Inventory systems with multi-class demands under continuous review : policies and algorithms

Ghosh, Sugoutam

2012

Ghosh, S. (2012). Inventory systems with multi-class demands under continuous review: policies and algorithms. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/48672>

<https://doi.org/10.32657/10356/48672>

NANYANG
TECHNOLOGICAL
UNIVERSITY

**INVENTORY SYSTEMS WITH MULTI-CLASS
DEMANDS UNDER CONTINUOUS REVIEW:
POLICIES AND ALGORITHMS**

SUGOUTAM GHOSH
SCHOOL OF MECHANICAL & AEROSPACE ENGINEERING
2012

**INVENTORY SYSTEMS WITH MULTI-CLASS
DEMANDS UNDER CONTINUOUS REVIEW:
POLICIES AND ALGORITHMS**

SUGOUTAM GHOSH

School of Mechanical and Aerospace Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2012

ACKNOWLEDGEMENTS

I would like to express my gratitude to both of my supervisors, A/P Rajesh Piplani and Prof S. Vishwanathan for their support throughout this research. Their motivation, patience, encouragement and knowledge have made this thesis possible. Their guidance helped me throughout the course of this research and writing of this thesis.

I would like to thank the lab technician, Mr Teo Hai Beng for providing the necessary help during this research. I would also like to thank all my lab mates who were always there with their constructive suggestions. I am also grateful to my friends who made me feel Singapore home away from home.

Last but not the least; I would like to thank my wife for managing things when I was not there and her patience during this research and my son, Rishaan for being a good boy and making things easier for me.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	vii
LIST OF TABLES	ix
ABSTRACT	xi
CHAPTER 1: INTRODUCTION	1
1.1 Introduction	1
1.2 Service differentiation	4
1.2.1 Separate stock policy	5
1.2.2 Common stock policy	6
1.3 Inventory rationing	7
1.3.1 Advantages of inventory rationing	10
1.3.2 Static and dynamic inventory rationing	11
1.3.3 Applications of inventory rationing	11
1.4 Motivationand research scope	14
1.5 Research objectives	16
1.6 Organization of the report	17
1.7 Summary	19
CHAPTER 2: LITERATURE REVIEW	20
2.1 Introduction	20
2.2 Inventory rationing	24
2.2.1 Continuous review systems	27

2.2.2 Periodic review systems	31
2.3 Make-to-order systems	33
2.4 Dynamic inventory rationing	36
2.5 Joint replenishment problem	37
2.6 Research plan	39
2.7 Summary	43
CHAPTER 3: INVENTORY RATIONING WITH DETERMINISTIC DEMAND	44
3.1 Introduction	44
3.2 Model formulation	48
3.2.1 Cost expressions	53
3.3 Model analysis and optimization	55
3.4 Algorithm to determine T & α_i	58
3.5 Numerical examples	63
3.6 Summary	69
CHAPTER 4: INVENTORY RATIONING FOR CONTINUOUS REVIEW SYSTEM WITH MULTIPLE CUSTOMER CLASSES	71
4.1 Introduction	71
4.2 Inventory rationing model with multiple customer classes	75
4.2.1 Evaluation of the policy cost	90
4.2.2 Solution algorithm	97
4.2.2.1 Bounds on $\bar{S} = \sum_{i=1}^n S_i$ and Q	98
4.3 Numerical results	99
4.4 Summary	106

CHAPTER 5: A NEW CLASS OF MULTI-BIN POLICIES FOR INVENTORY RATIONING IN A CONTINUOUS REVIEW ENVIRONMENT	107
5.1 Introduction	107
5.2 The new two-bin policy	116
5.2.1 Evaluation of the policy cost	124
5.2.2 Solution algorithm	137
5.2.2.1 Bounds on $\bar{S}$ and Q	138
5.3 A general 3-bin policy	139
5.3.1 Cost calculation for the 3-bin policy	141
5.4 Numerical results	143
5.4.1 Comparison of expected total cost	143
5.5 Summary	152
CHAPTER 6: CONCLUSIONS	154
6.1 Conclusions	155
6.2 Future research directions	158
PUBLICATIONS	161
REFERENCES	162
APPENDIX A: Detailed explanation of Deshpande et al. (2003)'s policy using a 2-bin framework, with threshold clearing	167
APPENDIX B: Detailed derivation of cost expression for 3-bin policy	169
APPENDIX C: Computational results for a 5-class model into an aggregated 3-class and 2-class problem	175
APPENDIX D: Computational results for a 3-class model into an aggregated 2-class model	178

LIST OF FIGURES

Figure 1.1 Inventory rationing	8
Figure 3.1 Typical replenishment cycle with inventory rationing	51
Figure 3.2 Algorithm to determine the optimal cycle time, T^*	62
Figure 3.3 Effect of varying demand on total cost	66
Figure 3.4 Effect of demand on number of backorders	67
Figure 3.5 Total cost variation vs. stock out cost ratio	68
Figure 3.6 Effect of varying stock out cost on rationing co-efficients	69
Figure 4.1 Inventory rationing for multiple classes	79
Figure 4.2 Inventory rationing with threshold clearing fro 3 classes	80
Figure 4.3 Inventory rationing as n -bin policy	82
Figure 4.4 Flowchart for n -bin policy	84
Figure 4.5 Pseudo code for clearing of the i^{th} class demand in the l^{th} order cycle in the rationing policy represented by n -bin	85
Figure 5.1 Deshpande et al. (2003)'s inventory rationing with threshold clearing	112
Figure 5.2 Proposed 2 –bin policy (d_1 stands for class 1 demand, d_2 stands class 2 demand)	113
Figure 5.3 Deshpande et al.'s policy represented as a different 2 –bin policy (d_1 stands for class 1 demand, d_2 stands class 2 demand)	114
Figure 5.4 Proposed 3 –bin policy (d_1 stands for class 1 demand, d_2 stands class 2 demand)	115
Figure 5.5 Flowchart of the proposed 2-bin policy	120
Figure 5.6 Representation of inventory position changes	125

LIST OF TABLES

Table 3.1 Notations used	49
Table 3.2 Optimal parameters for (T, α) and (Q, r, K) policies	65
Table 4.1 Notations for multiple-class rationing model	75
Table 4.2 Example of inventory rationing with threshold clearing for 3 classes	88
Table 4.3 Optimal parameters for 5 customer classes	101
Table 4.4 Cost increase from using the solution of an aggregated 3-class model for the 5-class problem	103
Table 4.5 Cost increase from using the solution of an aggregated 2-class model for the 5-class problem	105
Table 5.1 Notations for chapter 5	117
Table 5.2 Example of the proposed policy with threshold clearing	122
Table 5.3 Effect on holding and penalty cost for different cases of l_1 and l_2	128
Table 5.4 Different possibilities of demand arrival sequence for Case 4	134
Table 5.5 Comparison of expected cost of the two policies	144
$[\lambda_1 = \lambda_2 = 5, p_1 = 500, p_2 = 500, \pi_1 = \pi_2 = 0, A = 500, L = 1]$	
Table 5.6 Comparison of expected cost of the two policies	145
$[\lambda_1 = \lambda_2 = 5, p_1 = 500, p_2 = 100, \pi_1 = 100, \pi_2 = 50, A = 500, L = 1]$	
Table 5.7 Comparison of expected cost of the two policies parameters	146
$[\lambda_1 = \lambda_2 = 10, p_2 = 600, \pi_1 = \pi_2 = 0, A = 100, L = 0.25, h = 250]$	
Table 5.8 Comparison of expected cost of the two policies parameters	148
$[\lambda_1 = \lambda_2 = 10, p_1 = 6000, \pi_1 = \pi_2 = 0, A = 100, L = 0.25, h = 250]$	
Table 5.9 Comparison of service levels for the two policies	150

$$[\lambda_1 = \lambda_2 = 10, \pi_1 = \pi_2 = 0, A = 100, L = 0.25, h = 250]$$

Table 5.10 Comparison of service levels for the two policies	151

$$[\lambda_1 = \lambda_2 = 10, \pi_1 = \pi_2 = 0, A = 100, L = 0.25, h = 250]$$

Table 6.1 Combination of inventory levels for 3 classes' problem	159

Table B.1 Different cases of l_1 and l_2 for a 3-bin policy	170

Table C.1 Cost comparison (Case 1)	175

Table C.2 Cost comparison (Case 2)	176

Table C.3 Cost comparison (Case 3)	177

Table D.1 Cost comparison	178

ABSTRACT

As the after-market service industry's prospects grow, it is becoming imperative for the OEMs to satisfy their customer's varying service requirements and do so economically. Differentiating the demand into different classes according to their service requirement and using inventory rationing to fulfill the demand is one technique to accomplish this. In military materials management, when the same spare part is requisitioned from various divisions, an inventory control policy that prioritizes the request for fulfillment is needed so that the division with the most mission-critical needs obtain the part first. Critical level inventory rationing, where the lower class demand is either backordered or not satisfied after on hand inventory reaches a pre-determined critical level (called rationing level), is the most cost effective inventory policy used for this type of problems. Since the revenue generated from aftermarket service and sales of spare parts is huge, firms are more concerned to solve this type of problems with an objective of minimizing the inventory cost and maximizing the service level imparted to different customer classes.

In the first part of the dissertation, inventory rationing policies are considered for ' n ' customer classes, and a model is developed to find the optimal policy parameters, assuming deterministic demand. A general penalty cost structure including delay cost and stock out is considered in developing this model; which has not been considered in the other studies. It is assumed that demands are backordered from a customer class after its run-out time. Run-out time of a customer class is defined as the time after which demand from that class is rejected and satisfied when next replenishment arrives. The total inventory cost to be minimized consists of inventory carrying cost, ordering cost and penalty cost. An algorithm is developed to determine

the optimal cycle time and consequently, the run-out time for each customer class. Finally some numerical examples are provided to demonstrate the effectiveness of the algorithm.

The part second part analyses an inventory system with multiple customer classes and considered rationing policies. Under the assumption of Poisson demand, a model is developed and the expression for expected cost is derived. Apart from developing an algorithm to solve the multiple-class problem, the impact of collapsing an n -class model into a 2-class or 3-class model (by aggregating several classes together) is investigated, numerically. In most of the cases, an aggregated 2-class or 3-class model achieves most of the benefits of rationing with a cost that is only marginally higher than an n -class model.

The final part of this dissertation reports a new class of two-bin policy where two separate bins are kept for the two classes. When demand arrives from a class, it is satisfied from the bin designated for it, if inventory is available. The higher class can use the lower class's bin when its own stock runs out but not vice versa. Demand follows a Poisson process and is differentiated by the penalty cost of not satisfying that class' demand. The policy reserves some stock for both the classes, but the higher class is given more priority by allowing it to use the lower class' stock. The exact expression for the expected cost of a policy as well as search algorithm for determining the optimal policy is developed. Although critical level inventory rationing provides an inventory policy with a lower cost, its shortcoming is in respect to the service levels it provides to the different classes. Numerical results show that this policy provides a much higher service level to the lower class, compared to inventory rationing, at a slightly higher cost, for most of the cases. It was also found that the proposed policy outperforms the critical level rationing policy when the holding cost is low and lead time is relatively higher.

Chapter 1

Introduction

1.1 Introduction

With increasing customer awareness and environmental concerns of the used products, the need for aftermarket service (AMS) has gone up tremendously. It is estimated that “aftermarket parts and services have a profit margin as much as ten times that of the initial product sales and account for 20% to 30% of revenues and 40% of profits for most manufacturers” (Frontline Solutions, July 2004) and thus it is one of the most profitable areas for the firms. Original equipment manufacturers (OEM) generally provide aftermarket service to their customers for the products or services they sell to them. One facing challenge of the OEMs is supplying the variety of products/systems they sell by making the key parts available in the right quantity, at the right location.

Since keeping large inventory is an expensive proposition for the firm, and there may be thousands of spare parts required for aftermarket service, it is important to design an appropriate inventory policy that minimizes the cost of carrying the inventory and also provide a desirable service level to the customers.

In many cases the firm may wish to provide different levels of service to different customers/orders depending on their requirements. Sometimes the firms may offer the customers different options to choose from. This is natural as not all the customers are of equal importance to the firm and satisfying more important customers realizes higher profits as well as goodwill for the firm than a less important customer. A customer with whom the firm has a long standing relation may be considered more important than a customer who has just started business with the firm. Satisfying demand from a big Multi-national company may be more critical than satisfying that from a local company. It also helps the firm to plan their production/services according to different service requirement e.g. a customer's order with low service requirement can be delayed to satisfy a customer with high service requirement. The firms sometimes use service contracts to differentiate these customers. Several options may be available for a service contract e.g. on site repair within 2hr, 4hr or 1 day. Another type of service contract specifies the cost of service and urgency of service (e.g. level and time of service). The customers choose between these available service contracts and are served according to their service contracts. A “*gold contract*” might provide a 99% fill rate within twenty-four hours, while a “*silver contract*” promises a 90% fill rate within two days. Also sometimes the firms differentiate the orders from one another rather than differentiating the

customers. In this case, orders are segregated into different ‘*order classes*’. One example of this is normal and emergency orders from same customer. All the emergency orders form the *Emergency* order class even though the orders are from different classes.

In other cases the firm allows the customer to decide on the delivery time and based on that the firm can provide a discount to those customers who choose a more relaxed delivery window. The customers with same service needs or service requirement within a pre-defined range are treated as the same customer class and are served according to their service requirements. Sometimes the penalty cost of not providing the service on time also determines its importance. This penalty cost generally refers to the delay cost per unit per unit time and also stock out cost per unit. The customers with higher priority have the higher delay and stock out cost. In most of the literature on multiple customer classes, the classes are prioritized based on their penalty cost (Deshpande et al., Arslan et al.).

Satisfying these different types of customers is of prime importance as the revenue generated from aftermarket service can be huge and there is not much competition in the aftermarket service as the customers generally go to the same firm which sold them the product/service. But maintaining a service network for aftermarket service can be very costly as a large number of service engineers and large quantity of slow-moving spare parts have to be maintained to satisfy the demand for different customers. The proper trade-off between the total cost and the fulfillment of service requirement is the main concern of the firm. Thus the objective of this research is to develop optimal inventory policies for service differentiated customer classes.

1.2 Service differentiation

Most of inventory literature considers the demand from different customer (order) classes to be equally important, to be satisfied on a first-come-first-served (FCFS) manner. However, there are cases where customers need to be given different priority depending on their importance to the firm, as discussed in the last section. Generally the profit a customer brings to the firm, or the cost incurred in not satisfying the demand determines its importance. Sometimes it is the type of order (order class) which determines its importance to the firm. In this research, the customer class and order class are treated the same.

The following scenario is very common in the Maintenance Repair and Overhaul (MRO) industry. A common part may be installed in various equipments, but the criticality of the equipment may be different. The breakdown of some particular equipment might be more expensive than other equipments. Therefore, more critical equipment should be given higher priority, to minimize the chance of breakdown and these equipments or customers form a customer/order class with higher priority. One of the classic examples of this can be seen in Military establishment. Military has thousands of spare parts which are common to several military equipments, but the importance of all these equipments is not the same for the military. Since the military work is critical in nature, a break-down of a machine or a vehicle or artillery with high importance can have serious repercussions on their operations and thus the equipment with higher importance

or priority have to be attended to immediately whereas the less important equipments can wait.

In inventory control, a customer class can be represented by the penalty costs for not satisfying the demand from that particular class. This phenomenon of service differentiated customer classes has found wide application in the fixed capacity, perishable inventory systems, such as airline and hotel industry (Lee and Hersh (1993)). Lately, this approach has found its way into the MRO industry where the criticality of some orders makes it imperative to treat it differently. The main objective is to satisfy the service level of each customer class with least amount of inventory, incurring minimum total cost.

Firms have employed different techniques to design the inventory system when there is more than one customer class. The policies utilized to serve different customer classes are discussed next.

1.2.1 Separate stock policy: The most common policy to satisfy the demands from several customer classes is known as separate stock policy. In this policy the firm maintains separate stock for each customer class. If the firm has ' n ' number of customer classes it maintains separate stock for all the n customer classes. When demand comes from a particular customer class, it is satisfied from the stock/inventory designated for that particular class. If there is no inventory available for that particular class, the demand is satisfied only when next replenishment arrives (backorder) or demand is treated as lost

(lost sales). The advantage of this policy is that it improves the service level of the lower class. But it does not take advantage of inventory pooling and incurs higher cost.

1.2.2 Common stock policy: Another way to satisfy demands from several customer classes with different service needs is to keep a common stock for all the classes. When demand comes from any of the n class it is satisfied from the common stock on a FCFS basis as long as inventory is available. If the stock is empty, the demand is backordered or lost. This policy does benefit from inventory pooling and provides a higher service level to the customers. Kim (2002) showed the advantage of inventory pooling in a system where supply lead times are endogenously generated by a production system with finite capacity. Deshpande et al (2003b) studied US military and found that they keep a common stock for all the customers; demand from each class is then satisfied on a FCFS basis from the common stock. They also found that the inventory is kept based on the highest service level requirement out of the all classes i.e. service requirement of every customer class is rounded up to the highest level. They termed this policy as ‘*round-up*’ policy. If customer class A requires a service level of 85% and customer class B requires a service level of 90%, according to ‘round-up’ policy a common stock should be maintained such that an aggregate service level of 90% is provided to both the classes.

The main advantage of this policy is that it provides a higher service level to all the customers, even to the customers who do not require the higher service level (and are not paying for it). The main disadvantage of this policy is that the firm has to invest in higher

level of inventory for not so important customer classes. Secondly, in this policy lower priority customer classes can preempt higher priority classes, as this policy gives same priority to all the customer classes. Eppen (1979) considered a single-product inventory system and showed the advantage of centralization of inventory over separate stock. He showed that the expected holding and penalty costs in a system with separate stock exceeded the costs in a system with a common stock. He also showed that the savings generated from common stock policy over separate stock were dependent on the correlation of demands from different customer classes. Baker (1986) also showed the effects of commonality on safety stock in a simple inventory system subject to a service level constraint. He assumed two types of components: unique and common, and showed that a reduction in common stock should be accompanied by an increase in unique stock to exploit the full benefits of commonality.

1.3 Inventory rationing

Another technique to address the service differentiated customer classes is inventory rationing. This policy is widely applied in the industries where service level requirement of different customer classes vary considerably. The main feature of this policy is that when there is not enough inventory to satisfy all the demands coming from different customer classes, it reserves some inventory for the future high priority demands and low priority demands are rejected. The policy works as follows: all the demands from different customer classes are satisfied on a FCFS basis until on hand inventory drops to a threshold level (called critical level or rationing level); after on hand

inventory falls below rationing level demand from lower priority customer classes is rejected and high priority demands continue to be satisfied. The rejected demands can either be treated as lost sales or backordered to the next cycle. The inventory rationing policy is characterized by the critical level/ inventory level to start rejecting lower priority class. If the inventory system has multiple ordering opportunities to replenish the inventory, the ordering policy also interacts with the rationing policy. In that case the key decision variables are order quantity (Q); reorder point (r) and critical levels (or rationing levels, K) and can be denoted as (Q, r, K) inventory system. The graphical representation of inventory rationing policy for two customer classes is shown in Figure 1.1.

Figure 1.1 Inventory rationing

Figure 1.1 shows an inventory system which follows a continuous (Q, r) review inventory policy. When the inventory position (on hand inventory plus inventory on-order minus backorders) reaches reorder point (at time t), r , an order of quantity Q is placed which arrives after lead time L (at time $t + L$). All the demands from both the customer classes are satisfied on a FCFS basis as long as on hand inventory level is greater than or equal to the rationing level, K . Define t_{B_i} as the time when i^{th} class demands begin to backorder. K is called critical level or rationing level for the lower priority customer class. After on hand inventory level falls below K , only demand from higher priority customer class i.e. class 1 is satisfied as long as inventory is available. From time t_{B_2} , all the demand from lower priority class, class 2, is rejected until the next replenishment arrives. At time t_{B_1} the on hand inventory level becomes zero and demand from higher priority class is also rejected. If it is assumed that the demands not satisfied immediately are satisfied when replenishment arrives, then BO_1 and BO_2 are the total number of higher and lower class backorders, respectively, before replenishment arrives.

Sometimes rejected demands are never satisfied and can be treated as lost sale. But when demand is backorders, there need to have some clearing mechanism to clear the backorders. One of the most important clearing mechanisms used in literature is threshold clearing. This clearing mechanism would be discussed in Chapter 2.

Some firms follow an inventory policy with periodic inventory review where, unlike continuous review, inventory is checked only periodically, such as weekly or monthly. The difference in inventory rationing, in case of periodic review, is that at each

review epoch the decision is made whether to place an order to replenish the stock. Initial research on rationing was done assuming periodic inventory review. Only during recent years have researchers considered continuous review policy which is more realistic and practical. With the advent of computers and software, it has become easier to keep track of the inventory on a regular basis and now most of the firms follow continuous review of inventory. Consequently, in this research, a continuous inventory review environment is considered.

1.3.1 Advantage of inventory rationing

The advantage of rationing policy is that it overcomes the shortcomings of both the separate stock policy and the round up policy. Inventory rationing makes use of inventory pooling. It increases the service level of the higher priority class to a great extent and also cut down the penalty cost. In case of inventory rationing, unlike round-up policy, lower priority demands cannot preempt higher priority demands. Some inventory is always reserved for the higher priority demands and thus fill rate for higher priority demands increases. Since fulfilling demands from higher priority class realizes more profit than the lower priority classes, the total revenue generated is greater. Sometimes, if the profit from all the customers is same, satisfying high priority customers more than the low priority customers increases the goodwill of the firm and also future business prospects. Effectively, inventory rationing removes the shortcomings of separate stock

policy by keeping lower stock but also tries to provide higher service level to the higher priority customer by reserving inventory for it.

1.3.2 Static and dynamic inventory rationing

In the literature on differentiated service, two types of rationing policies are used: static critical level rationing policy and dynamic critical level rationing policy. Critical level for a particular customer class is defined as the value of the on hand inventory below which demands from that class are rejected. In static rationing, the critical level does not change with time whereas in dynamic rationing policy the critical level may change over time. The critical level depends on inventory position in case of dynamic rationing, whereas in static rationing it is fixed in advance. In dynamic rationing, the rationing level changes as time for replenishment arrival gets closer. If the replenishment is due in the very near future, the rationing level may decrease, sometimes even approaching zero, as there is no need to reserve inventory for future higher class demand.

1.3.3 Applications of inventory rationing

Inventory rationing has found widespread applications in industries where multiple customer classes require different level of service, such as airline and hotel. The inventory in such system is perishable, besides being very expensive. This research, however, focuses on the area of spare parts inventory management. A part may be installed in various equipments, but the criticality of the equipment may vary. The breakdown of some equipments are more costly than others. In this case, the more critical

equipment should be given higher priority than the equipment with lower criticality. These critical equipments or machines thus form a customer class with higher priority and demand for spare parts from these customers are to be satisfied immediately. Inventory rationing solves this type of problem by satisfying the demand for spare parts at the lowest cost.

Another application area of inventory rationing is replenishing stock of a warehouse. The warehouse may place routine order with the suppliers; at times it may also have to place emergency orders for out-of-stock products. The urgency for fulfilling these two types of orders is not the same as an emergency order needs to be fulfilled immediately whereas normal orders may wait. The orders from the warehouse can be divided into two demand classes and rationing policy may be employed to satisfy the demand so that total profit of fulfilling the orders is maximized.

In section 1.2, some examples of service differentiation were presented. For all of these examples, inventory rationing is the most useful technique to satisfy demand from service differentiated classes. One of the classic examples of application of inventory rationing is revenue management. The problem is characterized by some customers' willingness to pay more for the product or the services than others. Room reservation in a hotel and seat reservation on an airplane are considered as revenue management problems. The main characteristic of this problem is that the capacity (seats or hotel rooms) is fixed and perishable (cannot be replenished). Inventory rationing is very effective in solving this problem by reserving (rationing) some of the seats (or rooms) a certain time before the actual date of the flight (or date of reservation of the room) and can increase the profit

for the company. Researchers have shown great interest in revenue management problems. Belobaba (1989) examined the booking limits for airline seats with different price classes. Lee and Hersh (1993) develop a discrete-time dynamic programming model for finding the optimal booking policy.

Distribution of products from a warehouse to different demand centers is another area ripe for application of inventory rationing. The cost of transporting the products from the warehouse to different demand centers is different but the price of the product is the same everywhere. Products at different demand centers earn different profit for the company. If enough inventory is not available to satisfy all the demand from different demand centers, the company may have to implement rationing policies to reserve inventory for more profitable markets.

The service contracts also force firms to implement inventory rationing. Since different contracts realize different profits to the firm, it is obvious that some demands from the less lucrative contracts have to be sacrificed so that demands from the more lucrative contracts can be satisfied. Inventory rationing technique can be used for these situations. The firms, who have service contracts with their customers, have started to implement inventory rationing in their inventory control systems. Urban (2000) and Bassok and Anupindi (1999) have studied this scenario.

Inventory rationing is widely used in Military materials management. It is very useful in deciding when to serve a particular division or machine and when not to. Torkis (1968) presented a problem faced by Army Material Command and suggested inventory

rationing to solve the problem. Recently, Deshpande (2003) carried out a similar study for US military.

1.4 Motivation and research scope

As discussed in section 1.3.3, it is clear that inventory rationing has wide applications in a variety of fields, ranging from military materials management to airlines industry. Since inventory rationing overcomes the problems of separate stock policy and ‘round-up’ policy (Deshpande et al. (2003)), and reduces a company’s cost, organizations have increasingly started to implement inventory rationing.

In order to design an effective inventory rationing policy, it is essential to know the rationing parameters, such as reorder point, order quantity and critical levels, which minimize the total inventory cost. Many researchers have considered inventory rationing problems and tried to determine the optimal rationing parameters under different settings. But most of them considered two customer class problem which is restrictive and cannot be applied in all the cases. Others have considered more than two customer classes but have only derived the approximate expressions for the optimal parameters.

Static inventory rationing has found more attention than dynamic inventory rationing. Although static rationing policy is easy to understand and simple to implement, dynamic rationing policies may be more economical, but difficult to implement as they consider the time left to the next replenishment arrival. In this research, a static rationing policy will be considered because of its ease of use and applicability.

The inventory rationing models mostly consider stochastic demand. Only one researcher considered deterministic demand but did not consider the general penalty cost structure which includes both delay cost per unit time per unit and stock out cost per unit. A general cost structure considering both time based delay cost and one time stock out cost represents the real life scenario better than only considering delay cost as witnessed in the automobile industry and is considered in this research. In this research, the optimal rationing policy for ' n ' customer classes is developed under the assumption of deterministic demand and a general cost structure. When there is little demand fluctuation over the horizon, demand can be assumed to be deterministic.

For stochastic demand, most of the research considered two customer classes only. No researcher has derived the exact cost expressions when the firm has more than two customer classes. During a study conducted in an automobile industry for this research, it was found that they consider three order classes based on their importance to the firm: regular order, emergency order and the most important, VOR (vehicle off road). Because n -class problem is more general and realistic, an inventory rationing problem with ' n ' customer classes and stochastic demand are also considered in this research.

Critical level inventory rationing is found to be the most useful tool to satisfy customers when their service requirement varies. It reserves inventory in anticipation of the higher class demand arrival by rejecting demand from the lower class. But when there are fewer higher class demands at the end of the replenishment cycle, the lower class demand is backordered even when there is inventory. So, inventory rationing provides a very high service level to the higher priority class at the expense of lower priority class'

service level. No research has tried to develop any other inventory policy which provides better service level to the lower class, without compromising the service level for the higher class. In this research, a new policy is presented which attempts to increase the service level of the lower priority class. The objectives of this research are presented in the next section.

1.5 Research objectives

Design of inventory policies for service differentiated customer classes has been researched for a very long time. Most researchers have looked at inventory rationing and developed models for it, mainly for two customer classes.

This research aims at determining the optimal inventory policy which minimizes the expected inventory cost for a differentiated service system; the research objectives are as follows:

1. Review the literature pertaining to service differentiation and inventory rationing, and determine the research gaps.
2. Develop a model for inventory rationing assuming deterministic demand (for n customer classes) and develop the total cost expression for the same under a general cost structure.
3. Develop an inventory rationing model for ' n ' customer classes under stochastic demand.
4. Examine the effect of consolidating ' n ' customer classes into fewer classes.

5. Propose a new inventory policy which provides higher service level to lower priority class without compromising the service level of higher priority class.

1.6 Organization of the report

The remainder of this report is organized as follows:

In chapter 2, the literature on inventory rationing is presented. The literature reviewed is divided in two parts –continuous review systems and periodic review systems. Literature for dynamic inventory rationing as well as make-to-order systems is also discussed. Literature related to joint replenishment problem is reviewed next. The literature on the solution algorithm for different inventory problems is also presented. A research plan to close the gap between this research and the relevant literature is discussed at the end of the chapter.

In chapter 3, static inventory rationing policies for service differentiated customer classes in a continuous review environment are considered and expressions for total inventory cost are derived assuming fixed lead time and general cost structure. ‘ n ’ customer classes are considered whose service requirement are different. Demand from each class is assumed to be deterministic and known. It is also assumed that all the rejected demands are backordered to the next cycle. An algorithm is developed to determine the optimal parameters, namely cycle time and rationing coefficients (fraction of cycle time after which demand is backordered from a class). Finally, some numerical examples are presented to demonstrate the efficacy of the policy.

In chapter 4, an expression for expected inventory cost is developed for ' n ' customer classes considering inventory rationing. It is assumed that demand from all the classes follow Poisson process and lead time of replenishment is fixed. Inventory is continuously reviewed and replenished by base stock (Q, S) policy. Demand is backordered when it cannot be fulfilled from the stock and like the previous model, incurs both delay cost and stock out cost. Also an algorithm is presented to determine the optimal parameters under variable bounds. Next, the cost of considering five classes and that of two and three classes is computed based on the expressions developed and a comparative study is presented. Also some managerial insights are presented for consolidating different classes together.

In chapter 5, a new inventory policy is proposed to solve the service differentiation problem, but only two customer classes are considered. Two separate bins are kept for the two classes and when demand arrives from one class, it is satisfied from its bin. When bin for higher class runs out, the demand from the higher class can be satisfied from the lower class' bin but lower class' demand is not allowed to be filled from the higher class' bin. Demand from both the classes is backordered if lower class' bin is empty. When a replenishment order arrives, higher class returns back all the items it borrowed from the lower class' bin. The inventory is replenished by a (Q, S) policy where quantity Q is ordered when demand from the last order becomes Q and the base stock of each bin is raised to S_i . The expression for expected inventory cost for this policy is developed under Poisson demand. It is assumed that lead time for replenishment is fixed. The backordered demand incurs two penalty costs: delay cost per unit per time and

a onetime stock out cost. Based on the cost expression, an algorithm is presented to determine the optimal parameters. Finally, some numerical experiments are conducted to calculate the expected cost and the service levels of the two classes for the proposed policy and rationing policy. The condition under which the proposed policy outperforms the rationing policy is also mentioned.

In chapter 6, conclusions from this research and directions for future work are presented.

1.7 Summary

In this chapter, background of the research problem is presented. Different policies to tackle the problem are discussed. The most widely used technique inventory rationing is discussed in detail along with its application area. Then the research scope for this problem along with research objectives of this research is presented. The contribution expected from this research is also presented at the end of the chapter. In the next chapter, the literature about service differentiation and inventory rationing as well as joint replenishment policies will be reviewed and research plan will be presented.

Chapter 2

Literature review

2.1 Introduction

In this chapter the literature on inventory policies for differentiated service as well as the algorithms to solve those policies is reviewed. The problem of allocating inventory to different customer classes with different service needs is similar to the yield management problem. The classic example of an yield management problem is the allocation of seats (e.g. economy class on a flight) by airlines to different customers who are willing to pay different prices for the same seat. The main characteristics of this type of problem are that the inventory (of seats) is limited and is perishable. Many studies have been carried out in this area. Belobaba (1989) uses rationing levels to allocate seats on an airplane to different customers at different prices. There is also some research available where inventory is dynamically allocated to the different customers.

The main difference between yield management and spare parts inventory management problem is that spare parts inventory can be replenished when required. Since inventory is fixed, in case of yield management problems, the decision variables are rationing levels (when to stop selling a lower priced seat) and prices for different customer types. But for spare parts inventory management, as parts can be ordered and replenished, the ordering policy becomes of paramount importance. Secondly, unlike yield management problems, it is possible to hold inventory, but at a cost, for the spare part inventory management. So even though the problems look similar, spare part inventory management problem for differentiated service is far more difficult to analyze. Also, inventory control for differentiated service is significantly different from classical inventory problem where all customers are treated same; but in case of differentiated service, customers are classified according to their service level requirement or the price they are willing to pay for the service. If there are two customer classes, sometimes it is possible to find the exact solution; in case of multiple customer classes, the problem is very difficult to analyze and solve.

A common method to address the problem of allocating inventory to different customer classes with service level requirement is to maintain separate stock for each of the customer classes (separate stock policy). As mentioned in Chapter 1 this method has the disadvantage of not pooling the inventory, and in some cases it is very difficult to separate the inventory due to operational problems. This policy reduces the probability of rejection for lower class demand but it does not take advantage of inventory pooling and incurs higher inventory cost.

A more useful and popular way to serve differentiated customers is to maintain a common stock for all the classes and satisfy them on a FCFS basis. This policy does benefit from inventory pooling and provides a higher service level to the customers, even to those who do not require the higher service level. Many researchers have considered part commonality problem where different products require the same component. This problem is similar to the differentiated service problem where different customer classes require the same item; Eppen (1979) considered single-product inventory system and showed the advantage of centralization of inventory over keeping separate stock. Baker et al. (1986) also showed the effect of commonality on safety stock in a two-product two-level inventory model with a service constraint. Kim (2002) showed the advantage of inventory pooling in a system where supply lead times are endogenously generated by a production system with finite capacity. He also showed the advantages of consolidating inventory from multiple locations to a single location. But when there are multiple customer classes it is difficult to decide how much inventory to keep in the common stock, considering different classes require different service levels. Deshpande et al. (2003) studied US military and found that they keep common stock based on the highest level of service requirement and demand from each class is satisfied on a FCFS basis from the common stock. The downside to this policy is that lower priority customer classes may preempt higher priority classes as the policy gives same priority to all the customer classes. They numerically showed that this policy was inferior to inventory rationing.

Kleijn and Dekker (1998) studied different cases where several demand classes can be easily distinguished. They only considered single location inventory systems and

presented real life examples like airline industry, spare parts industry, and a case where there is one warehouse and a number of retailers. They also provided an overview of the literature considering inventory systems with several customer classes.

The most researched inventory policy to address this problem of satisfying service differentiated customer classes is inventory rationing. In inventory rationing, inventory is reserved in anticipation of the arrival of higher class demand after inventory level reaches some pre-determined rationing level. Since there are multiple ordering opportunities; this rationing policy is dependent on the ordering policy and, when inventory rationing is considered with an ordering policy, it is very difficult to analyze. Also for a given ordering policy, in most cases, it is not very easy to find an optimal inventory policy. Due to the difficulty in tackling the problem analytically many researchers consider the rationing problems for a given ordering policy and only two customer classes with different service needs.

The remainder of this chapter is organized as follows: in section 2.2 the literature on inventory rationing policy is discussed. First the literature which considers continuous review system is discussed and then the literature for periodic review system is discussed. The research on make-to-orders systems are discussed in section 2.3. Next in section 2.4, the literature on dynamic rationing policy is presented briefly. The literature on the joint replenishment problem and research related to the algorithms for determining the optimal parameters for different inventory systems is presented in section 2.5. In section 2.6 the research gaps as well as the research plans are highlighted. Finally a summary is presented in section 2.7.

2.2 Inventory rationing

The problem of allocating inventory to different customer classes has found wide attention from the researchers for a long time. Veinott (1965) was the first to consider this problem for ' n ' demand classes in a multi-period, single product, non-stationary inventory environment. Demands are differentiated into different classes, based on either their occurrence time during a period or the relative importance to fill them, and a periodic review policy with zero lead time is considered. Demands from different customer classes are satisfied, starting from the highest priority customer class, from the on hand inventory and any unmet demand is backordered. The backorders are fulfilled according to demand class priority when replenishment arrives at the end of the period. It was assumed that different customer classes incur different backorder cost. The author also assumed non stationary cost structure with the ordering cost being proportional to the amount ordered. Based on these assumptions, the reorder point and order quantity were derived, but without considering rationing levels. It was also showed that it was optimal to replenish inventory with base stock policy. Furthermore, Veinott proposed a critical level rationing policy where instead of satisfying as much demand as possible from one customer class before satisfying demands from another class, demands from all the customer classes are satisfied until inventory level reaches certain level (rationing or critical level); thereafter only demand from higher priority class is satisfied and demand from lower priority is backlogged. However, the author did not analyze this critical level policy (or rationing policy) in the paper.

Topkis (1968) was the first to analyze the concept of inventory rationing as suggested by Veinott and showed how to allocate inventory among multiple classes by

rationing. The author assumed periodic review of inventory and that all the un-met demands were either entirely backlogged or entirely lost and incurred different penalty cost. The author modeled the scenario where in any given period the optimal rationing policy is such that demand from a given class is satisfied from the existing stock as long as there exists no unsatisfied demand from a higher class and the stock level does not drop below the critical level for that class. The author analyzed the scenario by breaking the review period into finite number of sub periods, and at the end of each sub period the decision is made whether to fill the demand from the lower priority class in the current sub period or reserve the inventory for the higher priority class for the next sub period. The author proved that there exists optimal, non negative rationing level for each demand class within a single review interval. The main limitation of this model is that there is only one replenishment opportunity at the beginning of the review period. The author later extended the model for multiple replenishment opportunities where inventory can be replenished at the start of each period. The author assumed that the backlogged demands at the end of the period are cleared by the next replenishment and no backlogged demands are carried forward to the next period. Using this assumption, the author concluded that the optimal rationing policy in each period is the same as described under a single period model. He also showed that the optimal rationing and procurement (ordering) policies may be determined myopically if the unsatisfied demand in each sub period is completely backlogged to the next sub period, and a base stock policy is optimal for ordering in each period. The main shortcoming of this model is that it does not allow demands to be backordered between periods and thus cannot be treated as a general case.

Evans (1968) and Kaplan (1969) independently derive the same results as Topkis (1968), but considered only two demand classes. Evans (1968) analyzed the same multi period model under the settings of Topkis (1968) and assumed that unsatisfied demand is lost. The author also showed that the optimal ordering policy is a base stock policy and optimal rationing policy is a critical level policy with single critical level (i.e. only demands from higher customer class are satisfied when inventory level drops below the critical level). Kaplan (1969) also considered similar scenario as Evans (1968) but considered that unsatisfied demands are backlogged. The author used the example of Army Material Command and, under certain assumptions, proved the existence of critical levels (reserve levels). Unlike Evans (1968) he confined his analysis to critical level policy only and showed the existence of optimal critical levels and proposed an algorithm to find them.

As can be seen, many researchers have shown interest in the area of inventory rationing and have tried to determine the policy parameters under different environments. As such the literature in this field can be categorized as follows:

1. *Number of customer classes*: Two demand classes (e.g. high and low, normal and emergency) or multiple classes (N classes).
2. *Treatment of the shortages*: Unsatisfied demands can be backordered (i.e. satisfied at a later date) or treated as lost sales (i.e. not satisfied at all but incurs penalty cost).
3. *Inventory review policy*: Periodic review policy (inventory is reviewed periodically e.g. weekly or monthly basis) or continuous review policy (inventory is reviewed regularly e.g. daily).

4. *Type of rationing policy*: Static rationing levels (static rationing policy) where the levels do not change with time, or dynamic rationing levels (dynamic rationing policy) where the rationing levels change with time, depending on certain parameters.

Key parameters in inventory rationing policies are ordering policy (when to order and how much to order) and the rationing level. These decision variables are dependent on the inventory review policy used. Most of the initial studies in inventory rationing considered periodic review, but recently some studies have considered continuous review. In the initial phase of rationing literature, researchers (Veinott, Topkis, Evans and Kaplan) considered periodic inventory review; the first contribution in a continuous review environment was made by Nahmias and Demmy (1981). In the next section, literature which considers service differentiation in continuous inventory review environment is discussed.

2.2.1 Continuous review systems

Nahmias and Demmy (1981) assumed a (Q, r) inventory replenishment policy with Poisson demand and two demand classes. Inventory is rationed according to a critical level policy as suggested by Veinott (1965). They termed the critical level as support level (K) and assumed it fixed; all the low priority demand is backordered while high priority demand continues to be filled. They also assumed that un-met demands are completely backordered, a fixed lead time and only one outstanding order. The last assumption implies that when an order is placed the inventory level and inventory position are identical. The main contribution of the research is the derivation of approximate expression for expected backorders for both the demand classes. First they

considered a single period model and derived the expressions for expected number of backorders for a periodic review system. Then they extended the results of the model to a multi-period model with zero lead time and presented an example with exponentially distributed demands. They also considered a continuous (Q, r) policy with Poisson demands and positive lead times and developed an approximate expressions for expected backorders. Lastly, they showed how these approximate results can be used to develop tables that will help to choose the various system parameters to achieve the required service level. The main problem with Nahmias and Demmy (1981)'s work is that the results are only approximate due to the assumption of one outstanding order.

Moon and Kang (1998) extended the result derived by Nahmias and Demmy (1981) for Compound Poisson process and derived the approximate expressions for fill rates for both the demand classes. They first extended the single period model with two demand classes to that of multiple demand classes (i.e. multiple support levels or critical levels). They assumed that all the demands occur simultaneously at the end of the period and obtained the expected backorders for each demand class. This assumption is not as reasonable as the assumption of demands occurring uniformly during the period, but it helps to compute the analytical solutions to the problem. They then presented a simulation model under the assumption that demands occur uniformly during the period. They also developed a single period cost minimization model analytically, assuming constant demand rate for each class and finally extended the model considering compound Poisson process demand. They presented a continuous review (Q, r) model with compound Poisson process.

Dekker et al. (1998) first considered the case of more than one outstanding order. They assumed Poisson demand process for the customer classes and lot-for-lot inventory policy i.e. every time a demand occurs, a replenishment order is placed and the order arrives after a non-zero lead time. They assumed two levels - base stock level and critical level. Every time inventory is depleted, an order is placed to raise the inventory up to the base stock level and when inventory level drops to critical level, inventory is reserved for higher priority class (they termed it as critical demand) and lower priority class (non-critical) demands get backordered until the inventory again exceeds critical level. The main contribution of the paper is the derivation of an approximation of the service level for both the demand classes.

Dekker et al. (2002) later considered an inventory problem with several demand classes prioritized according to their importance and $(S-1, S)$ inventory model with lost sales. They assumed that for each demand class, there is one critical level, and demand for that class is not satisfied if the inventory level is at or below the critical level and the unsatisfied demands are lost. The main contribution of this paper is the expressions for service levels for each demand class and the average cost per unit time. They assumed Poisson demand for each demand class, generally distributed lead time and presented an efficient method for determining the optimal critical level, lot-for-lot policy, with or without service level constraints. In addition to this, they also presented a heuristic approach for the problem without a service level constraint.

Melchioris et al. (2000) later considered the same lost sales case, but for two demand classes and in a continuous (Q, r) environment. Based on the approximate expressions derived by Nahmias and Demmy (1981), they derived the exact formulation

of average inventory cost. They also assumed that demands from both the classes follow Poisson distribution and deterministic lead time. Similar to Nahmias and Demmy (1981) they also assumed that there is at most one order outstanding but, unlike Nahmias and Demmy, they assumed that unsatisfied demands are lost; this helped them derive the exact expressions for the total cost. They also presented a simple optimization procedure based on enumeration and bounds.

The research most relevant to this research was by Deshpande et al. (2003) and Arslan et al. (2007). Deshpande et al. (2003) analyzed the continuous review (Q, r) policy with two demand classes. Similar to Nahmias and Demmy (1981) they also assumed that demands follow Poisson distribution and differ in delay and shortage cost and considered a static critical level policy to ration inventory, backordering the unfilled demands. The objective of the study was to optimize ordering and rationing parameters. But unlike Nahmias and Demmy (1981) they did not make the assumption of at most one outstanding order. Allowing more than one outstanding order generates a problem—how to fill backorders when replenishment arrives. To solve this problem they proposed a new clearing mechanism called ‘*threshold clearing*’ to fill backorders, rather than using priority clearing mechanism where higher priority backorders are cleared first. The basic idea of threshold clearing mechanism is to clear the backorders in the same manner they would have been filled if there were more inventory available at the time of the demand arrival. With this mechanism it is possible to compute the expected number of backorders for both the demand classes. Based on the expressions, they developed an algorithm to calculate the optimal ordering and rationing parameters defined as (Q, r, K) , K being the threshold or rationing level. Finally, they showed the effectiveness of their model over

other clearing mechanisms like separate stock and ‘round up’, numerically. The most important contribution of this work is development of a mechanism to fill the backorders when there is more than one order outstanding.

Arslan et al. (2007) considered the same setting as Deshpande et.al. (2003) but allowed more than 2 demand classes under the same ‘threshold’ clearing mechanism. They showed the equivalence between this inventory rationing system and a serial inventory system. Based on the equivalence they developed a model for cost evaluation and optimization, proposed an efficient heuristic and developed a bound on its performance. But they did not derive the cost expression for the system rather they presented this problem as a service-level problem where the objective is to minimize the total inventory subject to fill rate requirement for each customer class and provided an algorithm to solve the problem.

All the initial research in inventory rationing considered periodic review. In the next section, the literature which considered inventory rationing under periodic review environment is discussed.

2.2.2 Periodic review systems

Veinott (1965), Evans (1968), Topkis (1968) and Kaplan (1969) considered periodic review system for their analysis. Although continuous review is more realistic in today’s industries, some researchers considered periodic review environment also. Cohen et al. (1988) considered the lost sales scenario, but did so under periodic review (s, S) policy. They assumed deterministic lead time and two demand classes and used a simple priority mechanism to allocate inventory in a multi-echelon inventory system. They

considered two replenishment modes—normal and emergency, with different lead times. At the end of each period, direct customer and emergency shipment orders are satisfied first and normal replenishment orders are satisfied later. An approximate renewal - based model was derived and a greedy heuristic was developed to minimize the expected annual cost subject to a fill rate constraint. In this model critical level for inventory rationing was not used.

Katircioglu and Atkins (1996) considered a periodic review inventory system with several demand classes each with its minimum service level and a fixed positive lead-time. Like Cohen et al. (1988) they allowed unfilled demands to be fully backlogged to the next period. Since there is no one-to-one mapping between a model with service levels and a model with linear backordering costs, they assumed quadratic backordering cost. The objective of the study was to find a policy to allocate available inventory among the demand classes and an ordering policy for inventory replenishment to achieve pre-specified customer service levels with minimum inventory holding cost. They developed a heuristic allocation-order policy for both finite and infinite horizon cases. For infinite horizon, they verified the performance of the heuristic using simulation and found that the heuristic delivered service levels very close to the target and costs close to the lower bound.

Frank et al. (2003) considered periodic review inventory system with two types of customers: one with long-term service contract with the firm and another which requests goods only occasionally. Since the orders from the long-term customer are known in advance, they have to be filled completely without any delay; this demand can be assumed to be deterministic. The orders from the other class are unknown (and sporadic)

and can be assumed to be stochastic. The deterministic demand is satisfied immediately in each period, but stochastic demand can be rejected, and is treated as lost. Demands from the two classes are observed at the beginning of each period and decision is made whether to place an order and its size, as well as how much demand from stochastic class to fill. The model allowed rationing inventory for deterministic demand without satisfying the stochastic demand. They first categorized the structure of the optimal policy and showed that the optimal ordering and rationing policies are state dependent and do not have a simple structure. Then they proposed a simple (s, k, S) policy where s and S are reorder point and order up to level respectively, while k denotes the level of stochastic demand to satisfy. They assumed a fixed ordering cost and zero replenishment lead-time. This study is different from the previous studies in that the objective of this study is to avoid high fixed ordering cost rather than ration inventory for high priority demand.

The research in inventory rationing reviewed so far has considered stochastic demand only. The only study which considers deterministic demand is by Ding et al. (2007). They considered multiple demand classes and deterministic demand with holding, backorder, delay cost (backorder cost per unit per unit time) and setup costs. They also included price discount for demand that is willing to wait and be satisfied from the next cycle.

2.3 Make-to-order systems

Ha (1997a) first considered a make-to-order production system with a single production facility. So far all the studies had assumed that the inventory supply is

exogenous and there was no capacity restriction and did not deal with the effect of capacity on inventory policies and rationing policies. He considered this scenario for several demand classes assuming exponential production time and Poisson demand i.e. a single-product M/M/1 make-to-order queue. Any demand not satisfied immediately from the on hand inventory is assumed as lost. He showed that the optimal policy can be characterized by a sequence of monotone and stationary rationing levels. He modeled a production facility which at any given time chooses to produce or stop production and the production capacity can either satisfy the incoming demand or reject it. He analyzed this queuing model and found the operating costs of a rationing policy. He also showed that optimal policy is a static rationing policy. The main shortcoming of the model was that it only considered exponentially distributed production times which may not be realistic.

Ha (1997b) later extended his model by allowing backorder, but with only two demand classes. As in Ha (1997a) he assumed there is a single production facility that produces a single item with exponential production times. The demands from the two demand classes follow Poisson process and have different backordering costs. He formulated the problem as a queuing model and showed that the optimal rationing policy has a monotone switching curve structure. The structure of the optimal policy is induced by the trade-off in the production and rationing decisions. He suggested that control policies of similar form should perform reasonably well for more general systems. He also suggested that a policy with a base stock level and a fixed rationing level for each class could be a very useful future research direction.

Ha (2000) removed the limitation of Ha (1997a) assuming Erlangian production time. This assumption helped to model more realistic systems since rationing decisions

depend on the on hand inventory and the status of production facility. The work storage level (inventory level plus work in process) is used to capture the information regarding inventory level and status of current production. The optimal rationing policy can be characterized by a sequence of monotone critical work storage levels, and the optimal production policy can also be characterized by a critical work storage level.

Gayon et al. (2004) considered the same problem as Ha (2000) but assumed that unsatisfied demands are backordered. They also showed that optimal rationing policy is also a stationary critical work storage level policy, and that certain structural properties of optimal stock and capacity allocation policies exist when production may be interrupted and restarted. They presented a complete characteristic of the optimal policy in case of uninterrupted production and when extra production can be salvaged.

Vericourt et al. (2002) extended the work of Ha (1997b) by allowing more than two demand classes. Using dynamic programming, they showed that the optimal allocation policy has a simple and intuitive structure and optimal rationing policy is a static critical level policy. They also presented an efficient algorithm to compute the parameters of the optimal policy and illustrated that ignoring the stock allocation dimension can lead to incorrect decisions.

Since the work of Veinott (1965) which considers dynamic rationing, the research in this field has mainly considered static rationing policy. The primary reason for this is that static inventory rationing policy is easy to comprehend and simple to implement. However, in some of the cases, static rationing policy may not be optimal and dynamic rationing is best suited. In the next section, literature on dynamic rationing is discussed.

2.4 Dynamic inventory rationing

There exists another type of inventory rationing policies where rationing levels are not decided in advance, but they change with time and these are called dynamic rationing policies. In case of static rationing, the critical levels are always fixed and do not change with time, which makes them easier to understand and implement. But because of its dynamic nature, dynamic rationing policies are very difficult to analyze and are not easy to implement.

One of earliest paper on dynamic inventory rationing was by Topkis (1968). As discussed earlier, Topkis (1968) considered an inventory system in which demands are of ' n ' classes of varying importance. He assumed general demand processes and built a dynamic programming model to solve the problem. The model can be applied to different scenarios of shortages or unfilled demands –complete backordering, complete lost sales or partial backordering. He showed that optimal rationing policy was a dynamic critical level policy. Evans (1968) and Kaplan (1969) also showed results similar to Topkis (1968) but considered only two demand classes.

Melchoirs (2003) extended his previous research (Melchoirs et al. 2000) considering dynamic rationing. He assumed a continuous (Q, r) policy with Poisson demand and deterministic lead time. As in the previous study, he assumed that any unsatisfied demand is lost and that there is at most one order outstanding. Due to this assumption, his dynamic critical level policy is a restricted one, called restricted time-remembering policy. He developed expressions for the expected average cost, for fixed ordering and rationing policies. Based on empirical observations a neighborhood search heuristic was developed to find appropriate values for policy parameters.

As discussed at the beginning of the chapter, one of the applications of dynamic critical level policy is in the airline seat inventory allocation problem. The problem is of selling similar seats to different customers willing to pay different prices. This problem cannot be dealt with by a static critical level policy because of the obsolescence problem. What is needed are non-stationary control policies that dynamically adjust as time of inventory expiration (in this case, flight time) approaches. MacGill and Ryzin (1999) presented a research overview of airline revenue management in which seat control is the focus and give the prospects of research in this area.

One of the most relevant papers on the problem of seat booking in an airline industry using dynamic policy is by Lee and Hersh (1993). They assumed that demand from each class follows Poisson distribution and developed a discrete-time dynamic programming model to find an optimal booking policy. They also incorporated the issue of multiple bookings in the model. The model is almost the same as Topkis (1968); the only difference is that there is no holding cost associated with airline seats. Bitran and Mondschein (1995) also considered the same problem, but for the hotel industry. They developed optimal strategies as functions of capacity and the time left until the end of the planning horizon. Subramanian et al. (1999) and Zhao and Zheng (2001) have made valuable contributions in the area of dynamic allocation. In the next section, the literature that deals with the joint replenishment problems is presented.

2.5 Joint replenishment problem (JRP)

Many researchers considered joint replenishment problems where several different types of items are replenished in a single order and tried to develop inventory

polices to serve them. Service differentiation problem is similar to JRP; the only difference is that the items are same in case of service differentiation problem. Pantumsinchai (1992) evaluated the performance of a joint inventory policy for n items. He considered an inventory system where an order is placed when the total inventory position of all the ' n ' items drops to the group reorder point, to raise the inventory positions of all the items to their base stock levels. Under unit demand case, for an order quantity Q , reorder point is reached when the total usage of all the items becomes Q . He empirically compared this policy with two other policies: can (S, s, c) and periodic policies like (Q, S) policy. The can-order policy has a must-order point(s), a can-order point (c) and an order-up-to level (S). In periodic policy an item is replenished up to R , every T periods. He found that no one policy is superior to the other but the (Q, S) policy sometimes performs surprisingly better than the other. In the stochastic model developed in this research, it is assumed that items are replenished based on a (Q, S) policy. When the total usage for both the classes reaches Q , an order is placed to raise the inventory position of both the classes to their base stock.

Viswanathan (1997) proposed a new policy, namely $P(s, S)$ policy, which uses independent, periodic review (s, S) policies for each item. He showed computationally that his policy is superior to other classes of policies. He also suggested that this policy can be improved by monitoring the total consumption continuously and reviewing the individual items when total consumption reaches or exceeds the order quantity, Q . For each individual item same (s, S) policy is followed. He also predicted that this policy should perform better than (Q, S) policy but computing the policy parameters is hard.

Research related to algorithms for computing optimal parameters for both continuous review and periodic review inventory systems are also reviewed. The most useful and efficient algorithm for determining the optimal parameters in (s, S) policy was proposed by Zheng and Federgruen (1991). They proposed a new algorithm which is simple and easy to understand and the computation time is far less than the other algorithms. Their algorithm is applicable for both periodic and continuous review systems. Federgruen and Zheng (1992) made the most important contribution for continuous review (Q, r) policy by developing an efficient algorithm to compute optimal parameters. The computational complexity of the algorithm is linear in Q . Based on the unimodality of the loss function they presented the algorithm to identify the reorder point and order quantity, Q . Deshpande et.al (2003) later used this algorithm to derive the optimal parameters for their rationing policy.

2. 6 Research gap

In this section, the research gaps in the literature are presented along with the research plan. From the study of the available literature, it is evident that the problem of allocating inventory among different customer classes with different service needs has been the focus of research for many years. Motivated by the usefulness and ease of implementation of the static rationing policy, static rationing is the focus in this research also.

This research is divided into three parts. The first part considers static inventory rationing for deterministic demand for ' n ' customer classes. The second part focuses on stochastic demand (Poisson) for two customer classes and a new type of rationing policy

is proposed to satisfy the demand from the classes. The third part considers inventory rationing for ' n ' classes facing stochastic demand.

A review of the literature revealed that almost no research has considered deterministic and known demand from the customer classes. The only research which considers deterministic demand (Ding et al. (2007)) did not consider a general penalty cost structure. So in this research a static inventory rationing policy for ' n ' customer classes in a continuous review environment is considered first. It is assumed that the demands from the customer classes are deterministic and constant; the lead time is also assumed to be fixed. The rejected demands are backordered, and customer classes are prioritized according to the shortage or penalty cost they incur if demand is not fulfilled from stock immediately. All the backordered demands are satisfied when next replenishment arrives. Although stochastic demand is more realistic, deterministic demand can be used when there is little or no variation over the planning horizon, and can lead to useful insights for stochastic demand case.

Most of the researchers consider only the penalty cost of delay (i.e., penalty cost per unit per unit time) for backorders, whereas in this research a general penalty cost for backorders is considered, which includes delay cost and stock out cost (per unit). This general penalty cost more precisely represents the real-life scenario. Whenever an aircraft is grounded because of the unavailability of a spare part, it not only incurs a one-time penalty cost (stock out cost), but also incurs another penalty cost (delay cost) based on how long it is grounded. In airline industry this cost can be thousands of dollars per hour. To the best of our knowledge this penalty cost structure has not been considered in the literature for deterministic demand.

The key contribution in this part would be the development of an algorithm to find the optimal cycle time, the optimal order quantity and the rationing levels for ' n ' customer classes. The effectiveness of the algorithm would be shown with the help of numerical examples.

From the insights from the deterministic demand model, a continuous review policy with non-zero lead time is considered with the assumption that the demand from different customer classes follows Poisson process. Deshpande et al. (2003) derived an exact expression for the total cost under the same assumptions but for two classes. They also developed an algorithm, based on Federgruen and Zheng (1992) to derive the optimal parameters of the policy. The third part of this research would consider inventory rationing policy when there are more than two customer classes facing stochastic demand. No research has derived the exact cost expression for more than two classes. The exact cost expression for ' n ' classes facing stochastic demand and with general penalty cost would be developed in this part of the research. It would also be shown, computationally, the effect of segregating the customers into many different classes than into only few classes.

Though inventory rationing provides the lowest inventory cost for service differentiated problem, the main drawback of this policy is that it reserves inventory in anticipation of higher class demand arrival. But if the demand from the higher priority class is less than the inventory reserved for it, demand from the lower priority class will be backordered in spite of having inventory on hand and will incur some penalty cost. This situation is more likely to happen when demand from the lower priority class arrives

late in the cycle i.e. most of the lower priority class demand arrives when inventory is already reserved for higher priority class and cannot be satisfied.

To solve this problem, a new policy is proposed where two separate bins are maintained for the two classes. When demand comes from a particular class it is satisfied from the bin designated for that class. When the stock for the higher priority class runs out it can use the stock designated for the lower priority class. But if the stock for lower priority class runs out it cannot use the higher priority class' stock and has to be backordered. By allowing it to use the lower class' stock, the higher priority class is getting more priority than the lower priority class. At the same time some inventory is kept for the lower priority class demand arrival as well. The total stock for both the classes can be used by the higher priority class whereas the lower priority class only has access to its own stock. Thus, the new policy would increase the service level of the lower priority class without compromising the service level of higher priority class.

A model is developed for this 2-bin policy that allows higher class to use lower class' stock; the expressions for the total cost are derived for the two classes. Although the policy can be used for more than two classes where the higher class can use all the lower class' stock starting from lowest class, the expressions will be very complicated. This derivation of cost expression for more than two customer classes is the future research direction. Like Deshpande et al. (2003) it is assumed that demand is backordered when it cannot be satisfied from the stocks and a general penalty cost consisting of delay cost and stock out cost is considered. Numerical experiments are carried out to show the effectiveness of the proposed policy over traditional inventory rationing policy. A

comparative study on the optimal cost and service level for the two classes is also conducted between the proposed 2-bin policy and Deshpande's policy.

Continuous review inventory policy is considered in this research as, with the advent of computers, it is very easy to review the inventory status continuously and almost all the firms follow continuous review policy; as such continuous review environment is more realistic. Non- zero lead time is considered because in reality lead time is always greater than zero in case of spare part inventory. Assuming zero lead time limits the application of the research to some specific cases where spare parts arrive instantly. The demand arrival is assumed to follow Poisson process as it most appropriately depicts the random arrival of spare part demand.

2.7 Summary

In this chapter the literature related to inventory problems where customers are segregated based on their service requirement is studied in detail. The literature on inventory rationing is reviewed extensively. Also literature related to joint replenishment problem (JRP) is reviewed. The literature on algorithms to determine the optimal parameters in different inventory settings is also studied. From the review of literature, the research gaps are identified and research plan for this research is presented. In the next chapter, inventory rationing model for ' n ' customer classes facing deterministic demand will be considered under continuous review and general cost structure.

Chapter 3

Inventory rationing with deterministic demand

3.1 Introduction

As discussed in Chapter 2 there is abundant literature on static rationing policies but most of it considers inventory policies for two customer classes differentiated by service requirement or shortage cost. Researchers have developed algorithms, but no one has developed exact expressions for the parameters of optimal rationing policies facing stochastic demand. Only one study has, so far, considered deterministic demand. Ding et al. (2007) considered multiple demand classes and deterministic demand with holding, backorder, lost demand (sales) and setup costs. They considered delay cost (backorder cost per unit per unit time) but did not consider stock out cost (fixed penalty cost per unit backordered). The customer classes are differentiated based on the price they pay and their willingness to incur delay in fulfillment of their demand. When demand arrives the firm either fulfills the demand or offers a price discount to the customer for waiting till

the next replenishment order. Thus their study is basically related to dynamic pricing for differentiated customer classes. They also showed that dynamic pricing is a combination of a pricing and allocation mechanism.

Frank et al. (2003) assumed two customer classes, one being deterministic and known and the other being stochastic. The deterministic demand is satisfied immediately whereas the stochastic demand can be rejected and assumed as lost if the amount reserved for stochastic demand runs out. They proposed a simple (s, k, S) policy where s , k and S are reorder point, amount of stochastic demand to fill and order quantity, respectively. However, the objective of this study was to minimize the ordering cost rather than reserve inventory for the higher priority class.

The deterministic demand scenario can occur when the customers have different service contracts with the firm. A customer can choose the contract based on the service level it requires. It can choose a contract with lower price and a lower service level or a contract with higher price and higher service level. This price also determines the penalty cost incurred if demand from the customer with a particular service contract is not satisfied immediately. Rejection of demand from a customer with higher price contract is more costly than one from a customer with lower price contract. If the firm has a long-term contract (as in Frank et al. (2003)) with the customer, the orders from the class are known in advance and can be assumed as deterministic. Moreover, for some classes, when there is not much variation in demand across the planning horizon, demand from customer classes can be assumed to be deterministic.

Unlike Ding et al. (2007), a general penalty cost structure is considered in this research. Most of the researchers consider only the penalty cost of delay (i.e., penalty cost

per unit per unit time) for backorders, whereas a general penalty cost for backorders also includes fixed stock out cost (per unit). Although the general penalty cost more precisely represents the practical cases, it is more difficult to analyze. Whenever an aircraft is grounded because of the unavailability of a spare part, it not only incurs a one-time penalty cost (stock out cost), but also incurs another penalty cost (delay cost) based on how long it is grounded. In airline industry this stock-out cost can be a few thousand dollars, and may be much higher than the delay-cost. To the best of our knowledge this penalty cost structure has not been considered in the literature for deterministic demand. Secondly, unlike Frank et al. (2003) the objective of this research is to minimize the total inventory cost, which consists of ordering, inventory carrying and penalty costs.

The cost structure considered in this chapter closely resembles the one followed at the after-market service (AMS) unit of a leading automobile manufacturer that was studied for this research. The AMS unit also faces different order classes and segregates them as regular, emergency and VOR (vehicle off-road) orders, VOR being the highest priority order class. They satisfy the regular demands from the inventory until it falls below safety stock. After that the regular demands are backordered, but emergency and VOR orders are satisfied as long as inventory is available. For VOR orders parts can even be shipped from the global DC if the inventory in the regional DC runs out. The company also follows continuous review reorder point policy. Placing an order incurs a nominal ordering cost. In Singapore, it rents a warehouse managed by a 3PL. It was found that it uses different mode of transport for different type of orders with different cost. Regular orders are replenished by sea which takes about seven days to reach the customer. For emergency orders, air-shipment is used which takes just over a day. The associated cost

of air shipment is approximately ten times that of sea freight. For VOR, it uses the next available flight to ship the item from the regional DC to the customer; which is costlier than regular airfreight. These additional costs closely resemble the penalty cost structure considered in our research. It was found that regular and emergency orders do not have any delay cost (per unit per time); these orders only have stock out cost (per unit) which is basically the difference between the shipping cost for the emergency order and the shipping cost for regular orders; for regular orders, it can be assumed that the penalty cost is zero. VOR orders, however, have both delay cost as well as stock out cost. Stock out cost is the difference between the cost of shipping by next flight and the regular shipping charges; this extra cost of shipping over the regular sea freight can be treated as the penalty cost for that order class. The delay cost rate is based on the time weighted contracted charges incurred when the vehicle is off road and out of service (represented by the cost of replacement or rental charges).

In this chapter, a static inventory rationing policy for ' n ' customer classes facing deterministic demand in a continuous review environment is considered. The customers are prioritized based on the penalty cost they incur if demand is not fulfilled upon arrival. It is assumed that the demand rates from the customer classes are deterministic and constant. The lead time for replenishment is assumed to be fixed. The demand from a particular class is satisfied up to a certain time (its run-out time), any demand arriving after the run-out time is reached is backordered to the next cycle. It is assumed that all the backorders are cleared when the next replenishment arrives.

The objective of the study is to determine the optimal parameters of the rationing policy, namely, order quantity (Q), rationing co-efficient (α_i) (defined as the fraction of

cycle time up to which demand from a class i is satisfied) by minimizing the average total cost per unit time. The key contribution of this chapter is the development of an algorithm to find the optimal cycle time, the optimal order quantity and the rationing levels for ' n ' customer classes.

The remainder of this chapter is organized as follows. In section 3.2, the model and the exact cost expression for inventory rationing for ' n ' customer classes with deterministic demand is developed. Then an algorithm is developed to determine the optimal cycle time and the rationing levels. In section 3.3, empirical examples are presented and results of a numerical study are discussed. Finally a summary and presented in section 3.4.

3.2 Model formulation

The notation used throughout this chapter is summarized in Table 3.1. A single facility inventory system is considered where demands from ' n ' customer classes ($i=1, 2, 3, \dots, n$) are satisfied from inventory. Holding costs are incurred at a rate of h per unit of inventory per unit time. Each replenishment order from the supplier incurs a fixed cost A . Replenishments for a fixed order quantity is placed when inventory level reaches the reorder point. Replenishments arrive after a fixed and constant lead time. Demand that is not fulfilled immediately is backordered and satisfied upon arrival of the next replenishment. Customer classes are differentiated based on the penalty cost incurred if the demand from that class is not fulfilled immediately from stock. It is assumed that the backordered demand incurs two types of penalty costs—a fixed stock-out cost per unit (π_i)

and a time-weighted delay cost incurred at the rate of p_i . It is also assumed that $p_1 \geq p_2 \geq p_3 \dots \geq p_n$ and $\pi_1 \geq \pi_2 \geq \pi_3 \geq \dots \geq \pi_n$. Class 1 is considered as 'highest priority' class and class n as 'lowest priority' class.

Table 3.1 Notations used

Symbols	Meaning
i	Class index
n	Number of classes
λ_i	Demand rate for class i
λ	Total demand rate of all the classes
p_i	Time-weighted delay cost per unit for class i
π_i	Fixed stock-out cost per unit for class i
h	Holding cost per unit per unit time.
A	Fixed ordering cost per order
L	Lead time for replenishment arrival
T	Cycle time
α_i	Rationing co-efficient for class i
Q	Order quantity
r	Reorder point

The demand arrivals are deterministic and for each customer class, i , the demand occur at a constant rate of λ_i . Without loss of generality, it can be assumed that either p_i or

$\pi_i > 0$, for all i . Otherwise if both $p_i = \pi_i = 0$, the fulfillment of demand can be delayed indefinitely. The inventory rationing policy works as follows: the inventory is replenished according to a continuous review (Q, r) policy which means that whenever inventory level reaches reorder point r , an order of quantity Q is placed which arrives after a lead time L . After the replenishment arrives, and the previous backorders are fulfilled, the inventory level at the commencement of the cycle is I_0 . From I_0 until level K_n all the demand from the n classes are satisfied immediately. Below this inventory level, demand from class n is backordered. When the inventory level decreases further to a level K_{n-1} , demand from both class n and class $n-1$ are backordered for the rest of the cycle. In general, demand from class i (along with class $i+1, \dots, n$) are backordered for inventory level $I \leq K_i$. As the demands for all the classes occur at a constant rate, the threshold levels, $K_n, \dots, K_1$, can be easily converted to specific times in the inventory cycle. Let the total cycle time be T and let $t_i = \alpha_i T$ be the time at which the inventory level in the cycle reaches level K_i . For each class $i = 1, 2, \dots, n$, the total inventory required in the cycle is $t_i \lambda_i$. Therefore, $K_i = \sum_{j=1}^{i-1} \lambda_j (t_j - t_i)$ as only demand from classes $1, \dots, i-1$ is fulfilled once inventory level drops below K_i . α_i ($0 \leq \alpha_i \leq 1$) is denoted as the fraction of cycle time (T) over which demands from customer class ' i ' are satisfied and called the rationing co-efficient for class i . Figure 3.1 illustrates a typical replenishment cycle with inventory rationing.

Figure 3.1: Typical replenishment cycle with inventory rationing

Therefore,
$$I_0 = \sum_{j=1}^n \lambda_j t_j$$

By definition, $K_I=0$. So, the value of the rationing levels can be given by

$$K_i = \begin{cases} \sum_{j=1}^{i-1} \lambda_j (t_j - t_i) & 1 < i \leq n \\ 0 & i=1 \end{cases} \quad (3.1)$$

On hand inventory depletes at the rate of $\sum_1^n \lambda_i$ until time t_n , because demand from all the classes are satisfied until t_n ; then at the rate of $\sum_1^{n-1} \lambda_i$ until time t_{n-1} ; and finally at the rate of λ_1 (after time t_2) until no inventory is left or the next replenishment arrives. We denote the total number of backorders, accumulated at the end of the cycle, for the i^{th} class as B_i . Clearly,

$$\begin{aligned} B_i &= \lambda_i (T - t_i) \\ &= (1 - \alpha_i) \lambda_i T \end{aligned}$$

The order quantity, $Q = \sum_{i=1}^n \lambda_i T$

since all the demands is fulfilled.

As already stated

$$\begin{aligned} I_0 &= \sum_{i=1}^n \lambda_i t_i \\ &= \sum_{i=1}^n \lambda_i \alpha_i T \\ &= Q - \sum_{i=1}^n B_i \end{aligned}$$

The objective is to determine the policy parameters: cycle time T (or, equivalently the order quantity Q) and rationing co-efficients for the classes, α_i 's which minimize the average inventory cost.

The problem can be stated as:

$$\text{Min } TC(T, \alpha_i) = O(T, \alpha_i) + H(T, \alpha_i) + P(T, \alpha_i),$$

subject to:

$$0 \leq \alpha_i \leq 1$$

$$T \geq 0.$$

where, for a given (T, α_i) policy

$TC(T, \alpha_i)$ = total average inventory cost per unit time,

$O(T, \alpha_i)$ = total average ordering cost per unit time,

$H(T, \alpha_i)$ = total average holding cost per unit time, and

$P(T, \alpha_i)$ = total average penalty cost per unit time.

3.2.1 Cost expressions

The holding cost over a cycle can be determined by multiplying the unit holding cost with the area under the inventory-time curve (Figure 3.1). The total area is the summation of all the trapezoids (A, B,...) and the triangle X.

Area of the trapezoid A can be written as $\frac{1}{2} \left(I_0 + \left(I_0 - \sum_1^N \lambda_i \alpha_N T \right) \right) \alpha_N T$

Similarly, Area of the trapezoid B =

$$\frac{1}{2} \left(\left(I_0 - \sum_1^N \lambda_i \alpha_N T \right) + \left(I_0 - \sum_1^N \lambda_i \alpha_N T - \sum_1^{N-1} \lambda_i (\alpha_{N-1} - \alpha_N) T \right) \right) (\alpha_{N-1} - \alpha_N) T$$

Other trapezoid areas can be determined similarly.

The area of the triangle X = $\frac{1}{2} \lambda_1 (\alpha_1 - \alpha_2)^2 T^2$

Adding the expressions for all the areas and after simplification, total area can be written

as $\frac{1}{2} \sum_1^n \lambda_i \alpha_i^2 T^2$; therefore the holding cost for the duration of the cycle, T , is given by

$$\frac{h}{2} \sum_1^n \lambda_i \alpha_i^2 T^2$$

This is the holding cost over an order cycle of duration T . Therefore, the average holding

$$\text{cost per unit time, } H(T, \alpha_i) = \frac{\frac{h}{2} \sum_1^n \lambda_i \alpha_i^2 T^2}{T} = \frac{h}{2} \sum_1^n \lambda_i \alpha_i^2 T \quad (3.2)$$

The expression (3.2) can also be derived by considering inventory for class i , as a

separate item. Clearly, $I_0 = \sum_{i=1}^n \lambda_i t_i$, and for each class i , the inventory $\lambda_i t_i$ is expended

over time $\alpha_i t_i$. Hence, the average inventory over the duration of the cycle for class i is

$$\frac{1}{2} \frac{\lambda_i \alpha_i^2 t_i}{T} \text{ which leads to expression (3.2).}$$

As already mentioned,

$$B_i = (1 - \alpha_i) T \lambda_i \quad i=1, 2, \dots, n \quad (3.3)$$

Since demand occurs at a continuous constant rate, the average number of

backorders is $\frac{B_i}{2}$, which occur over time period $(1 - \alpha_i) T$.

Therefore, the delay cost per cycle for class i is

$$\frac{1}{2} B_i (1 - \alpha_i) T \times p_i$$

The stock-out cost per cycle for class i is $B_i \times \pi_i$

Total penalty cost for class ' i ' for an order cycle can be written as

$$P_i = \left[\frac{1}{2} B_i (1 - \alpha_i) T \times p_i \right] + (B_i \times \pi_i) \quad (3.4)$$

Substituting (3.2) into (3.3), summing over all classes i , and dividing by the cycle time T , the average penalty cost per unit time can be obtained as

$$P(T, \alpha_i) = \frac{1}{2} \sum_1^n (1 - \alpha_i)^2 \lambda_i T p_i + \sum_1^n (1 - \alpha_i) \lambda_i \pi_i \quad (3.5)$$

There is only one replenishment order during each cycle and as such the ordering cost per unit time is

$$O(T, \alpha_i) = A/T \quad (3.6)$$

Now using equations (3.2), (3.5) and (3.6), the total average inventory cost per unit time can be written as

$$\begin{aligned} TC(T, \alpha_i) &= (T, \alpha_i) + H(T, \alpha_i) + OP(T, \alpha_i) \\ &= \frac{A}{T} + \frac{h}{2} \sum_1^n \lambda_i \alpha_i^2 T + \frac{1}{2} \sum_1^n (1 - \alpha_i)^2 \lambda_i T p_i + \sum_1^n (1 - \alpha_i) \lambda_i \pi_i \end{aligned} \quad (3.7)$$

3.3 Model analysis and optimization

The optimal values of the parameters can be derived by minimizing the total cost per unit time. Differentiating $TC(T, \alpha_i)$ with respect to α_i , and using the first order condition, we get

$$\frac{\partial TC(T, \alpha_i)}{\partial \alpha_i} = h\alpha_i \lambda_i T - (1 - \alpha_i) \lambda_i p_i T - \lambda_i \pi_i = 0 \text{ for } \forall i \quad (3.8)$$

The second derivative of TC with respect to α_i ,

$$\frac{\partial^2 TC(T, \alpha_i)}{\partial \alpha_i^2} = \lambda_i T(h + p_i) > 0$$

Hence, the total cost function is convex with respect to α_i for a fixed T .

At $\frac{LT}{\alpha_i \rightarrow 0^+}$, the first derivative is negative as either p_i or π_i strictly > 0 . This implies that rationing co-efficient, $\alpha_i > 0$.

Solving equation (3.7) and considering the fact that $\alpha_i \leq 1$, we get

$$\alpha_i = \begin{cases} \frac{p_i + \pi_i / T}{h + p_i} & \text{when } hT \geq \pi_i \\ 1 & \text{otherwise} \end{cases} \quad (3.9a)$$

When $hT < \pi_i$, the value of α_i satisfying equation (3.7) becomes greater than 1. Since the value of α_i cannot be greater than 1 ($0 \leq \alpha_i \leq 1$), combined with the fact that the cost function is convex with respect to α_i , its value can only be 1. So, in this case, the optimal $\alpha_i = 1$.

Let all the customer classes be divided into two sets: the set M comprising of the classes that satisfy the condition $hT \geq \pi_i$, and the other comprising of the classes that do not, i.e. define $M = \{i; hT \geq \pi_i\}$. The set M consists of the customer classes for which demand is partly backordered during the replenishment cycle. For the customer classes $i \notin M$, $\alpha_i = 1$, and all demand is fulfilled immediately upon arrival.

Equation (3.9), can then be written as

$$\alpha_i = \begin{cases} \frac{p_i + \pi_i / T}{h + p_i} & i \in M \\ 1 & i \notin M \end{cases} \quad (3.9b)$$

Using the values of α_i from equation (3.9b) in equation (3.7), the expression for total average inventory cost per unit time becomes

$$\begin{aligned} \text{TC}(T) = & \frac{A}{T} + \sum_{j \in M} \frac{hT}{2} \lambda_j \left(\frac{p_j + \pi_j / T}{h + p_j} \right)^2 + \sum_{j \in M} \frac{T}{2} \lambda_j p_j \left(\frac{h - \pi_j / T}{h + p_j} \right)^2 + \\ & \sum_{j \in M} \lambda_j \pi_j \left(\frac{h - \pi_j / T}{h + p_j} \right) + \sum_{j \notin M} \frac{hT}{2} \lambda_j \end{aligned} \quad (3.10)$$

After further manipulation, equation (3.10) can be written as

$$\begin{aligned} \text{TC}(T) = & \frac{A}{T} + T \sum_{j \in M} \frac{h \lambda_j p_j}{2(h + p_j)} - \frac{1}{T} \sum_{j \in M} \frac{\lambda_j \pi_j^2}{2(h + p_j)} + \sum_{j \in M} \frac{h \lambda_j \pi_j}{h + p_j} + T \sum_{j \in M} \frac{h \lambda_j}{2} \\ = & \frac{1}{T} \left[A - \sum_{j \in M} \frac{\lambda_j \pi_j^2}{2(h + p_j)} \right] + T \left[\sum_{j \in M} \frac{h \lambda_j p_j}{2(h + p_j)} + \sum_{j \notin M} \frac{h \lambda_j}{2} \right] + \sum_{j \in M} \frac{h \lambda_j \pi_j}{h + p_j} \end{aligned} \quad (3.11)$$

The expression for average inventory cost per unit time (3.11) is of the form $\frac{X}{T} + YT + C$

$$\begin{aligned} \text{where} \quad X = & A - \sum_{j \in M} \frac{\lambda_j \pi_j^2}{2(h + p_j)} \\ Y = & \sum_{j \in M} \frac{h \lambda_j p_j}{2(h + p_j)} + \sum_{j \notin M} \frac{h \lambda_j}{2} \\ C = & \sum_{j \in M} \frac{h \lambda_j \pi_j}{h + p_j} \end{aligned}$$

It is evident that $Y > 0$.

In case $X > 0$, the cost function (equation 3.11) is convex with respect to T and optimal T can be determined by solving the first order condition.

$$\frac{\partial ATC(T)}{\partial T} = -\frac{\left[A - \sum_{j \in M} \frac{\lambda_j \pi_j^2}{2(h + p_j)}\right]}{T^2} + \left[\sum_{j \in M} \frac{h\lambda_j p_j}{2(h + p_j)} + \sum_{j \notin M} \frac{h\lambda_j}{2}\right] = 0$$

This gives

$$T = \sqrt{\frac{\left[A - \sum_{j \in M} \frac{\lambda_j \pi_j^2}{2(h + p_j)}\right]}{\sum_{j \in M} \frac{h\lambda_j p_j}{2(h + p_j)} + \sum_{j \notin M} \frac{h\lambda_j}{2}}} \quad (3.12)$$

However, when $X < 0$, the total cost is minimized by having $T \approx 0^+$. But, when $T \approx 0^+$, $hT \approx 0$ and as a result $hT \leq \pi_i$ for all i , π_i because $\pi_i \geq 0$. Then no customer class will satisfy the condition $hT \geq \pi_i$ and correspondingly $a_i = 1$ for all i . Therefore, $M = \{\emptyset\}$, but in this case the second term of X vanishes and X becomes equal to $K \geq 0$. Therefore, in the optimal solution X cannot be less than 0.

3. 4 Algorithm to determine T & α_i

Next an algorithm is presented to determine the optimal cycle time, T and rationing co-efficients, α_i which minimize the average total cost, TC . As the cycle time and the rationing co-efficients depend on the members of the set M , they cannot be determined in a straight-forward manner from equation (3.12). Initially it is assumed that all the classes are to be rationed i.e. $M = \{1, 2, \dots, n\}$ meaning that all the classes satisfy the

condition $hT \geq \pi_i$. Then the cycle time T is computed and if, based on the computed value of T , it is found that the initial assumption (all the classes are rationed) is correct, the algorithm stops and the computed T becomes the optimal cycle time T^* . Otherwise, the highest priority class is excluded from the set M and the cycle time is recalculated. If the new T satisfies the condition $hT \geq \pi_i \forall i \in M = \{2, \dots, n\}$ then it is optimal. Otherwise, again the next higher priority class is excluded and the cycle time recalculated. The algorithm stops only when T satisfies the condition $hT \geq \pi_i$ for $\forall i \in M$. Based on this cycle time T^* , the values of run-out times can then be computed from equation (3.9b). Note that when item i , with the highest π_i value is removed from the current set M , the resulting value of T increases (see equation 3.12). However the new value of T will still be less than the π_i/h value of all the items excluded from set M , i.e. for $i \notin M$ the condition $hT \geq \pi_i$ will always be satisfied by the procedure followed in the algorithm. This is proved below.

Theorem 3.1: Let $M = \{1, 2, \dots, n\}$ and if the highest priority customer class, i is removed from set M , the value of T increases but cannot be greater than π_i/h . i.e. $T_{old} < T_{new} < \pi_i/h$ where i is the class that has been removed from M .

Proof: For $M = \{i, i+1, i+2, \dots, n\}$

$$\begin{aligned}
 T_{old}^2 &= \frac{K - \sum_{j=i}^n \frac{\lambda_j \pi_j^2}{2(h + p_j)}}{\sum_{j=i}^n \frac{h \lambda_j p_j}{2(h + p_j)} + \sum_{j=1}^{i-1} \frac{h \lambda_j}{2}} \\
 &= \frac{V}{W}
 \end{aligned} \tag{3.13}$$

where $V = A - \sum_{j=i}^n \frac{\lambda_j \pi_j^2}{2(h + p_j)}$ and $W = \sum_{j=i}^n \frac{h \lambda_j p_j}{2(h + p_j)} + \sum_{j=1}^{i-1} \frac{h \lambda_j}{2}$

If element i is excluded from set M i.e. $M = \{i+1, i+2, \dots, n\}$, T_{old} becomes T_{new} and

$$\begin{aligned}
 T_{new}^2 &= \frac{A - \sum_{j=i+1}^n \frac{\lambda_j \pi_j^2}{2(h + p_j)}}{\sum_{j=i+1}^n \frac{h \lambda_j p_j}{2(h + p_j)} + \sum_{j=1}^i \frac{h \lambda_j}{2}} \\
 &= \frac{A - \sum_{j=i}^n \frac{\lambda_j \pi_j^2}{2(h + p_j)} + \frac{\lambda_i \pi_i^2}{2(h + p_i)}}{\sum_{j=i}^n \frac{h \lambda_j p_j}{2(h + p_j)} + \sum_{j=1}^{i-1} \frac{h \lambda_j}{2} - \frac{h \lambda_i p_i}{2(h + p_i)} + \frac{h \lambda_i}{2}} \\
 &= \frac{V + \frac{\lambda_i \pi_i^2}{2(h + p_i)}}{W - \frac{h \lambda_i p_i}{2(h + p_i)} + \frac{h \lambda_i}{2}}
 \end{aligned}$$

Let $C = \frac{\lambda_i \pi_i^2}{2(h + p_i)}$ and $D = -\frac{h \lambda_i p_i}{2(h + p_i)} + \frac{h \lambda_i}{2}$ (3.14)

Therefore, $T_{new}^2 = \frac{V + C}{W + D}$ (3.15)

Using (3.14) and simplifying further, $\frac{C}{D} = \frac{\pi_i^2}{h^2}$ (3.16)

The element ' i ' was excluded from the set M because it did not satisfy the condition to be

in set M i.e. $\pi_i > h T_{old}$

or $\pi_i^2 > h^2 T_{old}^2$ or $T_{old}^2 < \frac{\pi_i^2}{h^2}$ (3.17)

From (3.14) and (3.13), using (3.17), $T_{old}^2 < \frac{C}{D}$ or $\frac{V}{W} < \frac{C}{D}$

If $\frac{V}{W} < \frac{C}{D}$, then a basic rule algebra is

$$\frac{V}{W} < \frac{V+C}{W+D} < \frac{C}{D} \quad (3.18)$$

Using (3.13), (3.15), (3.16) and (3.18)

$$T_{old}^2 < T_{new}^2 < \frac{\pi_i^2}{h^2} \quad \square$$

Theorem 3.1 ensures that the classes that are removed from set M in the algorithm need not be considered again for inclusion back into the set.

A flowchart summarizing the algorithm is presented in Figure 3.2.

The algorithm gives the optimal value of the cycle time T^* and the optimal values of rationing co-efficients, α_i^* . The following parameters can also be determined using the T^* and α_i^* .

Figure 3.2: Algorithm to determine the optimal cycle time, T^*

The optimal order quantity is

$$Q^* = \sum_1^n \lambda_i T^* = \lambda T^*$$

$$\text{where } \lambda \text{ is the total demand rate} = \sum_{j=1}^n \lambda_j$$

The run-out time for customer class i , $t_i^* = \alpha_i^* T^*$

The amount of i^{th} class demand that is backordered in a cycle

$$B_i = (T^* - t_i^*) \lambda_i$$

Remarks

The threshold inventory levels $K_1, K_2, \dots, K_n$ can be obtained by substituting optimal t_i^* , $i = 1, \dots, n$ into the equation (3.1).

$$K_i = K_{i-1} + (\alpha_{i-1} - \alpha_i) T^* \sum_1^{i-1} \lambda_j \quad \forall \quad i = 2 \text{ to } n$$

and $K_1 = 0$.

The reorder point, r , depends on the lead time L for replenishment. If $L = T$, then

$$\text{Reorder point, } r = I_0$$

Otherwise, ' r ' can be calculated appropriately.

3.5 Numerical examples

Next, some numerical results obtained using the algorithm are discussed. The algorithm was coded in 'C' programming language and the program was run using different combinations of parameters.

The rationing model was solved for three different scenarios: one for three customer classes, second for four and the third for two customer classes. The optimal cycle time and the rationing co-efficients were calculated using the algorithm of section 3.4. The run-out time (t_i) of each class was also calculated using (3.18). The parameters for critical level (Q, r, K) policy, namely order quantity, Q , reorder point, r and rationing levels, K_i were then calculated. The problem parameters as well as the results are listed in Table 3.2.

In the first example, class 1 demand is always fulfilled on arrival and demand from other two classes is rationed i.e. backordered for part of the cycle. In the second example the demand from the top three classes are always fulfilled on arrival; only demand from the lowest priority class is rationed.

Table 3.2: Optimal parameters for (T, α) and (Q, r, K) policies

Parameter	Example 1	Example 2
No of Customer Classes(i)	3	4
Demand rate (λ_i)	2,2,2	2,2,2,2
Delay Cost (p_i)	3,2,1	5,3,2,1
Stockout Cost (π_i)	10,8,5	15,10,8,5
Holding cost (h)	1.0	1.0
Ordering Cost (K)	200	200
Optimal Cycle Time (T*)	8.76	7.32
Rationing co-efficient (α_i^*)	1, 0.97, 0.78	1,1,1,0.84
Run-out time (t_i^*)	8.76, 8.51, 6.88	7.32, 7.32, 7.32, 6.16
Order Quantity (Q*)	52.54	58.55
No of Backorders (B _i)	0, 0.5, 3.76	0,0,0, 2.32
Critical Level (K _i)	0, 0.51, 7.03	0,0,0, 6.94

The threshold inventory levels for rationing are also provided in Table 3.2. When demand from a particular class is always fulfilled without rationing, the corresponding critical inventory level, $K_i = 0$.

The effect of different parameters on the system cost is studied using an example with two demand classes. First the demand rate of class 2 is changed, keeping the total demand fixed at 8 and effect on the total cost and the number of backorders studied.

Figure 3.3: Effect of varying demand on total cost

Figure 3.3 depicts the TC-Demand ratio curve; the X-axis is the ratio of class 2 demand to total demand. It is evident that total cost decreases as the class 2 demand increases. This is understandable since class 2 demand has lower stock out and delay cost. It shows that the cost of the system becomes lesser when there is more class 2 demand and the firm has to invest lesser amount for inventory related cost.

Figure 3.4: Effect of demand on number of backorders

Figure 3.4 shows the variation of the number of backorders for the classes with the demand. The backorder for class 2 shoot up rapidly as more class 2 demand arrives. Number of class 1 backorders, on the other hand, decrease as there is lower class 1 demand, since class 2 demand is increasing.

In the next experiment, the demands in example 3 are kept same as in the base case (4.4) and the stock out cost varied. Initially it is assumed that the stock out cost for both the classes is the same (\$10); later the stock out cost for class 1 is increased in increments of 5, keeping the stock out cost for class 2 fixed at \$10. The other problem parameters are the same as in the base case (Table 3.2). Figure 3.5 shows the variation of total cost with increasing stock out cost ratio. The x-axis represents the ratio of stock out costs of class 1 and class 2. The total cost increases as the stock-out cost ratio increases, since π_2 is kept fixed and π_1 increases as the ratio increases.

Figure 3.5: Total cost variation vs. stock out cost ratio

Figure 3.6 shows the change in rationing co-efficient (α) as stock out cost ratio increases. Since the stock out cost for class 2 is fixed (\$10) the rationing co-efficient of class 2 will not change much, but the rationing co-efficient of class 1 increases rapidly as the stock-out ratio increases. When stock out cost for class 1 increases, it is more beneficial to satisfy class 1 demand; accordingly its rationing co-efficient (α) increases. Managerially, it shows that more items should be reserved for the higher class if stockout cost of the higher class increases with respect to the lower class.

Figure 3.6: Effect of varying stock out cost on rationing co-efficients

3.6 Summary

In this chapter inventory rationing for ‘ n ’ customer classes differing in their service level requirement is considered where demand rates for the classes is deterministic. The classes differ in the penalty cost they incur if their demand is not satisfied from stock. Unlike previous literature on deterministic demand, a general penalty cost structure is considered, rather than time based delay cost only. The expression for total annual cost is developed and an algorithm is developed to determine the cycle time, T and, in turn, the run-out time, $t_i (= \alpha_i T)$ for each class. The threshold inventory level K_i for rationing of class i demand can be determined using T and α_i .

A numerical study was carried out consisting of three examples. The optimal run-out times in the numerical study show that the demand from the higher classes need

to be fulfilled immediately in most of the cases, whereas the demand from lower classes can be rationed and back-ordered. Although most demand patterns in general are stochastic, this model can still be useful when there isn't much variation in demand over the planning horizon. There are also cases when the firm has a service contract with the customers and demand is known in advance. In those cases, the model presented in this chapter would be quite useful. Moreover, the insights from the deterministic demand model are used to develop a new class of policies for the stochastic demand case in the next chapter.

Chapter 4

Inventory rationing for continuous review system with multiple customer classes

4.1 Introduction

With increasing focus on aftermarket services even among manufacturing firms, service differentiation is seen as means to achieve higher profits. Inventory policies that can tackle differentiated service requirements effectively and efficiently are therefore very important in the management of service parts logistics. Policies based on inventory rationing are one of the best techniques to minimize the inventory cost and to provide appropriate service levels to the different customer classes. In inventory rationing, the lower priority demand is not fulfilled immediately on arrival after the on hand inventory reaches a pre-defined rationing level. Even if inventory is available when the demand

arrives (but below the rationing level), it is reserved for satisfying higher priority demand and lower priority demand is backordered. Deshpande et al. (2003) derived an optimal rationing policy in a continuous review system for service differentiated customers, but only for two classes.

While it is unrealistic to expect to have a large number of different customer/demand classes in a service system in practice, it is quite common to have more than two classes. For example, in interactions with a leading automobile company, revealed that they use three demand classes for the purposes of service differentiation. In this chapter, a model is developed to extend the rationing policy of Deshpande et al. (2003), for ' n ' customer classes where demand from each class is backordered after on hand inventory falls below the critical level of that class. The contribution of this research is the development of exact expression for the expected total cost. As in the Deshpande et al. (2003), it is assumed that demand from the customer classes follow a Poisson process. Holding cost is incurred at a constant rate for inventory held on hand and a fixed ordering cost is incurred for every replenishment. There is a fixed lead time for replenishment of orders. When a demand is backordered it incurs two costs: *delay cost* per unit time per unit incurred for the time duration the demand is backordered and a one-time *stock out cost* per unit. No other work has derived the exact cost expression for multiple customer classes.

Veinott (1965) was the first to propose the concept of inventory rationing. Topkis (1968) analyzed inventory rationing but considered dynamic rationing where rationing levels change with time. He considered periodic review of inventory and also assumed a

general demand process. Many others also considered periodic review and tried to solve the problem of service differentiation. Nahmias and Demmy (1981) were the first to consider continuous review systems with rationing and derived approximate expressions for the expected backorders for two customer classes. Moon and Kang (1998) later extended that to find the approximate expression for fill rate assuming compound Poisson process. Evans (1968), Melchioris et al. (2000) considered continuous review systems with lost sales. Dekker et al. (2002) also considered lost sale case and derived the expression for expected cost. Lost sales are more appropriate in a retail setting where rationing might not always be appropriate or practical. In industrial or B2B settings such as spare parts inventory systems, it might be more appropriate to assume that unsatisfied demand is backordered rather than lost.

The most relevant paper to this research is the one by Deshpande et al. (2003). They assumed that the demand from the two classes follow Poisson processes and are satisfied from the inventory on a first come first serve (FCFS) basis, until on hand inventory is equal or above the rationing level. Demand from the lower priority class (defined as class 2) is backordered after on hand inventory falls below rationing level, K , but demand from the higher priority class (defined as class 1) is still satisfied as long as inventory is available on hand. Demand from class 1 is also backordered after the inventory runs out. The inventory is replenished according to a (Q, r) policy where an replenishment order of quantity Q is placed whenever inventory position drops to the level r . For a (Q, r) policy, with a fixed rationing level, Deshpande et al. (2003) derived the expression for the average expected cost. For a fixed rationing level, K , they were

able to use the algorithm of Federgruen and Zheng (1992) to determine the optimal Q and r . the optimal rationing policy can then be determined by searching exhaustively over all possible values of K . They showed the benefit of inventory rationing over other policies such as a separate stock policy and a policy based on rounding up the service levels.

Following Deshpande et al. (2003), Arslan et al. (2007) considered inventory systems with more than two customer classes. However, their main focus was on formulating the problem as a serial stage inventory system and then demonstrating the equivalence to Deshpande et al. (2003)'s model. Moreover, instead of including penalty costs into the model, they attempted to minimize the inventory holding cost subject to service level constraints.

In this research, the problem with multiple ($n > 2$) customer classes is considered, but with the objective of minimizing the total costs including penalty cost, as in Deshpande et al. (2003). Apart from developing an algorithm to solve the multiple-class problem, the impact of collapsing an n -class model into a 2-class or 3-class model (by aggregating several classes together) is investigated, numerically, for the purpose of calculating and implementing an inventory policy. The major findings of the computational experiments of this research are as follows: the computational time required to find the optimal policy increases significantly with the number of customer classes in the model. However, in most of the cases, an aggregated 2-class or 3-class model achieves most of the benefits of rationing with a cost that is only marginally higher than an n -class model.

The rest of the chapter is organized as follows. In section 4.2 the threshold inventory rationing model with multiple ($n > 2$) customer classes is developed. The algorithm for determining the optimal policy parameters is also developed in this section. Section 4.3 discusses the results of the computational experiments comparing the n -class model with a (aggregated) 2-class and 3-class model. Section 4.4 provides a brief summary and the concluding remarks on this chapter.

4.2 Inventory rationing model with multiple customer classes

The notation used in this chapter is presented in table 4.1.

Table 4.1 Notations for multiple-class rationing model

Symbols	Meaning
i	Class index
n	Number of classes
λ_i	Demand rate for class i
λ	Total demand rate of all the classes
p_i	Time-weighted delay cost per unit for class i
π_i	Fixed stock-out cost per unit for class i
h	Holding cost per unit per unit time.
A	Fixed ordering cost per order
L	Lead time for replenishment arrival
Q	Order quantity
r	Reorder point
S_i	Base stock for class i
K_i	Rationing level for class i

The problem being considered is a single item inventory system. The demand for the item arises from n different customer classes. The demand from each customer class i follows a Poisson process with mean λ_i . The total demand therefore also follows a Poisson process with mean $\lambda = \sum_{i=1}^{i=n} \lambda_i$. Every replenishment order incurs a fixed ordering cost, A . Inventory held on hand incurs holding cost per unit time at a rate h . The (fixed) lead time for replenishment is $L > 0$. Inventory is replenished according to a (Q, r) policy. When the inventory position (on hand inventory + on order – backorders) drops to r or below, a replenishment order for a fixed quantity Q is placed. All demands that are not satisfied immediately on arrival are backordered and incur two types of penalty cost – delay cost per unit per time (p_i), and stock out cost per unit (π_i). Customer class 1 is assumed to be the highest priority class and class n is the lowest priority class. Thus by definition $p_1 \geq p_2 \geq \dots \geq p_{n-1} \geq p_n$ and $\pi_1 \geq \pi_2 \geq \dots \geq \pi_{n-1} \geq \pi_n$.

The rationing policy operates as follows. Demands from all classes are satisfied on a first come first served (FCFS) basis as long as on hand inventory is above K_n , the rationing level for the n^{th} class. After that, when on hand inventory drops to a level equal to K_n or below, only demand from class 1 to class $n-1$ is satisfied until the on hand inventory reaches K_{n-1} (the rationing level of $n-1^{th}$ class) and demand from n^{th} class is backordered. Similarly demand from $n-1^{th}$ class is backordered after on hand inventory drops to K_{n-1} or below and demand from class 1 to class $n-2$ is satisfied. In general, demand from class i to class n is backordered when inventory level is equal to or below K_i . By definition, $K_1 = 0$. The policy is illustrated in figure 4.1.

As in Deshpande et al. (2003), upon receipt of a replenishment order, the backorders are cleared according to a threshold clearing mechanism. Essentially this mechanism determines the sequence in which backorders are cleared once a new replenishment arrives after lead time L . The threshold clearing mechanism clears backorders as if the inventory level immediately after placing the replenishment order was $(r+Q)$ and rationing was carried out according to the rationing policy. This means all the class i demand that occurred until the $(r+Q-K_i)^{\text{th}}$ demand arrival in order cycle is cleared upon the receipt of the replenishment for that cycle. Any class i demand that occurs subsequently in that cycle will only be cleared upon receipt of the replenishment in the next order cycle. As stated Deshpande et al. (2003), the threshold clearing mechanism is necessary to ensure that the inventory positions are regenerated in every order cycle and this in turn facilitates the calculation of the steady state probability distribution of the inventory levels.

The rationing policy can be effectively represented as an equivalent n -bin policy. Firstly, this equivalent n -bin policy along with the threshold clearing mechanism is explained for a system with two customer classes (2-bin policy) for the l^{th} order cycle. After the placement of the l^{th} order of quantity Q the inventory position is $(r + Q)$. Of this, $(r+Q-K_2)$ unit is allocated to BIN₂, and the remaining inventory K_2 to BIN₁. When a new demand (for both class 1 and class 2) arrives, it is first attempted to be satisfied from BIN₂. If BIN₂ inventory position is > 0 , it is assumed to be satisfied from that BIN₂. If the BIN₂ inventory position is ≤ 0 , and if the demand is from class 2, it is backordered and will be cleared (i.e. satisfied) only upon arrival of $(l+1)^{\text{th}}$ replenishment order. If BIN₂

inventory position > 0 , but the actual inventory level (on-hand inventory) is zero, then if the demand is for class 2, it is backordered but cleared upon arrival of the l^{th} replenishment order. [Note that since inventory position is > 0 , there will be sufficient inventory upon arrival of the l^{th} replenishment order]. However, if the demand is for class 1, and if the inventory level in BIN_2 is zero, then the demand is satisfied from BIN_1 and the BIN_1 inventory level is reduced by 1; the demand is backordered if BIN_1 inventory level is also zero. If the inventory position BIN_2 was ≥ 0 , then technically it should have been satisfied from BIN_2 , therefore inventory position of BIN_2 is reduced by 1, and correspondingly qt_{12} , the quantity borrowed from BIN_1 by BIN_2 , is updated by 1; this quantity would be returned to BIN_1 upon receipt of l^{th} replenishment order. If BIN_2 inventory position is ≤ 0 , but BIN_1 inventory position is > 0 , then a demand for class 1 would be satisfied immediately if BIN_1 inventory level is > 0 , or backordered and cleared upon arrival of l^{th} replenishment order. As BIN_1 inventory position is ≥ 0 , there would be sufficient inventory to clear the backlogs upon arrival of l^{th} replenishment order.

If BIN_1 and BIN_2 have inventory position ≤ 0 , then class 1 demand is also backordered and can be cleared only upon arrival $(l+1)^{\text{th}}$ replenishment order. The inventory level as well as the backorders for both the customer classes after a lead time L with the 2-bin system will be the same as under the rationing policy with threshold clearing, since only the class 2 demand arriving within first $(r + Q - K_2)$ demands are cleared upon arrival of the replenishment order.

Figure 4.1 Inventory rationing for multiple classes

Figure 4.2 Inventory rationing with threshold clearing for 3 classes

In a similar manner the n -class problem with rationing and threshold clearing mechanism can be equivalently represented as an n -bin policy. Immediately after ordering, the inventory position (or the total base stock) becomes $(r + Q)$ and this inventory is allocated to the n bins as follows:

Base stock S_1 allocated to $\text{BIN}_1 = K_2 - K_1 = K_2$

(Note that by definition, $K_1=0$)

Base stock S_2 allocated to $\text{BIN}_2 = K_3 - K_2$

Base stock S_i allocated to $\text{BIN}_i = K_{i+1} - K_i$

Base stock S_n allocated to $\text{BIN}_n = K_{n+1} - K_n$

(where $K_{n+1} = r + Q$ by definition)

Based on the above allocation, it is easy to show that

$$K_i = \sum_{j=1}^{i-1} S_j \quad i=2,3,\dots,n$$

$$S_i = K_{i+1} - K_i \tag{4.1}$$

and
$$r = \sum_{i=1}^n S_i - Q$$

This is evident from figure 4.2, for a problem with three customer classes.

Figure 4.3 Inventory rationing as n -bin policy

The decision variables (policy parameters) for n -bin rationing policy can be represented as $(Q, \mathbf{S})$, where Q is the order quantity and $\mathbf{S} = \{S_1, S_2, \dots, S_n\}$ is the base stock value for each bin. Note that once Q and $\mathbf{S}$ are known, the threshold level K_i and recorder point r can be calculated using (4.1). The n -bin policy that is equivalent to the rationing policy (without the intricacies of the threshold clearing mechanism) is summarized in Figure 4.3. The intricacies of the threshold clearing mechanism for the n -bin policy (for a particular order cycle) summarized through a flowchart presented in Figure 4.4. The corresponding pseudo code (for clearing of orders for i^{th} class demand in the l^{th} order cycle) is given in Figure 4.5. The inventory position and inventory level of BIN_i is represented as $\text{IP}(\text{BIN}_i)$ and $\text{IL}(\text{BIN}_i)$, respectively in these figures.

Figure 4.4 Flowchart for n -bin policy

```

j = n;
while (IP(BINj) ≤ 0), j-;
if (j < i), (BOiCR)++;
else
{
    IP (BINj)-;
    if IL(BINj) > 0
    {
        IL(BINj)-;
    }
    else
    {
        j = j*;
        while (IL(BINj) ≤ 0), j*-;
        if j* < i, (BOiCL)++;
        else
        {
            IL(BINj*)-;
            qtj*j++;
        }
    }
}

```

Figure 4.5 Pseudo code for clearing of the i^{th} class demand in the l^{th} order cycle in the rationing policy represented by n -bin

Note that the order quantity for each BIN_i for the l^{th} order cycle is $q_i = S_i - IP_i^{\tau_l}$, where $IP_i^{\tau_l}$ is the inventory position of BIN_i (at time τ_l) when the order is placed. The total order quantity is $Q = \sum_{i=1}^n q_i$ (since $\sum_{i=1}^n S_i = r + Q$ and $\sum_{i=1}^n IP_i = r$).

Based on the clearing mechanism used for satisfying/clearing the demand, the quantity that is received for the l^{th} order would be sufficient to clear all backorders for the l^{th} order cycle i.e. BO_i^{CL} as well as the quantity loaned from BIN_i to BIN_j , qt_{ij} , $i=1,2,...,n$; $j=i+1,i+2,...,n$. When the replenishment is received, first the inventory levels IL_i for BIN_i is updated according to the quantity received, thereafter BO_i^{CL} is cleared from this IL_i and finally the quantity qt_{ij} is returned back to BIN_j . Note that all backorders that occur between the placement of l^{th} and $l+1^{th}$ order and not cleared with the l^{th} replenishment, will certainly be cleared with $l+1^{th}$ replenishment as the inventory position at $l+1^{th}$ order are updated to ensure that the previous backorders are cleared.

The proposed n -bin policy with the threshold clearing mechanism above, effectively works similar to policy of Arslan et al. (2007) that used a serial inventory system framework. However, the analysis in this chapter and cost evaluation is different in the sense that a multi-bin policy is used instead of a multi-stage framework. Moreover, the focus of this chapter is on the minimization of total expected cost (comprised of ordering, holding and penalty costs) unlike Arslan et al. (2007) who minimizes the average inventory level for a fixed ordering quantity, and subject to a service level (or order fill rate) constraint. Further the computational experiments carried out in this research are focused on evaluating the benefit of using an n -class model as opposed to a 2

or 3-class model (where some of the customer classes are aggregated together). The proposed n -bin rationing policy with the threshold clearing mechanism is now illustrated using an example, as shown in Table 4.2. The example is for 3 classes with base stock $S_1 = 2$, $S_2 = 3$ and $S_3 = 4$ and order quantity $Q = 5$. qt_{ij} denotes the units of demand that have been satisfied from BIN_i but should have been satisfied from BIN_j since inventory position of BIN_j is positive, $i < j$. These units have to be returned to BIN_i when BIN_j receives the replenishment. In this example when the 6th demand arrives, it is satisfied from BIN_2 whereas it should have been satisfied from BIN_3 because inventory position of BIN_3 was 3 at that time, but on hand inventory was 0. So, BIN_3 returns this unit to BIN_2 when it receives the replenishment for the l^{th} cycle. BO_i^{CL} denotes the backorder from class i which will be cleared when l^{th} replenishment arrives and BO_i^{CR} denotes the backorder which will be carried forward and cleared in the $l+1^{\text{th}}$ replenishment cycle. In the above example, when the 8th demand (from class 3) arrives, it is backordered since there was no inventory available but the inventory position was positive. Therefore it should have been satisfied if there was inventory. This backorder is cleared when l^{th}

Table 4.2 Example of inventory rationing with threshold clearing for 3 classes

Demand from class	IP ₁	IP ₂	IP ₃	OH ₁	OH ₂	OH ₃	Action	
	2	3	4	2	3	4		
1	2	3	3	2	3	3	Satisfied from BIN ₃	
2	2	3	2	2	3	2	Satisfied from BIN ₃	
1	2	3	1	2	3	1	Satisfied from BIN ₃	
3	2	3	0	2	3	0	Satisfied from BIN ₃	
1	2	2	0	2	2	0	Satisfied from BIN ₂ 1 st Order Placed q ₁ =0, q ₂ =1 and q ₃ =4	
	2	3	4	2	2	0	Raise IP to S _i	
2	2	3	3	2	1	0	Satisfied from BIN ₂	qt ₂₃ =1
1	2	3	2	2	0	0	Satisfied from BIN ₂	qt ₂₃ =2
3	2	3	1	2	0	0	Backorder	BO ₃ ^{CL} =1
1	2	3	0	1	0	0	Satisfied from BIN ₁	qt ₁₃ =1
3	2	3	-1	1	0	0	Backorder 2 nd Order Placed q ₁ =0, q ₂ =0 and q ₃ =5	BO ₃ ^{CR} =1
	2	3	4	1	0	0	Raise IP to S _i	
	2	3	4	1	1	4	1 st Order Received	
				2	3	1	Return q ₃₂ and q ₃₁	qt ₂₃ =0, qt ₁₃ =0
				2	3	0	Clear BO	BO ₃ ^{CL} =0 BO ₃ ^{CR} =1

The proposed n -bin rationing policy with threshold clearing can be summarized as follows:

1. *When a demand from class i , $i=1,2,3,\dots,n$ arrives:* Satisfy the demand as per the clearing mechanism explained above and summarized in the flowchart (Figure 4.4) and pseudo code (Figure 4.5).
2. *Ordering Policy:* When total demand arrival of all the classes since the last replenishment order is Q (or when $\sum_{i=1}^n IP_i$ drops to r), place an order of Q and raise the inventory position of all the bins to S_i ; $i=1, 2, \dots, n$. The quantity ordered for each BIN $_i$, $q_i = S_i - IP_i$ and $Q = \sum_{i=1}^n q_i$.

3. *When replenishment order arrives*

For all BIN $_i = n$ down to 1

Update $IL_i = IL_i + q_i$ and $IL_i = IL_i - (BO_i^{CL})_i - (BO_i^{CL})_{i-1}$

For all $k < i$, $IL_k = IL_k + qt_{ki}$ and $IL_i = IL_i + qt_{ki}$

Note that when the lead time is much longer than the order cycle, there could be several replenishment orders outstanding. Therefore, the backorders BO_i^{CL} and BO_i^{CR} have to be tracked separately for each order cycle. Hence, the notations BO_i^{CL} and BO_i^{CR} are used in the discussions above.

4.2.1 Evaluation of the policy cost

The decision variables in this model are Q and $\mathbf{S} = \{S_1, S_2, \dots, S_n\}$. The objective is to determine the policy $(Q, \mathbf{S})$ that minimizes the expected cost (comprised of ordering, inventory holding and penalty cost).

The total cost for a given $(Q, \mathbf{S})$ can be written as

$$TC(Q, \mathbf{S}) = \frac{\lambda A}{Q} + G(Q, \mathbf{S}) \quad (4.2)$$

where $\frac{\lambda A}{Q}$ is the expected ordering cost per unit time and $G(Q, \mathbf{S})$ is the expected inventory holding and penalty cost.

Under the $(Q, \mathbf{S})$ policy the inventory position follows a regenerative process. Let the l^{th} order is placed at time τ_l and the $l+1^{\text{th}}$ order is placed at τ_{l+1} . Let t be any point in the l^{th} replenishment cycle such that $\tau_l \leq t < \tau_{l+1}$. Knowing the probability distribution of

inventory position at time t (or total demand in the interval $[\tau_l, t)$), and the demand during period $[t, t+L)$, the cost $G(Q, \mathbf{S})$ can be evaluated. The order placed at τ_l would be received by $\tau_l + L$. Therefore under the threshold clearing mechanism the inventory represented by inventory position S_i , $i = 1, 2, \dots, n$, at τ_l would be fully available at $\tau_l + L$,

and this inventory position, less the demand during $(\tau_l, t+L)$, would accurately represent the inventory level/stock out levels at time $t+L$. Also the backorders from the previous order cycle is already accounted for in the inventory position S_i (even though they may be cleared only at $\tau_l + L$) and hence these can be ignored in computing the inventory levels and cost $G(Q, S)$ at time $t+L$. However, due to the way threshold clearing mechanism is designed, the sequence of demand arrivals during the period $(\tau_l, t+L)$ is important in arriving at the correct penalty and holding costs. Given the Poisson demands and the use of a (Q, r) policy, the total demand during the period $[\tau_l, \tau_{l+1})$ is $Q-1$ (the Q^{th} demand arrives exactly at time τ_{l+1}) and the total demand for all the classes, D_t , during the period $[\tau_l, t)$ follows a uniform distribution with

$$g(D_t) = \frac{1}{Q} \quad D_t = 0, 1, 2, \dots, Q-1 \quad (4.3)$$

The total demand for all the classes during the period $[t, t+L)$, D_L follows a Poisson distribution

$$f(D_L = j) = \frac{e^{-\lambda} (\lambda L)^j}{j!} \quad j = 0, 1, \dots, \infty \quad (4.4)$$

The probability distribution of the total demand, D_{t+L} , during the period $(\tau_l, t+L)$ can be computed as a convolution of the uniform distribution and the Poisson distribution. Specifically,

$$r_j = \Pr(D_{t+L} = j) = \sum_{i=0}^{\min(j, Q-1)} \left(\frac{1}{Q} \frac{e^{-\lambda L} (\lambda L)^{j-i}}{(j-i)!} \right) \quad j = 0, 1, \dots, \infty \quad (4.5)$$

If the total demand, $j \leq S_n$, then all the demand is satisfied from the n^{th} bin and there will be $(\sum_i^n S_i - j)$ units of inventory left. Therefore

$$G(Q, \mathbf{S}) = r_j h(\bar{S} - j), \quad j \leq S_n \quad (4.6)$$

where $\bar{S} = \sum_{i=1}^n S_i$.

However, if the demand is more than S_n , then rationing comes into play. If the demand j during $[\tau_l, t+L)$ is such that $S_n < j \leq S_n + S_{n-1}$, then clearly only the demand from class n is rationed. In order to calculate the penalty cost for the n^{th} demand class, the sequence of demand arrivals need to be taken into account.

Clearly, all the arrivals of the n^{th} class demand in the first S_n demand arrivals would be fulfilled as there would be sufficient inventory in the n^{th} bin. So it is immaterial how many n^{th} class demands arrived in the first S_n demands. What is important is the number of n^{th} class demands in the demand arrivals $S_n+1, S_n+2, \dots, j$. Since all the n classes face Poisson demands, the probability that an arriving demand belongs to class i is $\alpha_i = \frac{\lambda_i}{\lambda}$. So, the number of n^{th} class demands in the demand arrivals $S_n+1, S_n+2, \dots, j$ follows a binomial distribution. Therefore, if the demand is such that $S_n < j \leq S_n + S_{n-1}$, the cost would be

$$G(Q, \mathbf{S}) = r_j \sum_{i_n=0}^{j-S_n} B(\alpha_n, j-S_n, i_n) [h(\bar{S} - j + i_n) + i_n p_n + \lambda_n \pi_n] \quad (4.7)$$

where $B(\alpha_k, M, i) = {}^M C_i \alpha_k^i (1 - \alpha_k)^{M-i}$ is the binomial distribution for i outcomes of type k , in M trials, α_j is the probability of type k outcome in a single trial. $B(\alpha_n, j - S_n, i_n)$ denotes the probability of i_n units of n^{th} class demand out of total $j - S_n$ units of demand. These i_n units would be backordered and incur delay cost at rate $i_n p_n$. [Note that due to PASTA property the expected stock out cost when demand class k is not served is $\lambda_k \tau_k$]. Thus the units of demand (for all classes) fulfilled out of j units, in this case, would be $j - i_n$ and the inventory left would be $\bar{S} - (j - i_n)$. Therefore, the holding cost would be $h(\bar{S} - j + i_n)$.

When $S_n + S_{n-1} < j \leq S_n + S_{n-1} + S_{n-2}$, the cost expression gets more complicated. In the demand arrivals between the $(S_n + S_{n-1} + 1)^{\text{th}}$ to the j^{th} demand, both n^{th} and $(n-1)^{\text{th}}$ class demand can possibly be backordered whereas between $(S_n + 1)^{\text{th}}$ to $(S_n + S_{n-1})^{\text{th}}$ demand only the n^{th} class is backordered. The cost expression now involves nested binomials. The n^{th} class demand would be backordered for all demand arrivals from $S_n + 1, S_n + 2, \dots, j$. Out of these $(j - S_n)$ demand arrivals let i_n be the number of n^{th} class demand. Therefore, i_n follows a binomial distribution $B(\alpha_n, j - S_n, i_n)$. The possible values i_n can take is 0 to $(j - S_n)$. Since i_n demands of n^{th} class are backordered (not fulfilled immediately), the $(n-1)^{\text{th}}$ bin would be empty only after the $(S_n + S_{n-1} + i_n)^{\text{th}}$ demand. All $(n-1)^{\text{th}}$ class demands that arrive upto the $(S_n + S_{n-1} + i_n)^{\text{th}}$ demand would be fulfilled. Backordering of the $(n-1)^{\text{th}}$ class demand can only occur for the demand arrivals

$S_n + S_{n-1} + i_n + 1, \dots, j$. Refer to figure 4.2 for a 3-class problem. Due to the memoryless property of the Poisson process, each of these $j - (S_n + S_{n-1} + i_n)$ demand arrivals are independent events (or trials) with probability $\alpha_{n-1} = \frac{\lambda_{n-1}}{\lambda}$ that the particular demand belongs to $(n-1)^{th}$ class. Therefore, when $S_n + S_{n-1} < j \leq S_n + S_{n-1} + S_{n-2}$, and the number of backordered n^{th} class demand = i_n , the number of backordered $(n-1)^{th}$ class demand i_{n-1} follows a Binomial distribution

$$B(\alpha_{n-1}, j - (S_n + S_{n-1} + i_n), i_{n-1}) \quad (4.8)$$

Note that when $i_n > j - (S_n + S_{n-1})$, the $(n-1)^{th}$ bin would still have inventory at time $t+L$, and therefore $i_{n-1} = 0$, as there will be no backordered demand for the $(n-1)^{th}$ class. Let $B(\alpha_k, x, y)$ be defined as

$$B(\alpha_k, x, y) = \begin{cases} 0 & \forall x < 0 \text{ and } y \geq 0 \\ 0 & \forall x = 0 \text{ and } y > 0 \\ 1 & \forall x = 0 \text{ and } y = 0 \end{cases} \quad (4.9)$$

Then, the distribution given in (4.8) would still accurately represent the probability distribution of i_{n-1} backorders of demand class $n-1$ in j demand arrivals, given that there were i_n backorders of class n . Therefore, when $S_n + S_{n-1} < j \leq S_n + S_{n-1} + S_{n-2}$

$$G(Q, \mathbf{S}) = \begin{cases} r_j \sum_{i_n=0}^{j-S_n} B(\alpha_n, j-S_n, i_n) \left[h(\bar{S} - j + i_n) + i_n p_n + \lambda_n \pi_n \right] & \text{if } j - (S_n + S_{n-1} + i_n) < 0 \\ r_j \sum_{i_n=0}^{j-S_n} \left[B(\alpha_n, j-S_n, i_n) \sum_{i_{n-1}=0}^{j-(S_n+S_{n-1}+i_n)} B(\alpha_{n-1}, j-(S_n+S_{n-1}+i_n), i_{n-1}) \right. \\ \quad \left. \left[h(\bar{S} - j + i_n + i_{n-1}) + i_n p_n + \lambda_n \pi_n + i_{n-1} p_{n-1} + \lambda_{n-1} \pi_{n-1} \right] \right] & \text{otherwise} \end{cases} \quad (4.10)$$

Combining the two conditions in (4.10), the expression can be rewritten as

$$G(Q, \mathbf{S}) = r_j \sum_{i_n=0}^{j-S_n} \left[\begin{aligned} & \left[B(\alpha_n, j-S_n, i_n) \right] \left[h(\bar{S} - j + i_n) + i_n p_n + \lambda_n \pi_n \right] + \\ & \left[B(\alpha_n, j-S_n, i_n) \sum_{i_{n-1}=0}^{j-(S_n+S_{n-1}+i_n)} B(\alpha_{n-1}, j-(S_n+S_{n-1}+i_n), i_{n-1}) \right] \\ & \left[h i_{n-1} + i_{n-1} p_{n-1} + \lambda_{n-1} \pi_{n-1} \right] \end{aligned} \right] \quad (4.11)$$

In a similar manner, the expression for $G(Q, \mathbf{S})$ can be written for specific ranges j ,

$$\text{i.e. } 0 \leq j < S_n, S_n \leq j < S_n + S_{n-1}, S_n + S_{n-1} \leq j < S_n + S_{n-1} + S_{n-2}, \dots, \sum_{m=2}^n S_m \leq j < \sum_{m=1}^n S_m$$

Note, however, that when $j > \sum_{m=1}^n S_m$, there will be situations when the inventory at time

$t+L$ is 0 and all demands are backordered (i.e. $i_1, i_2, \dots, i_n > 0$). In that case (i.e. for

$$j - \left[\sum_{m=1}^n S_m + \sum_{m=2}^n i_m \right] \leq 0), \text{ there will be no holding cost term in the expression for } G(Q, \mathbf{S}).$$

The expression (4.11) has to be modified to take that into account.

Using (4.11) and the discussions above $G(Q, \mathbf{S})$ can be written as a recursive function, as shown below:

$$G(Q, \mathbf{S}) = \sum_{j=0}^{\infty} \left[r_j \phi_n(j - S_n) \right] \quad (4.12)$$

where

$$\phi_k(x) = \begin{cases} h(-x + \sum_{m=1}^{k-1} S_m) + \sum_{m=k+1}^n (i_m p_m + \lambda_m \pi_m) & \text{if } x < 0 \\ \sum_{i_k=0}^x B(\alpha_k, x, i_k) \phi_{k-1}(x - (S_{k-1} + i_k)) & \text{if } x \geq 0 \end{cases} \quad (4.13)$$

$$\phi_1(x) = \begin{cases} h(-x) + \sum_{m=2}^n (i_m p_m + \lambda_m \pi_m) & \text{if } x < 0 \\ x p_1 + \lambda_1 \pi_1 + \sum_{m=2}^n (i_m p_m + \lambda_m \pi_m) & \text{if } x \geq 0 \end{cases} \quad (4.14)$$

Note that by definition, $i_{n+1}, p_{n+1}, \lambda_{n+1}$ and $\pi_{n+1} = 0$.

For a three class problem, after expanding the recursive expression given in (4.12),

$G(Q, \mathbf{S})$ can be written as follows:

$$\begin{aligned} G(Q, \mathbf{S}) = & \sum_{j=0}^{S_3-1} r_j h(\bar{S} - j) + \sum_{j=S_3}^{S_3+S_2-1} \left[r_j \left(\sum_{i_3=S_3}^{j-S_3} B(\alpha_3, j - S_3, i_3) [h(\bar{S} + i_3 - j) + i_3 p_3 + \lambda_3 \pi_3] \right) \right] \\ & + \sum_{j=S_3+S_2}^{\bar{S}-1} r_j \left[\sum_{i_3=0}^{j-(S_3+S_2)} B(\alpha_3, j - S_3, i_3) \left[\sum_{i_2=0}^{j-(S_3+S_2+i_3)} B(\alpha_2, j - (S_3 + S_2 + i_3), i_2) \right. \right. \\ & \left. \left. \left(h(\bar{S} + i_3 + i_2 - j) + \sum_{k=2}^3 (i_k p_k + \lambda_k \pi_k) \right) \right] \right. \\ & \left. + \sum_{i_3=j-(S_3+S_2)+1}^{j-S_3} B(\alpha_3, j - S_3, i_3) [h(\bar{S} + i_3 - j) + i_3 p_3 + \lambda_3 \pi_3] \right] \\ & + \sum_{j=\bar{S}}^{\infty} r_j \left[\sum_{i_3=0}^{j-\bar{S}} B(\alpha_3, j - S_3, i_3) \left[\sum_{i_2=0}^{j-(\bar{S}+i_3)} \left((j - (\bar{S} + i_3 + i_2)) p_1 + \sum_{k=2}^3 (i_k p_k + \lambda_k \pi_k) \right) \right. \right. \\ & \left. \left. + \sum_{i_2=j-(\bar{S}+i_3)+1}^{j-(S_3+S_2+i_3)} \left(h(\bar{S} + i_3 + i_2 - j) + \sum_{k=2}^3 (i_k p_k + \lambda_k \pi_k) \right) \right] \right. \\ & \left. + \sum_{i_3=j-\bar{S}+1}^{j-S_3} B(\alpha_3, j - S_3, i_3) [h(\bar{S} + i_3 - j) + i_3 p_3 + \lambda_3 \pi_3] \right] \end{aligned} \quad (4.15)$$

Following the procedure similar to the one above, the expressions for $G(Q, \mathbf{S})$ can be developed for any problem with a certain specific number of classes. An algebraic expression for anything more than three classes or for a general n -class problem would look very messy. However, for computational purposes to evaluate $G(Q, \mathbf{S})$, the recursive expressions given by (4.12), (4.13) and (4.14) can be used.

4.2.2 Solution algorithm

The method for deriving the cost expression for multiple customer classes and the exact cost expression for 3 customer classes is provided in Section 4.2.3. In this section, a method is proposed to determine the optimal value of Q and the base stocks S_i , $i = 1, 2, \dots, n$ that minimize the expected cost. Deshpande et al. (2003) showed that for two customer classes and for a fixed value of S_1 ; Q and S_2 can be determined (using an algorithm similar to that of Federgruen and Zheng (1992)) by exploiting the convexity of the underlying cost terms. However, they still needed to enumerate the cost exhaustively for all possible values of S_2 . Unfortunately with a multi-bin policy formulation for n classes, no such structural results on the cost function can be obtained to use efficient algorithms like that of Federgruen and Zheng (1992). Therefore, exhaustive enumeration of the cost function $TC(Q, \mathbf{S})$ across all possible values of Q , S_i , $i = 1, 2, \dots, n$ is required to determine the optimal rationing policy. However, some bounds for Q and S_i can be developed to restrict the search. These bounds are derived below.

4.2.2.1 Bounds on $\bar{S} = \sum_{i=1}^n S_i$ and Q

Clearly a lower bound $\bar{S}_{L1}$ on $\bar{S}$ can be obtained by solving a single customer class problem with demand $\lambda = \sum_{i=1}^n \lambda_i$, $\pi = \pi_n$ and $p = p_n$. The corresponding Q_U will also be an upper bound on Q . An upper bound $\bar{S}_U$ on $\bar{S}$ and a lower bound Q_{L1} on Q can be obtained by solving a single customer class problem with $\lambda = \sum_{i=1}^n \lambda_i$, $\pi = \pi_1$ and $p = p_1$.

If the different customer classes were to be treated as different items (i.e. BIN_i can only be used to satisfy demand for class i), the problem reduces to a joint replenishment problem (JRP). The (Q, S) policy developed by Pantumsinchai (1992) can be used to solve this problem. As the inventory of each bin is segregated to satisfy only a specific demand class, there is no substitution between the bins. Therefore, $\bar{S} = \sum_{i=1}^n S_i$ of this solution should serve as an upper bound $\bar{S}_{U2}$ on $\bar{S}$ of the original problem. The upper bound then can be decided as $\bar{S}_U = \min(\bar{S}_{U1}, \bar{S}_{U2})$. The value of S_n obtained in this JRP problem would be a lower bound on S_n of the original problem. The inventory holding and penalty costs in the JRP would be higher as there is no substitution between the bins. Therefore, Q corresponding to the JRP solution would be a lower bound Q_{L2} on Q . Finally, the lower bound on Q can be found by $\bar{Q}_L = \max(Q_{L1}, Q_{L2})$.

In the rationing policy, the effective demand handled by BIN_n is $> \lambda_n$ since it can be used to satisfy demand from all the classes. Similarly the effective demand handled by $BIN_n + BIN_{n-1}$ is $> \lambda_n + \lambda_{n-1}$. Therefore, lower bounds on S_n , $S_n + S_{n-1}$, $S_n + S_{n-1} + S_{n-2}, \dots$ etc can be obtained by solving single item single class inventory problems with respective demands equal to λ_n , $\lambda_n + \lambda_{n-1}$, $\lambda_n + \lambda_{n-1} + \lambda_{n-2}, \dots$ etc. with delay cost and stock out cost equal to π_n and p_n , respectively. This can be used to derive lower bounds on S_n , $S_{n-1}, \dots, S_1$.

An upper bound on S_n , $S_n + S_{n-1}$, $S_n + S_{n-1} + S_{n-2}, \dots$ etc can be obtained by repeatedly solving the JRP combining the n^{th} class, the n^{th} and $(n-1)^{\text{th}}$ class (with penalty cost corresponding to the $(n-1)^{\text{th}}$ class), the n^{th} , $(n-1)^{\text{th}}$ and $(n-2)^{\text{th}}$ class (with penalty cost corresponding to the $(n-2)^{\text{th}}$ class) and so on. This can be used to derive upper bounds on S_n , $S_{n-1}, \dots, S_1$.

The algorithm for determining the optimal threshold rationing policy then involves searching for $TC(Q, S)$ within the lower and upper bounds of Q , S_i , $i=1, 2, \dots, n$ as

well as $\bar{S} = \sum_{i=1}^n S_i$.

4.3 Numerical results

The algorithm developed in section 4.2.4 can be used to determine the optimal rationing policy and the corresponding expected cost. The expected cost of a policy is succinctly expressed as a recursive function. However, the computational effort required to evaluate it increases significantly as the number of customer classes

increases. And the cost has to be evaluated several times in search algorithm for the optimal solution. The experience from the computational experiments in this research suggests that it is unrealistic to expect to solve problems involving more than 5 customer classes in reasonable time.

Managerially, it is also cumbersome to implement a system involving too many customer classes in practice. So a relevant practical question is, what is the cost increase if the several customer classes are aggregated together so that the system itself implemented with two or three differentiated service classes. In the numerical study, the answer to the above question is sought.

In the numerical study, first an example is provided which shows the optimal cost and optimal parameters for a 5-class problem under certain parameters combinations. Then, the cost increase from implementing a (aggregated) 3-class rationing policy for a 5-class model is first evaluated. The cost increase from implementing a 2-class rationing policy for a 5-class model is also evaluated.

The algorithm developed for determining optimal parameters of a multi-bin policy is used to determine the optimal parameters for 5 customer classes. Since an exhaustive search for different parameters has to be performed within parameter bounds, the computational complexity increases with the number of classes, n , as the total number of decision variables are $n+1$. Thus, the problem with more than 5 classes is computationally prohibitive. The optimal cost and parameters for different settings are given in table 4.3.

Table 4.3 Optimal Parameters for 5 customer classes

L	p_1	p_2	p_3	p_4	p_5	Cost	Q, S Parameters	Q, r, K Parameters
							$Q, S_1, S_2, S_3, S_4, S_5$	Q, r, K_2, K_3, K_4, K_5
0.5	5000	4000	3000	2000	1000	2309.25	7,0,0,0,2,19	7,14. 0,0,0,2
0.5	10000	5000	3000	2000	1000	2429.93	7,0,0,1,1,20	7,15,0,0,1,2
1	10000	7000	3000	2000	1000	3118.86	8,0,1,1,1,34	8,29,0,1,2,3
1	10000	5000	3000	2000	1000	3081.73	7,0,1,0,2,33	7,29,0,1,1,3

For the problem with 5 customer classes, the decision variables are order quantity Q and base stock, S_i , $i = 1, 2, \dots, 5$. Different combinations of delay cost and lead time are considered and optimal parameters are calculated based on the algorithm presented in the last section. For the all the parameter sets, the demand rate, λ_i is assumed to be same and equal to 5. The results do not change when the demand rates are not equal. The input parameter values are as follows: $h = 250$ and $A = 100$, $L = 0.25$. It is also assumed that there is no stock out cost involved although the results are similar when there is a positive stock out cost. The first example in table 4.4 shows that the optimal order quantity, Q to be 7 and only two bins are needed to satisfy the demands from 5 classes; with base stocks 2 and 19 respectively. Demand from all the 5 classes are satisfied from the bin with base stock 19 first and when it runs out, only demand first 4 classes are satisfied from the other

bin with base stock 2. There is no need to reserve inventory for higher classes e.g. 1, 2 or 3.

Now, two examples are provided for the 5-class problem. For the first example $\lambda_j = 5$ for $\forall j=1,2,\dots,5$, $A = 100$, $h = 250$, $L = 0.5$, $\pi_j = 0$ $\forall j=1,2,\dots,5$, $p_1 = 5000, p_2 = 4000, p_3 = 3000, p_4 = 2000$ and $p_5 = 1000$. For the second example, all the parameter values are same as in the first example, except the delay cost. In the first example, the delay costs were spread apart in an equidistant fashion. In the second example, the delay costs were $p_1 = 10000, p_2 = 4000, p_3 = 3000, p_4 = 2000$ and $p_5 = 1000$.

There are several ways in which the 5-class can be collapsed into 3-classes problem. Clearly since $p_1 \geq p_2 \geq \dots \geq p_{n-1} \geq p_n$ and $\pi_1 \geq \pi_2 \geq \dots \geq \pi_{n-1} \geq \pi_n$, all the customer classes that are aggregated into a single class have to be contiguous. This reduces the number of possible ways to combine the 5-classes into 3 distinct classes. All the possible combinations are listed in Table 4.4.

In Table 4.4, for each (aggregated) class in the 3-class model, the penalty cost used for calculating the optimal rationing policy is the weighted average penalty cost (weighted by respective demand) for the underlying demand classes for that aggregate class. However, once the optimal solution is obtained from the 3-class model, the expected cost for implemented rationing policy is calculated using the actual penalty cost for the 5 classes. Table 4.5 reports the cost increase in percentage by implementing the solution obtained using a 3-class model.

Table 4.4 Cost increase from using the solution of an aggregated 3-class model for the 5-class problem

Example	Problem parameters	Expected Cost using 5-class model	Cost Increase from using an aggregated 3-class model (%)					
			(1)(2)(3,4,5)	(1)(2,3)(4,5)	(1)(2,3,4)(5)	(1,2)(3)(4,5)	(1,2)(3,4)(5)	(1,2,3)(4)(5)
1	$\lambda_j = 5$ for $j= 1,2,..5$, $A=100$, $h=250$, $L=0.5$, $\pi_j=0$, $j=1,2,..5$, $p_1=5000, p_2=4000, p_3=3000, p_4=$ 2000 and $p_5=1000$	2309.25	1.23%	0.27%	0%	0.27%	0%	0%
2	$\lambda_j = 5$ for $j= 1,2,..5$, $A=100$, $h=250$, $L=0.5$, $\pi_j=0$, $j=1,2,..5$, $p_1=10000, p_2=5000, p_3=3000, p_4$ $=2000$ and $p_5=1000$	2429.93	0.66%	0.47%	0.54%	0.47%	0.20%	0%

In similar manner, for the same two example problems, the cost increase when using an aggregated 2-class model was also investigated. The results for this case are reported in Table 4.5.

As can be seen from the two numerical examples, the worst case increase in the cost percentage for using a 3-class model was 1.23% and for using a 2-class model was 2.16%.

Some more experiments were carried out for an aggregated 3-class and 2-class model instead of a 5-class model. They are presented in Appendix C. Also a more exhaustive computational study was carried out to find out the impact of implementing a 2-class rationing policy for a 5-class model. The problem parameters used for this study were as follows: $\lambda_j = 5$ for $\forall j = 1, 2, 3$, $A = 0, 100, 200, \dots, 400, 500$, $h = 0, 50, 100, \dots, 450, 500$, $L = 0, 0.5, \dots, 2.5, 3.0$, $\pi_j = 0 \forall j = 1, 2, 3$, $p_1 = 3000, p_2 = 2000, p_3 = 1000$. The average cost increase is approximately 0% when using an aggregated 2-class model instead of a 3-class model. The results are presented in appendix D.

Table 4.5 Cost increase from using the solution of an aggregated 2-class model for the 5-class problem

Example	Problem parameters	Expected Cost using 5-class model	Cost Increase from using an aggregated 3-class model (%)			
			(1)(2,3,4,5)	(1,2)(3,4,5)	(1,2,3)(4,5)	(1,2,3,4)(5)
1	$\lambda_j = 5$ for $j = 1, 2, \dots, 5$, $A = 100$, $h = 250$, $L = 0.5$, $\pi_j = 0$, $j = 1, 2, \dots, 5$, $p_1 = 5000, p_2 = 4000, p_3 = 3000, p_4 = 2000$ and $p_5 = 1000$	2309.25	1.23%	1.23%	0.27%	0%
2	$\lambda_j = 5$ for $j = 1, 2, \dots, 5$, $A = 100$, $h = 250$, $L = 0.5$, $\pi_j = 0$, $j = 1, 2, \dots, 5$, $p_1 = 10000, p_2 = 5000, p_3 = 3000, p_4 = 2000$ and $p_5 = 1000$	2429.93	2.16%	0.66%	0.47%	0.54%

4.4 Summary

Most of the prior research for inventory rationing for service-differentiated customer classes considered only two demand classes. In this chapter, the exact cost expression for inventory rationing in a continuous review system for ' n ' customer classes is developed. An algorithm is developed to determine the optimal rationing policy that minimizes the expected cost. While the expected cost of a particular rationing policy can be expressed succinctly as a recursive function, evaluating them computationally becomes hard for problems involving large number of customer classes. Also, in practice it is cumbersome to implement rationing policies involving a large number of classes. Numerical studies revealed that implementing a 2-class or 3-class rationing policy for a 5-class model (and a 2-class rationing policy for 3-class model) does not increase the cost too much. The average cost increase was only 0.46 % and the maximum cost increase was only 1.33% for a 3-class model. For a 2-class model, the average cost increase was only 1.06 % and the maximum cost increase was only 3.3%

The expected cost for the rationing policy has been evaluated under the assumption that the backorders are cleared using the "threshold" clearing mechanism. A priority clearing mechanism, where the backorders for the higher priority class is always cleared first will obviously result in lower cost. However, based on the currently available mathematical tools, it does not seem possible to analytically evaluate the expression for the expected cost under the priority clearing mechanism. This could be a possible topic for future research.

Chapter 5

A new class of multi-bin policies for inventory rationing in a continuous review environment

5.1 Introduction

Inventory rationing is the most effective strategy when there are multiple customer classes requiring different service levels. Many industries use service differentiation to segregate critical and non-critical customer/orders. In military materials management, when the same spare part is requisitioned from various divisions, an inventory control policy is needed which can serve the most important demand first. Critical level inventory rationing, where the lower class demand is either backordered or not satisfied after on hand inventory reaches a pre-determined critical level (called rationing level) is the most widely studied and recommended policy in the literature for

this type of situations. Since the revenue generated from aftermarket service and sales of spare parts is huge, firms are keen to solve this type of problems with the objective of minimizing the inventory cost and maximizing service level provided to different customer classes.

Although critical level inventory rationing provides an inventory policy with a lower cost, as has been shown by many researchers (Deshpande et al. (2003)), its major shortcoming is that it provides a higher service level to the higher priority class at the expense of lower priority class. It is conceivable that inventory reserved for the higher class demand may not get fully utilized if there is only little demand for the higher class after rationing kicks in. The lower class demand therefore may not be fulfilled on time even when there is a large stock of inventory available. Thus the critical level rationing policy (Deshpande et al. (2003)'s rationing policy) might result in a low fill rate for the lower class demands. In this chapter, new classes of 2-bin and 3-bin policies are proposed for rationing of inventory in a system with two customer classes. The proposed policy effectively solves the problem of low service level to the lower class and is also more advantageous than the critical level rationing policy under certain conditions.

Over the years many researchers have worked on inventory rationing. Veinott (1965) was the first to propose inventory rationing but Topkis (1968) first analyzed, theoretically, the concept of inventory rationing and showed how to allocate inventory among multiple classes. Evans (1968) and Kaplan (1969) independently derived the same results as Topkis (1968), for two customer classes. Evans (1968) assumed that unsatisfied demand is lost whereas Kaplan (1969) assumed that it is backordered to the next cycle.

Recent works on inventory rationing in a continuous review environment include Deshpande et al. (2003) and Arslan et al. (2007). Deshpande et al. (2003) considered backordering for two customer classes and proposed a new clearing mechanism to clear the backorders when the replenishment arrives. Arslan et al. (2007) considered the same clearing mechanism as Deshpande et al. (2003) but considered ' n ' classes. They developed an equivalent serial-stage inventory system framework for solving the problem. In their problem formulation, the objective was to minimize the expected average inventory level subject to constraints on the pre-defined service levels for the different customer classes.

Deshpande et al. (2003) assumed that demand from the two customer classes follow Poisson process and is satisfied from inventory on a first-come-first-serve (FCFS) basis, as long as on hand inventory is equal to or above the rationing level. The inventory is replenished according to a (Q, r) policy where a replenishment order of quantity Q is placed whenever inventory position drops to the level r . Demand from the lower priority class (*class 2*) is backordered once on hand inventory falls below the rationing level but demand from higher priority class (*class 1*) is still satisfied as long as there is on-hand inventory. Demand from the higher priority class is backordered only after the inventory runs out. Backorders are cleared when the next replenishment arrives. If the replenishment quantity is not sufficient to clear all the backorders, they are cleared according to a mechanism termed as '*threshold*' clearing. The threshold clearing mechanism clears backorders as if the inventory level immediately after placing the replenishment order was $(r + Q)$ and thereafter rationing was carried out. This means all

the lower class demand that occurs after the placement of replenishment order, until the $(r + Q - K)^{\text{th}}$ demand arrival is cleared upon receipt of the replenishment, where K is the rationing level. The concept is depicted in Figure 5.1. The threshold clearing mechanism facilitates the analytical derivation of the probability distribution of inventory level easily as compared to other clearing mechanisms, such as priority clearing where backorders are cleared based on the priority class of the demand. For a (Q, r) policy, with a fixed rationing level, K , Deshpande et al. (2003) derived the expression for long run expected cost. Using the algorithm by Federgruen and Zheng (1992) the optimal (Q, r) policy for a fixed K can be determined, and the optimal rationing level can then be determined by exhaustive search over all possible rationing levels.

Although Deshpande et al. (2003)'s inventory rationing reserves inventory in anticipation of higher class demand arrival and consequently provides a very high service level to the higher class, numerical experiments show that the service level or fill rate for the lower priority customer class suffers dramatically. In order to solve this disparity of service levels, in this research a 2-bin inventory rationing policy is proposed which attempts to provide a higher service level to the lower priority class without compromising the service level of the higher priority class. In this policy, two separate bins, BIN_1 and BIN_2 , are kept for the two customer classes. Demand from each class is satisfied from its respective bins. The policy reserves inventory for both the classes separately as in a separate stock policy; however, demand from class 1 can be satisfied from BIN_2 (if available) once BIN_1 becomes empty, on a FCFS basis.

The lower priority class, however, cannot use BIN_1 when BIN_2 runs out. This mechanism ensures that the lower class demand is not backordered as long as there is inventory left in BIN_2 . At the same time the proposed policy does not undermine the priority of higher class, as it allows it to use the lower class' stock once its own stock runs out. Thus this policy gives some protection to the lower class while still

Figure 5.1 Deshpande et al. (2003)'s inventory rationing with threshold clearing

providing higher priority to class 1 demand. The policy is depicted in Figure 5.2.

Figure 5.2: Proposed 2 –bin policy,
(d_1 stands for class 1 demand, d_2 stands class 2 demand)

Apart from providing a higher service level to the lower priority class, the proposed policy outperforms the rationing policy by Deshpande et al. (in terms of expected cost) when holding cost is very low and lead time is short. The cost difference is sometimes as high as 7%. Numerical experiments show that the proposed rationing policy, discussed later in the chapter, increases the service level of the lower class by around 25% on average, with a slight increase in cost (around 3-4%).

Deshpande et al. (2003) policy can also be represented as a different type of 2-bin policy as depicted in Figure 5.3. In their policy, both class 1 and class 2 demands are first satisfied from BIN_2 . At the time ordering, of the inventory position $(r+Q)$, $(r+Q-K)$ is assigned to BIN_2 and the remaining K to BIN_0 , where K is the critical inventory level in their policy. Once BIN_2 is empty, class 2 demands are backordered and class 1 demands are satisfied from BIN_0 .

Figure 5.3 Deshpande et al.'s policy represented as a different 2 –bin policy, (d_1 stands for class 1 demand, d_2 stands class 2 demand)

A combined 3-bin policy which incorporates both the proposed 2-bin policy as well as Deshpande et al. (2003)'s rationing policy is also developed in this chapter. In this policy, there are three bins: BIN_0 , BIN_1 and BIN_2 . BIN_1 and BIN_2 are used in the same manner as in the proposed 2-bin policy. When both these bins become empty, only class 1 can be satisfied from BIN_0 and class 2 is backordered.

Figure 5.4 Proposed 3 –bin policy
(d_1 stands for class 1 demand, d_2 stands class 2 demand)

The remainder of this chapter is organized as follows. In section 5.2 the proposed, new 2-bin policy for inventory rationing is explained in detail. The model for this policy is formulated and the exact expression for the expected cost, including ordering, holding and penalty cost, is then developed. The algorithm to determine the optimal parameters is described next. In section 5.3 a discussion is provided on the development of the 3-bin policy which generalizes and incorporates both the proposed 2-bin policy and Deshpande et al.'s rationing policy. Detailed derivation of the cost expression for this policy is provided in the appendix B. Section 5.4 provides the results of a computational study comparing the proposed 2-bin rationing policy with the Deshpande et al. (2003)'s rationing policy. Concluding remarks and summary are provided in section 5.5.

5.2 The new two-bin policy

Before analyzing the proposed 2-bin policy, the model assumptions are provided. The notation used throughout this chapter is presented in Table 5.1.

The model being considered is a single item continuous review inventory system that follows a (Q, r) policy. Under the (Q, r) policy, a replenishment order of quantity Q is placed when the inventory position (inventory on hand + inventory on order - backorders) of the system drops to r or below. The demand for the item comes from two different customer classes that vary in their service requirement. The demand from each customer class, i , follows a Poisson process with intensity λ_i ; thus the total demand from two classes also follows a Poisson process with intensity $\lambda = \lambda_1 + \lambda_2$. Each replenishment order incurs a fixed ordering cost A . Holding cost is incurred at a rate of h per unit of on-hand inventory per unit time. The (fixed) lead time for replenishment is $L > 0$.

The demand that is not satisfied immediately upon arrival is backordered and incurs two type of penalty costs – delay cost per unit per time (p_i) for the duration the demand is not fulfilled and stock out cost per unit (π_i). The customers are prioritized according to the penalty costs they incur. It is assumed that class 1 demand has higher priority than class 2 demand; accordingly penalty cost for class 1 is higher than class 2 i.e. $p_1 \geq p_2$ and $\pi_1 \geq \pi_2$.

Table 5.1 Notation for chapter 5

Symbol	Meaning
I	Class index
λ_i	Demand rate for class i
λ	Total demand rate of all the classes
p_i	Time-weighted delay cost per unit for class i
π_i	Fixed stock-out cost per unit for class i
H	Holding cost per unit per unit time.
A	Fixed ordering cost per order
L	Lead time for replenishment arrival
Q	Order quantity
R	Reorder point
S_i	Base stock for class i
K_i	Rationing level for class i

The inventory rationing model with threshold clearing proposed by Deshpande et al. (2003) can be viewed a different type of two bin inventory system. A detailed explanation of this policy using a 2-bin framework, along with the threshold clearing mechanism is provided in appendix A.

The 2-bin inventory policy proposed in this research, works as follows. Two separate bins are kept for the two customer classes, BIN_1 and BIN_2 . As a (Q, r) policy is followed and demand follows a Poisson process, the total inventory position immediately after an order is placed is $r + Q = S_1 + S_2$ where S_1 and S_2 are the base stock allocated to BIN_1 and BIN_2 respectively. The inventory positions of the two bins are raised to S_1 and S_2 , respectively upon placement of an order. Let $IP_1(\tau_l)$ and $IP_2(\tau_l)$ be the inventory positions of two bins at the time of placing the l^{th} order at time τ_l . The quantity ordered

for each bin is $q_1 = S_1 - IP_1(\tau_1)$ and $q_2 = S_2 - IP_2(\tau_1)$, with $Q = q_1 + q_2$. Let $IL (BIN_i)$ and $IP (BIN_i)$ are the inventory level and inventory position of BIN_i . Whenever a demand comes from class i , it is satisfied from BIN_i ($i=1, 2$) provided it has positive on-hand inventory (i.e. $IL_i > 0$). If BIN_i inventory position, $IP (BIN_i) > 0$, it is assumed to be satisfied from BIN_i . If $IL (BIN_i)$ is also > 0 , then the demand is actually satisfied from BIN_i and $IL (BIN_i)$ is reduced by one. If the $IP (BIN_2)$ is ≤ 0 , and if the demand is from class 2 it is backordered and is cleared only upon arrival of $(l+1)^{th}$ replenishment order. The variable, BO_2^{CR} (number of backorders that will be carried forward and cleared in the $l+1^{th}$ replenishment cycle from BIN_2) is incremented by 1. If $IP (BIN_2) > 0$, but the on-hand inventory level, $IL (BIN_2) = 0$ and the demand is for class 2, it is backordered but cleared upon arrival of l^{th} replenishment order and the variable, BO_2^{CL} (number of backorders that will be cleared from BIN_2 by the l^{th} replenishment arrival) is incremented by 1. [Note that since inventory position is > 0 , there will be sufficient inventory upon arrival of l^{th} replenishment order].

If $IP (BIN_1) > 0$, but the actual on-hand inventory, $IL (BIN_1) = 0$ and the demand is for class 1, it is attempted to satisfy from BIN_2 . If $IL (BIN_2) > 0$, it is satisfied from BIN_2 and $IL (BIN_2)$ is reduced by 1. But as $IP (BIN_1) > 0$, it is assumed that the demand is satisfied from BIN_1 and $IP (BIN_1)$ is reduced by one and qt_{12} , the number of units borrowed from BIN_2 by BIN_1 is increased by one. These qt_{12} units will be returned to BIN_2 when replenishment arrives. Now if $IL (BIN_2)$ is also zero, then the demand is backordered and cleared from BIN_1 upon arrival of l^{th} replenishment order; correspondingly BO_1^{CL} is incremented by 1. In this case also, since $IP (BIN_1) > 0$, it is assumed that the demand is satisfied from BIN_1 and $IP (BIN_1)$ is reduced by one. Now

when $IP (BIN_1) \leq 0$ and $IP (BIN_2) > 0$, and the demand from class 1, it is assumed to be satisfied from BIN_2 . If $IL (BIN_2) > 0$, the demand will be satisfied immediately from BIN_2 . The demand is backordered and cleared upon arrival of l^{th} replenishment order, from BIN_2 if $IL (BIN_2) = 0$ and BO_2^{CL} is incremented by 1. The policy is represented as a flowchart in Figure 5.5.

When replenishment order arrives, after lead time, each bin is incremented by q_i units, $IL(BIN_i) = IL(BIN_i) + q_i$ where $i = 1, 2$. Then the units borrowed from BIN_2 by BIN_1 , qt_{12} are returned back to BIN_2 as much as possible. Finally the backorders BO_1^{CL} and BO_2^{CL} are cleared from BIN_1 and BIN_2 , respectively. BO_1^{CR} and BO_2^{CR} are carried forward and cleared upon arrival of the next replenishment.

In Deshpande et al.'s rationing policy, demand for both the classes are initially satisfied from BIN_2 . When BIN_2 is empty, class 2 demand is backordered and class 1 demand is satisfied from BIN_0 . BIN_0 is the reserve inventory that is made available only for class 1 after the BIN_2 is empty.

5.5 Flowchart for the proposed 2-bin policy

In the proposed 2-bin policy, the two demand classes are initially allocated to separate bins. Class 1 demand is initially satisfied from BIN_1 and class 2 demand from BIN_2 . When BIN_1 becomes empty, class 1 demand can still be satisfied from BIN_2 ; however when BIN_2 becomes empty, class 2 demand will be backordered even if inventory is available in BIN_1 .

When the proportion of class 1 demand in a cycle is high, the proposed policy does not keep as much reserve inventory exclusively for class 1 as does Deshpande et al.'s policy. While this can sometimes result in slightly higher expected cost, the service level provided for both the classes are good, unlike Deshpande et al. Table 5.2 illustrates how the proposed policy works using an example. In the example, $Q = 6$, $r = 2$, $S_1 = 5$ and $S_2 = 3$. Starting inventory level for BIN_1 and BIN_2 are respectively are $OH_1 = 5$ and $OH_2 = 3$.

Table 5.2 Example of the proposed policy with threshold clearing

Demand Sequence	Demand from class	IP_1	IP_2	OH_1	OH_2	Action	Total demand since last order placed	Item borrowed	BO cleared	BO carried forward
		5	3	5	3	Starting Condition				
1	1	4	3	4	3	Satisfy from BIN_1	1			
2	1	3	3	3	3	Satisfy from BIN_1	2			
3	2	3	2	3	2	Satisfy from BIN_2	3			
4	1	2	2	2	2	Satisfy from BIN_1	4			
5	1	1	2	1	2	Satisfy from BIN_1	5			
6	2	1	1	1	1	Satisfy from BIN_2 Order placed $q_1=4, q_2=2$	6			
		5	3	1	1	Raise the IP to S_i				
7	1	4	3	0	1	Satisfy from BIN_1	1			

Table 5.2 Example of the proposed policy with threshold clearing (continued)

Demand Sequence	Demand from class	IP_1	IP_2	OH_1	OH_2	Action	Total demand since last order placed	Item borrowed	BO cleared	BO carried forward
8	1	3	3	0	0	Loan from BIN_2	2	1		
9	2	3	2	0	0	Backorder but clear, when replenishment arrives, from BIN_2	3	1	$BO_2^{CL}=1$	
10	1	2	2	0	0	Backorder but clear, when replenishment arrives from BIN_1	4	1	$BO_1^{CL}=1$ $BO_2^{CL}=1$	
		2	2	4	2	Replenishment received		1	$BO_1^{CL}=1$ $BO_2^{CL}=1$	
		2	2	3	3	Return loaned items		0	$BO_1^{CL}=1$ $BO_2^{CL}=1$	
		2	2	2	2	Clear the backorder			$BO_1^{CL}=0$ $BO_2^{CL}=0$	
11	1	1	2	1	2	Satisfy from BIN_1	5			
12	1	0	2	0	2	Satisfy from BIN_1 Order placed $q_1=5, q_2=1$	6			
		5	3	0	2	Raise IP_i				

5.2.1 Evaluation of the policy cost

The following assumptions are made in the development of the model:

- 1) There are 2 customer classes, prioritized according to the penalty costs of not satisfying the demand from the class; class 1 is the higher priority class i.e. $p_1 \geq p_2$ and $\pi_1 \geq \pi_2$
- 2) The rate of demand arrival of different customer classes follow Poisson process with rate, λ_i ($i = 1, 2$); total demand rate is $\lambda = \sum_{i=1}^{i=2} \lambda_i$
- 3) There is fixed positive replenishment lead time L , $L > 0$.
- 4) Demand from different customer classes is satisfied on a FCFS basis.
- 5) Inventory holding cost is h per unit per unit time, $h > 0$.
- 6) An ordering cost A is incurred whenever an order is placed.
- 7) All the unsatisfied demand is backordered and satisfied once the next replenishment arrives, based on threshold clearing.
- 8) The unsatisfied demand incurs two types of penalty costs
  - i) Stock out cost per unit backordered (p_i)
  - ii) Delay cost per unit per period of delay (π_i)

The decision variables for the proposed policy are the order quantity Q and S_1 and S_2 , the base stock position of BIN_1 and BIN_2 , respectively. Note that the reorder point, $r = S_1 + S_2 - Q$. Let $C(Q, S_1, S_2)$ denote the long-run expected cost of the policy for a given Q , S_1 and S_2 . The expected cost comprises of three components – expected ordering cost, expected holding cost and expected penalty cost. Since the order quantity is Q for every

order, the expected ordering cost is $\frac{A\lambda}{Q}$. Therefore, the expected cost can be written as

$$C(Q, S_1, S_2) = \frac{A\lambda}{Q} + G(Q, S_1, S_2) \quad (5.1)$$

where $G(Q, S_1, S_2)$ is the expected holding and penalty cost for a given policy.

Given that a (Q, r) policy is being used, the total inventory position of the two bins follows a regenerative process. Let the l^{th} order be placed at time τ_l and the $l+1^{\text{th}}$ order be placed at τ_{l+1} . Let t be any point in the l^{th} replenishment cycle such that $\tau_l \leq t < \tau_{l+1}$ (see Figure 5.6 below). Knowing the probability distribution of inventory

Figure 5.6 Representation of Inventory position changes

position at time t (or total demand in the interval $[\tau_l, t)$), and the demand during period $[t, t+L)$, the cost $C(Q, S_1, S_2)$ can be evaluated. The order placed at τ_l would be received by $\tau_l + L$. Therefore the inventory represented by inventory position (S_1, S_2) at τ_l would be fully available at $\tau_l + L$, and this inventory position minus the demand during $[\tau_l, t+L)$ would accurately represent the inventory level/stock out level at time $t+L$. Also the backorders from the previous replenishment are already accounted for in the inventory position S_i (even though they may be cleared only at $\tau_l + L$) and can be ignored for

computing the inventory levels and cost at time $t+L$. However, because of the way the threshold clearing mechanism is designed, the sequence of demand arrivals during the period $[\tau_l, t+L)$ is important in arriving at the correct penalty and holding costs. Since a (Q, r) policy is used, and the demand follows a Poisson distribution, the total demand for the two classes, D_t during the period $[\tau_l, t)$ follows a uniform distribution with probability

$$u(D) = \frac{1}{Q}, \quad D = 0, 1, 2, \dots, Q-1 \quad (5.2)$$

The total demand for the two classes during the period $[t, t+L)$, D_L follows a Poisson distribution

$$f(D_L = j) = \frac{e^{-\lambda L} (\lambda L)^j}{j!} \quad j = 0, 1, \dots, \infty \quad (5.3)$$

Let D_{t+L} be the total demand during the period $[\tau_l, t+L)$. Then the probability distribution of D_{t+L} can be computed as a convolution of the uniform and the Poisson distribution as follows

$$f_{t+L}(j) = \Pr(D_{t+L} = j) = \sum_{i=0}^{\min(j, Q-1)} \left(\frac{1}{Q} \frac{e^{-\lambda L} (\lambda L)^{j-i}}{(j-i)!} \right) \quad j = 0, 1, \dots, \infty \quad (5.4)$$

Since the demands follow Poisson process, the probability that a particular demand arrives from class i ($i=1, 2$) is $\alpha_i = \frac{\lambda_i}{\lambda}$ and the number of i^{th} class demand out of a total demand of D follows a binomial distribution given by

$$B(D_i, D, \alpha_i) = {}^D C_{D_i} \alpha_i^{D_i} (1 - \alpha_i)^{D - D_i} \quad (5.5)$$

where D_i is the demand for the i^{th} class.

If k_1 and k_2 are the number of demand arrivals for class 1 and class 2 respectively during time period $[\tau_1, t+L)$, the probability distribution of k_1 and k_2 can be computed as a convolution of the Poisson distribution, given by (5.4) and the binomial distribution, given by (5.5). The joint probability distribution of demand for the two classes D_1 and D_2 during the time period $[\tau_1, t+L)$ is given by

$$\begin{aligned} g(k_1, k_2) &= \Pr(D_1 = k_1, D_2 = k_2) \\ &= f_{t+L}(k_1 + k_2) B(k_1, k_1 + k_2, \alpha_1) \end{aligned} \quad (5.6)$$

As class 1 demand can also be satisfied from BIN₂ if BIN₁ is empty, the expected holding and penalty costs depend on the sequence of demand arrivals. Let $l_1 = S_1 - D_1$ and $l_2 = S_2 - D_2$. When both l_1 and l_2 are greater than zero, they represent the inventory levels in BIN₁ and BIN₂ at time $(t+L)$. For all the other cases Table 5.3 provides a list of possibilities for the holding and penalty costs.

Table 5.3 Effect on holding and penalty cost for different cases of l_1 and l_2

Case	l_1	l_2		Description	Associated Cost
1	> 0	> 0		Demands for both classes are satisfied from their respective bins	Holding cost for both BIN ₁ & BIN ₂ .
2	> 0	≤ 0		Demand for class 1 is satisfied from its bin, BIN ₂ runs out, and class 2 demands are backordered	Holding cost for BIN ₁ & penalty cost for class 2/BIN ₂ .
3a	≤ 0	> 0	$ l_1 < l_2$	All demands are satisfied, class 1 demand is partly satisfied from BIN ₂ . BIN ₁ has zero inventory.	Holding cost for BIN ₂ .
3b			$ l_1 \geq l_2$	Both BIN ₁ and BIN ₂ have zero inventory. Demands for both classes are only partly satisfied.	Penalty cost for class 1/BIN ₁ and/or class 2/BIN ₂ , depending on sequence of demand arrivals.
4	≤ 0	≤ 0		Both BIN ₁ and BIN ₂ have zero inventory. Demands for both classes are only partly satisfied.	Penalty cost for class 1/BIN ₁ and/or class 2/BIN ₂ , depending on sequence of demand arrivals.

Case (1) and case (2) do not require class 1 to use BIN₂ because in both these cases, inventory in BIN₁ is enough to satisfy all the demand from class 1. In Case (3a), BIN₁ runs out of inventory but it can borrow all the extra units required from BIN₂. However, in cases (3b) and (4) the inventory available in BIN₂ is not enough to satisfy all the demands from the two classes. In these cases, the number of units of class 1 demands (consequently class 2 demands) that are backordered depends on the sequence of demand arrivals.

Let $G_j(Q, S_1, S_2)$ represent as the expected holding and penalty cost for case $j, j=1, 2, 3a, 3b, 4$.

Case 1

In this case, all demands for each class are satisfied from their respective bins and there are no backorders at time $(t+L)$. The units remaining after satisfying the demands incur holding cost; the expected penalty and holding cost for this case is

$$G_1(Q, S_1, S_2) = \sum_{l_1=1}^{\infty} \sum_{l_2=1}^{\infty} h(l_1 + l_2) g(S_1 - l_1, S_2 - l_2) \quad (5.7)$$

Case 2

The inventory in BIN₁ is enough to satisfy the demand from class 1 in the period $[t_\tau, t+L)$, but the inventory in BIN₂ is not enough to satisfy all the demand from class 2. Therefore there is inventory leftover in BIN₁, but penalty cost is incurred for some of the class 2 demand arrivals. The expected penalty and holding cost for this case is

$$G_2(Q, S_1, S_2) = \sum_{l_1=1}^{\infty} \sum_{l_2=-\infty}^0 (hl_1 + p_2 |l_2| + \lambda_2 \pi_2) g(S_1 - l_1, S_2 - l_2) \quad (5.8)$$

The units left in BIN₁ after satisfying all the demand from class 1 is l_1 and the holding cost is hl_1 . Since there are some backorders for class 2 they incur two types of penalty costs - delay cost (p_2) per unit per unit time and stock out cost (π_2) per unit. The long-run average stock out cost is $\lambda_2 \pi_2$ and delay cost is $|l_2| p_2$ where $|l_2| = -l_2$ denotes the number of backorders for class 2.

Case 3a

In this case the inventory available in BIN₁ is not enough to satisfy the demand from class 1 in the period $[t_\tau, t+L)$ but the inventory in BIN₂ is enough to satisfy all demand from class 2 during that period plus the demand from class 1 which could not be

satisfied from BIN₁. Hence, there are no backorders for either class at time $t+L$. The cost for this case is therefore

$$G_{3a}(S_1, S_2, Q) = \sum_{l_1=-\infty}^0 \sum_{l_2=|l_1|+1}^{\infty} h(l_1 + l_2) g(S_1 - l_1, S_2 - l_2) \quad (5.9)$$

As an example, suppose at time $t+L$, $l_1 = -5$ and $l_2 = 8$. This means that 5 units of class 1 cannot be satisfied from BIN₁, but after satisfying all demand from class 2, BIN₂ still has 8 units left. So, these units left in BIN₂ can be used to fill the unsatisfied class 1 demand and then the units leftover will be $(l_1 + l_2)$ and associated cost will be $h(l_1 + l_2)$.

Case 3b

In this case, the inventory available in BIN₂ after satisfying all the class 2 demand is not enough to satisfy the class 1 demand that spills over from BIN₁. Even though class 1 can satisfy a maximum number of $|l_2|$ units of demand from BIN₂, the actual number of units fulfilled from BIN₂ depends on the sequence of demand arrivals. When BIN₁ runs out, both class 1 and class 2 can use BIN₂ to satisfy their demand on an FCFS basis. This case is best illustrated with an example.

Let $S_1 = 7$ and $S_2 = 6$, $l_1 = -3$ and $l_2 = 1$. The demand for the two classes during the period $[t, t+L)$ is 10 ($S_1 - l_1$) and 5 ($S_2 - l_2$), respectively and total demand during this period is 15. Out of the 10 units of demand from class1, 7 would be satisfied from BIN₁. Of the total demand of 15 for both classes, a maximum of 13 ($S_1 + S_2$) can only be satisfied. The exact number of units of class 1 demand (in excess of 7) that would be

satisfied depend on the sequence of arrival of the class 1 and class 2 demands. There are three possible ways these demand could arrive.

a) Of the first 13 units of demand arrivals, 10 units are from class 1 and 3 units are from class 2 (maximum possible demand from class 1 is 10) and the last 2 (14th and 15th) demand arrivals are from class 2.

b) Of first 13 units of demand arrivals, 9 units are from class 1 and 4 units are from class 2. Of last 2 demand arrivals, 1 is from class 1 and the other from class 2.

c) Of first 13 units of demand arrivals, 8 units are from class 1 and 5 units are from class 2 and the last 2 demand arrivals are from class 1.

No other case is possible because demand from class 2 cannot exceed 5. In the first case, all the class 1 demands will be satisfied and there will be no backorder for class 1. Also 3 units of class 2 demands will be satisfied and 2 units of demand which arrived last will be backordered. In the second case, 1 unit of each class demand will be backordered. In the third case, all the class 2 demands are satisfied and 2 units of class 1 will be backordered.

The probability of the first case is $\frac{{}^{13}C_3 {}^2C_2}{{}^{15}C_5}$. The numerator denotes number of ways in which there can be 3 class 2 demands out of first 13 demands and 2 class 2 demands out of last 2 demands. The denominator denotes the number of ways in which there can be 5 class 2 demands out of total 15 demands. Similarly, the probability for the

second case is $\frac{{}^{13}C_4 {}^2C_1}{{}^{15}C_5}$ and third case is $\frac{{}^{13}C_5 {}^2C_0}{{}^{15}C_5}$. Since in the first case, there are no class 1 backorders and 2 units of class 2 backorders, the penalty cost is $0*p_1+2*p_2+ \lambda_1\pi_1 +\lambda_2\pi_2$. Similarly, for the second and third case it will be $1*p_1+1*p_2+ \lambda_1\pi_1+\lambda_2\pi_2$ and $2*p_1+0*p_2+ \lambda_1\pi_1$ respectively. So, the expected penalty cost is $\frac{{}^{13}C_3 {}^2C_2}{{}^{15}C_5}$

$$(0*p_1+2*p_2+\lambda_2\pi_2)+ \frac{{}^{13}C_4 {}^2C_1}{{}^{15}C_5} (1*p_1+1*p_2+\lambda_1\pi_1+\lambda_2\pi_2)+ \frac{{}^{13}C_5 {}^2C_0}{{}^{15}C_5} (2*p_1+0*p_2+ \lambda_1\pi_1).$$

For the general case with $l_1 \leq 0, 0 < l_2 \leq |l_1|$, the expected holding and penalty cost for case 3b is

$$G_{3b}(Q, S_1, S_2) = \sum_{l_1=-\infty}^0 \sum_{l_2=1}^{|l_1|} \frac{\sum_{j=0}^{\min(S_2-l_2, |l_1+l_2|)} {}^{S_1+S_2}C_{S_2-l_2-j} {}^{|l_1+l_2|}C_j \left\{ (|l_1+l_2|-j)p_1 + \right.}{S_1+S_2-l_1-l_2 {}^{S_2-l_2}C_{S_2-l_2}} \left. j p_2 + \lambda_1\pi_1 + \lambda_2\pi_2 \right\} g(S_1-l_1, S_2-l_2) \quad (5.10)$$

This is because out of total demand $(S_1+S_2-l_1-l_2)$ only (S_1+S_2) units of demand can be satisfied and the remaining $|l_1+l_2|$ units of demand will be backordered. Since class 1 demand can also use BIN₂ after finishing BIN₁, demand is satisfied from both classes on an FCFS basis and only the last $|l_1+l_2|$ demands will be backordered. Let j be number of the units of class 2 demand that arrives within the last $|l_1+l_2|$ demand arrivals. Then, out of first S_1+S_2 demands, S_2-l_2-j units would come from class 2. The probability of this

event is $\frac{S_1+S_2 C_{S_2-l_2-j}^{|l_1+l_2|} C_j}{S_1+S_2-l_1-l_2 C_{S_2-l_2}}$. The units of class 1 demand within last $|l_1+l_2|$ demands

will be $|l_1+l_2|-j$. So, the penalty cost is $\{(|l_1+l_2|-j)p_1 + jp_2 + \lambda_1\pi_1 + \lambda_2\pi_2\}$. However,

the total class 2 demands in the time interval is S_2-l_2 and hence j cannot be more than

either S_2-l_2 or $|l_1+l_2|$. This leads to the expression given in (4.10).

Case 4

In this case, the demand for both the classes in the period $[\tau_l, t+L)$ is greater than the allocated inventory position in the respective bins at time τ_l , i.e. $D_1 \geq S_1$ and $D_2 \geq S_2$.

While some the class 1 demand can possibly be satisfied from BIN₂, as in case 3b, there can be backorders for each class; the exact amount of backorders for each class would depend on the sequence of demand arrivals. Again, a numerical example is used to illustrate the different possibilities in this case.

Let $S_1 = 7$ and $S_2 = 8, l_1 = -3$ and $l_2 = -2$. Thus, the demands for the two classes in the period $[\tau_l, t+L)$ are $D_1 = 10$ and $D_2 = 10$ and total demand for the two classes is 20. Since $S_1 + S_2 = 15$, only 15 units of the 20 units of demand arrival can be satisfied. The remaining 5 units will have to be backordered. Since class 1 can also use BIN₂, all these available units can be used to satisfy class 1 demand whereas only 8 units can be used to satisfy class 2 demands. Out of the 15 units of demand that is satisfied, the number of units of class 1 and class 2 that is satisfied would depend on the sequence in which the

demands arrive. There are several possibilities this can happen, which are presented in Table 5.4.

Table 5.4 Different possibilities of demand arrival sequence for Case 4

Possibility	Class 1 out of first S_1+S_2 demand	Class 2 out of first S_1+S_2 demand	Class 1 out of last $ l_1+l_2 $ demand	Class 2 out of last $ l_1+l_2 $ demand
A	10 F(7,3)	5 F(0,5)	0 B(0)	5 B(5)
B	9 F(7,2)	6 F(0,6)	1 B(1)	4 B(4)
C	8 F(7,1)	7 F(0,7)	2 B(2)	3 B(3)
D	7 F(7,0)	8 F(0,8)	3 B(3)	2 B(2)
E	6 F(6,0)	9 F(0,8),B(1)	4 F(1,0),B(3)	1 B(2)
F	5 F(5,0)	10 F(0,8)B(2)	5 F(2,0),B(3)	0 B(2)

There can be no other possibility as the maximum demand arrival from either class is 10 each. The second and third columns in Table 5.4 represent the number of class 1 and class 2 demand, respectively, out of the first S_1+S_2 (15 in this example) units of demand and the last two columns represent the number of class 1 and class 2 demand, respectively, out of the last $|l_1+l_2|$ units of demand (5 in this example). $F(x, y)$ denotes that x and y units out of the demand is fulfilled from BIN_1 and BIN_2 for the two customer classes, respectively. Similarly, $B(z)$ indicates that z units of demand are backordered for

the particular customer class. In the first case, all units of the class 1 demand (7 units from BIN_1 and 3 units from BIN_2) are fulfilled and only 5 units of class 2 demand are fulfilled, leading to 5 class 2 backorders. In the second case, out of 10 units of class 1 demand, only 9 units (7 from BIN_1 and 2 from BIN_2) are fulfilled and 1 unit is backordered. Out of 10 units of class 2 demands only 6 units are fulfilled and the remaining 4 backordered. Similar logic holds for cases (c) and (d).

Case e, however, is different as in this case, 9 units of demand are from class 2 but there are only 8 units designated for it. Since class 2 cannot use BIN_1 to satisfy its demand, only 8 units of class 2 demands can be satisfied and the 9th class 2 demand cannot be satisfied, in spite of being the first 15 demand arrivals. Thus there would be 2 units of backorders for class 2; the 10th Class 2 demand arrives after BIN_2 is already empty. BIN_1 can be used to satisfy 7 units of class 1 demand and 3 units of class 1 demand will be backordered as there are no units left in the BIN_2 when these demands arrive. It should be noted that the 7th class 1 demand is satisfied in spite of the fact that it arrives after the first 15 demands.

Case (f) is similar to case (e). 7 units of class 1 demand and 8 units of class 2 demands are satisfied, respectively, from BIN_1 and BIN_2 . So, 3 units of class 1 demand and 2 units of class 2 demand will be backordered. For cases (d), (e) and (f) all the BIN_2 stock is used only to fulfill class 2 demands. For all these cases, there would be 3 and 2 units of backorders for class 1 and class 2, respectively.

As explained in case (3b), the probability of the first case is $\frac{{}^{15}C_5 {}^5C_5}{{}^{20}C_{10}}$. Similarly, the probabilities for the other cases are $\frac{{}^{15}C_6 {}^5C_4}{{}^{20}C_{10}}$, $\frac{{}^{15}C_7 {}^5C_3}{{}^{20}C_{10}}$, $\frac{{}^{15}C_8 {}^5C_2}{{}^{20}C_{10}}$, $\frac{{}^{15}C_9 {}^5C_1}{{}^{20}C_{10}}$ and $\frac{{}^{15}C_{10} {}^5C_0}{{}^{20}C_{10}}$, respectively. Since in the first case, there are no class 1 backorders and 5 class 2 backorders, the penalty cost involved is $0*p_1 + 5*p_2 + \lambda_2\pi_2$. Similarly, for the other cases the penalty cost will be $1*p_1 + 4*p_2 + \lambda_1\pi_1 + \lambda_2\pi_2$, $2*p_1 + 3*p_2 + \lambda_1\pi_1 + \lambda_2\pi_2$, $3*p_1 + 2*p_2 + \lambda_1\pi_1 + \lambda_2\pi_2$, $3*p_1 + 2*p_2 + \lambda_1\pi_1 + \lambda_2\pi_2$ and $3*p_1 + 2*p_2 + \lambda_1\pi_1 + \lambda_2\pi_2$, respectively. It can be noted that the penalty costs for the last three cases are the same as all the BIN₂ stock is used to fulfill class 2 demands only. Penalty cost remains unchanged once the class 2 demands during the first $(S_1 + S_2)$ demands are equal to or more than S_2 . So, the expected penalty cost is $\frac{{}^{15}C_5 {}^5C_5}{{}^{20}C_{10}} (0*p_1 + 5*p_2 + \lambda_2\pi_2) + \frac{{}^{15}C_6 {}^5C_4}{{}^{20}C_{10}} (1*p_1 + 4*p_2 + \lambda_1\pi_1 + \lambda_2\pi_2) + \frac{{}^{15}C_7 {}^5C_3}{{}^{20}C_{10}} (2*p_1 + 3*p_2 + \lambda_1\pi_1 + \lambda_2\pi_2) + (\frac{{}^{15}C_8 {}^5C_2}{{}^{20}C_{10}} + \frac{{}^{15}C_9 {}^5C_1}{{}^{20}C_{10}} + \frac{{}^{15}C_{10} {}^5C_0}{{}^{20}C_{10}}) (3*p_1 + 2*p_2 + \lambda_1\pi_1 + \lambda_2\pi_2)$.

In general, the cost for case 4, $G_4(Q, S_1, S_2)$ is given by

$$\sum_{j=\max(|l_2|+1, -S_1-l_2)}^{\min(S_2-l_2, |l_1+l_2|)} S_1+S_2 C_{S_2-l_2-j} |l_1+l_2| C_j \left\{ (|l_1+l_2|-j)p_1 + \right. \\ \left. jp_2 + \lambda_1\pi_1 + \lambda_2\pi_2 \right\} + \sum_{l_1=-\infty}^0 \sum_{l_2=-\infty}^0 \frac{\sum_{j=\max(0, -S_1-l_2)}^{|l_2|} S_1+S_2 C_{S_2-l_2-j} |l_1+l_2| C_j (|l_1|p_1 + |l_2|p_2 + \lambda_1\pi_1 + \lambda_2\pi_2)}{S_1+S_2-l_1-l_2 C_{S_2-l_2}} g(S_1-l_1, S_2-l_2) \quad (5.11)$$

The first term in the numerator represents the case when the number of class 2 demand arrivals, j , in the last $|l_1 + l_2|$ arrivals is $|l_2| + 1$ or more. Clearly, j cannot be more than $\min(S_2 - l_2, |l_1 + l_2|)$ as in cases (a), (b), (c) above. Also if $D_2 > S_1 + S_2$, then at least

$D_2 - (S_1 + S_2)$ units of class 2 demand have to occur within the last $|l_1 + l_2|$ demand arrivals. Therefore, $j > D_2 - (S_1 + S_2) = -S_1 - l_2$. In these cases, part of class 1 demand is satisfied from BIN₂ and, as a result, class 2 has more than $|l_2|$ backorders. The second term represents the case when BIN₂ is not used at all for satisfying class 1 demand i.e. when the number of class 2 demands, j , that arrive during the last $|l_1 + l_2|$ demands are $|l_2|$ or less, or the number of class 2 demands during the first $(S_1 + S_2)$ demand arrivals is S_2 or more. As already discussed when $D_2 > S_1 + S_2$, $j > -S_1 - l_2$, therefore, minimum value of j for this second summation is $\max(0, -S_1 - l_2)$. When all class 2 demands, D_2 , arrive within the first $S_1 + S_2$ demand arrivals (which is only possible when $D_2 < S_1 + S_2$), then $j=0$. This implies that summation for the second term in numerator of (5.11) is from $j = \max(0, -S_1 - l_2)$ to $j = |l_2|$.

Finally, using equations (5.7) to (5.11), the expected total cost can be obtained as

$$C(Q, S_1, S_2) = \frac{\lambda A}{Q} + G_1(Q, S_1, S_2) + G_2(Q, S_1, S_2) + G_{3a}(Q, S_1, S_2) + G_{3b}(Q, S_1, S_2) + G_4(Q, S_1, S_2) \quad (5.12)$$

Equation (5.12) gives the exact expression for the expected total cost of the proposed 2-bin policy.

5.2.2 Solution algorithm

Now, a method is developed to derive the optimal policy parameters Q , S_1 and S_2 , which minimize the total expected cost for the proposed 2-bin policy. Deshpande et al.

(2003) showed that for their rationing policy, for a fixed value of S_1 ; S_2 and Q can be determined using an algorithm similar to Federgruen and Zheng (1992) by exploiting the convexity of the underlying cost terms. However, they still need to enumerate the cost exhaustively for all possible values of S_2 . Unfortunately for the proposed 2-bin policy, no such structural results on the cost function could be obtained to facilitate application of efficient algorithms like that of Federgruen and Zheng (1992). Therefore, exhaustive enumeration of the cost function $C(Q, S_1, S_2)$ across all possible values is required to determine the optimal inventory policy. However, efficiency of the search can be improved by developing bounds on the decision variables Q , S_1 and S_2 .

5.2.2.1 Bounds on $\bar{S} = S_1 + S_2$ and Q

Clearly a lower bound $\bar{S}_{L1}$ on $\bar{S}$ can be obtained by solving a single customer class problem with demand $\lambda = \sum_{i=1}^2 \lambda_i$, $\pi = \pi_2$ and $p = p_2$. The corresponding Q_U will also be an upper bound on Q . An upper bound $\bar{S}_U$ on $\bar{S}$ and a lower bound Q_{L1} on Q can be obtained by solving a single customer class problem with $\lambda = \sum_{i=1}^2 \lambda_i$, $\pi = \pi_1$ and $p = p_1$.

If this was treated as a pure separate stock policy (i.e. BIN₂ can only be used to satisfy demand from class 2), then the problem reduces to a joint replenishment problem (JRP) and the (Q, S) policy developed by Pantumsinchai (1992) can be used to solve it. As inventory of each BIN_{*i*} is segregated to fulfill only a specific demand class, the value

of $S_1 + S_2$ of this JRP solution should serve as an upper bound $\overline{S_{U2}}$ on $\overline{S}$ of the original problem. One can then choose the upper bound on $\overline{S}$ as $\overline{S_U} = \min(\overline{S_{U1}}, \overline{S_{U2}})$. The inventory holding and penalty costs in the JRP would be higher as each bin only fulfills one class of demand. Therefore, the Q corresponding to the JRP solution would be a lower bound Q_{L2} on the Q . The lower bound on Q can then be $\overline{Q_L} = \max(Q_{L1}, Q_{L2})$.

The algorithm for determining the optimal policy then involves searching for $C(Q, S)$ within the lower and upper bounds of Q and $\overline{S}$.

5.3 A general 3-bin policy

In the previous section, a 2-bin policy was developed where the higher priority class 1 can use the bin designated for the lower priority class, BIN_2 when inventory in its own bin runs out of stock. In this section, a new 3-bin policy is proposed which has the characteristics of the 2-bin policy presented in the last section as well as that of the Deshpande et al. (2003)'s inventory rationing policy. The 3-bin policy works as follows:

- There are 3 bins, BIN_0 , BIN_1 and BIN_2 with base stock positions S_0 , S_1 and S_2 respectively.
- Ordering is done according to a (Q, r) policy with quantity $r + Q = S_0 + S_1 + S_2$, allocated to the base stock levels in the 3 bins.
- Demand from class 2 (lower priority class) is fulfilled only from BIN_2 .
- Demand from class 1 (higher priority class) is fulfilled from BIN_1 first. If BIN_1 is empty then the demand is fulfilled from BIN_2 (same as our proposed 2-bin policy).

When BIN_2 also becomes empty, then demand is satisfied from BIN_0 . BIN_0 is thus reserved to fulfill only class 1 demand after both BIN_1 and BIN_2 have run out of stock.

- Demand from class 2 is backordered when BIN_2 is empty and demand from class 1 is backordered when all the three bins are empty.

Note that when $S_0 = 0$, this 3-bin policy reduces to the proposed 2-bin policy presented in the previous section. When $S_I = 0$, this 3-bin policy transforms into the inventory rationing policy proposed in Deshpande et al. (2003).

The detailed derivation of the expected cost of the 3-bin policy is provided in the appendix B. All the assumptions made to develop the expected cost function are similar to that of 2-bin policy. In all the numerical experiments that was carried out, the optimal 3-bin policy either had $S_0 = 0$ or $S_I = 0$. That is the optimal 3-bin policy was either a policy belonging to the class of inventory rationing policies proposed by Deshpande et al. (2003) or the class of 2-bin policies proposed in this research. Hence, no numerical results are provided for this 3-bin policy and only the proposed 2-bin rationing policy is compared with the inventory rationing policy of Deshpande et al. (2003) in the next section. However note that by implementing the 3-bin policy, results which are the better of the proposed 2-bin policy and Deshpande et al.'s policy can always be achieved.

5.3.1 Cost calculation for the 3-bin policy

Let $l_1 = S_1 - D_1$ and $l_2 = S_2 - D_2$, where D_1 and D_2 are the demand, for class 1 and class 2, respectively during the period $[\tau_l, t + L)$. Then, the joint probability distribution is D_1 and D_2 is given by (5.6).

In a manner similar to the derivation of the cost expression for the 2-bin policy, the expected penalty and holding cost, $G(Q, S_0, S_1, S_2)$ for this 3-bin policy can also be derived and is given next.

$$\begin{aligned}
 G(S_0, S_1, S_2, Q) &= \sum_{l_1=1}^{S_1} \sum_{l_2=1}^{S_2} h(l_1 + l_2 + S_0) g(S_1 - l_1, S_2 - l_2) \\
 &+ \sum_{l_1=1}^{S_1} \sum_{l_2=-\infty}^0 [h(l_1 + S_0) - p_2 l_2 + \pi_2 \lambda_2] g(S_1 - l_1, S_2 - l_2) \\
 &+ \sum_{l_1=-S_2+1}^0 \sum_{l_2=-l_1+1}^{S_2} h(l_1 + l_2 + S_0) g(S_1 - l_1, S_2 - l_2) \\
 &+ \sum_{l_1=-\infty}^0 \sum_{l_2=0}^{\min(S_2, -l_1)} g(S_1 - l_1, S_2 - l_2) \left[\sum_{k=0}^{\min\{(S_2-l_2), (|l_1|-l_2-S_0)\}} \{p_2 k + p_1(|l_1| - l_2 - k - S_0) + \pi_1 \lambda_1 + \pi_2 \lambda_2\} \frac{S_1+S_2 C_{S_2-l_2-k} |l_1|-l_2 C_k}{S_1+S_2-l_1-l_2 C_{S_2-l_2}} \right. \\
 &+ \left. \sum_{k=\max(0, |l_1|-l_2-S_0+1)}^{\min\{(S_2-l_2), (|l_1|-l_2)\}} \{p_2 k + \pi_2 \lambda_2 + h(S_0 - (|l_1| - l_2 - k))\} \frac{S_1+S_2 C_{S_2-l_2-k} |l_1|-l_2 C_k}{S_1+S_2-l_1-l_2 C_{S_2-l_2}} \right] \\
 &+ \sum_{l_1=-\infty}^{-S_0} \sum_{l_2=-\infty}^{-1} g(S_1 - l_1, S_2 - l_2) \left[\sum_{k=|l_2|+1}^{\min\{(S_2-l_2), (|l_1+l_2|-S_0)\}} \{p_2 k + p_1(|l_1+l_2| - k - S_0) + \pi_1 \lambda_1 + \pi_2 \lambda_2\} \frac{S_1+S_2 C_{S_2-l_2-k} |l_1+l_2| C_k}{S_1+S_2-l_1-l_2 C_{S_2-l_2}} \right. \\
 &+ \left. \sum_{\substack{k=\max(0, |l_1+l_2|-S_0+1) \\ k \neq |l_2|}}^{\min\{(S_2-l_2), |l_1+l_2|\}} \{p_2 k + \pi_2 \lambda_2 + h(S_0 - |l_1+l_2| - k)\} \frac{S_1+S_2 C_{S_2-l_2-k} |l_1+l_2| C_k}{S_1+S_2-l_1-l_2 C_{S_2-l_2}} \right. \\
 &+ \left. \{p_2 |l_2| + p_1(|l_1| - S_0) + \pi_1 \lambda_1 + \pi_2 \lambda_2\} \frac{\sum_{\rho=S_2}^{\min\{(S_1+S_2), (S_2-l_2)\}} S_1+S_2 C_{\rho} |l_1+l_2| C_{S_2-l_2-\rho}}{S_1+S_2-l_1-l_2 C_{S_2-l_2}} \right] \\
 &+ \sum_{l_1=-S_0+1}^0 \sum_{l_2=-\infty}^{-1} g(S_1 - l_1, S_2 - l_2) \left[\sum_{k=|l_2|+1}^{\min\{(S_2-l_2), (|l_1+l_2|-S_0)\}} \{p_2 k + p_1(|l_1+l_2| - k - S_0) + \pi_1 \lambda_1 + \pi_2 \lambda_2\} \frac{S_1+S_2 C_{S_2-l_2-k} |l_1+l_2| C_k}{S_1+S_2-l_1-l_2 C_{S_2-l_2}} \right. \\
 &+ \left. \sum_{\substack{k=\max(0, |l_1+l_2|-S_0+1) \\ k \neq |l_2|}}^{\min\{(S_2-l_2), |l_1+l_2|\}} \{p_2 k + \pi_2 \lambda_2 + h(S_0 - |l_1+l_2| - k)\} \frac{S_1+S_2 C_{S_2-l_2-k} |l_1+l_2| C_k}{S_1+S_2-l_1-l_2 C_{S_2-l_2}} + \right. \\
 &+ \left. \{p_2 |l_2| + h(S_0 - |l_1|) + \pi_2 \lambda_2\} \frac{\sum_{\rho=S_2}^{\min\{(S_1+S_2), (S_2-l_2)\}} S_1+S_2 C_{\rho} |l_1+l_2| C_{S_2-l_2-\rho}}{S_1+S_2-l_1-l_2 C_{S_2-l_2}} \right] \tag{5.13}
 \end{aligned}$$

5.4 Numerical results

The performance of the two policies, the proposed 2-bin policy and the rationing policy by Deshpande et al. (2003), is now compared for several problem parameter combinations. Firstly, the optimal cost of the proposed 2-bin policy is compared with the optimal cost of Deshpande et al. (2003)'s policy. This is followed by a comparison of the service level the optimal policies provide to each class.

5.4.1 Comparison of expected total cost

First, a set of experiments were carried out by varying the holding cost over a wide range, starting with a very low value. The rest of the problem parameters were kept fixed as follows: demand rate, $\lambda_1 = \lambda_2 = 5$, delay cost, $p_1 = 500$, $p_2 = 100$, stock out cost, $\pi_1 = \pi_2 = 0$ ordering cost, $A = 500$ and lead time, $L = 1$. From Table 5.5, it can be seen that the proposed 2-bin policy performs better than Deshpande et al. when the holding cost is very low. It can be noted that when $h = 10$, the expected cost of the proposed policy was 7% lower than that of Deshpande et al. As the holding cost rate increases the benefit of the proposed policy decreases and for $h \geq 30$, the cost of Deshpande et al.'s policy is lower, although the maximum cost difference is only 3%. In this experiment, stock-out

Table 5.5 Comparison of expected cost of the two policies

$$[\lambda_1 = \lambda_2 = 5, p_1 = 500, p_2 = 100, \pi_1 = \pi_2 = 0, A = 500, L = 1]$$

h	Results for optimal Deshpande et al.policy (D)		Results for optimal 2-bin proposed Policy (P)		Cost difference [(D - P)/P] *100
	TC	Parameters	TC	Parameters	
		Q,r,K		Q,S ₁ ,S ₂	
10	381.857	20 ,11, 1	355.505	25,11,25	7.41
20	485.188	20 ,9, 2	482.623	25,12,22	0.53
30	576.585	20 ,8, 2	588.262	21,11,18	-1.98
40	658.061	19 ,7, 2	674.17	18,10,16	-2.39
50	728.17	17 ,7, 2	748.776	17,10,14	-2.75
60	790.816	17 ,6, 2	811.811	16,9,14	-2.59
70	844.676	16 ,6, 2	869.271	15,9,13	-2.83
80	893.705	15 ,6, 2	921.838	14,8,13	-3.05
90	938.993	14 ,6, 2	970.447	14,8,12	-3.24
100	981.942	13 ,6, 2	1013.67	14,8,12	-3.13

cost, π_1 and π_2 was kept at zero, but if stock out costs were included, the nature of results does not change much. Table 5.6 provides the cost comparison for two policies when positive values of π_1 and π_2 are considered. For the results reported in Table 5.6, $\pi_1 = 100$ and $\pi_2 = 50$. In this case also the benefit of the proposed policy is high when holding cost is low. When holding cost is high, Deshpande et al. (2003)'s policy begins to perform

better, however the cost is different by only a few percentage points. Since the proposed policy results in a slightly higher inventory, its superior cost performance at low values of h is to be expected.

Table 5.6 Comparison of expected cost of the two policies

$$[\lambda_1 = \lambda_2 = 5, p_1 = 500, p_2 = 100, \pi_1 = 100, \pi_2 = 50, A = 500, L = 1]$$

h	Results for optimal Deshpande et al.policy (D)		Results for optimal 2-bin proposed Policy (P)		Cost difference [(D - P)/P] *100
	TC	Parameters	TC	Parameters	
		Q,r,K		Q,S ₁ ,S ₂	
10	402.68	20 ,13, 1	374.60	25,4,34	7.50
20	528.39	20 ,11, 1	521.03	25,7,29	1.41
30	641.06	20 ,10, 1	645.54	21,8,23	-0.69
40	743.63	18 ,10, 1	749.28	18,7,21	-0.75
50	831.89	17 ,9, 1	841.75	16,6,20	-1.17
60	911.24	16 ,9, 1	921.56	16,7,18	-1.12
70	983.98	16 ,8, 2	995.21	15,7,17	-1.13
80	1048.75	15 ,8, 2	1063.68	14,7,16	-1.40
90	1109.10	14 ,8, 1	1126.43	14,7,15	-1.54
100	1165.64	13 ,8, 1	1183.86	13,7,14	-1.54

When the holding cost rate is higher and lead time for replenishment, L is lower, Deshpande et al. (2003)'s policy resulted in lower cost than the proposed 2-bin policy, though the maximum cost difference was still less than 4.1%. The results for this set of

Table 5.7 Comparison of expected cost of the two policies parameters

$$[\lambda_1 = \lambda_2 = 10, p_2 = 600, \pi_1 = \pi_2 = 0, A = 100, L = 0.25, h = 250]$$

P_1	Results for optimal Deshpande et al.policy (D)		Results for optimal 2-bin proposed Policy (P)		Cost difference [(D - P)/P] *100
	Optimal Cost	Optimal Parameters	Optimal Cost	Optimal Parameters	
		Q,r,K		Q,S ₁ ,S ₂	
600	1167.2	6,3,0	1167.2	6,0,9	0.00
1200	1308.02	6,4,0	1308.02	6,0,10	0.00
1800	1397.19	6,4,1	1399.23	6,1,9	-0.15
2400	1458.49	6,4,1	1481.46	5,1,9	-1.55
3000	1505.42	5,5,1	1534.57	6,1,10	-1.90
3600	1538.85	6,5,1	1581.21	6,1,10	-2.68
4200	1570.22	6,5,1	1626.53	6,1,10	-3.46
4800	1601.59	6,5,2	1670.22	5,1,10	-4.11
5400	1630.53	6,5,2	1696.58	5,1,10	-3.89
6000	1651	6,5,2	1721.87	5,4,7	-4.12

experiments are given in Table 5.7 and Table 5.8. For these set of problems, $\lambda_1 = \lambda_2 = 10$, $p_2 = 600$, $\pi_1 = \pi_2 = 0$ ordering cost, $A = 100$, $h = 250$ and lead time, $L = 0.25$. p_1 was varied between 600 to 6000. In Table 5.8, the results are reported for the case where $p_1 = 6000$ and p_2 was varied from 600 to 6000.

From the results in Table 5.7 and Table 5.8, it is evident that the proposed policy does not outperform Deshpande et al.'s policy in terms of cost when the lead time is shorter and the holding cost is higher. It also shows that when the delay costs of the two classes are very close, the two policies perform the same and in those cases there is no need to keep separate inventory for the two classes. Also as the difference between delay costs increases, the cost difference between the policies increases. However, the maximum cost difference is still less than 4.1%.

To summarize, the proposed 2-bin policy is beneficial in terms of expected cost when holding cost rate is very low and the lead times are relatively longer. But in a majority of the cases the rationing policy of Deshpande et al. (2003) outperforms the proposed 2-bin policy in terms of expected cost though the maximum cost difference in the experiments was only 4.1%.

Moreover, as shown in the results in the next section, the proposed policy provides higher service levels than Deshpande et al. (2003), especially for the lower demand class. The service level for the two classes is determined based on the PASTA property which states that for Poisson demand the service level (fill rate) is equal to the probability that the steady state inventory level is positive. For the Deshpande et al.'s

Table 5.8 Comparison of expected cost of the two policies parameters

$$[\lambda_1 = \lambda_2 = 10, p_1 = 6000, \pi_1 = \pi_2 = 0, A = 100, L = 0.25, h = 250]$$

P ₂	Results for optimal Deshpande et al. policy (D)		Results for optimal 2- bin proposed Policy (P)		Cost difference [(D - P)/P] *100
	Optimal Cost	Optimal Parameters	Optimal Cost	Optimal Parameters	
		Q,r,K		Q,S ₁ ,S ₂	
600	1651	6,5,2	1721.87	5,4,7	-4.12
1200	1723.19	5,6,1	1755.72	5,2,9	-1.85
1800	1777.41	6,6,1	1780.18	6,1,11	-0.16
2400	1803.45	6,6,0	1803.45	6,0,12	0.00
3000	1826.72	6,6,0	1826.72	6,0,12	0.00
3600	1849.98	6,6,0	1849.98	6,0,12	0.00
4200	1873.25	6,6,0	1873.25	6,0,12	0.00
4800	1887.12	5,7,0	1887.12	5,0,12	0.00
5400	1899.71	5,7,0	1899.71	5,0,12	0.00
6000	1912.31	5,7,0	1912.31	5,0,12	0.00

policy the fill rates for the two demand classes SL_1 and SL_2 , respectively can be calculated as

$$SL_1 = Pr(IL > 0)$$

$$SL_2 = Pr(IL \geq K), \text{ where } K \text{ is the rationing level.}$$

where IL is the steady state inventory level.

For the proposed 2-bin policy the fill rates for the two demand classes can be calculated as follows:

$$SL_1 = Pr(l_1 > 0) + Pr(l_1 \leq 0 \& l_2 > 0) \quad (5.14)$$

$$SL_2 = Pr(l_2 > 0) \quad (5.15)$$

where l_1 and l_2 are as defined in section 5.2.

For the problem set whose expected cost was reported in Table 5.7, the service level comparison is provided in Table 5.9.

It is clear from the Table 5.9 that the service levels the proposed policy provides are higher than that of Deshpande et al. (2003)'s policy. The service level for class 2 provided by the proposed policy is as much as 28% higher without compromising the class 1 service level; in fact the service level of class 1 is also higher albeit to a lesser extent. The computational results show that even though the proposed policy is slightly costlier than the rationing policy, it provides a much higher service level, especially for the lower customer class and especially when $p_1 \gg p_2$. The average service levels for above experiments for Deshpande et al.'s policy are 80.4% and 69% for class 1 and class 2 respectively; whereas the average service levels for proposed 2-bin policy are 82.9% and 82.5% respectively.

Table 5.9 Comparison of service levels for the two policies

$$[\lambda_1 = \lambda_2 = 10, \pi_1 = \pi_2 = 0, A = 100, L = 0.25, h = 250]$$

p ₁	p ₂	SL difference between by proposed policy and Deshpande et al.s policy (%) (SL ^P -SL ^D)	
		SL ₁	SL ₂
600	600	0.00	0.00
1200	600	0.00	0.00
1800	600	0.01	11.70
2400	600	6.49	18.17
3000	600	2.63	13.17
3600	600	0.00	9.09
4200	600	0.00	9.09
4800	600	4.79	25.59
5400	600	4.79	25.59
6000	600	5.59	22.79

For the problem set whose expected cost was reported in Table 5.8, the service levels comparison is provided in Table 5.10. The results in Table 5.10 are similar to Table 5.9. The average service levels for above experiments for Deshpande et al.'s policy are 91.9% and 88.5% for class 1 and class 2 respectively; whereas the average service levels for proposed 2-bin policy are 92.6% and 92.2% respectively. It can be seen that the increase in service level for class 2 is much higher, and the service level of class 1 also increases. Note that when there is no cost difference between the two policies, the service

Table 5.10 Comparison of service levels for the two policies

$$[\lambda_1 = \lambda_2 = 10, \pi_1 = \pi_2 = 0, A = 100, L = 0.25, h = 250]$$

delay1 P ₁	delay2 P ₂	SL difference between by proposed policy and Deshpande et al.s policy (%) (SL ^P -SL ^D)	
		SL ₁	SL ₂
6000	600	5.59	22.79
6000	1200	0.04	7.31
6000	1800	0.001	6.31
6000	2400	0	0
6000	3000	0	0
6000	3600	0	0
6000	4200	0	0
6000	4800	0	0
6000	5400	0	0
6000	6000	0	0

levels are same as well. These represent cases where rationing is not required for both policies and a single stock is sufficient to serve the two classes. A Pareto chart is presented in Figure 5.7 to capture the cost and service levels provided by the two policies for the problem presented in Table 5.7 and 5.9.

Figure 5.7 Pareto chart showing cost and service level

5.5 Summary

In this chapter, a new class of 2-bin inventory rationing policies is developed for a continuous review system with two service differentiated customer/demand classes. The expression for the expected cost of the policy under a threshold clearing mechanism is derived; a search algorithm based on exhaustive enumeration within the variable bounds is then used to determine the optimal policy. When the holding cost rate is low and the replenishment lead-time is relatively higher, the proposed 2-bin policy gives a lower expected cost. For most of the problems in the numerical experiments, with larger holding cost, the expected cost of the proposed policy was higher, albeit by less than

4.1%. However, the proposed policy provides a much higher service level (especially for the lower customer class) compared to Deshpande et al.

A more general 3-bin policy which incorporates both the proposed 2-bin policy as well as the Deshpande et al. policy is also developed. If cost minimization is the objective, then the 3-bin policy will always perform as well or better than Deshpande et al. policy. The proposed 2-bin policy results in a slightly higher cost than Deshpande et al. policy but provides much higher service levels.

Chapter 6

Conclusions

Service differentiation is very common in industries these days especially in MRO industry where providing higher service at lowest cost is the key. In today's competitive business environment it is necessary to look for newer strategies to compete and many firms are using inventory rationing to achieve service differentiation. Providing a high-level of service is becoming a business imperative as it facilitates customer retention and also helps in market penetration. Inventory rationing gives an edge to the firm by allowing it to provide the required level of service to the customers at lower cost. It minimizes the total inventory cost and thus maximizes the profit for the firm. It also enhances the goodwill of the firm by giving required service level or fulfilling the commitment stated in the contract.

How to ration inventory or when to replenish or how much to order are the key decisions the management has to take, such that the resulting inventory cost is minimum. Many researchers have considered inventory rationing due to its potential in cost saving. But no analytical results have been derived yet for the cost expression for more than two customer classes. Motivated from the effectiveness of inventory rationing and lack of a generalized model, inventory rationing for several classes is considered in this research.

Although inventory rationing is most the widely used tool to tackle service differentiation problem, it has some shortcomings. Inventory rationing reserves inventory to satisfy future demand from the higher priority class. But if there is little demand from the higher priority class, the reserved inventory may not get utilized and lower class' demand may be rejected although inventory is available. Inventory rationing, thus, provides very high service level to the higher priority class and as a result provides lower service level to the lower priority class. An inventory policy is developed to overcome this shortcoming in this research. In the next section 6.2, the conclusions from this research are presented. In section 6.3, research directions for the future are discussed.

6.1 Conclusions

In the first part of this research, inventory rationing for multiple classes under deterministic demand is considered. It is assumed that the unsatisfied demand is backordered to the next cycle and the lead time is fixed. Whereas most of the literature consider only the time based delay cost, a model under general cost structure considering both stock out cost and delay cost, is developed for the first time. It was

found that the cost function is convex in cycle time, T . An algorithm was developed to determine the optimal cycle time and rationing co-efficients (the fraction of cycle time after which a demand is backordered) for all the classes. A program is written in 'C' programming language, based on the algorithm and numerical experiments are conducted. The optimal run-out times (time after which demand is backordered from a customer class) show that the higher classes need to be satisfied in most of the cases whereas demand from the lower classes can be backordered. The contribution of this chapter is the development of a model for rationing of multiple customer classes operating under general penalty cost, and the development of an efficient algorithm to determine the cycle time and rationing co-efficients. Although the demand in general is stochastic, this model can still be useful when there little or no variation in demand over the planning horizon. There are also cases when the firm has a service contract with the customers and demand is known in advance and, in those cases, the model would be quite useful.

In the second part of this research, traditional inventory rationing policy was extended for ' n ' number of classes. Most of the research related to inventory rationing for service-differentiated customer classes considered only two demand classes. Only few considered more than two classes but none of them derived the exact cost expression for inventory rationing when there are more than two customer classes. In this research, the exact cost expression for ' n ' customer classes is developed under the assumption of Poisson demand. Rather than considering only time-based delay cost a general cost structure is considered which includes both time-based delay cost and one time stock out

cost. The model is developed for a continuous review (Q, r) inventory system. It was found that the cost expression derived is of recursive nature. Also an algorithm is presented to determine the optimal parameters, order quantity, Q and base stock for class i , S_i , which minimizes the total expected cost. Although the rationing policy can be represented as a recursive function, evaluating it computationally becomes very difficult for more than 5 customer classes. Also from managerial point of view, it is very cumbersome to use too many classes in practice. The numerical experiments suggest that implementing a consolidated 2-class or 3-class model for a 5-class model does not increase the cost much and can be used, for all practical purposes

In the third part of this research, a new inventory policy is proposed which eliminates the shortcomings of critical level inventory rationing model. At first, only two customer classes under stochastic demand are considered. Almost all the contemporary research relating to service differentiation has focused on inventory rationing; none, however, has been able to develop any other policy which can solve the service differentiation problem effectively. Motivated by the fact that the inventory rationing provides very low service level to the lower class, a 2-bin policy is developed where separate inventory is kept for the two classes but the higher class gets more priority by allowing it to use the lower class's stock when its own stock runs out. However, lower priority class is not allowed to use the high class' stock when its own stock runs out. The backorders are cleared by 'threshold 'clearing mechanism i.e. they are cleared in the same manner they would have if more inventory were available to satisfy the demand at the time it arrives. The exact expression for the proposed policy is derived

assuming Poisson demand. The model was developed under continuous (Q, S) policy. An algorithm was developed to determine the optimal parameters which minimize the total inventory cost under variable bounds. Numerical experiments show that when the holding cost is low, the proposed policy gives a lower expected cost. For all the experiments the proposed policy is slightly costlier (4.1%) than the rationing policy but it provides much higher service levels to all the customer classes, specially the lower class.

A more general 3-bin policy which incorporates both the proposed 2-bin policy as well as the Deshpande et al. (2003) policy is also developed. The 3-bin policy always performs as well or better than Deshpande et al. (2003) policy, in terms of cost. In the next section, future research directions are discussed.

6.2 Future research directions

In this research, the new multi-bin inventory policy is presented for two classes only and exact cost expression is determined. Determination of exact cost expression for more than two classes for this policy will be the future direction of research. For the three class case, three separate bins, BIN_1 , BIN_2 and BIN_3 are kept and demand from each class is satisfied from the respective bin. When BIN_1 runs out of stock demand from class 1 can be satisfied from BIN_2 and BIN_3 . It is first attempted to satisfy from BIN_3 and if BIN_3 is empty it can be satisfied from BIN_2 , if available. If BIN_2 is also empty, then the demand is backordered. Similarly when BIN_2 runs out of stock demand from class 2 can be satisfied from BIN_3 but not from BIN_1 . If BIN_3 is empty, the demand is backordered. If BIN_3 runs out the demand from class 3 is backordered immediately. The expected cost

of this 3-bin policy can be derived in the same manner as the two-bin policy. The following table gives the combination of three inventory levels without borrowing from other class' stock. The table shows the outcomes of different situations and cost type associated with each situation.

Table 6.1 Combination of inventory levels for 3 classes' problem

Case	I_1	I_2	I_3	Description	Cost Associated
1	+	+	+	Demand for both classes is satisfied from the respective bins	Holding cost for BIN_1, BIN_2 & BIN_3
2	+	+	-	Demand for class 1 and 2 is satisfied from its bin, BIN_3 runs out, class 3 demands are backordered	Holding cost for BIN_1 & BIN_2 , penalty cost for BIN_3
3	+	-	-	Demand from class 1 is satisfied from its bin, class 2 can borrow from BIN_3 but depends on the sequence of demand arrival, class 3 demand is backordered	Holding cost for BIN_1 . Penalty cost for both BIN_2 & BIN_3 but dependent on demand sequence
4	+	-	+	Demand from class 1 is satisfied from its bin, class 2 can borrow from BIN_3 but depends on the sequence of demand arrival	Holding cost for BIN_1 . Penalty cost/holding cost for both BIN_2 & BIN_3 but dependent on demand sequence
5	-	+	+	Class 1 can borrow from BIN_2 and/or BIN_3 ,but the amount borrowed depends on the sequence of demand	Penalty cost/holding cost for BIN_1, BIN_2 and BIN_3 but dependent on demand sequence
6	-	+	-	Class 1 can borrow from BIN_2 and/or BIN_3 ,but the amount borrowed depends on the sequence of demand	Penalty cost/holding cost for BIN_1, BIN_2 and BIN_3 but dependent on demand sequence
7	-	-	+	Class 1 can borrow from BIN_2 and/or BIN_3 ,but the amount borrowed depends on the sequence of demand	Penalty cost/holding cost for BIN_1, BIN_2 and BIN_3 but dependent on demand sequence
8	-	-	-	Class 1 can borrow from BIN_2 and/or BIN_3 ,but the amount borrowed depends on the sequence of demand	Penalty cost for BIN_1, BIN_2 and BIN_3 but dependent on demand sequence

The above table shows under what conditions higher classes can borrow from the lower class. The first four cases are straightforward; however, when inventory level of class 1 becomes negative, things start to get messy. This derivation for three classes will be dealt with in future. The finding of the two classes problem could be used to solve the three classes problem e.g. two classes out of three classes can be combined and the cost can be determined. For the general ' n ' classes problem there will be 2^n possible combinations but can be simplified by using the recursive process.

In Chapter 4, an inventory rationing model is developed for multiple classes and an algorithm to derive the optimal parameters is provided. However the algorithm is computationally very exhaustive. Developing a better algorithm or heuristic to determine the optimal parameters is one of the future directions of work. Also the model is developed under the assumption of 'threshold' clearing but '*priority*' clearing is more attractive and easy to understand. Total cost expression for ' n ' classes under 'priority' clearing will be dealt with in the future also.

PUBLICATIONS

- S Ghosh, R Piplani and S Viswanathan (2009) “Inventory models for differentiated service,” *IS DSI Conference, January 3-5, 2009, IIT Mumbai, India* (Technical presentation)
- S Ghosh, R Piplani and S Viswanathan (2009) “ Inventory Policies for service differentiation with deterministic demand,” *9th International Workshop on Closed-Loop Supply Chain Management, September 17-19,2009, Braunschweig, Germany*
- R Piplani, S Vishwanathan and S Ghosh (2011) “A new two-bin policy for inventory control of multi-priority demands in a continuous review environment.” *International Conference on Operations Research, August 30 –September 2, 2011, Zurich, Switzerland.*(Accepted)

REFERENCES

1. Arslan, H., Graves, S. C., and Roemer, T., 2007. "A single-product inventory model for multiple demand classes," *Management Science*, 53(9), 1486-1500.
2. Baker, K. R., Magazine, M. J., Nuttle, H. L., 1986. "The effect of commonality on safety stock in a simple inventory model," *Management Science*, 32, 982-988.
3. Bassok, Y., Anupindi, R., 1997. "Analysis of supply contracts with total minimum commitment," *IIE Transactions*, 29(5), 373-381.
4. Belobaba, P., 1989. "Application of a probabilistic decision model to airline seat inventory control," *Operations Research*, 37(2), 183-197.
5. Bitran, G.R., and Mondschieen, S.V., 1995. "An application of yield management to the hotel industry with multiple day stays," *Operations Research*, 43, 427-443.
6. Cattani, K.D., and Souza, G.C., 2002. "Inventory rationing and shipment flexibility alternatives for direct market firms," *Production and Operations Management*, 11(4), 441-457.
7. Chen, F. and Zheng, Y.S., 1993. "Inventory models with general backorder costs," *European Journal of Operations Research*, 65, 175-186.
8. Cohen, M.A., Kleindorfer, P., and Lee, H. L., 1988. "Service constrained (s, S) inventory systems with priority demand classes and lost sales," *Management Science*. 34, 482-499.
9. Dekker, R., Hill, R.M., and Kleijn, M.J., 2002. "On the $(S - 1, S)$ lost sales inventory model with priority demand classes," *Naval Research Logistics*, 49(6), 593-610.
10. Dekker, R., Kleijn, M.J., de Rooij, P.J., 1998. "A spare part stocking policy based on equipment criticality," *International Journal of Production Economics*, 56-57, 69-77.
11. Deshpande, V., Cohen, M.A., and Donohue, K., 2003. "A threshold inventory rationing policy for service-differentiated demand classes," *Management Science*, 49(6), 683-703.
12. Deshpande, V., Cohen, M.A., and Donohue, K., 2003b. "An empirical study of service differentiation for weapon system service parts," *Operations Research*, 51(4),

- 518-530.
13. Ding, Q., Kouvelis, P., Milner, J.M., 2007. Dynamic pricing for multiple class deterministic demand fulfillment. *IIE Transactions*, 39(11), 997-1013.
  14. Eppen, G., 1979. "Effects of centralization on expected costs in multilocation newsboy problems," *Management Science*, 25(5), 498-501.
  15. Evans, R.V., 1968. "Sales and restocking policies in a single inventory system," *Management Science*, 14(7), 463-473.
  16. Frank, K.C., Zhang, R.Q., and Duenyas, I., 1998. "A simple (s,z,S) inventory policy for inventory systems with priority demand classes," *INFORMS Montreal*.
  17. Frank, K.C., Zhang, R.Q., and Duenyas, I., 2003. "Optimal policies for inventory systems with priority demand classes," *Operations Research*, 51(6), 993-1002.
  18. Federgruen, A., Zheng, Y., 1992. "An efficient algorithm for computing an optimal (r, Q) policy in continuous review stochastic inventory systems," *Operations Research*, 40(4) 808–813.
  19. Gayon, F. Véricourt, D., Karaesmen, F., and Dallery, Y., 2004. "Stock rationing in a multi-class make-to-stock queue with information on the production status". Technical report, Laboratoire Génie Industriel, Paris.
  20. Gerchak, Y. and Henig M., 1986. "An inventory model with component commonality," *Operations Research Letters*, 5,157-160.
  21. Gerchak, Y.,Magazine M., and Gamble B,1988. "Component commonality with service level constraints," *Management Science*, 34(6) 753–760.
  22. Graves, S.C., 1985. "A multi-echelon inventory model for a repairable item with one-for-one replenishment," *Management Science*, 42(1), 1-18.
  23. Graves, S.C., 1996. "A multi-echelon inventory model for a repairable item with one-for-one replenishment. 31, 1247-1256.
  24. Ha, A.Y., 1997a. "Inventory rationing in a make-to-stock production system with several demand classes and lost sales," *Management Science*, 43(8), 1093-1103.
  25. Ha, A.Y., 1997b. "Stock rationing policy for a make-to-stock production system with two priority classes and backordering," *Naval Research Logistics*, 44(5), 457-472.

26. Ha, A. Y., 2000. "Stock rationing in an $M/E_k/1$ make-to-stock queue". *Management Science*, 46 77-87.
27. Hadley, G., and Whitin, T., 1963, "Analysis of inventory systems", Prentice Hall, Englewood Cliffs, NJ.
28. Kaplan, A., 1969. "Stock rationing," *Management Science*, 15(5), 260-267.
29. Katircioglu, K. and Atkins, D., 1996. "Managing inventory for multiple customers requiring different levels of service," Working Paper, University of British Columbia, Vancouver, Canada, B.C.
30. Kim, J.S., 2002. "On the benefits of inventory-pooling in production-inventory systems," *Manufacturing & Service Operations Management*, 4(1) 12-16.
31. Kleijn, M. and Dekker. R., 1998. "An overview of inventory systems with several demand classes. Econometric Institute report 9838/A, Erasmus University, Rotterdam, The Netherlands.
32. Lee, T. C. and Hersh, M., 1993. "A model for dynamic airline seat inventory control with multiple seat bookings," *Transportation Science*, 27(3), 252-265.
33. McGill, J. I. and Ryzin, G. J. V., 1999. "Revenue management: research overview and prospects," *Transportation Science*, 33(2), 233-256.
34. Melchior P, Dekker, R., and Kleijn, M.J., 2000. "Inventory rationing in an (s, q) inventory model with lost sales and two demand classes," *Journal of Operations Research Society*, 51(1), 111-122.
35. Melchior. P. M., 2003. "Restricted time-remembering policies for the inventory rationing problem," *International Journal of Production. Economics*, 81-82, 461-468.
36. Moon, I. and Kang, S., 1998. "Rationing policies for some inventory systems," *Journal of the Operations Research Society*, 49(5), 509-518.
37. Nahmias, S. and Demmy, S., 1981. "Operating characteristics of an inventory system with rationing," *Management Science*, 27(11), 1236-1245.
38. Pantumsinchai, P., 1992." A comparison of three joint ordering inventory policies," *Decision Sciences*, 23(1), 111-127.

39. Subramanian, J., Stidham, S., and Lautenbacher, C., 1999, "Airline yield management with overbooking, cancellations and no-shows," *Transportation Science*, 33(2), 147-167.
40. Teunter, R. H. and Haneveld, W. K. K., 2008."Dynamic inventory rationing strategies for inventory systems with two demand classes, Poisson demand and inventory rationing," *European Journal of Operational Research*, 190(1), 156-178
41. Topkis, D., 1968. "Optimal ordering and rationing policies in a non-stationary dynamic inventory model with n demand classes," *Management Science*, 15(3), 160-176.
42. Tsay, A., Nahmias, S., and Agrawal, N., 1999. "Modeling supply chain contracts: a review," *Quantitative models for supply chain management*, Kluwer Academic Publishers, Norwell, MA.
43. Urban, T.L., 2000. "Supply contracts with periodic, stationary commitment," *Production and Operations Management*, 9(4), 400-413.
44. Veinott, A.F., 1965. "Optimal policy in a dynamic, single product, non stationary inventory model with several demand classes," *Operations Research*, 13(5), 761-778.
45. de Véricourt, F., Karaesmen, F., Dallery, Y., 2002. "Optimal stock allocation for a capacitated supply system," *Management Science*, 48(11), 1486-1501.
46. de Véricourt, F., Karaesmen, F., Dallery, Y., 2000. "Dynamic scheduling in a make-to-order system: a partial characterization of optimal policies," *Operations Research*, 48(5), 811-819.
47. Viswanathan, S., 1997." Periodic review (s,S) policies for joint replenishment inventory systems," *Management Science*, 43(10), 1447-1454.
48. Wang, Y., Cohen, M.A., and Zheng, Y.S., 2002. "Differentiating customer service on the basis of delivery lead times," *IIE Transactions*, 34(11), 979-989.
49. Zhao, W., and Zheng, Y. S., 2001. "A dynamic model for airline seat allocation with passenger diversion and no-show," *Transportation Science*, 35(1), 80-98.
50. Zheng, Y.S. and Federgruen, A., 1991, "Finding optimal (s,S) policies is as simple as evaluating a single policy," *Operations Research*, 39(4), 654-665.

48. Zipkin, P., 1986a. "Stochastic lead-times in continuous time inventory models,"
Naval Research Logistics Quarterly, 33, 763–774.
49. Zipkin, P., 1986. "Inventory service-level measures: convexity and approximation,"
Management Science, 12(8) 975–981.
50. Zipkin, P.H., 2000. "Foundations of inventory management," McGraw-Hill, New
York.

Appendix A

Detailed explanation of Deshpande et al. (2203)'s policy using a 2-bin framework, with threshold clearing

After the placement of the l^{th} order of quantity Q the inventory position becomes $(r + Q)$. Allocate inventory position $(r+Q - K)$ to BIN_2 , and the remaining inventory K to BIN_0 . When a new demand arrives (from either class), it is attempted to be fulfilled from BIN_2 first. If the BIN_2 inventory position is > 0 , the demand is assumed to be satisfied from BIN_2 . If, however, BIN_2 inventory position is ≤ 0 , and if the demand is from class 2 it is backordered and is cleared (satisfied) only upon arrival of $(l+1)^{th}$ replenishment order. If BIN_2 inventory position > 0 , but the actual inventory level (on-hand inventory) is zero, and if the demand is for class 2, it is backordered but cleared upon arrival of l^{th} replenishment order. [Note that since inventory position is > 0 , there will be sufficient inventory upon arrival of l^{th} replenishment order]. If the demand is from class 1 and inventory level of BIN_2 is zero, the demand is satisfied from BIN_0 , provided BIN_0 is not empty. The demand is backordered if the inventory level of BIN_0 is also zero. Moreover, if the inventory position of BIN_2 is > 0 then theoretically it should have been satisfied from BIN_2 and thus the inventory position of BIN_2 is reduced by one and inventory level of BIN_0 is reduced by one as well if it is satisfied immediately (as in reality the demand is satisfied from BIN_0). Correspondingly qt_{02} , the quantity borrowed by BIN_2 from BIN_0 is incremented by 1. This quantity qt_{02} is returned to BIN_0 upon receipt of l^{th} replenishment order.

If BIN_2 inventory position is ≤ 0 , but BIN_0 inventory position is > 0 , then a demand for class 1 would be satisfied immediately if BIN_0 inventory level is > 0 ; else it is backordered and cleared upon arrival of l^{th} replenishment order. [Note as BIN_0 inventory position is ≥ 0 , there would be sufficient inventory to clear the backlogs upon arrival of l^{th} replenishment order]. If the inventory position of both BIN_2 and BIN_0 is ≤ 0 , then demand from class 1 is backordered and only satisfied upon receipt of $l+1^{th}$ replenishment order. It should be noted here that when the inventory position of a bin is ≤ 0 , the inventory level is clearly ≤ 0 . It is clear from the above discussion that the inventory level as well as the number of backorders for each class after lead time L with the 2-bin policy described above is same as that under threshold clearing since only class 2 demand arriving within the first $(r + Q - K)$ demand is cleared upon replenishment and any other demand from class 2 later on can only be cleared from the next $(l+1^{th})$ replenishment.

Appendix B

Detailed derivation of the cost expression for the 3-bin policy

The derivation of the expression for expected holding and penalty cost for a 3-bin policy is similar to that of a 2-bin policy; the only difference is in case of 3-bin policy there is one more bin available. The inventory position of this bin (i.e. BIN₀) has to be included in the cost expression. If k_1 and k_2 are the number of demand arrivals for class 1 and class 2 respectively during time period $[\tau_1, t+L)$, the probability distribution of k_1 and k_2 can be computed as a convolution of the Poisson distribution, given by (4.4) and the binomial distribution, given by (4.5). The joint probability distribution of demand for the two classes D_1 and D_2 during the time period $[\tau_1, t+L)$ is given by

$$\begin{aligned} g(k_1, k_2) &= \Pr(D_1 = k_1, D_2 = k_2) \\ &= f_{t+L}(k_1 + k_2) B(k_1, k_1 + k_2, \alpha_1) \end{aligned} \quad (\text{B.1})$$

Let $l_1 = S_1 - D_1$ and $l_2 = S_2 - D_2$. As class 1 and class 2 demand is attempted to satisfy from BIN₁ and BIN₂ respectively, l_1 and l_2 represent the inventory levels in BIN₁ and BIN₂ at time $(t+L)$ when both l_1 and l_2 are greater than zero. When demand cannot be satisfied from their respective bins, class 1 demand can be fulfilled from BIN₂ or BIN₀ (if BIN₂ is empty). The following table provides the possible cases of l_1 and l_2 .

Table B.1. Different cases of l_1 and l_2 for a 3-bin policy

Case	l_1	l_2		Description	Associated Cost
1	> 0	> 0		Demands for both classes are satisfied from their respective bins, BIN_0 untouched	Holding cost for both BIN_1 , BIN_2 & BIN_0 .
2	> 0	≤ 0		Demand for class 1 is satisfied from its bin, BIN_2 runs out, and class 2 demands are backordered, BIN_0 untouched	Holding cost for BIN_1 & BIN_0 but penalty cost for class 2/ BIN_2 .
3a	≤ 0	> 0	$ l_1 < l_2$	All demands are satisfied, class 1 demand is partly satisfied from BIN_2 . BIN_1 has zero inventory, BIN_0 untouched	Holding cost for BIN_2 and BIN_0
3b			$ l_1 \geq l_2$	Both BIN_1 and BIN_2 have zero inventory. Class 1 demand can be satisfied from BIN_0 .	Penalty cost for class 1 and/or class 2, depending on sequence of demand arrivals and number of class 1 demand.
4	$-\infty \leq l_1 \leq -S_0$	≤ 0		Both BIN_1 and BIN_2 have zero inventory. Class 1 demand can be satisfied from BIN_0 .	Penalty cost for class 1 and/or class 2, depending on sequence of demand arrivals and number of class 1 demand.
5	$-S_0+1 \leq l_1 \leq 0$	≤ 0		Both BIN_1 and BIN_2 have zero inventory. Some of the Class 1 demand can be satisfied from BIN_0 .	Penalty cost for class 1 and/or class 2, depending on sequence of demand arrivals and number of class 1 demand.

The first three cases are similar to the 2-bin policy. BIN_0 is not needed to satisfy class 1 demand. For the first two cases, it can be fulfilled from BIN_1 only and for the case 3a, after BIN_1 runs out, all the remaining class 1 demand can be fulfilled from BIN_2 . The $G_j(Q, S_0, S_1, S_2)$ for the first three cases would be same as 2-bin policy plus the holding cost for BIN_0 . $G_j(Q, S_0, S_1, S_2)$ denotes the expected holding and penalty cost for case j in the above table.

Therefore, the expected costs are given by

$$G_1(S_0, S_1, S_2, Q) = \sum_{l_1=1}^{S_1} \sum_{l_2=1}^{S_2} h(l_1 + l_2 + S_0) g(S_1 - l_1, S_2 - l_2) \quad (\text{B.2})$$

$$G_2(S_0, S_1, S_2, Q) = \sum_{l_1=1}^{S_1} \sum_{l_2=-\infty}^0 [h(l_1 + S_0) - p_2 l_2 + \pi_2 \lambda_2] g(S_1 - l_1, S_2 - l_2) \quad (\text{B.3})$$

$$G_{3a}(S_0, S_1, S_2, Q) = \sum_{l_1=-S_2+1}^0 \sum_{l_2=-l_1+1}^{S_2} h(l_1 + l_2 + S_0) g(S_1 - l_1, S_2 - l_2) \quad (\text{B.4})$$

The Cases 3b, 4 and 5 are different from the 2-bin policy as all of the class 1 demand cannot be fulfilled from BIN₁ and BIN₂ i.e. BIN₀ has to be used to fulfill the demand. The holding and penalty cost depend on the number of the class 1 demand that cannot be satisfied from BIN₁ and BIN₂. Below the expression for $G(Q, S_0, S_1, S_2)$ is derived for the cases 3b, 4 and 5.

Case 3b

The class 1 demands cannot be fully satisfied from BIN₁ and it is attempted to be satisfied from BIN₂; but BIN₂ is also not enough to satisfy all the class 1 demands that cannot be satisfied from BIN₁. So, the demands that cannot be satisfied from BIN₁ and BIN₂, are attempted to satisfy from BIN₀. If the number demands unsatisfied (from BIN₁ and BIN₂) is less than S_0 , there will be only holding cost for BIN₀ and no penalty cost for class 1; but if it is more than S_0 , there will be penalty cost for class 1. The number of unsatisfied demand depends on the sequence of demand arrival. As explained in case 3b for 2-bin policy, the holding cost would be $h(S_0 - (|l_1| - l_2 - k))$ where units that cannot be

satisfied from BIN₁ and BIN₂. is $(|l_1| - l_2 - k)$ and expected holding and penalty cost will be

$$\sum_{k=\max(0, |l_1| - l_2 - S_0 + 1)}^{\min\{(S_2 - l_2), (|l_1| - l_2)\}} \{p_2 k + \pi_2 \lambda_2 + h(S_0 - (|l_1| - l_2 - k))\} \frac{S_1 + S_2 C_{S_2 - l_2 - k} |l_1| - l_2 C_k}{S_1 + S_2 - l_1 - l_2 C_{S_2 - l_2}}$$

Similarly, when number of unsatisfied demands is more than S_0 , the penalty cost will be

$$\sum_{k=0}^{\min\{(S_2 - l_2), (|l_1| - l_2 - S_0)\}} \{p_2 k + p_1 (|l_1| - l_2 - k - S_0) + \pi_1 \lambda_1 + \pi_2 \lambda_2\} \frac{S_1 + S_2 C_{S_2 - l_2 - k} |l_1| - l_2 C_k}{S_1 + S_2 - l_1 - l_2 C_{S_2 - l_2}}$$

Therefore, the expected holding and penalty cost for case 3b is

$$\begin{aligned} G_{3b}(S_0, S_1, S_2, Q) = & \sum_{l_1=-\infty}^0 \sum_{l_2=0}^{\min(S_2, -l_1)} g(S_1 - l_1, S_2 - l_2) \left[\sum_{k=0}^{\min\{(S_2 - l_2), (|l_1| - l_2 - S_0)\}} \{p_2 k + p_1 (|l_1| - l_2 - k - S_0) + \pi_1 \lambda_1 + \pi_2 \lambda_2\} \frac{S_1 + S_2 C_{S_2 - l_2 - k} |l_1| - l_2 C_k}{S_1 + S_2 - l_1 - l_2 C_{S_2 - l_2}} \right. \\ & \left. + \sum_{k=\max(0, |l_1| - l_2 - S_0 + 1)}^{\min\{(S_2 - l_2), (|l_1| - l_2)\}} \{p_2 k + \pi_2 \lambda_2 + h(S_0 - (|l_1| - l_2 - k))\} \frac{S_1 + S_2 C_{S_2 - l_2 - k} |l_1| - l_2 C_k}{S_1 + S_2 - l_1 - l_2 C_{S_2 - l_2}} \right] \end{aligned} \quad (B.5)$$

Case 4

The number of class 1 and class 2 demands during time $[\tau_l, t + L)$ is more than the inventory available in BIN₁ and BIN₂. So, class 1 demands have to be satisfied from BIN₀. Depending on the sequence of demand arrivals, some class 1 demands might be fulfilled from BIN₂. As before, the cost depends on how many units are unsatisfied from BIN₁ and BIN₂ and holding and penalty cost is similar to the case 3b. But there is one scenario which is different from these scenarios: when none of the class 1 demand is satisfied from BIN₂ i.e. only class 2 demands are satisfied from BIN₂ (because they came before class 1 demand). In this case, all the unsatisfied class 1 demand from BIN₁ (i.e. $|l_1|$)

has to be satisfied from BIN_0 and since for case 4 the $-\infty \leq l_1 \leq -S_0$, only S_0 units will be satisfied from BIN_0 and $|l_1| - S_0$ units will be backordered. The penalty cost for this scenario will be $p_2 |l_2| + p_1 (|l_1| - S_0) + \pi_1 \lambda_1 + \pi_2 \lambda_2$. This scenario happens when the number of class 2 demand out of first $S_1 + S_2$ demand is S_2 or more and the probability for this case

i.e. only satisfying class 2 demands from BIN_2 is
$$\frac{\sum_{\rho=S_2}^{\min\{(S_1+S_2), (S_2-l_2)\}} S_1+S_2 C_{\rho}^{|l_1+l_2|} C_{S_2-l_2-\rho}}{S_1+S_2-l_1-l_2 C_{S_2-l_2}}.$$

Combining all the possibilities the expected penalty and holding cost for case 4 is given by

$$\begin{aligned} G_4(S_0, S_1, S_2, Q) = & \sum_{l_1=-\infty}^{-S_0} \sum_{l_2=-\infty}^{-1} g(S_1 - l_1, S_2 - l_2) \left[\sum_{k=|l_2|+1}^{\min\{(S_2-l_2), (|l_1+l_2|-S_0)\}} \{p_2 k + p_1 (|l_1 + l_2| - k - S_0) + \pi_1 \lambda_1 + \pi_2 \lambda_2\} \frac{S_1+S_2 C_{S_2-l_2-k}^{|l_1+l_2|} C_k}{S_1+S_2-l_1-l_2 C_{S_2-l_2}} \right. \\ & + \sum_{\substack{k=\max(0, |l_1+l_2|-S_0+1) \\ k \neq |l_2|}}^{\min\{(S_2-l_2), |l_1+l_2|\}} \{p_2 k + \pi_2 \lambda_2 + h(S_0 - |l_1 + l_2| - k)\} \frac{S_1+S_2 C_{S_2-l_2-k}^{|l_1+l_2|} C_k}{S_1+S_2-l_1-l_2 C_{S_2-l_2}} \\ & \left. + \{p_2 |l_2| + p_1 (|l_1| - S_0) + \pi_1 \lambda_1 + \pi_2 \lambda_2\} \frac{\sum_{\rho=S_2}^{\min\{(S_1+S_2), (S_2-l_2)\}} S_1+S_2 C_{\rho}^{|l_1+l_2|} C_{S_2-l_2-\rho}}{S_1+S_2-l_1-l_2 C_{S_2-l_2}} \right] \end{aligned} \quad (\text{B.6})$$

Case 5

This case is similar to the case 4; the only difference is when class 1 demands cannot be fulfilled from BIN_2 (because class 2 demands finish the entire inventory in BIN_2) and are attempted to satisfy from BIN_0 ; they can be satisfied fully from BIN_0 . This is because in this case $-S_0 + 1 \leq l_1 \leq 0$. There will be holding cost instead of penalty cost like case 4.

The expected holding and penalty cost is given by

$$\begin{aligned}
G_5(S_0, S_1, S_2, Q) = & \sum_{l_1=-S_0+1}^0 \sum_{l_2=-\infty}^{-1} g(S_1-l_1, S_2-l_2) [\sum_{k=|l_2|+1}^{\min\{(S_2-l_2), (|l_1+l_2|-S_0)\}} \{p_2 k + p_1(|l_1+l_2|-k-S_0) + \pi_1 \lambda_1 + \pi_2 \lambda_2\} \frac{S_1+S_2}{S_1+S_2-l_1-l_2} \frac{C_{S_2-l_2-k}}{C_{S_2-l_2}} C_k \\
& + \sum_{\substack{k=\max(0, |l_1+l_2|-S_0+1) \\ k \neq |l_2|}}^{\min\{(S_2-l_2), |l_1+l_2|\}} \{p_2 k + \pi_2 \lambda_2 + h(S_0 - |l_1+l_2|-k)\} \frac{S_1+S_2}{S_1+S_2-l_1-l_2} \frac{C_{S_2-l_2-k}}{C_{S_2-l_2}} C_k + \\
& \{p_2 |l_2| + h(S_0 - |l_1|) + \pi_2 \lambda_2\} \frac{\sum_{\rho=S_2}^{\min\{(S_1+S_2), (S_2-l_2)\}} S_1+S_2 C_{\rho}^{|l_1+l_2|} C_{S_2-l_2-\rho}}{S_1+S_2-l_1-l_2 C_{S_2-l_2}}] \tag{B.7}
\end{aligned}$$

Appendix C

Computational results for a 5-class model into an aggregated 3-class and 2-class problem

Case 1

Input Parameters

$\lambda_i = 2$ for $i = 1, 2, 3, 4, 5$

$p_1 = 1000, p_2 = 800, p_3 = 600, p_4 = 400, p_5 = 200$

$h = 100, LT = 0.5, A = 100$ $\pi_i = 0$ for all i

Table C.1 Cost comparison (Case 1)

	Breakup	Penalty Cost	TC	Parameters	Difference
N classes	(1)(2)(3)(4)(5)	(1000)(800)(600)(400)(200)	620.398	0,0,0,1,10,7	
3 classes	(1)(2)(3,4,5)	(1000)(800)(400)	626.149	0,0,11,7	-0.927
	(1)(2,3)(4,5)	(1000)(700)(300)	624.191	0,1,10,7	-0.611
	(1)(2,3,4)(5)	(1000)(600)(200)	620.398	0,1,10,7	0.000
	(1,2)(3)(4,5)	(900)(600)(300)	624.191	0,1,10,7	-0.611
	(1,2)(3,4)(5)	(900)(500)(200)	620.398	0,1,10,7	0.000
	(1,2,3)(4)(5)	(800)(400)(200)	620.398	0,1,10,7	0.000
2 classes	(1)(2,3,4,5)	(1000)(500)	626.149	0,11,7	-0.927
	(1,2)(3,4,5)	(900)(400)	626.149	0,11,7	-0.927
	(1,2,3)(4,5)	(800)(300)	624.191	1,10,7	-0.611
	(1,2,3,4)(5)	(700)(200)	620.398	1,10,7	0.000

Case 2

$\lambda_i=5$ for all i

$p_1=10000, p_2=7000, p_3=3000, p_4=2000, p_5=1000$

$h=250, LT=1, A=100,$

$\pi_i=0$ for all i

Table C.2 Cost comparison (Case 2)

	Breakup	Penalty Cost	TC	Parameters	Difference
N classes	(1)(2)(3)(4)(5)	(10000)(7000)(3000)(2000)(1000)	3118.86	0,1,1,1,34,8	
3 classes	(1)(2)(3,4,5)	(10000)(7000)(2000)	3153.17	0,1,36,7	-1.100
	(1)(2,3)(4,5)	(10000)(5000)(1500)	3146.18	0,2,35,8	-0.876
	(1)(2,3,4)(5)	(10000)(4000)(1000)	3160.48	0,3,34,7	-1.334
	(1,2)(3)(4,5)	(8500)(3000)(1500)	3129.88	1,1,35,8	-0.353
	(1,2)(3,4)(5)	(8500)(2500)(1000)	3121.19	1,2,34,8	-0.075
	(1,2,3)(4)(5)	(6666.7)(2000)(1000)	3133.63	2,1,34,8	-0.474
2 classes	(1)(2,3,4,5)	(10000)(3250)	3222.68	0,37,7	-3.329
	(1,2)(3,4,5)	(8500)(2000)	3153.17	1,36,7	-1.100
	(1,2,3)(4,5)	(6666.7)(1500)	3146.18	2,35,8	-0.876
	(1,2,3,4)(5)	(5500)(1000)	3160.48	3,34,7	-1.334

Case 3

$\lambda_i=5$ for all i

$p_1=10000, p_2=5000, p_3=3000, p_4=2000, p_5=1000$

$h=250, LT=1, A=100$

$\pi_i=0$ for all i

Table C.3 Cost comparison (Case 3)

	Breakup	Penalty Cost	TC	Parameters	Difference
N classes	(1)(2)(3)(4)(5)	(10000)(5000)(3000)(2000)(1000)	3081.73	0,1,0,2,33,7	
3 classes	(1)(2)(3,4,5)	(10000)(5000)(2000)	3112.25	0,1,36,8	-0.990
	(1)(2,3)(4,5)	(10000)(4000)(1500)	3101.36	0,2,35,8	-0.637
	(1)(2,3,4)(5)	(10000)(3333.3)(1000)	3113.54	0,3,34,8	-1.032
	(1,2)(3)(4,5)	(7500)(3000)(1500)	3093.16	1,1,34,7	-0.371
	(1,2)(3,4)(5)	(7500)(2500)(1000)	3081.73	1,2,33,7	0.000
	(1,2,3)(4)(5)	(6000)(2000)(1000)	3091.28	1,2,33,7	-0.310
2 classes	(1)(2,3,4,5)	(10000)(2750)	3171.91	0,37,7	-2.926
	(1,2)(3,4,5)	(7500)(2000)	3112.26	1,36,8	-0.991
	(1,2,3)(4,5)	(6000)(1500)	3101.36	2,35,8	-0.637
	(1,2,3,4)(5)	(5000)(1000)	3113.54	3,34,8	-1.032

Appendix D

Computational results for a 3-class model into an aggregated 2-class problem

Table D.1 Cost comparison

λ_i	p_i	p_{ai}	A	h	LT	3-Class	2-class	Diff
5,5,5	3000,2000,1000	0	0	250	1	1715.38	1715.38	0.00
			100			2104.67	2104.67	0.00
			200			2325.05	2325.05	0.00
			300			2512.55	2512.55	0.00
			400			2677.24	2677.24	0.00
			500			2821.60	2821.6	0.00
5,5,5	3000,2000,1000	0	500	50	1	1133.11	1161.93	28.82
				100		1683.5	1683.5	0.00
				150		2117.65	2117.65	0.00
				200		2494.38	2494.38	0.00
				250		2821.6	2821.6	0.00
				300		3119.33	3119.33	0.00
				350		3398.19	3398.19	0.00
				400		3656.54	3656.54	0.00
				450		3884.1	3884.1	0.00
				500		4111.66	4111.66	0.00
5,5,5	3000,2000,1000	0	500	250	0	1812.5	1812.5	0.00
					0.5	2439.27	2439.27	0.00
					1	2821.6	2821.6	0.00
					1.5	3136	3136	0.00
					2	3400.28	3400.28	0.00
					2.5	3640.12	3640.12	0.00
					3	3859.62	3859.62	0.00