Static and Dynamic Scheduling in a Two Station Mixed Queuing Network

by

Kamala P. Murti

M.A., Oxford University (1990)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science
in
Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1994

Author ..... Kamala Murti .....
Department of Electrical Engineering and Computer Science
May 13, 1994

Certified by ..... Viên Nguyen .....
Viên Nguyen
Assistant Professor, Sloan School of Management
Thesis Supervisor

Accepted by ..... Richard C. Larson .....
Richard C. Larson
Co-Director, Operations Research Center and
Professor, Department of Electrical Engineering

Eng.

Static and Dynamic Scheduling in a Two Station Mixed Queuing Network

by

Kamala P. Murti

Submitted to the Department of Electrical Engineering and Computer Science
on May 13, 1994, in partial fulfillment of the
requirements for the degree of
Master of Science
in
Operations Research

Abstract

This thesis studies static and dynamic scheduling in a two-station mixed queuing network in which two classes of items, make-to-order and make-to-stock, are produced at one production facility. Make-to-order items are modeled as 'open' jobs which enter the production facility (station 1), are processed and then leave the system. Make-to-stock items are modeled as 'closed' jobs which are produced at the same production facility and are held in finished goods inventory (station 2) from which demand is met. Production of make-to-stock products follows a one-for-one replenishment policy. Demand for make-to-stock items cannot be backlogged; that is, demand that cannot be met from finished goods inventory is simply lost. Demands follow a Poisson process, service times are exponentially distributed and there are no setup costs or delays when switching from one product to the other. In the static scheduling scenario we develop expressions for the optimal stock level to hold in finished goods inventory which would either minimize the average cost incurred by the system per unit time, or alternatively which would enable a certain percentage of the demand to be met. In the dynamic scheduling case we formulate the optimal policy which determines the priority in which to produce the items in order to minimize the long term discounted cost of the system. We find that the optimal policy takes the form of a switching curve.

Thesis Supervisor: Viên Nguyen
Title: Assistant Professor, Sloan School of Management

Acknowledgements

I would like to thank my thesis advisor, Viên Nguyen, for sparking my interest in queuing theory and for her valuable insight and inspiration throughout the course of this research. I also wish to acknowledge the Leaders for Manufacturing Program for its support of this work.

My thanks to all the students, faculty and staff at the Operations Research Center who have made these past two years enjoyable. Special thanks to Sarah, Rodrigo, Elaine, Christian and Chung Piaw for memorable times.

My heartfelt thanks go to my parents and my parents-in-law who have sacrificed a great deal to help me succeed in this endeavour. Thanks to Indra and Vaidhyanath who have been extremely tolerant of their mother who would, at times, read queuing literature to them instead of bedtime stories. Finally, to my husband, Vasu, who encouraged me to embark on this Master's program - many, many thanks, I couldn't have done it without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 A Make-to-Order and Make-to-Stock Production System

The question of when and how much to produce is of utmost importance in any production facility. Many manufacturing systems schedule production as soon as an order is placed for a specific product. This would make sense particularly if the product is perishable, or customized. Such a system is a make-to-order system, where an order placed triggers production of the good. However, in a pure make-to-order system, the customer would have to wait from the moment of ordering until the product is manufactured to have the order fulfilled. This may not be the best situation from the customer's point of view, especially if the lead time is long. From the production system point of view it may also be economically unsound to schedule production only after an order is placed, for several reasons. For example, there may be uncertainty in demand, leading to periods of idleness and then periods of overload when all demand cannot be met. There may be uncertainty in the delivery of raw material, seasonal fluctuations in the price of raw materials and, of course, uncertainty in machine breakdowns. Considering these factors, for a product that is not customized, it may be advantageous for the facility to hold some finished goods inventory from which to satisfy demand, and to schedule production appropriately to replenish the stock. Such a system is a make-to-stock production system.

Figure 1-1 is a schematic diagram of the production facility to be studied in this

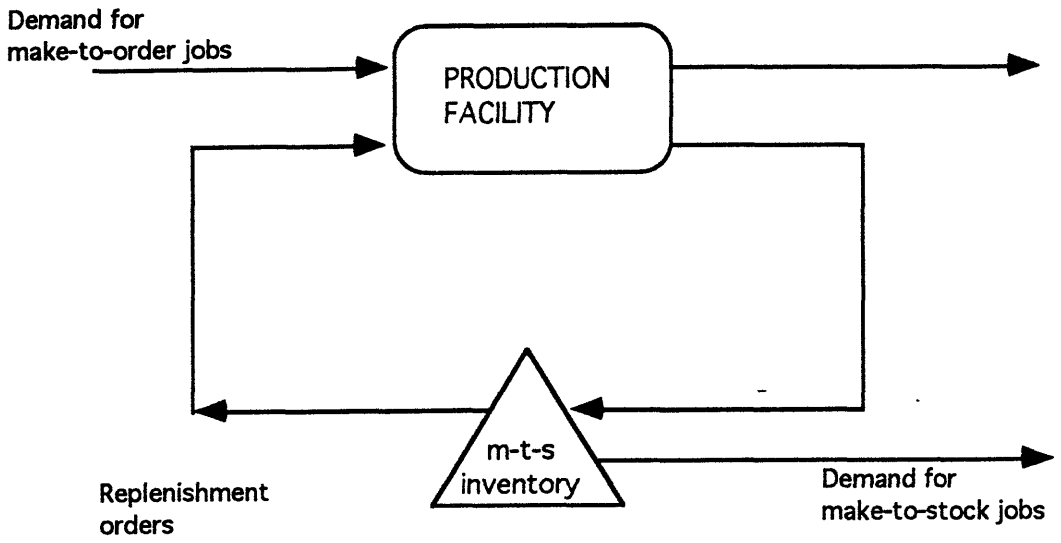

Figure 1-1: The Production System

thesis. The facility produces two types of items, one made to order and the other made to stock. Each demand for a make-to-order item initiates production of that item. Production of the make-to-stock items, on the other hand, follows a base-stock policy in the following sense. Demands for make-to-stock products are filled from inventory. Each filled demand triggers a replenishment order to restore the inventory to a target base-stock level. (Such a production policy has also been described as an (S-1,S) inventory policy, or a "one-for-one" replenishment system).

If a make-to-stock demand occurs when the inventory is empty, that request is simply lost; that is, sales that cannot be filled from inventory are turned away rather than backlogged. Observe that in this scenario, the sum of items in inventory and outstanding replenishment orders remains constant over time and is equal to the target base-stock level.

1.2 A Queuing Network Model

Queuing networks are extensively utilized to aid in the mathematical analysis of production facilities. We model the system described in the previous section as the mixed queuing network depicted in Figure 1-2. Two types of jobs are to be processed in this system. Demand for a make-to-order product is modeled as an arrival of an

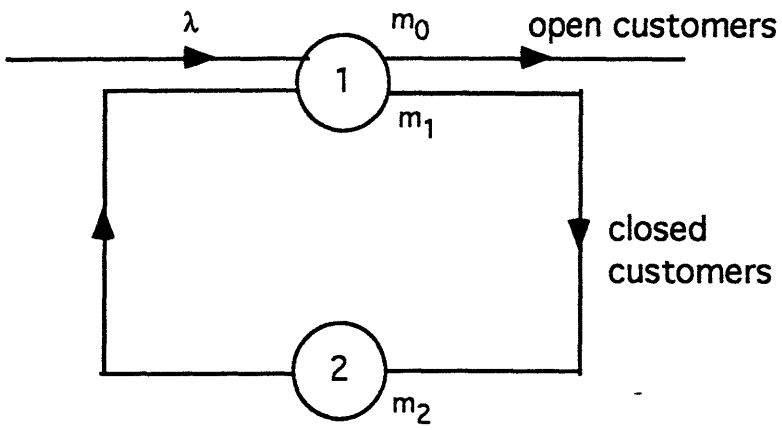

Figure 1-2: A Two-Station Mixed Queuing Network

"open" job to station 1, and that demand is fulfilled when that job completes service at station 1. Thus a queue of open jobs at station 1 represents the number of make-to-order products waiting to be processed. Make-to-stock products are modeled as a fixed number of "closed" jobs circulating between the two stations. A service at station 2 means that one unit of demand has been filled from finished goods inventory, which in turn initiates a request for a make-to-stock product at station 1. The fixed number of closed jobs represents the target base-stock level to be maintained in finished goods inventory. Thus, a queue at station 2, including the job in service, represents the number in finished goods inventory, from which demand for make-to-stock products has to be met. A queue of closed jobs at station 1 represents replenishment orders; i.e., the number of make-to-stock products waiting to be processed in order to restore the level of the finished goods inventory to the target level.

Let us now introduce the notation. Open jobs arrive at station 1 at rate λ and are processed at station 1 with a mean time of m_0. There are N closed jobs circulating in the system which are processed at station 1 with a mean time of m_1 and at station 2 with mean time m_2.

Throughout this thesis we assume Poisson arrivals and exponential service at each station. Note that exponential service at station 2 indicates a Poisson arrival process for the demand of make-to-stock products. As stated in Section 1.1, we assume no backlogged orders, so any demand arriving when there is zero finished goods inventory will be lost. This poses the question of whether the service epochs

at station 2 faithfully model the demand process. As in Nguyen [12], consider a time period $[t_1, t_2]$ during which the finished goods inventory is empty. All demand arriving during $[t_1, t_2]$ will be lost. The first demand to be fulfilled would have arrived after t_2. If this order arrived at t^* then $t^* - t_2$ may have a different distribution from the other interarrival times. However, because of the memoryless property of the exponential distribution, we know that even if τ units of time has elapsed since the last demand arrival until t_2, the distribution of the remaining time $t^* - t_2$ is still an exponential random variable with the same mean. Thus, services at station 2 accurately model the demand process.

The traffic intensity, ρ, at a station is defined as the ratio of the rate at which work enters the station to the rate at which work can be processed at that station. We are only interested in cases where $\rho < 1$, that is, the facility can cope with the amount of incoming work. Thus at station 1 we require that $\rho_1 = \lambda m_0 + m_1/m_2 < 1$.

We will also only consider cases where $m_1 < m_2$ for the following reason. Consider if $m_1 > m_2$, then station 1 will perpetually be a bottleneck since demand at station 2 for make-to-stock products will be greater than the rate of their production from station 1 to maintain the finished goods inventory. So, if we ensure $m_1 < m_2$, then make-to-stock demands can be filled.

1.3 Related Literature

Queuing models are valuable tools for analyzing manufacturing systems. Open queuing networks modelling make-to-order production systems have been studied extensively, originating from Jackson's paper [8]. Buzacott and Shantikumar [2] present a variety of results stemming from these studies. The original work modelling make-to-stock production systems as closed queuing networks is by Morse [11]. Subsequently much analysis of production-inventory systems have been carried utilizing these models, following the fundamental work of Gordon and Newell [4]. Silver and Peterson [14] give a good overview of such models.

Relatively less work has been carried out on mixed queuing networks modelling both make-to-order and make-to-stock production. Baskett, Chandy, Muntz and Palacios extended the work by Jackson [8] and Gordon and Newell [4] by illustrating

the product form nature of a class of queuing networks, including mixed queuing networks. More recently, Heffes [6] and Akyildiz and Strelen [1] have developed algorithms for computing moments of queue size distributions in mixed queuing networks.

Carr, Gullu, Jackson and Muckstadt [3] study a mixed queuing network where they partition the products into three classes: A items are highest demand, B medium demand and C low demand. Carr *et al.* propose that the optimal production strategy is to hold a base stock of A items, and hold no B/C items in inventory. Additionally B/C items are given strict priority over A items in production. Thus, drawing a parallel with the production facility studied in this thesis, A items would be make-to-stock and B/C items the make-to-order products. Carr *et al.* model their production center as an $M/D/1$ queue. That is, demands for items arrive as a Poisson distribution; service times for any product is deterministically one unit of time; and the production center is one machine or resource which can process items only one at a time. They find that their "No B/C strategy" performs best compared to a first in first out (FIFO) policy when the traffic intensity at the production center is 0.95, in which case its efficiency (100 x cost of "No B/C"/cost of FIFO) is 60

Nguyen [12] presents a heavy traffic analysis of a two station mixed queuing network, with two classes of jobs, open and closed. Station 1 uses a FIFO service discipline. Nguyen finds that as the traffic intensity at station 1 tends to unity, the workload at station 1 is a reflected Brownian motion on a finite interval, and the partial workloads due to each type of job is a fixed proportion of the total workload of the station. This result may seem counterintuitive in that the workload process at station 1 need not be bounded. Refinements to the Brownian approximation are presented and three performance measures are tested against exact solutions for the special case of a mixed network with product form probability distributions. The study finds that the approximation for the throughput rate is within 1% of the exact solution for $N \geq 15$; the estimates for queue length of open customers are within 10% of the exact solutions for $N \geq 5$; and, the estimates for queue length of closed customers are within 10% of the exact solutions for $N \geq 15$.

1.4 Organization of the Thesis

Chapter 1 of this thesis provides a framework for modeling a make-to-order and make-to-stock production system as a mixed queuing network. A review of related current literature on this subject matter is also presented.

In Chapter 2 we determine that this queuing network has product form steady state probability distributions under certain conditions and derive expressions for various performance measures.

Chapter 3 is devoted to setting a target base-stock level for the make-to-stock products with the objective of either minimizing the average cost incurred by the system per unit time, or of achieving a desired fill rate for the items under different priority schemes for production.

Chapter 4 develops a dynamic model of the production facility in order to explore various scheduling policies which would minimize the long-term discounted cost of the system. The optimal policy is described and its performance compared with other easily implementable policies.

Chapter 5 provides concluding remarks and suggestions for future research directions.

Chapter 2

Analysis of the Queuing Network

2.1 Problem Description

We are interested in analyzing the queuing network described in Section 1.2, finding closed form solutions when possible and identifying the conditions under which such solutions hold. For this section, we assume that each station operates under the First-In-First-Out (FIFO) discipline.

2.2 Quasi-Reversibility

A network of quasi-reversible queues has product-form solutions; Kelly [9]. If we can determine conditions under which each station is quasi-reversible, then the network will have product form stationary distributions.

Definition 1 *A queue is quasi-reversible if its state $x(t)$ is a stationary Markov process with the property that the state of the queue at time t_0, $x(t_0)$, is independent of*

1. *the arrival times of type i customers subsequent to time t_0;*
2. *the departure times of type i customers prior to time t_0.*

We prove the following theorem in the next two sub-sections:

Theorem 2.2.1 *Suppose $m_0 = m_1 = m$. The queuing network shown in Figure 1-2 is quasi-reversible, and thus has product form stationary distribution.*

2.2.1 Station 1

Because we assume the same service rates for both types of jobs at station 1, the state of the queue, x, can be completely described by the number of each customer type in the station (in queue or service). Let

n_0 = number of open customers at station 1

n_1 = number of closed customers at station 1, thus

n_2 = N - n_1 = number of closed customers at station 2,

n = total number of customers at station 1 = $n_0 + n_1$ -

$x = (n_0, n_1)$.

For convenience, the system at station 1, that is, the queue and the service station together, will be referred to as a queue. Open jobs are referred to as "type 0" and closed jobs as "type 1". Following the notation from Kelly [9] we can describe the queue in the following way:

1. Each customer requires an amount of service which is a random variable exponentially distributed with mean m.
2. A total service effort is supplied at rate $\phi(n)$ when there are n customers in the queue.
3. A proportion $\gamma(l,n)$ of this effort is directed to the customer in position l (l = 1,2,...,n) when there are n customers in the queue.
4. When a customer arrives at the queue, with n customers already in queue,he moves into position l (l = 1,2,...,$n+1$) with probability $\delta(l, n+1)$.

For both stations:

$$\phi(n) = 1 \qquad n = 1,2,\ldots$$

$$\gamma(l,n) = \begin{cases} 1 & l = 1 \\ 0 & \text{otherwise} \end{cases} \qquad n = 1,2,\ldots$$

$$\delta(l,n+1) = \begin{cases} 1 & l = n+1 \\ 0 & \text{otherwise} \end{cases} \qquad n = 1,2,\ldots$$

The equilibrium distribution for the queue is:

$$\pi(x) = b \prod_{l=1}^{n} \frac{a(i(l))}{\phi(l)},$$

where

$a(i(l))$ = arrival rate of customer type i in position l

b = normalizing constant, chosen such that the probability distribution sums to unity

In the case of station 1,

$$a(0(l)) = \lambda$$

$$a(1(l)) = \begin{cases} 1/m_2 & 0 \leq n_1 < N \\ 0 & n_1 = N \end{cases}$$

Let $\mathcal{S}(i,x)$ be the set of states in which the queue contains one more customer of type i than in state x, with the same number of customers of the other type. Let $x' \in \mathcal{S}(i,x)$ and the transition rates be denoted by $q(x,x')$.

The possible transitions are:

I) Arrivals to Station 1, $x \rightarrow x' \in \mathcal{S}(i,x)$:

(1) An open customer arrives: $x=(n_0,n_1) \rightarrow x'=(n_0+1,n_1)$

$$q(x,x') = \lambda\delta(n+1,n+1) = \lambda \qquad n_0 \geq 0$$

(2) A closed customer departs station 2 and arrives at station 1: $x=(n_0,n_1) \rightarrow x'=(n_0,n_1+1)$

$$q(x,x') = a(1(l))\gamma(1,n+1)\delta(n+1,n+1) = \begin{cases} 1/m_2 & \text{for } 0 \leq n_1 < N \\ 0 & \text{for } n_1 = N \end{cases}$$

II) Departures from Station 1, $x' \in \mathcal{S}(i,x) \rightarrow x$:

(3) An open customer departs: $x'=(n_0+1, n_1) \rightarrow x=(n_0, n_1)$

$$q(x', x) = \phi(n+1)\gamma(1, n+1) = 1/m \qquad n_0 \geq 0$$

(4) A closed customer departs station 1 and arrives at station 2: $x'=(n_0, n_1+1) \rightarrow x=(n_0, n_1)$

$$q(x', x) = \phi(n+1)\gamma(1, n+1)\delta(n+1, n+1) = 1/m \qquad 0 \leq n_1 < N$$

Any non-zero transition rate of the process x is one of the four forms given above.

Since we are assuming Poisson input to station 1, we need only to verify the following partial balance equations to determine whether the queue exhibits the property of quasi-reversibility:

$$\pi(x) \sum_{x' \in \mathcal{S}(i,x)} q(x, x') = \sum_{x' \in \mathcal{S}(i,x)} \pi(x') q(x', x).$$

There are two cases to consider:

Case 1: $0 \leq n_1 < N$

left hand side:

$$b(m\lambda)^{n_0} \left(\frac{m}{m_2}\right)^{n_1} \left(\lambda + \frac{1}{m_2}\right)$$

right hand side:

$$\begin{aligned} & b(m\lambda)^{n_0+1} \left(\frac{m}{m_2}\right)^{n_1} \left(\frac{1}{m}\right) + b(m\lambda)^{n_0} \left(\frac{m}{m_2}\right)^{n_1+1} \left(\frac{1}{m}\right) \\ = \; & b(m\lambda)^{n_0} \left(\frac{m}{m_2}\right)^{n_1} (\lambda + \frac{1}{m_2}) \\ = \; & \text{left hand side.} \end{aligned}$$

Case 2: $n_1 = N$

In this case, we would only have transitions (1) and (3).

left hand side:

$$b(m\lambda)^{n_0} \left(\frac{m}{m_2}\right)^{N} (\lambda)$$

right hand side:

$$b(m\lambda)^{n_0+1}\left(\frac{m}{m_2}\right)^N\left(\frac{1}{m}\right)$$
$$= \text{ left hand side.}$$

Since the partial balance equations are verified, we have shown that station 1 is quasi-reversible.

Consider the situation that $m_0 \neq m_1$, then the transition rate for transition (3) (open customer departing station 1) would be $1/m_0$, and the rate for transition (4) (closed customer departing station 1) would be $1/m_1$. Then in Case 1 above, the partial balance equations would only hold if $m_0 = m_1$. Thus, for station 1 to exhibit the property of quasi-reversibility, we require the condition that $m_0 = m_1 = m$.

2.2.2 Station 2

Since quasi-reversibility is an "input-output" property, we will have Poisson output from station 1 and into station 2.

Station 2 is an M/M/1 FIFO queue with a single class of customer. From Burke's theorem this is a reversible and a quasi-reversible queue.

From Kelly [9] theorems 3.7 and 3.12, a network of quasi-reversible queues has product form solutions, thus we have proved Theorem 2.2.1.

2.3 Steady-state Distributions

We have established in the previous section that if we assume

1. Poisson arrivals
2. exponential service time distributions at each station, with the additional constraint that both open and closed customers are served at the same average rate at station 1,

then the network has product-form steady-state distributions. In addition, we require $\lambda m < 1$ and $N < \infty$ for steady state to be reached. (Note that because the "closed"

portion of the network "self-regulates" the arrival process to station 1, we actually do not require $\rho_1 = \lambda m + m/m_2 < 1$ for stability of the station.)

We now turn to the task of calculating this distribution. Let us first introduce the notion of the throughput rate of closed customers, which we denote by α, the rate at which closed customers arrive at either station 1 or equivalently at station 2. (Because we are in equilibrium, on average closed customers must proceed from station 1 to station 2 at the same rate as from station 2 to station 1.) The maximum rate at which jobs can circulate throughout the network is the minimum of the service rates at the two stations, namely, $1/m_2$. Due to periodic idleness at station 2, however, the actual throughput rate will be strictly less than $1/m_2$.

Theorem 2.3.1 *Let $a = \frac{m/m_2}{1-\lambda m}$, then the steady state probability distribution, π, for the network is*

- *for $a \neq 1$,*

$$\pi(n_0, n_1) = \begin{pmatrix} n_0 + n_1 \\ n_1 \end{pmatrix} \left(\frac{1-a}{1-a^{N+1}}\right)(1-\lambda m)\left(\frac{m}{m_2}\right)^{n_1}(\lambda m)^{n_0}$$

- *for $a = 1$,*

$$\pi(n_0, n_1) = \begin{pmatrix} n_0 + n_1 \\ n_1 \end{pmatrix} \left(\frac{1}{N+1}\right)(1-\lambda m)\left(\frac{m}{m_2}\right)^{n_1}(\lambda m)^{n_0}.$$

We prove this theorem in the following subsection.

2.3.1 Probability Distribution

Following Walrand [15], the steady-state probability $\pi(n)$ of having n customers in an M/M/1 queue is

$$\pi(n) = \rho^n(1-\rho)$$

where ρ is the traffic intensity at the station, and a customer is of type i with probability

$$p_i = \frac{\lambda_i m_i}{\rho}$$

At station 1, the probability of having n customers, of which n_0 are open customers and n_1 are closed customers is

$$\rho_1^{n_0+n_1}(1-\rho_1)\begin{pmatrix} n_0+n_1 \\ n_1 \end{pmatrix}\left(\frac{\lambda m}{\rho_1}\right)^{n_0}\left(\frac{\alpha m}{\rho_1}\right)^{n_1}$$

where

ρ_1 = traffic intensity at station 1

α = throughput rate of the closed customers.

At station 2, the traffic intensity is $\rho_2 = \alpha m_2$. So, the probability of having $n_2 = N - n_1$ customers at station 2 is $(\alpha m_2)^{N-n_1}(1-\rho_2)$. Hence, putting the two expressions together, the steady-state probability distribution of having n_0 open customers and n_1 closed customers at station 1 is

$$\pi(n_0, n_1) = G\begin{pmatrix} n_0+n_1 \\ n_1 \end{pmatrix}(\lambda m)^{n_0}(\alpha m)^{n_1}(\alpha m_2)^{N-n_1}$$

where G is the normalizing constant to ensure that

$$\sum_{n_0=0}^{\infty}\sum_{n_1=0}^{N}\pi(n_0, n_1) = 1.$$

So,

$$\begin{aligned} 1 &= G\sum_{n_0=0}^{\infty}[(\alpha m_2)^N + (n_0+1)(\alpha m)(\alpha m_2)^{N-1} \\ &\quad + \frac{(n_0+2)(n_0+1)}{2!}(\alpha m)^2(\alpha m_2)^{N-2} \\ &\quad \ldots + \frac{(n_0+N)(n_0+N-1)(n_0+N-2)\ldots(n_0+1)}{N!}(\alpha m)^N](\lambda m_0)^{n_0} \\ &= G(\alpha m_2)^N\left[\frac{1}{1-\lambda m} + \frac{m/m_2}{(1-\lambda m)^2} + \ldots + \frac{(m/m_2)^N}{(1-\lambda m)^{N+1}}\right]. \end{aligned}$$

For ease of notation, let

$$a = \frac{m/m_2}{1-\lambda m},$$

with which we have

$$1 = G\frac{(\alpha m_2)^N}{1-\lambda m}\left[1 + a + a^2 + \ldots + a^N\right]$$

When $a \neq 1$,

$$G = \frac{1-\lambda m}{(\alpha m_2)^N}\left(\frac{1-a}{1-a^{N+1}}\right),$$

and the steady state distribution is

$$\begin{aligned}\pi(n_0, n_1) &= G\begin{pmatrix} n_0+n_1 \\ n_1 \end{pmatrix}(\lambda m)^{n_0}(\alpha)^N(m)^{n_1}(m_2)^{N-n_1} \\ &= \begin{pmatrix} n_0+n_1 \\ n_1 \end{pmatrix}\left(\frac{1-a}{1-a^{N+1}}\right)(1-\lambda m)\left(\frac{m}{m_2}\right)^{n_1}(\lambda m)^{n_0}.\end{aligned}$$

From this we can see that, as expected, the steady-state distribution is not a function of α, the throughput rate of closed jobs.

When $a = 1$, the traffic intensities at both stations are equal to unity and the queuing network is considered "balanced". Using the summation result that for $a = 1$, $\sum_{i=0}^{N} a^i = N + 1$,

$$G = \frac{1-\lambda m}{(\alpha m_2)^N}\left(\frac{1}{N+1}\right),$$

and the steady state distribution reduces to

$$\pi(n_0, n_1) = \begin{pmatrix} n_0+n_1 \\ n_1 \end{pmatrix}\left(\frac{1}{N+1}\right)(1-\lambda m)\left(\frac{m}{m_2}\right)^{n_1}(\lambda m)^{n_0}.$$

One interpretation for a is *the rate of incoming make-to-stock work per unit of time available to the production facility for processing make-to-stock products.* In Section 2.3 we mentioned the condition that $\lambda m < 1$. This allows for all values of $a > m/m_2$. However, if $a > 1$, then the traffic intensity at station 1, $\rho_1 = \lambda m + m/m_2 > 1$. This would not be an interesting situation to analyze for an inventory model, and hence we will restrict our study to values of a such that $m/m_2 < a < 1$.

2.3.2 Average Queue Length

We can use the steady-state distribution to determine the average queue length at each station. At station 1, let $E[N_0]$ denote the expected number of open customers, and $E[N_1]$ the expected number of closed customers.

Corollary 2.3.2 *From Theorem 2.3.1 the expected queue length at station 1 is given by:*

- *for* $a < 1$,

$$\begin{aligned} E[N_0] &= \left(\frac{am_2}{m} - 1\right)\left(\frac{1-(N+2)a^{N+1}\left(1-a\frac{N+1}{N+2}\right)}{(1-a)(1-a^{N+1})}\right) \\ E[N_1] &= a\left(\frac{1-(N+1)a^{N}\left(1-a\frac{N}{N+1}\right)}{(1-a)(1-a^{N+1})}\right) \end{aligned}$$

- *for* $a = 1$,

$$\begin{aligned} E[N_0] &= \left(\frac{\lambda m}{1-\lambda m}\right)\left(\frac{N+2}{2}\right) \\ E[N_1] &= \left(\frac{\lambda m}{1-\lambda m}\right)\left(\frac{N}{2}\right) \end{aligned}$$

Proof. Following similar steps as in the previous section,

$$\begin{aligned} E[N_0] &= \sum_{n_0=0}^{\infty}\sum_{n_1=0}^{N} n_0\pi(n_0,n_1) \\ &= G\sum_{n_0=0}^{\infty} n_0[(\alpha m_2)^N + (n_0+1)(\alpha m)(\alpha m_2)^{N-1} \\ &\quad + \frac{(n_0+2)(n_0+1)}{2!}(\alpha m)^2(\alpha m_2)^{N-2} + \ldots \\ &\quad \ldots + \frac{(n_0+N)(n_0+N-1)(n_0+N-2)\ldots(n_0+1)}{N!}(\alpha m)^N](\lambda m_0)^{n_0} \\ &= G(\alpha m_2)^N\left[\frac{\lambda m}{(1-\lambda m)^2} + \frac{2(\lambda m)(m/m_2)}{(1-\lambda m)^3} + \ldots + \frac{(N+1)(\lambda m)(m/m_2)^N}{(1-\lambda m)^{N+2}}\right] \\ &= G\frac{(\lambda m)(\alpha m_2)^N}{(1-\lambda m)^2}[1+2a+\ldots+(N+1)a^N] \end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{\lambda m}{1-\lambda m}\right)\left(\frac{1-a}{1-a^{N+1}}\right)[1+2a+\ldots+(N+1)a^N] \\
&= \left(\frac{\lambda m}{1-\lambda m}\right)\left[\frac{1-(N+2)a^{N+1}+(N+1)a^{N+2}}{(1-a)(1-a^{N+1})}\right] \\
&= \left(\frac{am_2}{m}-1\right)\left(\frac{1-(N+2)a^{N+1}\left(1-a\frac{N+1}{N+2}\right)}{(1-a)(1-a^{N+1})}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
E[N_1] &= \sum_{n_0=0}^{\infty}\sum_{n_1=0}^{N} n_1\pi(n_0,n_1) \\
&= G\sum_{n_0=0}^{\infty}(\lambda m_0)^{n_0}[0+(n_0+1)(\alpha m)(\alpha m_2)^{N-1} \\
&\quad +\frac{2(n_0+2)(n_0+1)}{2!}(\alpha m)^2(\alpha m_2)^{N-2}+\ldots \\
&\quad \ldots+\frac{N(n_0+N)(n_0+N-1)(n_0+N-2)\ldots(n_0+1)}{N!}(\alpha m)^N] \\
&= G(\alpha m)(\alpha m_2)^{N-1}\sum_{n_0=0}^{\infty}(\lambda m_0)^{n_0}[(n_0+1)+\frac{(n_0+2)(n_0+1)}{1!}(m/m_2) \\
&\quad \ldots+\frac{(n_0+N)(n_0+N-1)\ldots(n_0+1)}{(N-1)!}(m/m_2)^{N-1}] \\
&= G(m/m_2)(\alpha m_2)^N\left[\frac{1}{(1-\lambda m)^2}+\frac{2(m/m_2)}{(1-\lambda m)^3}+\ldots+\frac{N(m/m_2)^{N-1}}{(1-\lambda m)^{N+1}}\right] \\
&= G\frac{(m/m_2)(\alpha m_2)^N}{(1-\lambda m)^2}[1+2a+\ldots+Na^{N-1}] \\
&= \left(\frac{m/m_2}{1-\lambda m}\right)\left(\frac{1-a}{1-a^{N+1}}\right)[1+2a+\ldots+Na^{N-1}] \\
&= \left(\frac{m/m_2}{1-\lambda m}\right)\left[\frac{1-(N+1)a^N+Na^{N+1}}{(1-a)(1-a^{N+1})}\right] \\
&= a\left(\frac{1-(N+1)a^N\left(1-a\frac{N}{N+1}\right)}{(1-a)(1-a^{N+1})}\right).
\end{aligned}
$$

When $a=1$, we have the following:

$$
\begin{aligned}
E[N_0] &= G\frac{(\lambda m)(\alpha m_2)^N}{(1-\lambda m)^2}\left(\frac{(N+1)(N+2)}{2}\right) \\
&= \left(\frac{\lambda m}{1-\lambda m}\right)\left(\frac{N+2}{2}\right)
\end{aligned}
$$

$$E[N_1] = G\frac{(m/m_2)(\alpha m_2)^N}{(1-\lambda m)^2}\left(\frac{N(N+1)}{2}\right)$$
$$= \left(\frac{\lambda m}{1-\lambda m}\right)\left(\frac{N}{2}\right)$$

□

Let us look at the form of these functions. Consider the case where $a << 1$, where the rate of incoming make-to-stock work is much smaller than the fraction of time available to the production facility to process make-to-stock products. Then we can take the terms in a^N, for large closed job populations N, to be approximately equal to zero:

$$\begin{aligned} E[N_0] &\approx \left(\frac{m_2 a}{m} - 1\right)(1-a)^{-1} \\ &\approx \left(\frac{m_2 a}{m} - 1\right)(1 + a + \ldots) \\ &\approx \frac{m_2 a}{m} - a - 1 \\ &\approx a\left(\frac{m_2}{m} - 1\right) - 1, \end{aligned}$$

and

$$E[N_1] \approx a(1 + a + \ldots) \approx a.$$

Taking $m_2 = 1$, $m = 0.01$ and varying values of λ, Figures 2-1 and 2-2 show the behavior of $E[N_0]$ and $E[N_1]$ versus a for values of N=5, 30, and 100. For $a << 1$, $E[N_0]$ and $E[N_1]$ are quite insensitive to N, and seem to increase linearly with a.

2.3.3 Throughput Rate

Corollary 2.3.3 *The throughput rate, α, of closed jobs is given by:*

- *for $a < 1$,*
$$\alpha = \left(\frac{1}{m_2}\right)\left(\frac{1-a^N}{1-a^{N+1}}\right)$$
- *for $a = 1$,*
$$\alpha = \left(\frac{1}{m_2}\right)\left(\frac{N}{N+1}\right).$$

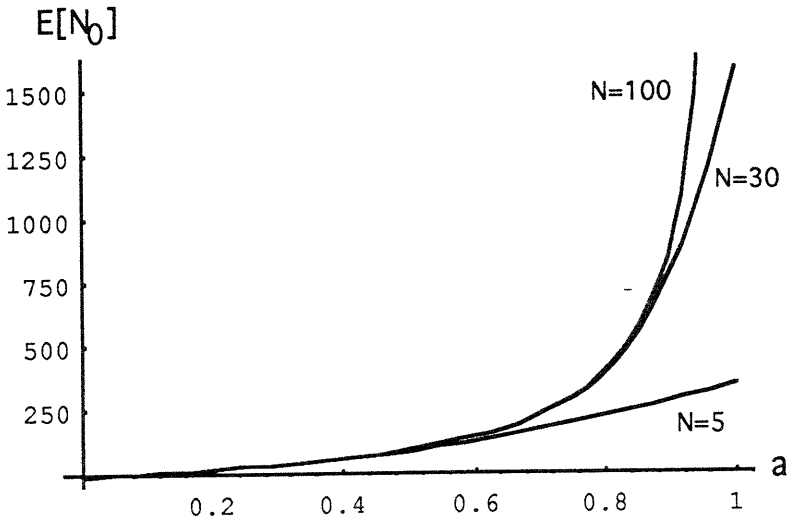

Figure 2-1: Behavior of $E[N_0]$ with a

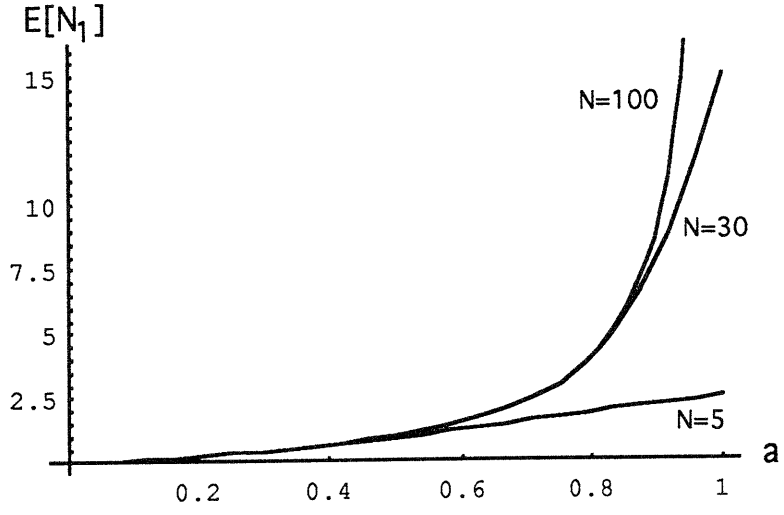

Figure 2-2: Behavior of $E[N_1]$ with a

Proof. In order to calculate α, we examine the fraction of time that station 2 is busy. Denoting this quantity by β_2 and noting that $\beta_2 = \alpha m_2$, we have

$$\begin{aligned}
\beta_2 &= 1 - \text{probability of being idle} \\
&= 1 - \text{probability of all } N \text{ closed customers being at station 1} \\
&= 1 - \sum_{n_0=0}^{\infty} \pi(n_0, N) \\
&= 1 - \sum_{n_0=0}^{\infty} G \begin{pmatrix} n_0 + N \\ N \end{pmatrix} (\lambda m)^{n_0} (\alpha m)^N \\
&= 1 - G(\alpha m)^N [1 + (N+1)\lambda m + (N+2)(N+1)\frac{(\lambda m)^2}{2!} + \ldots] \\
&= 1 - G(\alpha m)^N [(1 - \lambda m)^{-(N+1)}] \\
&= 1 - a^N [1 + a + a^2 + \ldots + a^N]^{-1}.
\end{aligned}$$

When $a < 1$,

$$\begin{aligned}
\beta_2 &= 1 - \frac{a^N(1-a)}{1-a^{N+1}} \\
&= \frac{1-a^N}{1-a^{N+1}},
\end{aligned} \tag{2.1}$$

and the throughput

$$\alpha = \left(\frac{1}{m_2}\right)\left(\frac{1-a^N}{1-a^{N+1}}\right).$$

When $a = 1$,

$$\begin{aligned}
\beta_2 &= 1 - \frac{1}{N+1} \\
&= \frac{N}{N+1},
\end{aligned} \tag{2.2}$$

and the throughput

$$\alpha = \left(\frac{1}{m_2}\right)\left(\frac{N}{N+1}\right).$$

□

2.3.4 Sojourn Time of a Typical Job

We define the sojourn time of a job as the sum of its waiting time at station 1 (time spent in queue) and its service time at station 1 (time to process the job). The sojourn times for open and closed jobs are denoted by W_0 and W_1 respectively.

Corollary 2.3.4 *The expected sojourn time for a typical job is:*

- *for* $a < 1$,

$$\begin{aligned} W_0 &= \left(\frac{m}{1-\lambda m}\right)\left[\frac{1-(N+2)a^{N+1}+(N+1)a^{N+2}}{(1-a)(1-a^{N+1})}\right] \\ W_1 &= \left(\frac{m}{1-\lambda m}\right)\left[\frac{1-(N+1)a^{N}+(N)a^{N+1}}{(1-a)(1-a^{N})}\right] \end{aligned}$$

- *for* $a = 1$,

$$\begin{aligned} W_0 &= \left(\frac{m}{1-\lambda m}\right)\left(\frac{N+2}{2}\right) \\ W_1 &= \left(\frac{m}{1-\lambda m}\right)\left(\frac{N+1}{2}\right). \end{aligned}$$

Proof. Because we have independent Poisson arrivals, we can use PASTA (Poisson Arrivals See Time Averages) and Little's Law [10] to show that the expected sojourn time, that is, the time for queuing and service completion, is $W_0 = \frac{E[N_0]}{\lambda}$ and $W_1 = \frac{E[N_1]}{\alpha}$ for open and closed jobs respectively. From these it is simple to derive the desired results. □

2.3.5 Asymptotic Behavior

What can be said about the expected sojourn times of jobs as N becomes large? In the limit as $N \to \infty$ the terms in a^N and higher powers will be approximately equal to zero, since $a < 1$. Thus we have the following limits

$$\lim_{N\to\infty} E[N_0] = \left(\frac{\lambda m}{1-\lambda m}\right)\left(\frac{1}{1-a}\right)$$

$$
\begin{aligned}
&= \frac{\lambda m}{1-\lambda m - m/m_2}, \\
\lim_{N\to\infty} E[N_1] &= \frac{a}{1-a} \\
&= \frac{m/m_2}{1-\lambda m - m/m_2}, \\
\lim_{N\to\infty} \alpha &= 1/m_2,
\end{aligned}
$$

and so,

$$
\lim_{N\to\infty} E[W_0] = \frac{m}{1-\lambda m - m/m_2},
$$

$$
\lim_{N\to\infty} E[W_1] = \frac{m}{1-\lambda m - m/m_2}.
$$

The result for the asymptotic values for the average sojourn times is as expected, since $N \to \infty$ would mean essentially Poisson arrivals of closed customers from an infinite pool. Therefore, at station 1, we would have the equivalent of two independent Poisson arrival processes of 'open' jobs, and we would expect that the system time for one type of job to be the same as for the other.

Chapter 3

Static Control of the Production System

3.1 Problem Description

This section is devoted to the analysis of static control policies of the make-to-order make-to-stock production system. We assume that a priori, a service discipline is specified for the production facility (e.g. FIFO) and that the facility follows a base-stock policy (or equivalently, a one-for-one replenishment policy). The problem that we study is how to set the base-stock level so as to achieve one of two objectives:

- satisfy a service level constraint for the make-to-stock demands filled from inventory, or
- minimize the average cost incurred per unit time.

In analyzing the latter problem, we assume a linear cost structure with the following costs:

1. c = cost of holding one unit of WIP of open jobs at the production center per unit time, measured in \$/unit/time
2. l = cost of lost sale of one unit of make-to-stock product, measured in \$/unit, and
3. h = cost of holding one unit of finished goods inventory for make-to-stock products per unit time, measured in \$/unit/time.

The WIP in the first cost item c refers to Work-In-Process of make-to-order products at the production facility, in other words, demand waiting to be processed, or being processed. At station 1 this is represented by the queue length of open jobs. The second cost, l, is incurred when there is a lost sale of make-to-stock products due to insufficient stock in finished goods inventory.

It may be difficult to quantify accurately the costs involved, particularly in the loss of sales, where one has the cumulative effect of immediate loss of profits and the long term effect of loss of good will. Hence we also consider an alternate approach which is to set a service level constraint for meeting demand for the make-to-stock products. In other words, we decide what service level we would like to achieve in fulfilling orders for the make-to-stock products, and use that to determine the inventory level that would be sufficient to meet the demand. In trying to achieve this service level, we also consider the effect of employing different policies for production at station 1, such as, giving priority to one or other type of product to see what effect that has on the optimal inventory level.

3.2 Service Level Constraints

Suppose we want to achieve a certain service level for the make-to-stock products. In order to meet the target, we need to hold stock in the finished goods inventory. From Section 2.3.3 we have β_2, the fraction of time that station 2 is busy. This is also the service level, or "fill rate", that is the fraction of demands for make-to-stock products filled from inventory. Denoting the desired fill rate by β_2^* we can determine the critical value of N required to achieve this level of service.

3.2.1 First In First Out

We first consider the system where station 1 uses a first in first out (FIFO) discipline to process the two types of products.

Theorem 3.2.1 *In order to achieve a fill rate of β_2^* of make-to-stock products, when jobs are processed at station 1 using a FIFO policy, the critical value of N is given by:*

- *for* $a < 1$,

$$N = \frac{ln\left(\frac{1-\beta_2^*}{1-\beta_2^* a}\right)}{ln\ a}$$

- *for* $a = 1$,

$$N = \frac{\beta_2^*}{1-\beta_2^*}.$$

Proof. When $a < 1$, from equation (2.1) we have

$$\beta_2^* = \frac{1-a^N}{1-a^{N+1}}.$$

Rearranging the terms,

$$\beta_2^*(1-a^{N+1}) = 1-a^N$$

$$a^N(1-\beta_2^* a) = 1-\beta_2^*$$

$$N = \frac{ln\left(\frac{1-\beta_2^*}{1-\beta_2^* a}\right)}{ln\ a}.$$

When $a = 1$, from equation (2.2) we have

$$\beta_2^* = \frac{N}{N+1}.$$

Again, rearranging the terms gives

$$N = \frac{\beta_2^*}{1-\beta_2^*}.$$

□

Consider, for example, that a 95% service level is desired. That is, we require $\beta_2^* = 0.95$

$$\frac{N}{N+1} = 0.95$$

$$\Rightarrow N = 19$$

Thus, in order to meet 95% of the demand for make-to-stock products, in a balanced network, we would need to keep a finished goods inventory of 19.

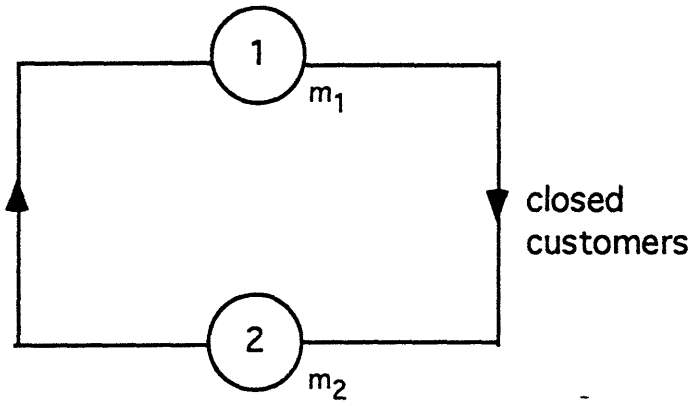

Figure 3-1: Network as Experienced by Closed Customers

3.2.2 Priority for Make-to-Stock Products

Consider the situation if we were to give priority to one or the other type of customer at station 1. Then we would no longer have product-form solutions. Without product-form solutions what can we say about the system?

Consider a priority discipline at station 1 such that if there are any orders waiting in queue then make-to-stock products are always processed before make-to-order products. If it happens that a make-to-order product is being processed when a replenishment order comes in for a make-to-stock product then the make-to-order product is ejected and sent to the head of the queue, and the make-to-stock product is processed, after which, service for the make-to-order product restarts where it left off. Such a priority is known as *FIFO preemptive resume* priority for make-to-stock products.

Let us give such priority for make-to-stock products. Then, although there are no product form solutions, we can still say something about the system and possibly calculate a threshold value for N in order to achieve a certain service level. With such a priority the closed customers experience an M/M/1 queue at both station 1 and station 2, as if no open customers were present, as in Figure 3-1. The state of the system can be represented by a birth and death Markov chain, as in Figure 3-2, where the numbers within the circles represent n_1, or the number of closed customers at station 1.

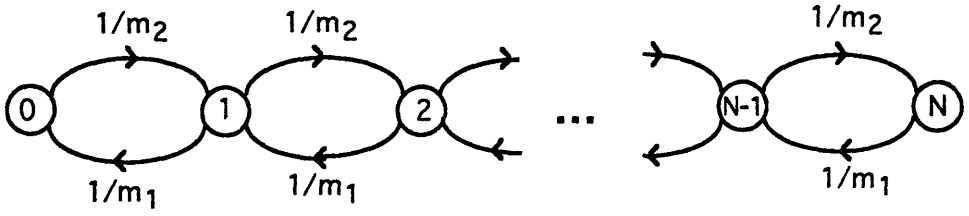

Figure 3-2: Birth and Death Markov Chain

Let p_{n_1} = probability$\{n_1$ closed customers at station 1 and $N - n_1$ at station 2$\}$ then,

$$p_{n_1} = \left(\frac{m}{m_2}\right)^{n_1} p_0; \qquad \text{for } 0 \leq n_1 \leq N.$$

Using the normalization $\sum_{n_1=0}^{N} p_{n_1} = 1$, we get

$$p_{n_1} = \left(\frac{m}{m_2}\right)\left(1 - \frac{m}{m_2}\right)\left[1 - \left(\frac{m}{m_2}\right)^{N+1}\right]^{-1}.$$

Following the earlier procedure for calculating the expected number of closed customers at station 1, we get,

$$E[N_1] = \left(\frac{m/m_2}{1 - m/m_2}\right)\left(\frac{1 + N(m/m_2)^{N+1} - (N+1)(m/m_2)^N}{1 - (m/m_2)^{N+1}}\right).$$

As we would expect with such a priority discipline, the probability distribution and expected queue length for closed jobs is not a function of λ, and therefore a, because closed jobs are not affected by open jobs in the system.

Theorem 3.2.2 *In order to achieve a fill rate of β_2^* of make-to-stock products, when closed jobs are given FIFO preemptive resume priority over open jobs at station 1, the critical value of N is given by:*

$$N = \frac{ln[\frac{m_2/m - \beta_2^*}{1 - \beta_2^*}]}{ln(m_2/m)} - 1.$$

Proof. Consider the fraction of time station 2 is busy:

$$\begin{aligned}
\beta_2 &= 1 - \text{probability}\{\text{all } N \text{ closed customers at station 1}\} \\
&= 1 - \left(1 - \frac{m_2}{m}\right)\left[1 - \left(\frac{m_2}{m}\right)^{N+1}\right]^{-1} \\
&= \left(\frac{m_2}{m}\right)\frac{[1-(m_2/m)^N]}{[1-(m_2/m)^{N+1}]}.
\end{aligned}$$

For a particular service level, β_2^*, we can solve for N:

$$\begin{aligned}
\beta_2^*\left[1 - \frac{m_2}{m}^{N+1}\right] &= \frac{m_2}{m} - \left(\frac{m_2}{m}\right)^{N+1} \\
\left(\frac{m_2}{m}\right)^{N+1}[1-\beta_2^*] &= \frac{m_2}{m} - \beta_2^* \\
(N+1)ln\left(\frac{m_2}{m}\right) &= ln\left(\frac{\frac{m_2}{m} - \beta_2^*}{1-\beta_2^*}\right) \\
N &= \frac{ln[\frac{m_2/m-\beta_2^*}{1-\beta_2^*}]}{ln(m_2/m)} - 1.
\end{aligned}$$

□

Behavior of the critical stock level

Consider the case where $m << m_2$, then N seems to be well approximated by $N \approx A[ln(1/1-\beta_2^*)]$, where A is some constant. Figure 3-3 shows a plot of N versus β_2^* for $m/m_2 = 0.01$ and also of $A[ln(1/1-\beta_2^*)]$ with $A = 0.2$. Figure 3-4 shows a similar plot but with $A = 0.2152$ illustrating that the approximation is very good. The plots show that even a large change in the level of service required would only mean a small change in N. This is as expected since $m << m_2$ means that the rate of production from station 1 is much greater than the demand rate for make-to-stock products. Hence in the case shown, for example, in order to meet a service level requirement of 95%, one would need a finished goods inventory of just one.

As $m/m_2 \rightarrow 1$, the curve resembles $N \approx (1/1-\beta_2^*)$ more closely. However, approximating the function with $N \approx (1/1-\beta_2^*)$ leads to an overestimate for N, as in Figure 3-5. A plot of $N = 0.8/(1-\beta_2^*)$ seems to give a better approximation, as in Figure 3-6.

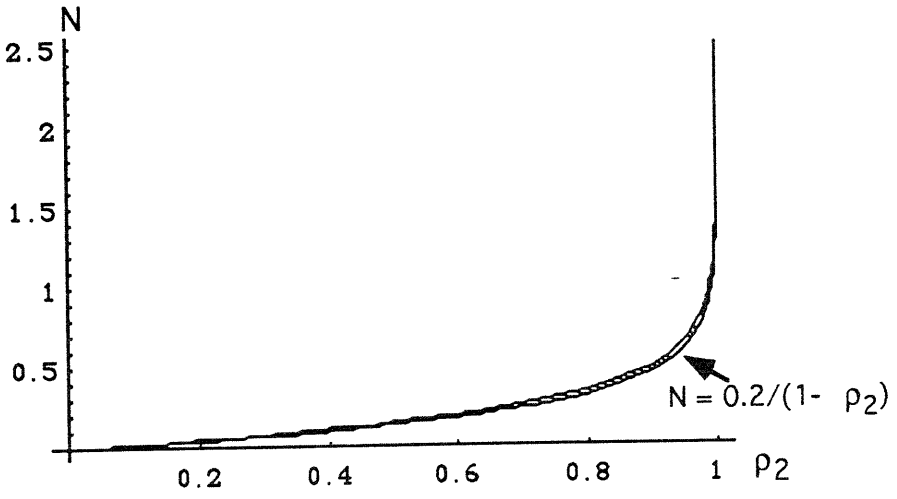

Figure 3-3: N versus β_2^* for $m/m_2 = 0.01$

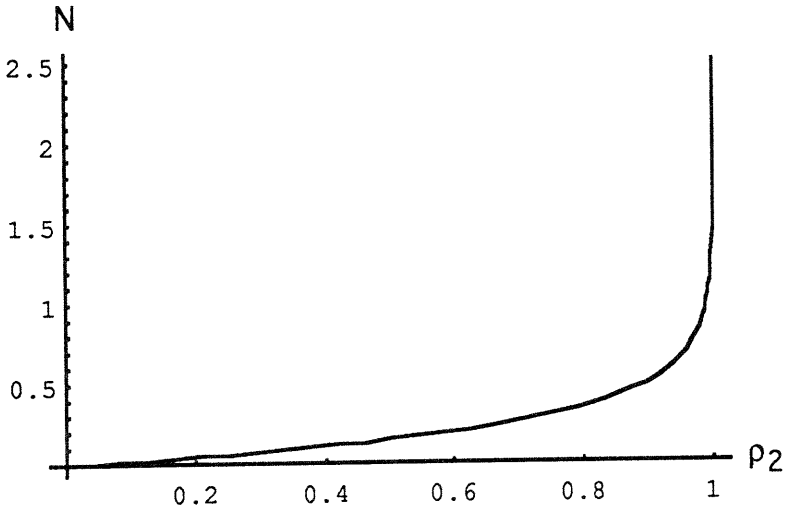

Figure 3-4: N versus β_2^* for $m/m_2 = 0.01$

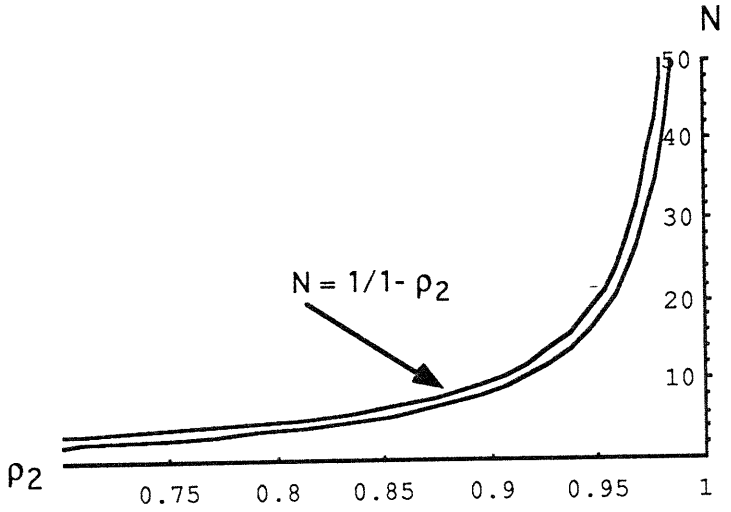

Figure 3-5: N versus β_2^* for $m/m_2 = 0.99$

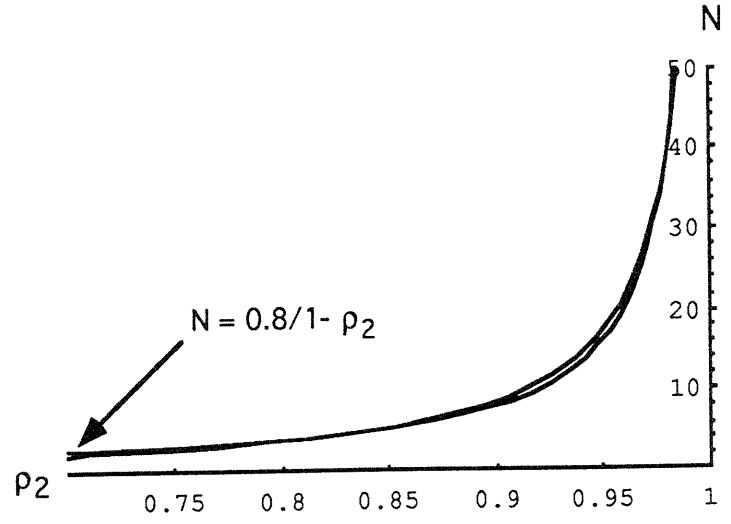

Figure 3-6: N versus β_2^* for $m/m_2 = 0.99$

For β_2^* very small ($\beta_2^* << 1$), N grows linearly with β_2^*:

$$N \approx \frac{(1 - m/m_2)}{ln(m_2/m)} \beta_2^*$$

3.2.3 Priority for Make-to-Order Products

We now consider giving FIFO preemptive resume priority to make-to-order products at station 1. As in the previous section, we no longer have product-form solutions. This time the open customers experience an M/M/1 queue at station 1, as if no closed customers were present. Thus we can calculate the average Work-in-Process for make-to-order products and the average number of replenishment orders for make-to-stock products at station 1. However since the departure process from station 1 is not a Poisson process, we cannot determine a critical stock level to maintain in order to achieve a certain service level at station 2.

3.3 Minimum Cost Objective

3.3.1 Problem Description

We now turn to the task of setting an optimal base-stock level in order to minimize the average cost incurred by the production system per unit time. As described in Section 3.1, we assume that the costs are linear. We would expect that $c < h$, since c represents just the cost of an order waiting to be processed. In other words, for c, we need only to consider the cost of holding raw material, whereas for h, we also need to consider labor costs and capital tied up in holding the products as finished goods. We would also expect that $h < l$, that the cost of losing a sale because demand cannot be satisfied from the finished goods inventory is greater than holding a product in the finished goods inventory. Thus l would represent the selling price of a finished product, plus possibly some penalty for loss of good will for not meeting demand.

We chose three different combinations of $[c, l, h]$, which might realistically model the costs incurred in the production facility. They are:

1. $c = 1, \qquad l = 10, \qquad h = 2$
2. $c = 1, \qquad l = 100, \qquad h = 2$

3. $c = 1, \qquad l = 250, \qquad h = 4$

3.3.2 General Solution for the Production System

Theorem 3.3.1 *The average cost for the production system per unit time is given by:*

- *for* $a < 1$,

$$\begin{aligned} E[S] &= \left(\frac{m_2 a}{m} - 1\right) c \left[\frac{1 - (N+2)a^{N+1} + (N+1)a^{N+2}}{(1-a)(1-a^{N+1})}\right] \\ &+ \frac{l}{m_2}\left[\frac{a^N(1-a)}{(1-a^{N+1})}\right] \\ &+ h\left[\frac{N(1-a) - a + a^{N+1}}{(1-a)(1-a^{N+1})}\right] \end{aligned}$$

- *for* $a = 1$,

$$E[S] = c\left(\frac{m_2}{m} - 1\right)\left(\frac{N+2}{2}\right) + \frac{l}{m_2}\left(\frac{1}{N+1}\right) + h\left(3 - \frac{m_2}{m}\right)\left(\frac{N}{2}\right)$$

Proof. The cost due to WIP at station 1 is c multiplied by the expected queue length of open customers. The cost due to holding finished goods inventory is h multiplied by the expected queue length of closed customers at station 2. Consider the cost due to lost orders at station 2. The fraction of time that inventory is zero is the fraction of time that the station is idle, which is, $1 - \beta_2$. At this time, orders still come in at a rate of $1/m_2$. Therefore, the number of lost orders per unit time is $(1 - \beta_2)/m_2$. Hence the expected cost per unit time for the network is,

$$E[S] = cE[N_0] + l(1 - \beta_2)/m_2 + h(N - E[N_1])$$

For $a < 1$, substituting the expressions for $E[N_0]$, $E[N_1]$ and β_2, from Sections 2.3.2 and 2.3.3 and simplifying, gives

$$1 - \beta_2^* = 1 - \frac{1 - a^N}{1 - a^{N+1}}$$

$$= \frac{a^N(1-a)}{(1-a^{N+1})}$$

and,

$$\begin{aligned} N - E[N_1] &= N - \left[\frac{a-(N+1)a^{N+1}+Na^{N+2}}{(1-a)(1-a^{N+1})}\right] \\ &= \frac{N[1-a^{N+1}-a+aN+2]-[a-(N+1)a^{N+1}+Na^{N+2}]}{(1-a)(1-a^{N+1})} \\ &= \frac{N(1-a)-a+a^{N+1}}{(1-a)(1-a^{N+1})} \end{aligned}$$

therefore,

$$\begin{aligned} E[S] &= \left(\frac{m_2 a}{m}-1\right) c\left[\frac{1-(N+2)a^{N+1}+(N+1)a^{N+2}}{(1-a)(1-a^{N+1})}\right] \\ &\quad + \frac{l}{m_2}\left[\frac{a^N(1-a)}{(1-a^{N+1})}\right] \\ &\quad + h\left[\frac{N(1-a)-a+a^{N+1}}{(1-a)(1-a^{N+1})}\right] \end{aligned} \tag{3.1}$$

This expression, as it stands, does not afford a clear intuitive understanding of the cost function. So, we must look for simplifications or approximations which will clarify the form of the function. We pursue this strategy in the next section in order to get a good approximation for the optimal N^* which would minimize $E[S]$, when $a < 1$.

For the case of a balanced network, $a = 1$, again using the expressions for $E[N_0]$, $E[N_1]$ and β_2 from Sections 2.3.2 and 2.3.3, we have

$$\begin{aligned} N - E[N_1] &= N - \left(\frac{m_2}{m}-1\right)\frac{N}{2} \\ &= \left(3-\frac{m_2}{m}\right)\frac{N}{2} \end{aligned}$$

Thus,

$$E[S] = c\left(\frac{m_2}{m}-1\right)\left(\frac{N+2}{2}\right) + \frac{l}{m_2}\left(\frac{1}{N+1}\right) + h\left(3-\frac{m_2}{m}\right)\left(\frac{N}{2}\right)$$

□

Corollary 3.3.2 *For $a = 1$ the optimal N^* which minimizes the average cost is given by:*

$$N^* = \sqrt{\frac{2l/m_2}{c\left(\frac{m_2}{m} - 1\right) + h\left(3 - \frac{m_2}{m}\right)}} - 1$$

Proof. In order to determine the critical value of N which minimizes the cost function, we take the derivative with respect to N and set $E'[S]$ equal to zero.

$$E'[S] = \frac{c}{2}\left(\frac{m_2}{m} - 1\right) + \frac{h}{2}\left(3 - \frac{m_2}{m}\right) - \frac{l/m_2}{(N+1)^2} = 0$$

which yields

$$N^* = \sqrt{\frac{2l/m_2}{c\left(\frac{m_2}{m} - 1\right) + h\left(3 - \frac{m_2}{m}\right)}} - 1.$$

□

We also have that

$$E''[S] = \frac{2l/m_2}{(N+1)^3} > 0.$$

Thus we have determined the N^* which minimizes the cost function when $a = 1$. Note that N^* must be greater than or equal to one. Substituting for N^* we have

$$\begin{aligned}
E[S]_{min} &= \frac{1}{2}\left[c\left(\frac{m_2}{m} - 1\right) + h\left(3 - \frac{m_2}{m}\right)\right]\left[\sqrt{\frac{2l/m_2}{c\left(\frac{m_2}{m} - 1\right) + h\left(3 - \frac{m_2}{m}\right)}} - 1\right] \\
&\quad + \frac{l}{m_2}\sqrt{\frac{c\left(\frac{m_2}{m} - 1\right) + h\left(3 - \frac{m_2}{m}\right)}{2l/m_2}} + c\left(\frac{m_2}{m} - 1\right) \\
&= \sqrt{\left(\frac{l}{2m_2}\right)\left[c\left(\frac{m_2}{m} - 1\right) + h\left(3 - \frac{m_2}{m}\right)\right]} \\
&\quad - \frac{1}{2}\left[c\left(\frac{m_2}{m} - 1\right) + h\left(3 - \frac{m_2}{m}\right)\right] \\
&\quad + \sqrt{\left(\frac{l}{2m_2}\right)\left[c\left(\frac{m_2}{m} - 1\right) + h\left(3 - \frac{m_2}{m}\right)\right]} + c\left(\frac{m_2}{m} - 1\right) \\
&= \sqrt{\left(\frac{2l}{m_2}\right)\left[c\left(\frac{m_2}{m} - 1\right) + h\left(3 - \frac{m_2}{m}\right)\right]} + \frac{1}{2}\left[c\left(\frac{m_2}{m} - 1\right) + h\left(3 - \frac{m_2}{m}\right)\right].
\end{aligned}$$

The above expression for N^* makes intuitive sense in that we would expect to hold less finished goods inventory if either the WIP holding costs or finished goods

holding costs were to increase, and, if just the cost due to lost sales was to increase then we would expect to hold more in the finished goods inventory.

3.3.3 Approximations: Low Demand for Make-to-Stock Products

Approximation 3.3.3 *For $a < 1$, given a low demand for make-to-stock products, the optimal N^* which minimizes the average cost is given by:*

$$N^* \approx \sqrt{\frac{2l/m_2}{c\left(\frac{m_2}{m}-1\right)+h\left(3-\frac{m_2}{m}\right)}} - 1.$$

Turning to the case where $a < 1$, let us consider that there is low demand for make-to-stock products, say, one unit per week, which would mean $m_2 = 1$. In this case we find that the critical value of the inventory level determined for the balanced network case, N^*_{bal}, provides a good approximation to both the N^*_{actual} and $E[S]_{min}$. That is,

$$N^* \approx \sqrt{\frac{2l/m_2}{c\left(\frac{m_2}{m}-1\right)+h\left(3-\frac{m_2}{m}\right)}} - 1. \tag{3.2}$$

Tables 3.1, 3.2 and 3.3 compare the actual values of N^* and $E[S]_{min}$ with the values for a balanced network.

3.3.4 Approximations: High Demand for Make-to-Stock Products

It may be more realistic to look at a demand rate of, say, a thousand units of make-to-stock products per week. Taking $m_2 = 10^{-3}$ and $m = 0.8 \text{ x } 10^{-3}$ we find the balanced network simplification a poor model for the cost function and so we must find a better approximation.

E[S]					
m2 = 1					
m = 0.8					
[c, l , h]	N-actual	E[S]-actual	N-bal.ntwk	E[S]forNbal	% off E[S]act
[1, 10, 2]					
a=0.9	2	5.37	2	5.37	0.00
a=0.95	2	5.6	2	5.6	0.00
a=0.99	2	5.79	2	5.79	0.00
[1,100,2]					
a=0.9	8	16.95	7	16.95	0.00
a=0.95	8	18.53	7	18.53	0.00
a=0.99	8	19.98	7	19.98	0.00
[1,250,4]					
a=0.9	9	36.87	8	37.45	1.57
a=0.95	9	40.27	8	40.43	0.40
a=0.99	10	43.41	8	43.41	0.00

Table 3.1: Performance of Balanced Network Approximation

E[S]					
m2 = 5					
m = 3					
[c, l , h]	N-actual	E[S]-actual	N-bal.ntwk	E[S]forNbal	% off E[S]act
[1, 10, 2]					
a=0.9	1	2.74	1	2.74	0.00
a=0.95	1	2.87	1	2.87	0.00
a=0.99	1	2.97	1	2.97	0.00
[1,100,2]					
a=0.9	3	8.69	3	8.69	0.00
a=0.95	3	9.17	3	9.17	0.00
a=0.99	3	9.57	3	9.57	0.00
[1,250,4]					
a=0.9	4	18.24	4	18.24	0.00
a=0.95	4	19.1	4	19.1	0.00
a=0.99	4	19.82	4	19.82	0.00

Table 3.2: Performance of Balanced Network Approximation

E[S]					
m2 = 10					
m = 7					
[c, l , h]	N-actual	E[S]-actual	N-bal.ntwk	E[S]forNbal	% off E[S]act
[1, 10, 2]					
a=0.9	1	1.29	1	1.29	0.00
a=0.95	1	1.36	1	1.36	0.00
a=0.99	1	1.41	1	1.41	0.00
[1,100,2]					
a=0.9	2	5.68	2	5.68	0.00
a=0.95	2	5.93	2	5.93	0.00
a=0.99	2	6.13	2	6.13	0.00
[1,250,4]					
a=0.9	3	12.22	2	12.3	0.65
a=0.95	3	12.62	2	12.75	1.03
a=0.99	3	13	2	13.1	0.77

Table 3.3: Performance of Balanced Network Approximation

Approximation 3.3.4 *For $a < 1$, given a high demand for make-to-stock products, the optimal N^* which minimizes the average cost is given by:*

$$N^* \approx \frac{ln\left[\frac{ah-(\frac{m_2 a}{m}-1)c}{(l/m_2)(1-a)^2}\right]}{ln\ a}$$

Since we have $m_2 << 1$, and $c < h << l$, for small N's, $E[S]$ can be approximated by the term

$$E[S] \approx \left(\frac{l}{m_2}\right)\frac{(1-a)a^N}{(1-a^{N+1})} \tag{3.3}$$

However, for $m_2 << 1$, the number we must hold in finished goods inventory will not be small. So let us examine the function as N becomes large. The dominant term then is the finished goods holding cost. Since $N(1-a) >> a^{N+1} - a$ we have

$$E[S] \approx \frac{hN}{(1-a^{N+1})} \tag{3.4}$$

In order to compute the N^*, we looked at the intersection of these two curves, that is, we equated equations (3.3) and (3.4). This did not provide a satisfactory approximation. However, the intersection of the second curve, equation (3.4), with the actual $E[S]$ function, equation (3.1), did give a good estimate, especially when $h << l$. Equating the two gives

$$\left(\frac{m_2 a}{m} - 1\right) c \left[\frac{1 - Na^{N+1}(1-a) + a^{N+1}(a-2)}{(1-a)(1-a^{N+1})}\right] + \frac{l}{m_2}\left[\frac{a^N(1-a)}{(1-a^{N+1})}\right] - h\frac{a}{(1-a)(1-a^{N+1})} \approx 0 \qquad (3.5)$$

For large N and with $a < 1$, we have $1 >> a^{N+1}(a-2) - Na^{N+1}(1-a)$, and equation (3.5) reduces to:

$$\left(\frac{m_2 a}{m} - 1\right) c + \frac{l}{m_2} a^N (1-a)^2 - ha \approx 0$$

$$\left(\frac{l}{m}\right)(1-a)^2 a^N \approx ah - \left(\frac{m_2 a}{m} - 1\right) c$$

which yields

$$N^*_{approx} = \frac{ln\left[\frac{ah - (\frac{m_2 a}{m} - 1)c}{(l/m_2)(1-a)^2}\right]}{ln\ a}$$

The conditions under which this approximation holds are:

1. $a < 1$
2. $m_2 << 1$, so that $N^* >> 1$
3. $ah - (\frac{m_2 a}{m} - 1)c < (l/m)(1-a)^2$

The results of this approximation are given in Tables 3.4, 3.5 and 3.6 and as can be seen, even if N^* is far off, the value of $E[S]_{min}$ predicted by this approximate model is very close to the actual value.

E[S]					
m2=0.001					
m = 0.0008					
[c, l, h]	N-actual	E[S]-actual	N-approx	E[S]forNaprx	% off E[S]act
[1, 10, 2]					
a=0.9	38	79.02	39	79.04	0.02
a=0.95	54	113.38	52	113.71	0.29
a=0.99	80	179.67	*	*	*
[1,100,2]					
a=0.9	60	121.43	61	121.62	0.16
a=0.95	96	195.57	97	195.68	0.06
a=0.99	205	433.96	174	448.13	3.26
[1,250,4]					
a=0.9	62	250.01	62	250.01	0
a=0.95	100	404.35	100	404.35	0
a=0.99	223	916	189	947.24	3.41
* Negative value is obtained since condition 1 is violated					

Table 3.4: Performance of Second Approximation

E[S]					
m2=0.005					
m = 0.003					
[c, l, h]	N-actual	E[S]-actual	N-approx	E[S]forNaprx	% off E[S]act
[1, 10, 2]					
a=0.9	25	55.04	26	55.22	0.33
a=0.95	31	71.85	27	72.89	1.45
[1,100,2]					
a=0.9	45	95	48	96.22	1.28
a=0.95	66	143.58	71	145.14	1.09
[1,250,4]					
a=0.9	47	193.63	49	194.69	0.55
a=0.95	70	293.02	72	293.45	0.15

Table 3.5: Performance of Second Approximation

E[S]					
m2=0.01					
m = 0.007					
[c, l, h]	N-actual	E[S]-actual	N-approx	E[S]forNaprx	% off E[S]act
[1, 10, 2]					
a=0.9	20	42.88	18	43.3	0.98
a=0.95	23	55.4	10	82.41	48.75
[1,100,2]					
a=0.9	39	82.52	40	82.81	0.35
a=0.95	54	119.92	55	119.98	0.05
[1,250,4]					
a=0.9	41	168.36	42	168.9	0.32
a=0.95	58	245	57	245.11	0.04

Table 3.6: Performance of Second Approximation

Chapter 4

Dynamic Control of the Production System

4.1 Problem Description

In this section, we are analyzing the mixed queuing network with a fixed number, N, of make-to-stock jobs, to determine the best scheduling policy for processing the two types of jobs. We again have Poisson arrival of make-to-order jobs with rate λ to station 1, and exponential service rate at each station, with rate μ at station 1 (same rate for both types) and rate μ_2 at station 2. Let $u(t)$ be the action to be taken at time t. The allowable actions are:

- process make-to-order jobs, denoted by 0
- process make-to-stock jobs, denoted by 1

Let X_t^i be the queue length at station 1 of type i customers, including the one in service. Then $X_{t=0} = (x_0, x_1)$ represents the initial queue length of open and close jobs respectively. The state space of the problem is Z^2 where Z is the set of integers. The objective is to determine the best policy for processing jobs at station 1 so as to minimize the total discounted cost:

$$E\left\{\int_0^\infty e^{-\alpha t}[cX_t^0 + h(N - X_t^1)]dt + \sum_{k=1}^{\infty} l\, e^{-\alpha T_k}\right\}$$

where

$$\alpha = \text{discount rate}$$

$$T_k = \text{instants of lost sales}$$

Since it is most expensive to have lost sales, we could conjecture that station 1 should process make-to-stock jobs if their number in finished goods inventory falls below a threshold value, and process make-to-order jobs otherwise.

We can show that the problem is equivalent to a discrete time problem, as in Walrand [15].

- Let potential transitions be all arrivals and all the service completions (that would occur if the queue were never empty), at station 1. These potential transitions occur at a rate:

$$\lambda + \mu + \mu_2$$

- Consider the discounting as resulting from terminating the process after a random time, exponentially distributed with parameter α. Once the process is terminated we incur no additional cost, so the cost to be paid at time t is the original cost provided that the process is running. Thus the average cost at time t given the evolution of the system is equal to $e^{-\alpha t}$ times the original cost. Hence, the discounted cost for the original process is the undiscounted cost for the terminated process.

- We can consider the termination of the process as an additional potential transition. Therefore, potential transitions occur at rate:

$$\lambda + \mu + \mu_2 + \alpha$$

  By scaling time, one can assume that:

$$\lambda + \mu + \mu_2 + \alpha = 1$$

  thus potential transitions occur at each unit of time.

Letting τ_k be the kth potential transition time, the cost can be written as:

$$E\left\{\sum_{k=0}^{\infty}\int_{\tau_k}^{\tau_{k+1}}[cX_t^0+h(N-X_t^1)]dt+\sum_{k=1}^{\infty}l\,1\{\text{lost sales at time }\tau_k\}\right\}$$

Decision epochs are the starts of services of either type of job. The transition probabilities are:

$$P_{(x_0,x_1)(x_0',x_1')}(u)=\begin{cases}\lambda & x_0'=x_0+1, \quad x_1'=x_1\\ \mu_2' & x_0'=x_0, \quad x_1'=x_1+1\\ \mu & x_0'=x_0-1, \quad x_1'=x_1\\ \mu & x_0'=x_0, \quad x_1'=x_1-1\\ \alpha & \text{terminated process}\\ 0 & \text{otherwise}\end{cases}$$

where

$$\mu_2'=\begin{cases}\mu_2 & 0\le x_1\le N\\ 0 & x_1=N\end{cases}$$

Suppose that there are T time units remaining before termination of the process. Let $V_T(x_0,x_1)$ denote the expected minimum cost for the T time units given an initial queue length at Station 1 of (x_0,x_1). This cost would be a function of x_0,x_1 and V_{T-1} where V_{T-1} represents the"future minimum cost" or the expected minimum cost with $T-1$ time units remaining. Then the dynamic programming equations to be solved for the different initial conditions (x_0,x_1) are the following:

Case 1: $x_0>0$ and $0<x_1<N$

$$\begin{aligned}V_T(x_0,x_1) &= cx_0+h(N-x_1)+\lambda V_{T-1}(x_0+1,x_1)+\mu_2 V_{T-1}(x_0,x_1+1)\\ &\quad +\mu\min\{V_{T-1}(x_0-1,x_1),V_{T-1}(x_0,x_1-1)\}\end{aligned}\tag{4.1}$$

Case 2: $x_0>0$ and $x_1=N$

$$\begin{aligned}V_T(x_0,x_1) &= cx_0+h(N-x_1)+\lambda V_{T-1}(x_0+1,x_1)+\mu_2\{l+V_{T-1}(x_0,x_1)\}\\ &\quad +\mu\min\{V_{T-1}(x_0-1,x_1),V_{T-1}(x_0,x_1-1)\}\end{aligned}\tag{4.2}$$

Case 3: $x_0 > 0$ and $x_1 = 0$

$$\begin{aligned} V_T(x_0, x_1) &= cx_0 + h(N - x_1) + \lambda V_{T-1}(x_0 + 1, x_1) + \mu_2 V_{T-1}(x_0, x_1 + 1) \\ &\quad + \mu V_{T-1}(x_0 - 1, x_1) \end{aligned} \tag{4.3}$$

Case 4: $x_0 = 0$ and $0 < x_1 < N$

$$\begin{aligned} V_T(x_0, x_1) &= cx_0 + h(N - x_1) + \lambda V_{T-1}(x_0 + 1, x_1) + \mu_2 V_{T-1}(x_0, x_1 + 1) \\ &\quad + \mu V_{T-1}(x_0, x_1 - 1) \end{aligned} \tag{4.4}$$

Case 5: $x_0 = 0$ and $x_1 = N$

$$\begin{aligned} V_T(x_0, x_1) &= cx_0 + h(N - x_1) + \lambda V_{T-1}(x_0 + 1, x_1) + \mu_2 \{l + V_{T-1}(x_0, x_1)\} \\ &\quad + \mu V_{T-1}(x_0, x_1 - 1) \end{aligned} \tag{4.5}$$

Case 6: $x_0 = 0$ and $x_1 = 0$

$$V_T(x_0, x_1) = cx_0 + h(N - x_1) + \lambda V_{T-1}(x_0 + 1, x_1) + \mu_2 V_{T-1}(x_0, x_1 + 1) \tag{4.6}$$

Consider case 1: the first term is the WIP holding cost; the second term is the finished goods holding cost; the third term is the future minimum cost if the next transition is a demand for a make-to-order product; the fourth term is the future minimum cost if the next transition is a demand for a make-to-stock product; the fifth term is the future minimum cost if the next transition is a service. This last term indicates that one has the choice of processing either type of product, depending on which would represent the greater savings. In case 2, if the next transition is a "service" at station 2, we incur a lost sale and then the system remains in the same state.

4.2 Properties of the Optimal Policy

We follow Ha's [5] method for analyzing the model as follows. Let $\mathcal{S}$ be the set of functions defined on Z^2 such that if $v \in \mathcal{S}$, then v exhibits the following properties:

- Convexity

$$v(x_0 - 1, x_1) - v(x_0, x_1) \text{is decreasing in } x_0$$

$$v(x_0, x_1 - 1) - v(x_0, x_1) \text{is decreasing in } x_1$$

- Supermodularity

$$v(x_0 - 1, x_1) - v(x_0, x_1) \text{is decreasing in } x_1$$

$$v(x_0, x_1 - 1) - v(x_0, x_1) \text{is decreasing in } x_0$$

- Diagonal dominance

$$v(x_0, x_1 - 1) - v(x_0 - 1, x_1) \text{is decreasing in } x_1, \text{and increasing in } x_0.$$

Note that supermodularity and diagonal dominance together imply convexity; see Appendix A.

We define the optimal operator G as

$$\begin{aligned} Gv(x_0, x_1) = \; & cx_0 + h(N - x_1) + \lambda v(x_0 + 1, x_1) + \mu_2 v(x_0, x_1 + 1) \\ & + \mu \min\{v(x_0 - 1, x_1), v(x_0, x_1 - 1)\} + \text{constant} \end{aligned}$$

where

$$\text{constant} = \begin{cases} 0 & \text{case 1} \\ \mu_2 l & \text{case 2} \end{cases}$$

Proposition 4.2.1 *If $v \in \mathcal{S}$, then*

1. *$min[v(x_0 - 1, x_1), v(x_0, x_1 - 1)] \in \mathcal{S}$*
2. *$Gv \in \mathcal{S}$*

The proof for Proposition 4.2.1 is given in Appendix B. Based on this Proposition we can prove the following theorem:

Theorem 4.2.2 *1. The optimal cost function V is convex, supermodular and has diagonal dominance.*

2. There exists an optimal stationary policy.

3. Given the WIP levels x_0 and x_1, there exists a function $B(x_0)$ such that it is optimal to produce make-to-stock products if $x_1 > B(x_0)$ and produce make-to-order products otherwise.

4. $B(x_0)$ is nondecreasing in x_0.

The proof of Theorem 4.2.2 is given in Appendix B. Part (2) states that $B(x_0)$ is stationary over time. Part (3) states that make-to-stock jobs should be processed if their inventory level falls below $[N - B(x_0)]$, and also gives the form of the optimal policy. The function $B(x_0)$ is defined by

$$B(x_0) = \min\{x_1 : V(x_0 - 1, x_1) - V(x_0, x_1 - 1) > 0\}$$

This function is a switching function which determines the priority for production. Part (4) shows that if it is optimal to produce product i in preference to product j, it remains optimal to do so if either the number of orders for i increases or the number of orders for j decreases. This follows from the diagonal dominance of V.

4.3 Computation of the Optimal Policy

We compute the optimal policy using the value iteration method, as discussed in Howard [7]. As mentioned in Section 4.1, we keep a fixed number, N, of make-to-stock jobs. Since the number of open jobs in the system could be infinite, we need to truncate the state space in order to make the computing feasible. We let the maximum number of open jobs be M. Then the term $\lambda V_{T-1}(x_0 + 1, x_1)$ in equations (4.1), (4.2) and (4.3) is approximated at the boundary $x_0 = M$ by the following term:

$$\lambda\{c + V_{T-1}(M, x_1)\}.$$

The reasoning for this follows the same logic as for the lost sales cases (equations (4.2) and (4.5)), that is, if we have a demand for a make-to-order product when $x_0 = M$,

we incur a cost c for that time period and then the system remains in the same state. We computed the costs for values of $M = 70, 75$ and 80. We found that the costs for the latter two cases differed only in the first decimal place and the optimal policy in all three cases is the same and hence chose the value of $M = 80$. The costs used are $c = 1$, $l = 100$ and $h = 2$, and the target base stock level, $N = 10$. The iteration process was stopped when $V_T(x_0, x_1) - V_{T-1}(x_0, x_1) < 0.0001 V_T(x_0, x_1)$.

The parameters to be used for these computations were determined bearing in mind the conditions that $\lambda + \mu + \mu_2 + \alpha = 1$ and $\rho_1 = \frac{\lambda}{\mu} + \frac{\mu_2}{\mu} < 1$. As mentioned in Section 1.2, we will only consider cases where $\mu > \mu_2$. Additionally we will not take μ to be much greater than μ_2 since this would mean that make-to-stock items would be produced much faster than they were being requested, thus the case would not be so interesting. Taking $\alpha = 0.01$,

- for $\rho_1 = 0.8$

$$\mu = 0.55,\ \lambda = 0.04,\ \mu_2 = 0.4$$

- for $\rho_1 = 0.98$

$$\mu = 0.50,\ \lambda = 0.09,\ \mu_2 = 0.4.$$

The optimal policy for these values is shown in Figures 4-1 and 4-2. These figures illustrate that as stated in Theorem 4.2.2, the switching function, $B(x_0)$, is indeed increasing in x_0. The interpretation of Figure 4-1, for example, is that if all 10 closed jobs are at station 1 and there are less than 56 open jobs waiting, then production priority should be given to closed jobs. Similarly, if there are 9 closed jobs and less than 19 open jobs waiting to be processed, then priority should be given to the closed jobs.

Note that if the system starts in the shaded region, the optimal policy dictates serving closed jobs. This would push the system towards the curve $B(x_0)$ in a westerly or north-westerly direction. Conversely if the system starts in the unshaded region, optimally open jobs should be served and the system is pushed towards the $B(x_0)$ curve in an southerly or south-easterly direction.

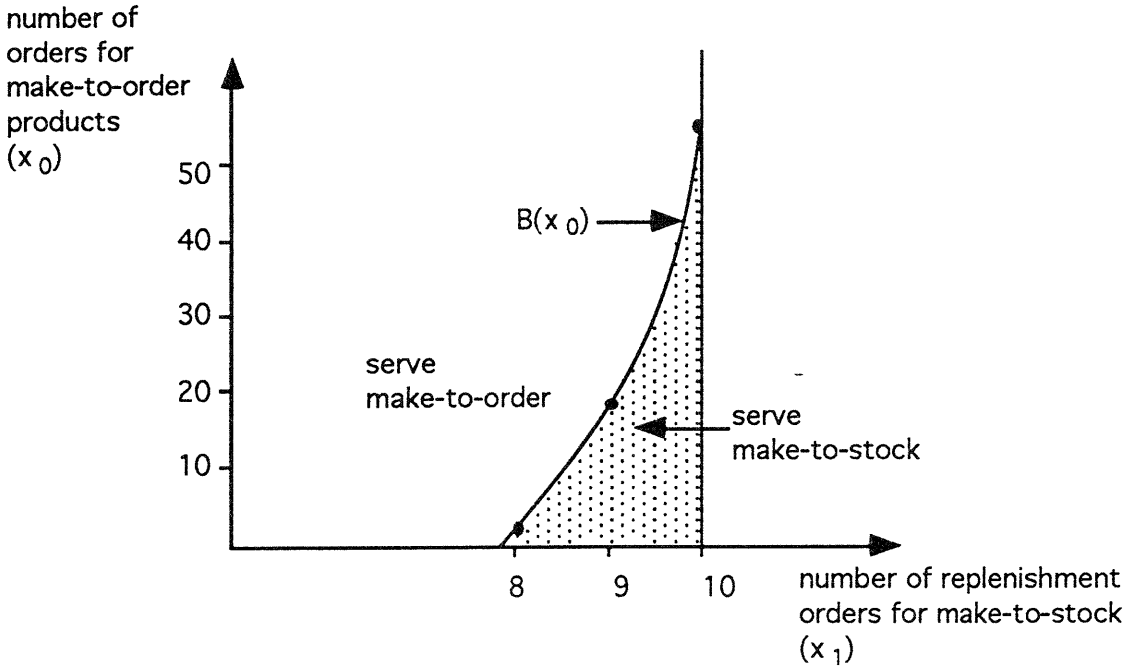

Figure 4-1: The Optimal Policy for $\rho = 0.8$

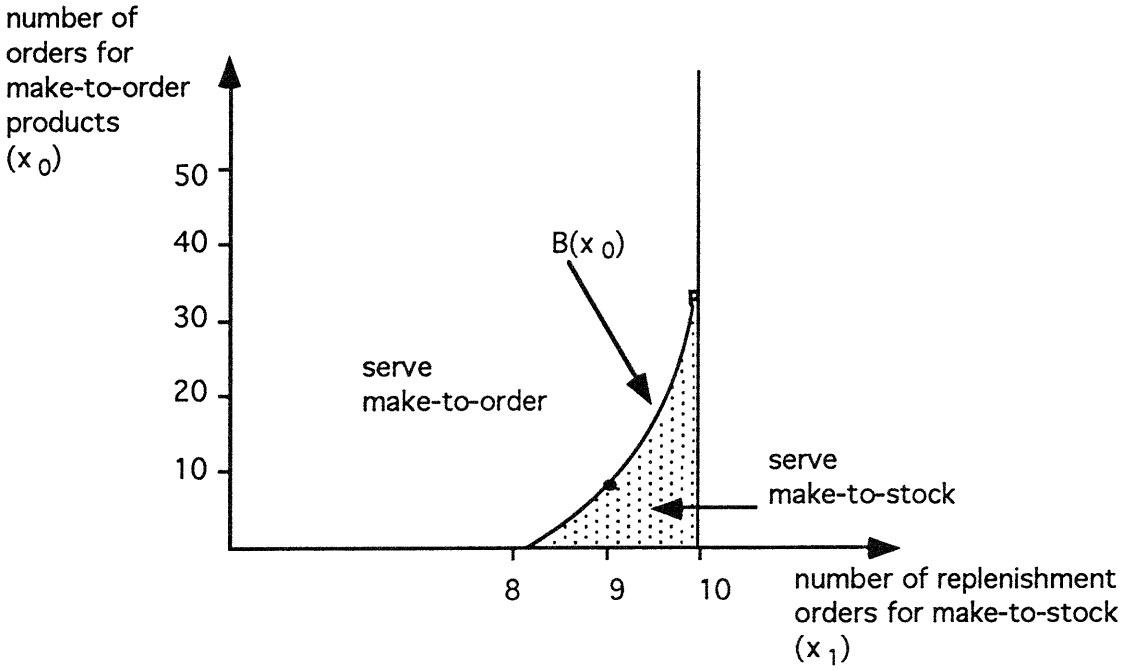

Figure 4-2: The Optimal Policy for $\rho = 0.98$

Case	[c,l,h]	ρ1	(μ2, λ, μ)
A1	[1,100,2]	0.8	(0.55,0.04,0.4)
A2		0.9	(0.52,0.07,0.4)
A3		0.98	(0.50,0.09,0.4)
B1	[1,250,4]	0.8	(0.55,0.04,0.4)
B2		0.9	(0.52,0.07,0.4)
B3		0.98	(0.50,0.09,0.4)

Table 4.1: Scenarios for Comparing Scheduling Policies

	Initial Condition (x0,x1) = (0,0)			
	Efficiency = 100 x (cost of optimal policy)/(cost of policy)			
	Policy (1)	Policy (2)	Policy (3)	Policy (4)
Case	Priority open	Priority closed	Priority open unless x1=N	Priority open unless x1≥N-1
A1	99.69%	91.65%	99.87%	99.44%
A2	99.08%	82.15%	99.64%	98.41%
A3	98.48%	75.15%	99.47%	97.40%
B1	99.40%	93.64%	99.64%	99.17%
B2	98.06%	86.78%	98.85%	97.53%
B3	96.53%	82.14%	97.94%	95.84%

Table 4.2: Comparison of the Optimal Policy With Other Scheduling Policies

4.4 Comparison of the Optimal Policy With Other Scheduling Policies

We would like to compare the performance of the optimal policy with other scheduling priorities. We tested the optimal policy against the following four policies:

(1) Non-premptive priority for open jobs.

(2) Non-premptive priority for closed jobs.

(3) Serve closed jobs if their number at station 1 (replenishment orders) is N, otherwise serve open jobs.

(4) Serve closed jobs if their number at station 1 is $N-1$ or N, otherwise serve open jobs.

	Worst Initial Condition (x0,x1) for policy			
	Efficiency = 100 x (cost of optimal policy)/(cost of policy)			
	Policy (1)	Policy (2)	Policy (3)	Policy (4)
Case	Priority open	Priority closed	Priority open unless x1=N	Priority open unless x1≥N-1
A1	(10,10)	(11,4)	(7,9)	(8,9)
	84.10%	53.37%	94.91%	85.83%
A2	(9,10)	(10,4)	(6,9)	(7,9)
	87.83%	52.30%	96.33%	87.22%
A3	(7,10)	(9,3)	(4,9)	(6,9)
	90.26%	51.89%	97.30%	88.21%

Table 4.3: Comparison of the Optimal Policy With Other Scheduling Policies

We used two different cost cases and tested them under three values of the traffic intensity at station 1. These scenarios are summarized in Table 4.1. The results comparing the optimal policy with these four scheduling policies are presented in Table 4.2. Policy (2) performs the worst and policy (3) performs the closest to the optimal policy. Note that the extent to which the optimal policy outperforms the other policies depends upon the initial condition of the system. For example, if the system were to start with all N closed jobs at station 1 and yet production priority was always given to open jobs, we would expect this policy to do significantly worse than the optimal policy. However, if the system were to start with station 1 empty and employ such a scheduling policy we would expect a better performance. This is illustrated in Table 4.3 which indicates, for each policy, the initial condition with the highest cost relative to the optimal policy, and the percentage by which it is higher. Thus, following policy (1) when $\rho_1 = 0.8$ and the system starts with station 1 empty the total discounted cost is 0.27% worse than the optimal policy. However, if there are ten open jobs and ten closed jobs initially waiting at station 1 then policy (1) performs 18.83% worse than the optimal policy. As one might expect, given the chosen parameters, a policy of always giving priority to closed jobs performs the worst. Consider the case when $\rho_1 = 0.9$ and the system starts with (9,3) and policy (2) is chosen. This means that although seven make-to-stock items are already in finished goods inventory, make-to-stock orders are given priority in production. The chance that there will be more than seven demands for make-to-stock items before another

Initial condition (0,0)		Worst initial condition (x0,x1)	
Policy	Efficiency for N=2*	Policy	Efficiency for N=2*
Optimal	100%	Optimal	100%
Policy (3)	100%	Policy (3)	99.23%
Policy (2)	99.70%	Policy (2)	96.87%
Policy (1)	98.32%	Policy (1)	94.63%
Policy (4)	94.35%	Policy (4)	87.89%
* the ranking remains the same for N=1			

Table 4.4: Ranking of the Policies for $\bar{N} = 2$

Initial condition (0,0)		Worst initial condition (x0,x1)	
Policy	Efficiency for N=3*	Policy	Efficiency for N=3*
Optimal	100%	Optimal	100%
Policy (3)	99.92%	Policy (3)	99.29%
Policy (1)	98.14%	Policy (1)	94.02%
Policy (2)	97.20%	Policy (2)	91.80%
Policy (4)	94.30%	Policy (4)	88.39%

Table 4.5: Ranking of the Policies for $N = 3$

one is produced is so small that the optimal policy actually performs 92.34% better in this case.

Looking at Tables 4.2 and 4.3 one is struck by the dismal performance of policy (2). The reason for this is the choice of N. If N is much smaller than the optimal value for minimum cost of the system it would make little sense to give priority to closed jobs since we would just be increasing the holding costs for finished goods inventory. However, if N were much smaller, then policy (2) may fare better. Consider case A2 for example, with $\rho_1 = 0.9, \mu = 0.52, \lambda = 0.07$ and $\mu_2 = 0.4$ and $[c, l, h] = [1, 100, 2]$. From equation (3.2) we get $N^*_{FIFO} \approx 4$ which gives us an idea for the size of $N_{optimal}$. The results of comparing the optimal policy with the other four scheduling policies for different values of N are presented in Tables 4.4 - 4.7. The policies are ranked in order of performance. As one can see from Table 4.4, for small N policy (2) and (3) are very close to optimal. However, as N increases, policy (2) quickly assumes last position in the ranking; Tables 4.6 and 4.7.

Initial condition (0,0)		Worst initial condition (x0,x1)	
Policy	Efficiency for N=4*	Policy	Efficiency for N=4*
Optimal	100%	Optimal	100%
Policy (3)	99.69%	Policy (3)	99.30%
Policy (1)	98.00%	Policy (1)	93.13%
Policy (4)	94.69%	Policy (4)	88.39%
Policy (2)	93.23%	Policy (2)	84.23%
* the ranking remains the same for 3<N<14			

Table 4.6: Ranking of the Policies for $N = 4$

Initial condition (0,0)		Worst initial condition (x0,x1)	
Policy	Efficiency for N=14*	Policy	Efficiency for N=14*
Optimal	100%	Optimal	100%
Policy (3)	99.77%	Policy (3)	94.37%
Policy (1)	99.74%	Policy (4)	86.07%
Policy (4)	99.31%	Policy (1)	84.83%
Policy (2)	83.42%	Policy (2)	44.21%
* the ranking remains the same for N≥14			

Table 4.7: Ranking of the Policies for $N = 14$

As discussed in Section 4.3 the state space of the problem is divided into two regions and the optimal policy determines the best priority for scheduling depending upon the region. For example, in Figure 4-1 if the system were to start in the shaded region and production priority was given to open jobs then we would be moving away from th optimal strategy and incurring higher costs. Interestingly, this contradicts the policy described by Carr *et al.* [3], who propose a "No B/C strategy". That is, as discussed in Section 1.3, the B/C items would be our equivalent of open jobs and would always be given production priority over A items, or closed jobs. our optimal policy clearly states that priority for B/C items would be a sub-optimal strategy to pursue if the system were to start in the "shaded region".

Chapter 5

Conclusions and Further Research

In this thesis we model a production facility which produces two types of items, one make-to-stock and the other make-to-order, as a mixed queuing network. We first consider a static model in the sense that a service discipline is specified for the production facility and a base-stock policy for the make-to-stock items is followed. The problem of setting a target base-stock level, N^*, was tackled with one of the following two objectives in mind: the first was to achieve a desired fill rate for the make-to stock products, or alternatively, the second was to minimize the average cost incurred by the system per unit time. In trying to achieve a particular fill rate we derived expressions for N^* given either a FIFO priority scheme at the production facility or a pre-emptive resume priority for the make-to-stock products.

For the second objective of minimizing the average cost per unit time, we considered a FIFO service discipline. We derived an exact expression for N^* when the traffic intensity at station 1, ρ_1, is unity. When $\rho_1 < 1$ we develop approximations for N^* for the case where demand rate per week for make-to-stock products is low and also for the case where their demand rate per week is high. We note that for the low demand case, both the N^*_{approx} and the corresponding minimum cost are very close to the actual values. However, for the high demand case, although the N^*_{approx} is not very close to N^*_{actual}, the minimum cost evaluated using this approximation is very close to the actual minimum cost.

Next we turned to a dynamic model of the system to determine the best scheduling policy at the production center. We found that the optimal policy is a switching

function. The function divides the state space into two distinct regions, such that if the system starts in one region it would be optimal to give production priority to make-to-oder items, whereas starting in the other region would indicate giving production priority to make-to-stock items. We compared the optimal policy with four other scheduling policies and found that the cost of the optimal policy can be upto 44.21% less than the cost of the chosen policy depending upon the initial state of the system.

We have modeled the production facility as a mixed queuing network and in order to analyze this network a number of simplifying assumptions were made. This work could be extended to relax some of these assumptions. For example:

- each type of product could have a different service rate at the production center,
- a different service time distribution could be considered since the exponential distribution may not accurately represent what actually happens at a production facility,
- demand could be batch arrivals,
- due dates and machine failure could be a consideration,
- set-up costs could be included in the analysis when switching production of the items,
- a different cost structure could be considered for the system.

However, by pursuing these recommendations we may render the problem intractable. Nevertheless, using this work as a basis, simulations could be run to gain further insight.

Returning to the comment in section 4.4 regarding the article by Carr *et al.* [3], we note that the optimal policy developed in this thesis may be in contradiction to the policy in their paper. Further investigation is required, however, using the parameters of their work and applying it to a suitable situation which would validate either policy.

Appendix A

Proof of Convexity of v

To show convexity of v, we need to show:

(1) $v(x_0 - 1, x_1) - v(x_0, x_1)$ is decreasing in x_0

(2) $v(x_0, x_1 - 1) - v(x_0, x_1)$ is decreasing in x_1.

For part (1), if $v(x_0 - 1, x_1) - v(x_0, x_1)$ is decreasing in x_0, then we must have

$$v(x_0 - 2, x_1) - v(x_0 - 1, x_1) \geq v(x_0 - 1, x_1) - v(x_0, x_1).$$

From supermodularity of v, we have

$$v(x_0 - 1, x_1 - 1) - v(x_0 - 1, x_1) \geq v(x_0, x_1 - 1) - v(x_0, x_1)$$

$$\Rightarrow v(x_0 - 1, x_1 - 1) - v(x_0, x_1 - 1) \geq v(x_0 - 1, x_1) - v(x_0, x_1).$$

From diagonal dominance of v, we have

$$v(x_0, x_1 - 1) - v(x_0 - 1, x_1) \geq v(x_0 - 1, x_1 - 1) - v(x_0 - 2, x_1)$$

$$\Rightarrow v(x_0 - 2, x_1) - v(x_0 - 1, x_1) \geq v(x_0 - 1, x_1 - 1) - v(x_0, x_1 - 1).$$

Thus

$$v(x_0 - 2, x_1) - v(x_0 - 1, x_1) \geq v(x_0 - 1, x_1) - v(x_0, x_1).$$

For part (2), if $v(x_0 - 1, x_1) - v(x_0, x_1)$ is decreasing in x_1, then we must have

$$v(x_0, x_1 - 2) - v(x_0, x_1 - 1) \geq v(x_0, x_1 - 1) - v(x_0, x_1).$$

From supermodularity of v, we have

$$v(x_0 - 1, x_1 - 1) - v(x_0, x_1 - 1) \geq v(x_0 - 1, x_1) - v(x_0, x_1)$$

$$\Rightarrow v(x_0 - 1, x_1 - 1) - v(x_0 - 1, x_1) \geq v(x_0, x_1 - 1) - v(x_0, x_1).$$

From diagonal dominance of v, we have

$$v(x_0, x_1 - 2) - v(x_0 - 1, x_1 - 1) \geq v(x_0, x_1 - 1) - v(x_0 - 1, x_1)$$

$$\Rightarrow v(x_0, x_1 - 2) - v(x_0, x_1 - 1) \geq v(x_0 - 1, x_1 - 1) - v(x_0 - 1, x_1).$$

Thus

$$v(x_0, x_1 - 2) - v(x_0, x_1 - 1) \geq v(x_0, x_1 - 1) - v(x_0, x_1)$$

□

Appendix B

Proof of Optimal Policy

Proof of Proposition 4.2.1

Define function m by $m(x_0, x_1) = \min[v(x_0 - 1, x_1), v(x_0, x_1 - 1)]$. We want to show that if $v \in \mathcal{S}$ then $m \in \mathcal{S}$ also; that is, m is convex, supermodular and has diagonal dominance. We take an arbitrary point $(x_0, x_1) \in Z^2$ and consider all possible values of m at that point and its neighboring points and show that m exhibits these properties in each case.

Let v_{ij} denote $v(x_0 - i, x_1 - j)$. Figure B-1 shows the function v evaluated at (x_0, x_1) and its neighbors. Since $v \in \mathcal{S}$, we have the following inequalities:

Convexity

$$\begin{aligned} v_{00} - v_{10} &\geq v_{10} - v_{20} \quad \geq v_{20} - v_{30} \\ v_{00} - v_{01} &\geq v_{01} - v_{02} \quad \geq v_{02} - v_{03} \\ v_{01} - v_{11} &\geq v_{11} - v_{21} \\ v_{10} - v_{11} &\geq v_{11} - v_{12} \end{aligned}$$

Supermodularity

$$\begin{aligned} v_{00} - v_{10} &\geq v_{01} - v_{11} \quad \geq v_{02} - v_{12} \\ v_{01} - v_{02} &\geq v_{11} - v_{12} \\ v_{00} - v_{01} &\geq v_{10} - v_{11} \quad \geq v_{20} - v_{21} \end{aligned}$$

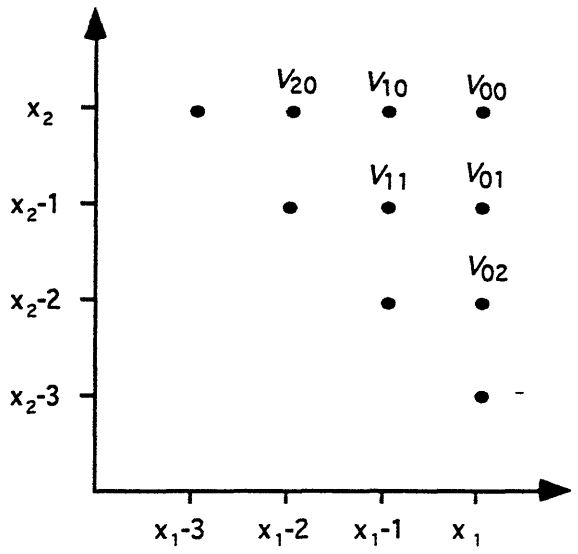

Figure B-1: Values of v evaluated at points in Z^2

Diagonal Dominance

$$\begin{aligned}
v_{02} - v_{12} &\geq v_{11} - v_{21} \geq v_{20} - v_{30} \\
v_{20} - v_{21} &\geq v_{11} - v_{12} \geq v_{02} - v_{03} \\
v_{01} - v_{11} &\geq v_{10} - v_{20} \\
v_{10} - v_{11} &\geq v_{01} - v_{02}
\end{aligned}$$

For convenience, denote

$$\begin{aligned}
m_{ij} &= m(x_0 - i, x_1 - j) \\
&= \min[v_{i-1,j}, v_{i,j-1}]
\end{aligned}$$

Let a_{ij} = indicator variable that indicates the optimal action when the state is $(x_0 - i, x_1 - j)$. Then $a_{ij} = 1$, if $v_{i,j-1} \leq v_{i-1,j}$. Thus,

$$a_{ij} = \begin{cases} 0 & \text{if } m_{ij} = v_{i-1,j} \\ 1 & \text{if } m_{ij} = v_{i,j-1} \end{cases}$$

Consider the following matrix of indicator variables:

$$\begin{bmatrix} a_{01} & a_{11} \\ a_{00} & a_{10} \end{bmatrix}$$

The possible cases for this matrix are:

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

If $a_{10} = 0$, open customers are being served, then when there is 1 more open customer in the system we must still serve open customers. Then $a_{00} \neq 1$, and the following configuration is not possible:

$$\begin{bmatrix} & \\ 1 & 0 \end{bmatrix}$$

Similarly, if $a_{01} = 1$, closed customers are being served, then when there is 1 more closed customer in the system we must still serve closed customers. Then $a_{00} \neq 0$, and the following configuration is not possible:

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Then the remaining cases are:

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Supermodularity

Consider the following matrix of indicator variables:

$$\begin{bmatrix} a_{01} & a_{11} \\ a_{00} & a_{10} \end{bmatrix}$$

We want to show that

$$m_{00} - m_{10} \geq m_{01} - m_{11}$$

Case 1: $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$

$$\begin{aligned} m_{01} = v_{11} &\quad \rightarrow v_{11} < v_{02} \\ m_{00} = v_{10} &\quad \rightarrow v_{10} < v_{01} \\ m_{10} = v_{20} &\quad \rightarrow v_{20} < v_{11} \\ m_{11} = v_{21} &\quad \rightarrow v_{21} < v_{12} \end{aligned}$$

$$\begin{aligned} m_{00} - m_{10} &= v_{10} - v_{20} \\ &\geq v_{11} - v_{21} \quad \text{from supermodularity of } v \\ &= m_{01} - m_{11} \end{aligned}$$

Case 2: $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$

$$\begin{aligned} m_{01} = v_{11} &\quad \rightarrow v_{11} < v_{02} \\ m_{00} = v_{10} &\quad \rightarrow v_{10} < v_{01} \\ m_{10} = v_{11} &\quad \rightarrow v_{11} \leq v_{20} \\ m_{11} = v_{21} &\quad \rightarrow v_{21} < v_{12} \end{aligned}$$

$$\begin{aligned} m_{00} - m_{10} &= v_{10} - v_{11} \\ &\geq v_{01} - v_{11} \quad \text{since } v_{10} < v_{01} \\ &\geq v_{11} - v_{21} \quad \text{from convexity of } v \\ &= m_{01} - m_{11} \end{aligned}$$

Case 3: $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$

$$\begin{aligned} m_{01} = v_{11} &\quad \rightarrow v_{11} < v_{02} \\ m_{00} = v_{10} &\quad \rightarrow v_{10} < v_{01} \\ m_{10} = v_{20} &\quad \rightarrow v_{20} < v_{11} \\ m_{11} = v_{12} &\quad \rightarrow v_{12} \leq v_{21} \end{aligned}$$

$$\begin{aligned} m_{00} - m_{10} &= v_{10} - v_{20} \\ &> v_{10} - v_{11} \quad \text{since } v_{20} < v_{11} \\ &\geq v_{11} - v_{12} \quad \text{from convexity of } v \\ &= m_{01} - m_{11} \end{aligned}$$

Case 4: $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$

$$\begin{aligned} m_{01} = v_{11} &\quad \rightarrow v_{11} < v_{02} \\ m_{00} = v_{01} &\quad \rightarrow v_{01} \leq v_{10} \\ m_{10} = v_{11} &\quad \rightarrow v_{11} \leq v_{20} \\ m_{11} = v_{21} &\quad \rightarrow v_{21} < v_{12} \end{aligned}$$

$$\begin{aligned} m_{00} - m_{10} &= v_{01} - v_{11} \\ &\geq v_{11} - v_{21} \quad \text{from convexity of } v \\ &= m_{01} - m_{11} \end{aligned}$$

Case 5: $\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$

$$\begin{aligned} m_{01} = v_{11} &\quad \rightarrow v_{11} < v_{02} \\ m_{00} = v_{10} &\quad \rightarrow v_{10} < v_{01} \\ m_{10} = v_{11} &\quad \rightarrow v_{11} \leq v_{20} \\ m_{11} = v_{12} &\quad \rightarrow v_{12} \leq v_{21} \end{aligned}$$

$$\begin{aligned} m_{00} - m_{10} &= v_{10} - v_{11} \\ &\geq v_{11} - v_{12} \quad \text{from convexity of } v \\ &= m_{01} - m_{11} \end{aligned}$$

Case 6: $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$

$$\begin{aligned} m_{01} = v_{02} &\quad \rightarrow v_{02} \leq v_{11} \\ m_{00} = v_{01} &\quad \rightarrow v_{01} \leq v_{10} \end{aligned}$$

$$m_{10} = v_{11} \quad \rightarrow v_{11} \leq v_{20}$$
$$m_{11} = v_{21} \quad \rightarrow v_{21} < v_{12}$$

$$\begin{aligned} m_{00} - m_{10} &= v_{01} - v_{11} \\ &\geq v_{02} - v_{12} \quad \text{from supermodularity of } v \\ &\geq v_{02} - v_{21} \quad \text{since } v_{21} < v_{12} \\ &= m_{01} - m_{11} \end{aligned}$$

Case 7: $\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$

$$m_{01} = v_{11} \quad \rightarrow v_{11} < v_{02}$$
$$m_{00} = v_{01} \quad \rightarrow v_{01} \leq v_{10}$$
$$m_{10} = v_{11} \quad \rightarrow v_{11} \leq v_{20}$$
$$m_{11} = v_{12} \quad \rightarrow v_{12} \leq v_{21}$$

$$\begin{aligned} m_{00} - m_{10} &= v_{01} - v_{11} \\ &\geq v_{11} - v_{12} \quad \text{from convexity of } v \\ &= m_{01} - m_{11} \end{aligned}$$

Case 8: $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

$$m_{01} = v_{02} \quad \rightarrow v_{02} \leq v_{11}$$
$$m_{00} = v_{01} \quad \rightarrow v_{01} \leq v_{10}$$
$$m_{10} = v_{11} \quad \rightarrow v_{11} \leq v_{20}$$
$$m_{11} = v_{12} \quad \rightarrow v_{12} \leq v_{21}$$

$$\begin{aligned} m_{00} - m_{10} &= v_{01} - v_{11} \\ &\geq v_{02} - v_{12} \quad \text{from supermodularity of } v \\ &= m_{01} - m_{11} \end{aligned}$$

Since (x_0, x_1) is arbitrary, the following holds for any (x_0, x_1):

$$m(x_0 - 1, x_1 - 1) - m(x_0 - 1, x_1) \geq m(x_0, x_1 - 1) - m(x_0, x_1)$$

Therefore m is supermodular. □

Diagonal Dominance

Consider the following matrix of indicator variables:

$$\begin{bmatrix} a_{01} & a_{11} & \\ & a_{10} & a_{20} \end{bmatrix}$$

We want to show that

$$m_{01} - m_{11} \geq m_{10} - m_{20}$$

Case 1: $\begin{bmatrix} 0 & 0 & \\ & 0 & 0 \end{bmatrix}$

$$\begin{aligned} m_{01} = v_{11} &\quad \rightarrow v_{11} < v_{02} \\ m_{11} = v_{21} &\quad \rightarrow v_{21} < v_{12} \\ m_{10} = v_{20} &\quad \rightarrow v_{20} < v_{11} \\ m_{20} = v_{30} &\quad \rightarrow v_{30} < v_{21} \end{aligned}$$

$$\begin{aligned} m_{01} - m_{11} &= v_{11} - v_{21} \\ &\geq v_{20} - v_{30} \quad \text{from diagonal dominance of } v \\ &= m_{10} - m_{20} \end{aligned}$$

Case 2: $\begin{bmatrix} 0 & 0 & \\ & 0 & 1 \end{bmatrix}$

$$\begin{aligned} m_{01} = v_{11} &\rightarrow v_{11} < v_{02} \\ m_{11} = v_{21} &\rightarrow v_{21} < v_{12} \\ m_{10} = v_{20} &\rightarrow v_{20} < v_{11} \\ m_{20} = v_{21} &\rightarrow v_{21} \leq v_{30} \end{aligned}$$

$$\begin{aligned} m_{01} - m_{11} &= v_{11} - v_{21} \\ &\geq v_{20} - v_{30} \quad \text{from diagonal dominance of } v \\ &\geq v_{20} - v_{21} \quad \text{since } v_{21} \leq v_{30} \\ &= m_{10} - m_{20} \end{aligned}$$

Case 3: $\begin{bmatrix} 1 & 0 & \\ & 0 & 0 \end{bmatrix}$

$$\begin{aligned} m_{01} = v_{02} &\rightarrow v_{02} \leq v_{11} \\ m_{11} = v_{21} &\rightarrow v_{21} < v_{12} \\ m_{10} = v_{20} &\rightarrow v_{20} < v_{11} \\ m_{20} = v_{30} &\rightarrow v_{30} < v_{21} \end{aligned}$$

$$\begin{aligned} m_{01} - m_{11} &= v_{02} - v_{21} \\ &\geq v_{02} - v_{12} \quad \text{since } v_{21} < v_{12} \\ &\geq v_{20} - v_{30} \quad \text{from diagonal dominance of } v \\ &= m_{10} - m_{20} \end{aligned}$$

Case 4: $\begin{bmatrix} 0 & 0 & \\ & 1 & 1 \end{bmatrix}$

$$\begin{aligned} m_{01} = v_{11} &\quad \rightarrow v_{11} < v_{02} \\ m_{11} = v_{21} &\quad \rightarrow v_{21} < v_{12} \\ m_{10} = v_{11} &\quad \rightarrow v_{11} \leq v_{20} \\ m_{20} = v_{21} &\quad \rightarrow v_{21} \leq v_{30} \end{aligned}$$

$$\begin{aligned} m_{01} - m_{11} &= v_{11} - v_{21} \\ &= m_{10} - m_{20} \end{aligned}$$

Case 5: $\begin{bmatrix} 1 & 0 & \\ & 0 & 1 \end{bmatrix}$

$$\begin{aligned} m_{01} = v_{02} &\quad \rightarrow v_{02} \leq v_{11} \\ m_{11} = v_{21} &\quad \rightarrow v_{21} < v_{12} \\ m_{10} = v_{20} &\quad \rightarrow v_{20} < v_{11} \\ m_{20} = v_{21} &\quad \rightarrow v_{21} \leq v_{30} \end{aligned}$$

$$\begin{aligned} m_{01} - m_{11} &= v_{02} - v_{21} \\ &> v_{02} - v_{12} \quad \text{since } v_{21} < v_{12} \\ &\geq v_{11} - v_{21} \quad \text{from diagonal dominance of } v \\ &\geq v_{20} - v_{21} \quad \text{since } v_{20} < v_{11} \\ &= m_{10} - m_{20} \end{aligned}$$

Case 6: $\begin{bmatrix} 1 & 0 & \\ & 1 & 1 \end{bmatrix}$

$$m_{01} = v_{02} \quad \rightarrow v_{02} \leq v_{11}$$

$$m_{11} = v_{21} \quad \rightarrow v_{21} < v_{12}$$
$$m_{10} = v_{11} \quad \rightarrow v_{11} \leq v_{20}$$
$$m_{20} = v_{21} \quad \rightarrow v_{21} \leq v_{30}$$

$$\begin{aligned} m_{01} - m_{11} &= v_{02} - v_{21} \\ &> v_{02} - v_{12} \quad \text{since } v_{21} < v_{12} \\ &\geq v_{11} - v_{21} \quad \text{from diagonal dominance of } v \\ &= m_{10} - m_{20} \end{aligned}$$

Case 7: $\begin{bmatrix} 0 & 1 & \\ & 1 & 1 \end{bmatrix}$

$$m_{01} = v_{11} \quad \rightarrow v_{11} < v_{02}$$
$$m_{11} = v_{12} \quad \rightarrow v_{12} \leq v_{21}$$
$$m_{10} = v_{11} \quad \rightarrow v_{11} \leq v_{20}$$
$$m_{20} = v_{21} \quad \rightarrow v_{21} \leq v_{30}$$

$$\begin{aligned} m_{01} - m_{11} &= v_{11} - v_{12} \\ &\geq v_{11} - v_{21} \quad \text{since } v_{12} \leq v_{21} \\ &= m_{10} - m_{20} \end{aligned}$$

Case 8: $\begin{bmatrix} 1 & 1 & \\ & 1 & 1 \end{bmatrix}$

$$m_{01} = v_{02} \quad \rightarrow v_{02} \leq v_{11}$$
$$m_{11} = v_{12} \quad \rightarrow v_{12} \leq v_{21}$$
$$m_{10} = v_{11} \quad \rightarrow v_{11} \leq v_{20}$$
$$m_{20} = v_{21} \quad \rightarrow v_{21} \leq v_{30}$$

$$\begin{aligned} m_{01} - m_{11} &= v_{02} - v_{12} \\ &\geq v_{11} - v_{21} \quad \text{from diagonal dominance of } v \\ &= m_{10} - m_{20} \end{aligned}$$

Since (x_0, x_1) is arbitrary, the following holds for any (x_0, x_1):

$$m(x_0, x_1 - 1) - m(x_0 - 1, x_1) \geq m(x_0 - 1, x_1 - 1) - m(x_0 - 2, x_1)$$

Therefore m has diagonal dominance. □

Convexity The minimum of two convex functions is also convex, therefore m is convex.
Thus we have shown that if $v \in \mathcal{S}$ then $m \in \mathcal{S}$. □

For part (2), since the other terms in the dynamic programming equations are linear, they are also convex, supermodular and have diagonal dominance. Since $\mathcal{S}$ is closed under addition, and multiplication by positive scalars we have the desired result that $Gv \in \mathcal{S}$. □

Proof of Theorem 1
For $v \in \mathcal{S}$ define the function

$$B(x_0) = \min\{x_1 : v(x_0 - 1, x_1) - v(x_0, x_1 - 1) > 0$$

We prove Theorem 1 based on Ha [5] and Porteus [13]. Let a structured decision rule be one that chooses to process closed jobs if their number in finished goods inventory falls below a threshhold value, as in part (2) of Theorem 1. As shown by Porteus, since $\mathcal{S}$ is complete the limit of any convergent sequence in $\mathcal{S}$ will be in $\mathcal{S}$. From the optimality equations one can see that the function $B(x_0)$ as defined above is optimal and from Theorem 5.1 of Porteus [13] parts (1), (2) and (4) hold.

For part (3), by definition of $B(x_0)$,

$$v(x_0 - 1, B(x_0) - 1) < v(x_0, B(x_0) - 2)$$

From diagonal dominance of v, since $v(x_0, x_1 - 1) - v(x_0 - 1, x_1)$ is increasing in x_0, we have

$$v(x_0 + 1, B(x_0) - 2) > v(x_0, B(x_0) - 1)$$

But in order for this to hold, since $v(x_0, x_1 - 1) - v(x_0 - 1, x_1)$ is decreasing in x_1, we must have

$$B(x_0 + 1) > B(x_0)$$

Thus $B(x_0)$ is increasing in x_0. □