1989

A Stochastic model of multi-stage pull production systems

Samia Siha
Iowa State University

Order Number 9003564

A stochastic model of multistage pull production systems

Siha, Samia, Ph.D.

Iowa State University, 1989

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

A Stochastic model of multi-stage pull production systems

by

Samia Siha

A Dissertation Submitted to the

Graduate Faculty in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major: Industrial Engineering

Approved:

Members of the Committee:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa
1989

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am deeply indebted to many people for their part in making this dissertation possible.

I would like to express my deepest appreciation and gratitude for my advisor, Dr. H. T. David, whose excellent guidance, stimulating discussions and inspiring encouragement helped make this research possible.

My gratitude also goes to Dr. Keith McRoberts for his valuable comments, suggestions and support. I wish to thank professor Vector Tamashunas for introducing the idea of "Pull System" to me , his helpful suggestions valuable comments and constant support are very much appreciated.

My gratitude also goes to other members of my committee, Dr. V. Sposito, Dr. D. Grosvenor and Dr. T. Barta. I am also grateful to Dr. R. Linn for his help.

I am especially grateful to my husband Magdy Riad for his continuous encouragement, support, understanding and patience. I am also very grateful to my children Lillian, Hania and Magd and my mother.

I would like to express my indebtedness to pastor Richard Kapfer for providing spiritual direction and support during hard times.

Finally, my thanksgiving to my Heavenly Father who blessed me with all these people.

1 INTRODUCTION

1.1 Overview

In this research we will study the pull production system based on the Just In Time doctrine with emphasis on stochastic modeling under Markovian assumptions.

1.1.1 The Push System

This is the traditional production system. The production requirement is met by issuing various production schedules to all stations. The stations produce the parts in accordance with their schedule, with the preceding station supplying the parts to its following station, and so on. According to Monden [25], it is difficult, under this system, to promptly adapt to changes caused by trouble at some stations or by demand fluctuations, without carrying high inventory at all stages of production.

1.1.2 Just-In-Time

"Just-In-Time" is a philosophy or a doctrine more than a method. It includes all the activities from market research to the design of product, and on through production to delivery to the customer. The objective of the JIT doctrine is to avoid waste in all aspects of a production system. The ideal JIT plant is visualized

as a series of stations whether physically located in series or not [31]. To implement the JIT doctrine, it is especially important to reduce setup time, reduce lot size and assure quality output upon demand.

1.1.3 The Pull System

The Pull system is a way of implementing the JIT doctrine. It is a revolutionary system according to Monden [25], in the sense that subsequent stations demand parts from preceding stations. In particular, the final assembly line is responsive to the timing and quantity of parts required, and goes to the preceding station to obtain the necessary parts in the necessary quantity at the necessary time for part assembly. Terada and Kimura [36] added to the above that the pull system is to hold the inventory at a certain level at each stage.

1.1.4 Kanban

This is a production control system that can help apply the pull system concept. It involves a card in a vinyl pocket which carries the following information: part number, quantity per container, preceding station and succeeding station. There are two kinds of Kanban cards: 1-in station Kanban. 2- inter-station Kanban. According to [29], [32], [36], the procedure of using the two Kanban cards is as follows:

1. To install the Kanban system the manager assigns a certain number of Kanban cards to each station of the process.

2. When a unit is to be used, an attached inter-station card is detached from the unit, and is taken to the preceding station. This constitutes the authorization

to pick up the required (replacement) unit.

3. The in-station card is removed from the unit, and the inter-station card is attached to it; it is then moved forward to where it is to be used.

4. The in-station card is hung on a board; this unattached in-station card authorizes the production of another unit.

The Kanban system is built on the following assumptions, which may be difficult to satisfy in particular instances:

- The setup and order costs are negligible, which is not always the case.
- Kanban depends on worker experience because it is controlled by constant visual monitoring.
- All operations are linked in chains which feed the final assembly line.

1.2 Previous Research

1.2.1 Pull Production Systems

Most of the published literature in this area is focused on the explanation of the system , its constraints, prerequisites and applications [28], [31], [35]. However, there are just two publications that offers mathematical modeling of the system. Terada and Kimura [36] provides several basic equations for the Kanban system in a multi-stage serial production system. He finds that, in case of the pull system, the production fluctuation will not be amplified in the preceding stages if the size of order is kept small enough. Bitran and Chang [5] present an optimization model for

the Kanban system in a multi-stage assembly production setting; they also offer a solution procedure. Both of the above mentioned papers consider the deterministic model with constant lead time.

1.2.2 Queuing Networks and Production System

The queuing network is a related area of study, and our model can be considered as a special open queuing network. The earliest work in this area belongs to Hunt [22] who obtained the maximum possible utilization and the expected number of customers in the system where the stations have exponential service times. Avi-itzhak [3] studied tandem queues with no intermediate queue and with arbitrarily distributed service times at both servers. Konheim and Reiser [23] developed an algorithm to find the steady state probabilities numerically in a system of two queues with finite buffer. Foster and Perros [13] derived exact and approximate bounds for the mean blocking time in queueing networks with exponential service times and finite buffer.

Hillier and Boling [20] have developed an approximate method to calculate the long-run mean output rate and the average number of customers in a system of finite queues in series having exponential or Erlang service time. An analytical approximation method for open networks of queues was proposed by Altiok [1]. His analysis is based on the method of decomposition where the total network is broken down into queues which are analyzed to find the steady state probabilities of the number of customers at each station.

A matrix solution for the steady state joint queue length distribution of two finite queues in tandem was proposed by Wong, Giffin and Disney [37] . Neuts [30]

used transform methods and matrix geometry to solve the problem of two station with finite waiting queue between them.

1.2.3 Production/Inventory Control Systems

The work in this area is diverse, but we will review the related literature only. Love [24] considered an inventory model consisting of two stages coupled together so that the reorder demand of stage 1 becomes the sales of stage 2. He presented this two-stage model as a finite Markov chain, and obtained the state equations. He presented an algorithm for finding the inventory policy that minimizes the expected cost. Barten [4] developed a queueing simulator for determining optimum inventory levels for any sequence of operations with finite storage. Elsayed and Hwang [10] analyzed a two-stage production line system with a buffer storage. Their concern was the reliability and efficiency of the production system. Solberg [34] presented a capacity planning model with stochastic processing time and compared it to a corresponding deterministic model.

1.3 Problem Definition

There is a growing interest among industrial corporations in adopting and applying the JIT doctrine. Therefore, the need has arisen to evaluate and analyze this new approach. The Japanese,and now more and more others, employ the pull system to implement the JIT doctrine, so that a good modeling of the system is timely. Since there are sources of uncertainty in the system due to customer demand fluctuation and production and transportation problems, conclusions drawn from a stochastic model of the pull production system will be especially relevant for real-

life decisions concerning allocation of units between stations and the optimizing of production rates. Such a model also will be of help in ascertaining how the pull system is related to the conventional push system.

In this dissertation we develop a stochastic model of the multi-stage pull system, primarily as a chain of stations in series. We consider the uncertainty in production and transportation time, as well as fluctuation in final demand, using a stochastic model of a multi-station system with finite capacities R_i at stations i (i=1,2,...,N).

We are interesting in analyzing and optimizing the following measures of performance and effectiveness for this model:

- The probability that the final station is out of units
- The mean units at the final station.
- The system responsiveness, measured as the ratio of the effective production rate of the first station to the required demand at the final station.

Although processing and demand times are assumed exponential, the limitation imposed by station capacity will cause the output process not to be poisson. For this reason, closed form solutions for stationary probabilities of the system are not available and approximations and numerical methods, among them a certain separation-of-variable technique traceable to Hunt, are used.

The first chapter introduces the pull production system and gives a survey of the relevant literature.

The mathematical model is derived and solved in the second chapter, and various methods of solution are utilized and developed. System performance is examined and an optimization of resource allocation is presented.

The third chapter presents pull production system variants comprising one-at-a-time service and larger-than-unit lot sizes.

The push production system is defined and modeled in the fourth chapter; here a new method for modeling the traditional push system is introduced, that reduces the Markovian state number

In the fifth chapter a duality phenomenon between the pull and a specially defined push system is presented and discussed.

Tree structures for the pull production systems are introduced and modeled in the sixth chapter.

The seventh chapter presents conclusion and ideas for research extensions.

2 MODELING THE MULTI-STAGE PULL PRODUCTION SYSTEM I

2.1 Introduction

It is a known fact in management that a smaller plant will outperform a larger plant in every important measure of performance [29]. Many Japanese businesses operate large physical facilities that are focused. Focus is accomplished by carving out many small plant activities within a large facility [29]. Each plant is visualized as a series of stations on an assembly lines, whether physically located in series or not [31]. The cellular manufacturing system (CMS) which is described by Black [6] emphasize the same idea. The CMS is composed of manufacturing cells, which process a group or family of parts. The cell has a U-shape so that even one worker can handle it. The cells are linked with Kanban, compatibly with the idea of Just-In-Time material flow. A variety of cell geometries can be used; this research will concentrate on modeling a series of stages in a production system, but will also address tree structure modeling, as presented in Chapter 6.

2.2 Modeling Approach

The production system studied here is a series of N work stations functioning as a pull system. Each work station is assumed to consist of a processing and a

local storage function. In the pull system, stations withdraw units from preceding stations at the required time, in the required quantity. The word "unit" here means production unit, which may be an individual workpiece, assembly, batch, pallet load and so forth. Whatever the system processes and moves as a together will be treated as a unit [34].

As mentioned in the first chapter, Kanban are used to implement the pull idea in a production system. See Figure 2.1 for model illustration.

In the beginning of the process, the manager assigns a few Kanban (units) to each station[1]. When a demand occurs, it will remove one unit[2] from the local storage of the furthest downstream station, the Kanban card is detached, and is sent to the production function of that station, triggering the processing of a replacement unit. In turn, this station will withdraw a unit from the preceding station and so on.

Three points are presented here to illustrate the congruence between the model developed in this thesis and the pull (Kanban) production system:

1. The total number of Kanbans circulating between stations is unchanged. Consequently, by controlling the number of circulating Kanbans and requiring that every unit has a Kanban attached to it, managers can fix the capacity of a station.

2. The movement of Kanbans is triggered by the withdrawal of units from a station by its immediate successor. In other words, a particular station will

[1]We are going to call this value the capacity of the station or just station capacity.
[2]We are assuming the ideal JIT with lot size equal to one. In a later chapter we will discuss the case of greater lot size.

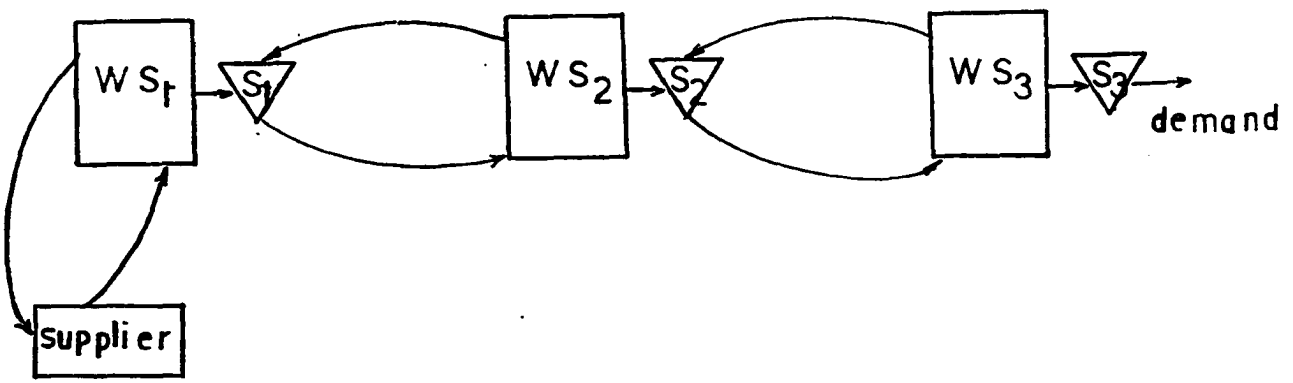


Figure 2.1: Kanban mechanism

produce to replenish what has been withdrawn by the successor one.

3. Circulating of Kanban causes all stations in a production system to be chained together.

2.2.1 Assumptions

The above described pull (Kanban) production system will be stochastically modeled under the following assumptions:

1. The system consists of N serial stations. Station 1 is the first station up stream and station N is the final station downstream.

2. The first station has an infinite supply.

3. Processing times are exponentially distributed with effective processing rate responsive to local storage level.

4. Time between demand is exponentially distributed with fixed rate.

5. The transportation time between stations and warmup time of machines are negligible[3].

6. The stations' processing times are statistically independent of each other.

7. The production capacity at each station is always sufficient [4] (In Section 3.1 we will consider the one-at-a-time model.)

[3]Due to the fact that under the JIT doctrine, the stations are close to each other with limited work-in-process space.

[4]This implies that there are enough machines or servers to handle the work simultaneously.

8. Orders can "cross" [18].

9. Station capacities are fixed

Before we model the system it is appropriate to state the analogy between queueing systems and the pull production systems.

2.2.2 Analogy Between Queueing Systems and Pull Production Systems

It is clear that we are dealing with a system which is related to queueing systems, but with different interpretation of the mathematical variables. The analogy between the two types of systems may be detailed as follows:

1. The system level (number of units) decreases with demand arrival in the pull system while the queue size increases with customer arrival in the queueing system.

2. The "service operation" in queueing corresponds to the process of producing a unit in the pull system.

3. The "service time" in the queueing system corresponds to the time required to replace the unit in the pull system. It is the interval between the time when the unit is withdrawn from the subsequent station and the time a new unit arrives to fill the vacancy caused by the earlier withdrawal.

4. The "queue" in the pull system is the number of the unfilled orders sent by the succeeding station to the subsequent one for units to replace those withdrawn.

5. A "blocking" state in a queuing system corresponds to a "starving/null" state in the pull system. To elaborate on this, the "blocking" means that a unit that finished service at a certain station cannot move forward because the queue is full, and has to block its service until a unit finishes service at the next station. On the other hand, "starving" means that a certain station has unfilled orders that equal its maximum capacity, so that the station will be in the null state, unable to release units to the subsequent station, until it receives a unit from its preceding station (an order is filled).

2.3 The Model

To derive the model, we assume that the system is Markovian. A Markov process can be roughly defined as [9]: "A process whose future probabilistic evolution after any time "t" depends only on the state of the system at time "t", and is independent of the history of the system prior to time "t"." Since we assume that all the production times as well as demand inter-arrival times are exponential, it is clear that the Markovian assumption applies. Moreover, the process has the additional birth-and-death property that the net change across an infinitesimal time interval can never be other than -1, 0, +1. The system has discrete state space which is the amount of units in each station, and continuous time parameter.

2.3.1 State Explanation and Model Formulation

The two-station model will be studied first and that will lead to the general model with N stations. The system is represented by a finite continuous Markov chain with state space $S = \{(n_1, n_2) : 0 \leq n_1 \leq R_1 < \infty, 0 \leq n_2 \leq R_2 < \infty\}$.

Where n_i is the number of units at station i and R_i is the maximum capacity of station i in units. This Markov chain has the following properties:

Homogeneous The probability of moving from one state to another in a fixed unit of time is independent of the starting time. Hence,

$$p_{ij}(t, t+h) = p_{ij}(0, h)$$

where, $p_{ij}(t, t+h) = P(\ x(t+h) = i \mid x(t) = j)$ and $x(t)$ is a Markov chain.

Irreducible There a positive probability that state j can be reached from state i in some finite number of transitions. In other words, all pairs af states intercommunicate.

Ergodic A homogeneous Markov chain is ergodic if

$$\lim_{t \to \infty} p_{ij}(t) = \pi_j$$

independent of i for all j and $\sum_{j=0}^{\infty} \pi_j = 1$, $\pi_j \geq 0$. The vector $\pi = (\pi_0, \pi_1, \ldots)$ is called the stationary distribution.

Recurrent The probability of ultimately returning to state i from state i is equal to 1.

When the Markov chain is irreducible, ergodic and recurrent, then the stationary probabilities are positive. Moreover, as indicated above, these stationary probabilities are limiting probabilities, independently of the initial condition. Since we are interested in the process after a long period of time, we we will concentrate on these stationary limiting probabilities of the process.

Figure 2.2 shows the transition diagram of the system. When the production time at station 1 is exponential with mean $1/\delta_1$, the production time at station 2 is exponential $1/\delta_2$ and the demand rate is Poisson with rate λ, the stationary probabilities of the system satisfy:

For $n_1 = R_1, R_1 - 1, \ldots, 0$ and $n_2 = 0$

$$\begin{aligned}[\delta_1(R_1 - n_1) + \delta_2 * min(R_2 - n_2, n_1)]\, P_{n_1,0} = \\ \delta_1\,(R_1 - (n_1 - 1))\, P_{n_1-1,0} \\ + \lambda P_{n_1,1}\end{aligned} \tag{2.1}$$

For $n_1 = R_1, R_1 - 1, \ldots, 0$ and $n_2 = R_2, R_2 - 1, \ldots, 1$

$$\begin{aligned}[\lambda + \delta_1(R_1 - n_1) + \delta_2 * min(R_2 - n_2, n_1)]\, P_{n_1,n_2} = \\ \delta_1(R_1 - n_1 + 1)P_{n_1-1,n_2} \\ + \delta_2 * min(R_2 - n_2 + 1, n_1 + 1)P_{n_1+1,n_2-1} \\ + \lambda P_{n_1,n_2+1}\end{aligned} \tag{2.2}$$

and the boundary conditions are:

$$\begin{aligned} P_{-1,n_2} &= 0 \\ P_{n_1,-1} &= 0 \\ P_{R_1+1,n_2} &= 0 \\ P_{n_1,R_2+1} &= 0 \end{aligned}$$

We can expand these equations to the general case with N stations as follows:

For $n_i = R_i, R_i - 1, \ldots, 0 \qquad [i = 1, 2, 3, \ldots, N-1]$

and $n_N = 0$

$$[\delta_1 * (R_1 - n_1) + \delta_2 * min(R_2 - n_2, n_1) + \cdots + \delta_N * min(R_N, n_{N-1})]\, P_{n_1,n_2,\ldots,0} =$$

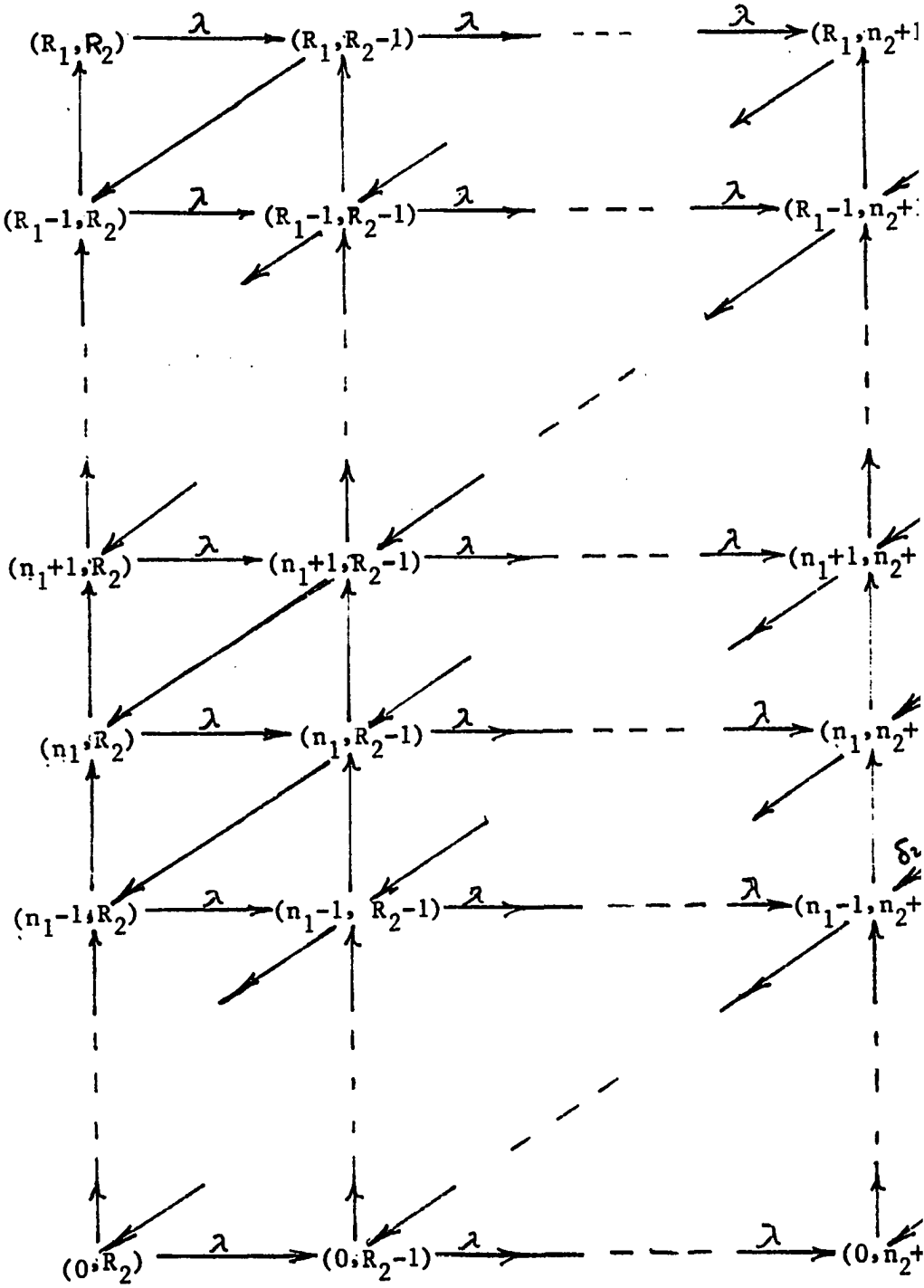
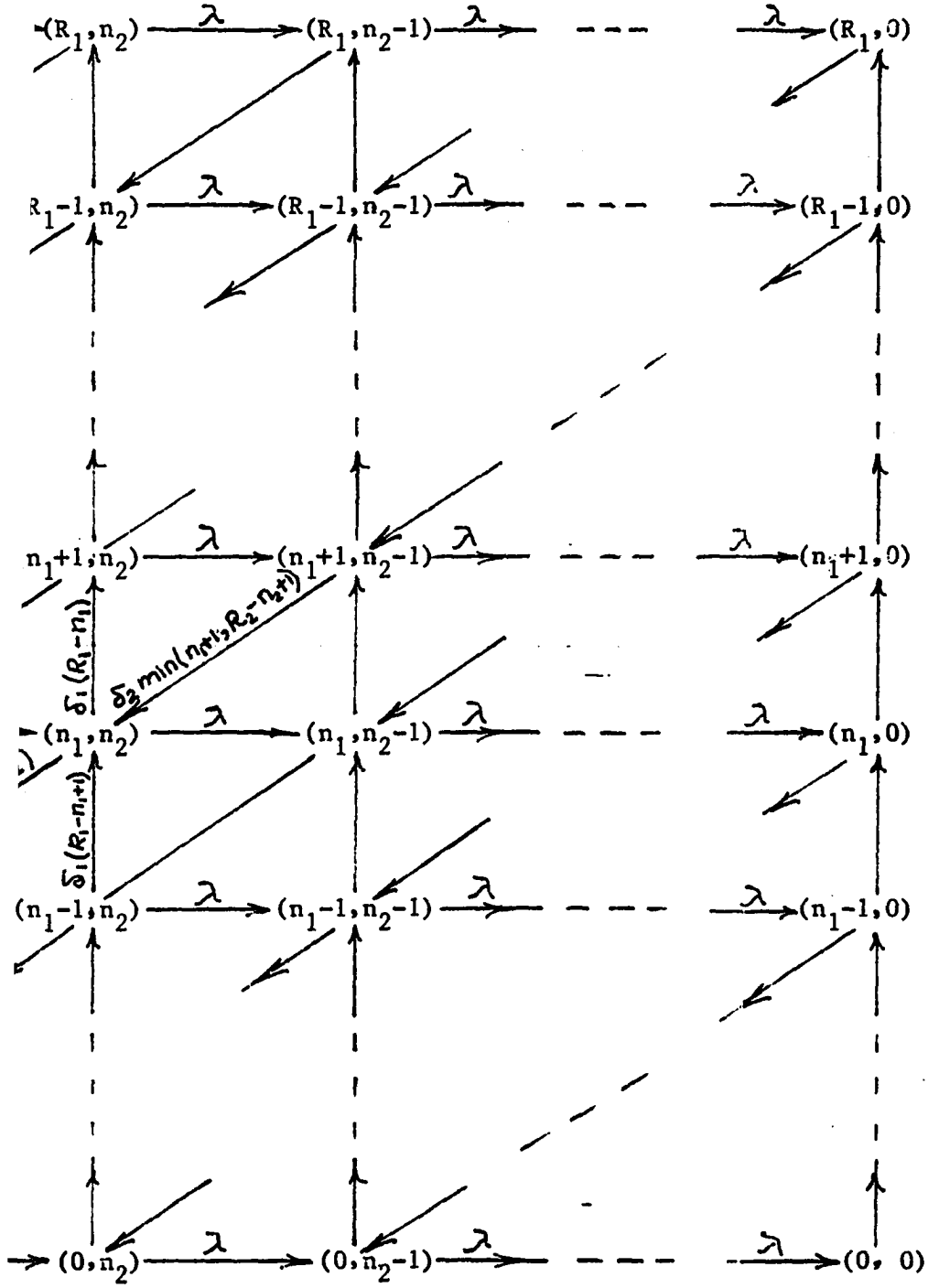


Figure 2.2: Transition diagram for two stations-pull syst


(R1,n2)
λ
(R1,n2-1)
λ
λ
(R1,0)
R1-1,n2)
λ
(R1-1,n2-1)
λ
λ
(R1-1,0)
n1+1,n2)
λ
(n1+1,n2-1)
λ
λ
(n1+1,0)
δ1(R1-n1)
δ2 min(n1+1,R2-n2+1)
(n1,n2)
λ
(n1,n2-1)
λ
λ
(n1,0)
δ1(R1-n1+1)
(n1-1,n2)
λ
(n1-1,n2-1)
λ
λ
(n1-1,0)
(0,n2)
λ
(0,n2-1)
λ
λ
(0, 0)

$$
\begin{aligned}
& \lambda P_{n_1,n_2,\ldots,n_N+1} \\
+\ & \delta_1 * (R_1 - n_1 + 1) P_{n_1-1,n_2,\ldots,n_N} \\
+\ & \delta_2 * min(R_2 - n_2 + 1, n_1 + 1) \\
& P_{n_1+1,n_2-1,\ldots n_N} \\
+\ & \\
& \vdots \\
+\ & \delta_{N-1} * min(R_{N-1} - n_{N-1} + 1, n_{N-2} + 1) \\
& P_{n_1,n_2,\ldots,n_{N-2}+1,n_{N-1}-1,n_N}
\end{aligned}
\tag{2.3}
$$

For $n_i = R_i, R_i - 1, \ldots, 0 \qquad [i = 1, 2, 3, \ldots, N-1]$

and $n_N = R_N, R_N - 1, \ldots, 1$

$$
\begin{aligned}
& [\lambda + \delta_1 * (R_1 - n_1) + \delta_2 * min(R_2 - n_2, n_1) + \cdots + \\
& \delta_N * min(R_N - n_N, n_{N-1})] P_{n_1,n_2,\ldots,n_N} = \\
& \qquad \lambda P_{n_1,n_2,\ldots,n_N+1} \\
& \qquad + \delta_1 * (R_1 - n_1 + 1) P_{n_1-1,n_2,\ldots,n_N} \\
& \qquad + \delta_2 * min(R_2 - n_2 + 1, n_1 + 1) P_{n_1+1,n_2-1,\ldots,n_N} \\
& \qquad + \\
& \qquad \vdots \\
& \qquad + \delta_N * min(R_N - n_N + 1, n_{N-1} + 1) \\
& \qquad P_{n_1,n_2,\ldots,n_{N-1}+1,n_N-1}
\end{aligned}
\tag{2.4}
$$

and the boundary conditions are:

$$
\begin{aligned}
P_{-1,n_2,\ldots,n_N} &= 0 \\
P_{n_1,-1,n_2,\ldots,n_N} &= 0 \\
&\vdots \\
P_{n_1,n_2,\ldots,-1} &= 0 \\
P_{R_1+1,n_2,\ldots,n_N} &= 0 \\
P_{n_1,R_2+1,\ldots,n_N} &= 0 \\
&\vdots \\
P_{n_1,n_2,\ldots,R_N+1} &= 0
\end{aligned}
$$

2.4 Methods of Solution

We solve the above system of equations for the stationary distribution, using various methods which include numerical analysis methods, a method using a related discrete time chain, and a method of variable separation which was presented first by Hunt [22].

2.4.1 Numerical Methods

The transition matrix for this system has a special structure and is sparse. For large systems, this matrix tend to be very large and needs a special computer storage method, so that most of the zero elements can be excluded. As a matter of fact, the size of the transition matrix grows exponentially as the system grows. The transition matrix size can be expressed as follows:

$$\text{matrix size} = MS = \prod_{i=1}^{N} (1 + R_i),$$

where N is the number of stations and R_i the maximum capacity for station i. However, out of the MS^2 elements, fewer than $[(1+N)*MS]$ are nonzero. Given the states and infinitesimal matrix Q, the problem is reduced to the solution of this system of linear equations

$$\begin{aligned} P.Q &= 0 \\ \sum_{i=1}^{N} P_i &= 1 \end{aligned} \tag{2.5}$$

where $P = (P_1, P_2, ..., P_N)$ and P_i is the stationary probability of being in the i^{th} state.

Since any one of the equations 2.5 is redundant, one of them should be replaced by the normalizing condition equation. This system of equations could be solved by the method of Gauss-Seidel,which is an efficient iterative method for solving a large system of linear equations with a high proportion of zero coefficients, and with the diagonal element of the coefficient matrix dominating. However, the Gauss-Seidel method is not guaranteed to converge when that diagonal dominance is "weak". Unfortunately, this is the case with the above set of equations. Nevertheless, the method can be made to lead to convergence after simple modification: The normalizing condition equation is eliminated from the system and P_n is assigned the value of "one" temporarily. The exact solution can then be obtained by a modification to satisfy the normalizing condition equation (method of Nekrasov [11]).

Another exact method of solution is Gauss elimination with partial pivoting, in which a factorization algorithm is used to factor the matrix Q in the form $A = LU$. This method was implemented here using the software package LINPACK. A Fortran computer program was developed to create the transition matrix Q for

two, three, four and five stations, to solve for the stationary probability distribution, and to calculate certain system performance measures detailed below. Both of these numerical methods were implemented, and the results for the two methods were in agreement up to the 8^{th} decimal point at least.

2.4.2 Related Discrete Time Markov Chain Method

In this section we will be looking at the state of the system at certain selected times, and turn our attention away from the original continuous-time process to an embedded discrete time Markov chain process. We are going to look at two discrete time chains determined by the infinitesimal matrix Q.

1. The embedded chain: To illustrate the above, let us consider the original process only at transition times, that is we consider an embedded stochastic process which becomes a discrete time Markov chain. Let us define the matrix R as follows,

$$r_{ij} = \begin{cases} q_{ij}/|q_{ii}| & \text{for } i \neq j \text{ and } q_{ii} \neq 0 \\ 0 & \text{for } i = j \text{ and } q_{ii} \neq 0 \\ 1 & \text{for } i = j \text{ and } q_{ii} = 0 \end{cases}$$

It is clear that R is a stochastic matrix, hence the powers of R generate a discrete time Markov chain which is called embedded Markov chain [33]. There are two problems involved with the embedded chain. First, the embedded chain is periodic for our problem so that, no steady state exists. Second, the embedded chain losses some properties possessed by the original one, such as transition rates, so that its limit distribution, even it does exist, does not

equal the continuous time limit distribution, and must be modified to take transition into account.

2. Another discrete time chain determined by Q was first introduced by Yong [38]. Let the matrix $\bar{P}$ be defined as follows,

   $\bar{P} = I + \frac{Q}{c}$ where,

   $\infty > c > sup_i\{|q_{ii}|\}$

   Hence, $\bar{P}$ is a positive stochastic matrix, and we may consider the discrete-time Markov chain corresponds to $\bar{P}$. It has been proved [33][5] that the original continuous time chain and the generated one have the same limiting distribution. That is,

$$\lim_{n\to\infty} \bar{P_{ij}}(n) = \lim_{t\to\infty} P_{ij}(t) = \pi_j \text{ and } \sum_{j=0}^{\infty} \pi_j = 1.$$

A computer program was written to compute the limiting distribution by taking the n^{th} power of the matrix $\bar{P}$. The stopping criterion used was the Cauchy criterion, namely, the calculations stops when $max_i|p_i^{(n+1)} - p_i^{(n)}| < \epsilon$. To speed up the process, the program computes P^2, P^4, P^8, P^{16} and so on. The program was used for different values of the station capacity and number of stations and the results were identical to those achieved by the numerical methods mentioned in the previous section. The convergence was achieved in most cases between P^{256} and P^{1024}, and, due to the vectorization process, which is very effective in matrix operations, the CPU time was not excessive. Although the above method seemed to work well, it is not to be recommend for solving for the limiting distribution, because serious round-off error can occur. One could also use the Gause-Seidel stepping method

[5]The solution to $X = X\bar{P} \Rightarrow X = X(\frac{Q}{c} + I) \Rightarrow 0 = X\frac{Q}{c} \Rightarrow 0 = XQ$.

on the discrete time Markov chain, in which the Markov chain recursive equation $\pi^{k+1} = \pi^k \bar{P}$ is used to solve for the steady state probability vector π [6].

2.4.3 Separation of Variables Method for the Back-Order Case

In this section, a separation of variables method is used to solve the set of multi-variable homogeneous difference equations with constant coefficients corresponding to the stationary probability equations. Hunt used a similar method to solve a sequential queueing problem. To apply this method, we will relax the assumption that the demand stops when the final station runs out of units. We will assume the demand will continue and a backorder policy will be utilized. Under this new assumption, the station capacity levels can be negative. To illustrate the method, a two-stage system with capacity $R_i = 2 \quad , \quad i = 1,2$ is used. The stationary probabilities of the system satisfy:

For $n_2 = 0, -1, \ldots, -\infty$

$$(\lambda + 2\delta_2)P_{2,n_2} = \lambda P_{2,n_2+1} + \delta_1 P_{1,n_2} \tag{2.6}$$

For $n_2 = 1, 0, -1, \ldots, -\infty$

$$\begin{aligned} (\lambda + \delta_1 + \delta_2)P_{1,n_2} &= \lambda P_{1,n_2+1} + 2\delta_1 P_{0,n_2} + 2\delta_2 P_{2,n_2-1} \\ (\lambda + 2\delta_1)P_{0,n_2} &= \lambda P_{0,n_2+1} + \delta_2 P_{1,n_2-1} \end{aligned} \tag{2.7}$$

And the boundary coditions are:

$$\begin{aligned} \lambda P_{2,2} &= \delta_1 P_{1,2} \\ (\lambda + \delta_2)P_{2,1} &= \lambda P_{2,2} + \delta_1 P_{1,1} \end{aligned}$$

[6] The solution for the discrete time may converge after fewer iterations [16].

$$\begin{aligned}(\lambda+\delta_1)P_{1,2} &= \delta_2 P_{2,1} + 2\delta_1 P_{0,2} \\ (\lambda+2\delta_1)P_{0,2} &= \delta_2 P_{1,1} \end{aligned} \tag{2.8}$$

Now apply the change of variable: $n = 2 - n_2$ to the above stationary equations, and obtain:

For $n = 2, 3, \ldots, \infty$

$$(\lambda+2\delta_2)P_{2,n} = \lambda P_{2,n-1} + \delta_1 P_{1,n} \tag{2.9}$$

For $n = 1, 2, 3, \ldots, \infty$

$$\begin{aligned}(\lambda+\delta_1+\delta_2)P_{1,n} &= \lambda P_{1,n-1} + 2\delta_1 P_{0,n} + 2\delta_2 P_{2,n+1} \\ (\lambda+2\delta_1)P_{0,n} &= \lambda P_{0,n-1} + \delta_2 P_{1,n+1}, \end{aligned} \tag{2.10}$$

for which the boundary conditions become

$$\begin{aligned}\lambda P_{2,0} &= \delta_1 P_{1,0} \\ (\lambda+\delta_2)P_{2,1} &= \lambda P_{2,0} + \delta_1 P_{1,1} \\ (\lambda+\delta_1)P_{1,0} &= \delta_2 P_{2,1} + 2\delta_1 P_{0,0} \\ (\lambda+2\delta_1)P_{0,0} &= \delta_2 P_{1,1} \end{aligned} \tag{2.11}$$

Using the vector notation:

$$\vec{P}_n = \begin{vmatrix} P_{2,n} \\ P_{1,n} \\ P_{0,n} \end{vmatrix}$$

equations 2.9 and 2.10 can be put in the form,

$$A_{-1}\vec{P}_{n-1} + A_0\vec{P}_n + A_1\vec{P}_{n+1} = 0 \tag{2.12}$$

Now, invoking the separation-of-variables assumption $P_{n_1,n} = c_{n_1} b^n$, one obtains,

$$A_{-1}\vec{C} + bA_0\vec{C} + b^2 A_1\vec{C} = 0 \tag{2.13}$$

$$\text{or } D\,\vec{C} = 0, \tag{2.14}$$

where D is a matrix of the same dimension as the A matrices, and is equal to

$$\begin{pmatrix} (\lambda + 2\delta_2)b - \lambda & -\delta_1 b & 0 \\ -2\delta_2 b^2 & (\lambda + \delta_1 + \delta_2)b - \lambda & -2\delta_1 b \\ 0 & -\delta_2 b^2 & (\lambda + 2\delta_1)b - \lambda \end{pmatrix}$$

and $\vec{C}$ is given by

$$\vec{C} = \begin{vmatrix} c_2 \\ c_1 \\ c_0 \end{vmatrix}.$$

As $\vec{C} > 0$, the matrix D is singular, determinant

$$\begin{aligned} &- \lambda^3 b^4 + \lambda^2(d_1 + d_2 + d_{12})b^3 - \lambda(d_1 d_2 + d_1 d_{12} + d_2 d_{12})b^2 \\ &+ (d_1 d_2 d_{12} + 4\delta_1\delta_2\lambda)b - 2\delta_1\delta_2(d_1 + d_2) = 0 \end{aligned} \tag{2.15}$$

equal to zero, where

$$\begin{aligned} d_1 &= \lambda + 2\delta_1 \\ d_2 &= \lambda + 2\delta_2 \\ d_{12} &= \lambda + \delta_1 + \delta_2 \end{aligned} \tag{2.16}$$

The symbolic machine "MACSYMA" was used to solve for the roots of this equation, which yielded

$$b_1 = \frac{\lambda}{d_{12}}$$

$$
\begin{aligned}
b_2 &= \frac{\lambda\sqrt{\lambda^2 + 4(\delta_1+\delta_2)\lambda + 4(\delta_1-\delta_2)^2} + \lambda^2 + 2(\delta 1+\delta_2)\lambda}{8\delta_1\delta_2} \\
b_3 &= \frac{-\lambda\sqrt{\lambda^2 + 4(\delta_1+\delta_2)\lambda + 4(\delta_1-\delta_2)^2} + \lambda^2 + 2(\delta 1+\delta_2)\lambda}{8\delta_1\delta_2} \\
b_4 &= 1
\end{aligned}
\tag{2.17}
$$

Note that all roots are positive, including b_3 since,

$\sqrt{(}\lambda^2 + 4(\delta_1 + \delta_2)\lambda + 4(\delta_1 - \delta_2)^2) = \sqrt{((}\lambda + 2(\delta_1 + \delta_2))^2 + 4(\delta_1 - \delta_2)^2 - 4(\delta_1 + \delta_2)^2)$

$= \sqrt{((}\lambda + 2(\delta_1 + \delta_2))^2 - 4\delta_1\delta_2) < (\lambda + 2(\delta_1 + \delta_2))$

Proceeding analogously to the development in [16], [24], and discarding the root $b_4 = 1$ for reasons of convergence, the general solution for the stationary probabilities are given by the linear combinations of the n^{th} power of the roots excepting b_4:

$$P_{n_1,n} = a[(c_{n_1})_1 b_1^n + (c_{n_1})_2 b_2^n + (c_{n_1})_3 b_3^n]. \tag{2.18}$$

Here, a is a normalizing factor. The number of the constants c_{n_1}'s are then reduced by substituting this solution in equations 2.9 and 2.10. The number of constants are reduced to three and the boundary constraints 2.11 are used to solve for the constants. An arbitrary boundary state probability is chosen to be equal to the normalizing factor a, and, finally, substitution of 2.18 into the normalizing equation $\Sigma n_1 \ \Sigma n \ P_{n_1,n} = 1$ yields the value of a.

Discussion Looking ahead to the next section, we look briefly at the two performance measures I_1 and I_2 below, for which the backorder analysis allows

specialized conclusions.

$$I_1 = \sum_{n_1=0}^{2} n_1 X_{n_1} \qquad where \;\; X_{n_1} = \sum_{n}^{\infty} P_{n_1,n}$$

$$I_2 = \sum_{n=0}^{\infty} n X_n \qquad where \;\; X_n = \sum_{n_1=0}^{2} P_{n_1,n}$$

Substituting for the values of the joint probabilities, and after some algebraic manipulation, we get

$$I_1 = 2 - \frac{\lambda}{\delta_1} \tag{2.19}$$

but the expression for I_2 does not turn out as simple as I_1. I_1 suggests the following considerations:

- Since the roots of the detrimental equation are summed in an infinite geometric series to compute the joint probabilities, each should be less than one. To get condition for summability, we solve for $B_3 = 1$ (orB_2), and obtain $\lambda = \frac{2\delta_1\delta_2}{\delta_1+\delta_2}$. Hence, $b_3 < 1 \rightarrow \lambda < \frac{2\delta_1\delta_2}{\delta_1+\delta_2}$. Furthermore, $\lambda < \frac{2\delta_1\delta_2}{\delta_1+\delta_2} \rightarrow E_1 = 2 - \frac{\lambda}{\delta_1} > \frac{2\delta_1}{\delta_1+\delta_2}$. Thus $b_3 < 1$ also guarantees that I_1 is positive.

- The critical rate of demand is equal to $\frac{2\delta_1\delta_2}{\delta_1+\delta_2}$, and the system will be unstable if $\lambda > \frac{2\delta_1\delta_2}{\delta_1+\delta_2}$ [7]

- If the production rates are equal, i.e., $\delta_1 = \delta_2$ the critical condition for stability will be $\lambda = \delta_1 = \delta_2$.

- In equation 2.19 the value "2" is the capacity of station #1. As is known in the analogous queueing, relation (2.19) also applied for an isolated station.

[7]The problem of instability does not occur for a finite system.

Thus, if back ordering is allowed at the final stage, the first stage will behave in autonomous fashion with regard to I_1.

2.5 Performance Measures

To evaluate a multi-stages production system well defined performance measures are needed. A survey of different production measures from the literature is presented here.

Maximum utilization [7]. This is defined as the fraction of time that the first station is free to service incoming units.

Steady State Mean Output Rate [20]. This is expressed as

$$\delta_N \sum_{j \in B} P_j$$

where B is the set of states corresponding to the last station being busy, and δ_N is the mean service rate of this station.

Production Capacity [7], or maximum arrival rate for which the system is stable.

System Responsiveness . This is defined as the probability that the system can successfully meet an operational demand when operated under specified conditions. System responsiveness is a term used in a broad context to reflect the system performance, and may be expressed differently depending on the specific application.

Because of the special nature of the pull system, and because it is a finite system, the above measures need to be redefined. Therefore the performance measures for the pull system will be defined as follows:

- System responsiveness, expressed as

$$SE = \frac{\delta_1 * (Prob \ \ n_1 < R_1)}{\lambda} \tag{2.20}$$

which measures the system response to the demand by comparing the demand rate λ (downstream) to the effective production rate of the first station (upstream)

- Mean number of units at the last station, expressed as

$$I_N = \sum_{n_N=0}^{R_N} n_N \sum_{n_{N-1}=0}^{R_{N-1}} \cdots \sum_{n_1=0}^{R_1} P_{n_1,n_2,\ldots,n_N} \tag{2.21}$$

which measures the ability of the system to response to the demand.

- Probability that the last station is out of stock, expressed as

$$P_{out} = \sum_{n_1=0}^{R_1} \cdots \sum_{n_{N-1}=0}^{R_{N-1}} P_{n_1,n_2,\ldots,0} \tag{2.22}$$

which measures the system success in being just-in-time

2.6 Study of Pull System Behavior

In this section we will study the effect of system parameters on the above system performance measures. See Figures 2.3 , 2.4.

2.6.1 System Responsiveness

2.6.1.1 Effect of Station Capacity In order to examine the effect of station capacity on the system responsiveness, several computer runs were made

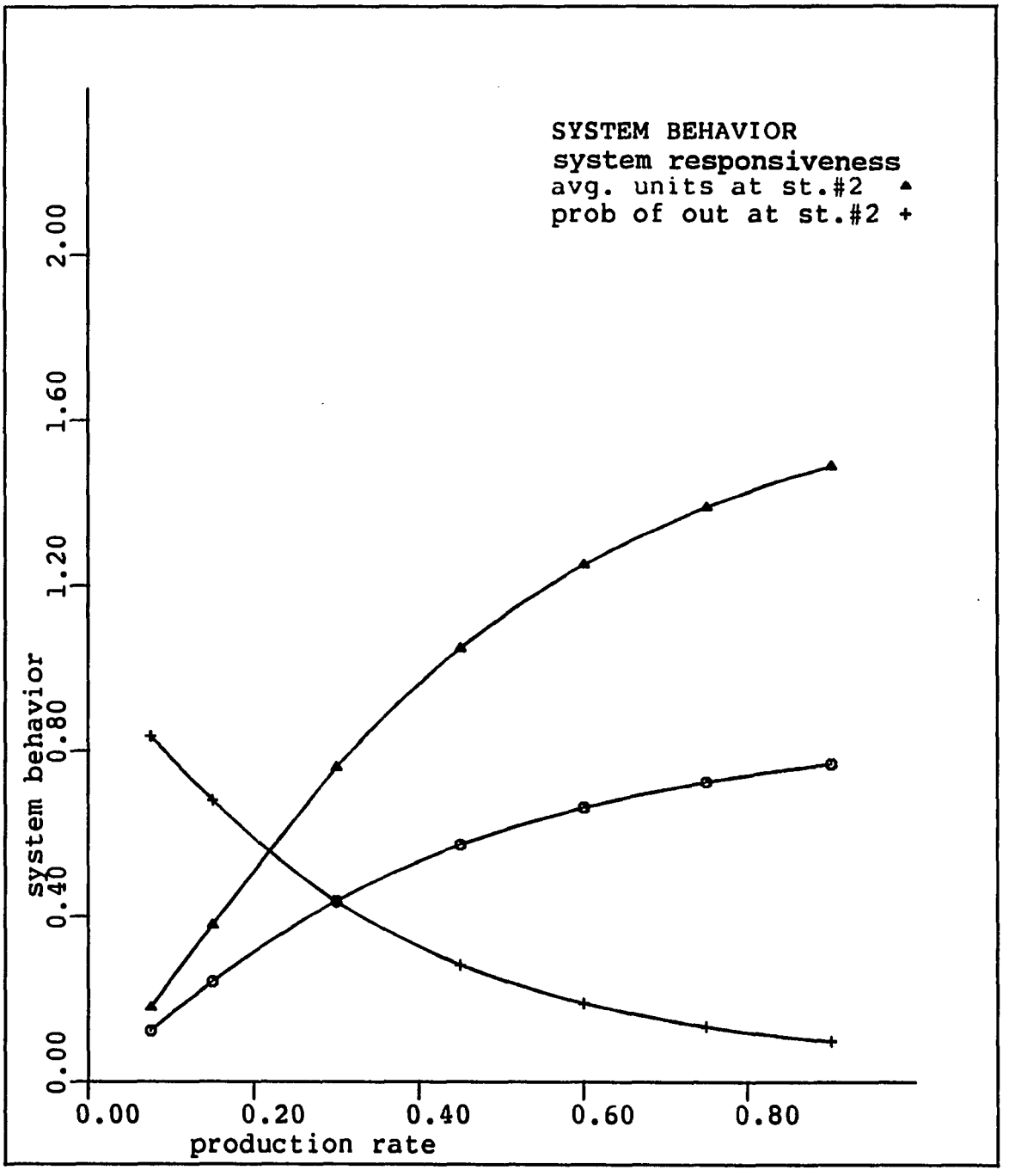


Figure 2.3: Effect of production rates

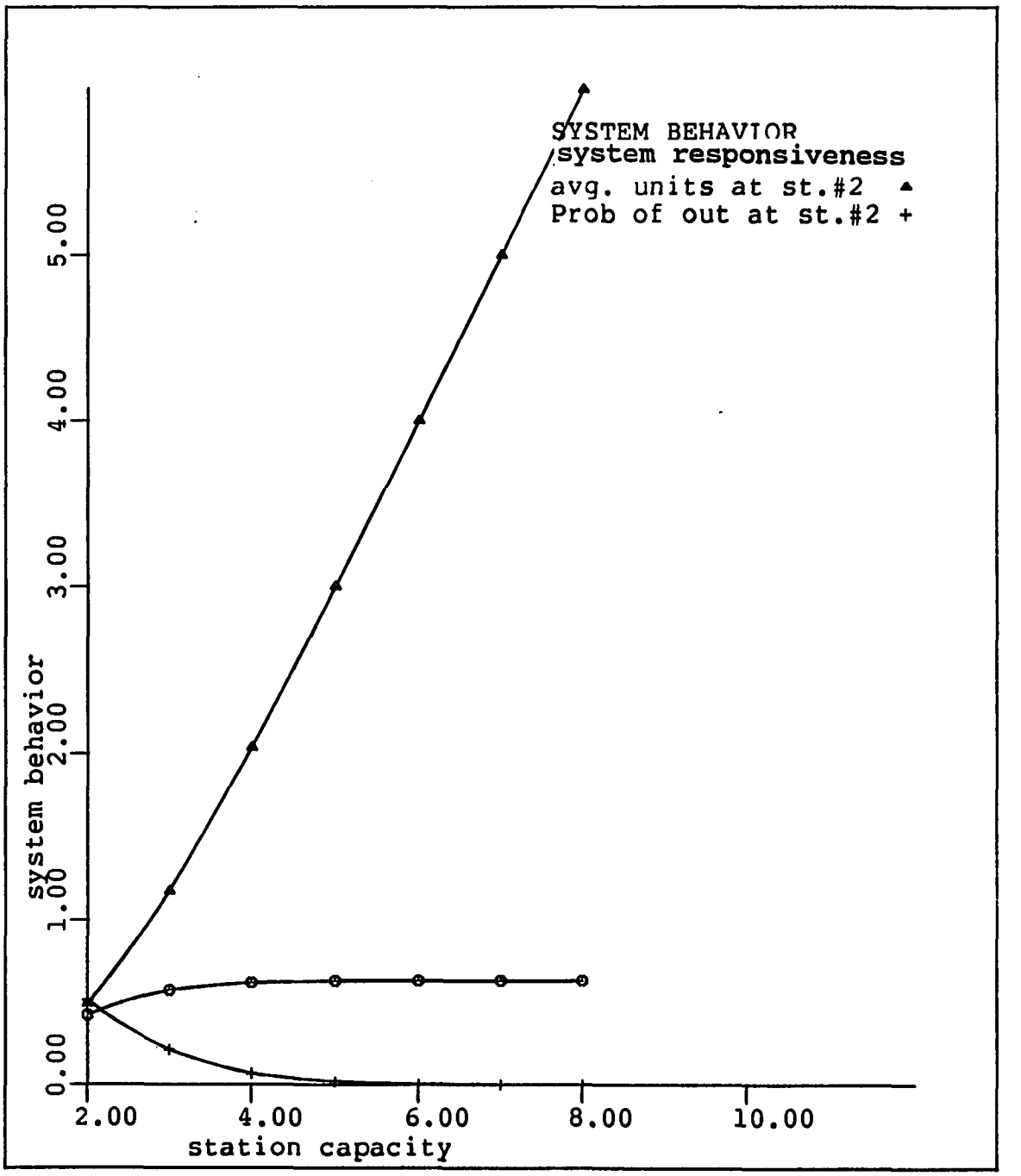


Figure 2.4: Effect of station capacity

for capacity equal to one, two, up to six. The results are shown in Figure 2.5. The curves in Figure 2.5 show the system responsiveness measured on the ordinate and the station capacity along the abscissa. The curves are almost parallel, differing in displacement as functions of the production rates. The curves shown in the figure appear to be approaching an equilibrium condition which is a function of the production rates, and which is approximately achieved at station capacities of two or three units. This phenomenon agrees with the JIT concepts that this system is modeling.

2.6.1.2 Effect of Demand Rate Figure 2.6 shows the effect of increasing the demand rate on the system responsiveness . As expected, increasing demand rate decreases the system responsiveness (see equation 2.20). The relationship is not linear and it tends to level off. The reason is: as the demand rate increases (the denominator), the probability of having less than maximum capacity at station one increase (the numerator). That is, an increase in denominator is counterbalanced by an increase in numerator.

2.6.2 Average Number of Units at Final Station

2.6.2.1 Effect of station Capacity Figure 2.7 shows that increasing the station capacity causes the average units at the final station to increase linearly. This is an expected results because adding more capacity means keeping a high buffer.

2.6.2.2 Effect of Production Rate The curves in Figure 2.8 show that increasing the production rates causes an increase in the average units at the final

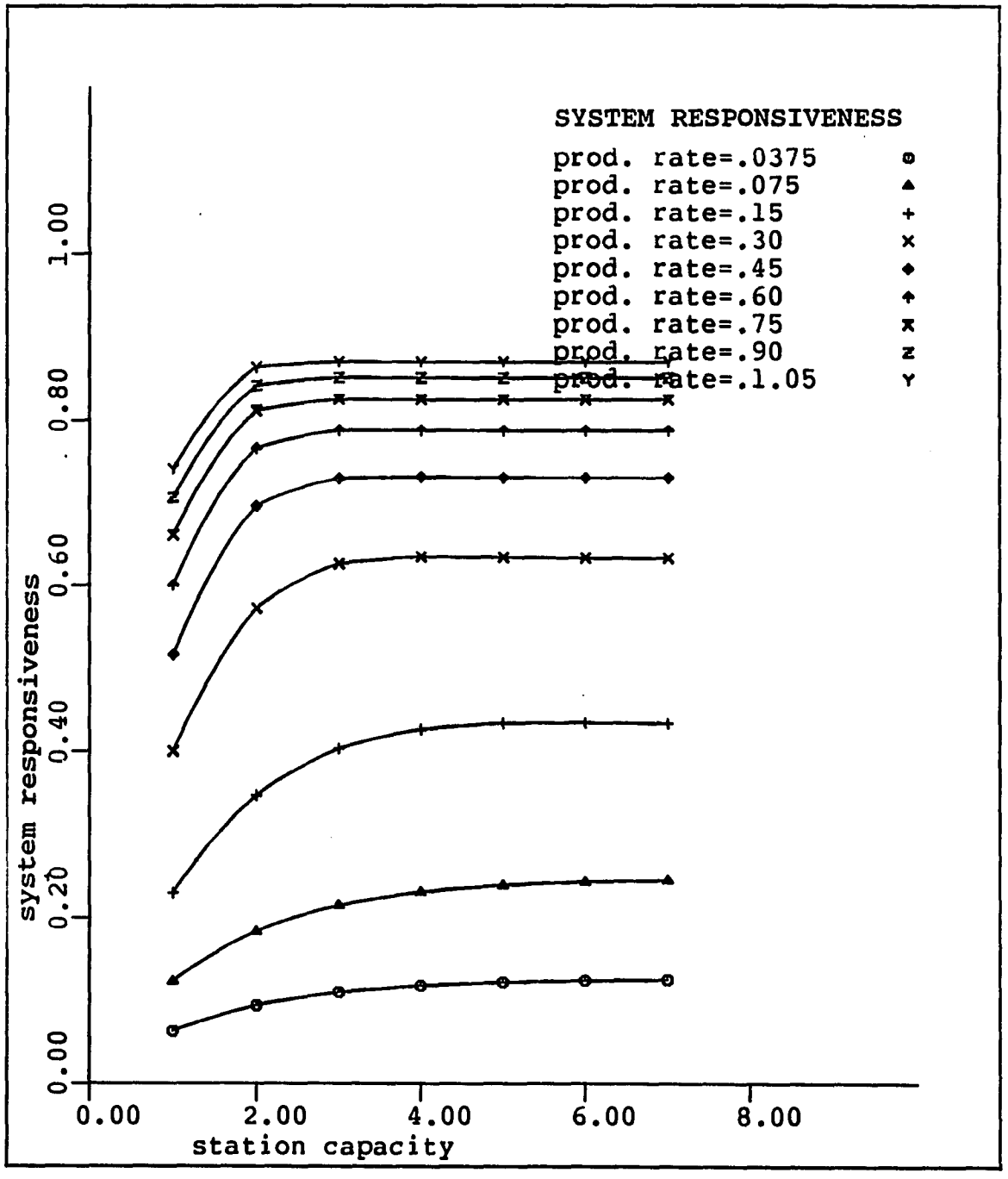


Figure 2.5: System responsiveness vs station capacity

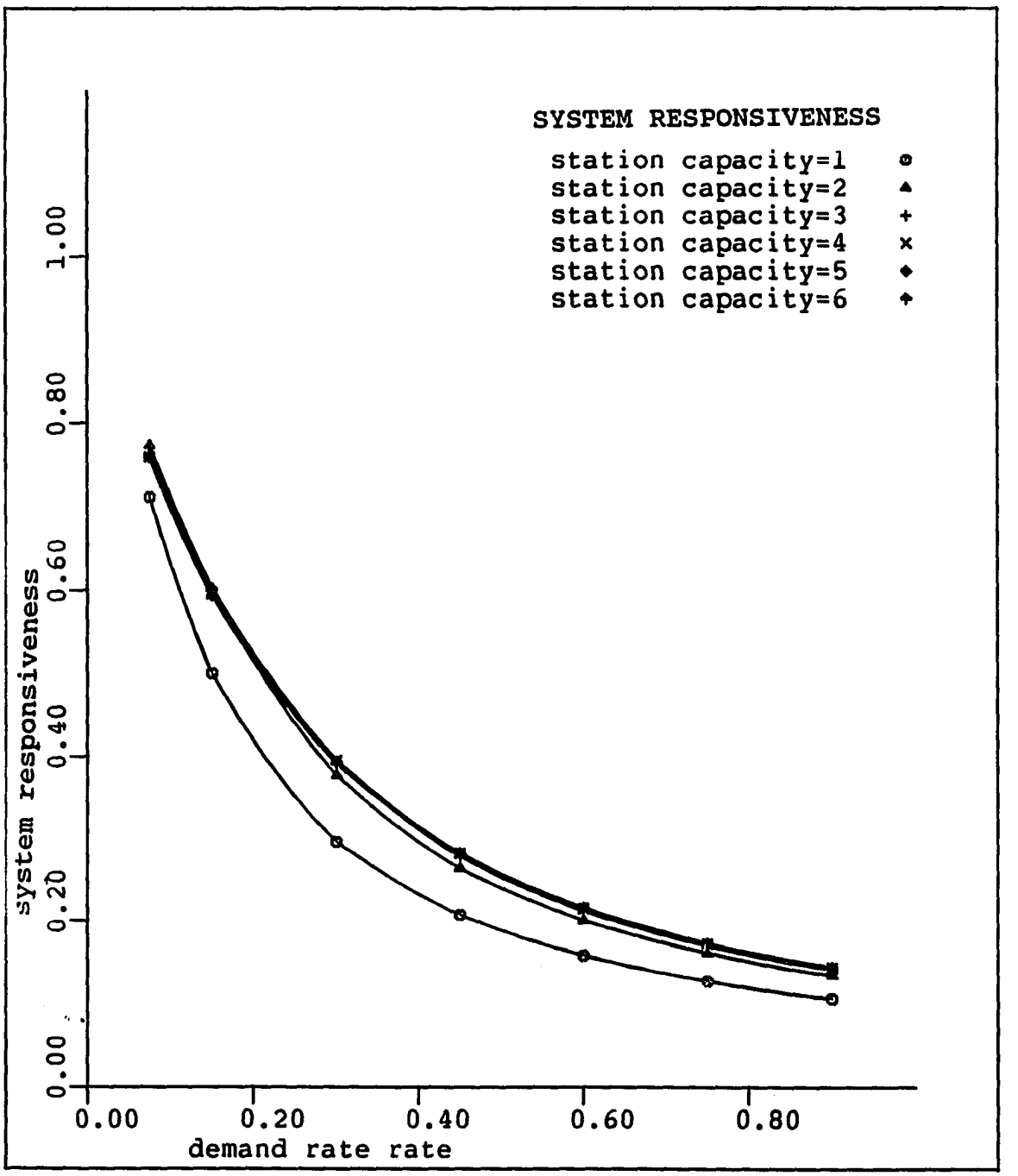


Figure 2.6: System responsiveness vs demand rate

station. At production rates of two times the demand rate or more, however, the increase of production rate has little effect on the average units at the final station. This phenomenon agrees with the concept of JIT which this system is modeling: "The subsequent station will not produce units if it is not needed by the succeeding station."

2.6.3 Probability that the Final Station is Out of Units P_{out_N}

2.6.3.1 Effect of station capacity As for the previous performance measures, Figure 2.9 presents a set of curves with P_{out_N} on the ordinate and the station capacity on the abscissa. It is immediately apparent that the probability that the final station is out of stock decreases (improves) as the production rates and the station capacities increase. This is to be expected because the higher production rate and/or the higher station capacity prevent the last station from "starving". At station capacity of three or more, and with production rate equal to or greater than the demand rate, the situation is entirely different. The station capacity has little effect on P_{out_N} and the curves converge.

2.7 Resource Allocation

In the previous section, we have studied the system in the light of its performance measures, without emphasis on resource allocation, assuming equal station capacity and equal production rates. In this section we will deal with resource allocation.

The important question facing any production planner is how to allocate resources for different stations. The answer is not simple, particularly in view of

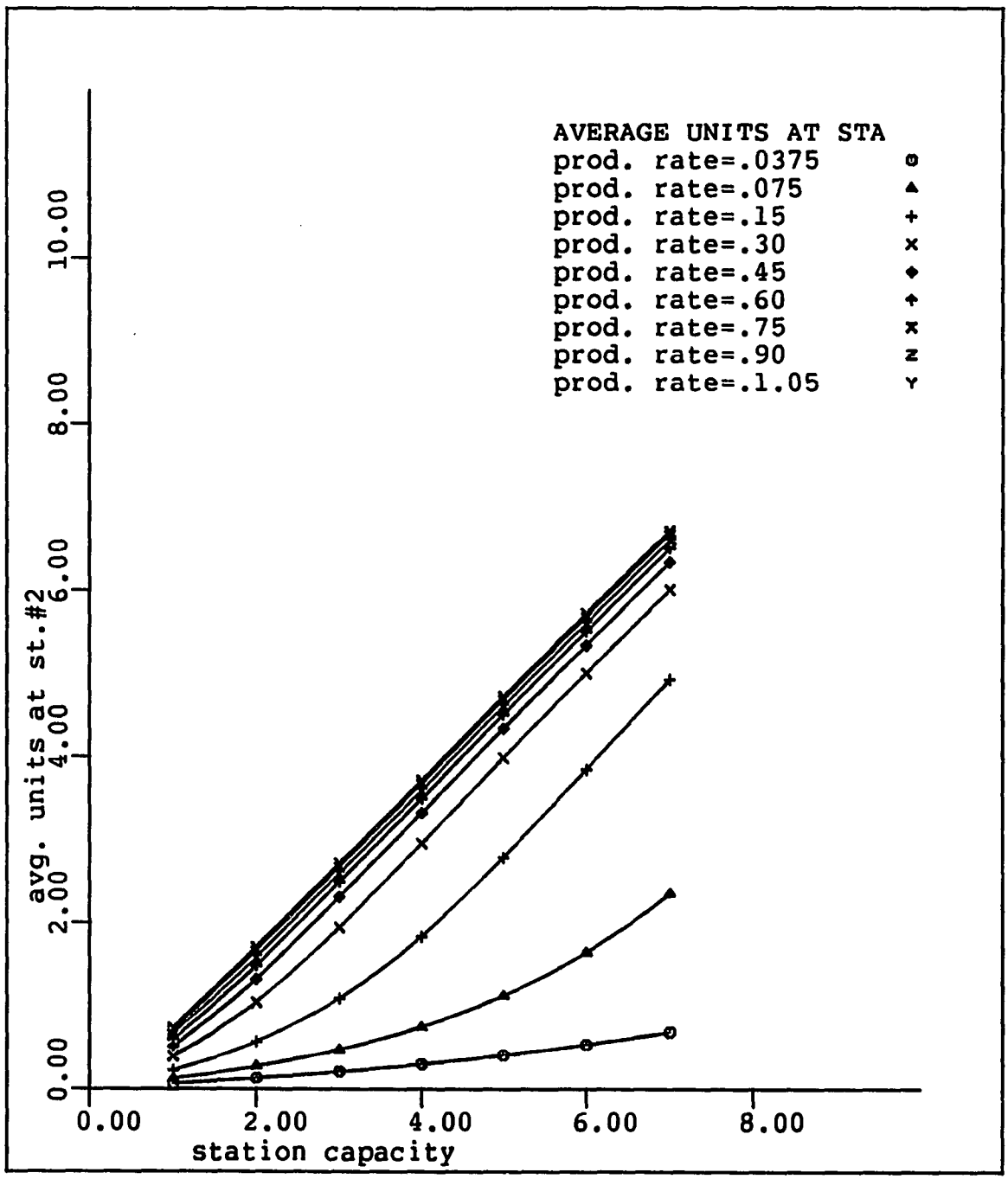


Figure 2.7: Average units vs station capacity

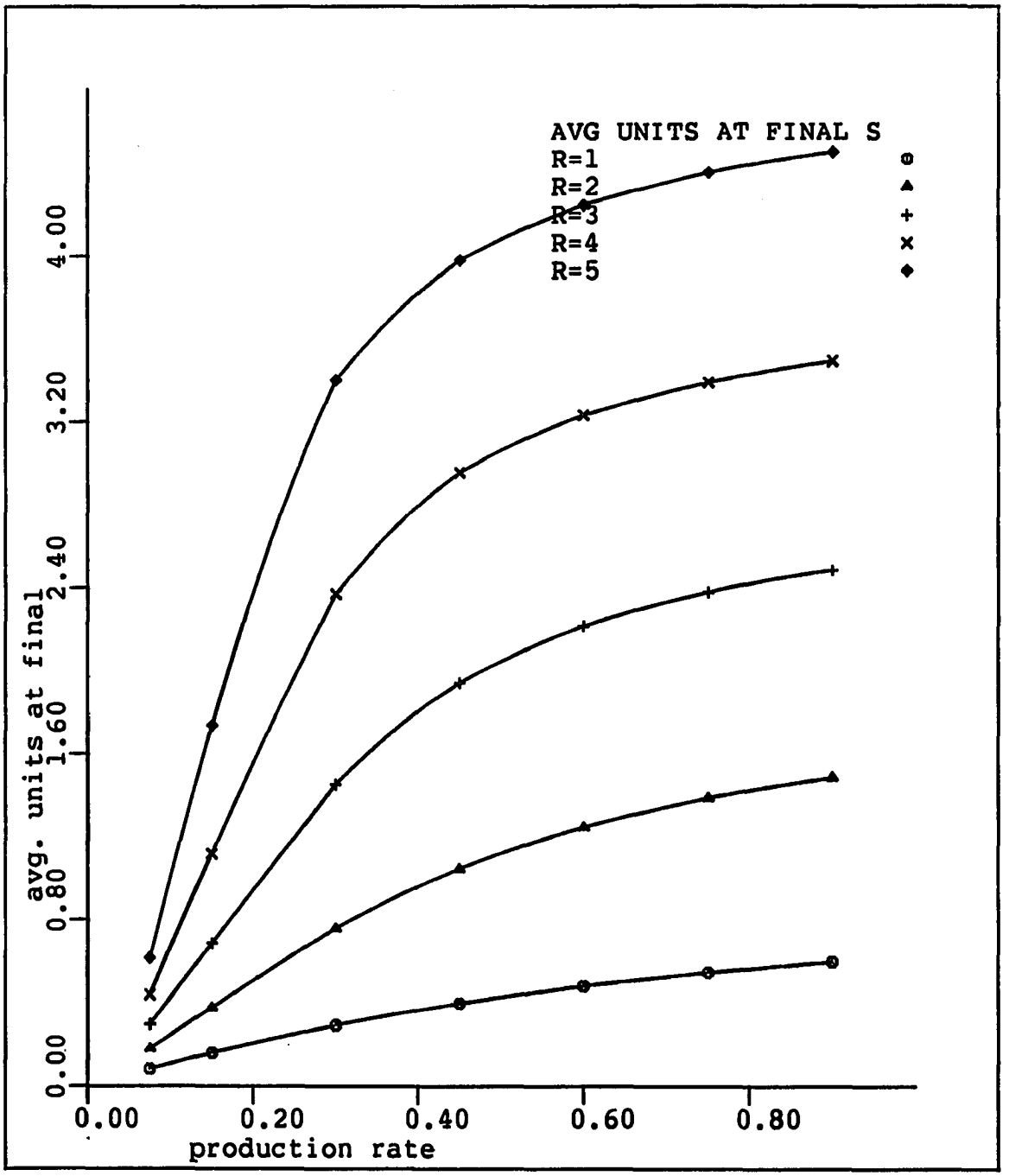


Figure 2.8: Average units vs production rate

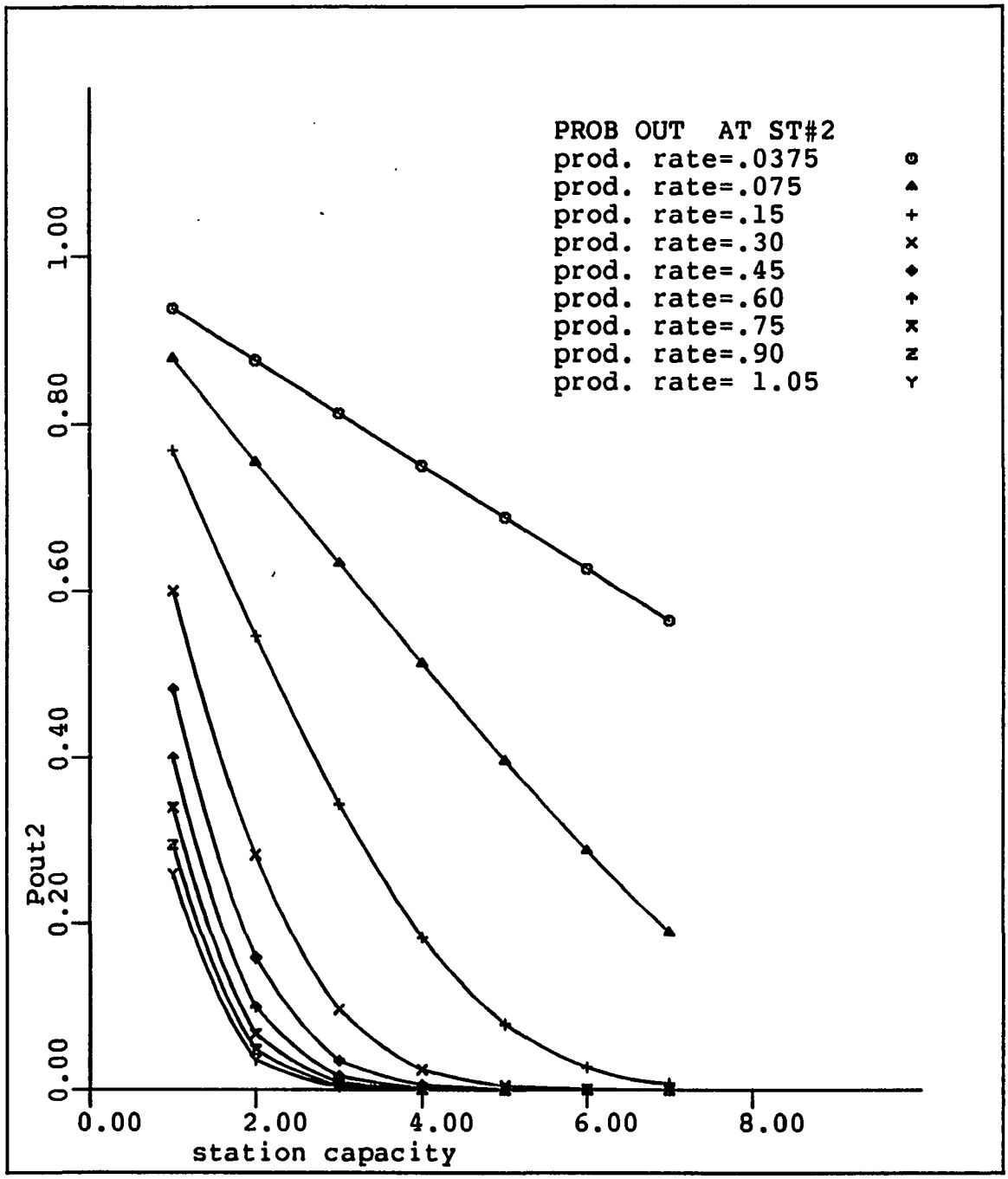


Figure 2.9: Pout vs station capacity

uncertainty about production and demand times. In an attempt to answer this question, we analyze the system and study its behavior in two ways:

1. The effect of allocating work elements and personnel, i.e., production rates, among stations.

2. The effect of allocating (Kanban), i.e., station capacites, among the stations.

This study has been done using the probability that the final station is out of stock,as the performance measure. The reason for this choice is our conviction that the most important concern for the JIT doctrine is the availability of the product when it is needed.

Computer programs were written to perform preliminary tests for the above two types of analysis, and the following two optimization problems were formulated.

2.7.1 Optimal allocation of station capacity(Kanban number)

The problems of finding a station-capacity allocation minimizes P_{out_N} is expressible, for an arbitrary number N of stations, in the form.

$$\begin{aligned} \text{Min} \quad & P_{out_N} = \sum_{n_1=0}^{R_1} \cdots \sum_{n_{N-1}=0}^{R_{N-1}} P_{n_1,n_2,\ldots,0} \\ \text{such that} \quad & \sum_{i=1}^{N} R_i = C_1. \end{aligned}$$

The constraint keeps storage cost fixed in a sense.

The problem, in the given form, has been addressed under the further conditions:

$$\begin{aligned} \delta_i &= \delta \qquad\qquad i = 1,2,3,\ldots,N \\ \delta/\lambda &= C_2. \end{aligned}$$

The first condition equates all production rates, to highlight and isolate the effect of the station capacities (Kanban allocation), while the second condition exploits the homogeneity of the model, in the parameters λ and δ. The minimization was done using enumeration for two , three and four stations. All the possible combinations of station capacities were tested for each chosen value of C_1 and C_2, yielding the optimum value of P_{out_N}, together with optimum allocation of capacities R_i. It turns out that there is a value of δ/λ above which the optimum allocation of R_i exhibits a "funnel pattern", with the higher capacities allocated downstream, whereas the allocation is essentially uniform otherwise.

To fine-tune the value of δ/λ (i.e, C_2) at which the optimum allocation of R_i changes from uniform to "funnel", we followed this simple process:

1. Identify two values (the interval) of the ratio δ/λ between which the change in character of optimum allocation occurs.

2. Divide this interval by five to get four intermediate values, and again check the optimum allocation at each and identify a new and smaller interval, nested in the previous one.

3. Repeat step 2.

4. Stop when the interval reaches a small value, ε, or less.

For $N \leq 5$, an approximate relation determining the value of the above critical δ/λ is:

$$\delta/\lambda = c * N^a * \bar{R}^b$$

where N = number of stages, $\bar{R} = \frac{\sum_{i=1}^{N} R}{N}$, and a, b, c are constants. A regression linear model was developed to calculate the constants, utilizing a large number of (C_1, C_2, N) combination with $N \leq 4$, and verified using several combinations with $N = 5$. Tables 2.1, 2.2 and 2.3 provide numerical illustration of the optimization. The values of the constant were found to be

$$a = -.849, \quad b = -1.908, \quad c = 28.576$$

From the above results, it may be seen that "underdesigned" production systems (low δ/λ), with correspondingly high values of P_{out_2}, do not feature the funnel pattern; on the other hand,"overdesigned" production systems (high δ/λ), with correspondingly low values ofP_{out_2}, do feature it.

It may also be noted that, in actual manufacturing contexts, storage expense often increases in the downstream direction, so that a *weighted* sum of the capacities R_i, with weights increasing in the downstream direction, might provide a more realistic constraint. One can expect that such a modification would pull the optimum funnel pattern back toward uniformity.

2.7.2 Optimal allocation of production capacity

The problem of finding a production-rate allocation minimizing P_{out_N} is expressible, for an arbitrary number N of stations, in the form:

$$\begin{aligned} \text{Min} \quad & P_{out_N} = \sum_{n_1=0}^{R_1} \cdots \sum_{n_{N-1}=0}^{R_{N-1}} P_{n_1,n_2,\ldots,0} \\ \text{such that} \quad & \sum_{i=1}^{N} \delta_i = C_3 \end{aligned}$$

The problem, in the given form, has been addressed under the further conditions

Table 2.1: Optimum number of Kanban for two stages

C_1	C_2	R_1	R_2	P_{out_2}	I_2	SE
4	5.0	1	3	.0154	2.6382	.9846
4	3.0	2	2	.0481	1.659	.8416
4	1.0	2	2	.0026	3.9234	.5728
6	3.0	2	4	.0026	3.9234	.8729
6	2.0	2	4	.0143	3.3091	.8232
6	1.5	3	3	.0343	2.3015	.7297
6	1.0	3	3	.0916	1.9411	.6259
8	2.0	3	5	.0008	4.4660	.7950
8	1.6	3	5	.0027	4.2979	.7560
8	1.0	4	4	.0238	2.9545	.6342
8	0.7	4	4	.0757	2.4532	.5324

Table 2.2: Optimum number of Kanban for three stages

C_1	C_2	R_1	R_2	R_3
7	10.	1	1	5
7	2	2	2	3
7	.66	2	3	2
7	0.5	2	3	2
9	10.	1	2	6
9	2.0	2	3	4
9	1.0	3	3	3
9	0.5	3	3	3
12	2.0	3	3	6
12	1.0	3	4	5
12	.66	4	4	4
12	0.5	4	4	4

Table 2.3: Optimum number of Kanban for four stages

C_1	C_2	R_1	R_2	R_3	R_4	P_{out_4}	I_4	SE
8	10.	1	1	2	4	.0000	3.38947	.9996
8	2.0	2	2	2	2	.1081	1.4656	.7620
8	1.0	2	2	2	2	.3365	0.9463	.5377
8	0.5	2	2	2	2	.62677	0.4577	.2966

$$R_i = R \qquad i = 1, 2, \ldots, N$$
$$\bar{\delta}/\lambda = C_4, \qquad \text{or.} \lambda = C_3/C_4 N.$$

Remark: The constraint is compatible with the philosophy of JIT manufacturing. To control the cycle time of a station and consequently the production rates, JIT calls for adding or removing workers at the station, and when those worker are multi-function [8], they may be moved from one station to another, changing the production rate. Thus, having the sum of production rates constant suggests constant number of worker and constant labor cost.

The decision variables here are the production rates, which are continuous variables, as opposed to the number of Kanban units which are discrete variables in the previous problem. The algorithm chosen for solving this optimization problem is the Hook and Jeeves pattern search method [21]. It is a search technique for solving unconstrained nonlinear minimization problems with multi-variables. The technique consists of searching the objective function locally and then moving in a direction favorable for reducing the objective function. Since the problem we have is a multi-variable nonlinear constrained problem, some modification is needed.

[8] Each worker can operate more than one kind of process or multiple version of the same process [6].

Table 2.4: "Optimum" production rates for two-stations system

$\sum_i \delta_i$	$\bar{\delta}/\lambda$	R_1	R_2	δ_1	δ_2
1.04	4.0	2	2	0.3400	0.7000
0.52	2.0	2	2	0.2025	0.3175
0.26	1.0	2	2	0.1150	0.1450
0.13	.50	2	2	0.0625	0.675
.065	.25	2	2	0.0322	0.0328
1.04	4.0	3	3	0.2875	0.7525
0.52	2.0	3	3	0.1800	0.3400
0.26	1.0	3	3	0.1075	0.1525
0.13	0.5	3	3	0.0600	0.0700
.065	.25	3	3	0.0320	0.0330

Since $Pout_N$ is a function of $(\delta_1, \delta_2, \ldots, \delta_N)$ the elimination method is used to eliminate one of the δ's, say δ_N, between the objective function and the constraint $\sum_{i=1}^{N} \delta_i = C_3$. That is done by substituting $C_3 - \sum_{i=1}^{N-1} \delta_i$ for δ_N in the objective function. After the search is done and the optimum values for δ's are obtained, the value of δ_N is calculated as $\delta_N = C_3 - \sum_{i=1}^{N-1} \delta_i$. The optimization results are presented in tables 2.4, 2.5, 2.6. As a general rule, for $\bar{\delta}/\lambda \leq 1$, $Pout_N$ is minimized when the production rates are approximately equal, which is known in manufacturing as the case of a balanced line. However, with $\bar{\delta}/\lambda > 1$, the optimum $Pout_N$ tends to be achieved by assigning higher production rates toward the downstream station, (the "funnel effect" already met in sub-section 4.1)

Tables 2.5 and 2.6 show a variety of distributions of δ's over the various stations that minimize $Pout_N$. Those distributions also depend on the value of C_4, as defined above.

Table 2.5: "Optimum" production rates for three-station system

$\sum_i \delta_i$	δ/λ	δ_1	δ_2	δ_3	P_{out_3}	I_3	SE
1.56	4.0	0.230	0.386	0.944	.0006	2.8582	.7699
0.78	2.0	0.158	0.236	0.386	.0087	1.9841	.6940
0.39	1.0	0.100	0.138	0.153	.0946	1.9841	.5660
.195	0.5	0.057	0.074	0.064	.3900	0.9858	.3608
.098	.25	0.030	0.038	0.030	.6246	0.4964	.3542

Table 2.6: "Optimum" production rates for four-station system

$\sum_i \delta_i$	δ/λ	δ_1	δ_2	δ_3	δ_4	P_{out_4}	I_4	SE
3.12	6.0	0.320	0.470	0.668	1.660	.0042	1.915	.8461
2.08	4.0	0.265	0.373	0.483	0.960	.0132	1.846	.8194
1.04	2.0	0.178	0.235	0.265	0.363	.0851	1.554	.7268
0.52	1.0	0.105	0.135	0.140	0.140	.3279	0.966	.5166
0.26	0.5	0.057	0.071	0.071	0.061	.6225	0.463	.2890
0.13	.25	0.029	0.036	0.036	0.029	.8074	0.214	.1474

3 MODELING THE MULTI-STAGE PULL PRODUCTION SYSTEM II

3.1 One-At-A-Time Pull Production Systems

In the last chapter , the multi-stage several-at-a-time pull production systems is modeled and analyzed. The assumption was that each station has "ample" servers (machines, indexed tools, operators, etc.) to handle units simultaneously. In this section we will model the one-at-a-time system and compare it with the several-at-a-time system. Because of the relative simplicity of the steady state probability equations of this system, a special mathematical method is presented to solve the set of equations. By inspecting Figure 3.1, it is noticed that the transition between states moving horizontally is similar. This means that moving from state (n_1, n_2) to state $(n_1, n_2 - 1)$is the same as moving from $(n_1, n_2 - 1)$ to $(n_1, n_2 - 2)$. That suggests the idea that a recursive approach is appropriate. The steady state equations for the system are:

For $n_2 = 1, 2, \ldots, R_2 - 1$

$$(\lambda + \delta_1) P_{R_1, n_2} = \delta_1 P_{R_1 - 1, n_2} + \lambda P_{R_1, n_2 + 1} \tag{3.1}$$

For $n_1 = 1, 2, \ldots, R_1 - 1$

and $n_2 = 1, 2, \ldots, R_2 - 1$

$$(\lambda + \delta_1 + \delta_2)P_{n_1,n_2} = \delta_1 P_{n_1-1,n_2} + \delta_2 P_{n_1+1,n_2-1} + \lambda P_{n_1,n_2+1} \tag{3.2}$$

For $n_2 = 1, 2, \ldots, R_2 - 1$

$$(\lambda + \delta_1)P_{0,n_2} = \delta_2 P_{1,n_2-1} + \lambda P_{0,n_2+1} \tag{3.3}$$

The boundary equations are:

$$\begin{aligned}
\delta 2 P_{R_1,0} &= \delta_1 P_{R_1-1,0} + \lambda P_{R_1,1} \\
&\vdots \\
(\delta_1 + \delta_2)P_{n_1,0} &= \delta_1 P_{n_1-1,0} + \lambda P_{n_1,1} \quad , n_1 = 1, 2, \ldots, R_1 - 1 \\
&\vdots \\
\delta_1 P_{0,0} &= \lambda P_{0,1}
\end{aligned} \tag{3.4}$$

And

$$\begin{aligned}
\lambda P_{R_1,R_2} &= \delta_1 P_{R_1-1,R_2} \\
&\vdots \\
(\lambda + \delta_1)P_{n_1,R_2} &= \delta_1 P_{n_1-1,R_2} + \delta_2 P_{n_1+1,R_2} \quad , n_1 = 1, 2, \ldots, R_1 - 1 \\
&\vdots \\
(\lambda + \delta_1)P_{0,R_2} &= \delta_2 P_{1,R_2-1}
\end{aligned} \tag{3.5}$$

The above steady state probability equations can be put matrix form using the notation:

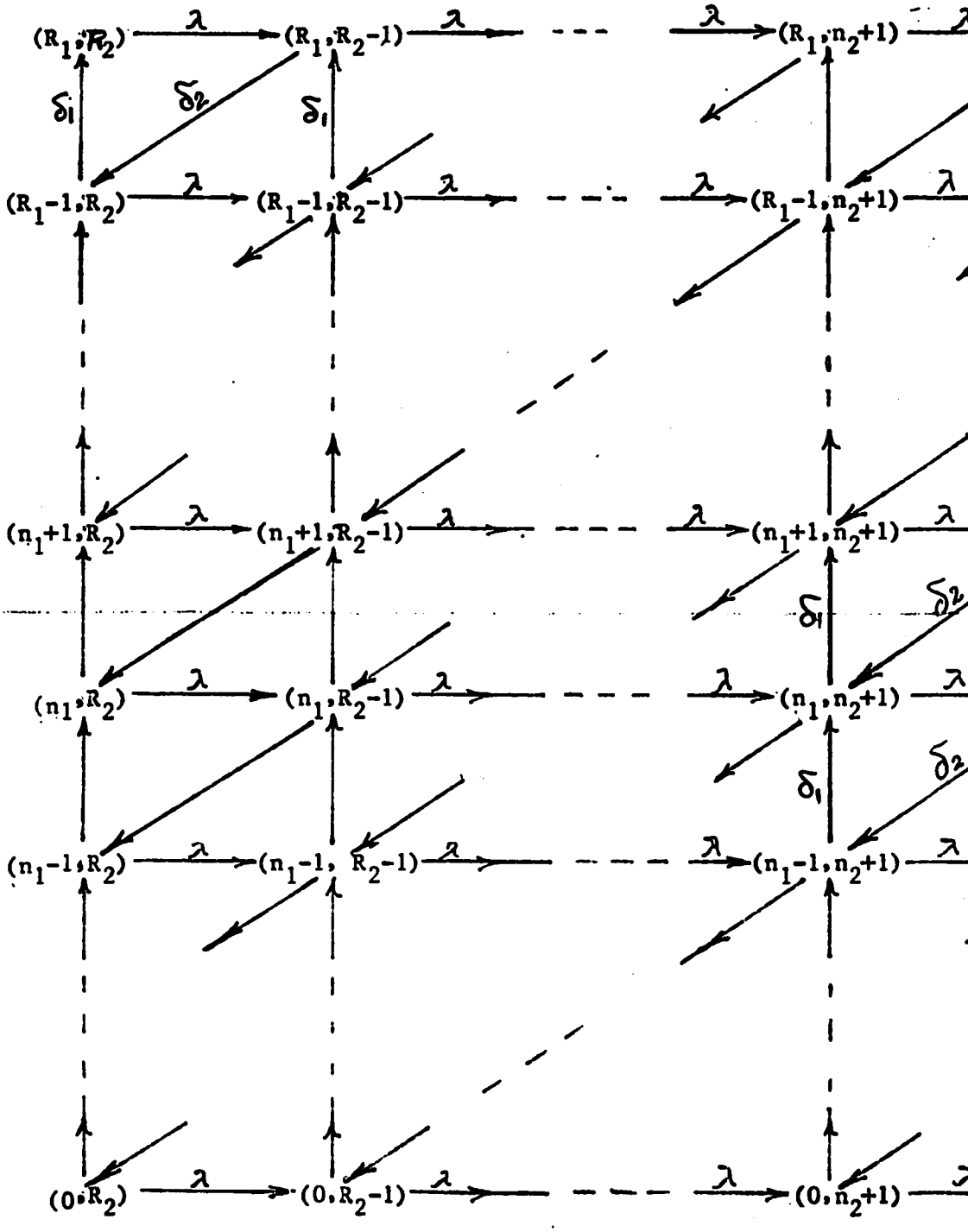
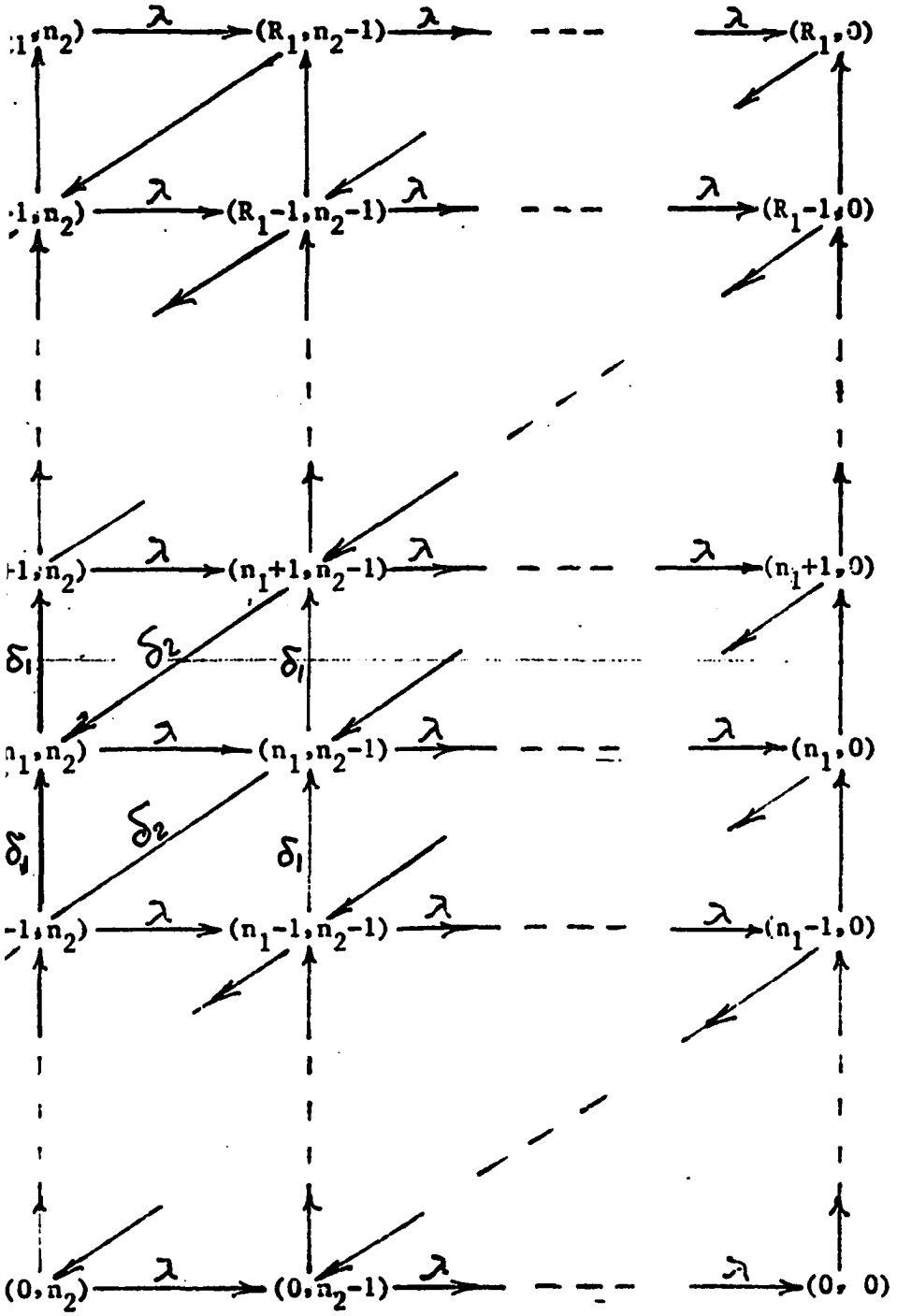


Figure 3.1: Transition diagram for two stations, one at-a-time


λ
δ1
δ2
(R1,n2-1)
(R1,0)
(R1-1,n2-1)
(R1-1,0)
(n1+1,n2-1)
(n1+1,0)
(n1,n2-1)
(n1,0)
(n1-1,n2-1)
(n1-1,0)
(0,n2)
(0,n2-1)
(0, 0)

$$P_i = \begin{vmatrix} P_{R_1,i} \\ P_{R_1-1,i} \\ \vdots \\ P_{0,i} \end{vmatrix}$$

so equations 3.1, 3.2, 3.3 will be,

$$\begin{aligned} (\lambda I)P_{I+1} &= \acute{M}P_i + \acute{N}P_{i-1} \\ P_{i+1} &= MP_i + NP_{i-1} \end{aligned} \tag{3.6}$$

where,

$$M = \begin{pmatrix} \frac{\lambda+\delta_2}{\lambda} & \frac{-\delta_1}{\lambda} & 0 & 0 & \cdots & 0 \\ 0 & \frac{\lambda+\delta_1+\delta_2}{\lambda} & \frac{-\delta_1}{\lambda} & 0 & \cdots & 0 \\ & & \ddots & & & \\ & & & \ddots & & \\ 0 & 0 & \cdots & 0 & \frac{\lambda+\delta_1+\delta_2}{\lambda} & \frac{-\delta_1}{\lambda} \\ 0 & 0 & 0 & \cdots & 0 & \frac{\lambda+\delta_1}{\lambda} \end{pmatrix}$$

and

$$N = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \frac{-\delta_2}{\lambda} & 0 & \cdots & 0 \\ 0 & \frac{-\delta_2}{\lambda} & \cdots & 0 \\ & & \ddots & \\ 0 & \cdots & \frac{-\delta_2}{\lambda} & 0 \end{pmatrix}$$

The boundary conditions are in the form:

$$P_1 = AP_0 \tag{3.7}$$

$$P_{R_2} = ZP_{R_2-1} \tag{3.8}$$

where,

$$A = \begin{pmatrix} \frac{\delta_2}{\lambda} & \frac{-\delta_1}{\lambda} & 0 & 0 & \cdots & 0 \\ 0 & \frac{\delta_1+\delta_2}{\lambda} & \frac{-\delta_1}{\lambda} & 0 & \cdots & 0 \\ & & \ddots & & & \\ & & & \ddots & & \\ 0 & 0 & \cdots & 0 & \frac{\delta_1+\delta_2}{\lambda} & \frac{-\delta_1}{\lambda} \\ 0 & 0 & 0 & \cdots & 0 & \frac{\delta_1}{\lambda} \end{pmatrix}$$

and

$$Z = \begin{pmatrix} \lambda & -\delta_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda+\delta_1 & -\delta_1 & 0 & \cdots & 0 \\ & & \ddots & & & \\ & & & \ddots & & \\ 0 & 0 & 0 & \cdots & 0 & \lambda+\delta_1 \end{pmatrix}^{-1} * \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \delta_2 & 0 & \cdots & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & \cdots & \delta_2 & 0 \end{pmatrix}$$

3.1.1 Superposition Method

To solve the above set of the state probability equations,

$$P_{i+1} = MP_i + NP_{i-1} \qquad i = 1, 2, \cdots, N-1 \tag{3.9}$$

with the two boundary conditions,

$$P_1 = AP_0 \tag{3.10}$$

$$P_{R_2} = ZP_{R_2-1} \tag{3.11}$$

and the normalizing equation,

$$\sum_i \sum_j p_{ij} = 1, \tag{3.12}$$

where

$$P_i = \begin{vmatrix} p_{0i} \\ p_{1i} \\ \vdots \\ p_{R_1,i} \end{vmatrix}$$

, we will employ the linearity and recursive structure of the system. The following steps illustrate the algorithm used:

- **Step 1:** Put P_0 in the form

$$P_0 = \begin{vmatrix} p_{00} \\ 0 \\ 0 \\ \vdots \\ 0 \end{vmatrix} + \begin{vmatrix} 0 \\ p_{10} \\ 0 \\ \vdots \\ 0 \end{vmatrix} + \cdots + \begin{vmatrix} 0 \\ \vdots \\ p_{j0} \\ \vdots \\ 0 \end{vmatrix} + \cdots + \begin{vmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ p_{R_1,0} \end{vmatrix}$$

- **Step 2:** For $p_{j0} = 1$ and $p_{k0} = 0$ for $k \neq j$, $j = 0, 1, \cdots, R_1$, solve recursively the equations 3.9 with the first boundary equation 3.11. Since the system is linear, superposition can be applied and the actual solution is a linear combination of these $R_1 + 1$ solutions, in the form:

$$P_i = (P_i)_0 * p_{00} + (P_i)_1 * p_{10} + \cdots + (P_i)_{R_1} * p_{R_1,0}, \tag{3.13}$$

with the weights $p_{00}, \cdots, p_{R_1,0}$ as yet undetermined, but determinable (in step 3)) , and with $(P_i)_j$, $j = 0, 1, \cdots, R_1$ equal to the value of P_i obtained above with $p_{j0} = 1$

The linearity required by superposition can be demonstrated as follows:

From equations 3.9 and 3.11, P_i can be put in the form

$$P_i = C_i P_0.$$

So,

$$(P_i)_j = C_i e_j, \qquad j = 0, 1, \cdots, R_1,$$

and, substituting in 3.13,

$$\begin{aligned} P_i &= C_i p_{00} e_0 + C_i p_{10} e_1 + \cdots + C_i p_{R_1,0} e_{R_1} \\ &= C_i (p_{00} e_0 + p_{10} e_1 + \cdots + p_{R_1,0} e_{R_1}) \\ &= C_i P_0 \end{aligned}$$

- **Step 3:** To find the actual value of P_0, substitute equation 3.13 in the second boundary equation 3.11 and the normalizing equation and solve for P_0 [1].

The above method allows us to solve for the state probabilities, by solving a reduced system of $R_1 + 1$, rather than $(R_1 + 1)(R_2 + 1)$ equation, which may constitute a remarkable reduction of computer time and /or memory. To illustrate the above method, an example of a two-stage system with station capacity of 2 at each, is presented below.

Example The steady state equations for the two-stage system with station capacity of 2 are:

$$\begin{aligned} (\lambda + \delta_2) p_{2,n_2} &= \delta_1 p_{1,n_2} + \lambda p_{2,n_2+1} \\ (\lambda + \delta_1 + \delta_2) p_{1,n_2} &= \delta_1 p_{0,n_2} + \delta_2 p_{2,n_2-1} + \lambda p_{1,n_2+1} \end{aligned}$$

[1] Eliminate one row of the set of the boundary equations.

$$(\lambda + \delta_1)p_{0,n_2} = \delta_2 p_{1,n_2-1} + \lambda p_{0,n_2+1}$$
$$n_2 = 1, 2 \tag{3.14}$$

with the boundary conditions,

$$\begin{aligned} \delta_1 p_{0,0} &= \lambda p_{0,1} \\ (\delta_1 + \delta_2)p_{1,0} &= \lambda p_{1,1} + \delta_1 p_{0,0} \\ \delta_2 p_{2,0} &= \lambda p_{2,1} + \delta_1 p_{1,0} \end{aligned} \tag{3.15}$$

$$\begin{aligned} \lambda p_{2,R_2} &= \delta_1 p_{1,R_2} \\ (\lambda + \delta_1)p_{1,R_2} &= \delta_1 p_{0,R_2} + \delta_2 p_{2,R_2-1} \\ (\lambda + \delta_1)p_{0,R_2} &= \delta_2 p_{1,R_2-1} \end{aligned} \tag{3.16}$$

In keeping with the algorithm, set $p_{20} = 1$ and $p_{10} = p_{00} = 0$. Substitute in equation 3.15 and solve for the state probability vector P_1, and repeat the same for $p_{10} = 1$ and $p_{20} = p_{00} = 0$, and for $p_{00} = 1$ and $p_{20} = p_{10} = 0$, and obtain the three corresponding solutions for P_1,

$$(P_1)_2 = \begin{vmatrix} \frac{\delta_2}{\lambda} \\ 0 \\ 0 \end{vmatrix}, (P_1)_1 = \begin{vmatrix} \frac{-\delta_1}{\lambda} \\ \frac{\delta_1+\delta_2}{\lambda} \\ 0 \end{vmatrix} \text{ and } (P_1)_0 = \begin{vmatrix} 0 \\ \frac{-\delta_1}{\lambda} \\ \frac{\delta_1}{\lambda} \end{vmatrix}.$$

To get P_2, substitute for P_0 and P_1 in equation 3.14, and the three corresponding solutions are:

$$(P_2)_j = \begin{vmatrix} (p_{22})_j \\ (p_{12})_j \\ (p_{02})_j \end{vmatrix} \quad \text{and } j = 0, 1, 2$$

where,

$$
\begin{aligned}
(p_{22})_j &= \sum_{i=0}^{2} m_{2l}(p_{l,1})_j + \sum_{l=0}^{2} n_{2l} \\
(p_{12})_j &= \sum_{i=0}^{2} m_{1l}(p_{l,1})_j + \sum_{l=0}^{2} n_{1l} \\
(p_{02})_j &= \sum_{i=0}^{2} m_{0l}(p_{l,1})_j + \sum_{l=0}^{2} n_{0l}
\end{aligned} \tag{3.17}
$$

and n_{ij} , m_{ij} is the $(ij)_{th}$ entry of the matrices N and M respectively. To get the actual value of p_{j0} , $j = 0, 1, 2$, substitute the above values in equation 3.16, which yields

$$
\sum_{j=0}^{2} (p_{k2})_j \; p_{j0} = \sum_{l=0}^{2} z_{kl} \; \sum_{j=0}^{2} (p_{l1})_j \; p_{j0} \tag{3.18}
$$

for k=0,1,2.

With one of the above equation replaced by the normalizing condition

$$
\sum_{i=0}^{2} \sum_{j=0}^{2} P_{ij} = 1,
$$

solve for $p_{j0}, j = 0, 1, 2$, and substitute back in equations 3.14 and 3.15 to get the whole set of the state probabilities.

3.1.2 Results and Comparison with Several-At-A-Time (ample) Server Systems

To illustrate the effect of single server on the system performance, several computer runs were made using several different values of station capacities and production rates, and the results are reported in Table 3.1. As expected, the probability that the final stage is out of stock is higher in the case of one server. The

Table 3.1: Comparison of one-at-a-time with several-at-a-time for two stages

Parameters				One-At-A-time			Several-At-a-Time		
R_1	R_2	δ_1	δ_2	I_1	I_2	P_{out_2}	I_1	I_2	P_{out_2}
2	2	.35	.35	1.5817	1.5240	.1070	1.6504	1.6196	.0587
4	6	.35	.35	3.4307	5.3932	.0020	3.6286	5.6272	.0000
6	4	.35	.35	5.4312	3.4432	.0121	5.6288	3.6288	.0006
2	2	.20	.50	1.1682	1.4975	.1249	1.3849	1.6657	.0537
4	6	.20	.50	2.5587	5.3581	.0075	3.3500	5.7363	.0000
6	4	.20	.50	4.3750	3.5750	.0098	5.3501	3.7400	.0002
2	2	.50	.20	1.7621	1.2660	.2084	1.7715	1.4093	.1210
4	6	.50	.20	3.6672	4.4965	.0280	3.7400	5.3465	.0001
6	4	.50	.20	5.6882	2.7991	.0707	5.7410	3.3525	.0038

average units per station are higher in the several-at-a-time case, the reason for which being that, in the several-at-a-time situation, the effective production rate is of course higher and the orders from subsequent stations are more quickly filled. On the other hand, the system responsiveness is higher for the one-at-a-time system, the reason for this being that our definition of SE has the term$(Prob\ n_1 < R_1)$ in the numerator and, as the out-of-stock probabilities will be higher for the single server system, and the number-in-stock distribution will be shifted to the left, causing SE to increase.

3.2 Lot Size Greater Than One

Under the JIT concept a part should continue to be processed or assembled at different stations without being held at any of them. In other words, the ideal work-in-process should not exceed one and so does the lot size. In the previous analysis we have held to the lot-size-of-one condition, and have relaxed the work-

Parameters						One-At-A-time				Several-At-a-Time			
R_1	R_2	R_3	δ_1	δ_2	δ_3	I_1	I_2	I_3	P_{out_3}	I_1	I_2	I_3	P_{out_3}
2	2	2	.3	.3	.3	1.5112	1.4274	1.3930	.1562	1.5905	1.5464	1.5286	.0865
2	4	7	.3	.3	.3	1.3926	2.8834	5.9844	.0058	1.5517	3.3970	6.5317	.0000
7	4	2	.3	.3	.3	6.3941	3.4186	1.4784	.1241	6.5808	3.5808	1.5806	.0650
4	3	1	.2	.3	.5	3.1221	2.5165	0.7871	.2129	3.6905	2.6905	0.6904	.3096
2	4	7	.2	.3	.5	1.0333	2.2073	5.9726	.0143	1.3501	3.2586	6.7015	.0001
4	3	1	.5	.3	.2	3.8255	2.6944	0.6051	.3949	3.8424	2.7374	0.6060	.3940
2	4	7	.5	.3	.2	1.6901	3.2346	5.3052	.0204	1.7400	3.5022	6.3344	.0000

Table 3.2: Compare one-at-a-time with several-at-a-time for three stages

in-process condition. Now, we will relax the lot size condition as well, and model the system with lot size greater than one. The two-stage pull production system will be modeled under the following assumptions, which agrees to some extent with Morse [26]:

1. The station capacity is a multiple of the lot size.

   That is $R_i = l_i\, Q_i \qquad (i = 1,2,\ldots,N)\ , \quad l_i$ is constant

2. Station i will send an order to withdraw from station $i-1$ if its capacity drops from $m_i\ Q_i + 1$ to $m_i\ Q_i$, $l_i > m_i \geq 0$. So if the number of units is in the range from $(l_i - n_i)Q_i$ to $(l_i - n_i - 1)Q_i + 1$, there are n_i orders outstanding (each order is for lot size of Q units).

3. The lot size of a station is equal to, or is a multiple of, the lot size of its subsequent station.

Summarizing the above,

$$R_i = Q_i l_i \qquad , \ l_i \geq 1, \qquad i = 1,2,\ldots,N$$

$$Q_i = K_i Q_{i+1} \ , \ K_i \geq 1, \qquad i = 1,2,\ldots,N$$

The state-transition rate diagram (for $Q_1 = 2Q_2$) is shown in figure 3.2, which shows every possible state and all the possible transitions from one state to another. The steady state equilibrium equations for this model are as follows:

For $n_1 = 0, Q_2, 2Q_2, \ldots, R_1$

$$[\delta_1 * \mid \frac{R_1 - n_1}{Q_1} \mid + \delta_2 . min(\mid \frac{R_2 - 0}{Q_2} \mid, \mid \frac{n_1}{Q_2} \mid) P_{n_1,0} =$$

$$\lambda P_{n_1,1}$$

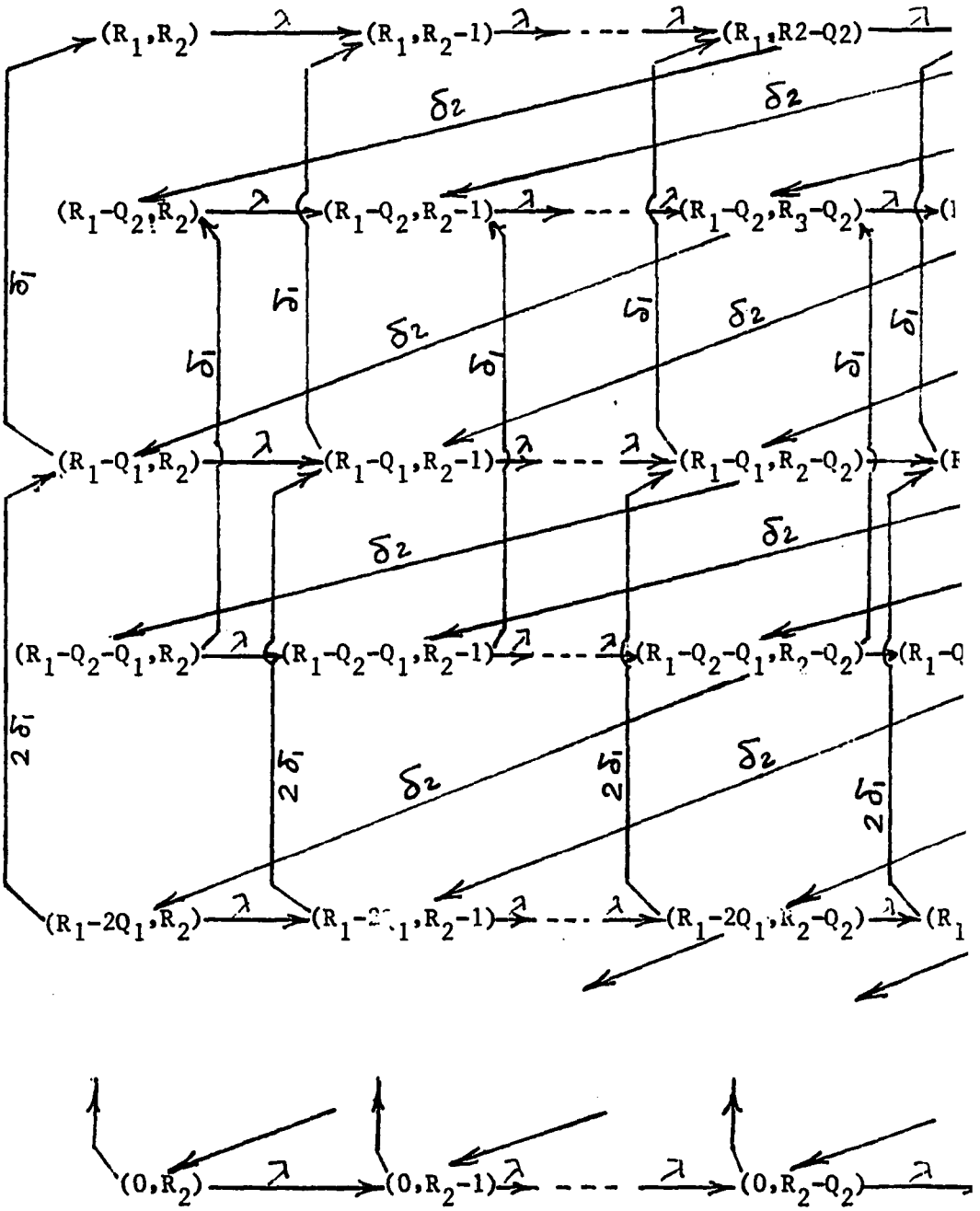
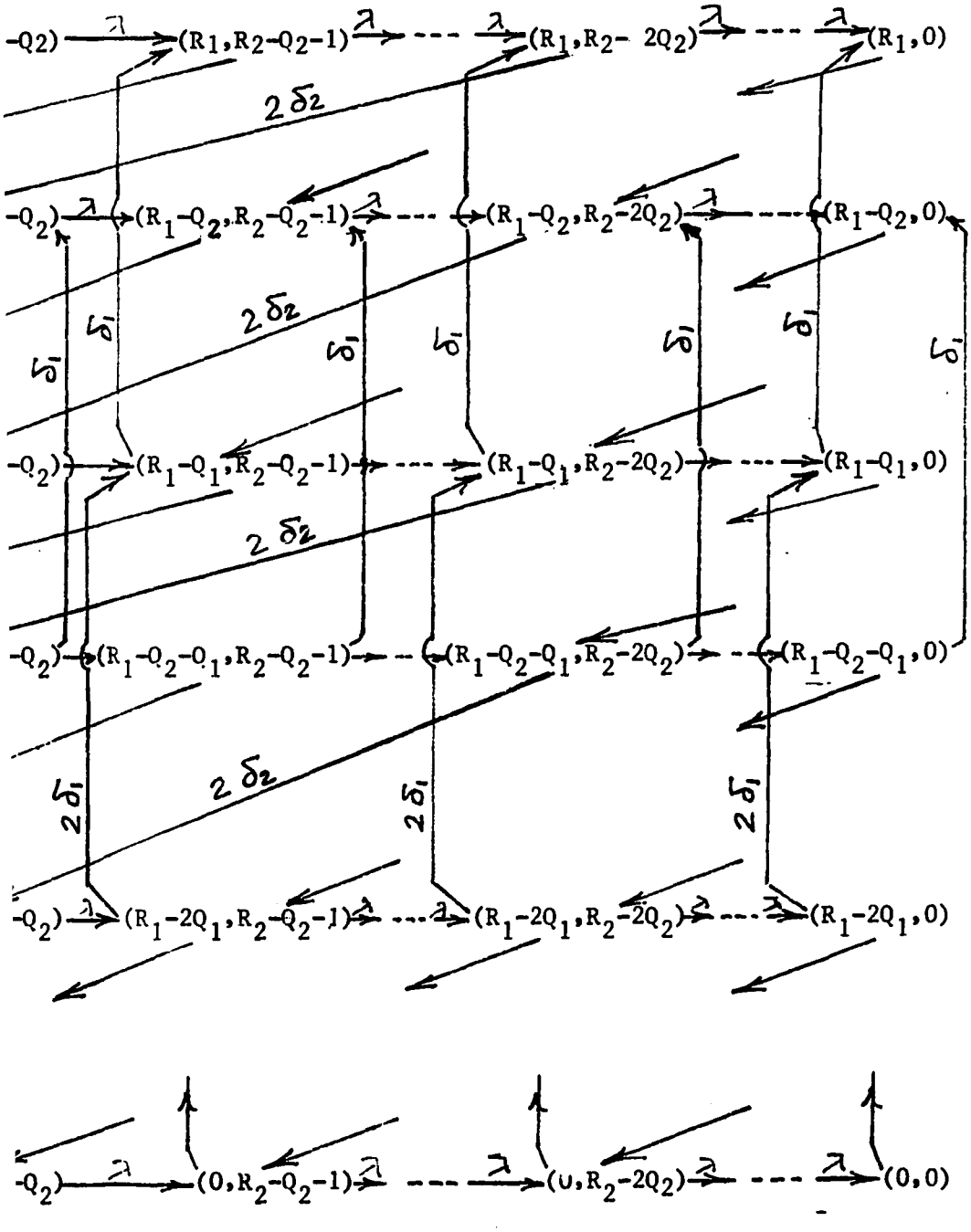


Figure 3.2: Transition diagram for lot size larger than 1


(R_1, R_2-Q_2-1)
(R_1, R_2-2Q_2)
$(R_1, 0)$
$2\delta_2$
(R_1-Q_2, R_2-Q_2-1)
(R_1-Q_2, R_2-2Q_2)
$(R_1-Q_2, 0)$
δ_1
(R_1-Q_1, R_2-Q_2-1)
(R_1-Q_1, R_2-2Q_2)
$(R_1-Q_1, 0)$
$(R_1-Q_2-Q_1, R_2-Q_2-1)$
$(R_1-Q_2-Q_1, R_2-2Q_2)$
$(R_1-Q_2-Q_1, 0)$
$2\delta_1$
(R_1-2Q_1, R_2-Q_2-1)
(R_1-2Q_1, R_2-2Q_2)
$(R_1-2Q_1, 0)$

$(0, R_2-Q_2-1)$
$(0, R_2-2Q_2)$
$(0, 0)$

$$+\delta_1 . \mid \frac{R_1-(n_1-Q_1)}{Q_1} \mid P_{n_1-Q_1,0} \tag{3.19}$$

For $n_1 = 0, Q_2, 2Q_2, \ldots, R_1$ and $n_2 = 1, 2, \ldots, R_2$

$$\begin{aligned}
&[\lambda + \delta_1 . \mid \frac{R_1-n_1}{Q_1} \mid + \delta_2 . min(\mid \frac{R_2-n_2}{Q_2} \mid, \mid \frac{n_1}{Q_2} \mid)] P_{n_1,n_2} = 0 \\
&\lambda P_{n_1,n_2+1} \\
&+ \delta_1 . \mid \frac{R_1-(n_1-Q_1)}{Q_1} \mid P_{n_1-Q_1,n_2} \\
&+ \delta_2 . min(\mid \frac{R_2-(n_2-Q_2)}{Q_2} \mid, \mid \frac{n_1+Q_2}{Q_2} \mid) P_{n_1+Q_2,n_2-Q_2}
\end{aligned} \tag{3.20}$$

where, $\mid \frac{a}{b} \mid$ is the largest integer less than or equal to $\frac{a}{b}$.

Using numerical methods as before, we solve the equilibrium equations and calculate the joint stationary probabilities and system performance measures, as defined above.

3.3 The Effect of Using Lot Size Greater Than One

To study the effect of large lot sizes, a number of computer runs were made for various values of lot sizes and station capacities. A comparison is made between this model and our standard model with $Q = 1$. Tableas 3.3, 3.4 show the system performance for various station capacities and lot sizes (including $Q = 1$). It is noticed that:

- The probability of stockout is smallest for lot size of 1 ,and generally is smaller for smaller lot size.

- The number of average units per station is greatest for lot size of 1, and generally is greater for smaller lot sizes.

Table 3.3: The system performance for various lot sizes, $R_1 = 12 \quad R_2 = 6$

Perf.	lot size	$\delta_1 = .5 \quad \delta_2 = .5$	$\delta_1 = .2 \quad \delta_2 = .5$
P_{out_2}	$Q_1 = 1$, $Q_2 = 1$	.5468E-04	.3308E-06
	$Q_1 = 4$, $Q_2 = 2$	.1116E-02	.2517E-04
	$Q_1 = 6$, $Q_2 = 3$	.6579E-02	.3811E-03
	$Q_1 = 6$, $Q_2 = 2$	.1116E-02	.2549E-04
	$Q_1 = 6$, $Q_2 = 6$	.1116E-02	.2516E-04
	$Q_1 = 3$, $Q_2 = 3$	.6579E-02	.3806E-03
	$Q_1 = 2$, $Q_2 = 2$	.9774E-01	.4153E-01
I_2	$Q_1 = 1$, $Q_2 = 1$	5.3500	5.7400
	$Q_1 = 4$, $Q_2 = 2$	4.8513	5.2400
	$Q_1 = 6$, $Q_2 = 3$	4.3609	4.7405
	$Q_1 = 6$, $Q_2 = 2$	4.8513	5.2400
	$Q_1 = 6$, $Q_2 = 6$	4.8513	5.2400
	$Q_1 = 3$, $Q_2 = 3$	4.3609	4.7405
	$Q_1 = 2$, $Q_2 = 2$	3.1579	3.3546
SE	$Q_1 = 1$, $Q_2 = 1$	0.8806	0.7353
	$Q_1 = 4$, $Q_2 = 2$	2.1548	0.9801
	$Q_1 = 6$, $Q_2 = 3$	2.0840	0.9247
	$Q_1 = 6$, $Q_2 = 2$	2.7183	1.1648
	$Q_1 = 6$, $Q_2 = 6$	0.4802	0.4529
	$Q_1 = 3$, $Q_2 = 3$	0.3266	0.3212
	$Q_1 = 2$, $Q_2 = 2$	0.1504	0.1595

- The system responsiveness is greater for larger Q_1/Q_2 .

However, there is a missing cost item here, which is the cost of moving items in bulk vs. one-at-a-time. As a partial answer, we should mention that, in the JIT factory, the stations are arranged closer to each other because there is no need for a WIP area and the transportation cost might not be a factor. To get more insight into the above considerations we are going to analyze the effect of lot size in the light of a traditional cost function. We will use profit/mean production time as our

Table 3.4: The system performance for various lot sizes, $R_1 = 6 \quad R_2 = 12$

Perf.	lot size	$\delta_1 = .5 \quad \delta_2 = .5$	$\delta_1 = .2 \quad \delta_2 = .5$
P_{out_2}	$Q_1 = 1$, $Q_2 = 1$	.1325E-09	.5932E-13
	$Q_1 = 6$, $Q_2 = 3$	.2920E-04	.3295E-05
	$Q_1 = 6$, $Q_2 = 2$	.1371E-04	.3067E-05
	$Q_1 = 6$, $Q_2 = 6$	.4064E-03	.7558E-05
	$Q_1 = 3$, $Q_2 = 3$	.1634E-05	.4322E-08
	$Q_1 = 2$, $Q_2 = 2$	.4755E-07	.5223E-10
I_2	$Q_1 = 1$, $Q_2 = 1$	11.3500	11.7400
	$Q_1 = 6$, $Q_2 = 3$	10.3130	10.7500
	$Q_1 = 6$, $Q_2 = 2$	10.7760	11.1750
	$Q_1 = 6$, $Q_2 = 6$	8.8470	9.2354
	$Q_1 = 3$, $Q_2 = 3$	10.3490	10.7400
	$Q_1 = 2$, $Q_2 = 2$	10.8500	11.2400
SE	$Q_1 = 1$, $Q_2 = 1$	0.8806	0.7353
	$Q_1 = 6$, $Q_2 = 3$	2.0784	0.9212
	$Q_1 = 6$, $Q_2 = 2$	2.7016	1.1562
	$Q_1 = 6$, $Q_2 = 6$	0.1666	0.1667
	$Q_1 = 3$, $Q_2 = 3$	0.3286	0.3214
	$Q_1 = 2$, $Q_2 = 2$	0.4807	0.4530

index of performance [26].

profit/mean prod. time = (gross profit − inv. cost − order cost)/mean prod. time

$$= G.D - C_i.I - C_o.N_R$$

where

G : Gross profit/unit

C_i : Inventory cost/unit/mean production time

C_o : Order cost/order

D : Mean number of units demanded in mean production time

I : Mean number of units

N_R : Mean number of orders per mean production time

For $Q = 1$,

$D = N_R$ and $profit_1 = D_1(G - C_o) - C_i.I_1$

For $Q > 1$,

$D = N_R\, Q$ and $profit_2 = D_2(G - C_o/Q) - C_i.I_2$

where,

$D_1 = \frac{\lambda}{\delta_N}[1 - (P_{out_N})_1]$

$D_2 = \frac{\lambda}{\delta_N}[1 - (P_{out_N})_2]$

Assuming the cost coefficients G, C_i, and C_o , as well as the production rates and demand rate, do not depend on lot size, we can deduce from the numerical results that:

- D_1 is greater than D_2 because $(P_{out_N})_1$ is less than $(P_{out_N})_2$. See Tables 3.3 and 3.4.

- The mean number of units "I_2" is smaller in the $Q > 1$ case, so that the inventory cost is smaller.

From the above, we conclude that the values of the cost coefficients G, C_i and C_o and their ratios are the determining factors in choosing the lot size.

4 CONTRASTING PULL AND PUSH

Many authors write about the pull systems and its superiority over the traditional push systems without dealing with the subject quantitatively. The exception for that is Terada and Kimura [36], who analyzed the pull systems and compare it to the push systems. The comparison object was the amplification of inventory/production fluctuation in the further precedent stations.

In this chapter , we will model the push system and compare the performance of both systems, using the performance measures explained before in Chapter 2.

4.1 Modeling the Push Systems

The push system , which is known as the traditional production systems, has been modeled in the literatures under different assumptions, see [1], [2], [22], [8]. In this research we will assume the demand rate is accurately forecast and units will be pushed by this rate λ to the upper stream station. For the comparative study, this rate is assumed to be the same actual demand rate used by the pull systems to pull units from the downstream station.

The push systems work as follows under the assumption of exponential distribution of the process and interarrival times:

A Part is pushed by rate λ into the first station whose capacity is R_1 (there

is room for R_1 units and there is R_1 "servers"). The flow of arrival stops [8] if the capacity reached R_1. The part is processed with rate δ_1 and leaves to station 2 if its capacity is less than R_2, otherwise it will block its server at station 1 and process continues the same way till final station. At the final station N, a part is processed with rate δ_N and is able to leave immediately to be shipped (no blocking is occurred at the final station).

4.1.1 The Model

As a simplified example, we will model a two-station system. The system is represented by a finite continuous Markov process with state space $S = \{(n_1, n_2, n_3) : 0 \leq n_1 \leq R_1 < \infty, 0 \leq n_2 \leq R_2 < \infty, 0 \leq n_3 \leq R_1\}$. Where n_i, $i = 1, 2$ is the number of units at station i and R_i is the maximum capacity of station i in units and n_3 is the number of blocked servers in station 1.

Figure 4.1 shows the transition diagram of the system and the stationary probability equations are as follows:

For $n_1 = 0, 1, \ldots, R_1 - 1$

$n_2 = 0, 1, \ldots, R_2 - 1$

$n_3 = 0$

$$\begin{aligned}[\lambda + \delta_1 * n_1 + \delta_2 * n_2] P_{n_1,n_2,0} = \\ & \delta_1 * (n_1 + 1) P_{n_1+1,n_2-1,0} \\ & + \delta_2 * (n_2 + 1) P_{n_1,n_2+1,0} \\ & + \lambda P_{n_1-1,n_2,0} \end{aligned} \tag{4.1}$$

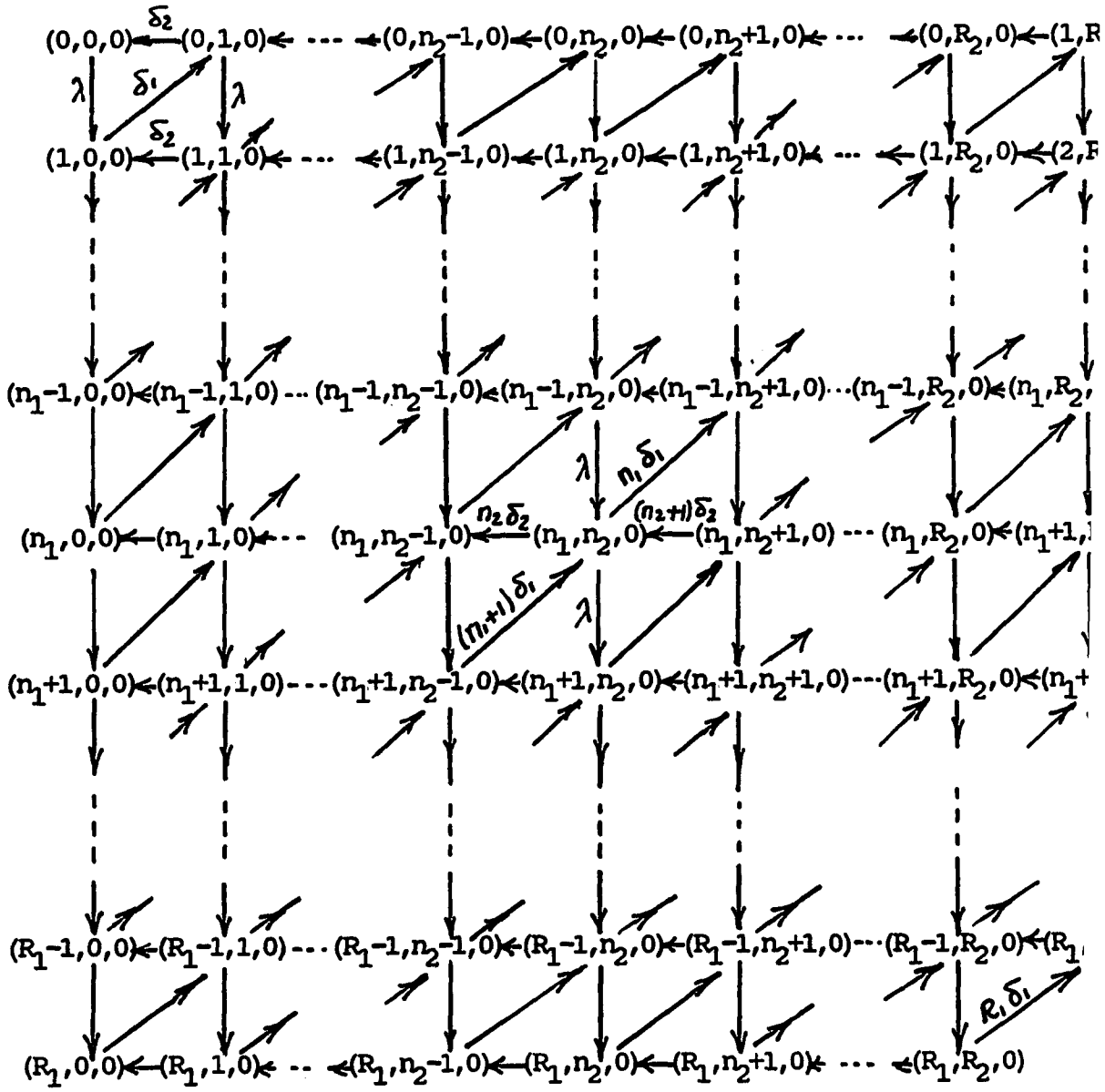
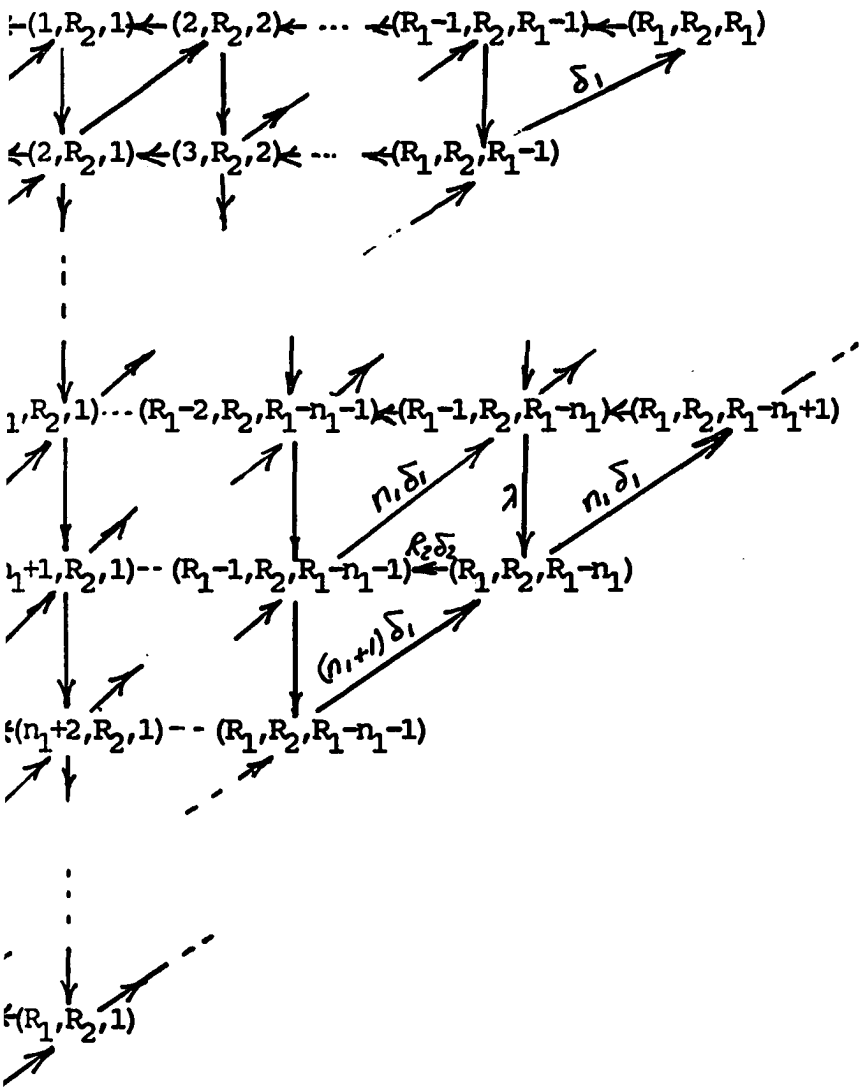


Figure 4.1: Transition diagram for two-station (usual model) push system


(1,R2,1)
(2,R2,2)
(R1-1,R2,R1-1)
(R1,R2,R1)
δ1
(2,R2,1)
(3,R2,2)
(R1,R2,R1-1)
(R1-2,R2,R1-n1-1)
(R1-1,R2,R1-n1)
(R1,R2,R1-n1+1)
n1δ1
λ
n1δ1
(R1-1,R2,R1-n1-1)
ρ2δ2
(R1,R2,R1-n1)
(n1+1)δ1
(n1+2,R2,1)
(R1,R2,R1-n1-1)
(R1,R2,1)

For $n_1 = 0, 1, \ldots, R_1 - 1$

$n_2 = R_2$

$n_3 = 0$

$$\begin{aligned}[\lambda + \delta_1 * n_1 + \delta_2 * R_2] P_{n_1,R_2,0} = \\ \delta_1 * (n_1 + 1) P_{n_1+1,R_2-1,0} \\ + \delta_2 * R_2 P_{n_1+1,R_2,1} \\ + \lambda P_{n_1-1,R_2,0}\end{aligned} \tag{4.2}$$

For $n_1 = R_1$

$n_2 = 0, 1, \ldots, R_2$

$n_3 = 0$

$$\begin{aligned}[\delta_1 * R_1 + \delta_2 * n_2] P_{R_1,n_2,0} = \\ \delta_2 * (n_2 + 1) P_{n_1,n_2+1,0} \\ + \lambda P_{R_1-1,n_2,0}\end{aligned} \tag{4.3}$$

For $n_1 = 1, 2, \ldots, R_1 - 1$

$n_2 = R_2$

$n_3 = 1, 2, \ldots, R_1 - 1$

$$\begin{aligned}[\lambda + \delta_1 * (n_1 - n_3) + \delta_2 * R_2] P_{n_1,R_2,n_3} = \\ \delta_1 * (n_1 - n_3 + 1) P_{n_1,R_2,n_3-1} \\ + \delta_2 * R_2 P_{n_1+1,R_2,n_3+1} \\ + \lambda P_{n_1-1,R_2,n_3}\end{aligned} \tag{4.4}$$

For $n_1 = R_1$

$n_2 = R_2$

$n_3 = 1, 2, \ldots, R_1$

$$[\delta_1 * (n_1 - n_3) + \delta_2 * R_2] P_{R_1,R_2,n_3} = \delta_1 * (R_1 - n_3 + 1) P_{R_1,R_2,n_3-1} + \lambda P_{R_1-1,R_2,n_3} \quad (4.5)$$

and the boundary conditions are:

$$P_{-1,n_2,n_3} = 0$$

$$P_{n_1,-1,n_3} = 0$$

$$P_{R_1+1,n_2,n_3} = 0$$

$$P_{n_1,R_2+1,n_3} = 0$$

$$P_{n_1,n_2,n_3} = 0 \quad \text{if } n_1 < n_3 \quad \text{or} \quad (n_2 < R_2 \text{ and } n_3 > 0)$$

The above equations can be solved simultaneously using one of the methods describes in Chapter 2 and the joint stationary probability distribution could be obtained. Consequently, the system performance measures as identified in Chapter 2 can be calculated.

4.2 Comparing Push and Pull Systems

The recent literature appears to favor "pull" over the more traditional "push", primarily on general and managerial, rather than quantitative, grounds ([28], [35], [25]). One exception to this is Terada and Kimura [36], who analyzed pull and push on the basis of the magnitude of inventory/production fluctuation in the furthest

downstream station. In this section, we pursue that sort of quantitative comparison [1], but with the first two performance measures indicated in Section 2.5. The comparisons are of course made fair by using the same station capacities, production rates and demand rate for both systems (see Section 4.1), although, as pointed out at the end of the present subsection, the added push assumption of perfect predictability of λ could be considered as a bias favoring push.

The results are given in both tabular (Table 4.1) and graphical (Figures 4.2 - 4.8) form. Specifically, Table 4.1 exhibits the differential effect, on a pull and a push system, of possible allocations of a fixed total of station capacities to the two stations of the system. Figures 4.2 - 4.5 exhibit the effect of changing station capacity on the performance index I_2 for a pull and a push system. Figures 4.6 - 4.8 exhibit the effect of changing production rate on performance index I_2 for a pull and a push system. In the case of Figures 4.2 - 4.8, both the station capacities and the production rates are kept equal among the two stations. It may be noted that, while results for the performance index P_{out_2} are not presented graphically, these exhibit precisely the same quantitative behavior as given in Figures 4.2 - 4.8 for the performance Index I_2.

General conclusions to be drawn from these tables and figures are as follows:

1. Pull systems tend to outperform push systems. Pull systems show lower values of stockout probability P_{out_2} and higher values of average number of units I_2 at the final station.

2. However, with low production rates δ (lower than demand (for pull) or supply

[1] For the special case in which the equations of sections 2.3.1 and 4.1 apply.

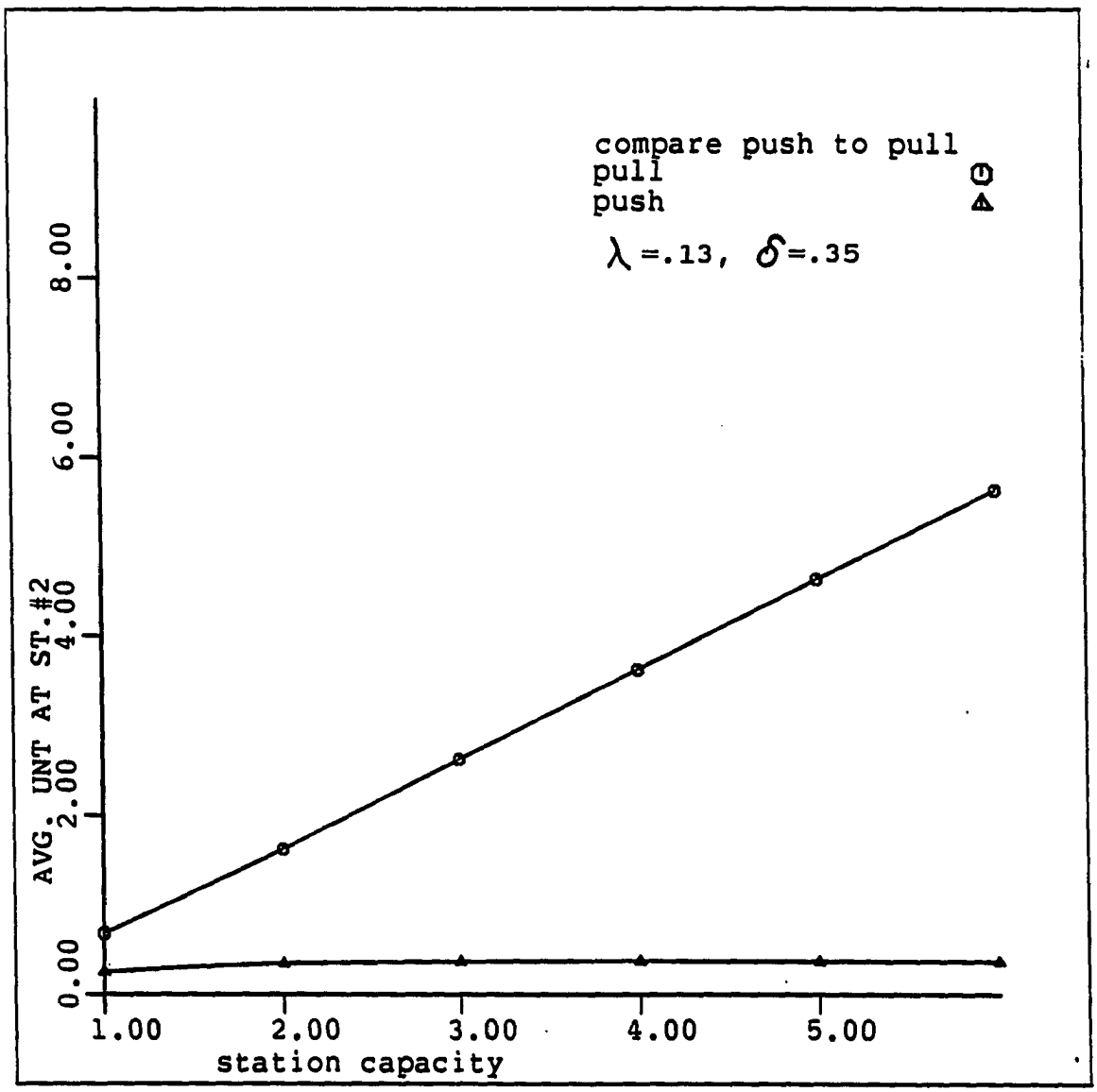


Figure 4.2: Effect of changing station capacity on a pull and a push system (a)

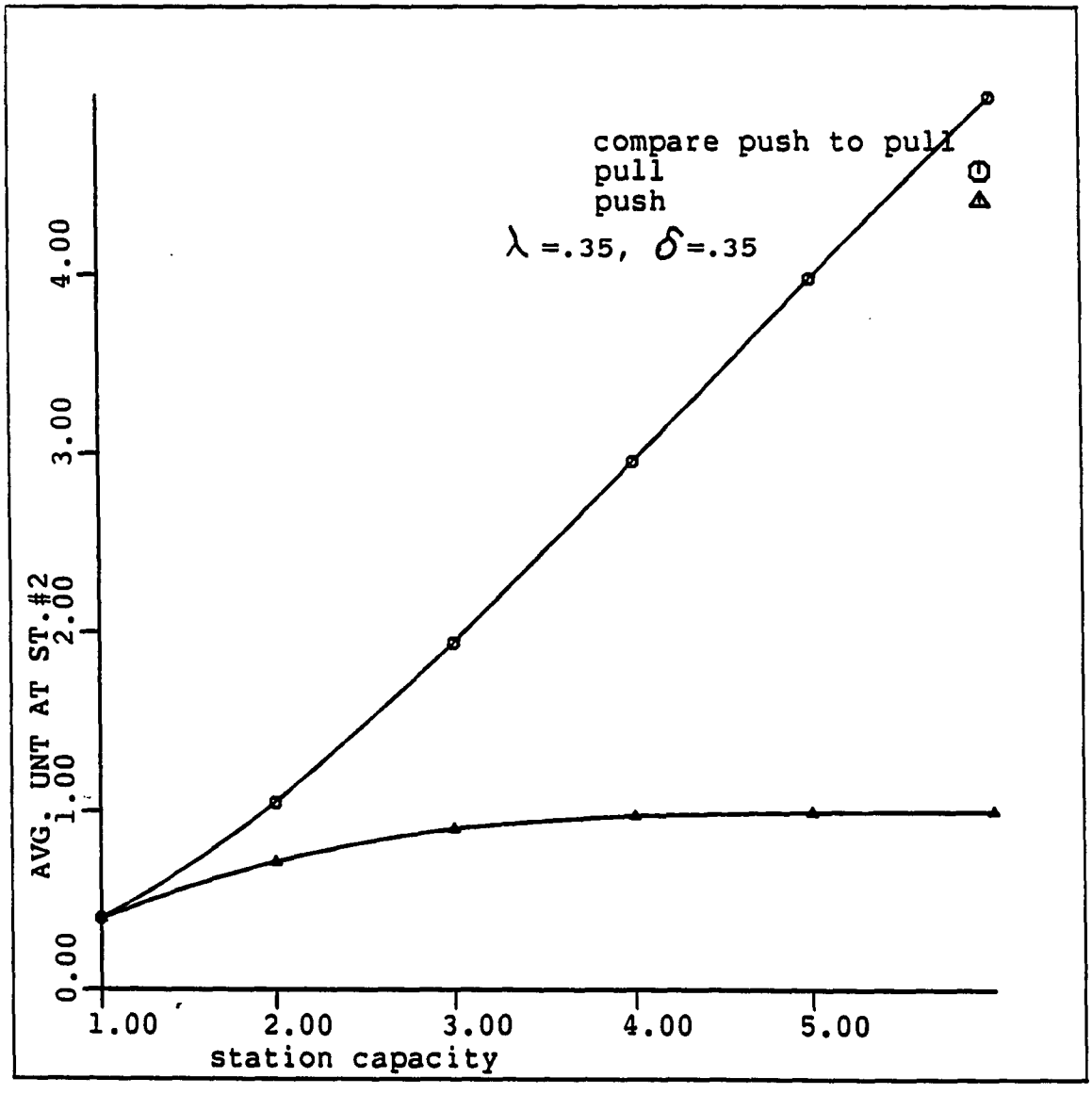


Figure 4.3: Effect of changing station capacity on a pull and a push system (b)

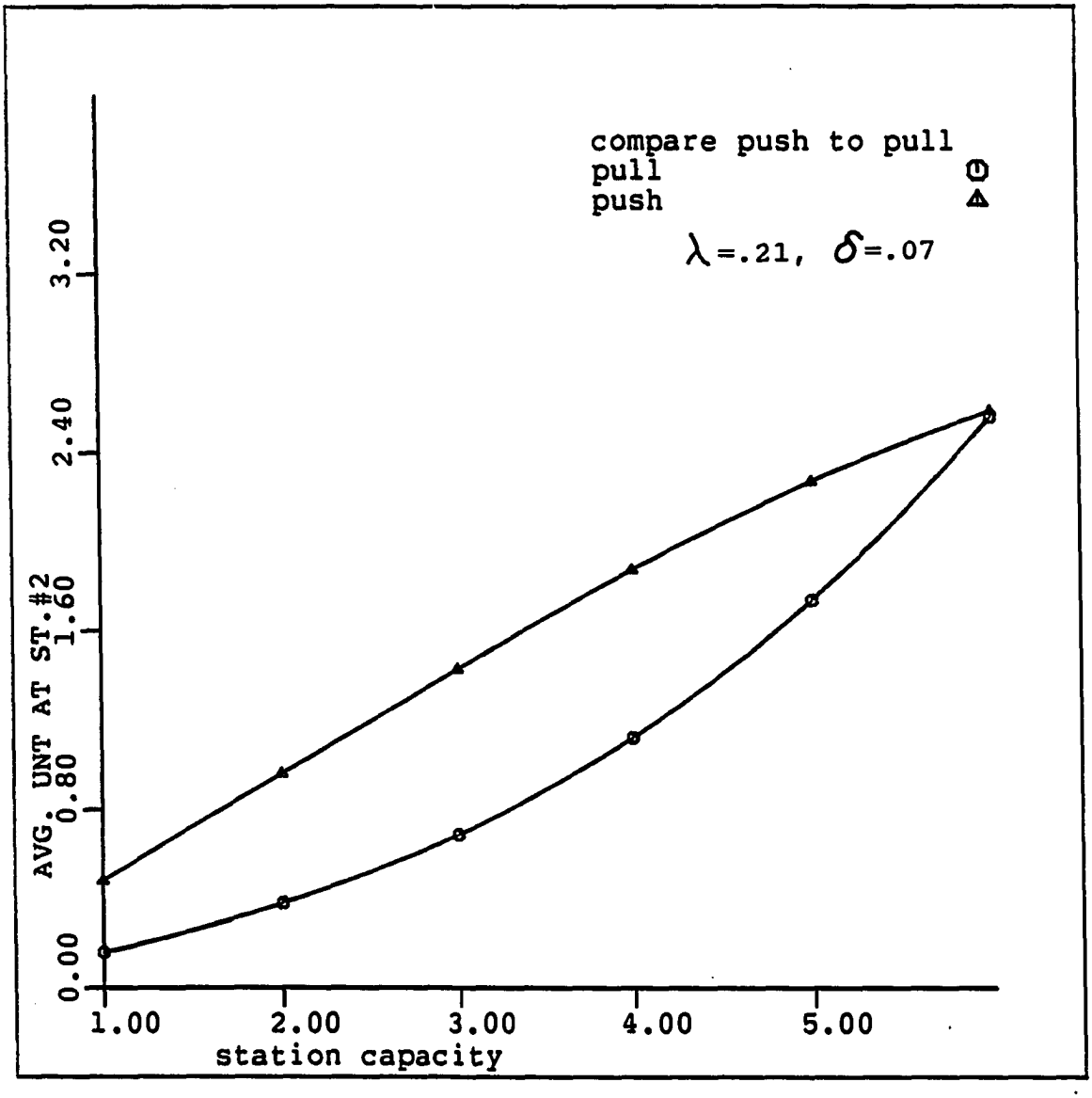


Figure 4.4: Effect of changing station capacity on a pull and a push system (c)

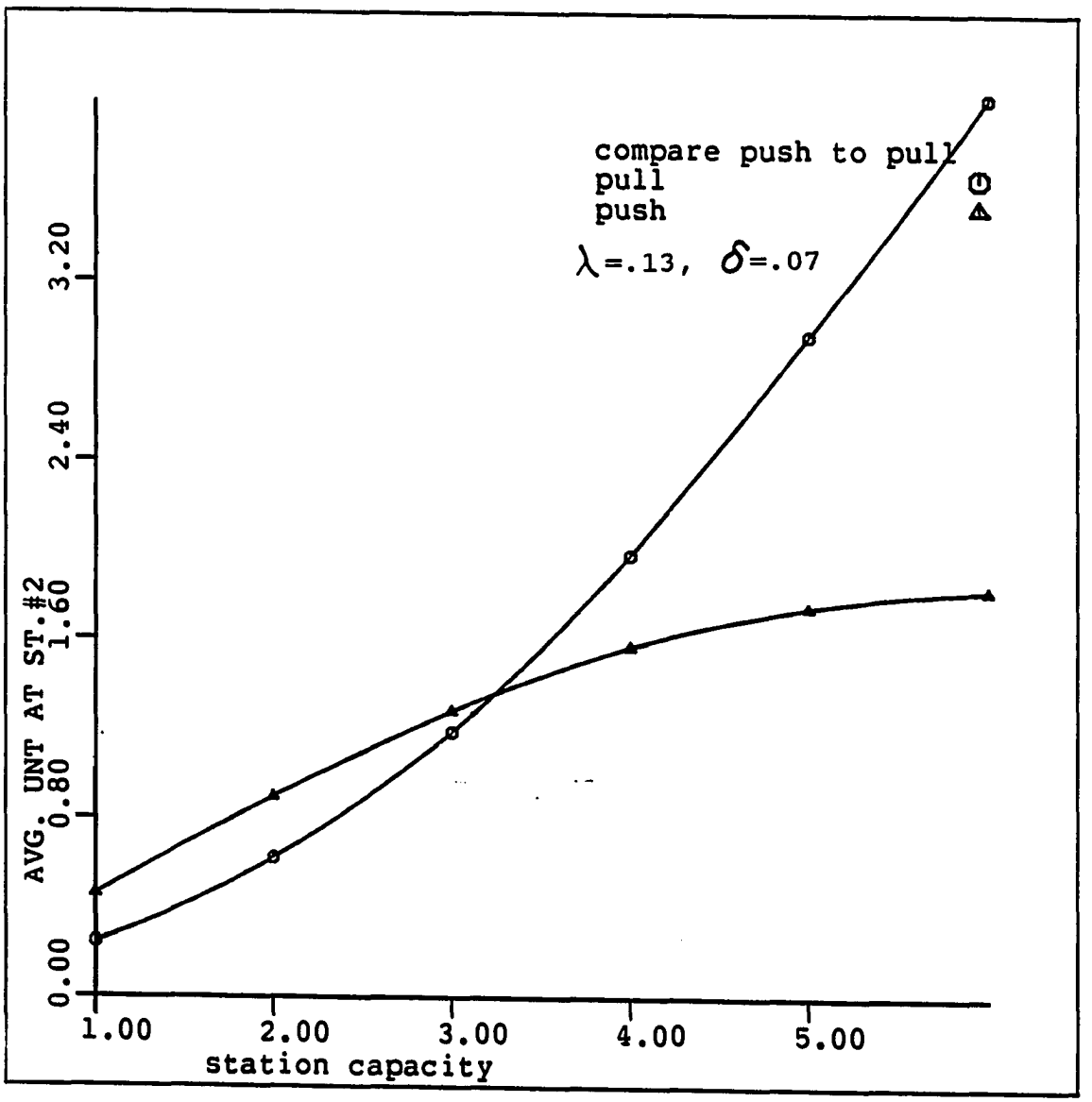


Figure 4.5: Effect of changing station capacity on a pull and a push system (d)

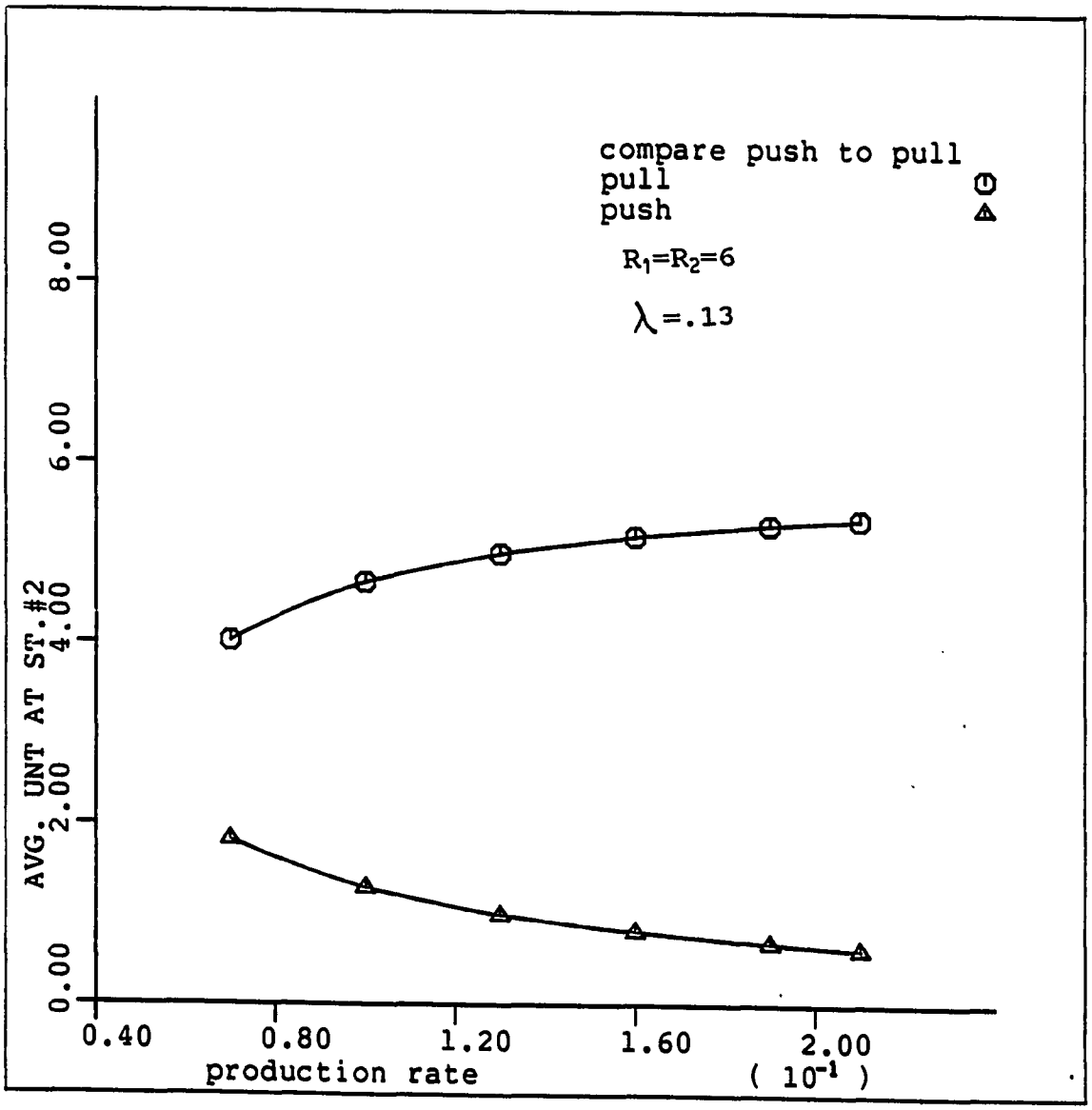


Figure 4.6: Effect of changing production rate on a pull and a push system (a)

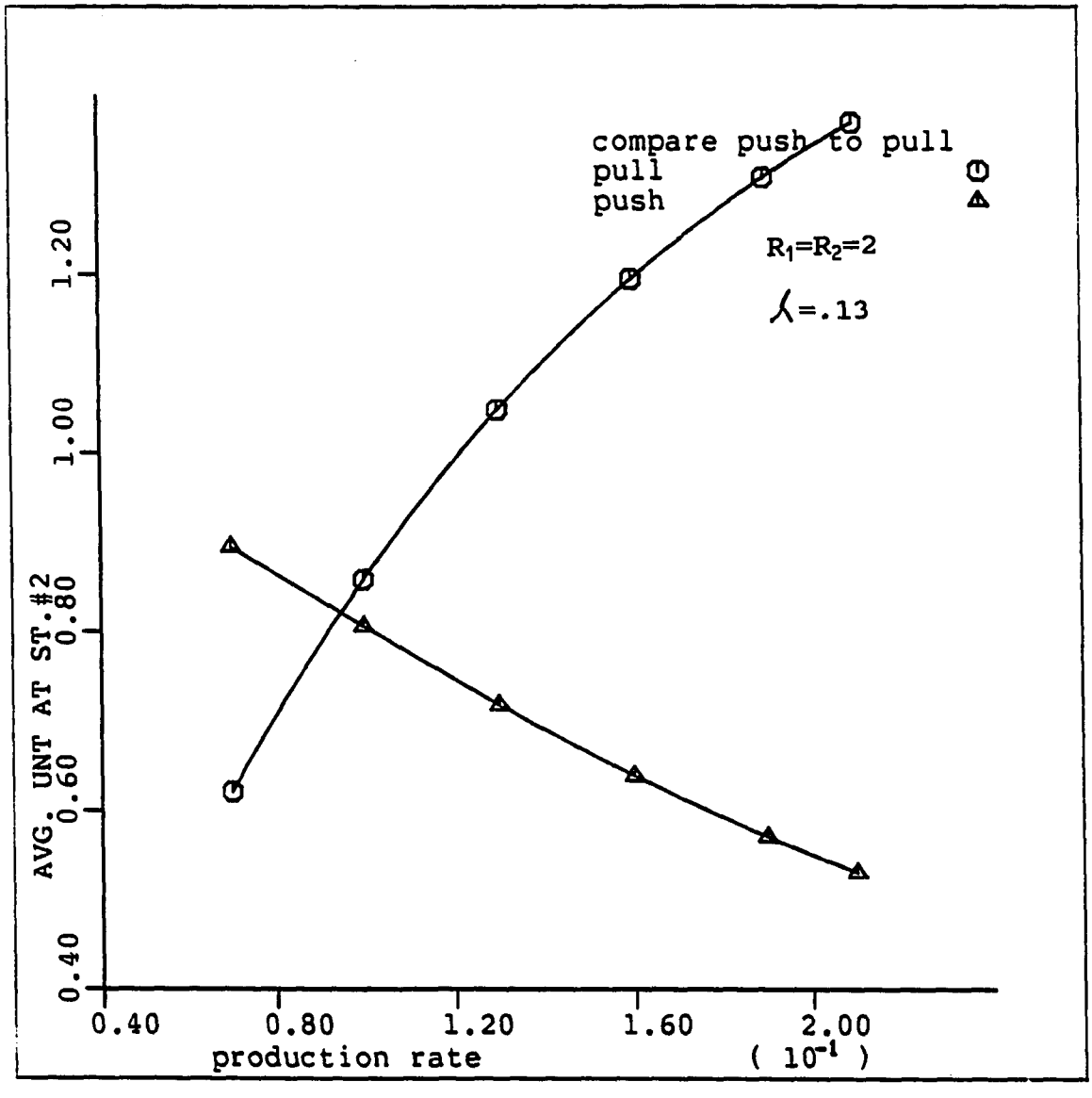


Figure 4.7: Effect of changing production rate on a pull and a push system (b)

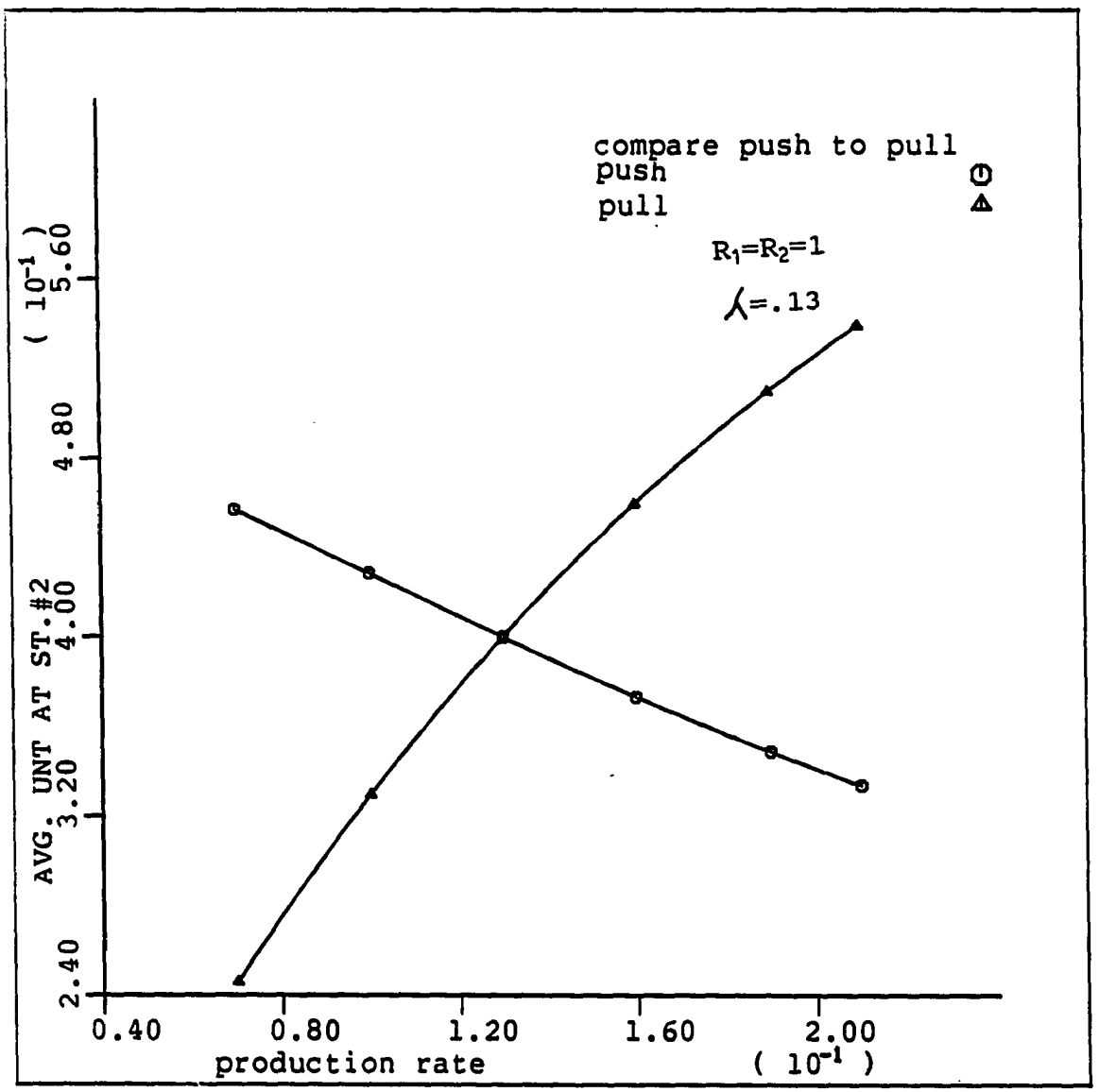


Figure 4.8: Effect of changing production rate on a pull and a push system (c)

Table 4.1: Performance of pull vs. that of push

Station capacity		Pull		Push	
R_1	R_2	$Pout_2$	I_2	$Pout_2$	I_2
1	3	0.5122	0.7073	0.5768	0.4993
3	1	0.5007	0.4993	0.5122	0.4878
2	6	0.1116	3.1046	0.4268	0.8000
6	2	0.2000	1.2000	0.3202	0.8884
4	6	0.0036	4.8842	0.3703	0.9843
6	4	0.0157	3.0135	0.3598	0.9964

(for push)rate λ), that is, in the case of "under powered" systems, and for small station capacities R, push systems can outperform pull systems.

3. For push systems, increasing station capacities R, in the presence of approximately equal δ and λ, does not increase the average number of units I_2 as it does in the case of pull systems.

These general conclusions are made plausible by leaning on the ordinary interpretations that one places on the terms *inflow*, *capacity* and *outflow*. In the case of pull systems, the rate of inflow is the effective production rate of the first station upstream, proportional to the difference between that station's capacity and the number of units currently in storage (see the equations of Subsection 2.3.1) , while the rate of outflow is the fixed demand rate. In push systems, the rate of inflow is the fixed rate of supply, while the outflow rate is the effective production rate of the last station downstream, proportional to the number of units currently in storage.

Now, with regard to the second above conclusion, i.e., that push outperforms pull when production rates δ are lower than the demand/supply rate λ, that is

simply a matter, under push, of material inflow more rapid than material outflow, resulting in high average number of units at the final downstream station and lower probability of stockout. On the other hand, under pull, the correspondence, respectively, of δ and λ with inflow and outflow is reversed. This phenomenon is affected by station capacities R, since the effective production rate (inflow in pull or outflow in push) is a function of these capacities.

With regard to the third above conclusion, we can similarly explain why I_2 does not increase under push with increasing the station capacity. The reason is that increasing station capacity will at the same time increase the effective rate of outflow, with the ensuing balance keeping the level of material essentially constant.

A final comment on the above comparative study, already anticipated at the beginning of the present subsection, is this: While push has been identified as superior to pull in certain parametric situations, all comparisons has been based on the assumption of perfect prediction of the demand rate λ under push, which favors push, and thus makes the case for pull stronger than this study might suggest.

4.3 A New Simplified Method for Modeling Push Systems

As has been pointed out in Subsection 4.1, the usual model involves certain "blocking" states which add restrictions to the model equations, and complexity to their solution.

We have developed an alternative method of modeling queueing systems with possible blockage, that has reduced the number of states, and in general simplified the process of keeping track of blocking as it occurs. The proposed modeling can be described as follows:

- There are N stations, each station having limited capacity R_i at station i, and being equipped with R_i servers.

- A part arrives from supplier (time between arrivals is exponentially distributed) at a rate λ and the supply will stop if the capacity R_1 of the first station is reached.

- The part is processed (processing times are exponentially distributed) with rate δ_1 at the first station, if and only if, it (the part) can proceed immediately, after processing, to the second station. That is, the first station will look ahead to the second station and process only parts that can proceed to the second station without blocking the first station.

- The process continues in the analogous way up to the final station. At the final station N, a part is processed with rate δ_N and is able to leave immediately to shipping.

Analogously to our exposition of the previos model, Figure 4.9 shows the proposed model transaction diagram for a two-station system. The new model, also modeled by finite state-space over continuous time Markov process, with state space $S = \{(n_1, n_2) : 0 \leq n_1 \leq R_1 < \infty,\ 0 \leq n_2 \leq R_2 < \infty\}$. This model involves $(1 + R_1)(1 + R_2)$ states, and is therefore correspondingly simpler to analyze.

The stationary probabilities of these fewerstates now satisfy:

For $n_2 = 0, \ldots, R_2$ and $n_1 = R_1$

$$[\delta_2 * n_2 + \delta_1 * min(R_2 - n_2, R_1)]\, P_{R_1, n_2} =$$

$$\lambda P_{R_1 - 1, n_2}$$

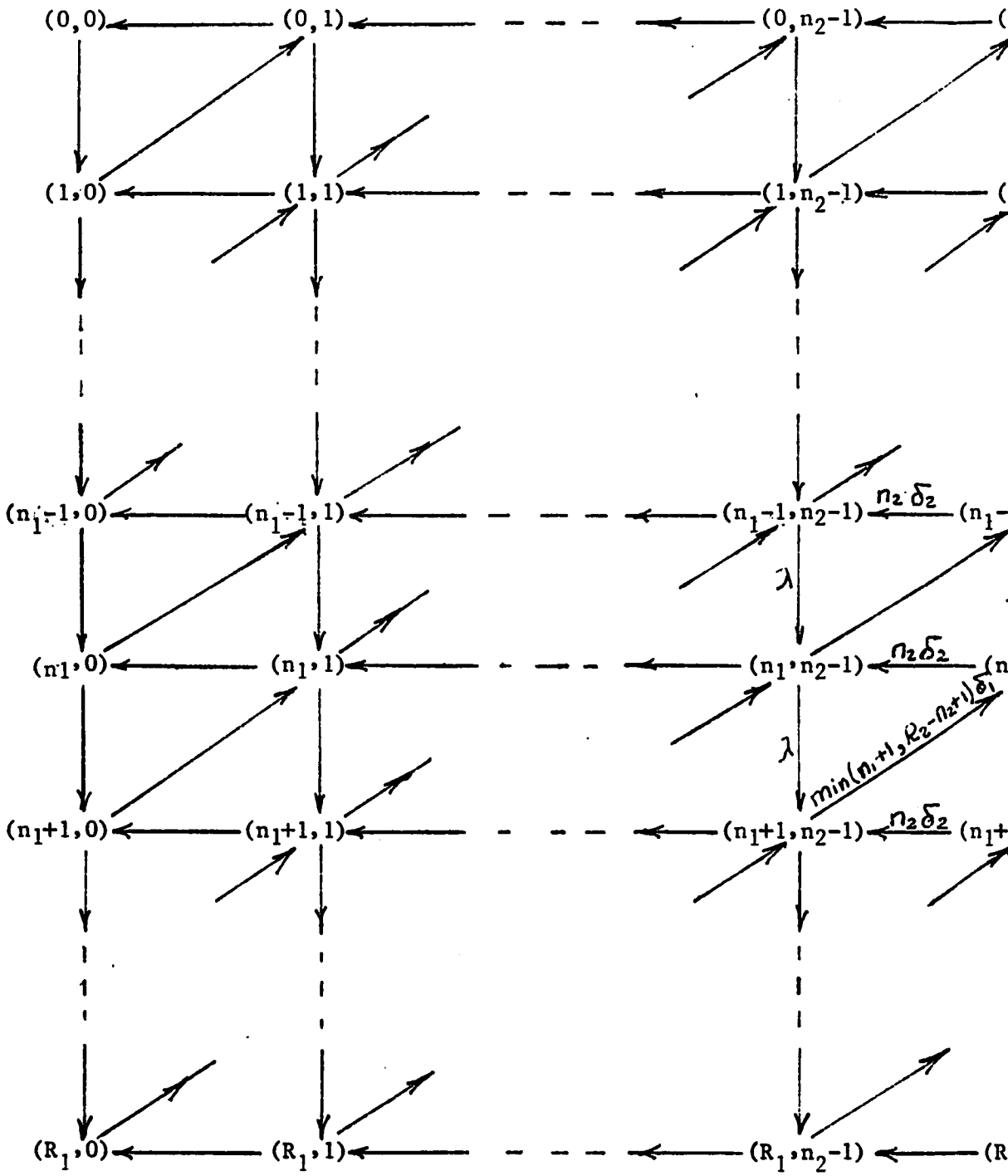
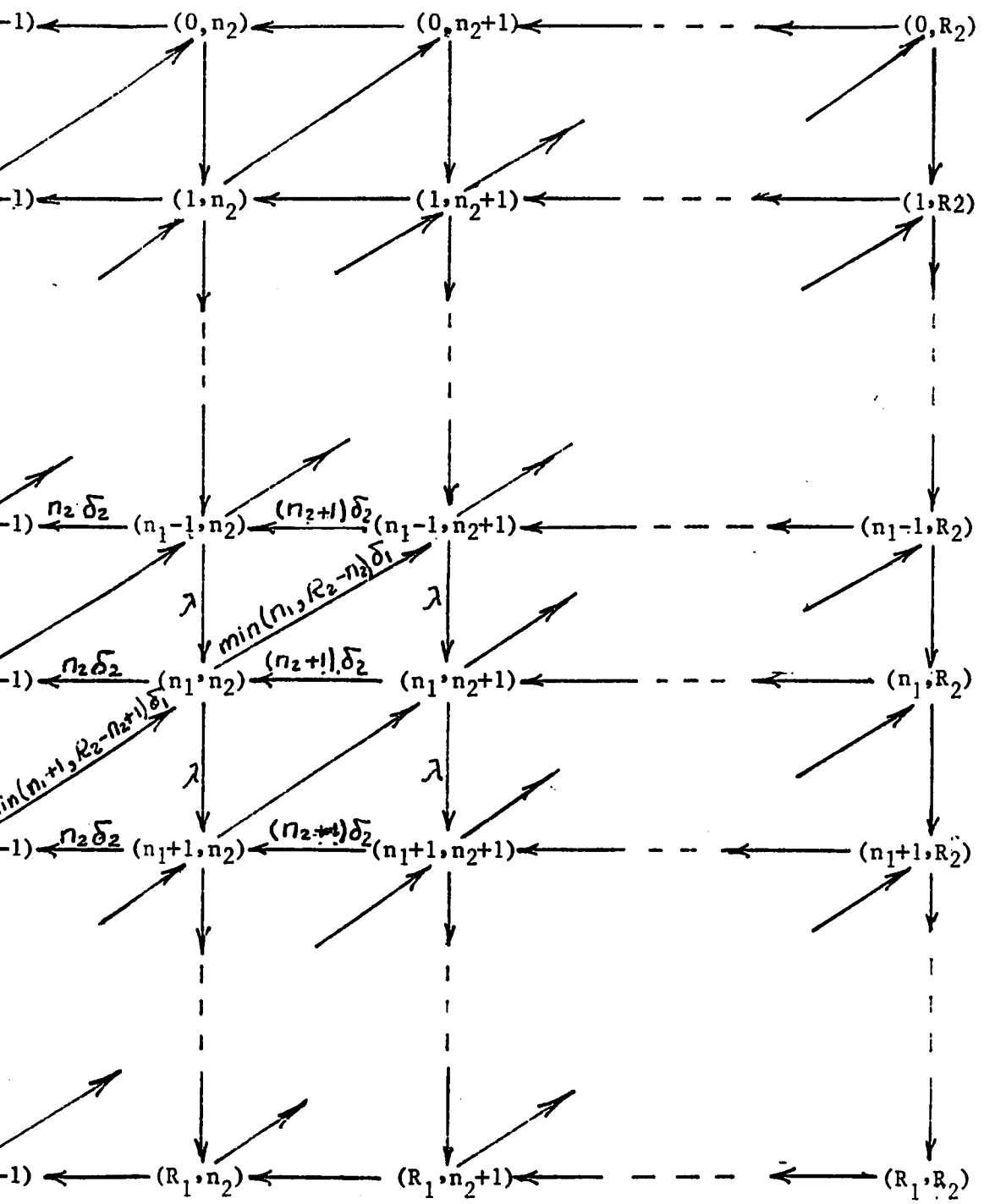


Figure 4.9: Transition diagram for two-station (proposed model) push system


(0, n2)
(0, n2+1)
(0, R2)
(1, n2)
(1, n2+1)
(1, R2)
(n1-1, n2)
(n1-1, n2+1)
(n1-1, R2)
(n1, n2)
(n1, n2+1)
(n1, R2)
(n1+1, n2)
(n1+1, n2+1)
(n1+1, R2)
(R1, n2)
(R1, n2+1)
(R1, R2)
n2 δ2
(n2+1) δ2
λ
min(n1, R2-n2) δ1
n2 δ2
(n2+1) δ2
λ
n2 δ2
(n2+1) δ2

ystem

$$+ \delta_2 * (n_2 + 1) P_{R_1, n_2+1} \tag{4.6}$$

For $n_1 = 0, 1, \ldots, R_1 - 1$ and $n_2 = 0, 1, \ldots, R_2$,

$$\begin{aligned}
&[\lambda + \delta_2 * n_2 + \delta_1 * min(R_2 - n_2, n_1)] P_{n_1, n_2} = \\
&\quad \lambda P_{n_1 - 1, n_2} \\
&\quad + \delta_2 * (n_2 + 1) P_{n_1, n_2+1} \\
&\quad + \delta_1 * min(R_2 - n_2 + 1, n_1 + 1) P_{n_1+1, n_2-1}.
\end{aligned} \tag{4.7}$$

The boundary conditions are

$$\begin{aligned}
P_{-1, n_2} &= 0 \\
P_{n_1, -1} &= 0 \\
P_{R_1+1, n_2} &= 0 \\
P_{n_1, R_2+1} &= 0
\end{aligned}$$

As in the case of the previous model, the stationary probabilities can be derived and the system performance measures of Section 2.5 calculated.

4.3.1 Results

In this section we will present a sample of our numerical results to show that this new method is not only very efficient in terms of computer time and storage, but also provides very accurate results.

As is shown in Table 4.2, there is excellent agreement between the performance measures based on the two models, under a variety of parametric conditions, at least in the case of two stations, and so our simpler modeling of the open queueing systems with blocking appears to be supportable on the basis of the evidence.

Table 4.2: System performance under a variety of parameters for the usual and proposed models

Parameters					Usual model			Proposed		
R_1	R_2	λ	δ_1	δ_2	P_{out_2}	I_2	SE	P_{out_2}	I_2	SE
4	6	.13	.40	.30	.6484	.4332	.8114	.6484	.4332	.8114
4	6	.13	.65	.05	.0736	2.5943	.3563	.0736	2.5943	.3563
1	3	.13	.20	.50	.8473	.1576	.9324	.8473	.1576	.9324
1	3	.13	.10	.60	.9070	.0942	.3345	.9070	.0942	.3345

5 DUALITY BETWEEN PUSH AND PULL

Duality in queueing theory has been discussed in the literature since the late 1950s. Foster[12] considered the number of customers as a dual to the number of waiting spaces, since the number of waiting spaces decrease by one as the number of customers increases by one. Gordon and Newell [15] defined the duality in tandem queueing as reversing the order of service in the primary system. Muth [27] discussed the reversibility property of production lines, proving that the total time for processing n dissimilar items through k dissimilar stations does not change when the order of the station is reversed.

In this chapter, the dual relationship between the pull system and a corresponding "specially defined" push system is investigated. A definition of the term "dual" pertaining to this case is presented and discussed.

5.1 Specially Defined Push system: (Dual Pull)

As stated before, units are "pushed" to the first "supplier side" station with pre-planned rate λ and processed in the subsequent stations till they leave the system at the final "customer side" station. The specially defined push system is built on the following assumptions:

1. The supply will stop if the capacity of the upper-stream station is reached.

2. A particular station will only process units that can be transferred directly after processing to the subsequent station. That is to avoid producing units that are not needed (the JIT concept).

3. The time between supplies and the processing times are exponentially distributed.

4. The last station can process any number of units without constraint (units will be removed and shipped, which is another JIT concept).

5. The transportation time between stations and the warm-up time of machines are negligible.

6. The supply time distribution is independent of the station processing time distributions.

7. The stations processing times are statistically independent of each other.

Now, the duality concept between the pull and push system is explained in the form of a binary comparison (see Table 5.1).

Figure 5.1 shows the transition diagram of a two-station specially defined push system, which could be compared to Figure 2.2 to illustrate the above comparison.

The equilibrium equations for this system are:

For $n_1 = 0, \ldots, R_1$

$$\begin{aligned}[\delta_1 * n_1 + \delta_2 * min(R_1 - n_1, R_2)]\, P_{n_1,R_2} = \\ \lambda P_{n_1,R_2-1} \\ + \delta_1 * (n_1 + 1) P_{n_1+1,R_2}\end{aligned} \tag{5.1}$$

Table 5.1: Duality concept between Pull and Push

PULL	PUSH
• The supply is infinite.	• The final storage is infinite.
• Demands stops when customer side station is empty.	• Supply stops when supplier side station is full.
• Units are "pulled" from final station at rate λ	• Units are "pushed" to the first station at rate λ
• The direction of flow of information is opposite to the flow of material.	• The direction of flow of information and material are the same.
• Supplier side station is station number one.	• Customer side station is station number one.
• The effective rate of replenishment of an intermediate station depends on the deficit at this station and the number of units at the previous one.	• The effective rate of prodction of an intermediate station depends on the number of units at this station and the deficit (free space) at the next one.
• The storage is positioned at the front of the processing facility.	• The storage is positioned at the back of the processing facility.

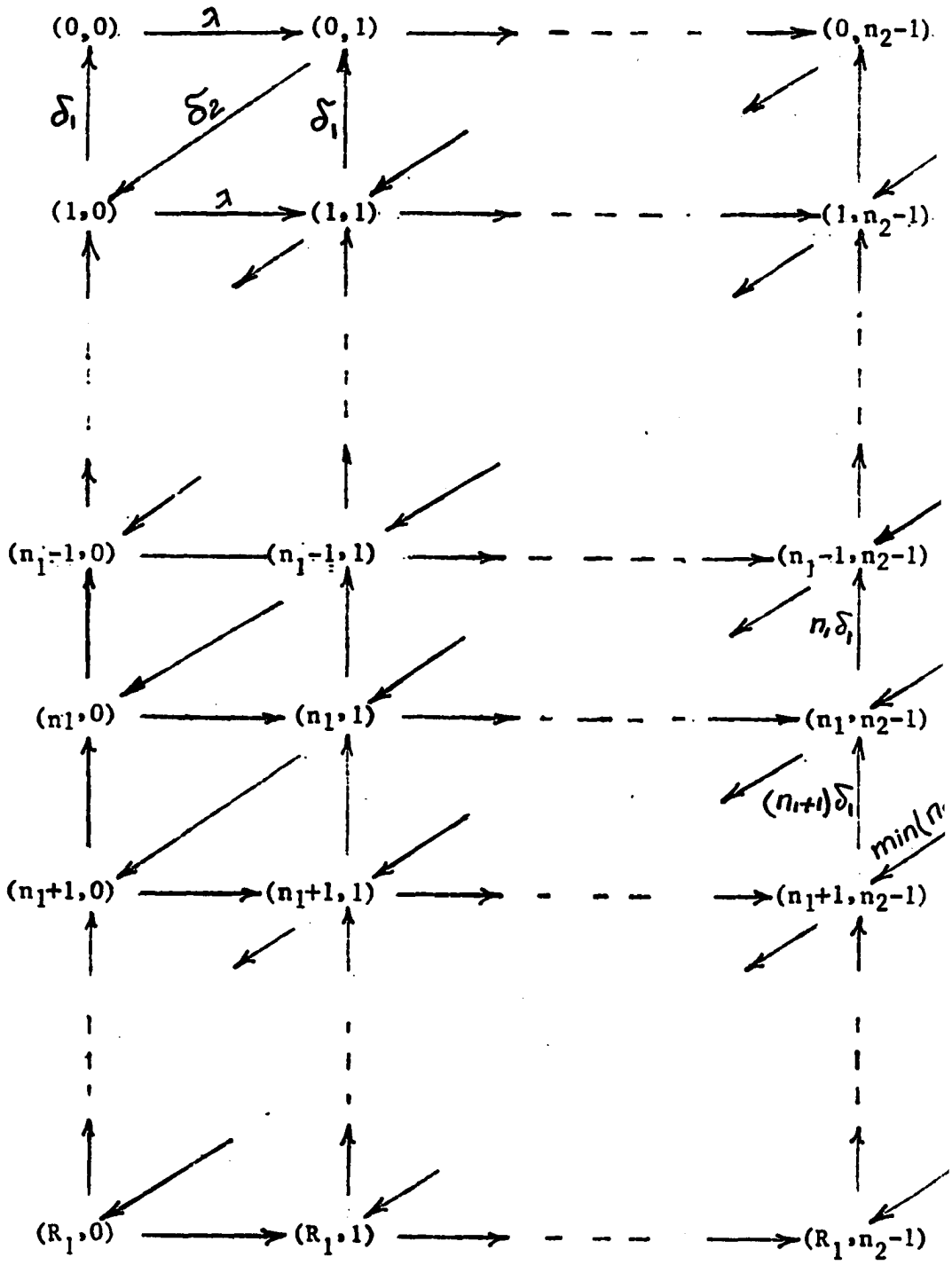
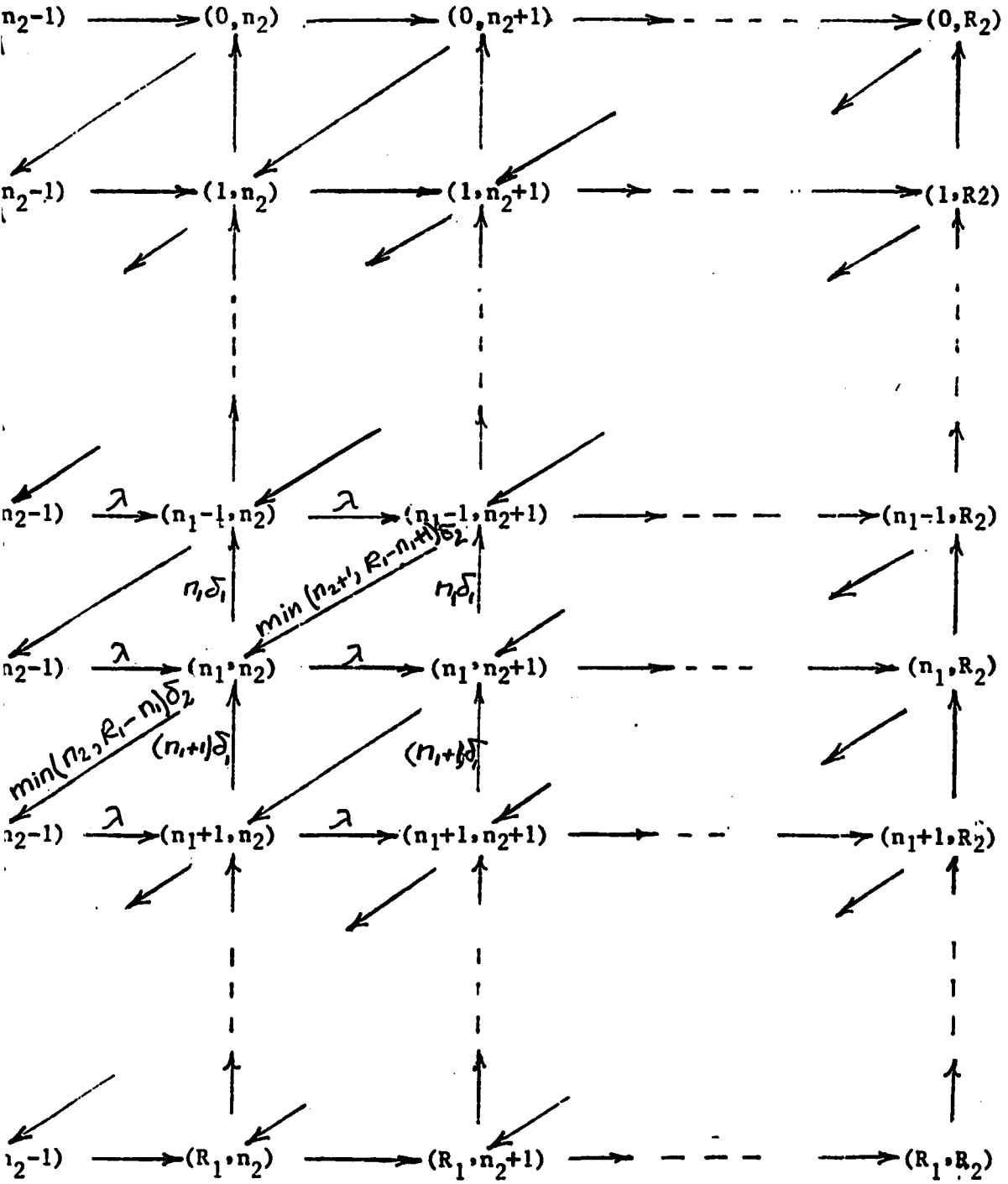


Figure 5.1: Transition diagram of two-station push (dual) system


$n_2-1)$ → $(0,n_2)$ → $(0,n_2+1)$ → - - - → $(0,R_2)$
$n_2-1)$ → $(1,n_2)$ → $(1,n_2+1)$ → - - - → $(1,R2)$
$n_2-1)$ λ → (n_1-1,n_2) λ → (n_1-1,n_2+1) → - - - → (n_1-1,R_2)
$n_1\delta_1$
$\min(n_2+1, R_1-n_1+1)\delta_2$
$n_1\delta_1$
$n_2-1)$ λ → (n_1,n_2) λ → (n_1,n_2+1) → - - - → (n_1,R_2)
$\min(n_2, R_1-n_1)\delta_2$
$(n_1+1)\delta_1$
$(n_1+1)\delta_1$
$n_2-1)$ λ → (n_1+1,n_2) λ → (n_1+1,n_2+1) → - - - → (n_1+1,R_2)
$n_2-1)$ → (R_1,n_2) → (R_1,n_2+1) → - - - → (R_1,R_2)

For $n_1 = 0, 1, \ldots, R_1$

and $n_2 = 0, 1, \ldots, R_2 - 1$

$$\begin{aligned}[\lambda + \delta_1 * n_1 + \delta_2 * min(R_1 - n_1, n_2)]P_{n_1,n_2} = \\ \lambda P_{n_1,n_2-1} \\ + \delta_1 * (n_1 + 1)P_{n_1+1,n_2} \\ + \delta_2 * min(R_1 - n_1 + 1, n_2 + 1)P_{n_1-1,n_2+1} \end{aligned} \tag{5.2}$$

The boundary equations are

$$\begin{aligned} P_{-1,n_2} &= 0 \\ P_{n_1,-1} &= 0 \\ P_{R_1+1,n_2} &= 0 \\ P_{n_1,R_2+1} &= 0 \end{aligned}$$

The above system is solved for the stationary joint probability distribution using any of the previously mentioned methods.

Comparing Figure 2.2 for the pull system with Figure 5.1 for the push system, and also comparing the joint probability distributions, We can deduce that the state joint probabilities for the two systems are complimentary in the sense that,

$$P_{n_1,n_2,\ldots,n_N} = P_{R_1-n_1,R_2-n_2,\ldots,R_N-n_N}$$

Complementary is extended to performance measures in the following three sub-sections.

5.1.1 The Extremal Probabilities at Extremal Stations

In the pull system this measure is expressed as:

$$P_{out_N} = \sum_{n_{N-1}} \cdots \sum_{n_1} P_{n_1,n_2,\ldots,0}$$

Substituting their compliments for the joint probabilities yields

$$\begin{aligned} P_{out_N} &= \sum_{n_{N-1}} \cdots \sum_{n_1} P_{R_1-n_1,R_2-n_2,\ldots,R_N} \\ &= \sum_{n_{N-1}} \cdots \sum_{n_1} P_{n_1,n_2,\ldots,R_N} \end{aligned}$$

So, we can say

$$P_{full_N} = \sum_{n_{N-1}} \cdots \sum_{n_1} P_{n_1,n_2,\ldots,R_N}$$

which is the equivalent measure for the push system. The interpretation here is that just as we can't fill any demand if the final station is out of stock in the pull system, we can't accept any "shipment" if the first station is full in the push system.

5.1.2 The Average Number of Units at Extremal Stations

In the pull system this measure is expressed as:

$$I_N = \sum_{n_N} \cdots \sum_{n_1} n_N P_{n_1,n_2,\ldots,n_N}$$

Substituting their compliments for the joint probabilities yields,

$$\begin{aligned} I_N &= \sum_{n_N} \cdots \sum_{n_1} n_N P_{R_1-n_1,R_2-n_2,\ldots,R_N-n_N} \\ &= R_N - \sum_{n_N} n_N \cdots \sum_{n_1} (R_N - n_N) P_{R_1-n_1,R_2-n_2,\ldots,R_N-n_N} \end{aligned}$$

So, we can say

$$I_N = R_N - \sum_{n_N} \cdots \sum_{n_1} n_N P_{n_1,n_2,\ldots,n_N}$$

The second term in the above equation is the average number of units at the first staion for the push system, which is complementary of the average number of units at the final station for the pull system.

5.1.3 System Responsiveness

This measure is defined in the pull system as:

$$SE = \frac{\delta_1(Prob\ \ n_1 < R_1)}{\lambda}$$

$$= \frac{\delta_1}{\lambda} \cdot \sum_{n_1=0}^{R_1-1} \sum_{n_2=0}^{R_2} \cdots \sum_{n_N=0}^{R_N} P_{n_1,n_2,\ldots,n_N}$$

Substituting their compliments for the joint probabilities yields,

$$SE = \frac{\delta_1}{\lambda} \sum_{n_1=0}^{R_1-1} \sum_{n_2=0}^{R_2} \cdots \sum_{n_N=0}^{R_N} P_{R_1-n_1,R_2-n_2,\ldots,R_N-n_N}$$

$$= \frac{\delta_1}{\lambda} \sum_{n_1=0}^{R_1-1} \sum_{n_2=0}^{R_2} \cdots \sum_{n_N=0}^{R_N} P_{R_1-n_1,n_2,\cdots,n_N}$$

$$= \frac{\delta_1(Prob\ \ n_1 > 0)}{\lambda}$$

which is the equivalent measure for the push system. It is the ratio of the effective production rate at station one to the supply rate.

6 TREE STRUCTURES UNDER PULL

In the previous chapters, the series structure was employed in all our modeling. The series structure models, conveniently, the transfer and the assembly lines where machine or assembly processes take place in sequence.

However, there also are situations, often traceable to management decision to make certain parts in-house, where machine and assembly processes in effect form confluent tree-shaped structures, in which the parts successively are made, and fed to sub-assembly, and then assembly, stations (see Figure 6.1).

6.1 The Model

The above situation will be modeled under the same basic assumptions as those for the series model (see Section 2.2.1). Additional assumptions concerning the tree structure are as follows:

1. If the product associated with station i is called product i, then assume that one unit (container) of product i and one of product j are required to make one unit (container) of product ij. The same is true for products ij and k with respect to product ijk. See [5].

2. Assume that when a demand withdraws a unit from station ijk, a signal (Kanban) is sent to both station k and ij requiring replenishment. The two

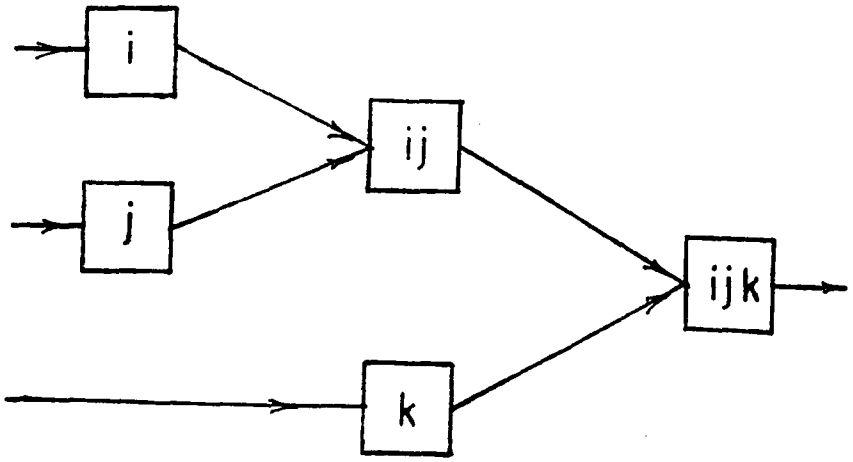


Figure 6.1: Tree structure of 5 stages

corresponding units are moved without delay to the production area of station ijk, and after an average time of $1/\delta_{ijk}$, the newly assembled unit is deposited in the local storage of station ijk. The analogous modeling assumption applies to the assembly of products i and j at station ij. (A technical feature of the stochastic modeling here is that the two component units of a unit under assembly are seen as belonging to the two origin stations until assembly is complete.)

Figure 6.3 illustrates the transition diagram for the simple tree structure shown in Figure 6.2, with maximum station capacity of R equal to two at each.

The general stationary equations for this simple tree structure are shown below.

For $n_1 = R_1, R_1 - 1, \ldots, 0$, $n_2 = R_2, R_2 - 1, \ldots, 0$

and $n_3 = 0$

$$\begin{aligned}[\delta_1(R_1 - n_1) + \delta_2(R_2 - n_2) + \delta_3 min(n_1, n_2, R_3 - n_3)]\, P_{n_1,n_2,0} = \\ \delta_1 (R_1 - (n_1 - 1))\, P_{n_1-1,n_2,0} \\ + \delta_2 (R_2 - (n_2 - 1))\, P_{n_1,n_2-1,0}\end{aligned}$$

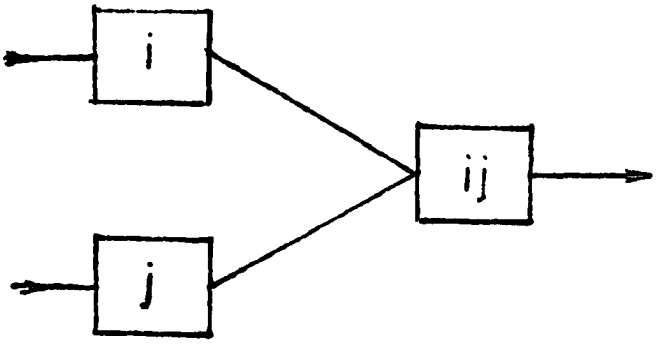


Figure 6.2: Simple tree structure of 3 stages

$$+ \lambda P_{n_1,n_2,1} \tag{6.1}$$

For $n_1 = R_1, R_1 - 1, \ldots, 0 \quad , n_2 = R_2, R_2 - 1, \ldots, 0$

and $n_3 = R_3, R_3 - 1, \ldots, 1$

$$\begin{aligned}[\lambda + \delta_1(R_1 - n_1) + \delta_2(R_2 - n_2) + \delta_3 min(n_1, n_2, R_3 - n_3)]\, P_{n_1,n_2} = \\ \delta_1(R_1 - n_1 + 1)P_{n_1-1,n_2,n_3} \\ + \delta_2(R_2 - n_2 + 1)P_{n_1,n_2-1,n_3} \\ + \delta_3 min(n_1 + 1, n_2 + 1, R_3 - n_3 + 1)Pn_1 + 1, n_2 + 1, n_3 - 1 \\ + \lambda P_{n_1,n_2,n_3+1} \end{aligned} \tag{6.2}$$

and the boundary conditions are:

$$\begin{aligned} P_{-1,n_2,n_3} &= 0 \\ P_{n_1,-1,n_3} &= 0 \\ P_{n_1,n_2,-1} &= 0 \\ P_{R_1+1,n_2,n_3} &= 0 \\ P_{n_1,R_2+1,n_3} &= 0 \\ P_{n_1,n_2,R_3+1} &= 0 \end{aligned}$$

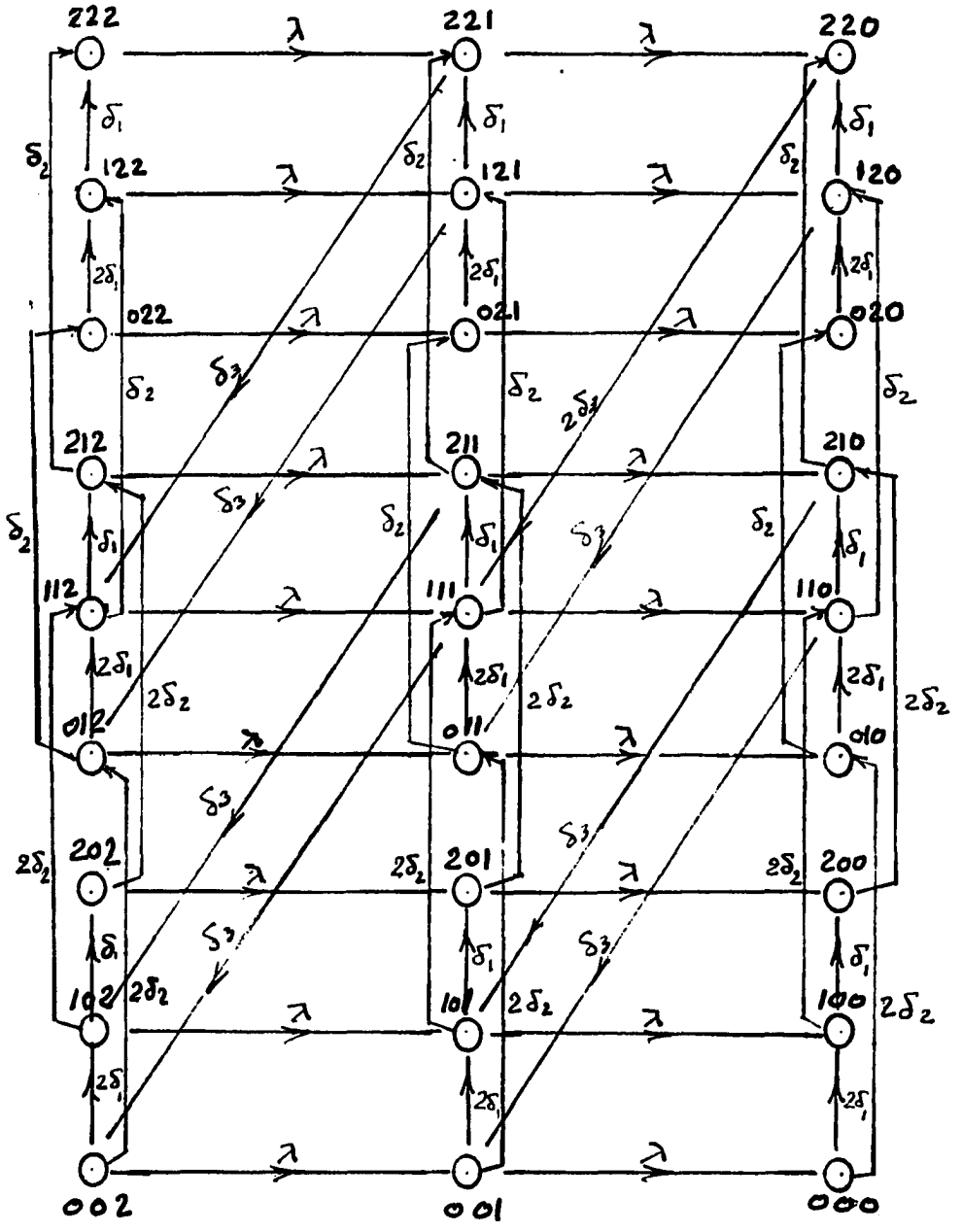


Figure 6.3: Transition diagram of the simple Tree structure

6.2 Production Optimization of Confluent Configurations

Given these fixed capacities, the question remains of assigning production rates to each station. Well designed production rates will assure the smooth flow of product, with the final station's local storage rarely out of stock. As in Section 2.7.2, the constraint of constant sum of production rates is used, with the Hook and Jeeves algorithm employed for finding optimum rates, and the results verified using a quasi-Newton algorithm. Following are the results and comments pertaining to the configuration given in Figures 6.1, 6.2, 6.4 and 6.5.

6.2.1 Symmetric Configuration

Generally, with a simple system as in Figure 6.2, or a more branched "symmetric" one, as in Figure 6.4, the probability of stockout is minimized when the production rates of stations on the same level (as i and j) are "balanced" (i.e., equal if the station capacities are equal, and, in the case of unequal capacities at the origin stations, a lower optimum production rate assigned to the larger-capacity station).

6.2.2 Asymmetric Configurations

In the case of asymmetric layouts (see Figures 6.1, 6.5), the relationships are slightly different. The optimum production rates at the equal-capacity stations ij and k of Figure 6.1, or at stations m and n of Figure 6.5, are not equal as before. The reason is to be found in the relative position of these stations in the system. The stations that are preceded by other stations in the line need higher production rates to overcome fluctuations in the duration of production in preceding stations.

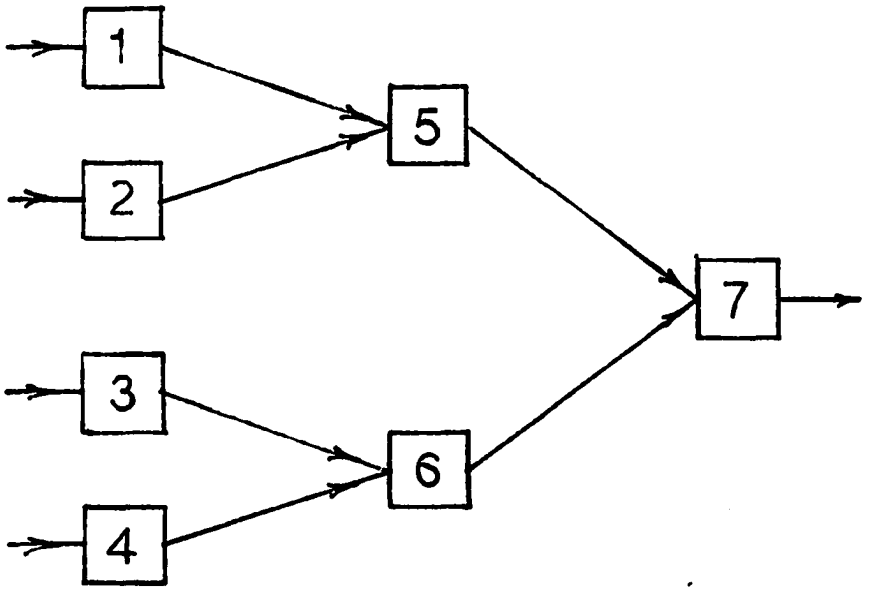


Figure 6.4: Seven stages tree structure

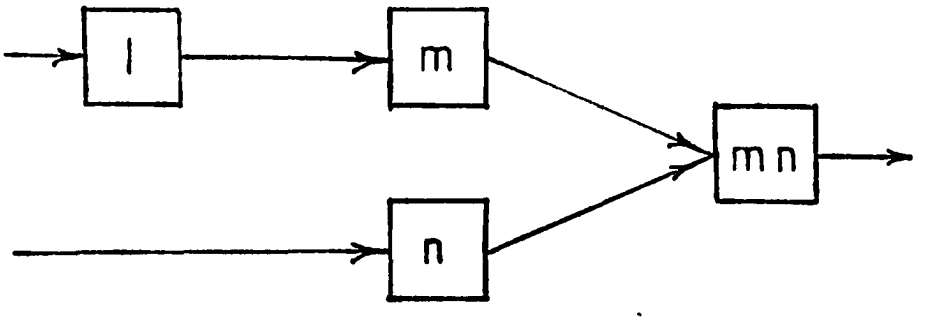


Figure 6.5: Four stages tree structure

Table 6.1 shows the optimum production rates for the five stations of Figure 6.1.

6.2.3 Sub-Configuration Modules

"Separability" is always to be hoped for. To what extend, then, is it possible to analyze sub-configurations or "modules", in order to predict their behavior as part of the larger configuration? To explore this effect, the optimum allocations of production rates among the five stations of Figure 6.1 and the four stations of Figure 6.5, are found, using the previously mentioned algorithms. The second step is to fix the sum of the optimum production rates of stations i, j and ij (Figure 6.1) at its optimum value, and then to optimize the three production rates of the corresponding (i, j, ij)-module (Figure 6.2) The analogous steps are taken for two-station (l, m)-module of Figure 6.5, and the results are recorded in Tables 6.2, 6.3. From these results we deduce that:

1. the Modular analysis appear, at least for small systems, to reproduce reasonably well the character of the underlying non-modular analysis. In particular,

2. Symmetric modules appear to retain the "balance" features of optimal production rates obtained under non-modular analysis.

3. Series modules appear to retain the "funnel effect" obtained under non-modular analysis.

Table 6.1: Optimum production rates for five station tree configuration

	$\delta_1 = \delta_2$	δ_3	δ_4	δ_5
$R_1 = R_2 = R_3 = R_4 = R_5 = 1$	.116	.153	.115	.150
$R_1 = R_2 = R_3 = R_4 = 1,\ R_5 = 2$	.119	.155	.116	.142
$R_1 = R_2 = 2,\ R_3 = R_4 = R_5 = 1$	.072	.163	.147	.196
$R_1 = R_2 = R_3 = 2,\ R_4 = R_5 = 1$	.080	.120	.163	.206
$R_1 = R_2 = 1,\ R_3 = R_4 = R_5 = 2$	.147	.178	.075	.103
$R_1 = R_2 = R_5 = 1.\ R_3 = R_4 = 2$	.135	.164	.068	.147

Table 6.2: Modular and non-modular analysis for a three-station module of a five-station configuration

				Five Stages			Three Stages		
R_1	R_2	R_3	$\sum_i \delta_i$	δ_1	δ_2	δ_3	δ_1	δ_2	δ_3
1	1	1	.3928	.1188	.1188	.1553	.1203	.1203	.1522
1	1	1	.3847	.1159	.1159	.1528	.1178	.1178	.1491
2	2	2	.2800	.0803	.0803	.1200	.0828	.0828	.1144
2	2	1	.3072	.0722	.0722	.1628	.0753	.0753	.1566
2	2	1	.3200	.0769	.0763	.1669	.0781	.0781	.1638
1	1	2	.4712	.1469	.1469	.1775	.1481	.1481	.1753

Table 6.3: Modular and non-modular analysis for a two-station module of a four-station configuration

			Four Stages		Two Stages	
R_1	R_2	$\sum_i \delta_i$	δ_1	δ_2	δ_1	δ_2
1	2	.4641	.2231	.2409	.2216	.2425
2	1	.3034	.0941	.2094	.0975	.2059
2	1	.3766	.1216	.2550	.1156	.2609
2	2	.3491	.1503	.1988	.1469	.2022
1	1	.4425	.2025	.2400	.1966	.2459

7 CONCLUSIONS AND EXTENSIONS

In this thesis we have developed a stochastic model of the multi-stations pull production system. Attention was restricted to Markov continuous-time modeling, with work stations imagined to consist of a processing function and a local storage function of given capacity (R). General stochastic modeling assumptions were as follows:

1. Processing times are exponentially distributed, with effective processing rate responsive to local storage levels.

2. Time between demand pulls or supply pushes is exponentially distributed with fixed rate.

3. The transportation time between stations, and warmup time of machines, are negligible.

Modeling variants studied within the above shared general frame work were:

1. work station discipline

   - One-at-a-time
   - Several-at-a-time

2. process configuration

- Series
- Confluent (tree structure)

3. Lot size
   - Lot size $= 1$
   - Lot size ¿ 1

Although processing and demand times are assumed exponential, the limitation imposed by storage capacity will cause the output process not to be poisson. For this reason, closed form solutions for equilibrium probabilities of the system were not available and numerical methods were investigated, and compared.

The model was then studied and analyzed in the light of the following measures of system performance.

- Probability that the last station is out of stock,
- Mean number of units at the last station,
- System responsiveness, which is the effective production rate at the last station in the case of pull, and at the first station in the case of push.

Since allocation of resources is an important issue in production planning, optimal allocation of station capacity R_i (Knaban units) and production capacity were investigated. Regarding station capacities, the optimum allocation of R_i exhibits a "funnel pattern", with higher capacities allocated downstream. Further, uniform

allocation can be achieved by controlling the ratio of production to demand rates. Regarding production rates (work elements and personnel), high system performance is achieved by assignments of work elements and/or personnel uniformly among stations if the ratio of the average production rate to the demand rate is less or equal to 1. However, with the same ratio greater than one, the optimum allocation tend to be achieved by assigning higher production rates toward the downstream station (another "funnel effect").

When the several-at-a-time discipline is compared to one-at-a-time, the first showed lower probabilities of stockout and a higher average number of units at the downstream station, while the second showed better system responsiveness.

With respect to the effect of lot size, is studied, it was found that economic considerations must be taken into account in choosing the proper lot size.

In this study, a "confluent" production system was taken to be one where "made" parts are processed in-house into subassemblies, and then into final assemblies. Optimum resource allocations were studied for confluent systems, and the possibility explored of analyzing a confluent production system through its sub-system "modules."

The traditional "push" system, was also modeled under the same stochastic assumptions. Here it was assumed that the demand rate is accurately forecast, and that material is pushed at a corresponding rate to the stations downstream. In comparing the push and the pull systems, it was concluded that:

1. Pull systems tend to outperform push systems.

2. However, with low production rates (lower than demand (for pull) or supply (for push)rate), that is, in the case of "under powered" systems, and for small station capacities R, push systems can outperform pull systems.

A duality phenomenon between the pull and a specially defined push model was presented and discussed.

Extensions During the course of this study numerous other questions arose and remain unanswered. Some of these are given below:

1. Investigation of models with non-exponential production time distributions.

2. When optimum resource allocations were studied, these were separated into two problems, one a discrete problem concerning the optimum allocation of station capacities and the other a continuous problem concerning optimum allocation of production rates. An algorithm to combine both optimization problems in one might be a possibility. See [14].

3. Define other performance measures, in terms of system utilization or waiting times and investigate optimum resource allocations in terms of these, as well as comparing push to pull systems in the light of these new performance measures.

4. Expand on the idea of modularity in tree structures, and how it might be utilized in simplifying the solution of branched systems.

8 BIBLIOGRAPHY

[1] Altiok, T. "Approximate Analysis of Exponential Tandem Queues with Blocking." European Journal of Operation Research 11, 390-398, 1982.

[2] Altiok, T. and Stidham, S. "The Allocation of Interstage Buffer Capacities in Production Lines." IIE Transactions 13, No. 4, 292-299, 1983.

[3] Avi-itzhak, B. "A Sequence of Service Stations with Arbitrary Input and Regular Service Time." Management Sci. 11, 565-571, 1965.

[4] Barten, K., "Queueing Simulator for Determining Optimum Inventory Levels in a Sequential Process." The Journal of Industrial Engineering 13, No.4, 245-252, 1962.

[5] Bitron, G. R. and Chang, L. " An Optimization Approach to the Kanban System." WP No. 1635-85. MIT, Cambridge, Mass., 1985.

[6] Black, J. T. "The Design of Manufacturing Cells (Step One to Integrated Manufacturing Systems)." Proc. of Manufacturing International '88 vol iii, 143-158, Atlanta, Ga, Apr. 19, 1988.

[7] Buzacott, J. A. "Modeling Automated Manufacturing System." Fall Industrial Engineering Conference Proceedings, 1983.

[8] Buzacott, J. A. and Kostelski, D. "Matrix-geometric and Recursive Algorithm Solution of a Two-Stage Unreliable Flow Line." IIE Transactions 29, No. 4, 429-438, 1987.

[9] Cooper, R. B. "Introduction to Queueing Theory." 2nd ed. North Holland, Inc., New York, 1981.

[10] Elsayed, E. A. and Hwang, C. C. "Analysis of Manufacturing Systems with Buffer Storage and Redundant Machines." Proceedings of 7th International Conference on Production Research, 204-210, 1985.

[11] Faddev, D. K. and Faddeeva, V. N. "Computational Methods of Linear Algebra". Translated by Robert C. Williams. W. H. Freeman and company, San Francisco, 1963.

[12] Foster, F. G. "A Unified Theory for stock, storage and queue control." Operation Research Q., 10, 121-130, 1959.

[13] Foster, F. G. and Perros, H. G. "On the Blocking process in Queue Network." European Journal of Operational Research 5, 276-283, 1980.

[14] Gopal, K.; Aggarwal, K. K. and Gupta, J. S. " A New Method for Solving Reliability Optimization Problem." IEEE Transactions on Reliability, R-29, 36-37, April 1980.

[15] Gordon, W. J. and Newell, G. F. "Cyclic Queueing Systems with Restricted length queues." Operations Research, 15, 266-277, 1967.

[16] Gross, D. and Harris, C. " Fundamental of Queueing Theory." John Wiley & Sons, New York, 1985.

[17] Hall,R. W. "Zero Inventory Crusade- Much More Than Materials Management." Production Inventory Management 24, No. 3, third Quarter, 1-9, 1983.

[18] Hadley, G. and Whitin, T. M. "Analysis of Inventory Systems." Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

[19] Heard, R. "Pull Can Always Use a Push." Zero Inventory Philosophy and Practices Seminar Proceedings, St. Louis, Missouri, 1984.

[20] Hillier, F. S. and Boling, R. W. "Finite Queues in Series with Exponential or Erlang Service Times- A Numerical Approach." Operation Research 15, 286-303, 1967.

[21] Hook, R. and Jeeves, T. A. "Direct Search Solution of Numerical and Statistical Problems." Assoc. for Comp. Mach. J. 8, 212, 1961.

[22] Hunt, G. C. "Sequential Arrays of Waiting lines." Operation Research 6, 674-683, 1956.

[23] Konheim, G. and Reiser, M. "A Queueing Model with Finite Waiting Room and Blocking." J. ACM 23, 328-341, 1976.

[24] Love, R. F. "A Two Station Stochastic Inventory Model with Exact Methods of Computing Optimal Policies." Naval Research Logistic Quarterly 14 , 185-217, 1967.

[25] Monden, Y. "Toyota Production System." IE Press, Norcross, GA., 1983.

[26] Morse,P. M. "Queues, Inventories, and Maintenance." John Wiley, New York, 1958.

[27] Muth, E. J. "The Reversibility Property of Production lines." Management Science, 25, 152-158, 1979.

[28] Nakane, J. and Hall, R. W. "Transferring Production Control Methods Between Japan and the United States.", APICS, 24th Annual International Conference Proceedings, Boston, MA, 124-128.

[29] Nellemann, D. O. "MRP vs Kanban? Combining the Best of the East and West." 25th Annual International conference Proceedings, Chicago, Ill, 124-128.

[30] Neuts, M. F. "Two Queues in series with a Finite Intermediate Waiting Room." J. Appl. Probability 5, 123-142, 1968

[31] Schoberger, R. J. "Kanban (Just-In-Time) Applications at Kawasaki, USA", APICS, 24th Annual International Conference Proceedings, Boston, MA, 1981.

[32] Schroer, B. J.; Black, J. T. and Zhang, S. X. "JIT, with Kanban, Manufacturing System Simulation on a Microcomputer." Simulation 3, 62-70, Aug. 1985.

[33] Scott, M. "Characterization of Strong Ergodicity for Continuous Time Markov Chains". Ph.D. dissertation, Iowa State University, Ames, Iowa, 1979.

[34] Solberg, J. J. "Capacity Planning with a Stochastic Workflow Model." AIIE Transaction 13, No. 2, 116-122, 1981.

[35] Sugimri, Y.; Kusunoki, K.; Cho, F. and Uchikawa, S. " Toyota Production System and Kanban System: Materialization of a Just-In-Time and Respect-for-Human System." International Journal of Production Research 15(6), 553-564, 1977.

[36] Terada, H. and Kimura, O. " Design and Analysis of Pull System, A Method of Multi-Stage Production Control." International Journal of Production Research 19, No. 3, 241-253, 1981.

[37] Wong, B., Giffin, W. and Disney, R. "Two Finite M/M/1 Queue in Tandem: A Matrix Solution for the Steady State." Opsearch 14, No. 1, 1-18, 1977.

[38] Yong, P. L. "Some Results Related to Q-Bounded Markov Proces." Nanta Mathematica 8:34-41, 1976.