi_k\,(b \otimes e_l) - \pi_{k-1}\,(e_m \otimes a) = 0\ , \qquad 1 < k < K\ \ , \qquad (3.6.4)$$

while postmultiplication of (3.6.4) with $(\beta \otimes \alpha)$ yields

$$\pi_k\,(b\,\beta \otimes e_l\,\alpha) = \pi_{k-1}\ (e_m\,\beta \otimes a\,\alpha)\ ,\ \ 1 < k < K\ \ . \qquad (3.6.5)$$

Now, postmultiplication of (3.6.3b)-(3.6.3d) with $(e_m\,\beta \otimes I_l)$ and use of (3.6.3b)-(3.6.3d) imply

$$\pi_1\ [(B + b\beta) \otimes I_l + (I_m - e_m\,\beta) \otimes A] = 0_{lm} \qquad (3.6.6a)$$

$$\begin{aligned}\pi_k\ &[(B + b\beta) \otimes I_l + (I_m - e_m\,\beta) \otimes A]\\ &+\pi_{k-1}\ [(I_m - e_m\,\beta) \otimes a\,\alpha] = 0_{lm}\ ,\quad 1 < k < K\ \ .\end{aligned} \qquad (3.6.6b)$$

Postmultiplication of (3.6.3b) and (3.6.3c) by $(I_m \otimes (e_l\,\alpha - I_l))$ yields

$$\begin{aligned}&\pi_{k+1}\ [b\,\beta \otimes (I_l - e_l\,\alpha)]\\ &+\pi_k\ [B \otimes (I_l - e_l\,\alpha) + I_m \otimes (A + a\,\alpha)] = 0_{lm}\ ,\quad 1 \le k < K - 1\ .\end{aligned} \qquad (3.6.7)$$

Both (3.6.6) and (3.6.7) provides a relation between the vectors $\pi_k$ and $\pi_{k-1}$ for $1 < k < K$. However, neither of these relations yield a recursive solution for the vectors $\pi_k$, $1 \leq k < K$, since all the matrices in these equations that are in square brackets are singular as discussed in the discrete-time case.

Recall the definitions of the matrices $\widetilde{M}$ and $\widetilde{N}$ given in (2.3.6). Equation (3.6.6b) can now be rewritten as

$$\begin{aligned}
\pi_k\ \widetilde{N} &= \pi_k\ [\,(I_m - e_m\,\beta) \otimes A + B \otimes I_l] \\
&= -\pi_k\,[b\,\beta \otimes I_l] - \pi_{k-1}\ [(I_m - e_m\,\beta) \otimes a\,\alpha]\ ,\quad 1 < k < K\ ,\ \ (3.6.8) \\
&= -\pi_k\ [b\,\beta \otimes (I_l - e_l\,\alpha)] - \pi_{k-1}[I_m \otimes a\,\alpha]\ ,\quad 1 < k < K\ ,\qquad (3.6.9)
\end{aligned}$$

where the last equality is obtained by direct substitution of (3.6.5) into (3.6.8). Similarly, use of (3.6.6a) and (3.6.3a) gives

$$\pi_1\ \widetilde{N} = -\pi_1\ [b\,\beta \otimes I_l] = \pi_0\ [\beta \otimes A]\quad . \qquad (3.8.10)$$

Moreover, (3.6.1a) and (3.6.7) combine to readily yield

$$\pi_k\ \widetilde{M} = -\pi_k\ \ [I_m \otimes a\,\alpha] - \pi_{k+1}\ \ [b\,\beta \otimes (I_l - e_l\alpha)]\qquad 1 \leq k < K-1\ ,\qquad (3.6.11)$$

whence

$$\pi_k\ \widetilde{N} = \pi_{k-1}\ \widetilde{M}\ ,\qquad 1 < k < K\quad . \qquad (3.6.12)$$

by direct inspection of (3.6.9) and (3.6.11).

Since the matrices $\widetilde{N}$ and $\widetilde{M}$ are both invertible owing to Lemma 2.3.19., it follows from (3.6.10), (3.6.3a) and (3.6.3e) that

$$\begin{aligned}
\pi_1 &= \pi_0\ (\beta \otimes A)\ \widetilde{N}^{-1} \\
\pi_k &= \pi_{k-1}\ \widetilde{M}\,\widetilde{N}^{-1} \qquad\qquad 1 < k < K \qquad\qquad (3.6.13) \\
\pi_K &= -\pi_{K-1}\ (I_m \otimes a)\,B^{-1}\quad .
\end{aligned}$$

Postmultiplication of (3.6.3b)–(3.6.3d) by $(e_m \otimes I_l)$ and of (3.6.3d) by $(e_m \otimes \alpha)$ and summing over $k$, $1 < k < K-1$, readily give the relation

$$\left( \pi_0 + \sum_{k=1}^{K-1} \pi_k (e_m \otimes I_l) \right) \; (A + a\,\alpha) = 0 \quad . \tag{3.6.14}$$

Equations (3.6.13) and (3.6.14) combine to give the main result which is summarized in the following theorem, whose proof goes along the same lines as the proof of Theorem 3.3.7.

**Theorem 3.6.2 :** *If the $1 \times l$ row vector $x$ is the unique solution to the equation*

$$x\,(A + a\alpha) = 0_l \ , \ \ x e_l = 1 \ , \tag{3.6.15}$$

*then the invariant probabilty distribution vector $\pi = (\pi_0, \ldots, \pi_K)$ has the following* **matrix geometric** *form*

$$\pi_k = \pi_0 \ S \ R^{k-1} \qquad\qquad 1 \le k < K \tag{3.6.16a}$$

$$\pi_K = \pi_0 \ S \ R^{K-2} \ U \ , \tag{3.6.16b}$$

*where the vector $\pi_0$ satisfies the equation*

$$\pi_0 \ Z = x \quad . \tag{3.6.17}$$

*The matrices R,S,U and Z are defined through the equations*

$\widetilde{M} = B \otimes (I_l - e_l\,\alpha) + I_m \otimes A$ — $m\,l \times m\,l$ matrix,

$\widetilde{N} = (I_m - e_m\,\beta) \otimes A + B \otimes I_l$ — $m\,l \times m\,l$ matrix,

$R = \widetilde{M} \ \widetilde{N}^{-1}$ — $m\,l \times m\,l$ matrix,

$S = [\beta \otimes A] \ \widetilde{N}^{-1}$ — $l \times m\,l$ matrix,

$U = -(I_m \otimes a) \ B^{-1}$ — $m\,l \times l$ matrix,

$W = I_l + \left[\sum_{k=1}^{K-1} \ S \ R^{k-1}\right] \ (e_m \otimes I_l)$ — $l \times l$ matrix,

$Z = W + S\,R^{K-2} \ U\,e_m\,x$ — $l \times l$ matrix.

# CHAPTER IV

# TWO NODE SYSTEM WITH GEOMETRIC AND PHASE TYPE SERVERS

## IV.1. Introduction :

In this chapter, a special case of the model studied in Chapter III is considered when one of the servers is *geometric*. The discussion focuses on the situation where the *first* and *second* node servers have *geometric* and *phase type* service distributions, respectively. The case when the order of the servers is reversed is only briefly outlined as the results of the previous chapter clearly hold in this case. A system state process is introduced and its probabilistic structure is carefully described. For this special case, unlike for the general model of Chapter III, the underlying Markov chain is shown to *always* have a *single* ergodic class, whence a unique invariant probability vector. The possibility of having transient states is established and these transient states are identified in terms of the PH-representation of the PH-server. Necessary and sufficient conditions are given for the irreducibility of this Markov chain. Explicit expressions for the unique invariant probability vector are obtained in *matrix-geometric* form and attempts are made to get an expression for the (average) system throughput, in order to study its behaviour as a function of the intermediate buffer capacity. The discussion of the discrete-time model concludes by applying the results to some specific PH-type distributions. Finally, the continuous-time formulation of the model is briefly discussed and similar results are obtained: As in Chapter III, the irreducibility of the underlying continuous-time Markov process is established and the unique invariant vector is given in matrix geometric form.

## IV.2. The discrete-time model (Geometric/PH) :

The model to be discussed consists of two nodes separated by a *finite* intermediate buffer of capacity $K$, i. e., there are exactly $K$ positions in the buffer, inclusive of the one taken by the job in service at the second node server. There is an *infinite* supply of job units available in front of the first node server, which has *geometric* service time distribution of parameter $\mu$ with $0 < \mu < 1$. Moreover, the second node server never gets blocked and has a PH-type service time distribution with representation $(\alpha, Q)$ of order $m$, the notation being the one introduced in Chapter II. *The immediate* blocking strategy is again adopted.

The state space of this system is naturally defined to be the set $E = \{0\} \cup \{1,2,\ldots,K\} \times \{1,2,\ldots,m\}$, i. e.,

$$E = \begin{cases} (k,j) \ , & 1 \leq k \leq K, \ 1 \leq j \leq m, \\ 0 \ , & k = 0, \end{cases}$$

where $k$ and $j$ indicate the buffer content and the service phase of the second node server, respectively. When the buffer is empty, the second node server has *no* jobs to process and its phase is thus not defined; the corresponding state is indicated simply by 0. Note that $E$ is the state-space of a Markov chain, the so-called *system process* which is different from the one given in Chapter III owing to the "memoryless" nature of geometric distribution.

The invariant probability vector is denoted by the $1 \times (K \cdot m + 1)$ row vector $\pi$, which is partitioned into $K+1$ blocks of components, say, $\pi = (\pi_0, \pi_1, \ldots, \pi_K)$, with first entry $\pi_0$ being scalar and $\pi_k$, $1 \leq k \leq K$, being $1 \times m$ row vectors. The $j^{th}$ entry of $\pi_k$, denoted by $\pi_{kj}$, represents the invariant probability of the state $(k,j)$, $1 \leq k \leq K$, $1 \leq j \leq m$, while $\pi_0$ is the invariant probability of having an empty buffer.

## IV.3. Analysis of the discrete-time model :

If the states are ordered lexicographically, say in the order $0, (1,1), (1,2), \ldots, (1,m), (2,1), \ldots, (2,m), \ldots, (K,1), \ldots, (K,m)$, the corresponding one-step transition matrix $T$ is a $(K \cdot m + 1) \times (K \cdot m + 1)$ *block tridiagonal* matrix. If $T_{k_1,k_2}$ again denotes the block corresponding to transitions from queue size $k_1$ to queue size $k_2$, then

$$\begin{array}{ll} A_{k,k-1} = \overline{\mu} p \alpha & 1 < k < K, \\ A_{k,k+1} = \mu Q & 1 \leq k < K, \\ A_{k,k} \quad = \overline{\mu} Q + \mu p \alpha & 1 \leq k < K, \end{array}$$

where each block is an $m \times m$ matrix and $\overline{\mu}$ is used for $1 - \mu$. For the *boundary* states, the entries are

$$\begin{array}{ll} A_{0,0} = \overline{\mu} & \text{scalar,} \\ A_{1,0} = \mu p & m \times 1 \text{ column vector,} \\ A_{0,1} = \mu \alpha & 1 \times m \text{ row vector,} \\ A_{K,K-1} = p \alpha & m \times m \text{ matrix,} \\ A_{K,K} = Q & m \times m \text{ matrix.} \end{array}$$

The resulting matrix is thus

$$T = \begin{pmatrix} \overline{\mu} & \mu\alpha & & & & & & \\ \overline{\mu}p & \overline{\mu}Q + \mu p\alpha & \mu Q & & & & & \\ & \overline{\mu}p\alpha & \overline{\mu}Q + \mu p\alpha & \mu Q & & & & \\ & & \cdot & \cdot & \cdot & & & \\ & & & \cdot & \cdot & \cdot & & \\ & & & & \cdot & \cdot & \cdot & \\ & & & & & \overline{\mu}p\alpha & \overline{\mu}Q + \mu p\alpha & \mu Q \\ & & & & & & p\alpha & Q \end{pmatrix} . \quad (4.3.1)$$

The next theorem gives a necessary and sufficient condition for the irreducibility of the matrix $T$ and identifies its transient and ergodic states in terms of the properties of the matrix $Q$. The corresponding invariant probability vector is shown to be *unique*, and is obtained explicitly in *matrix geometric* form.

**Theorem 4.3.1.** : *The Markov chain with one-step probability transition matrix $T$ is irreducible if and only if the index set*

$$\Gamma := \{1 \leq j \leq m : \ Q_{ij} = 0 \ for \ all \ 1 \leq i \leq m\}$$

*is empty. Moreover, if $\Gamma \neq \emptyset$ then the set of states $\Theta := \{(K,j) \ : \ j \in \Gamma\}$ will be transient and the rest of the states will form a* single *irreducible class.*

**Proof :** As in Chapter III, let the sets $E_k$, $0 \leq k \leq K$, be defined by

$$E_k := \begin{cases} \{x \in E : x = (k,j), \ 1 \leq j \leq m\}, & 0 < k \leq K, \\ \{0\}, & k = 0. \end{cases}$$

From the structure of the matrix $T$, the transitions among the states within $E_k, 0 < k < K$, are clearly governed by the directed graph $G(\overline{\mu}Q + \mu p\alpha)$. By the irreducibility assumption on the PH-representation, the matrix $Q + p\alpha$ is irreducible, and so is the matrix $\overline{\mu}Q + \mu p\alpha$ by Lemma 2.3.3; the states in each component $E_k$, $0 < k < K$, therefore communicate with each other, without leaving $E_k$. On the other hand, $Q > 0$ and $p\alpha > 0$, since both $\alpha > 0$ and $p > 0$ as discussed in Section 2.2. Therefore the sets $E_k$, $0 \leq k < K$, are reachable from each other, whence the set $E \setminus E_K$ is composed of states that communicate with each other without leaving it.

Since the transitions to buffer size $K$ are due to a service completion in the geometric server *only*, the set $\Theta$ by its very definition can never be reached from the set $E \setminus \Theta$. On the other hand, all the states in the set $E_K \setminus \Theta$ communicate with all the states in the set $E_{K-1}$ through a service completion in either of the servers owing to the irreducibility of the PH-representation. Therefore, the set $E \setminus \Theta$ form a single irreducible class for the state set $E$, whereas the set $\Theta$ is the transient class of the Markov chain, and the statement of the theorem readily follows.

$\triangle$

The next corollary is now an immediate consequence of Theorems 4.3.1 and 2.3.14.

**Corollary 4.3.2.** : *The invariant probability distribution of the Markov chain with one-step transition matrix $T$ is* unique.

In order to emphasize the simplicity of this model as it compares to the model of Chapter III, the unique invariant probability vector is obtained by direct manupilations, without specializing Theorem 3.3.7 to the case where $A = \overline{\mu}$, $\alpha = 1$ and $a = \mu$, although this would give the same result.

In order to obtain this unique solution vector, the equation

$$\pi\, T = \pi \tag{4.3.2}$$

is rewritten explicitly in terms of the block entries of the matrix $T$ and of the vector

$\pi$, in the form

$$\overline{\mu}\pi_0 + \overline{\mu}\pi_1 p = \pi_0 \tag{4.3.3a}$$

$$\mu\ \pi_0\ \alpha + \pi_1\ [\overline{\mu}\, Q + \mu\, p\, \alpha] + \overline{\mu}\ \pi_2\ p\ \alpha = \pi_1 \tag{4.3.3b}$$

$$\mu\pi_{k-1}Q + \pi_k[\overline{\mu}Q + \mu p\alpha] + \overline{\mu}\pi_{k+1}p\alpha = \pi_k \qquad 1 < k < K-1 \tag{4.3.3c}$$

$$\mu\pi_{K-2}Q + \pi_{K-1}[\overline{\mu}Q + \mu p\alpha] + \pi_K p\alpha = \pi_{K-1} \tag{4.3.3d}$$

$$\mu\pi_{K-1}Q + \pi_K Q = \pi_K \quad . \tag{4.3.3e}$$

Postmultiplication of equation (4.3.3c) by $e_m$ and use of the simple identities

$$Qe_m = e_m - p \qquad \text{and} \qquad \alpha e_m = 1 \tag{4.3.4}$$

give

$$\overline{\mu}\pi_{k+1}p - \mu\pi_k(e_m - p) = \overline{\mu}\pi_k p - \mu\pi_{k-1}(e_m - p) \quad , \quad 1 < k < K-1. \tag{4.3.5}$$

With the notation,

$$\gamma_k := \overline{\mu}\pi_k p - \mu\pi_{k-1}(e_m - p) \quad , \quad 1 \leq k \leq K,$$

the relation (4.3.5) takes the equivalent form

$$\gamma_2 = \gamma_3 = \ldots = \gamma_{K-2}. \tag{4.3.6}$$

Now, upon post multiplying (4.3.3b) with $e_m$ and using (4.3.3a), it follows that

$$\gamma_2 = \overline{\mu}\pi_2 p - \mu\pi_1(e_m - p) = 0. \tag{4.3.7}$$

Similar calculations in equations (4.3.3d) and (4.3.3e) lead to

$$\gamma_{K-1} = 0 \ , \tag{4.3.8}$$

and (4.3.6),(4.3.7) and (4.3.8) thus combine to yield

$$\gamma_k = 0 \ , \qquad 1 < k < K \ . \tag{4.3.9}$$

Consequently,

$$\gamma_k \alpha = 0_m \quad or \quad \overline{\mu}\pi_{k+1} p\alpha = \mu\pi_k(e_m - p)\alpha, \quad 1 < k < K, \tag{4.3.10}$$

and substitution of (4.3.10) into (4.3.3c) gives

$$\mu\pi_{k-1}Q = \pi_k[I_m - \overline{\mu}Q - \mu e_m\alpha] \ , \quad 1 < k < K \ . \tag{4.3.11}$$

Note that (4.3.11) agrees with (3.3.18) since for this special case, $M = \mu Q$ and $N = (I_m - \overline{\mu}Q - \mu e_m \alpha)$.

As readily observed, the $m \times m$ matrix $(I_m - \overline{\mu}Q - \mu e_m\alpha)$ is *stable*, thus nonsingular, and for ease of notation, it is convenient to pose

$$S := \mu\left(I_m - \overline{\mu}Q - \mu e_m\alpha\right)^{-1} \tag{4.3.12$a$}$$

$$R := QS \ . \tag{4.3.12$b$}$$

With this notation, (4.3.11) takes the equivalent form

$$\pi_k = \pi_{k-1} \ R, \qquad 1 < k < K. \tag{4.3.13}$$

so

$$\pi_k = \pi_1 \ R^{k-1}, \qquad 1 \leq k < K \ , \tag{4.3.14}$$

Solving (4.3.3e) for $\pi_K$ yields

$$\pi_K = \mu\pi_{K-1}Q(I_m - Q)^{-1} \quad , \tag{4.3.15}$$

while (4.3.7) and (4.3.3b) lead to

$$\pi_1 = \pi_0\alpha S \ . \tag{4.3.16}$$

As usual, the scalar $\pi_0$ is determined through the normalization condition

$$\pi_0 + \sum_{k=1}^{K} \pi_k e_m = 1 \quad . \tag{4.3.17}$$

Upon substituting (4.3.13)–(4.3.16) into (4.3.17), it is a simple matter to check that

$$\pi_0 = (1 + \alpha S T(K) e_m)^{-1}, \tag{4.3.18}$$

where

$$T(K) = \sum_{k=1}^{K-1} R^{k-1} + \mu R^{K-1}(S - R)^{-1}. \tag{4.3.19}$$

Nonsingularity of the matrix $(S - R)$ in (4.3.19) follows from the nonsingularity of the matrix $(I_m - Q)$ and (4.3.12b).

This discussion is now summarized in the following theorem.

**Theorem 4.3.3 :** *The unique invariant probability vector* $\pi = (\pi_0, \pi_1, \ldots, \pi_K)$ *is given by*

$$\pi_0 = (1 + \alpha S T(K) e_m)^{-1} \quad , \tag{4.3.20a}$$

$$\pi_k = \pi_0 \alpha S R^{k-1} \,, \qquad 1 \leq k < K, \tag{4.3.20b}$$

$$\begin{aligned} \pi_K &= \mu \pi_{K-1} Q (I_m - Q)^{-1} \\ &= \mu \pi_0 \alpha S R^{K-2} Q (I_m - Q)^{-1} \,, \end{aligned} \tag{4.3.20c}$$

*where* $S, R$ *and* $T(K)$ *are defined by equations (4.3.12a), (4.3.12b) and (4.3.19), respectively.*

From the Neuman expansion

$$S = \mu \left[ I_m - (\bar{\mu}\, Q + \mu\, e_m\, \alpha) \right]^{-1} = \mu \sum_{i=0}^{\infty} (\bar{\mu}\, Q + \mu\, e_m\, \alpha)^i \,, \tag{4.3.21}$$

and from the fact that $(\overline{\mu}\, Q + \mu\, e_m\, \alpha)$ is a positive matrix, it follows that $S$ is also a positive matrix, and so is the matrix $R$. A similar reasoning shows that $(I_m - Q)^{-1}$ is a positive matrix. Also $1 \times m$ row vector $\alpha S >> 0$. This can be shown by the following steps: Since $0 < \mu < 1$, it is an easy exercise to check that elementwise $\sum_{i=0}^{\infty}(\overline{\mu}Q + \mu e_m \alpha)^i > \sum_{i=0}^{\infty}(\overline{\mu}Q)^i + \mu^i e_m \alpha$. In view of this observation and (4.3.21)

$$\frac{1}{\mu}\alpha S > \sum_{i=0}^{\infty} \overline{\mu}^i (\alpha Q^i) + \mu^i \alpha \; . \tag{4.3.22}$$

If $\alpha_j > 0$ then $(\alpha S)_j > 0$ owing to (4.3.22). On the other hand if $\alpha_j = 0$, $1 \leq j \leq m$ then

$$\frac{1}{\mu}(\alpha S)_j > \sum_{l=1}^{m}\sum_{i=0}^{\infty} \overline{\mu}^i \alpha_l Q^i_{lj} = \sum_{i=0}^{\infty}\sum_{l \in L^*} \overline{\mu}^i \alpha_l Q^i_{lj} \; ,$$

where $L^* = \{1 \leq l \leq m : \alpha_l > 0\}$ and $L^* \neq \emptyset$ owing to the assumptions made throughout the thesis. Since $\overline{\mu} > 0$ and $Q > 0$, $(\alpha S)_j = 0$ if and only if $Q^i_{lj} = 0$ for every $l \in L^*$ and $i$, $0 \leq i < \infty$. But since $j$ is not in $L^*$, if $Q^i_{lj} = 0$ for every $l$ in $L^*$ and $i$, then the state $j$ of the PH-representation is not reachable from the set $L^*$ in *finite* number of transitions, thus contradicting the irreducibility of the PH-representation. Whence the inequality $\alpha S >> 0$ holds.

In view of the above conclusion the analytical results which were obtained in Theorem 4.3.3 readily check with the statement of Theorem 4.3.2: Indeed, a similar argument to the one given above shows that the nonnegative matrix $R = QS$ has no zero column and since $\alpha S >> 0$, from (4.3.20b) the inequality $\pi_k >> 0$, $0 \leq k < K$, easily follows. In particular, $\pi_{K-1} >> 0$ and if the $j^{th}$ column of the matrix $Q$ has all zero entries, then clearly every matrix $Q^i$, $i \geq 1$, will have all zero entries in the $j^{th}$ column, and so does the matrix $Q(I_m - Q)^{-1}$ since

$$Q(I_m - Q)^{-1} = Q + Q^2 + Q^3 + \ldots \; .$$

$x = 0$ for $(I_m - \overline{\mu}Q)$ is nonsingular. Therefore, the matrix $(I_m - \overline{\mu}Q - \mu e_m \alpha)$ is nonsingular, and so is the matrix $S$.

The effect of reversing the order of the servers discussed in Section III.5 for the more general model clearly still holds for this special case. Denoting the service phase of the PH-type server and the buffer size at steady state by the random variables $PH$ and $C$, respectively, equation (3.5.4) now takes the form

$$P\{PH = j,\ C = k\} = P_r\{PH = j,\ C = K - k\}\ ,\quad 1 \le j \le m,\ 0 < k \le K\ ,$$

$$P\{C = 0\} = P_r\{C = K\}$$

where $P_r$ again corresponds to the probability measure when the order of the servers is reversed.

## IV.4. System throughput and some concavity results :

From the exact analytical expressions for the steady state probabilities, an explicit expression for the average system throughput is developed as a function of the intermediate buffer size $K$.

In terms of the parameters defined earlier, the average system throughput $Tr(K)$, when the buffer size is $K$, is defined as the departure probability of a job from the system during a time slot, and is given by the expression

$$Tr(K) = \sum_{k=1}^{K} \pi_k p \quad . \tag{4.4.1}$$

From equation (4.3.20), it readly follows that

$$Tr(K) = \frac{\alpha\, S\, T(K)\, p}{1 + \alpha\, S\, T(K)\, e_m} \tag{4.4.2}$$

where $T(K)$ is defined by (4.3.19).

**Lemma 4.4.1 :** *If $T(K)$ is defined as in equation (4.3.19), then*

$$T(K+1) - T(K) = \mu\, R^{K-1}\, e_m\, \alpha\, (I_m - Q)^{-1} \ , \ K \geq 1, \tag{4.4.3}$$

*holds and $T(K)$ increases (elementwise) with $K$.*

**Proof :** From (4.3.21), simple algebra yields

$$\begin{aligned} T(K+1) - T(K) &= I_m - (I_m - R)\, T(K) \\ &= I_m - (I_m - R) \left[ \sum_{i=1}^{K-1} R^{i-1} + \mu\, R^{K-1}\, (S - R)^{-1} \right] \\ &= R^{K-1} \left[ I_m - \mu\, (I_m - R)\, (S - R)^{-1} \right] \\ &= R^{K-1} \left[ S - \mu\, I_m - \overline{\mu}\, R \right] (S - R)^{-1} \\ &= R^{K-1} \left[ I_m - \mu\, S^{-1} - \overline{\mu}\, Q \right] (I_m - Q)^{-1} \\ &= \mu\, R^{K-1}\, e_m\, \alpha\, (I_m - Q)^{-1} \ , \qquad (4.4.3) \end{aligned}$$

and the result readily follows since the right handside of equation (4.4.3) is a positive matrix.

△

To show the *integer-concavity* of the throughput as a function of the buffer size, $Tr(K+1) - Tr(K)$ must be shown to decrease as $K$ increases. From (4.4.2), it follows that

$$Tr(K+1) - Tr(K) \tag{4.4.4}$$
$$= \frac{\alpha S\ [T(K+1) - T(K)]\ p + \alpha S\, T(K+1)\left[p e_m^T - e_m p^T\right]\, T(K)^T\, (\alpha S)^T}{(1 + \alpha S\, T(K)\, e_m)\,(1 + \alpha S\, T(K+1)\, e_m)} ,$$

Although the integer concavity of the throughput is intutively plauisible, this could only be shown for the case when both servers are geometric, owing to algebraic difficulties. In this simple case since $e_1 = 1$ and $p$ is a scalar, $p e_1^T = e_1 p^T$ and the second term in the numerator of (4.4.4) is thus zero. The equation takes the form

$$Tr(K+1) - Tr(K) = \frac{\alpha S\ [T(K+1) - T(K)]\ p}{(1 + \alpha S\, T(K)\, e_m)\,(1 + \alpha S\, T(K+1)\, e_m)} , \tag{4.4.5}$$

$$\frac{Tr(K+1) - Tr(K)}{Tr(K) - Tr(K-1)} = \frac{\alpha S\ [T(K+1) - T(K)]\ p}{\alpha S\ [T(K) - T(K-1)]\ p} \cdot \frac{1 + \alpha S\, T(K-1)\, e_m}{1 + \alpha S\, T(K+1)\, e_m} . \tag{4.4.6}$$

The second factor of (4.4.6) is always less than one owing to Lemma 4.4.1

**Lemma 4.4.2 :** *For a two node blocking system with geometric servers at each node, the system throughput is integer-concave in the intermediate buffer size.*

**Proof :** When both servers are geometric, with $\lambda$ being the rate of the second node server, $p = \lambda$, $Q = \overline{\lambda}$, $\alpha = 1$, $S = \frac{\mu}{\lambda\,\overline{\mu}}$, and $R = \frac{\overline{\lambda}\,\mu}{\lambda\,\overline{\mu}}$, and (4.4.6) reduces to

$$\frac{Tr(K+1) - Tr(K)}{Tr(K) - Tr(K-1)} = R \cdot \frac{1 + \frac{\mu}{\lambda\,\overline{\mu}} T(K-1)}{1 + \frac{\mu}{\lambda\,\overline{\mu}}\, T(K+1)} . \tag{4.4.7}$$

When $\mu \leq \lambda$, or equivalently $R \leq 1$, the right hand side of equation (4.4.7) is less than one owing to Lemma 4.4.1. For $R > 1$, the recursion (4.3.22), after some cancellations, readily yields

$$\frac{Tr(K+1) - Tr(K)}{Tr(K) - Tr(K-1)} = \frac{(-\mu/\overline{\mu}) + \frac{\mu}{\lambda \overline{\mu}} T(K)}{1 + \frac{\mu}{\lambda \overline{\mu}} + \frac{\mu}{\lambda \overline{\mu}} R\,T(K)} \quad . \tag{4.4.8}$$

whence

$$\frac{Tr(K+1) - Tr(K)}{Tr(K) - Tr(K-1)} < \frac{(-\mu/\overline{\mu}) + \frac{\mu}{\lambda \overline{\mu}} T(K)}{1 + \frac{\mu}{\lambda \overline{\mu}} + \frac{\mu}{\lambda \overline{\mu}} T(K)} < 1 \quad ,$$

and the throughput is thus integer-concave.

$\triangle$

## IV.5 Numerical examples :

To see the effects of buffer size and stochastic variability of the PH-type server on the system throughput, (4.4.2) is evaluated for different types of distributions. The reader is referred to (2.2.5)-(2.2.7) for the representations of these distributions. In all cases the first node server had geometric service distribution with $\mu = 0.5$. The following distributions and numerical values were used for the second node server:

i. Deterministic (5 units),

ii. Negative Binomial with three transient phases ($NB_3$), with $q_{12} = 0.75$, $q_{23} = 0.5$, and $q_{34} = 0.6$, the forth phase being the absorbing phase;

iii. Geometric with rate $\lambda = 0.2$; and

iv. Hypergeometric with three transient phases ($HG_3$), with $q_{14} = 0.25$, $q_{24} = 0.15$ and $q_{34} = 0.2$, the forth phase being again the absorbing phase. The initialization vector was $\alpha = (0.5, 0.3, 0.2)$.

The cases are numbered in increasing order with the variance of the service time distribution of the PH-type server. Expected service time of the PH-type server is the same in all cases and is equal to 5 units. The resuts are summarized in Table 4.5.1. For each PH-type server, the columns indicate that the throughput is integer-concave in buffer size, whereas the rows indicate that for fixed buffer size, throughput increases as the service time distribution of the second node server becomes less variable.

THROUGHPUT AS A FUNCTION OF BUFFER SIZE

| buffer size | deterministic | $NB_3$ | geometric | $HG_3$ |
|---|---|---|---|---|
| 1 | 0.142857 | 0.142857 | 0.142857 | 0.142857 |
| 2 | 0.197531 | 0.195804 | 0.187500 | 0.187093 |
| 3 | 0.199907 | 0.199714 | 0.196970 | 0.196733 |
| 4 | 0.199997 | 0.199980 | 0.199248 | 0.199152 |
| 5 | 0.200000 | 0.199999 | 0.199812 | 0.199778 |
| 7 | 0.200000 | 0.200000 | 0.199988 | 0.199985 |
| 10 | 0.200000 | 0.200000 | 0.200000 | 0.200000 |

expected service time of the

first (geometric) server =2

second (PH-type) server =5

**Table 4.5.1**

## IV.6 The continuous-time formulation :

In this section, the continous-time formulation of the model of Section 4.2 is presented. Theorem 3.6.1 holds in that the underlying continuous-time Markov chain is *always irreducible.* The service distribution of the PH-type server is again represented by the pair $(\alpha, Q)$, while the exponential server operates at rate $\mu$.

The states of the Markov chain are defined as in Section 3. The states being considered in the usual lexicographical order, the corresponding infinitesimal generator matrix T takes the form

$$T = \begin{pmatrix} -\mu & \mu\,\alpha & & & & & \\ p & Q-\mu\,I_m & \mu\,I_m & & & & \\ & p\,\alpha & Q-\mu\,I_m & \mu\,I_m & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & p\,\alpha & Q-\mu\,I_m & \mu\,I_m \\ & & & & & p\,\alpha & Q \end{pmatrix} \quad , \qquad (4.6.1)$$

and the equation $\pi\,T = 0_{Km+1}$ satisfied by the invariant vector can be rewritten as

$$-\mu\pi_0 + \pi_1\,p = 0 \qquad (4.6.2a)$$

$$\mu\,\pi_0\,\alpha + \pi_1\,(Q-\mu\,I_m) + \pi_2\,p\,\alpha = 0_m \qquad (4.6.2b)$$

$$\mu\,\pi_{k-1} + \pi_k\,(Q-\mu\,I_m) + \pi_{k+1}\,p\,\alpha = 0_m \quad , \quad 1 < k < K \ , \qquad (4.6.2c)$$

$$\mu\,\pi_{K-1} + \pi_K\,Q = 0_m \quad . \qquad (4.6.2d)$$

As in the discrete-time situation, postmultiplication of each equation in (4.6.2) by $e_m$, after simplifications, gives

$$\pi_1\,p - \mu\pi_0 = 0 \qquad (4.6.3a)$$

$$\pi_{k+1}\,p - \mu\,\pi_k\,e_m = 0 \quad , \quad 1 < i < K \quad . \qquad (4.6.3b)$$

Postmultiplication of (4.6.3) by $\alpha$ and use of (4.6.2b)–(4.6.2c) and of the boundary equations (4.6.2a) and (4.6.2d) give

$$\mu\pi_0\alpha = \pi_1\,(\mu\, I_m - \mu\, e_m\,\alpha - Q)\ , \tag{4.6.4a}$$

$$\mu\,\pi_{k-1} = \pi_k\,(\mu\, I_m - \mu\, e_m\,\alpha - Q)\ ,\ 1 < k < K\ , \tag{4.6.4b}$$

$$\mu\,\pi_{K-1} = -\pi_K\, Q\ . \tag{4.6.4c}$$

By following the same arguments that lead to the nonsingularity of the matrix $S$ in discrete-time model of Section 3, it can be shown that the matrix $(\mu\, I_m - \mu\, e_m\,\alpha - Q)$ is nonsingular. If the matrix $R$ is defined by

$$R = \mu\,(\mu\, I_m - \mu\, e_m\,\alpha - Q)^{-1}\ , \tag{4.6.5}$$

then (4.6.4) has a solution in *matrix geometric* for given by

$$\pi_1 = \pi_0\,\alpha\, R\ , \tag{4.6.6a}$$

$$\pi_k = \pi_1\, R^{k-1}\quad ,\ 1 \leq k < K\ , \tag{4.6.6b}$$

$$\pi_K = -\mu\,\pi_{K-1}\, Q^{-1}\quad , \tag{4.6.6c}$$

The scalar $\pi_0$ is now obtained through the standard normalization condition as

$$\pi_0 = (1 + \alpha\, T(K)\, e_m)^{-1}\quad , \tag{4.6.7}$$

where

$$T(K) = \sum_{i=1}^{K-1} R^i - \mu\, R^{K-1}\, Q^{-1}\quad . \tag{4.6.8}$$

Since by Lemma 2.3.11 the matrix $-Q$ is a nonsingular M-matrix, its inverse is a nonnegative matrix [10] while the matrix $R$ can be seen to be a nonnegative matrix through its Neuman expansion. Therefore, $\pi_0$ given in (4.6.7) is strictly positive.

The discussion is summarized in the following theorem.

**Theorem 4.6.1.** : *The unique invariant probability vector* $\pi = (\pi_0, \ldots, \pi_K)$ *is given by*

$$\pi_0 = (1 + \alpha\, T(K)\, e_m)^{-1} \quad , \tag{4.6.9a}$$

$$\pi_i = \pi_0\, \alpha\, R^i \quad , \quad 1 \leq i < K \quad , \tag{4.6.9b}$$

$$\pi_K = -\mu\, \pi_0\, \alpha\, R^{K-1}\, Q^{-1} \quad , \tag{4.6.9c}$$

*where* $R$ *and* $T(K)$ *are defined by equations (4.6.5) and (4.6.8), respectively.*

# CHAPTER V

# TWO NODE SYSTEM WITH PHASE TYPE SERVERS AND AN ARRIVAL PROCESS TO A FINITE BUFFER

## V.1. Introduction :

In this chapter, a two node system with PH-type servers at *both* nodes is again considered. This time, a *finite* capacity queue is allowed in front of the first node server in order to capture the situation of a *Bernoulli* arrival stream to the system. In Section 2, a system state process is defined and the corresponding one-step probability transition matrix is given. In Section 3, necessary and sufficient conditions for the irreducibility of the Markov chain are obtained in terms of the system parameters. Unlike for the model of Chapter III, the invariant probability vector is shown to *always* be *unique*, even in the discrete-time formulation, and is obtained in *matrix geometric* form by grouping the states in pairs. The *joint* queue length distribution is then easily obtained as entries of this invariant probability vector. A continuous-time formulation of the same model is discussed in Section 4, where the underlying Markov process is shown to be *irreducible*, and the unique invariant vector is obtained in matrix geometric form by using the solution technique of Section 3.

### V.2. The discrete-time model :

The model consists of two nodes separated by *finite* capacity buffers of sizes $K_1$ and $K_2$ in front of the first and second node servers, respectively, including the jobs being served by these servers. Each node is attended by a *single* PH-type server with *irreducible* representations $(\alpha, A)$ and $(\beta, B)$, at the first and second node, respectively. The row vectors $\alpha$ and $\beta$ and the matrices $A$ and $B$ are of dimensions $1 \times l$, $1 \times m$, $l \times l$ and $m \times m$, respectively. The corresponding $l \times 1$ and $m \times 1$ column vectors of absorption probabilities for the first and the second node server are denoted by $a$ and $b$, respectively. A *Bernoulli* arrival stream with parameter $\eta$ feeds into the first buffer under the assumption that arrivals which see a full buffer are *lost*. It is also assumed that the second server is never blocked, and that the *immediate* blocking strategy is enforced for the first server.

The state-space of the system is naturally defined to be the set $E$ given by

$$E = \begin{cases} (k_1, k_2, i, j), & 1 \le k_1 \le K_1, \quad 1 \le k_2 < K_2, \\ (k_1, 0, i), & 1 \le k_1 \le K_1, \quad k_2 = 0, \\ (k_1, K_2, j), & 1 \le k_1 \le K_1, \quad k_2 = K_2, \\ (0, k_2, j), & k_1 = 0, \quad 1 \le k_2 \le K_2, \\ (0, 0)\ , & k_1 = 0, \quad k_2 = 0, \end{cases}$$

for $1 \le i \le l$ and $1 \le j \le m$. Here, $k_1$ and $k_2$ represent the numbers of jobs in buffer one and two, respectively, while $i$ and $j$ represent the service phases at the first and second servers, respectively. Note that the phase of the first server is not defined when it has no jobs to process or when it is blocked, and that the phase of the second server is not defined when the second buffer is empty.

For the sake of compactness, the notation

$$r := [(K_2 - 1)lm + l + m], \qquad s := K_2 m + 1,$$

is adopted in this chapter.

The invariant probability vector of these states is denoted by the $1 \times (K_1 r + s)$ row vector $\pi$, which is partitioned into $K_1 + 1$ blocks of components, say $\pi = (\pi_0, \pi_1, \ldots, \pi_{K_1})$, with $\pi_0$ being $1 \times s$ and $\pi_{k_1}$, $0 < k_1 \leq K_1$, being $1 \times r$ row vectors, respectively. Each entry $\pi_{k_1}$, $0 \leq k_1 \leq K_1$, is of the form $\pi_{k_1} = (\pi_{k_1 0}, \pi_{k_1 1}, \ldots, \pi_{k_1 K_2})$ where the vectors $\pi_{k_1 k_2}$ is of dimension

$$\begin{cases} lm\,, & 1 \leq k_1 \leq K_1, \ \ 1 \leq k_2 < K_2, \\ l\,, & 1 \leq k_1 \leq K_1, \ \ k_2 = 0, \\ m\,, & k_1 = 0, \ \ 1 \leq k_2 \leq K_2, \text{ or} \\ & 1 \leq k_1 \leq K_1, \ \ k_2 = K_2, \\ 1\,, & k_1 = 0, \ \ k_2 = 0, \end{cases}$$

and the entries of each $\pi_{k_1 k_2}$, $1 \leq k_1 \leq K_1$, $1 \leq k_2 < K_2$, are ordered as $(1,1), (2,1), \ldots, (l,1), (1,2) \ldots, (i,j), \ldots, (l-1,m), (l,m)$.

By lexicographically ordering the states as described above, the one-step transition matrix $P$ of the underlying Markov chain can be put in the form

$$P = \begin{pmatrix} C & D & & & & & \\ E & G & H & & & & \\ & F & G & H & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & F & G & H \\ & & & & & F & L \end{pmatrix}, \qquad (5.2.1)$$

where the block entries $C$, $D$, and $E$ are of dimensions $s \times s$, $s \times r$ and $r \times s$, respectively, and the matrices $F$, $G$, $H$ and $L$ are of dimensions $r \times r$. These matrices are given by

$$C = \overline{\eta} \begin{pmatrix} 1 & & & & & & \\ b & B & & & & & \\ & b\beta & B & & & & \\ & & \cdot & \cdot & & & \\ & & & \cdot & \cdot & & \\ & & & & b\beta & B & \\ & & & & & b\beta & B \end{pmatrix} ,$$

$$D = \eta \begin{pmatrix} \alpha & & & & & & \\ b \otimes \alpha & B \otimes \alpha & & & & & \\ & b\beta \otimes \alpha & B \otimes \alpha & & & & \\ & & \cdot & \cdot & & & \\ & & & \cdot & \cdot & & \\ & & & & b\beta \otimes \alpha & B \otimes \alpha & \\ & & & & & b\beta \otimes \alpha & B \end{pmatrix} ,$$

$$E = \overline{\eta} \begin{pmatrix} 0 & \beta \otimes a & & & & & \\ & b\beta \otimes a & B \otimes a & & & & \\ & & \cdot & \cdot & & & \\ & & & \cdot & \cdot & & \\ & & & & b\beta \otimes a & B \otimes a & \\ & & & & & b\beta \otimes a & B \otimes a \\ & & & & & & 0 \end{pmatrix} ,$$

$$F = \overline{\eta} \begin{pmatrix} 0 & \beta \otimes a\alpha & & & & & \\ & b\beta \otimes a\alpha & B \otimes a\alpha & & & & \\ & & \cdot & \cdot & & & \\ & & & \cdot & \cdot & & \\ & & & & b\beta \otimes a\alpha & B \otimes a\alpha & \\ & & & & & b\beta \otimes a\alpha & B \otimes a \\ & & & & & 0 & 0 \end{pmatrix} ,$$

$$G = \begin{pmatrix} \overline{\eta}A & \eta(\beta \otimes a\alpha) & & & & \\ \overline{\eta}(b \otimes A) & G_d & \eta(B \otimes a\alpha) & & & \\ & \overline{\eta}(b\beta \otimes A) & G_d & \eta(B \otimes a\alpha) & & \\ & & \cdot & \cdot & \cdot & \\ & & & \overline{\eta}(b\beta \otimes A) & G_d & \eta(B \otimes a) \\ & & & & \overline{\eta}(b\beta \otimes \alpha) & \overline{\eta}B \end{pmatrix} ,$$

$$H = \eta \begin{pmatrix} A & & & & & & \\ b \otimes A & B \otimes A & & & & & \\ & b\beta \otimes A & B \otimes A & & & & \\ & & \cdot & \cdot & & & \\ & & & \cdot & \cdot & & \\ & & & & b\beta \otimes A & B \otimes A & \\ & & & & & b\beta \otimes \alpha & B \end{pmatrix} ,$$

$$L = \begin{pmatrix} A & \eta(\beta \otimes a\alpha) & & & & \\ b \otimes A & L_d & \eta(B \otimes a\alpha) & & & \\ & b\beta \otimes A & L_d & \eta(B \otimes a\alpha) & & \\ & & \cdot & \cdot & \cdot & \\ & & & b\beta \otimes A & L_d & \eta(B \otimes a) \\ & & & & b\beta \otimes \alpha & B \end{pmatrix} ,$$

with the diagonal entries $G_d$ and $L_d$ being defined by

$$G_d = \overline{\eta}\,(B \otimes A) + \eta\,(b\beta \otimes a\alpha) \quad ,$$

$$L_d = B \otimes A + \eta\,(b\beta \otimes a\alpha) \ .$$

## V.3. Analysis of the discrete-time model :

A necessary and sufficient condition for the *irreducibility* of the Markov chain associated with the matrix $P$ of one-step transition probabilities is now studied and the *uniqueness* of the corresponding invariant probability vector is established.

If the sets $E_{k_1}$, $0 \leq k_1 \leq K_1$, are defined by

$$E_{k_1} := \left\{ e \in E : \; e = \left\{ \begin{array}{ll} (k_1, k_2, i, j), & 1 \leq k_2 \leq K_2 \\ (k_1, 0, i), & k_2 = 0 \\ (k_1, K_2, j), & k_2 = K_2 \end{array} \right. \right\}, 1 \leq k_1 \leq K_1,$$

$$E_0 := \{ e \in E : \; e = (0, k_2, j), \;\; 1 \leq k_2 \leq K_2 \} \cup \{(0,0)\} ,$$

then the sets $E_{k_1}$, $0 \leq k_1 \leq K_1$, form a partition of the state space $E$ for the Markov chain. Since $0 < \eta < 1$, the following three observations are easily verified:

(i) Lemma 2.3.3 quickly implies that the directed graphs of the matrices $G$, $L$ and $T$ have the same topological structure, with $T$ given by (3.3.1).

(ii) The set $E_{k_1 - 1}$ is reachable from *every* state in the set $E_{k_1}$ and vice versa, for $1 \leq k_1 \leq K_1$.

(iii) The state $(0, 0)$ is reachable from every state in the set $E_0$.

The next theorem readily follows from the observations (i)-(iii) by virtue of Theorem 3.4.5.

**Theorem 5.3.1.** : *If the matrix $T$ defined in Chapter III is irreducible, then so is the matrix $P$.*

Next, even when the matrix $P$ is not irreducible, it is now shown to have a *single* ergodic class, whence the invariant probability distribution vector is always unique. Let the $K_1 r \times K_1 r$ matrix $\widetilde{P}$ be obtained by deleting the first $s$ rows and columns of $P$, i. e., $\widetilde{P}$ is the submatrix of $P$ that governs the transition mechanism within the states in the set $E \setminus E_0$. Assume the directed graph $G(\widetilde{P})$ to induce several ergodic classes and a set of transient states, as would be the case if $T$ had several ergodic classes. In view of (ii) the set $E_0$ is reachable from all of the ergodic classes of $E \setminus E_0$, whence by (iii), the state $(0,0)$ is reachable from every ergodic class of the set $E \setminus E_0$. On the other hand, if $(0,0) \to e$ for a state $e$ in $E$, then $e$ will be in the communication *link* between the state $(0,0)$ and the other states of the chain that $(0,0)$ has access to, while if $(0,0) \not\to e$, then $e$ will be a transient state of the chain since $e \to (0,0)$ from the above argument. Therefore, in the directed graph $G(P)$, the set $E_{(0,0)} := \{e \in E : (0,0) \to e\} \cup \{(0,0)\}$ forms an irreducible class while the set $E \setminus E_{(0,0)}$ forms a transient class.

The next theorem follows from Theorem 2.3.14 and these remarks.

**Theorem 5.3.2.** : *The Markov chain with one-step transition matrix $P$ always has a* **single** *ergodic class, and the invariant probability vector $\pi$ as defined above is thus unique.*

In order to obtain a closed form expression for this unique vector $\pi$, it is convenient to group the vectors $\pi_{k_1}$, $0 \le k_1 \le K_1$, in pairs. With this in mind, the equation $\pi P = \pi$ can then be rewritten as

$$\pi_0 (I_s - C) = \pi_1 E \tag{5.3.1$a$}$$

$$\pi_0 D + \pi_1 G + \pi_2 F = \pi_1 \tag{5.3.1$b$}$$

$$(\pi_{k_1 - 1}, \pi_{k_1}) \begin{pmatrix} H & 0_{r\times r} \\ G & I_r \end{pmatrix} = (\pi_{k_1}, \pi_{k_1 + 1}) \begin{pmatrix} I_r & I_r \\ -F & 0_{r\times r} \end{pmatrix}, \quad 1 < k_1 < K_1 \tag{5.3.1$c$}$$

$$\pi_{K_1 - 1} H = \pi_{K_1} (I_r - L) \ . \tag{5.3.1$d$}$$

In order to get this explicit solution, nonsingularity of the matrices $(I_s - C)$

and $\begin{pmatrix} H & 0_{r\times r} \\ G & I_r \end{pmatrix}$ is needed. Since

$$I_s - C = \begin{pmatrix} \eta & & & & & \\ -\overline{\eta}b & I_m - \overline{\eta}B & & & & \\ & -\overline{\eta}b\beta & I_m - \overline{\eta}B & & & \\ & & \cdot & \cdot & & \\ & & & \cdot & \cdot & \\ & & & & -\overline{\eta}b\beta & I_m - \overline{\eta}B \end{pmatrix}$$

and $\eta > 0$, $I_s - C$ is invertible if and only if the matrix $I_m - \overline{\eta}B$ is invertible, a fact easily established by a direct application of Theorem 2.3.20. This can also be seen by noting that $\rho(B) < 1$ and hence 0 is not an eigenvalue of the matrix $I_m - \overline{\eta}B$. On the other hand, since

$$det \begin{pmatrix} H & 0_{r\times r} \\ G & I_r \end{pmatrix} = det(H) = \eta^r \; det(A)\, det(B)\, [det(B \otimes A)]^{K_2 - 1}$$

$$= \eta^r \; det(A)^{(K_2-1)m+1} \; det(B)^{(K_2-1)l+1} \; ,$$

where the last equality follows from (2.3.2d), it is easy to see that the matrix $\begin{pmatrix} H & 0_{r\times r} \\ G & I_r \end{pmatrix}$ is invertible if and only if *both* $A$ and $B$ are invertible.

As a result of this discussion, nonsingularity of both $A$ and $B$ will be assumed for the discrete time model. Although this may seem rather restrictive, many well-known discrete PH-type distributions enjoy this property, including the hyperexponential and negative binomial distributions, to name a few.

**Theorem 5.3.3.** : *If the matrices $A$ and $B$ are invertible, then the unique invariant probability vector $\pi = (\pi_0, \pi_1, \ldots, \pi_{K_1})$ has the following form*

$$\pi_0 = \pi_1 \, E \, (I_s - C)^{-1} \; , \tag{5.3.2}$$

$$\pi_{k_1} = \pi_{K_1} \; \left((I_r - L)H^{-1}, I_r\right) \; R^{K_1 - k_1 - 1} \; \begin{pmatrix} I_r \\ 0_{r\times r} \end{pmatrix} \; , \quad 1 \le k_1 < K_1 \tag{5.3.3}$$

$$\pi_{K_1} = (xw)^{-1} \; x \; , \tag{5.3.4}$$

*where* $1 \times r$ *vector* $x$ *satisfies*

$$x\,Z = 0_r \quad , \quad x > 0_r^T \quad (\textit{or } x < 0_r^T) \tag{5.3.5}$$

*with the* $2r \times 2r$ *and* $r \times r$ *matrices* $R$ *and* $Z$ *and the* $r \times 1$ *vector* $w$ *being given by*

$$R = \begin{pmatrix} I_r & I_r \\ -F & 0_{r\times r} \end{pmatrix} \begin{pmatrix} H & 0_{r\times r} \\ G & I_r \end{pmatrix}^{-1} \quad ,$$

$$Z = \left((I_r - L)\,H^{-1}, I_r\right)\; R^{K_1-2} \begin{pmatrix} E(I_s - C)^{-1}D + G - I_r \\ F \end{pmatrix} \quad ,$$

$$w = \left((I_r - L)H^{-1}, I_r\right)\left[R^{K_1-2}\begin{pmatrix} I_r \\ 0_{r\times r} \end{pmatrix} E(I_s - C)^{-1}e_s + \sum_{k_1=1}^{K_1-1} R^{K_1-k_1-1}\begin{pmatrix} e_r \\ 0_r^T \end{pmatrix}\right] + e_r .$$

**Proof :** Since $I_s - C$ is invertible, (5.3.2) follows from (5.3.1a). On the other hand, (5.3.1c) takes the form

$$(\pi_{k_1-1},\ \pi_{k_1}) = (\pi_{k_1},\ \pi_{k_1+1})\,R \quad , \quad 1 < k_1 < K_1 \quad ,$$

or equivalently, by induction,

$$(\pi_{k_1-1},\ \pi_{k_1}) = (\pi_{K_1-1},\ \pi_{K_1})\,R^{K_1-k_1} \quad , \quad 1 < k_1 \leq K_1 \quad . \tag{5.3.6}$$

From (5.3.1d), it follows that

$$(\pi_{K_1-1},\ \pi_{K_1}) = \pi_{K_1}\ \left((I_r - L)\,H^{-1}, I_r\right) \quad , \tag{5.3.7}$$

and therefore (5.3.3) readily obtains. Equations (5.3.4) and (5.3.5) are obtained by using (5.3.2), (5.3.6) and (5.3.7) in (5.3.1b) and the standard normalization condition $\pi_0 e_s + \sum_{k_1=1}^{K_1} \pi_{k_1}\, e_r = 1$.

△

### V.4. The continuous-time formulation :

In this section, a *complete* solution, without any additional assumptions on the service distributions, is presented for the continuous-time formulation. The notation and state description of Section 2 are again used. The arrival process is now assumed to be *Poisson* with rate $\eta$. The infinitesimal generator matrix $P$ is again of the form

$$P = \begin{pmatrix} C & D & & & & & \\ E & G & H & & & & \\ & F & G & H & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & F & G & H \\ & & & & & F & L \end{pmatrix}, \qquad (5.4.1)$$

where the block entries $C$, $D$, and $E$ have dimensions $s \times s$, $s \times r$, and $r \times s$, respectively, while the matrices $F$, $G$, $H$ and $L$ are of dimension $r \times r$. These matrices are now given by

$$C = \begin{pmatrix} -\eta & & & & & & \\ b & B-\eta I_m & & & & & \\ & b\beta & B-\eta I_m & & & & \\ & & \cdot & \cdot & & & \\ & & & \cdot & \cdot & & \\ & & & & b\beta & B-\eta I_m & \\ & & & & & b\beta & B-\eta I_m \end{pmatrix},$$

$$D = \eta \begin{pmatrix} \alpha & & & & \\ & I_m \otimes \alpha & & & \\ & & \ddots & & \\ & & & I_m \otimes \alpha & \\ & & & & I_m \end{pmatrix},$$

$$E = \begin{pmatrix} 0 & \beta \otimes a & & & & \\ & 0 & I_m \otimes a & & & \\ & & \ddots & \ddots & & \\ & & & 0 & I_m \otimes a & \\ & & & & 0 & I_m \otimes a \\ & & & & & 0 \end{pmatrix},$$

$$F = \begin{pmatrix} 0 & \beta \otimes a\alpha & & & & \\ & 0 & I_m \otimes a\alpha & & & \\ & & \ddots & \ddots & & \\ & & & 0 & I_m \otimes a\alpha & \\ & & & & 0 & I_m \otimes a \\ & & & & & 0 \end{pmatrix},$$

$$G = \begin{pmatrix} A - \eta I_l & & & & & \\ b \otimes I_l & G_d & & & & \\ & b\beta \otimes I_l & G_d & & & \\ & & \ddots & \ddots & & \\ & & & b\beta \otimes I_l & G_d & \\ & & & & b\beta \otimes \alpha & B - \eta I_m \end{pmatrix},$$

$$L = \begin{pmatrix} A & & & & & \\ b \otimes I_l & L_d & & & & \\ & b\beta \otimes I_l & L_d & & & \\ & & \ddots & \ddots & & \\ & & & b\beta \otimes I_l & L_d & \\ & & & & b\beta \otimes \alpha & B \end{pmatrix},$$

and $H = \eta\ I_r$. The diagonal entries of the matrices $G$ and $L$ are given by $G_d = B \otimes I_l + I_m \otimes A - \eta(I_m \otimes I_l)$ and $L_d = B \otimes I_l + I_m \otimes A$, respectively.

The observations (i)–(iii) made for the discrete-time model obviously still apply to the continuous-time model, and the conclusions of Theorem 5.3.1 thus combine with Theorem 3.6.1 to give the following result.

**Theorem 5.4.1.** : *The Markov process with infinitesimal generator matrix $P$ is* **irreducible,** *and the invariant probability vector $\pi$ as defined above is thus unique.*

In order to obtain an expression for the unique vector $\pi$, the vectors $\pi_{k_1}$, $0 \leq k_1 \leq K_1$, are again grouped in pairs. The equation $\pi P = 0_{K_1 r+s}$ can then be rewritten in the form

$$\pi_0 \, C = -\pi_1 \, E \tag{5.4.2a}$$

$$\pi_0 \, D + \pi_1 \, G + \pi_2 \, F = 0_r \tag{5.4.2b}$$

$$\eta \; (\pi_{k_1-1}, \pi_{k_1}) = (\pi_{k_1}, \pi_{k_1+1}) \begin{pmatrix} -G & \eta I_r \\ -F & 0_{r\times r} \end{pmatrix} \; , \quad 1 < k_1 < K_1 \; , \tag{5.4.3c}$$

$$\eta \pi_{K_1-1} = -\pi_{K_1} \; L \; . \tag{5.4.2d}$$

In this case the matrix $H$ is the *identity* matrix, only the nonsingularity of C is needed in order to get an explicit solution. The nonsingularity of $C$ again follows from the Gershgorin Circle Theorem given in Theorem 3.2.20. Note that no invertibility assumption on the matrices $A$ or $B$ is needed for the continuous-time model. Proceeding as in the proof of Theorem 5.3.3, the following similar result can be obtained.

**Theorem 5.4.2.** : *The unique steady state probability vector $\pi = (\pi_0, \pi_1, \ldots, \pi_{K_1})$ is given by the equations*

$$\pi_0 = -\pi_1 \, E \, C^{-1} \; , \tag{5.4.3}$$

$$\pi_{k_1} = \pi_{K_1} \; \left(-\eta^{-1} \, L, I_r\right) \; R^{K_1-k_1-1} \; \begin{pmatrix} I_r \\ 0_{r\times r} \end{pmatrix} \; , \quad 1 \leq k_1 < K_1 \tag{5.4.4}$$

$$\pi_{K_1} = (xw)^{-1} x \; , \tag{5.4.5}$$

*where* $1 \times r$ *vector* $x$ *satisfies*

$$x\,Z = 0_r \quad , \qquad x >> 0_r^T \ (\textit{or } x << 0_r^T) \tag{5.4.6}$$

*and the* $2r \times 2r$ *and* $r \times r$ *matrices* $R$ *and* $Z$ *and the* $r \times 1$ *vector* $w$ *being given by*

$$R = \begin{pmatrix} -\eta^{-1}\,G & I_r \\ -\eta^{-1}\,F & 0_{r\times r} \end{pmatrix} \ ,$$

$$Z = \left(-\eta^{-1}\,L, I_r\right)\ R^{K_1-2}\ \begin{pmatrix} G - E\,C^{-1}\,D \\ F \end{pmatrix} \ ,$$

$$w = \left(-\eta^{-1}L, I_r\right)\ \left[-R^{K_1-2}\ \begin{pmatrix} I_r \\ 0_{r\times r} \end{pmatrix}\ E\,C^{-1}\,e_s + \sum_{k_1=1}^{K_1-1} R^{K_1-k_1-1}\ \begin{pmatrix} e_r \\ 0_r^T \end{pmatrix}\right] + e_r.$$

# CHAPTER VI

# APPLICATIONS TO SERVERS SUBJECT TO FAILURES

## VI.1. Introduction :

In this chapter, a class of unreliable servers with *PH-type* service and repair time distributions is introduced. In Section 2, the *effective* service time distribution of such servers is shown to admit a PH-type represention of higher order so that the methods of the previous chapters apply. Under the irreducibility assumption on the service and repair PH-distributions, necessary and sufficient conditions for the irreducibility of the effective service representation are established. In Section 3, the situation where even idling servers may be subject to failure is considered: The transition probabilities among the boundary states are explicitly written out for this case and an explicit solution is then obtained in matrix geometric form for the invariant probabilities through algebraic manupulations similar to the ones given in the previous chapters. The case when only operational servers can fail is obtained as a special case of the discussion given for the case when idling servers can fail. Applications to other models are only briefly mentioned. Finally to illustrate the ideas, several numerical examples are considered in Section 4.

## VI.2. Representation of The Effective Service Time Distribution :

Consider the following model for a PH-type server subject to occasional failures: The service and repair distributions have irreducible PH-representations $(\alpha, A)$ and $(\beta, B)$ of order $m$ and $n$, respectively, with the corresponding $m \times 1$ and $n \times 1$ column vectors of absorption probabilities denoted by $a$ and $b$, respectively. As soon as a failure occurs, the repair process starts and proceeds according to a PH-type repair distribution. Let the sets $S$, $R$ and $E$ be defined by

$$S := \{s_i,\ 1 \leq i \leq m\}, \quad R := \{r_j,\ 1 \leq j \leq n\}, \quad E := S \cup R \ ,$$

where $s_i$ and $r_j$ are the $i^{th}$ service phase and the $j^{th}$ repair phase, respectively. Let $C$ and $D$ be $m \times n$ and $n \times m$ non-negative matrices with entries $C_{ij}$ and $D_{ij}$, $1 \leq i \leq m$, $1 \leq j \leq n$, respectively, with the property that $Ce_n = e_m$ and $De_m = e_n$. Similarly, let the $1 \times m$ column vector $f$ have non-negative entries $f_i$, $1 \leq i \leq m$.

It is assumed that when the server is up and in phase of *service* $i$, it can fail with probability $f_i$, $1 \leq i \leq m$, and the *repair* is initialized at phase $j$ with probability $C_{ij}$, $1 \leq i \leq m$, $1 \leq j \leq n$. Similarly, when the repair is completed, i.e., whenever there is a transition from a transient phase $j$ to the absorbing phase $n+1$ of the repair distribution, the service restarts at a phase $i$ with probability $D_{ji}$, $1 \leq j \leq n$, $1 \leq i \leq m$. Moreover, a failure at a repair completion epoch is not allowed. In addition to these fairly general transition mechanisms, the server is allowed to fail at a service completion epoch with probability $\phi$ and reinitialization of the repair phase is then done according to the probability distribution vector $\beta$. The natural conditions $\phi < 1$ and $f_i < 1$, $1 \leq i \leq m$, are imposed throughout.

Under the assumptions made, it is easy to see that the *effective* service time distribution of such a server admits a PH-type distribution. To that end, consider a discrete-time Markov chain on the state space $E \cup \{s_{m+1}\}$, where $s_{m+1}$ is interpreted as the instantenous state for the PH-representation of the (effective) service. The corresponding one-step transition matrix $P$ and the initialization probability vector $q$ are of the form

$$P = \begin{pmatrix} Q & p \\ 0_{m+n} & 1 \end{pmatrix} \ , \ (q, q_{m+n+1}) \ ,$$

$$q = (q_1, q_2, \ldots, q_{m+n}) \ ,$$

where $Q$, $p$ and $q$ are of dimension $(m+n) \times (m+n), (m+n) \times 1$ and $1 \times (m+n)$, respectively. The equalities

$$Q = \begin{pmatrix} \Lambda_{\overline{f}} \, A & \Lambda_f \, C \\ \\ \Lambda_b \, D & B \end{pmatrix} \ , \tag{6.2.1$a$}$$

$$p = \begin{pmatrix} \Lambda_{\overline{f}} \, a \\ 0_n^T \end{pmatrix} \ , \tag{6.2.1$b$}$$

$$q = (\overline{\phi} \ \alpha \ , \phi \, \beta) \ , \tag{6.2.1$c$}$$

readily follows, where the notation $\Lambda_x := diag(x_1, \ldots, x_l)$ is used for all $x$ in $I\!R^l$. The relations $\Lambda_f \, e_m = f$ and $\Lambda_b \, e_n = b$ easily follows.

The effective service distribution of such a server thus has a PH-representation $(q, Q)$ of order $n + m$. The assumption $q_{m+1} = \beta_{n+1} = 0$ imply that $q_{m+n+1} = 0$.

**Lemma 6.2.1.** : *The PH-representation $(q, Q)$ is irreducible if and only if every state $r_j$ in the set $R \backslash R^*$ is reachable from the set $R^*$ under the transition mechanism induced by $G(B)$, where*

$$R^* := \{ r_j \in R \ : r_j \ \textit{is reachable from the set } S \textit{ under } G(Q + pq) \ \} \ .$$

*Moreover, if a state $r_j$ in $R \backslash R^*$ is* not *reachable from $R^*$, $r_j$ is a* **transient phase** *for the PH-representation $(q, Q)$.*

**Proof :** By definition, the irreducibility of the PH-representation $(q, Q)$ is equivalent to the irreducibilty of the $(m+n) \times (m+n)$ matrix $Q + pq$, which is readily given by

$$Q + pq = \begin{pmatrix} \Lambda_{\overline{f}} A + \overline{\phi}\, \Lambda_{\overline{f}}\, a\, \alpha & \Lambda_f\, C + \phi\, \Lambda_{\overline{f}}\, a\, \beta \\ \Lambda_b\, D & B \end{pmatrix}.$$

Now, the property $s_i \rightarrow s_{i'}$, $1 \leq i, i' \leq m$, is a direct consequence of the irreduciblity of the matrix $\Lambda_{\overline{f}} A + \overline{\phi}\, \Lambda_{\overline{f}}\, a\, \alpha$, a fact which follows from Lemma 2.2.3 since by assumption $\phi < 1$, $f_i < 1$, $1 \leq i \leq m$, and the representation $(\alpha, A)$ is irreducible. Therefore, every state $s$ in $S$ communicates with any other state in $S$ without leaving $S$. On the other hand, since $(I_n - B)$ is assumed invertible, $B$ is substochastic and $b \neq 0_n^T$. This fact, together with the fact that the states $\{r_1, \ldots, r_n\}$ are all transient states for the representation $(\beta, B)$ yields the access relation $r_j \rightarrow s_i$ for *every* $r_j$ in $R$ and $s_i$ in $S$.

(Sufficiency) The sufficiency part of the first assertion of the Lemma now follows from the hypothesis, since every $r_j$ in $R \setminus R^*$ is reachable from $R^*$, thus they are communicating with the set $S \cup R^*$.

(Necessity) Follows trivally by the definition of irreducibility and the form of the matrix $Q + pq$ in that the transitions within the set $R$ are governed by the matrix $B$ *only*, due to the assumption that failures at repair completions are not allowed.

The second part of Lemma 6.2.1 also follows easily from the above discussion. Since $b \neq 0_n^T$, the underlying Markov chain will eventually leave the set $R$ and enter the set $S$. If $r_j$ in $R \setminus R^*$ is not reachable from the set $R^*$ then $r_j$ will never be revisited and will be a transient state for the Markov chain.

$\triangle$

Since, $(q, Q)$ is constructed from basic building blocks and is not given from the onset, no invertibility assumptions are made for the matrix $I_{m+n} - Q$. However, a sufficient condition is obtained in the following Lemma.

**Lemma 6.2.2.** : *If the PH-representation $(q, Q)$ is irreducible then the matrix $I_{m+n} - Q$ is invertible.*

**Proof :** The directed graph $G(Q)$ may in general induce several ergodic classes. However the irreducibility of the representation $(q, Q)$ implies that all of these classes communicate with each other through $s_{m+1}$ in the access relation induced by $G(Q+pq)$, and therefore, the states in $E$ are all transient for the Markov chain with one-step probability transition matrix $P$. On the other hand, if $G(Q)$ induces a single ergodic class, the states in $E$ are again all transient since $s_{m+1}$ is indeed an absorbing state for $P$, i. e., $p \neq 0_{m+n}^T$ from (6.2.1b), owing to the assumption that $a \neq 0_m^T$ and $f_i < 1,\ 1 \leq i \leq m$. The invertibility of $(I_{m+n} - Q)$ is now immediate since it is equivalent to the statement that the states in $E$ are transient for the Markov chain with one-step probability transition matrix $P$ [Neuts, 58 p. 45].

△

The representation of the effective service time given above for such a faliure type server subsumes the case when the server is *reliable*, i. e., $\phi = 0,\ f = 0_m^T$, in which case, the matrix $Q + pq$ takes the form $\begin{pmatrix} A + a\alpha & 0_{m\times n} \\ \Lambda_b D & B \end{pmatrix}$. This special case also provides a trivial example for the representation $(q, Q)$ *not* to be irreducible, although the representation $(\alpha, A)$ for the service duration is. If the representation $(q, Q)$ is not irreducible, Lemma 6.2.1 may provide a general guideline in identifying the corresponding irreducible representation. In the following sections it is assumed that the representation $(q, Q)$ is *irreducible* whence the matrix $(I_{m+n} - Q)$ is invertible.

In some applications the server may fail with a positive probability at a service completion epoch, i. e., $\phi > 0$. In that event, since $a \neq 0_m^T$, there exists some $i$, $1 \leq i \leq m$, such that $a_i > 0$, whence $s_i \longrightarrow r_j$ for every $r_j$ in the set $R_* := \{1 \leq j \leq n : \beta_j > 0\}$ as the repair phase is initialized according to $\beta$. Since the representation $(\beta, B)$ is irreducible, every state in $R \setminus R_*$ must be reachable from the set $R_*$ since otherwise $r$ will be a transient state for the matrix $B + b\beta$ thus contradicting the irreducibility of $(\beta, B)$. Therefore, similar arguments that lead to the sufficiency part of Lemma 6.2.1 show that the representation $(q, Q)$ is irreducible.

To conclude this section, it is noteworthy to keep in mind that the matrices $C$ and $D$, corresponding to transition probabilities between the service and repair phases, take special forms depending on the assumptions made. For instance, if upon completion of a repair, the phase of service is reinitialized according to $\alpha$, and similarly if upon a failure the phase of repair is initiazed according to $\beta$, then $C$ and $D$ take the special forms $C = e_m \beta$ and $D = e_n \alpha$. In the case when $C = e_m \beta$, regardless of the form of the matrix $D$, it is an easy exercise to see that the representation $(q, Q)$ given by (6.2.1) is irreducible.

### VI.3. Two Node System With PH-Type Failure Servers :

In this section model of Chapter III is considered in the event that the servers are subject to failures even when they are idling. The first and second node servers have irreducible PH-representations of the form (6.2.1), denoted by $(q_1, Q_1)$ and $(q_2, Q_2)$, respectively. The service and repair PH-representations are denoted by the pairs $(\alpha_i, A_i)$ and $(\beta_i, B_i)$, $i = 1, 2$. It is assumed that an idling server fails with probability $g_i$, $i = 1, 2$, and upon failure the phase of repair is initialized according to the initialization vector $\beta_i$. The row vectors $\alpha_i$ and $\beta_i$ and the matrices $A_i$ and $B_i$ are of dimension $1 \times m_i$, $1 \times n_i$, $m_i \times m_i$ and $n_i \times n_i$, respectively, for $i = 1, 2$. The corresponding $m_i \times 1$ and $n_i \times 1$ column vectors of absorption probabilities for the service and the repair distributions are denoted by $a_i$ and $b_i$, respectively, for $i = 1, 2$. For sake of compactness, the following notation is used hereafter.

$$\begin{aligned} h &= (m_1 + n_1)(m_2 + n_2) \ , & r_1 &= m_1 + n_1 \ , \\ h_1 &= (m_1 + n_1)(1 + n_2) \ , & r_2 &= m_2 + n_2 \ , \\ h_2 &= (1 + n_1)(m_2 + n_2) \ , & l &= n_2 + 1 \ . \end{aligned}$$

As in Chapter III it is assumed that the first server never starves while the second is never blocked, and the *immediate* blocking strategy is adopted. The states of the system are denoted by

$$\begin{cases} (i, 0) \ , & k = 0, \ 1 \le i \le r_1, \\ (i, 0, j) \ , & k = 0, \ 1 \le i \le r_1, \ m_2 + 1 \le j \le r_2, \\ (i, k, j) \ , & 1 \le i \le r_1, \ 0 < k < K, \text{ and } 1 \le j \le r_2, \\ (i, K, j) \ , & k = K, \ m_1 + 1 \le i \le r_1, \ 1 \le j \le r_2, \\ (K, j) \ , & k = K, \ 1 \le j \le r_2. \end{cases}$$

Here, $k$ indicates the buffer size, while $i$ and $j$ represent the service or repair phase in the first and the second node server, respectively. The phase of service of the second node server is not defined when it has no jobs to process and the phase of service of the first node server is not defined when the buffer is full as it is blocked. The pairs $(i,0)$, $1 \le i \le r_1$, and $(K,j)$, $1 \le j \le r_2$, correspond to states where the second and the first node server, respectively, are functional but idle.

The invariant probability vector of these states is denoted by the $1 \times [(K-1)\,h + h_1 + h_2]$ row vector $\pi$. As in Chapter III, this vector is partitioned into $K+1$ blocks of components, say $\pi = (\pi_0, \pi_1, \ldots, \pi_K)$, with $\pi_0$ being $1 \times h_1$, $\pi_i$, $0 < i < K$, being $1 \times h$, and $\pi_K$ being $1 \times h_2$, row vectors.

By ordering the states as in Chapter III, i. e., first varying the index of the first server, the one-step state transition matrix $T$ of the underlying Markov chain can be obtained in block tridiagonal form. Since the effective service time distribution of the failure service is still PH-type, the *intermediate* block entries of the matrix $T$ is still of the form given in Chapter III, i. e., if $T_{k_1,k_2}$ denote the $(k_1,k_2)$-th block of the matrix $T$, $0 \le k_1, k_2 \le K$, then

$$T_{k,k-1} = p_2\, \alpha_2 \otimes Q_1 \qquad\qquad 1 < k < K \;,$$

$$T_{k,k} = Q_2 \otimes Q_1 + p_2\, \alpha_2 \otimes p_1\, \alpha_1 \qquad\qquad 1 \le k < K \;,$$

$$T_{k,k+1} = Q_2 \otimes p_1\, \alpha_1 \qquad\qquad 1 \le k < K-1 \;,$$

where each block is an $h \times h$ matrix. For the *boundary* states the entries are given by

$$T_{00} = \begin{pmatrix} \overline{g}_2 & g_2\, \beta_2 \\ b_2 & B_2 \end{pmatrix} \otimes Q_1 \qquad\qquad h_1 \times h_1 \text{ matrix,}$$

$$T_{10} = p_2\; (\overline{\phi}_2 \,, \phi_2\, \beta_2) \otimes Q_1 \qquad\qquad h \times h_1 \text{ matrix,}$$

$$T_{01} = \begin{pmatrix} \overline{\phi}_2\, \alpha_2 & \phi_2\, \beta_2 \\ \Lambda_{b_2}\, D_2 & B_2 \end{pmatrix} \otimes p_1\, \alpha_1 \qquad\qquad h_1 \times h \text{ matrix,}$$

$$T_{K-1,K} = (\overline{\phi}_1\; Q_2 \otimes p_1 \;,\; \phi_1\; Q_2 \otimes p_1\, \beta_1) \qquad\qquad h \times h_2 \text{ matrix,}$$

$$T_{K,K-1} = \begin{pmatrix} p_2\,\alpha_2 \otimes \alpha_1 \\ p_2\,\alpha_2 \otimes (\Lambda_{b_1}\,D_1, B_1) \end{pmatrix} \qquad h_2 \times h \text{ matrix,}$$

$$T_{K,K} = \begin{pmatrix} \bar{g}_1\,Q_2 & g_1\,Q_2 \otimes \beta_1 \\ Q_2 \otimes b_1 & Q_2 \otimes B_1 \end{pmatrix} \qquad h_2 \times h_2 \text{ matrix.}$$

Since the effective service representations $(q_1, Q_1)$ and $(q_2, Q_2)$ are both assumed irreducible, the results of Section III.4 can be used to characterize the irreducibility of the Markov chain studied here.

As in Chapter III, the matrices $M$ and $N$ are defined by

$$M := I_{r_2} \otimes (I_{r_1} - e_{r_1}\,q_1) + Q_2 \otimes (e_{r_1}\,q_1 - Q_1)\ , \tag{6.3.1a}$$

$$N := (I_{r_2} - e_{r_2}\,q_2) \otimes I_{r_1} + (e_{r_2}\,q_2 - Q_2) \otimes Q_1\ , \tag{6.3.1b}$$

The invertibility of the matrix $I_{h_2} - T_{K,K}$ is needed. To see this, note that the matrix $\begin{pmatrix} \bar{g}_1 & g_1\beta_1 \\ b_1 & B_1 \end{pmatrix}$, being stochastic, has eigenvalues in the closed unit disc, while the matrix $Q_2$ has eigenvalues in the open unit disc in the complex plane. It is an easy exercise to show that if $(\lambda, u)$ and $(\mu, v)$ are right eigenpairs for the matrices $Q_2$ and $\begin{pmatrix} \bar{g}_1 & g_1\beta_1 \\ b_1 & B_1 \end{pmatrix}$, respectively, then $(\lambda\,\mu,\ u \otimes v)$ is a right eigenpair for the matrix $T_{K,K}$. The eigenvalues of $T_{K,K}$ are therefore all in the open unit disc, or equivalently the eigenvalues of $I_{h_2} - T_{K,K}$ have strictly positive real parts, and the matrix $I_{h_2} - T_{K,K}$ is invertible.

The invariant probability vector $\pi$ can now be obtained by following exactly the same steps as in Section III.4: First, rewrite the equation $\pi = \pi T$, in the form

$$\pi_0\,T_{00} + \pi_1\,T_{10} = \pi_0 (6.3.2a)$$

$$\pi_0\,T_{01} + \pi_1\,(Q_2 \otimes Q_1 + p_2\,q_2 \otimes p_1\,q_1) + \pi_2\,(p_2\,q_2 \otimes Q_1) = \pi_1 (6.3.2b)$$

$$\pi_{k-1}\,(Q_2 \otimes p_1\,q_1) + \pi_k\,(Q_2 \otimes Q_1 + p_2\,q_2 \otimes p_1\,q_1) + \pi_{k+1}\,(p_2\,q_2 \otimes Q_1) = \pi_k$$

$$1 < k < K-1 \tag{6.3.2c}$$

$$\pi_{K-2}\,(Q_2 \otimes p_1\, q_1) + \pi_{K-1}\,(Q_2 \otimes Q_1 + p_2\, q_2 \otimes p_1\, q_1) + \pi_K\, T_{K,K-1} = \pi_{K-1} \tag{6.3.2d}$$

$$\pi_{K-1}\, T_{K-1,K} + \pi_K\, T_{K,K} = \pi_K \tag{6.3.2e}$$

In this case the equations corresponding to (3.4.10)-(3.4.13) take the form

$$\pi_k\,[p_2\, q_2 \otimes (e_{r_1} - p_1)\, q_1] = \pi_{k-1}\,[(e_{r_2} - p_2)\, q_2 \otimes p_1\, q_1] \quad 1 < k < K\ , \tag{6.3.3}$$

$$\begin{aligned}&\pi_1\,[Q_2\,(e_{r_2}\, q_2 - I_{r_2}) \otimes Q_1 + (I_{r_2} - e_{r_2}\, q_2) \otimes I_{r_1}]\\ &\qquad\qquad + \pi_0\,(e_l\, q_2 \otimes p_1\, q_1 - T_{01}) = 0_h \end{aligned}\tag{6.3.4a}$$

$$\begin{aligned}&\pi_k\,[Q_2\,(e_{r_2}\, q_2 - I_{r_2}) \otimes Q_1 + (I_{r_2} - e_{r_2}\, q_2) \otimes I_{r_1}]\\ &\qquad\qquad \pi_{k-1}\,[Q_2\,(e_{r_2}\, q_2 - I_{r_2}) \otimes p_1\, q_1] = 0_h \quad 1 < k < K\ , \end{aligned}\tag{6.3.4b}$$

$$\begin{aligned}&\pi_k\,[Q_2 \otimes Q_1\,(e_{r_1}\, q_1 - I_{r_1}) - I_{r_2} \otimes (e_{r_2}\, q_1 - I_{r_1})]\\ &\qquad + \pi_{k+1}\,[p_2\, q_2 \otimes Q_1\,(e_{r_1}\, q_1 - I_{r_1})] = 0_h \quad 1 \le k < K - 1\ , \end{aligned}\tag{6.3.5}$$

$$\pi_k\, N = \pi_k\,[p_2\, q_2 \otimes (Q_1 - (e_{r_1} - p_1)\, q_1)] + \pi_{k-1}[Q_2 \otimes p_1\, q_1] \quad 1 < k < K\ , \tag{6.3.6}$$

Similarly use of (6.3.4a) and (6.3.1b) give

$$\pi_1\, N = \pi_1\,(p_2\, q_2 \otimes Q_1) + \pi_0\,(T_{01} - e_l\, q_2 \otimes p_1\, q_1) \tag{6.3.7}$$

On the other hand, postmultiplication of (6.3.2a) by $\begin{pmatrix} \alpha_2 & 0_{n_2} \\ 0_{n_2 \times m_2} & I_{n_2} \end{pmatrix} \otimes I_{r_1}$ readily yields

$$\pi_0\, L = \pi_1\,(p_2\, q_2 \otimes Q_1) \tag{6.3.8}$$

where the $h_1 \times h$ matrix $L$ is defined as

$$L := \begin{pmatrix} \alpha_2 & 0_{n_2} \\ 0_{n_2 \times m_2} & I_{n_2} \end{pmatrix} \otimes I_{r_1} - \begin{pmatrix} \bar{g}_2\, \alpha_2 & g_2\, \beta_2 \\ b_2\, \alpha_2 & B_2 \end{pmatrix} \otimes Q_1 \tag{6.3.9}$$

Therefore, (6.3.7) and (6.3.8) give

$$\pi_1 N = \pi_0 F \tag{6.3.10}$$

where the $h_1 \times h$ matrix $F$ is defined as

$$F := L + T_{01} - e_l\, q_2 \otimes p_1\, q_1 \quad . \tag{6.3.11}$$

It can easily be seen that when the second server is reliable (6.3.11) reduces to $F = \alpha_2 \otimes (I_{r_2} - Q_2)$ as in Chapter III.

Similarly (6.3.1a) and (6.3.5) yield

$$\pi_k M = \pi_k[Q_2 \otimes p_1\, q_1] + \pi_{k+1}[\pi_2\, q_2 \otimes (Q_1 - (e_{r_1} - p_1) q_1)] \quad 1 \leq k < K-1 \, . \tag{6.3.12}$$

and it now follows from (6.3.7) and (6.2.12) that

$$\pi_k N = \pi_{k-1} M \ , \quad 1 < k < K \ . \tag{6.3.13}$$

The system of equations (6.3.13) can be solved recursively if either the matrix $M$ or $N$ is invertible. The necessary and sufficient conditions for the invertibiltiy of $N$ are given in Lemma 2.3.16 and Corrolary 2.3.17. In the case when the matrix $N$ is invertible, this briefly outlined argument can be summarized as

$$\begin{aligned} \pi_1 &= \pi_0\, F\, N^{-1} \\ \pi_k &= \pi_{k-1}\, M\, N^{-1} \qquad\qquad 1 < k < K \\ \pi_K &= \pi_{K-1}\, T_{K-1,K}\, (I_{h_2} - T_{K,K})^{-1} \ , \end{aligned} \tag{6.3.14}$$

owing to nonsingularity of the matrix $(I_{h_2} - T_{K,K})$. Therefore, every component $\pi_k$, $1 \leq k \leq K$, of $\pi$ can be written in terms of $\pi_0$. To obtain $\pi_0$, the equations (6.3.2) is summed over $1 \leq k < K$ to yield

$$\begin{aligned} \sum_{k=1}^{K-1} \pi_k = \sum_{k=1}^{K-1} \ &\pi_k\, G + \pi_0\, T_{01} - \pi_1\, (p_2\, q_2 \otimes Q_1) \\ &+ \pi_K\, T_{K,K-1} - \pi_{K-1}\, (Q_2 \otimes p_1\, q_1) \end{aligned} \tag{6.3.15}$$

where the $h \times h$ matrix $G$ is given by

$$G := (Q_2 + p_2\, q_2) \otimes (Q_1 + p_1\, q_1) \ , \tag{6.3.16}$$

Equations (6.3.14) and (6.3.15) constitute the main result.

**Theorem 6.3.1 :** *If the matrix $N$ is invertible then the invariant probability vector $\pi = (\pi_0, \ldots, \pi_K)$ has the following* **matrix geometric** *form*

$$\pi_k = \pi_0 \ S \ R^{k-1} \qquad\qquad 1 \leq k < K \tag{6.3.17$a$}$$

$$\pi_K = \pi_0 \ S \ R^{K-2} \ U \ , \tag{6.3.17$b$}$$

*where the vector $\pi_0$ satisfies the equations*

$$\pi_0 \ Z = 0_{h_1} \quad , \quad \pi_0\, v = 1 \ . \tag{6.3.18}$$

*The matrices $R$,$S$,$U$, $Z$ and the vector $v$ are defined through the equations*

| | |
|---|---|
| $R = M \ N^{-1}$ | $h \times h$ matrix, |
| $S = F \ N^{-1}$ | $h_1 \times h$ matrix, |
| $U = T_{K-1,K} \ (I_{h_2} - T_{K,K})^{-1}$ | $h \times h_2$ matrix, |
| $W = \sum_{k=1}^{K-1} \ S \ R^{k-1}$ | $h_1 \times h$ matrix, |
| $Y = S\, R^{K-2}\, (U\, T_{K,K-1} - Q_2 \otimes p_1 q_1)$ | $h_1 \times h$ matrix, |
| $Z = [W - W\, G - T_{01} + S\, (p_2\, q_2 \otimes Q_1) - Y]$ | $h_1 \times h$ matrix, |
| $v = e_{h_1} + W\, e_{h_1} + S\, R^{K-2}\, e_{h_2}$ | $h_1 \times 1$ vector, |

*and the matrices $F$ and $G$ are given by (6.3.11) and (6.3.16), respectively.*

Similar calculations can be carried out for the continuous-time formulation and the analog of Theorem 3.6.3 would be obtained. Since in this case, the matrices $M$ and $N$ are *always nonsingular*, no assumptions are needed. Furthermore, the underlying Markov process is again *always irreducible.* Also, in both the continuous and discrete-time formulations, results similar to the ones of Section III.5 can be obtained when the order of the servers is reversed. Finally, by using the results of Section 2, the model of Chapter V can also be studied by the same solution technique.

**Special Case – When only operational failures are allowed :**

The solution for the case when idling servers cannot fail can be obtained as a special case of the discussion given above. In this case, even if the servers cannot fail when idling, the same state description has to be made as a server can still fail at the time epoch of a service completion. Therefore, the results of this case can be obtained by setting $g_2$ and $\phi_2$ equal to zero in matrices $T_{00}$ and $T_{01}$, respectively, and by setting $g_1$ equal to zero in matrices $T_{K,K-1}$ and $T_{K,K}$. Note that $\phi_2$ is set to zero *only* in $T_{01}$ since the second server is still allowed to fail with probability $\phi_2$ at a service completion epoch.

### VI.4 Numerical Examples :

In this section the effects of failures on the invariant probabilities for the model of Chapter III are illustrated through several numerical examples. For a buffer size of $K = 5$, the following three situations are considered:

(i) The case when both servers are reliable and have *Negative Binomial* service time distributions with PH-representations $(\alpha_i, A_i)$, $i = 1, 2$.

(ii) The case when both servers are subject to failures and have service distributions as in (i), and the first server has a *geometric* down time distribution with coefficient 0.8, while the second server has *hypergeometric* down time distribution with PH-representation $(\beta_2, B_2)$. In this case the idling servers are allowed to fail with probabilities $g_1 = 0.01$ and $g_2 = 0.5$.

(iii) In this case both servers are subject to failures and have service and repair distributions as in (ii), but idling servers are not allowed to fail.

The following numerical values are considered.

$$A_1 = \begin{pmatrix} 0.8 & 0.2 \\ 0 & 0.3 \end{pmatrix} , \qquad \alpha_1 = (1,0) ,$$

$$A_2 = \begin{pmatrix} 0.6 & 0.4 \\ 0 & 0.5 \end{pmatrix} , \qquad \alpha_2 = (1,0) ,$$

$$B_2 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.6 \end{pmatrix} , \qquad \beta_2 = (0.4,0.6) ,$$

The matrices $\Lambda_{f_i}$, $C_i$ and $D_i$, $i =, 1, 2$, are given by

$$\Lambda_{f_1} = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}, \quad \Lambda_{f_2} = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix} ,$$

and

$$C_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad C_2 = e_2 \beta_2 , \quad D_1 = \alpha_1 , \quad D_2 = e_2 \, \alpha_2 ,$$

while the probabilities of a failure at a service completion are $\phi_1 = 0.2$ and $\phi_2 = 0.3$. Therefore the effective service time representations $(q_i, Q_i)$, $i = 1,2$, are given by

$$Q_1 = \begin{pmatrix} 0.72 & 0.18 & 0.1 \\ 0 & 0.27 & 0.1 \\ 0.8 & 0 & 0.2 \end{pmatrix} , \qquad Q_2 = \begin{pmatrix} 0.48 & 0.32 & 0.08 & 0.12 \\ 0 & 0.4 & 0.08 & 0.12 \\ 0.5 & 0 & 0.5 & 0 \\ 0.4 & 0 & 0 & 0.6 \end{pmatrix} ,$$

and

$$q_1 = (0.8, 0, 0.2) , \qquad q_2 = (0.7, 0, 0.12, 0.18) .$$

If the random variables $S_i$, $R_i$ and $S_i^{eff}$, $i = 1,2$, denote the service time, repair time and the effective service time of server $i = 1,2$, respectively, then the above numerical values lead to following expected values: $E[S_1] = 7$, $E[S_2] = 4$, $E[R_1] = 1.25$, $E[R_2] = 2.3$, $E[S_1^{eff}] = 9.28$, and $E[S_2^{eff}] = 11.18$. Note that although $E[S_1 + R_1] > E[S_2 + R_2]$ the average effective service time of the second server is greater then the average effective service time of the first server since the second server is more likely to breakdown. The effect of breakdowns can easily be seen from the queue size probabilities in Table 6.4.1. As expected, in the first case, there are less then three jobs in the buffer for most of the time whereas in cases II and III there are more then three jobs in the buffer for most of the time. On the other hand the probabilities in cases II and III differ only slightly although the second server has a high probability of failure when it is idle. The reason for this is that the second server is idle with a very low probability and the probability of a failure when the first server is idle is very low. The invariant probability vectors obtained by using Theorem 6.3.1 are given in Table 6.4.2.

**Steady state queue size probabilities**

| Queue size | Case 1 | Case 2 | Case 3 |
|---|---|---|---|
| 0 | 0.3042 | 0.0619 | 0.0672 |
| 1 | 0.3950 | 0.1131 | 0.1143 |
| 2 | 0.1904 | 0.1495 | 0.1483 |
| 3 | 0.0758 | 0.1952 | 0.1936 |
| 4 | 0.0286 | 0.2512 | 0.2492 |
| 5 | 0.0059 | 0.2290 | 0.2275 |

**Table 6.4.1**

Case 1

$\pi_0 =(0.2181,0.0861)$

$\pi_1 =(0.2045,0.0433,0.1090,0.0382)$

$\pi_2 =(0.0736,0.0163,0.0817,0.0188)$

$\pi_3 =(0.0283,0.0063,0.0337,0.0075)$

$\pi_4 =(0.0110,0.0016,0.0132,0.0028)$

$\pi_5 =(0.0017,0.0042)$

Case 2

$\pi_0 =(0.0207,0.0061,0.0040,0.0075,0.0022,0.0014,0.0134,0.0040,0.0026)$

$\pi_1 =(0.0350,0.0087,0.0068,0.0151,0.0043,0.0028,0.0098,0.0025,0.0019,0.0181,0.0046,0.0035)$

$\pi_2 =(0.0441,0.0113,0.0086,0.0225,0.0058,0.0044,0.0128,0.0033,0.0025,0.0237,0.0061,0.0046)$

$\pi_3 =(0.0576,0.0147,0.0112,0.0294,0.0075,0.0057,0.0167,0.0043,0.0032,0.0309,0.0079,0.0060)$

$\pi_4 =(0.0788,0.0155,0.0122,0.0386,0.0097,0.0073,0.0225,0.0049,0.0038,0.0414,0.0093,0.0072)$

$\pi_5(0.0876,0.0561,0.0286,0.0567)$

Case 3

$\pi_0 =(0.0418,0.0124,0.0080,0.0012,0.0004,0.0002,0.0022,0.0006,0.0004)$

$\pi_1 =(0.0380,0.0087,0.0076,0.0164,0.0046,0.0031,0.0088,0.0025,0.0017,0.0157,0.0044,0.0030)$

$\pi_2 =(0.0437,0.0112,0.0085,0.0224,0.0057,0.0044,0.0127,0.0032,0.0025,0.0234,0.0060,0.0045)$

$\pi_3 =(0.0571,0.0146,0.0111,0.0292,0.0075,0.0057,0.0166,0.0042,0.0032,0.0307,0.0078,0.0060)$

$\pi_4 =(0.0783,0.0154,0.0119,0.0383,0.0096,0.0072,0.0223,0.0049,0.0037,0.0411,0.0093,0.0071)$

$\pi_5 =(0.0871,0.0557,0.0285,0.0563)$

**Table 6.4.2**

# CHAPTER VII

# AN APPROXIMATION METHOD FOR GENERAL TANDEM QUEUEING SYSTEMS SUBJECT TO BLOCKING

## VII.1. Introduction :

In this chapter an iterative approximation scheme is presented for finding the steady-state *marginal* probabilities of the queue sizes in general tandem queueing systems with *finite* capacity intermediate buffers and *PH-type* servers under the assumption that the steady-state exists. The algorithm is based on the analytical results obtained for two node systems and is presented under the *immediate* blocking policy. Effective two node representations of each buffer is first obtained and then approximations are made in order to express the effective representations recursively in terms of the quantities that can be obtained from the two node model. In Section 3, the accuracy of the algorithm is validated through numerical examples, both for continuous and discrete-time systems. The same approximation scheme also applies to tandem systems under *nonimmediate* blocking policy in view of the equivalence discussed for two node systems in Chapter I. Although no theoretical basis is provided for this case, comparison of the results against simulations indicates reasonable accuracy under both blocking policies even in the presence of significant blocking. In view of the results derived in Chapter VI, the approximation scheme is also applicable to failure type servers with PH-type service and repair distributions.

The algorithm uses the decomposition technique that has been used in many of the approximation algorithms. All the approximation algorithms known to date considers tandem lines with exponential servers, possibly subject to failures with exponential down time distributions. The present algorithm generalizes these algorithms to tandem lines where failure type servers with PH-type up and down time distributions are in attendance.

Most approximation algorithms use the *flow conservation principle* to decompose the tandem model into simple two node models. The approximation reported here uses the approach taken in Altıok [1], and represents the effective service distribution of a server by considering all the servers upstream of this server so as to capture the effect of blocking. The algorithm presented here differs from the method in [1] in that a similar effective representation is also considered in order to capture the effect of idling. Whereas in [1], the system is decomposed into several $M/PH/1/K$ queueing systems by approximating the arrivals to the $i^{th}$ buffer by a Poission process in view of the flow conservation principle. Once the decomposition is done, the iterative solution technique presented here uses a similar method to the one given by Bradwajn and Jow [13], where exponential servers with state dependent service rates and only immediate neighbors are considered in decomposing the tandem line.

## VII.2. Decomposition and Approximation Method :

Consider a tandem system as shown if Figure 7.2.1. There are $N$ PH-type servers in tandem with $N-1$ intermediate buffers with capacities $K_i$, $1 \leq i < N$, with the assumption that the last server is never blocked and the first server is *always busy*, i. e., infinite supply of exogenous jobs. The servers have PH-representations $(\alpha_i, Q_i)$, where $\alpha_i$ and $Q_i$ are $1 \times m_i$ and $m_i \times m_i$ matrices, respectively, and $m_i$ is the number of phases in the service distribution of the $i^{th}$ server, $1 \leq i \leq N$. The $m_i \times 1$ column vectors of absorption probabilities are denoted by $p_i$, $1 \leq i \leq N$.

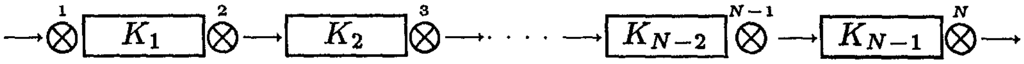

**Figure 7.2.1**

The steady state marginal probabilities for the queue sizes in the $i^{th}$ queue could be calculated from the results of Chapter III if the $(i+1)^{st}$ server were *not* subject to blocking and the $i^{th}$ server were *always* busy. However, at a service completion epoch at the $(i+1)^{st}$ server, this server may be blocked and remain blocked until there is a departure from the $(i+1)^{st}$ queue. Similarly, at a service completion at the $i^{th}$ server, this server may become idle.

In order to discuss the blocking and idling of the $i^{th}$ server the following events are defined for $1 \leq i, j \leq N$ and $t \geq 0$.

$B^i(t)$ : $i^{th}$ server is blocked at time $t$,

$\widetilde{B}_l^{i,j}(t)$ : $j^{th}$ server is not blocked and in phase of service $l$ at time $t$, and that all the servers $k$, $i \leq k < j$ are blocked,

$I^i(t)$ : $i^{th}$ server is idle at time $t$,

$\widetilde{I}_l^{i,j}(t)$ : $j^{th}$ server is not idle and in service phase $l$ at time $t$, and that all the servers $k$, $j < k \leq i$ are idle.

Assuming that the limits exists the following probabilities can now be defined:

$$P_B^i := \lim_{n\uparrow\infty} P[B^i(T_n^i)] , \qquad 1 \leq i \leq N$$

$$P_{\widetilde{B}_l}^{i,j} := \lim_{n\uparrow\infty} P[\widetilde{B}_l^{i,j}(T_n^i)] , \qquad 1 \leq i < j \leq N$$

$$P_I^i := \lim_{n\uparrow\infty} P[I^i(T_n^i)] , \qquad 1 \leq i \leq N$$

$$P_{\widetilde{I}_l}^{i,j} := \lim_{n\uparrow\infty} P[\widetilde{I}_l^{i,j}(T_n^i)] , \qquad 1 \leq j < i \leq N$$

where $T_n^i$ is the $n^{th}$ service completion time at node $i$, for $n = 0, 1, \ldots$. Note that, due to the assumption that the first server is never starved and that the last server is never blocked, some of the above defined probabilities are either 0 or 1 for some values of $i$ and $j$.

**Decomposition:** In order to capture the effect of blocking in the $(i+1)^{st}$ server, the *effective* service time distribution of this server is represented by a PH-distribution with $(m_{i+1} + \ldots + m_N)$ phases obtained by considering the service distributions of all the servers downstream from the $(i+1)^{st}$ server. Upon the $n^{th}$ service completion in the $(i+1)^{th}$ server, this server is either blocked with probability $P[B^{i+1}(T_n^{i+1})]$ or starts serving its next job (if any) according to the initialization vector $\alpha_{i+1}$. In the case of blocking, say the event $\widetilde{B}_l^{i+1,j}(T_n^{i+1})$ took place for some $j$, $i+1 < j \leq N$, and $l$, $1 \leq l \leq m_j$, the effective service time of the $i+1^{st}$ server will be equal to the residual service time in the $j^{th}$ server plus sum of the service times of all the blocked servers $j-1, \ldots, i+2$ (time to unblock), plus the service time in the $i+1^{st}$ server. Therefore, if at time $T_n^{i+1}$ the $i+1^{st}$ server is blocked the effective service phase of the $(i+1)^{th}$ server is reinitialized at phase $l$ of the $j^{th}$ server with probability $P[\widetilde{B}_l^{i+1,j}(T_n^{i+1})]$ and the transitions among these phases occur according to the matrix $Q_j$. Upon service completion in the $j^{th}$ server, service in the $(j-1)^{st}$ server is initialized according to $\alpha_{j-1}$, thus causing a transition from the phases corresponding to $j^{th}$ server to phases corresponding to the $(j-1)^{st}$ server in the effective representation of the $(i+1)^{th}$ server. Finally, upon service completion in the $(i+2)^{nd}$ server blocking period of the $(i+1)^{st}$ server will end and this server will resume its service according to the initialization vector $\alpha_{i+1}$. Therefore, if the $(m_{i+1} + \ldots + m_N) \times (m_{i+1} + \ldots + m_N)$ matrix $Q_2(i+1)$ denotes the one-step probability transition matrix among the transient phases of the effective representation of the $(i+1)^{st}$ server and if the corresponding $(m_{i+1} + \ldots + m_N) \times 1$ column vector of absorption probabilities from the effective service phases is denoted by $p_2(i+1)$, then

$$Q_2(i+1) = \begin{pmatrix} Q_{i+1} & & & & \\ p_{i+2}\,\alpha_{i+1} & Q_{i+2} & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & p_N\,\alpha_{N-1} & Q_N \end{pmatrix} , \quad p_2(i+1) = \begin{pmatrix} p_{i+1} \\ 0^T_{m_{i+2}} \\ \cdot \\ \cdot \\ 0^T_{m_N} \end{pmatrix} ,$$

$$1 \le i < N , \qquad (7.2.1)$$

and the $1 \times (m_{i+1} + \ldots + m_N)$ initialization probability vector $\alpha_2(i+1)$ of the effective representation of the $(i+1)^{st}$ server is given by

$$\alpha_2(i+1) := \left(\overline{P}_B^{i+1}\alpha_{i+1}, P_{\widetilde{B}}^{i+1,i+2}, \ldots, P_{\widetilde{B}}^{i+1,N}\right) \qquad (7.2.2)$$

for $1 \le i < N$, and the $1 \times m_j$ row vector $P_{\widetilde{B}}^{i+1,j}$ has entries $P_{\widetilde{B}_l}^{i+1,j}$, $1 \le l \le m_j$.

With this effective PH-representation the $(i+1)^{st}$ server is never blocked, but a service completion is allowed only from the phases that correspond to a service in the $(i+1)^{st}$ server. The subscript 2 indicates that the representation $(\alpha_2(i+1), Q_2(i+1))$

is the effective representation for the *second* node server when considering the $i^{th}$ buffer.

Similarly the effective service of the $i^{th}$ server can be represented by a PH-distribution with $(m_1 + \ldots + m_i)$ phases. A similar argument reveals that the effective representation $(\alpha_1(i), Q_1(i))$ of the $i^{th}$ server as the *first* node server in the two node equivalent representation of the $i^{th}$ buffer is given by

$$Q_1(i) = \begin{pmatrix} Q_i & & & & \\ p_{i-1}\,\alpha_i & Q_{i-1} & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & p_1\,\alpha_1 & Q_1 \end{pmatrix} \quad , \quad p_1(i) = \begin{pmatrix} p_i \\ 0^T_{i-1} \\ \cdot \\ \cdot \\ 0^T_1 \end{pmatrix}, \quad 1 \leq i \leq N \; , \tag{7.2.3}$$

where $p_1(i)$ is the $(m_1 + \ldots + m_i) \times 1$ column vector of absorption probabilities and

$$\alpha_1(i) := \left(\overline{P}^i_I \alpha_i, P^{i,i-1}_{\widetilde{I}}, \ldots, P^{i,1}_{\widetilde{I}}\right) \tag{7.2.4}$$

for $1 < i \leq N$, and the $1 \times m_j$ row vector $P^{i,j}_{\widetilde{I}}$ has entries $P^{i,j}_{\widetilde{I}_l}$, $1 \leq l \leq m_j$.

Therefore, for each queue i, $1 \leq i < N$, by considering the effective representations (7.2.1)-(7.2.4), the tandem system in Figure 7.2.1 can be decomposed into $N - 1$ two node systems of the type studied in ChapterIII. More preciesly the two node equivalent of buffer $i$ is as given in Figure 7.2.2.

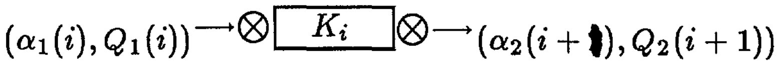

**Figure 7.2.2**

Let the $1 \times (m_1 + \ldots + m_i)$ and $1 \times (m_{i+1} + \ldots + m_N)$ row vectors $\pi_0(i)$ and $\pi_F(i)$ be the steady state probability vectors that the $i^{th}$ buffer is empty and full, respectively, obtained from the two node equivalent representation in Figure 7.2.2.

If $k = m_{i+1} + \ldots + m_{i+j} + l$ for $1 \leq j < N - i$ and $1 \leq l \leq m_{i+j+1}$, then the $k^{th}$ component of the vector $\pi_F(i)$ is the steady state probability that the servers $i, i+1, \ldots, i+j$ are all blocked and the $(i+j+1)^{st}$ server is in its $l^{th}$ phase of service. Similarly, if $k = m_i + m_{i-1} + \ldots + m_{i-j} + l$ , $0 \leq j < i-1$, $1 \leq l \leq m_{i-j-1}$, then the $k^{th}$ component of the vector $\pi_0(i)$ is the steady state probability that the buffers $i, i-1, \ldots, i-j$ are all empty and the $(i-j-1)^{th}$ server is in its $l^{th}$ phase of service.

**Approximation** : In order to put the equivalent model into the framework of Chapter III, the following approximations are made:

$$\begin{aligned} P_B^i &\approx \lim_{t \uparrow \infty} P[B^i(t)] , && 1 \leq i \leq N \\ P_{\widetilde{B}_l}^{i,j} &\approx \lim_{t \uparrow \infty} P[\widetilde{B}_l^{i,j}(t)] , && 1 \leq i < j \leq N \\ P_I^i &\approx \lim_{t \uparrow \infty} P[I^i(t)] , && 1 \leq i \leq N \\ P_{\widetilde{I}_l}^{i,j} &\approx \lim_{t \uparrow \infty} P[\widetilde{I}_l^{i,j}(t)] , && 1 \leq j < i \leq N \end{aligned} \tag{7.2.5}$$

In view of (7.2.5) and the above descriptions of the vectors $\pi_0(i)$ and $\pi_F(i)$, the initialization vectors $\alpha_1(i)$ and $\alpha_2(i+2)$ in the effective two node representation of Figure 7.2.2 are *approximated* by $\alpha_1^a(i)$ and $\alpha_2^a(i+1)$, respectively, given by

$$\alpha_1^a(i) := [\,(1 - P_0(i-1))\,\alpha_i \,, \pi_0(i-1)] \;, \tag{7.2.6a}$$

$$\alpha_2^a(i+1) := [\,(1 - P_F(i+1))\,\alpha_{i+1} \,, \pi_F(i+1)] \;. \tag{7.2.6b}$$

where the scalars $P_F(i)$ and $P_0(i)$ satisfy

$$P_F(i) = \pi_F(i)\, e_{r(i)} \quad , \quad P_0(i) = \pi_0(i)\, e_{s(i)} \quad , \quad 1 \leq i < N$$

with $r(i) = m_{i+1} + \ldots + m_N$ and $s(i) = m_1 + \ldots + m_i$.

This approximation leads to an iterative approach based on the two node analysis. The marginal probability distribution of the queue sizes at queue $i$ can be obtained once the vectors $\pi_0(i-1)$ and $\pi_F(i+1)$ are known. These vectors are calculated from the two node approximations of the $(i-1)^{st}$ and $(i+1)^{st}$ queues, respectively, during the past iteration step. More precisely, the steps of the iterative procedure, by abusing the notation and using a superscript to denote the iteration number, are :

1. Select an initial approximation for the vectors $\pi_F^0(i),\ 2 \le i < N$.
2. At iteration $n$, starting from the first queue, solves the effective two node systems of the form in Figure 7.2.2 for each buffer using $\pi_0^n(i-1)$ and $\pi_F^{n-1}(i+1)$ at queue $i$, for $2 \le i \le N-2$, and by using $\alpha_1^n(i) = \alpha_1$ at buffer 1 and $\alpha_2^{n-1}(N-1) = \alpha_N$ at buffer $N-1$. The solution for the $i^{th}$ queue will yield $\pi_0^n(i)$ and $\pi_F^n(i)$.
3. Test for a convergence criterion, if not satisfied set $n$ to $n+1$ and go to step 2.

Although there is no proof of convergence, the algorithm has always converged, up to $10^{-4}$, in less then 10 iterations in many examples. Since at each node all the servers are taken into account, the computational complexity of the algorithm grows quadratically with the number of servers in tandem. However, this growth can be made linear by considering only a few immediate neighboring servers in the effective representations. On the other hand, the algorithm can be made much faster as it is very suitable for parallel computation. At each node the effective representations $Q_1(i)$ and $Q_2(i)$, $1 \le i \le N$, are fixed, therefore a different processor can be assigned to each node, this processor needs information only from its immediate neighbors, namely from processors $i+1$ and $i-1$.

### VII.3. Numerical Examples :

The accuracy of the algorithm is tested against simulations with both relative and absolute errors in the approximations being listed. Although relative errors are usually a more important measure of accuracy of an approximation scheme, when determining the relative or absolute errors made when calculating *some* of the performance measures of interest, absolute errors play a more important role. To illustrate this point, *average queue sizes* at each buffer location are calculated. Both the relative and the absolute errors for this performance measure are better predicted by the *absolute* errors in the calculation of the probabilities. For instance, in example 3, in the approximation of the probability of having an empty buffer at the second queue 20% relative error is made, while the absolute error is only 2.3%. The relative error in the average queue size for this buffer is only 3.9%. In many cases the algorithm approximates the probabilities up to the third digit after the decimal point, but the relative error may be high when the true value of the probabilities are very small. Such an accuracy is generally considered reasonable since due to blocking it usually takes a long time for the system to reach steady state. Therefore, the simulations are believed to be accurate only up to the third digit after the decimal point.

The first four examples are for the *discrete* time model under *immediate blocking* strategy and involves only *geometric* servers. Examples 1 and 2 have the same parameters except the buffer capacities. The approximation gives better results in example 2 since it has smaller $P_0(i)$ and $P_F(i)$, $1 \leq i < N$. The examples indicate that the *reversibility* property discussed in Section 3.5 seem to hold for a general tandem line under *immediate blocking* strategy. In examples 1 and 2 since the tandem line is totally symmetric reversing the order of the servers yield the exact same probabilities. Example 4 is the same as example 3 except that the order of the servers and the intermediate buffers are reversed. The approximation scheme yields totaly *symmetric* probabilities. In the simulations the corresponding probabilities change in the third digit after the decimal point, which is believed to be the margin of inaccuracy in the simulations.

**Example 1 :** $N = 4,\ K_i = 2,\ i = 1,2,3$. Immediate blocking.

Geometric servers; $p_i = 0.5,\ i = 1,2,3,4$.

| Buffer | Queue Size | Approx. | Exact Sol'n. | Rel.Error | Abs.Error |
|---|---|---|---|---|---|
| 1 | 0 | 0.1738 | 0.1652 | -0.0521 | -0.0086 |
| | 1 | 0.4641 | 0.4596 | -0.0098 | -0.0045 |
| | 2 | 0.3621 | 0.3754 | 0.0354 | 0.0133 |
| Average queue length: | | 1.1183 | 1.2104 | 0.0183 | 0.0221 |
| 2 | 0 | 0.2629 | 0.2621 | -0.0031 | -0.0008 |
| | 1 | 0.4742 | 0.4757 | 0.0032 | 0.0015 |
| | 2 | 0.2629 | 0.2621 | -0.0031 | -0.0008 |
| Average queue length: | | 1.0000 | 0.9999 | -0.0001 | -0.0001 |
| 3 | 0 | 0.3621 | 0.3754 | 0.0354 | 0.0133 |
| | 1 | 0.4641 | 0.4596 | -0.0098 | -0.0045 |
| | 2 | 0.1738 | 0.1652 | -0.0521 | -0.0086 |
| Average queue length: | | 0.8117 | 0.7900 | -0.0275 | -0.0217 |

**Table 7.3.1**

**Example 2 :** $N = 4,\ K_i = 4,\ i = 1, 2, 3$. Immediate blocking.

Geometric servers; $p_i = 0.5,\ i = 1, 2, 3, 4$.

| Buffer | Queue Size | Approx. | Simulation | Rel.Error | Abs.Error |
|---|---|---|---|---|---|
| 1 | 0 | 0.081 | 0.079 | 0.025 | 0.002 |
| | 1 | 0.184 | 0.186 | 0.011 | 0.002 |
| | 2 | 0.238 | 0.235 | -0.013 | -0.003 |
| | 3 | 0.305 | 0.305 | 0.000 | 0.000 |
| | 4 | 0.193 | 0.185 | -0.043 | -0.008 |
| Average queue length: | | 2.347 | 2.311 | -0.016 | -0.036 |
| 2 | 0 | 0.130 | 0.127 | -0.024 | -0.003 |
| | 1 | 0.246 | 0.247 | 0.004 | 0.001 |
| | 2 | 0.247 | 0.247 | 0.000 | 0.000 |
| | 3 | 0.246 | 0.249 | 0.012 | 0.003 |
| | 4 | 0.130 | 0.130 | 0.000 | 0.000 |
| Average queue length: | | 1.998 | 2.008 | 0.005 | 0.010 |
| 3 | 0 | 0.193 | 0.195 | 0.010 | 0.002 |
| | 1 | 0.305 | 0.308 | 0.010 | 0.003 |
| | 2 | 0.238 | 0.238 | 0.000 | 0.000 |
| | 3 | 0.184 | 0.182 | -0.011 | -0.002 |
| | 4 | 0.081 | 0.076 | -0.066 | -0.005 |
| Average queue length: | | 1.657 | 1.634 | -0.014 | -0.023 |

**Table 7.3.2**

**Example 3 :** $N = 5,\ K_1 = K_3 = 2,\ K_2 = K_4 = 3$. Immediate blocking.

Geometric servers; $p_1 = 0.7,\ p_2 = 0.8,\ p_3 = 0.9,\ p_4 = 0.75,\ p_5 = 0.6$.

| Buffer | Queue Size | Approx. | Simulation | Rel.Error | Abs.Error |
|---|---|---|---|---|---|
| 1 | 0 | 0.169 | 0.176 | 0.040 | 0.007 |
| | 1 | 0.601 | 0.601 | 0.000 | 0.000 |
| | 2 | 0.230 | 0.223 | 0.031 | 0.007 |
| Average queue length: | | 1.061 | 1.047 | -0.013 | -0.014 |
| 2 | 0 | 0.095 | 0.118 | 0.195 | 0.023 |
| | 1 | 0.322 | 0.334 | 0.036 | 0.012 |
| | 2 | 0.408 | 0.377 | -0.082 | -0.031 |
| | 3 | 0.175 | 0.171 | -0.023 | -0.004 |
| Average queue length: | | 1.663 | 1.601 | -0.039 | -0.062 |
| 3 | 0 | 0.092 | 0.103 | 0.107 | 0.011 |
| | 1 | 0.592 | 0.589 | -0.005 | -0.003 |
| | 2 | 0.316 | 0.308 | -0.026 | -0.008 |
| Average queue length: | | 1.224 | 1.205 | -0.016 | -0.019 |
| 4 | 0 | 0.080 | 0.091 | 0.121 | 0.011 |
| | 1 | 0.297 | 0.305 | 0.026 | 0.008 |
| | 2 | 0.428 | 0.416 | -0.029 | -0.012 |
| | 3 | 0.195 | 0.188 | -0.037 | -0.007 |
| Average queue length: | | 1.738 | 1.701 | -0.022 | -0.037 |

**Table 7.3.3**

**Example 4 :** $N = 5,\ K_1 = K_3 = 3,\ K_2 = K_4 = 2.$ Immediate blocking.

Geometric servers; $p_1 = 0.6,\ p_2 = 0.75,\ p_3 = 0.9,\ p_4 = 0.8,\ p_5 = 0.7.$

| Buffer | Queue Size | Approx. | Simulation | Rel.Error | Abs.Error |
|---|---|---|---|---|---|
| 1 | 0 | 0.195 | 0.191 | -0.021 | -0.004 |
| | 1 | 0.428 | 0.413 | -0.036 | -0.015 |
| | 2 | 0.297 | 0.303 | 0.020 | 0.006 |
| | 3 | 0.080 | 0.093 | 0.140 | 0.013 |
| Average queue length: | | 1.262 | 1.298 | 0.028 | 0.036 |
| 2 | 0 | 0.316 | 0.308 | -0.025 | -0.008 |
| | 1 | 0.592 | 0.589 | -0.005 | -0.003 |
| | 2 | 0.092 | 0.103 | 0.107 | 0.011 |
| Average queue length: | | 0.795 | 0.776 | 0.024 | 0.019 |
| 3 | 0 | 0.175 | 0.172 | -0.017 | -0.003 |
| | 1 | 0.408 | 0.376 | -0.085 | -0.032 |
| | 2 | 0.322 | 0.333 | 0.033 | 0.011 |
| | 3 | 0.095 | 0.119 | 0.202 | 0.024 |
| Average queue length: | | 1.399 | 1.337 | 0.044 | 0.062 |
| 4 | 0 | 0.230 | 0.224 | -0.027 | -0.006 |
| | 1 | 0.601 | 0.604 | 0.005 | 0.003 |
| | 2 | 0.169 | 0.172 | 0.017 | 0.003 |
| Average queue length: | | 0.948 | 0.939 | 0.009 | 0.009 |

**Table 7.3.4**

The next example is intended to give some ideas of the accuracy of the algorithm developed in Section 2 under immediate blocking policy to tandem lines that operate under the *nonimmediate* blocking policy. The invariant probability vector for the model of Chapter III, with intermediate buffer capacity $K$, under nonimmediate blocking policy can also be obtained from Theorem 3.3.7 by considering the service location of the first node server as part of the intermediate buffer, thus incrementing the effective intermediate buffer capacity from $K$ to $K+1$. After the invariant probability vector $\pi = (\pi_0, \ldots, \pi_K, \pi_{K+1})$ is obtained in matrix-geometric form, as given by Theorem 3.3.7, the probability of having a full buffer is given by $\pi_K e_{ml} + \pi_{K+1} e_l$, where $m$ and $l$ are the number of phases in the PH-representations of the first and the second node servers, respectively.

As an example, a tandem line with three servers is considered. The first and the third node servers have *generalized Erlang* service distributions whereas the second node server has *hyperexponential* service distribution, all with two phases. The following numerical values are considered:

$$Q_1 = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix} , \quad \alpha_1 = (1,0) , \quad ES_1 = 1.5 ,$$

$$Q_2 \begin{pmatrix} -1 & 0 \\ 0 & -0.5 \end{pmatrix} , \quad \alpha_2 = (0.3, 0.7) , \quad ES_2 = 1.7 ,$$

$$Q_3 = \begin{pmatrix} -2.5 & 2.5 \\ 0 & -1 \end{pmatrix} , \quad \alpha_3 = (1,0) , \quad ES_3 = 1.4 ,$$

where $ES_i$ is the corresponding expected service time of server $i$, $1 \leq i \leq 3$. The results are summarized in Table 7.3.5. Four independent simulation runs are performed on the Performance Analysis Workstation (PAW) where the results obtained from each simulation were within 0.01 of each other. The resulsts listed in the simulation column in Table 7.3.5 are the arithmetic average of these four simulation runs. The approximation algorithm converged up to $10^{-4}$ in 5 iterations and the results are mostly within the accuracy range of the simulations.

**Example 5 :** $N = 3$, $K_1 = 2$, $K_2 = 3$. Nonimmediate blocking.

| Buffer | Queue Size | Approx. | Simulation |
|---|---|---|---|
| 1 | 0 | 0.167 | 0.157 |
| | 1 | 0.246 | 0.248 |
| | 2 | 0.587 | 0.597 |
| 2 | 0 | 0.316 | 0.350 |
| | 1 | 0.268 | 0.288 |
| | 2 | 0.196 | 0.194 |
| | 3 | 0.220 | 0.168 |

**Table 7.3.5**

## FUTURE DIRECTIONS

Throughout the thesis, long-run average results have been sought by assuming that the system will reach equilibrium fairly rapidly. However, in an actual production line, the system being subject to blocking and occasional failures, equilibrium may be achieved very slowly. Therefore, designs based on equilibrium calculations may not be adequate and a better understanding of the *transient* behavior of the system may be needed. It is thus of interest to study the transient behavior of the underlying Markov chain so as to get precise information on the *rate* at which the equilibrium is achieved and to understand how the transient performance measures behave.

Also, efforts should be made for better understanding of the approximation scheme presented in Chapter 7, as well as for developing new algorithmic methodologies to compute the quantities of interest. In particular, questions of convergence and in the event of convergence the rate of convergence and the possible limit points should be studied. Also as remarked in Chapter 7, parallel implementation of the algorithm should be explored.

# APPENDIX

## Annotated bibliography of blocking systems

Queueing systems subject to blocking have been studied by reaserchers from different research communities. Due to the wide applicability of these models there is an exhaustive list of related papers. The survey given here combines a portion of such papers. The rest of the papers, to the knowledge of the author, are included in the references.

The papers reviewed here are classified into two major classes depending on whether the servers are reliable or subject to breakdowns, and the papers in each class are presented in historical order as they appeared in the literature. A common assumption to all the models surveyed is that the last stage is never blocked, i.e., there is always space available into which the last stage server(s) can discharge a part. Also, as it is usually the case in many models, unless otherwise specified the stages have single servers.

## MODELS WITH RELIABLE SERVERS

[1] G.C. Hunt, " Sequential arrays of waiting lines," *Operations Res.*, 4, pp. 674-683, (1956).

**Model:** A tandem line of *exponential* servers with *Poisson* arrivals to an infinite buffer in front of the first server is considered. Four different models are discussed under non-immediate blocking assumption. The models involve; (i) infinite capacity buffers between the servers, (ii) no buffers between the servers, (iii) finite capacity buffers between the servers, and (iv) a line of M servers where the line moves at once as a unit (the unpaced belt production line) where no buffers and no vacant servers are allowed.
**Measure:** Mean number of jobs in the system and maximum utilization is considered.
**Results:** Maximum possible utilization is obtained in all cases and graphs of utilization vs. mean number of jobs in the system are displayed. It is reported that when utilization is less than 0.5 blocking has almost no effect.

[2] H.S. Hillier and R.W. Boling, " Finite queues in series with exponential or Erlang service times - A numerical approach," *Operations Res.*, 15, pp. 286-303, (1967).

**Model:** A tandem line of servers with *exponential* or *Erlang* service distributions seperated by finite capacity buffers is considered. Non-immediate blocking is assumed with infinite supply of job units in front of the first server.
**Measures:** Steady-state output rate and mean number of customers in the system are considered.
**Method:** When the service times have Erlang distribution the states are identified and the balance equations are solved numerically using Gauss-Seidel method. For exponential service times an approximate procedure which analyzes each stage individually as an $M/M/1/N$ queueing system is given.

[3] N.P. Rao, " Two-stage production systems with intermediate storage," *AIEE Trans.*, 7, pp. 414-421, (1975b).

**Model:** A two server tandem system with a finite capacity intermediate buffer is discussed. Service times are assumed *independent* and are *exponentially* distributed for the first server and have a *general* distribution for the second. Non-immediate blocking is assumed with infinite supply of job units in front of the first server.

**Measures:** Steady-state probabilities and the effect of unblancing is considered.

**Results:** The equations for the steady-state probabilities are shown to involve Laplace transforms of the density functions of the service time distributions and their derivatives. A recursive solution for the mean production rate is obtained and the effect of balancing is discussed. It is concluded that the balanced division of buffer capacity is not optimum for an unbalanced system and slightly higher buffer capacity is needed for the server with less variable service time distribution, and the effect of unbalancing increases with the difference in the variabilities of the service distributions. When service distributions are different, the idling times change slightly from the balanced condition with a greater rate of change at the server with less variable service time distribution.

[4] A.G Konheim and M. Reiser, " A queueing model with finite waiting room and blocking," *J. Assoc. Comput Mach.*, 23, pp. 328-341, (1976).

**Model:** A system with two servers in tandem and a finite capacity intermediate buffer is considered. Service times are assumed *independent* and *exponentially* distributed. There is a *Poisson* arrival stream to an infinite capacity buffer in front of the first server and a feedback path to this buffer for the serviced jobs in the second server. Immediate blocking strategy is assumed .

**Measure:** Steady-state probabilities are obtained.

**Method:** State of the system at time $t$ is defined by the pair $(X_t^1, X_t^2)$, where $X_t^1$ and $X_t^2$ denotes the number of jobs in the first and the second buffer at time $t$, respectively. Forward equations are written and solved for the steady state probabilities by using generating functions. Necessary and sufficient conditions are given for the stability of the system. Analytical results are then put into algorithmic form and some special cases are considered.

[5] A.B. Clarke, "A two-server queueing system with storage between servers," Math. Rep. 50, Western Michigan Univ., (1977).

**Model:** A tandem configuration of two servers with *independent* and *exponential* service distributions is considered. There is a finite capacity buffer between the servers. Non-immediate blocking is assumed with *Poisson* arrivals to an infinite capacity buffer in front of the first server. When both servers are idle an incoming job goes directly to the second server whereas if the second server is busy but the first one is not the idle is served in the first server and join the intermediate buffer. When a service completion occurs in the second server, all jobs in the intermediate buffer including the one that is blocked in the first server(if any) leave the system. If at the time of a service completion in the first server the second server is idle the job that has just been served leaves the system.

**Measure:** Steady-state probabilities of the system state process are studied.

**Method:** A system state process is defined and the block entries of the generator matrix is written explicitly. Then a matrix geometric solution is given for the steady state probabilities where the rate matrix is obtained as the minimal solution of a third order matrix equation. Some computational methods are also discussed.

[6] P. Caseau and G. Pujolle, " Throughput capacity of a sequence of transfer lines with blocking due to finite waiting room," *IEEE Trans. Software Engrg.*, SE-5, pp. 631-642, (1979).

**Model:** A tandem line of servers with *independent* and *exponential* service distributions is considered. The servers are seperated by finite capacity buffers. Immediate blocking is assumed with *Poisson* arrivals to an infinite capacity buffer in front of the first server. Extensions to the models where service times depend on queue sizes, and models that involve external arrivals to intermediate buffers are also studied.

In the latter case, interarrivals to intermediate buffers are assumed to be *Poisson* and all have same parameter. An intermediate arrival is rejected if the buffer is full, and if accepted, it leaves the system as soon as its service is completed.

**Measure:** Maximum system throughput is considered.

**Method:** The tandem line is replaced by several isolated $M/M/1/N$ queueing systems with equivalent arrival and service rates by using the equivalence relations between different types of blocking policies. Recursive relations for the utilizations of these subsystems are obtained and saturation conditions are given. Exact expressions for the maximum throughput is given for the case when there are only two servers, and also for the case when the servers are identical and the intermediate buffer capacities are equal.

[7] F.G. Foster and H.G. Perros, " On the blocking process in queueing networks," *European J. Operations Res.*, 5, pp. 276-283, (1980).

**Models:** Three different models that involves servers with *exponential* service times are discussed. In the first model, there are two such servers in tandem with a finite capacity intermediate buffer and a *Poisson* arrival stream to an infinite buffer in front of the first server. In the second model, there are several servers in parallel in the first stage and this stage is in tandem with a single server with no buffer space between them. There are *Poisson* arrival streams to infinite capacity buffers in front of the servers in the first stage. Third model is a generalization of the second model where there are two such models in parallel, tandem with a single server, again with no intermediate buffer. Non-immediate blocking strategy is adopted in all the models.

**Measure:** Mean blocking time is considered.

**Results:** For the first two models exact expressions for mean blocking time is obtained for cases when the first stage servers have infinite rates, and when they have minimum rates (to insure stability). Also conditions for the stability of the system are given. Only approximate results are given for the third model.

[8] J. Labetoulle and G. Pujolle, " Isolation method in a network of queues," *IEEE Trans. Soft. Engrg.*, SE-6, pp. 373-381, (1980).

**Model:** A general queueing network with *exponential* servers and finite capacity intermediate buffers is considered. There is an *Poisson* arrival process to the network and immediate blocking is assumed.

**Method:** An isolation method is presented where the network is subdivided into several subsystems so that each system can be studied independently. The method is said to give powerful approximations on any case where the service times or the arrival rates are dependent on the state of some part of the queueing network.

[9] G. Latouche and M.F. Neuts, " Efficient algorithmic solutions to exponential tandem queues with blocking," *SIAM J. Alg. Disc. Meth.*, 1, pp. 93-106, (1980).

**Model:** A two stage tandem system separated by a finite capacity intermediate buffer is considered. There are $r$ *identical* parallel servers in the first stage and $c$ *identical* parallel servers in the second stage. All the servers are assumed to have *independent* and *exponentially* distributed service times. There is a *Poisson* arrival stream to an infinite capacity buffer in front of the first stage. Non-immediate blocking strategy is adopted but also *full* blocking of all the first stage servers is considered in that when $r^*$ servers in stage I are blocked all the rest are also blocked. After full blocking of stage I when the number of departures at stage II reaches $k^*$, all the servers at stage I resume service again.

**Measure:** Stedy-state probabilities for the joint queue-length distribution are considered.

**Method:** A system state process is defined and for different choices of $r^*$ and $k^*$ explicit expressions for the block entries of the generator matrix are given. A recursive formula for the stationary probability distribution vector is obtained by using matrix-geometric methods. Some extensions and variants of the above model are briefly mentioned.

[10] H.G. Perros, " A symetrical exponential open queue network with blocking and feedback," *IEEE Trans. Software Engrg.*, SE-7, pp. 395-402, (1981).

**Model:** A two stage tandem system with feedback is considered. There are several servers in parallel in the first stage and only a single server in the second stage. *Identical* and *independent Poisson* arrival streams to infinite capacity buffers in front of the first stage servers is assumed. All the servers have *exponential* service distributions. Immediate blocking of a first stage server occurs each time it completes service and remains blocked until the job that it last served completes its service in the second stage server.

**Measure:** Queue-lenght distribution is studied.

**Method:** An approximate expression for the probability distribution of the number of blocked first stage servers is obtained. Based on this distribution, assuming processor sharing type of service, an approximate expression for the queue-length probability distribution is derived.

[11] M. Pinedo and R.W. Wolff, " A comparison between tandem queues with dependent and independent service times," *Operations Res.*, 30, pp. 464-479, (1982).

**Model:** Tandem queueing systems with *exponential* servers are considered. The service times are either *independent* at each server (case I) or *same* at each server, once generated according to an exponential distribution (case D). Specifically, a two server system with *Poisson* arrivals to an infinite capacity buffer is considered under light traffic conditions when there is infinite or no buffer space between the servers. Also, tandem configuration of servers with general service distributions with an infinite supply of jobs in front of the first server is considered with infinite or zero capacity intermediate buffers.

**Measures:** Expected waiting time, $E(W)$, mean and variance of departure epoch of a customer, $E(U)$ and $V(U)$, respectively, and the system capacity, $\lambda_{sup}$, is compared for cases (D) and (I). Also the effect of service regularity on the performance of the system is considered.

**Results:** For the two server case both $E(W(D))$ and $E(W(I))$ is computed when there is an infinite capacity intermediate buffer and also when there is no intermediate buffer, both under light traffic. It is concluded that under light traffic, average waiting time in case (D) is greater than in case (I) when the intermediate buffer has infinite capacity. Also for the two server system departure epochs are compared and it is shown that $E(U(D)) \leq E(U(I))$ for infinite capacity intermediate buffer and $E(U(D)) = E(V(I))$ and $V(U(D)) \geq V(U(I))$ when there is no intermediate buffer. Then, for a tandem configuration which involves servers with *general* service time distributions, expressions for $\lambda_{sup}$ are obtained. For the two server tandem system it is shown that $\lambda_{sup}(D) = \lambda_{sup}(I)$, while for a general system $\lambda_{sup}(D) < \lambda_{sup}(I)$, provided that the service times are not deterministic. Effect of service time regularity on the performance of the tandem systems is also considered. *Variability* of distribution functions is defined and some of its properties are given. It is shown that tandem systems that involves servers with less variable service time distributions has a larger capacity compared to ones with more variable service time distributions. Based on both analytical and simulation results it is concluded that *relative* performance of case (I) improves as (i) arrival rate decreases, (ii) arrival process becomes more regular, and (iii) service time distribution becomes less regular.

[12] T. Altıok, " Approximate analysis of exponential tandem queues with blocking," *European J. Operations Res.*, 11 , pp. 390-398, (1982).

**Model:** A tandem line of servers with *independent* and *exponential* service time distributions and finite capacity intermediate buffers is considered. Non-immediate blocking is assumed with *Poisson* arrivals to an infinite capacity buffer in front of the first server.

**Measure:** An approximate algorithm to obtain the steady-state probabilities of the queue sizes is studied.

**Method:** A decomposition procedure that revises the service time distribution at each server and decomposes the system into isolated simple queueing systems is presented. Decomposition is done by ignoring the interactions between the components of the system. Two basic assumptions are made; (1) the input process to each intermediate buffer is assumed to be Poisson, and (2) the blocking of a server occurs only due to the immediate successor buffer. The method decomposes the tandem line into queueing systems of the type $M/C_2/1/N$, where $C_2$ denotes a two stage Coxian distribution. A Markov chain is imbedded by looking at each system at departure points. Analytic results lead to systems of equations which are solved by iteration techniques. Numerical results and some extensions are mentioned.

[13] F.P. Kelly, " The troughput of a series of buffers," *Advances in Appl. Probability*, 14, pp. 633-653, (1982).

**Model:** Messages has to be transmitted through a tandem channel with M nodes. The nodes have finite but equal buffer capacities. Lengths of the messages are *i.i.d.* with a known common distribution. Time taken by a node to transmit a message (service time) is proportional to its message length and a given message has *same* transmission time at each node. Inputs to the first node are assumed instantaneous. Both immediate and non-immediate blocking are discussed. Also the model when transmission rates of a message at different nodes are not equal but independently distributed according to some distribution is discussed.

**Measure:** Asymptotic behavior of the throughput is studied as the number of channels increases.

**Results:** Throughput is defined as $\lim_{t\to\infty} E(\frac{N_t}{t})$, where $N_t$ is the number of messages which have been transmitted from the first node in the interval [0,t]. First, the rate at which the intermediate buffer sizes should grow to ensure that the throughput does not decline to 0 as $M \to \infty$ is investigated. Then, certain monotonicity relations between throughputs of the above models are obtained. Systems with no intermediate buffers are used to provide straightforward bounds on the degradation of throughput as the number of nodes increase. Then, more sophisticated bounds are obtained to see the effect of an increase in the buffer size. It is shown that for exponentially distributed message lengths either the transmission capacity or the buffer size should grow at rate *log M* and for distributions with tail parts proportional to $x^{-\rho}$, either the transmission rate should grow at a rate $M^{\frac{1}{\rho}}$ or the buffer size should grow at a rate $M^{\frac{1}{(\rho-1)}}$ to ensure that the throughput does not decline to zero as M increase to infinity. Therefore, the behavior of throughput for large M is determined by tail part of the distribution function of the message lenghts.

[14] H.G. Perros and T. Altıok, " Approximate analysis of open networks of queues with blocking: Tandem configurations," CS Rep. 83-11, NC State Univ., (1984).

**Model:** A tandem line of $M$ servers with *independent* and *exponential* service distributions and finite capacity intermediate buffers is considered. Non-immediate blocking is assumed with *Poisson* arrivals to an infinite capacity buffer in front of the first server. The case with *Poisson* arrivals to a *finite* capacity buffer in front of the first server is also considered.

**Measure:** An approximate algorithm to obtain the steady-state probabilities of the queue sizes is studied.

**Method:** Same decomposition method as in [12] is used but assumption (2) is relaxed and it is allowed to have blocking backlogged over any number of successive queues. Input process to each queue is still assumed to be Poisson. The system is again decomposed into several $M/C/1/N$ queueing systems in isolation, where now for the $i^{th}$ server $C$ is a $M - i + 1$ stage Coxian distribution, for $1 \leq i < M$. In the case when first buffer has finite capacity the effective arrival rate is estimated by successive iterations. Numerical examples are given and it is reported that the approximation is better for balanced systems.

[15] F.P. Kelly, " Segregating the input to a series of buffers," *Math Operations Res.,* 10, pp. 33-43, (1985).

**Model:** Two parallel systems of the type described in [13] is considered. The incoming message is directed into one of the two systems depending on its length.

**Measure:** System throughput is studied.

**Result:** It is shown that by using two systems in parallel, one dealing with long messages and the other with short messages the decay of throughput with the number of nodes $M$ can be improved from $(logM)^{-1}$ to $(loglogM)^{-1}$ when the message length distribution is exponential, and when the message length distribution is such that its tail part is proportional to $x^{-\rho}$, with $\rho > 1$, the improvement is from $M^{\frac{-1}{\rho}}$ to $M^{\frac{-1}{\rho^2}}$.

## BREAKDOWN MODELS

[16] M.C. Freeman, " The effects of breakdowns and interstage storage on production line capacity," *J. Industrial Engrg.*, 15, pp. 194-200, (1964).

**Model:** A tandem line of servers with *constant* and *equal* production times is considered. Non-immediate blocking is assumed with infinite supply of job units in front of the first server. The servers are subject to breakdown only when they are working on a job. Times between successive breakdowns and the duration of breakdowns are all assumed to be *independent* and *exponentially* distributed.

**Measure:** Efficiency of the line is defined as $\frac{P_D(0)-P_D(X)}{P_D(0)-P_D(\infty)}$, where $P_D(X)$ is the percentage of the time the line is down for a given storage capacity $X$.

**Results:** First, $P_D(0)$ and $P_D(\infty)$ are calculated. Then, the effect of system parameters on buffer capacity and gain efficiency is discussed through simulation results for three stage lines. Some general guidelines based on simulation results are given for storage allocation.

[17] E.J. Muth " A method for predicting system downtime," *IEEE Trans. on Reliability*, 17, pp. 97-102, (1968).

**Model:** A *single* machine subject to breakdowns is considered. It is assumed that the succesive failure and repair times are both *i.i.d* with known cumulative distribution functions.

**Measure:** System downtime is defined as the time the system is down during the time interval $(0,t)$ and is denoted by $D(t)$. Approximate distribution of $D(t)$ is obtained.

**Results:** It is shown that the Beta distribution is a suitable approximation for the conditional distribution of $\frac{D(t)}{t}$, given that at least one failure has occured prior to time $t$. Mean and variance of D(t) are calculated and it is shown that the distribution of $D(t)$ approaches to Normal distribution for large values of $t$.

[18] J. Masso and M.L. Smith " Interstage storages for three stage lines subject to stochastic failures," *AIIE Trans.*, 6, pp. 354-358, (1974).

**Model:** A tandem line of three servers with *equal* and *constant* service times and *independent* and *exponential* up and down times is considered. It is also assumed that there is an infinite supply of job units in front of the first server.
**Measure:** System utility is considered as a performance measure.
**Method:** Single and multiple regression techniques are used to approximate the minimal total buffer capacity required by the system to reach its maximal possible level of system utility. Also a technique to allocate a given quantity of total storage among individual interstage buffers is given.
**Results:** It is shown through simulations that when system utility is within 5% of its maximal value the effect of increasing the buffer capacity on system utility becomes insignificant.

[19] T.J. Sheskin "Allocation of interstage storage along an automatic production line," *AIIE Trans.*, 8, pp. 146-152, (1976).

**Model:** A tandem line of servers with *constant* and *equal* production times is considered. Immediate blocking is assumed with infinite supply of job units in front of the first server. Times between successive breakdowns and the duration of breakdowns are all assumed to be *independent* and *exponentially* distributed. Failures, repairs and transfer of jobs are all synchronized to time epochs.
**Measure:** Allocation of a fix total storage capacity so as to maximize the steady state output rate is considered.
**Method:** An exact *compression* algorithm is given by considering four servers in tandem. Discussion of the algorithm and guidelines for buffer allocation is given. For larger systems a much faster *approximate* decomposition algorithm is given. The algorithm analyzes each server seperately by ignoring the dependence between the arrivals and departures to a node.

[20] G.T. Artamanov, " Productivity of a two instrument discrete processing line in the presence of failures," *Kibernatika* 3, pp. 126-130; English trans. *Cybernetics*, 12 , pp. 464-468 (1977).

**Model:** Two servers in tandem with *equal* and *constant* service times and an intermediate finite capacity buffer is considered. Immediate blocking is assumed with infinite supply of job units in front of the first server. Times between successive breakdowns and the duration of breakdowns are all assumed to be *independent* and *exponentially* distributed.

**Measure:** Mean productivity is calculated using the steady state probabilities.

**Method:** A continuous time Markov chain is considered with states defined by the triple $(n, \alpha_1, \alpha_2)$ where $n$ is the number of jobs in buffer and $\alpha_1, \alpha_2$ are the up/down indicators for the first and second server, respectively. Balance equations for the steady state probabilities are written explicitly and closed form solutions are obtained.

[21] E. Ignall and A. Silver, " The output of a two-stage system with unreliable machines and limited storage," *AIIE Trans.*, 9, pp. 183-188, (1977).

**Model:** A two stage tandem system with a finite intermediate buffer and multiple servers in each stage is considered. The servers are assumed to have *constant* and *equal* service times and *independent* and *exponentially* distributed failure and repair times. Non-immediate blocking is assumed with infinite supply of job units in front of the first stage servers.

**Measure:** Estimating hourly line output by a computationally simple heuristic procedure is discussed.

**Method:** First, the model with a single server in each stage is considered and approximate line output is obtained by using known results for zero and infinite buffer capacities. Then, for multiple servers each stage is modelled as a single server with rate equal to the sum of the individual rates of the servers. An approximate formula for hourly line output is obtained as in the first case.

[22] K. Okamura and H. Yamashina, " Analysis of the effect of buffer storage capacity in transfer line systems," *AIEE Trans.*, 9, pp. 127-135, (1977).

**Model:** A line with two servers in tandem and a finite intermediate buffer is considered. The servers are assumed to have *constant* and *equal* service times, and *independent* and *geometrically* distributed failure and repair times. Non-immediate blocking is assumed with infinite supply of jobs in front of the first server.

**Measure:** The effect of buffer capacity on production rate and the mean number of jobs in the buffer is considered.

**Results:** The states of the system and the corresponding $4N+6$ by $4N+6$ transition matrix are explicitly written and the balance equations are solved for steady-state probabilities for several values of system parameters. Graphs for production rate and the mean number of jobs in the buffer are illustrated as a function of the intermediate buffer capacity. A classification of these graphs are made. The effect of variations in production times for unbalanced systems is considered. It is argued that the difference between the breakdown rates reduces the effect of installing buffer while the difference between the repair rates does not and although the effect of interchanging the servers is negligible, for large differerences it is better to put the faster server in front.

[23] J.A. Buzacott and L.E. Hanifin, " Models of automatic transfer lines with inventory banks- A review and comparison," *AIIE Trans.*, 10 , pp. 197-207, (1978).

**Model:** Compares the assumptions, method of derivation and the results of papers by Vladziewski/Sevastyanov, Koenigsberg, Buzacott and Sheskin for two stage blocking models with *independent* and *exponentially* distributed service, up and down times as common assumptions. The major difference in the assumptions is whether the idling machines can fail or not.

**Measure:** Line efficieny is compared for these models.

**Results:** In this mostly qualitative paper validity of the assumptions are tested by a real data from a transfer line and the predictions of the analytical models are compared with a simulation model which uses the actual data. The difference between analytical and simulation models is reported to be significant.

[24] R.A. Murphy, " Estimating the output of a series production system," *AIEE Trans.*, 10, pp. 139-148, (1978).

**Model:** A tandem line of servers with finite capacity intermediate buffers between some of the servers is considered. Up and down times are assumed mutually *independent* with *exponential* and *general* distributions, respectively. The servers are assumed to have *constant* but *different* service times. The failures are assumed operation dependent and no simultaneous repairs are permitted as there is a certain priority. Non-immediate blocking is assumed with infinite supply of job units in front of the first server.
**Measure:** Expected output rate is considered.
**Method:** First, results for a tandem line with no intermediate buffers are obtained. Then the case when there is only one intermediate buffer in the line is discussed by considering the servers in front of the buffer as an input block and the ones after the buffer as an output block. Effective(relative) up and down times are calculated analytically by viewing the effect of the buffer to increase the up time and decrease the down time of the output block as seen by the input block. Then, approximations are made to simplify the calculations so that sensitivity analysis and numerical optimization can be performed. The results are applied to the case where there are more than one intermediate buffer in the line by considering the effect of each buffer and finding the equivalent up and down times of the downstream servers, whence eliminating the buffers.

[25] Y.C. Ho, M.A. Eyler and T.T. Chien, " A gradient technique for general buffer storage design in a production line," *Proc.* 17th *IEEE Conf. on Control and Decision*, pp. 625-632, (1978).

**Model:** A tandem line of servers with *constant* but *different* production times is considered. Non-immediate blocking is assumed with infinite supply of job units in front of the first server. Times between successive breakdowns and the duration of breakdowns are all assumed to be *independent* and *exponentially* distributed.

**Measure:** Allocation of a given total buffer capacity to intermediate buffers to maximize the efficiency of the line is considered.

**Method:** First, the *forced-down* state of a server is classified as *irreducible* (caused by intrinsic differences in production rates of the servers) or *reducible* force-downs ( caused by random failures of the servers). An algorithm that computes the sensitivity (gradient) of the increase of line production per unit increase in buffer capacity at a buffer location and then allocates a buffer size to each location by a hill climbing procedure until all gradients become equal is presented and simulation results are displayed. In comparison with the brute force gradient approach, the algorithm can generate the gradients of all buffers in a single simulation run.

[26] S.B. Gershwin and M. Ammar, " Reliability in flexible manufacturing systems," *Proc.* 18th *IEEE Conf. on Control and Decision*, pp. 540-545, (1979).

**Model:** Tandem and merge configurations are considered for systems with three servers. Both *deterministic* and *exponential* service time distributions are considered. Repair and failure times are assumed either *geometric* or *exponential* and *independent* from the state of the system. Failures are assumed operational in that the servers can only fail when working on a job. Non-immediate blocking is considered with infinite supply of jobs in front of the first server.

**Measure:** Steady state probabilities are considered.

**Methods:** For tandem systems a state process is defined and the state transition equations are written for the internal states both for exponential and deterministic (and equal) service times. Also transition equations for a merge configuration are written for the internal states and they are shown to be similar to the equations for the tandem model. Comparisons and some speculations are made for more complex merge configurations. In all cases steady-state probabilities for the internal states are assumed to be in sum-of-products form.

[27] J. Wijngaard, " The effect of interstage buffer storage on the output of two unreliable production units in series, with different production rates," *AIIE Trans.*, 11, pp. 42-47, (1979).

**Model:** A line with two servers in tandem and a finite intermediate buffer is considered. The servers are assumed to have *constant* but *different* service times and *independent* and *exponentially* distributed failure and repair times. Immediate blocking is assumed with infinite supply of job units in front of the first server. The main difference from the other models is that the job units are modelled as a *continuous* random variable rather than being discrete.

**Measure:** System production rate is considered.

**Method:** The state of the system is defined as in [20]. Regeneration points are identified as enterence points to states that corresponds to empty buffer and a cycle is defined as the time between subsequent regenerations. Production rate is defined as quotient of the expected production per cycle and the expected duration of that cycle. For equal production rates, this lead to a differential equation and is solved explicitly. For different production rates, the problem gave rise to a system of three differential equations and only the form of the solution is given in this case. Simulation results for the effect of buffer on an unbalanced line are also briefly discussed.

[28] A.L. Soyster, J.W. Schmidt and M.W. Rohrer," Allocation of buffer capacities for a class of fixed cycle production lines," *AIEE Trans.*, 11, pp. 140-146, (1979).

**Model:** A tandem line of servers with finite buffers between the servers is considered. The servers are assumed to have *constant* but *different* service times. Down time distribution of each server is assumed to be an *independent Bernoulli* process, i.e., the probability that server $i$ is down during a given cycle is $p_i$, independent of the state of all other servers and previous state of server $i$. Non-immediate blocking is assumed with infinite supply of job units in front of the first server.

**Measure:** Allocation of a given set of buffer capacities to maximize the steady state output rate is considered.

**Method:** The problem is formulated as a nonlinear programming problem where the form of the objective function is not known. Objective function is approximated by using lower and upper bounds. First, for a two server system an exact expression is obtained. Then, for the general case approximate expressions are obtained by using two server subsystems. Upper and lower bounds are established for the steady-state system output, and certain concave, seperable programs are formulated to determine the optimal buffer capacities. Simulations showed that the objective function that is obtained through approximations is insensitive to modest changes in capacity allocation. Also, it is concluded that larger buffers should be allocated around servers with lower reliability.

[29] S.B. Gershwin and O. Berman, " Analysis of transfer lines consisting of two unreliable machines with random processing times and finite storage buffers," *AIIE Trans.*, 13 , pp. 2-11, (1981).

**Model:** A line with two servers in tandem with *exponential* service times and finite capacity intermediate buffer is considered. Immediate blocking is assumed with infinite supply of job units in front of the first server. The servers are subject to breakdown only when they are working on a job. Times between successive breakdowns and the duration of the breakdowns are all assumed to be *independent* and *exponentially* distributed.

**Measures:** System production rate, utilization and average in-process inventory are considered.

**Method:** A continuous time Markov chain is considered with states defined as in [20]. Explicit expressions for steady state probabilities of the states are found by using balance equations and assuming product-form solutions. Then, these probabilities are used to calculate the above performance measures. These measures are displayed as a function of the productivity of each server and limiting behaviors are obtained.

**Results:** The existence of a saturation effect is illustrated through numerical examples. The system saturates after a point in the sense that no further increase in the speed of the first server can improve the productivity of the system. It is illustrated that the system's production rate increases as the productivity of the servers increase. Also, the average in-process inventory increases as the first server becomes more productive and decreases as the second server becomes more productive.

[30] E.J. Muth and S. Yeralan, " Effect of buffer size on productivity of work stations that are subject to breakdowns," *Proc.* $20^{th}$ *IEEE Conf. on Decision and Control*, pp. 643-648, (1981).

**Model:** A line with two servers in tandem and a finite intermediate buffer is considered. The servers are assumed to have *constant* and *equal* service times and *independent* and *exponentially* distributed failure and repair times. The servers can only breakdown when working on a job. Non-immediate blocking is assumed with infinite supply of jobs in front of the first server.

**Measure:** Variation of productivity with the intermediate buffer capacity is considered.

**Method:** A system state process is defined and 4N+8 states are identified, where N is the capacity of the intermediate buffer. The states are ordered in such a way so that the transition matrix has a block tridiagonal structure with 4 by 4 blocks. This structured form leads to the solution of the steady-state probabilities by successively solving a system of 4 simultaneous equations. Moreover, probabilities for internal states are shown to have a *scalar* geometric property. Specifically, if X is the random variable that denotes the number of jobs in the buffer at steady-state then $P(X=k) = \lambda^k P(X=1)$, for $1 < k < N$, where the scalar $\lambda$ is an eigenvalue of a 4 by 4 matrix. Therefore, the probabilities are geometrically increasing or decreasing depending on the availabilities of individual servers. The system production rate is calculated by using these probabilities. Therefore, for *any* buffer size the computation of the production rate is equally simple. In general production rate depends on several system parameters but in order to study the behavior of the system production rate as a function of the intermediate buffer capacity a simpler approximate expression for the production rate that only depends on the buffer capacity is given by an empirical formula. This empirical formula is reported to be correct up to $10^{-10}$ over a wide range of parameters.

[31] T. Ohmi, " An approximation for the production efficiency of automated transfer lines with in-process storage," *AIEE Trans.*, 13, pp. 22-28, (1981).

**Model:** A tandem line of servers with *constant* and *equal* service times and finite capacity intermediate buffers is considered. Each server is composed of stations that are mechanically interlocked where each station is assumed to have a *constant* probability of breaking down. A server can only breakdown when working on a job and it breaksdown when one of its stations breakdown. Repair times of each station are *i.i.d.* with common *exponential* distribution. Two further assumptions are made; (i) only one server can be down at a given time, and (ii) at the time when a server breaksdown the number of stocked jobs in each buffer equals its line-averaged value, i.e., fluctuations of in-process inventories are ignored.

**Measure:** Production efficiency and capacity allocation are considered.

**Method:** First, lifetime of each server is shown to have *geometric* distribution. Then, a method for approximating the line efficiency is developed. Also, the optimal partitioning of the line and the method of allocating capacities for buffers are numerically investigated.

[32] M.F. Neuts, "A queue with server breakdowns and repairs" in *Matrix-geometric solutions in stochastic models - An algorithmic approach,* The John Hopkins Univ. Press, pp. 274-286, (1981).

**Model:** A system with $N$ parallel servers with *exponential* service times and *Poisson* arrivals to an infinite capacity buffer whose rates may depend on the number of operative servers is considered. Times between successive breakdowns and the duration of the breakdowns are all assumed to be *independent* and *exponentially* distributed. It is assumed that there are $C$ repairmen ($C < N$) available.

**Measure:** Computation of steady state probabilities is considered.

**Method:** The states of the system are defined and the generator matrix is given. The state transition process is shown to be a Quasi Birth and Death (QBD) process. For N=1 (M/M/1 queue) the steady-state probabilities are given explicitly. For the general case, recursive equations for the steady-state probabilities are obtained by matrix geometric methods and a numerical algorithm is briefly discussed. Also, computation of the conditional working time distribution is discussed and a numerical example is given.

[33] T. Altıok and S.Stidham, Jr. " A note on transfer lines with unreliable machines, random processing times and finite buffers," *IIE Trans.*, 14, pp. 125-127, (1982).

A comment to the effect that except the two node models the two type of blocking strategies (immediate and non-immediate) are not equivalent and it is necessary to insert a blocking indicator into the state description to study models which adopt non-immediate blocking strategy.

[34] T. Altıok and S. Stidham, Jr. " The allocation of interstage buffer capacities in production lines," *IIE Trans.*, 15 , pp. 292-299, (1983).

**Model:** A tandem line of servers with *independent* and *exponentially* distributed service, breakdown and repair times is considered. Non-immediate blocking is assumed with infinite supply of job units in front of the first server.

**Measures:** Optimal allocation of buffer capacities to maximize the average output rate is considered.

**Method:** It is shown that the effective service completion time has two stage Coxian type distribution, whence the system is transformed into a tandem line of *reliable* servers with two stage Coxian service distributions. States of a continuous time Markov chain are defined and the balance equations are solved by using the power method. A search technique for optimal buffer allocation is also discussed.

[35] S.B. Gershwin and I.C. Schick, " Modelling and analysis of three-stage transfer lines with unreliable machines and finite buffers, " *Operations. Res.*, 31, pp. 354-380, (1983).

**Model:** A tandem configuration of servers with *constant* and *equal* service times is considered. Failure and repair times are assumed *geometric* and *independent* from the state of the system. Failures are assumed operational in that servers can only fail when working on a job. Non-immediate blocking is considered with infinite supply of jobs in front of the first server.

**Measures:** In-process inventory and efficiency (production rate) are considered.

**Methods:** The states are defined as in [20] and sum-of-products form solution is assumed for the steady-state probabilities of the internal states. For the boundary states, expressions for the steady-state probabilities are derived by using the transition equations. Therefore the order of the system is reduced from $N^2$ to $N$, where $N$ is the total buffer capacity. Although the order is reduced the new system is not sparse and it may also become ill-conditioned. The general results are applied to a system with three servers, transient and boundary states are identified and a procedure is discussed for the solution of steady-state probabilities.

**Results:** The following results based on simulations are given for three server systems.

- Total average in-process inventory is proportional to the failure rate of the third server and inversely proportional to the failure rate of the second server.
- Efficiency increases with the total storage size, but so does the error in the calculations of the steady-state probabilities.
- Efficiency stays approximately constant when all probabilities are multiplied by a constant number and the storage capacity is divided by the same number.
- Production rate is not affected by the reversal of the data describing the system while the error in the numerical calculations is.
- For a balanced line, maximum efficiency is obtained when the intermediate buffers have equal capacities.
- When last machine is almost reliable, the system behaves like a two-server system.

[36] B. Vinod, " Unreliable queueing systems," IE Working Paper, 83-105,Dept. of Industrial Engrg. Rutgers Univ., (1983).

**Model:** Both a single server and several *identical* servers in *parallel,* subject to random breakdowns and repairs are considered. Since the servers are in parallel there is no blocking. System is modelled as a multiple customer class priority queueing system. Finite source of *failure* customers have pre-emptive resume priority over the *job* customers, and a queue of breakdowns is not permissible. Arrival process of failure customers are *Poisson* whose rate depends on the number of operative servers. Service times for failure customers (repair times) are *exponentially* distributed. Both operational and time dependent breakdowns are discussed for the single server case. In the latter case, extensions to multiple servers are also done.

**Measures:** Steady-state probabilities, marginal queue length distribution and its moments, covariance between failure and job class customers, utilizations and mean waiting times are considered.

**Method:** By lexicographically ordering the states, a continuous time Markov process with a block tri-diagonal transition matrix (a QBD process) is obtained. The stationary probability vector, $x$, of the process is partitioned into vectors $x_i$ and it is shown that these vectors have *matrix-geometric* form, i.e., $x_i = x_0 R$, where the irreducible, nonnegative rate matrix $R$ is the minimal solution to a non-linear matrix equation. A recurrence relation is given to compute R. For the single server case R and $x_0$ are obtained explicitly. For multiple servers $x_0$ is shown to be the solution of a system of linear equations. In both cases expressions for the above performance measures are obtained.

[37] T. Altıok, " Approximate analysis of production lines with general service and repair times and with finite buffers," IE Rep. 84-4, Rutgers Univ, (1984).

**Model:** A tandem line with finite capacity buffers between the servers is considered. The servers are assumed to have *independent Erlang* service time distributions. The up time distribution is assumed to be *exponential* while the down-time has a *general* distribution. Both infinite supply of job units in front of the first server and the case where there is a *Poisson* arrival stream to an infinite buffer in front of the first server is considered with non-immediate blocking assumption.

**Measures:** Average number of jobs in each buffer and utilization of each server is considered.

**Method:** Although an expression for the cumulative distribution of service completion time is obtained, it is very complicated to deal with. However, in the case of exponential repair times, by observing the corresponding Laplace-Stieltjes transforms it is seen that the service completion time distribution is the sum of several two-stage phase-type distributions. Then, a cumulative distribution function is obtained by assuming that at most one breakdown may occur during the process time of a job. In this way failures are incorporated into the service completion times which are approximated by specific phase-type distributions. For this approximation by phase-type distributions, first two or three moments of the derived service time distribution is used. Particular mixtures of the sum of exponential distributions are chosen by empirical observations. Then, by using the results of Perros and Altıok [14], the effective service of the $i^{th}$ server is represented with a phase structure involving $2 \times (M - i + 1)$ phases, where $M$ is the total number of nodes in the system. Numerical examples and some empirical observations are also given.

[38] B. Vinod and M. Sabbagh, " Optimal performance analysis of manufacturing systems subject to tool availability," preprint, (1984).

**Model:** A closed network of N jobs and M servers with *exponential* service times is considered. Routing probabilities are assumed to be *independent* from the state of the system. It is assumed that a job can be processed only if a tool is available and each server requires only one tool to process a job. The server is assumed *down* if a tool is not available. Up and down times of the servers are assumed to be *exponentially* distributed.

**Measure:** Optimal allocation of spare tool classes is discussed.

**Method:** The processing rate is modified to capture the interruptions caused by tool failures. Known results are then applied to these approximate, new processing times to obtain joint queue length distributions. Then, the problem of optimal allocation of spare tools at each server is formulated as a nonlinear integer programming problem. An algorithm is given to solve this mathematical programming problem.

# REFERENCES

[1] T.Altıok, " Approximate analysis of production lines with general service and repair times and with finite buffers," IE Rep. 84-4, Rutgers Univ, (1984).

[2] T. Altıok, " Approximate analysis of exponential tandem queues with blocking," *European J. Operations Res.*, 11 , pp. 390-398, (1982).

[3] T. Altıok and H.G. Perros, " Open netwoks of queues with blocking: Split and merge configurations," CS Rep. 83-10, NC State Univ., (1983).

[4] T. Altıok and S.Stidham, Jr. " A note on transfer lines with unreliable machines, random processing times and finite buffers," *IIE Trans.*, 14, pp. 125-127, (1982).

[5] T. Altıok and S. Stidham, Jr. " The allocation of interstage buffer capacities in production lines," *IIE Trans.*, 15 , pp. 292-299, (1983).

[6] G.T. Artamanov, " Productivity of a two instrument discrete processing line in the presence of failures," *Kibernatika* 3, pp. 126-130 ; English trans. *Cybernetics*, 12 , pp. 464-468 (1977).

[7] K.E. Atkinson, *An Introduction to Numerical Analysis*, Wiley, (1980).

[8] B. Avi-Itzhak, " A sequence of service stations with arbitrary input and regular service times," *Management Sci.*, 11, pp. 565-571, (1965).

[9] B. Avi-Itzhak and M. Yadin, " A sequence of two servers with no intermediate queue," *Management Sci.*, 11, pp. 553-564, (1965).

[10] A. Berman and R.J. Plemmons, *Nonnegative matrices in mathematical sciences*, Academic Press, (1979).

[11] P.P. Bocharov and V.A. Naumov, " Matrix-geometric stationary distribution for the $PH/PH/1/n$ queue," *INRIA Rapports de Recherche*, 304, (1984).

[12] O. Boxma and A. Konheim, " Approximate analysis of exponential queueing systems with blocking," *Acta Informatica*, 15, pp. 19-66, (1981).

[13] A. Brandwajn and Y.L. Jow, " An approximation method for tandem queues with blocking," preprint, (1985).

[14] J.W. Brewer, " Kronecker products and matrix calculus in systems theory," *IEEE Trans. on Circuits and Systems*, 25 , pp. 772-781, (1978).

[15] J.A. Buzacott, " Automatic transfer lines with buffer stocks," *Internat. J. Prod. Res.*, 5, pp. 183-200, (1967).

[16] J.A. Buzacott, " The effect of station breakdowns and random processing times on the capacity of flow lines with in-process storage," *AIEE Trans.*, 4, pp. 308-312, (1972).

[17] J.A. Buzacott and L.E. Hanifin, " Models of automatic transfer lines with inventory banks – A review and comparison," *AIIE Trans.*, 10 , pp. 197-207, (1978).

[18] P. Caseau and G. Pujolle, " Throughput capacity of a sequence of transfer lines with blocking due to finite waiting room," *IEEE Trans. Software Engrg.*, SE-5, pp. 631-642, (1979).

[19] A.B. Clarke, " Markovian queues with servers in tandem," Math. Rep. 49, Western Michigan Univ., (1977).

[20] A.B. Clarke, "A two-server queueing system with storage between servers," Math. Rep. 50, Western Michigan Univ., (1977).

[21] A.B. Clarke, " A multiserver general service time queue with servers in series," Math. Rep. 51, Western Michigan Univ., (1978).

[22] A.B. Clarke, " Waiting times for Markovian queue with servers in series," Math. Rep. 52, Western Michigan Univ., (1978).

[23] R.V. Evans, " Capacity of queueing networks," *Operations Res.*, 15, pp. 530-536, (1967).

[24] M.C. Freeman, " The effects of breakdowns and interstage storage on production line capacity," *J. Industrial Engrg.*, 15, pp. 194-200, (1964).

[25] F.G. Foster and H.G. Perros, " On the blocking process in queueing networks," *European J. Operations Res.*, 5, pp. 276-283, (1980).

[26] F.G. Foster and H.G. Perros, " Hierarchical queue networks with partially shared servicing," *J. Operations Res. Soc.*, 30, pp. 157-166, (1979).

[27] S.B. Gershwin and M. Ammar, " Reliability in flexible manufacturing systems," *Proc.* $18^{th}$ *IEEE Conf. on Control and Decision*, pp. 540-545, (1979).

[28] S.B. Gershwin and O. Berman, " Analysis of transfer lines consisting of two unreliable machines with random proccessing times and finite storage buffers," *AIIE Trans.*, 13 , pp. 2-11, (1981).

[29] S.B. Gershwin and I.C. Schick, " Modelling and analysis of three-stage transfer lines with unreliable machines and finite buffers," *Operations. Res.*, 31, pp. 354-380, (1983).

[30] W.J. Gordon and G.F. Newell, " Cyclic queueing systems with restricted length queues," *Operations Res.*, 15, pp. 266-278, (1967).

[31] J.M. Hatcher, " The effect of internal storage on the production rate of a series of stages having exponential service times," *AIIE Trans.*, 1, pp. 150-156, (1969).

[32] D.K. Hildebrand, " Stability of finite queue, tandem server systems," *J. Appl. Probability*, 4, pp. 571-583, (1967).

[33] D.K. Hildebrand, " On the capacity of tandem server, finite queue, service systems," *Operations Res.*, 16, pp. 72-82, (1968).

[34] H.S. Hillier and R.W. Boling, " Finite queues in series with exponential or Erlang service times - A numerical approach," *Operations Res.*, 15, pp. 286-303, (1967).

[35] Y.C. Ho, M.A. Eyler and T.T. Chien, " A gradient technique for general buffer storage design in a production line," *Proc.* $17^{th}$ *IEEE Conf. on Control and Decision*, pp. 625-632, (1978).

[36] G.C. Hunt, " Sequential arrays of waiting lines," *Operations Res.*, 4, pp. 674-683, (1956).

[37] E. Ignall and A. Silver, " The output of a two-stage system with unreliable machines and limited storage," *AIIE Trans.*, 9, pp. 183-188, (1977).

[38] F.P. Kelly, " The troughput of a series of buffers," *Advances in Appl. Probability*, 14, pp. 633-653, (1982).

[39] F.P. Kelly, " Segregating the input to a series of buffers," *Math Operations Res.* , 10, pp. 33-43, (1985).

[40] A.D. Knott, " The inefficiency of a series of workstations - A simple formula," *Intern. J. Prod. Res.*, 8, pp. 109-119, (1970).

[41] A.G Konheim and M. Reiser, " A queueing model with finite waiting room and blocking," *J. Assoc. Comput Mach.*, 23, pp. 328-341, (1976).

[42] A.G. Konheim and M. Reiser, " Finite capacity queuing systems with applications in computer modelling," *SIAM J. Comput.*, 7, pp. 210-229, (1978).

[43] J. Labetoulle and G. Pujolle, " Isolation method in a network of queues," *IEEE Trans. Soft. Engrg.*, SE-6, pp. 373-381, (1980).

[44] C. Langaris and B. Conolly, " On the waiting time of a two-stage queueing system with blocking," *J. Appl. Probability*, 21, pp. 628-638, (1984).

[45] G. Latouche and M.F. Neuts, " Efficient algorithmic solutions to exponential tandem queues with blocking," *SIAM J. Alg. Disc. Meth.*, 1, pp. 93-106, (1980).

[46] S.S. Lavenberg, " Stability and maximum departure rate of certain open queueing networks having finite capacity constraints," *RAIRO Informatique/ Computer Science*, 12, pp. 353-370, (1978).

[47] T. Makino, " On the mean passage time concerning some queueing problems of the tandem type," *J. Operations Res. Soc. Japan*, 7, pp. 17-47, (1964).

[48] J. Masso and M.L. Smith " Interstage storages for three stage lines subject to stochastic failures," *AIIE Trans.*, 6, pp. 354-358, (1974).

[49] B. Melamed, " A note on the reversibility and duality of some tandem blocking queueing systems " *Management Sci.,* to appear.

[50] E.J. Muth " A method for predicting system downtime," *IEEE Trans. on Reliability*, 17, pp. 97-102, (1968).

[51] E.J. Muth, " The production rate of work stations with variable service times," *Internat. J. Prod. Res.*, 11, pp. 155-169, (1973).

[52] E.J. Muth, " The reversibility property of production lines," *Management. Sci.*, 25, pp. 152-158, (1979).

[53] E.J. Muth and S. Yeralan, " Effect of buffer size on productivity of work stations that are subject to breakdowns," *Proc.* 20th *IEEE Conf. on Decision and Control*, pp. 643-648, (1981).

[54] M.F. Neuts, " Two queues in series with a finite, intermediate waitingroom," *J. Appl. Probability*, 5, pp. 123-142, (1968).

[55] M.F. Neuts, " Computational uses of the method of phases in the theory of queues," *Comput. Math. Appl.*, 1, pp. 151-166, (1975).

[56] M.F. Neuts, " The probabilistic significance of the rate matrix in matrix-geometric invariant vectors," *J. Appl. Probability*, 17, pp. 291-296, (1980).

[57] M.F. Neuts, " Explicit steady-state solutions to some elementary queueing models," *Operations Res.*, 30, pp. 480-489, (1982).

[58] M.F. Neuts, *Matrix-geometric solutions in stochastic models - An algorithmic approach* , The John Hopkins Univ. Press, (1981).

[59] G.F. Newell, *Approximate behavior of tandem queues*, Lecture Notes in Economics and Mathematical Systems 171, Spriger-Verlag, (1979).

[60] T. Ohmi, " An approximation for the production efficiency of automated transfer lines with in-process storage," *AIEE Trans.*, 13, pp. 22-28, (1981).

[61] K. Okamura and H. Yamashina, " Analysis of the effect of buffer storage capacity in transfer line systems," *AIEE Trans.*, 9, pp. 127-135, (1977).

[62] H.G. Perros, " A symmetrical exponential open queue network with blocking and feedback," *IEEE Trans. Software Engrg.*, SE-7, pp. 395-402, (1981).

[63] H.G. Perros and T. Altıok, " Approximate analysis of open networks of queues with blocking: Tandem configurations," CS Rep. 83-11, NC State Univ., (1984).

[64] M. Pinedo and R.W. Wolff, " A comparison between tandem queues with dependent and independent service times," *Operations Res.*, 30, pp. 464-479, (1982).

[65] N.U. Prabhu, " Transient behavior of a tandem queue," *Managment Sci.*, 13, pp. 631-639, (1967).

[66] N.P. Rao, " On the mean production rate of a two-stage production system of the tandem type," *Internat. J. Prod. Res.*, 13, pp. 207-217, (1975a).

[67] N.P. Rao, " Two-stage production systems with intermediate storage," *AIEE Trans.*, 7, pp. 414-421, (1975b).

[68] T.J. Sheskin, " Allocation of interstage storage along an automatic production line," *AIIE Trans.*, 8, pp. 146-152, (1976).

[69] A.L. Soyster, J.W. Schmidt and M.W. Rohrer," Allocation of buffer capacities for a class of fixed cycle production lines," *AIEE Trans.*, 11, pp. 140-146, (1979).

[70] T. Suzuki, " On a tandem queue with blocking," *J. Operations Res. Soc. Japan*, 6, pp. 137-157, (1964).

[71] Y. Takahashi, H. Miyahara and T. Hasegawa, " An approximation method for open restricted queueing networks," *Operations Res.*, 28, pp. 594-602, (1980).

[72] B. Vinod, " Unreliable queueing systems," IE Rep., 83-105,Dept. of Industrial Engrg. Rutgers Univ., (1983).

[73] B. Vinod and M. Sabbagh, " Optimal performance analysis of manufacturing systems subject to tool availability," preprint, (1984).

[74] D. Yao and J.A. Buzacott, " Modelling a class of flexible manufacturing systems with reversible routing," IE Rep. 83-02, Univ. of Toronto, (1983a).

[75] D. Yao and J.A. Buzacott, " Modelling the performance of flexible manufacturing systems," Manuscript, Columbia Univ., IE Dept., (1983b).

[76] J. Wijngaard, " The effect of interstage buffer storage on the output of two unreliable production units in series, with different production rates," *AIIE Trans.*, 11, pp. 42-47, (1979).

[77] B.Wong, W. Giffin and R. Disney, " Two finite $M/M/1$ queues in tandem: A matrix solution for the steady state," *Opsearch*, 14, pp. 1-18, (1977).

[75] D. Yao and J.A. Buzacott, " Modelling the performance of flexible manufacturing systems," Manuscript, Columbia Univ., IE Dept., (1983b).

[76] J. Wijngaard, " The effect of interstage buffer storage on the output of two unreliable production units in series, with different production rates," *AIIE Trans.*, 11, pp. 42-47, (1979).

[77] B.Wong, W. Giffin and R. Disney, " Two finite $M/M/1$ queues in tandem: A matrix solution for the steady state," *Opsearch*, 14, pp. 1-18, (1977).