Use of Advance Demand Information in Inventory Management with Two Demand Classes

Sourish Sarkar

Dissertation Submitted to the Faculty of the Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Industrial and Systems Engineering

John P. Shewchuk, Chair

Subhash C. Sarin

Roberta S. Russell

Ebru K. Bish

July 23, 2012

Blacksburg, Virginia

Keywords: production and inventory control, advance demand information, multiple demand classes

Use of Advance Demand Information in Inventory Management with Two Demand Classes

Sourish Sarkar

ABSTRACT

This work considers inventory systems with two demand classes, where advance demand information is available. Three related scenarios are presented: three-stage production-inventory systems are studied in first two, whereas pure inventory systems are studied in the last scenario.

In the first scenario, continuous review production-inventory systems are considered, where only one class provides advance demand information and early demand fulfillment is permitted. A new approach for production replenishment and order fulfillment in such systems is proposed, which combines the benefits of early fulfillment with Kanban-based pull systems. Simulation is used to compare the performance of the resulting policy with two other policies for a variety of scenarios (depending on the arrival rates, system utilizations, cost structures, arrival ratio, priority levels and amount of the advance demand information). A simulation-based lower bound on the optimal cost is established for some specific scenarios. The proposed policy outperforms the existing policies in every setting considered. Also, the proposed policy has added advantage of both retaining the benefit at high system utilizations and increasing the benefit up to the maximum level of advance demand information provided. A small fraction of customers providing advance demand information with early fulfillment acceptable is shown to have higher benefit than all customers providing same advance demand information with no early fulfillment.

In second scenario, both classes provide advance demand information in production-inventory systems, though only one class accepts early fulfillment. Different levels of system utilization, arrival ratio and backorder cost are considered in the simulation experiments to show the superiority of early fulfillment. Also, experiments suggest that lowering the expected supply lead time may be more beneficial than increasing the demand lead time by the same amount for production-inventory systems with utilization dependent supply lead times.

In third scenario, pure inventory systems are considered, where the demand classes provide different amount of advance demand information, and only one class accepts early fulfillment. The structure of an optimal policy is analytically characterized for periodic review systems under some specific conditions.

Acknowledgement

First and foremost, I would like to thank my advisor and Committee Chair, Dr. John Shewchuk for providing me with the opportunity to work with him. I truly appreciate his continued support and encouragement in the past five years of my PhD. I am also thankful to my Committee members Dr. Subhash Sarin, Dr. Roberta Russell and Dr. Ebru Bish for providing help and inputs when I needed them. The mentorship and guidance that I received from my committee members inspired me for a career in academia. I also want to thank my previous Committee member Dr. Yasemin Merzifonluoglu, who had been helpful during her brief stint in our department. Thanks to my previous Committee member, Dr. Raghu Pasupathy for his ideas and comments in my research.

I am thankful to the faculty members, staff and fellow students in the Grado Department of Industrial and Systems Engineering at Virginia Tech, who made me feel being at home. I will miss my time at Blacksburg.

My pursuit of higher education would not have been possible without the support and encouragement from my family. Thanks to my parents for inspiring me to achieve my goal. Finally, I thank my wife Bitan for always being there for me at good and bad times.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Motivation

For today’s global companies, creating and managing efficient supply chains are getting wide attention. Supply chain performance depends upon several fundamental factors, such as production and inventory capacities of its partners, delivery lead-time and service level. The capacity is inherently related to the cost of the supply chain; however, the last two quantities may depend upon effective managerial decisions and can be seen as opportunities for improvement without making any significant investment on the supply chain per se. Recent development in information technology and telecommunication infrastructure sets a new trend of sharing vast amount of data among supply chain partners. A retailer shares point-of-sale information with his suppliers; similarly, a huge amount of information is exchanged between the manufacturers and retailers and between the retailers and consumers. This information sharing helps to streamline supply chains by keeping a detail track of required items that minimizes replenishment order volume, which can effectively reduce inventory holding cost. Evidence can be provided in this context: In 1993 United States witnessed a period of economic growth (increasing business profits with relatively stable price, more workers and more high-technology investment) and yet decreasing productivity in the business sector. This conflict was attributed to significant shrinkage of inventories because of information sharing that had been possible through technological progress happened in that time (Woodward, 2000). The research on improving operational benefits through information sharing is still in an early stage, but it is rapidly developing. The key to operational benefit is to create a better match between supply and demand for all supply chain partners, which helps to reduce the amount of waste in form of lost revenue, surplus capacity or surplus inventory.

Matching supply with demand in a stochastic environment seems to be a never-ending challenge for supply chain managers. Apart from the recent trend of information sharing as described above, there are several ways to create a match. Traditionally, pricing has been used by the business managers to match supply with demand. When the demand is not as high as expected, a price discount or sales promotion is offered for stock clearance. A relatively recent strategy for a better match between supply and demand is to segregate customers among various classes and to offer different levels of service to these customer classes. Segregating customers is also effective for creating a flexible supply chain, which can cater to the needs of different requirements set by the customers.

Segregation of customers into different classes may be based upon the urgency of the customers to get the product (Chen, 2001). Different customers have different levels of urgency in reality because of the following reasons: (i) Many businesses may run an online store together with several brick-and-mortar stores. Selling items from brick-and-mortar stores needs immediate fulfillment. However, online customers need to wait for a few days, which automatically provide the firm with some advance demand information. (ii) Levels of urgency vary among the customers for the same product. A business executive who suddenly broke his laptop may pay extra for overnight delivery of a replacement. But a student who may buy his laptop from the same online store should prefer to wait a few days to get an inexpensive shipping rate. For example, Dell has four shipping options for its online customers − standard, priority, overnight and precision. Some other online retailers like Amazon sometimes order the products from the manufacturer after receiving the customer orders (Simchi-Levi, 2010). (iii) Some of the supply chain partners may have long-term contract with the manufacturer, while others may not. Because of the special relationship with long-term contract holders, the manufacturer may prioritize the order fulfillment for them. (iv) Levels of equipment criticality play a role to segregate customers. Critical backorders (e.g., parts needed to fix a breakdown) are much more expensive than non-critical backorders (e.g., parts needed for restocking). (v) A business may sell its products to other businesses as well as to the consumers. Delivery timeline for the businesses and consumers may be different. (vi) In the Just-In-Time environment, many intermediate partners focus on maintaining minimal inventory for their businesses. Although these partners may know the precise demand well in advance (and inform their vendors accordingly), still they may not accept delivery from the vendors until the items are due in the manufacturing floor. Thus, different buyers providing different amounts of advance demand information can be segmented into different classes. Catering to the needs of these customer classes and simultaneously maintaining an efficient system with low production and inventory cost is a practical challenging problem to the researchers.

1.2 Outline of the Document

The remainder of the Introduction provides some preliminary concepts, problem description and research objective. The next section briefly discusses some of the relevant papers on (i) inventory management with multiple demand classes and (ii) the use of advance demand information in inventory systems. Section 3 discusses the proposed approach for the above-mentioned problem. Section 4 describes the methodology used for production-inventory systems (simulation model, experimental design and implementation), followed by the results and discussion in Section 5. Section 6 discusses pure inventory systems, followed by the final section for conclusions.

1.3 Preliminaries

Some general concepts and related vocabulary are presented here.

Advance Demand Information (ADI): When customers announce their orders ahead in time of actual demand, a firm and precise information of future demands is obtained at the time of order arrival. Here, firm means non-cancellable and precise means detail information regarding order quantity and order due date. Clearly, orders with ADI are not due on arrival. In other words, there is a positive time gap between order arrival and order due date. When orders are not fulfilled within their due date, they are either lost or backordered. Because of the presence of ADI, the terms orders and demands may not be used interchangeably. The orders arrive when the inventory manager receives the ADI, but the demands are realized on the order due dates.

On-hand inventory: the number of items actually available in the inventory that are ready to serve the demand.

On-order inventory/ pipeline inventory: the number of units that have been ordered but yet to be received in the stock.

Inventory position: An inventory manager decides on order placement based upon the current inventory position. This term is indicated by the combined value of on-hand and on-order inventories minus the items demanded by customers, which are not yet delivered. Therefore, Inventory position = on-hand inventory + pipeline inventory – backorders.

Inventory level: Another important quantity in inventory management is inventory level, which determines the overall cost of maintaining inventory for given values of per unit holding and backorder costs. Inventory level = on-hand inventory – backorders.

Supply lead time: This is the time required to replenish an item in stock. For a production-inventory system, this indicates the production time using capacitated production line. For a pure inventory system, this is the time between order placement and item arrival.

Demand lead time: The gap between order arrival and order due is termed as demand lead time. This term is coined by Hariharan and Zipkin (1995). In their paper, the authors show that increasing the demand lead time has the exact same effect as reducing the supply lead time by the same amount, when early order fulfillment is not permitted.

Pure inventory system: Inventory replenishment is typically vendor driven, so the replenishment process is exogenous. Thus, the supply lead times for such systems are load-independent.

Production-inventory system: Inventory replenishment is typically done by manufacturing/assembling the units in-house. The replenishment process is therefore endogenous, and the supply lead time is load-dependent.

Risk pooling: When several demand classes require the same product and the demand is stochastic, it is beneficial to use a common inventory to serve all classes rather than maintaining separate stocks for each class. The strategy of combining inventories in this way is known as risk pooling. In reality, the aggregate forecast for stochastic demand tends to be more accurate than individual forecasts. Thus, the system that serves several classes through common inventory is less variable than the system that keeps separate inventories. Therefore, a smaller amount of safety stock is sufficient for a common inventory system.

Early fulfillment (EF): If an arrived order with ADI is fulfilled before its due date, then the fulfillment type is termed as early fulfillment (as opposed to exact fulfillment, where orders are fulfilled on their due dates). When early fulfillment is permissible, the inventory manager or the manufacturer has sole discretion to fulfill an order early or exactly on the due date without incurring any penalty. The possibility of early fulfillment may provide the inventory manager with greater flexibility and reduced inventory holding cost. However, from a customer's perspective, acceptance of early delivery may not be always beneficial. For example, if the customer serves as a supplier for its downstream supply chain, then she may prefer exact fulfillment to cut down excessive inventory holding cost. In contrast, when the customer is an end consumer, she should be more than happy to receive the delivery earlier than expected.

Order crossover: Order crossover describes the situation when an arriving order is due earlier than a previously arrived order. This is possible only when multiple demand classes exist with different demand lead times. Order crossover happens only when (i) the arrival of an order with ADI is followed by arrival of another order with shorter demand lead time and (ii) the order arrived later (i.e. the order with shorter demand lead time) is due before the due date of the order arrived earlier.

Base-stock policy: Under this policy, inventory is continuously monitored and whenever a demand occurs, a replenishment order is immediately placed (or a production is initiated) to restore inventory. Demands are fulfilled immediately on arrival if stock is available. Therefore, for a base-stock policy, at any instance of time, inventory position remains constant.

Order-base-stock policy: Under this policy, an order arrival, instead of a demand fulfillment triggers replenishment order placement (or production initiation). Therefore, the policy maintains a target value of modified inventory position. This policy works differently than a standard base-stock policy only when ADI is available at least for some of the customer orders.

(s, S) policy: Under this policy, an order is placed (or production is initiated) to raise the inventory position to the predetermined order-up-to level S whenever it falls to or below the predetermined reorder level *s*. Scarf (1960) uses the properties of *K*-convex functions to show that only a (*s*, *S*) type policy maximizes the expected discounted profit for finite horizon dynamic inventory problems where the holding and shortage costs are linear, and the ordering cost is a combination of a fixed and a unit variable cost.

CONWIP policy: Under this policy, the total number of work-in-process items and finished goods inventory is constant; hence, production can begin only when an item is consumed from final inventory. This policy helps to control cost by limiting the number of jobs present in the system.

Push and pull strategies: In a push-based supply chain strategy, production and distribution decisions are taken according to long-term forecasts. This system performs poorly against demand variability; hence, a large amount of safety stock is required to avoid stock-out. However, this system is considered as appropriate for large-scale production over a long-time period. In contrast, in a pull system, production and distribution decisions are demand-driven, which is effective to reduce inventory. But this system is not favorable for planning far ahead − thus, it cannot take advantage of economies of scale very often. A base-stock policy is an example of a pull-driven system.

1.4 Problem Description

We consider systems where ADI is available in the form of demand lead time (time gap between an order arrival and its due date) associated with some or all of the demands. Three different scenarios are investigated in this research. In each scenario, two demand classes are assumed, which are discussed later in this section. The following assumptions hold for all scenarios considered: (i) demands are interchangeable, hence same items can be used to fulfill demands from any class, (ii) no demand cancellation or modification after arrival, (iii) linear backorder and holding costs, (vi) stochastic arrival process for each demand class, (v) difference between the two classes is based upon the delivery requirement. Therefore, the demands from these classes are independent to each other (an example of online laptop store is provided earlier: in reality, demand of laptops from the executives are independent of that from the students), (vi) demand arrival and replenishment processes as well as the costs are stationary, (vii) all quantities (e.g., demand, inventory position) are integers, (viii) no setup or ordering cost, (ix) no machine failures, and (x) no raw material shortages.

Scenario 1

Continuous review, finite capacity production-inventory systems are considered, where only one demand class provides ADI (demand from the other class is due as soon as they arrive). Early fulfillment of demand is allowed– therefore the demands can be fulfilled anytime between their arrivals and the due dates.

A schematic of this scenario is shown in Figure 1, where the serial production-inventory system consists of three stages. The number of stages is restricted to three for fair comparison to the literature, where three stages are assumed. The three-stage system may represent an assembly-line process in reality. The three stage model is a sequence of queues with no restriction on queue length. However, because of the complexity of Kanban-based serial-production systems, results cannot readily be extended to the general N–stage case. The demands may provide ADI (depending upon the demand class) and ADI-providing demands may wait in the delay station (depending upon the fulfillment decision). Demand arrival and processing time distributions at the stages are known.

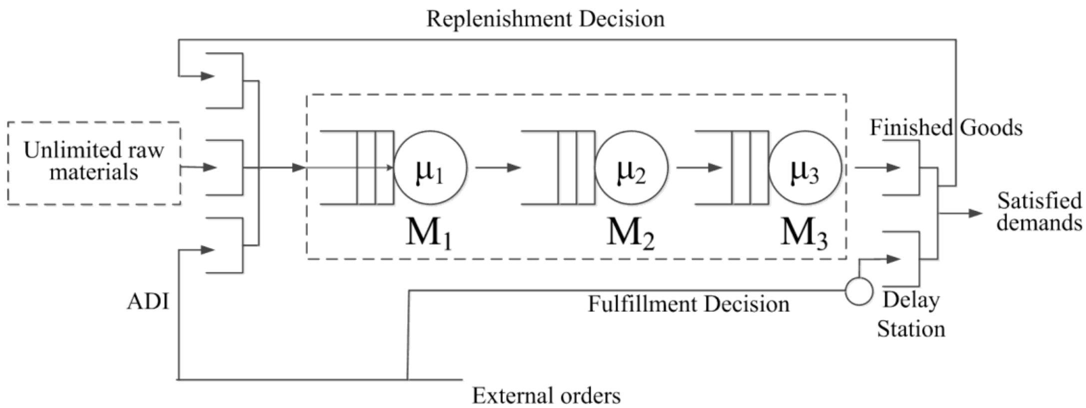


Figure 1. Production-inventory system with two demand classes.

As shown in Figure 2, there is a trade-off in early fulfillment of type 1 demands. Early fulfillment helps to save inventory holding cost, but increases the possibility of stock-out when type 2 demands arrive.

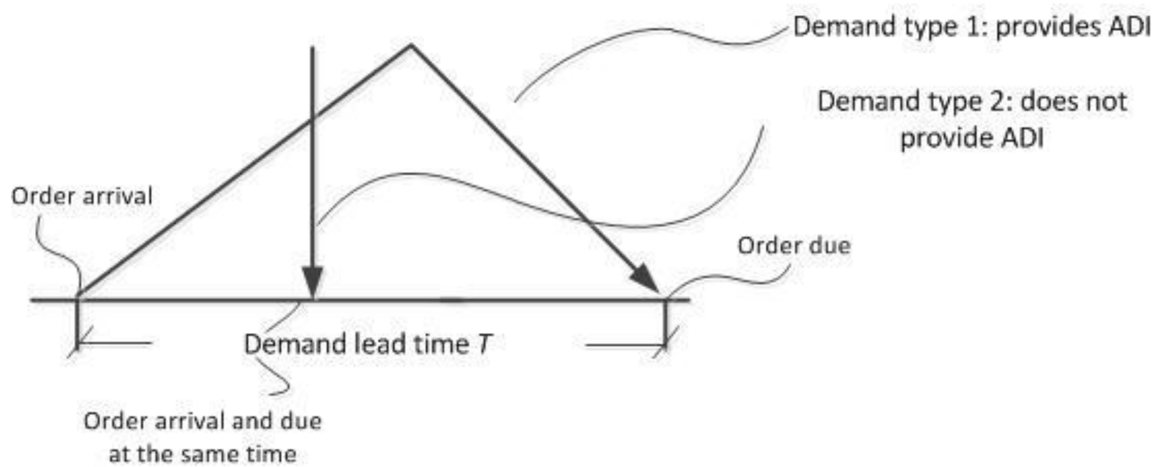


Figure 2. Two types of demands in a time diagram (Scenario 1).

<u>Scenario 2</u>

This scenario is the same as the previous one with the exception that the demands of both classes provide the same amount of ADI, but only one class accepts early fulfillment. The backorder costs for the demand classes may be different. Figure 3 shows two types of demands in this scenario in a time diagram. For both Scenarios 1 and 2, the replenishment process is endogenous (i.e., replenishment orders are pipelined and they arrive in the same order in which they are placed).

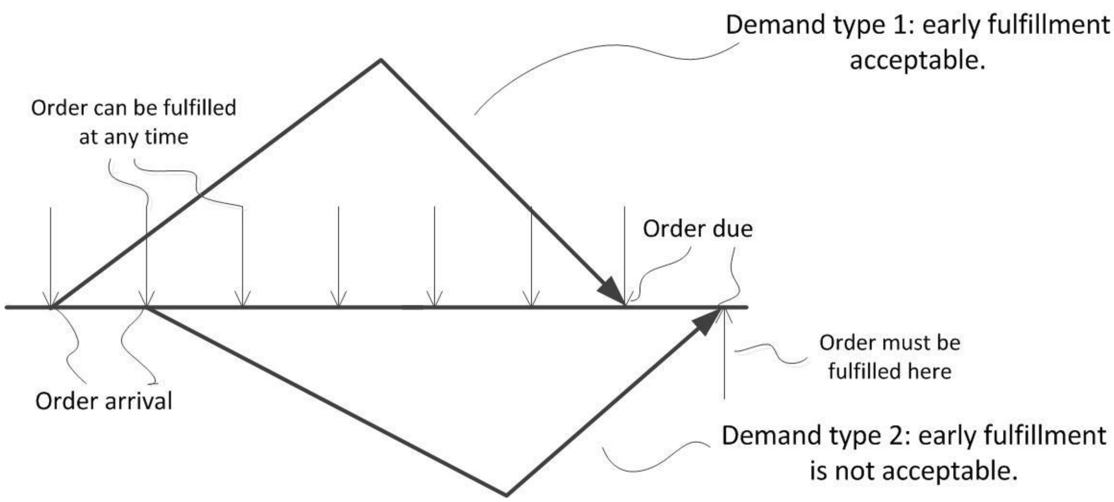


Figure 3. Two types of demands in a time diagram (Scenario 2).

For a production-inventory system, let w and h be the cost of holding WIP and FG inventory, respectively: b is the backorder penalty cost (all expressed in per item per unit time). The objective is to find a policy p that gives the minimum combined cost of holding and backordering for a given system. This is mathematically expressed by the following:

$$\min_p \lim_{T\to\infty} \frac{1}{T} \mathrm{E}_p \left[\int_0^T \left\{ w \cdot I_w(t) + h \cdot I_f^+(t) + b \cdot I_f^-(t) \right\} dt \right] \quad (1)$$

In above expression I_{w} and I_{f} are the WIP and inventory level respectively, at time t, (i.e., $I_{\mathrm{f}}^+(t) = \max(I_{\mathrm{f}}(t), 0)$ = FG inventory level and $I_{\mathrm{f}}^-(t) = \max(-I_{\mathrm{f}}(t), 0)$ = backorder level at time t). An effective policy for a given scenario balances the two costs – holding cost (summation of first two terms within the integral in (1)) and backorder penalty (the last term within the integral).

When the backorder costs (b_1, b_2) are different for different classes, the above expression becomes:

$$\min_p \lim_{T\to\infty} \frac{1}{T} \mathrm{E}_p \left[\int_0^T \left\{ w \cdot I_w(t) + h \cdot I_f^+(t) + b_1 \cdot I_{f1}^-(t) + b_2 \cdot I_{f2}^-(t) \right\} dt \right] \quad (2)$$

Here $I_{\mathrm{f}1}^-(t)$ and $I_{\mathrm{f}2}^-(t)$ are the backorder levels of two classes at time *t*. As with the previous case, an effective policy creates a balance between holding cost and backorder penalty.

Scenario 3

A periodic review pure inventory system is considered in this scenario, with two demand classes having different amounts of ADI and different fulfillment flexibility criteria (i.e., early fulfillment is not allowed for one class). Figure 4 shows a time diagram for this scenario. The demand class that accepts early fulfillment is assumed to provide shorter demand lead time.

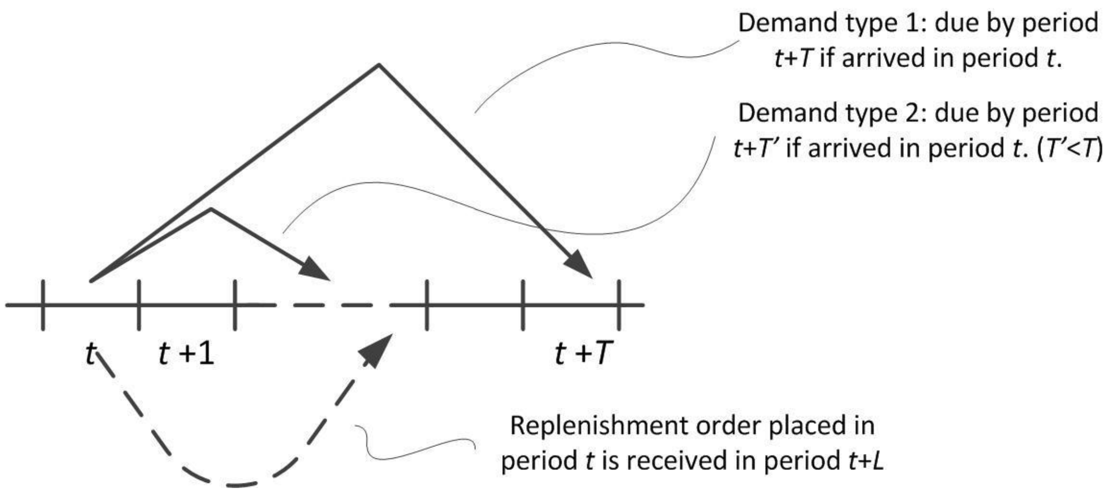


Figure 4. Demands and the inventory replenishment process in a time diagram (Scenario 3).

This scenario is different from the previous two production-inventory scenarios in the following aspects:

(1) This is also a two-demand-class scenario with both demand classes provide advance demand information. However, the demand lead time for Class 1 (T_1) is shorter than that of the Class 2 demand (T_2). Also, only the demands from Class 1 accept early fulfillment.

(2) For pure inventory systems supply lead time is not endogenous. Here the supply lead time (L) is assumed to be constant.

(3) This is a periodic-review system, where the ordering decision is made only at the beginning of each period (and this decision may depend on unfulfilled advance demand and available inventory). The sequence of events in each period is as follows: (a) inventory manager reviews the system state and (b) places order for replenishment quantity Z_t. (c) Then she receives the delivery of replenishment order (w_t), which was placed L periods earlier ($w_t = Z_{t-L}$). (d) Demands from both classes arrive through their respective advance information. (e) Decisions on demand fulfillment are taken by the inventory manager.

(4) The system is capacitated with the maximum order quantity in any period is a constant Q.

(5) Demand in a period is stochastic; however, no specific distribution of demand has been assumed. Single period demand distribution is denoted by ϕ, and the density by ø. The one-period discount factor in the context of classical inventory system is assumed to be 1 (i.e., future period costs are not discounted).

Table 1 summarizes the three scenarios.

Table 1. A summary of the three scenarios.

Type	Scenario	Class 1		Class 2		Remarks
		ADI	EF	ADI	EF	
Production -Inventory	1	√	√			
Production -Inventory	2	√	√	√		Both classes provide same amount of ADI
Pure Inventory	3	√	√	√		Class 2 provides higher amount of ADI

1.5 Research Objective

This research aims to find inventory-replenishment/production-control policies for Scenarios 1 and 2, which are expected to show benefits over conventional policies. A simulation-based lower bound on the optimal cost will also be established for some cases, which will help to measure the effectiveness of the proposed policy. Additionally, a sensitivity analysis will be conducted to observe the change in performance with the following variables: (i) for Scenario 1: system utilization, demand lead time, relative arrival rates of the demand classes, relative backorder costs; (ii) for Scenario 2: relative arrival rates, relative backorder costs, and a change in the demand and supply lead times. The optimal policy will be characterized for Scenario 3.

2. Literature Review

Previous work on the effect of ADI on inventory systems with multiple demand classes is related to two different streams of research. These are reviewed separately in the following sections.

2.1 Inventory Management with Multiple Demand Classes

Veinott (1965), Topkis (1968) and Kaplan (1969) present some of the earliest research concerned with multiple demand classes and rationing policies. Each of these works assumes periodic review inventory systems. The underlying notion of these papers is that of a threshold-type static rationing policy, which works to fulfill only the high-priority demands and all other demands becoming backorders once the on-hand inventory drops to a predetermined reserve level.

A continuous review production-inventory system with multiple demand classes is considered in Ha (1997a), Ha (1997b), among others. In his papers, Ha proves the optimality of stationary critical-level rationing policies in make-to-stock production systems when processing times are exponential, order arrivals are Poisson, and either lost-sales or backordering is allowed. The reader is referred to Teunter and Haneveld (2008) for an excellent review of the various rationing policies. Each work described there assumes that order fulfillment priority depends upon shortage cost or service level, which vary with the demand class. By contrast, the proposed research focuses on multiple classes that differ in the flexibility of early fulfillment or the amount of demand lead time provided.

A growing number of researchers investigate inventory management problem with multiple demand classes, where the demand lead time varies among the classes. These works are described in the following section, along with the single-class-version of the same problem.

2.2 Inventory Management Using Advance Demand Information

The literature on this research area can be categorized using some characteristics (e.g., system, review type, nature of supply lead time, etc.) as shown in Table 2. Some of the listed papers consider both categories of the characteristics. For example, Wang and Yan (2009) primarily consider uncapacitated inventory systems and then extend their analysis for capacitated systems. Hariharan and Zipkin (1995) focus on deterministic lead time and then extend the results on systems with stochastic lead times. Similar extensions are observed in the work of Wang and Toktay (2008) on demand classes, Karaesmen et al. (2004) on early fulfillment, Cattani and Souza (2002), and Wijngaard and Karaesmen (2007) on shortage treatment. Iravani et al. (2007) consider that the orders not fulfilled in due date are lost for one class and are backordered for the other class of demands. The table depicts that limited research has been conducted so far where early fulfillment is allowed: this is the focus of the current research. Some of the papers from the table are discussed in the following sections on uncapacitated and capacitated inventory systems.

Table 2. Relevant literature on inventory management with advance demand information.

	System		Review time		Supply lead time		Demand classes		Shortage cost		Early Fulfillment		Shortage treatment	
	Uncapacitated inventory	Capacitated inventory	Periodic review	Continuous review	Exogenous and deterministic	Stochastic	Single	Multiple	Different for different classes	Same for all	Allowed	Not allowed	Order loss	Backorder
Cattani and Souza (2002)		•		•		•		•	•			•	•	•
Claudio and Krishnamurthy (2009)		•		•		•	•			•		•		•
Duran et al. (2007)		•	•		•			•	•			•		•
Duran et al. (2008)		•	•		•			•	•			•		•
Gallego and Ozer (2001)	•		•		•			•		•		•		•
Gallego and ozer (2003)	•		•		•			•		•		•		•
Gayon et al. (2009)		•		•		•		•	•			•	•	
Hariharan and Zipkin (1995)	•		•		•	•	•			•		•		•
Iravani et al. (2007)		•		•		•		•	•			•	•	•
Karaesmen et al. (2004)		•		•		•	•			•	•	•		•
Kim et al. (2009)		•		•		•	•					•		•
Kocaga and Sen (2007)	•			•	•			•	•			•		•
Liberopoulos and Koukoumialos (2005)		•		•		•	•			•		•		•
Liberopoulos (2008)		•		•		•	•			•		•		•
Lu et al. (2003)	•			•		•	•			•		•		•
Ozer (2003)	•		•		•			•		•		•		•
Ozer and Wei (2004)	•		•		•			•		•		•		•
Tan et al. (2009)	•		•		•			•	•			•	•	
Wang and Yan (2009)	•	•		•	•			•	•		•		•	
Wang et al. (2002)	•			•		•		•		•		•		•
Wang and Toktay (2008)	•		•		•		•	•		•	•			•
Wijngaard and Karaesmen (2007)		•		•		•	•			•		•	•	•

2.2.1 Uncapacitated Inventory Systems

In their paper, Gallego and Ozer (2001) consider both positive and zero set up cost cases, and show the optimality of state-dependent (*s*, *S*) and state-dependent base-stock policies respectively. The state is represented by the advance demand vectors outside the supply lead time horizon. The second case is particularly relevant to this research, where the researchers focus on both finite and α-discounted infinite horizon problems. An important finding of their work is that the observed demand beyond the protection period (which is defined as the sum of supply lead time and a review period) has no influence on the optimal base-stock level.

Wang and Toktay (2008) extend the work of Gallego and Ozer for systems where early fulfillment is permitted. They consider homogeneous as well as heterogeneous demand lead time cases. For homogeneous demand lead time cases, again a state-dependent (*s*, *S*) policy is shown to be optimal, where advance demand vectors outside the supply lead time horizon defines the states. Additionally, they show that increasing the demand lead time is more cost effective than reducing the supply lead time by the same amount. When demand lead time is increased, flexibility due to early fulfillment can potentially be employed over a larger time horizon, which helps to reduce the cost. This finding is in contrast to the earlier finding by Hariharan and Zinpin (1995) considering early fulfillment forbidden. For heterogeneous demand lead time cases, the presence of order crossover makes joint optimization of replenishment ordering and inventory allocation decisions complicated. In the model of Gallego and Ozer (2001), system state is defined by the modified inventory position and observed demand profile beyond the protection period. But, allowing early fulfillment makes the model even more complicated, where the state is defined by three terms − inventory level, supply pipeline and unsatisfied advance demand profile. The dynamics of first and third terms are nonlinear, unlike the same terms for systems without early fulfillment. The researchers obtain the upper and lower bounds on the optimal cost by proposing protection-level heuristics and using allocation assumptions respectively. In both models mentioned above, the demand classes are formed based upon the amount of ADI comes with the orders; there is no difference in terms of the reward for fulfillment or the penalty cost for backordering among these classes. But the following works assume either different reward or different penalty cost among the demand classes.

Wang and Yan (2009) consider inventory systems serving both patient and impatient customers. Fulfillment of patient customer orders is flexible, as it can be deferred to the next cycle. However, demand from impatient customers must be fulfilled from the on-hand inventory. In this scenario, they prove the optimality of threshold-type inventory allocation and base-stock type replenishment ordering policies. Except for the models described in Wang and Toktay (2008) and Wang and Yan (2009), no other model in the literature on pure inventory management allows early fulfillment.

Cattani and Souza (2002) consider two different modes of shipment for direct market firms serving two demand classes with different demand lead times. Using a birth-and-death Markov decision model, they compare the performance of various rationing policies with a naïve first-come, first-serve policy for a variety of scenarios.

Kocaga and Sen (2007) also consider a two-demand class inventory management problem, where the demand classes are of different criticality. The criticality of a demand class is measured by the service level requirement for that class. Demand from the critical class must be satisfied immediately, but the non-critical class demands provide ADI. Under a one-for-one replenishment policy, they derive the expressions for service levels of both classes. As early fulfillment is not permitted in their model, the purpose of rationing becomes pointless when the demand classes are of equal criticality. A numerical study conducted by the researchers depicts as much as 14% savings on inventory cost through the use of demand lead time and rationing.

2.2.2 Capacitated Production Inventory Systems

Karaesmen et al. (2004) investigate the value of ADI for M/M/1 queuing systems, both with and without the considerations of early fulfillment. Assuming all orders provide the same amount of ADI, they propose the base-stock policy with advance demand information (BSADI) for the scenario when early fulfillment is not allowed. As opposed to the standard base-stock policy defined by a single parameter (the target inventory level), BSADI has two parameters: the target inventory level and the release lead time. Under the proposed policy, the optimal release lead time may postpone production release from the customer order arrival time, depending upon the availability of ADI, and the policy is claimed to achieve near-optimal performance. Introduction of the release lead time results in a partial transition to make-to-order system. When early fulfillment is permitted, the researchers use a simple property that the system with early fulfillment/order-base-stock policy is equivalent to the standard base-stock system with a modified backorder cost that starts to incur after a delay equal to the amount of demand lead time. For this equivalent system, ADI does not exist, and it is possible to determine the mathematical expression of the optimal base-stock level and the optimal cost assuming an M/M/1 configuration. This is the only paper on production-inventory systems that considers early fulfillment.

Liberopulos and Koukoumialos (2005) used simulation to investigate trade-off between optimal base-stock levels, number of Kanbans and planned supply lead times for single-stage and two-stage systems with ADI under base-stock and hybrid base-stock/Kanban policies.

Wijngaard ((2007) proves the optimality of order-base-stock policy for systems with deterministic production process, and shows that inducing ADI into the system would result in inventory cost savings of $c.(1-\rho).h\,.P$ (where c, ρ, h and P are the inventory holding cost per unit per unit time, system utilization rate, demand lead time and production speed respectively).

Liberopulos (2008) shows that the trade-off between the optimal order-base-stock level and demand lead time is linear for single-demand-class M/M/1 and M/D/1 production-inventory systems. It is also shown for generalized queuing systems that the optimal order-base-stock vanishes if the demand lead time is sufficiently long.

The literature discussed so far in this section considers that all orders provide the same amount of ADI, so these are single-demand-class problems. The following works address multiple demand class problems.

Iravani et al. (2007) consider a capacitated supplier serving both primary and secondary customers. The primary customer's demand needs to be fulfilled at the end of each cycle, but the order volume becomes known only after an elapse of a fixed amount of time in each cycle. The secondary customer's demand for a single item occurs randomly. The benefit of fulfilling the primary customer's demand is more than that for the secondary customers. If the orders are not fulfilled in due time, they are backordered for the primary customers, and they are lost for the secondary customers. The researchers show the optimality of threshold-type policies, which are monotone with primary customer's order volume for both average profit and discounted profit cases.

Duran et al. (2008) model the two-demand-class inventory allocation problem where the item prices and lost-sales costs vary with the demand class. It is possible to delay order fulfillment of one class till the next period, but the other class does not accept any delay. They model the problem as a Markov decision process and show the optimality of threshold type policies for production, reserving and backlogging decisions.

No work in the existing literature seems to consider the exact problem in production-inventory systems for either scenario (i.e., Scenario 1 — satisfying a mix of ADI and non-ADI orders with backordering, and Scenario 2 — satisfying a mix of flexible and inflexible ADI orders), which are addressed in this research.

In the context of pure inventory systems (Scenario 3), previous works characterize the optimal policies for the following cases: (i) multiple demand classes, but no early fulfillment (problem is trivial for uncapacitated systems with no ordering cost. The optimal policies can be characterized when either an ordering cost, or a capacity limit, or both are present); (ii) single demand class, but early fulfillment is acceptable. The literature also extends the analysis in case (ii) to a multiple-demand-class scenario (where the arriving orders in the current period may be due in the same period, or they may be due in any future period within the span of the maximum demand lead time set for a particular problem), but the optimal policy has not been characterized (though, protection-level heuristic policies are proposed). The literature assumes a general problem with virtually any number of classes for this scenario, whereas the current research considers only the two-class problems.

3. Approach

3.1 Production-Inventory Systems

Even in absence of early fulfillment, the development of an optimal policy for a single demand class system with endogenous replenishment time requires detailed information on future order arrivals and the timing of future due dates, and can be very complicated (Karaesmen et al., 2004). A simulation model will be developed in this context, and experiments will be performed to study how well the policy works under various conditions.

The first of the three policies described in this study is a Kanban-based make-to-order policy built upon Claudio and Krishnamurthy (2009). The second policy is a one-for-one replenishment CONWIP policy, where performance improvement is obtained solely from early fulfillment. This policy is based on order-base-stock policy proposed by several researchers (e.g., Karaesmen et al., 2004). The final one is the proposed policy that combines early fulfillment with partial make-to-order policy.

In each of the first two scenarios, two possible cases will individually be considered for two demand classes: (i) same priority/backorder-penalty and (ii) different priority/backorder-penalty. For each case, early fulfillment decisions should be governed by some sort of rationing policy. Otherwise, the service level of shorter demand lead time orders (or orders with no flexibility) will be severely affected. Moreover, a second kind of rationing is used in (ii) to prioritize the fulfillment of higher backorder cost demands.

3.1.1 Policy 1: Partial Make-to-order

In the partial make-to-order policy, production is initiated upon order arrival, as long as the current FG + WIP level is less than a specified limit. The available stock is used for immediate fulfillment of non-ADI orders. ADI orders are fulfilled on their due dates, if stock is available. A schematic representation is provided in Figure 5(a).

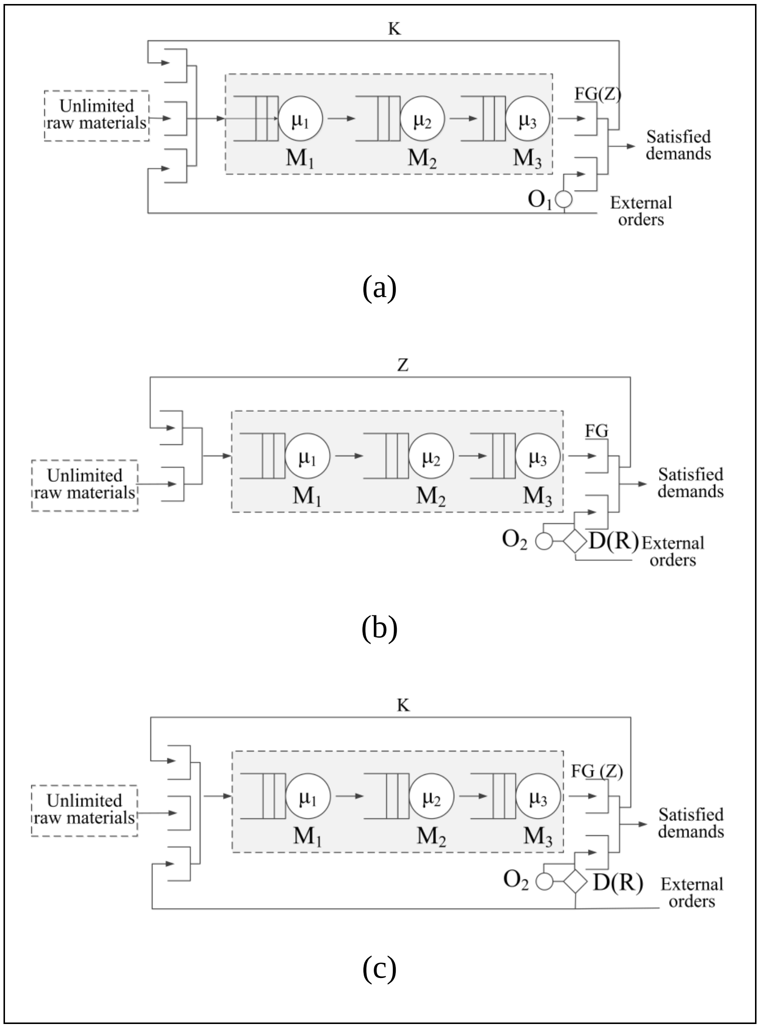


Figure 5. Production-inventory control policies for utilizing advance demand information.

To get the benefit of ADI, this policy employs a number of Kanban cards (K) for controlling production but maintains a separate initial base-stock level (Z); $K \geq Z \geq 0$. A Kanban card remains attached to every item in base-stock, thus initially ($K-Z$) Kanban card sets remain free. To release raw materials for production, both activated orders and free Kanban cards are necessary. Once an item is utilized to fulfill an order, the corresponding Kanban card becomes free. It is then capable of initiating further production upon future order arrivals. The quantity of Kanban cards equals the maximum FG + WIP limit, and the initial base-stock level Z is an integer in $[0,K]$. The original policy (Claudio and Krishnamurthy, 2009) starts production in max (0, τ_1–LT) time after an ADI-order arrival (LT is the average flow-time for jobs in the facility when ADI is not utilized but assumed available). This delayed production release policy is seen to work only for systems with low utilization, low backorder cost and high demand lead time. However, implementation of this policy in the current work starts production immediately upon order arrival, as it is observed that immediate production release is always better than delaying production. No ADI order is satisfied before spending time τ_1 in delay station O_1.

3.1.2 Policy 2: Early Fulfillment with Fixed-level Threshold Rationing

In this policy, ADI orders are fulfilled early when the current inventory level is sufficiently high (i.e., above threshold). Otherwise, exact fulfillments of those orders are carried out. Figure 5(b) provides a schematic representation of this policy.

The policy starts production only upon actual consumption of items from the inventory. Thus, it is a CONWIP-type policy, where (FG + WIP) level is always constant and the system has no way to make items to order. The policy is defined by two parameters: the base-stock level Z and the rationing level R, $Z \geq R \geq 0$. Each non-ADI order is fulfilled immediately upon arrival, as long as stock is available. Each ADI order is also fulfilled immediately if the current FG inventory level is above R; otherwise, the order waits for the first of the two events to occur: (i) the order becomes due, or (ii) the FG inventory level goes above R. This determines the waiting time of ADI orders at delay station O_2, as shown by decision block D(R) in Figure 5(b). The probability of early fulfillment in this policy increases with a reduced value of R. When R = 0, no rationing is used and all items in stock are available for early fulfillment.

3.1.3 Policy 3: Pairing of Previous Two Approaches

A policy is proposed, where early fulfillment complements the make-to-order strategy. As shown in Figure 5(c), this Kanban-based partial-make-to-order policy is governed by three parameters: K, Z and R (all as previously described). The relationships mentioned previously (i.e., $K \geq Z \geq 0$ and $Z \geq R \geq 0$) also hold for this policy.

As with Policy 1, this policy also initiates new production upon order arrival, as long as a free Kanban card is available. Reduced base-stock levels can be used, as $K \geq Z$ and some items are produced in make-to-order mode to meet demand. Again, ADI orders are fulfilled early, when the current FG level is above R (as with Policy 2). Therefore, O_2 and D(R) are exactly the same as those in Policy 2. By pairing the two strategies, it is possible to take advantage of both. Each of the first two policies reduces inventory holding cost, but in different ways. Early fulfillment clears stock sooner, whereas make-to-order replenishes stock later. The synergy of these two approaches should thus result in a well-performing policy.

Although the three policies are operationally different, all reduce to the same conventional system (no make-to-order, no early fulfillment) under following configurations: (i) Policy 1 with $K = Z = c$, (ii) Policy 2 with $Z = R = c$, and (iii) Policy 3 with $K = Z = R = c$, where c is any positive integer. Also, Policy 3 reduces to the classical base-stock policy when $K = R = \infty$. Therefore, the selection of parameter values for these policies will play a pivotal role in policy performance.

3.1.4 Determination of a Lower Bound on Total Cost

Early fulfillment is beneficial since it lowers average holding cost due to items spending less time in FG inventory. However, early fulfillment may also create stock-outs and thus enhance the risk of backordering in the presence of order crossovers. These two contradicting objectives make it difficult to derive an expression for a lower bound on total cost. By employing a modified version of the *allocation assumption* made by Wang and Toktay (2008), however, a simulation-based approach to a lower bound can be established. The assumption is very simple: items already consumed for early fulfillment can be recalled and utilized to fulfill urgent orders. Therefore, the lower bound policy (termed as LB) would fulfill all ADI orders as soon as they arrive, then retrieve a portion of already consumed items to fulfill urgent orders, which otherwise can no longer be fulfilled on their respective due dates. While not practically possible, the assumption nevertheless allows early fulfillment only when it is advantageous to do so. The approach is relatively straightforward for the case of continuous review with endogenous replenishment time, when the demand lead times provided by two customer classes are either positive constants or zero (as assumed in this research).

The operating procedure of LB is the same as Policy 3, with following two exceptions: (i) rationing level $R = 0$ and (ii) when a non-ADI order arrives at time t_{noADI} during stock-out, the policy searches for any previously fulfilled ADI order such that the arrival time of that ADI order (t_{ADI}) satisfies the condition $t_{\text{ADI}} \le t_{\text{noADI}} < t_{\text{ADI}} + \tau_1$. If one such ADI order is found, the allocation assumption is used to take back the item from ADI order for fulfilling the non-ADI order. Later, the ADI order can be fulfilled again when the stock becomes available. Swapping items this way between ADI orders and non-ADI orders reduces the combined backorder cost. In case several ADI orders satisfy the above condition, the last arrived ADI order is chosen to for retrieving the item (this keeps the combined backorder cost at a minimum). If no crossover between an early fulfilled ADI order and the current non-ADI order occurs, however, no item is retrieved from previously fulfilled orders: rather, the non-ADI demand is backordered until replenishment stock becomes available. For fulfilling backordered demand, the earliest-due-date rule is followed irrespective of the class of backorder. It can be noted here that the applicability of the lower bound evaluation is limited to the case, where backorder costs of the demands are same. For the equal-backorder-cost case, as items are reallocated to newly-arrived non-ADI orders, production replenishment time should be recalculated in a strict sense. Avoiding this complex calculation makes the analysis simpler (but not incorrect as it gives only a lower bound cost). Simulation, using this assumption, can then be used to establish the bounds for any given setting.

3.2 Pure Inventory Systems

The pure inventory scenario described in Section 1.4 is analytically more tractable than the previous two scenarios. Thus, the approach is to analytically characterize the optimal policy in Scenario 3, rather than employing simulation. In the presence of multiple demand classes, allowing early fulfillment makes the analysis more complicated than that for a system without early fulfillment. Wang and Toktay (2008) show that the characterization of optimal policy for any number of classes is difficult when early fulfillment is considered. The problem with two demand classes is the simplest multiple-class problem. An analytical model based upon dynamic programming may provide some general insights, which helps the decision maker to choose a suitable class of policies.

The intention is to use backward recursion to characterize the structure of the optimal policy. Without doing any numerical evaluations, here the optimal decision rules for each period will be found.

In reality, customers should receive some incentives both for providing larger ADI and for accepting early fulfillment. In this scenario, Class 1 demand provides smaller amount of DLT, but agrees to accept early fulfilment at the discretion of the inventory manager. The Class 2 demand does not accept early fulfilment, but it provides a larger ADI.

4. Production-Inventory Systems: Methodology

A simulation model was developed to evaluate the performance of three policies (along with their lower bounds) in Scenarios 1 and 2. We assume the demand lead time is τ_1, and order arrival is Poisson process for both demand classes. Processing time is exponential with mean μ^{-1} at each stage. Poisson arrival processes and exponential processing times are assumed by many researchers in past (e.g., Bonvik et al., 1997, Liberopoulos and Koukoumialos, 2005). The assumptions described in section 1.4 hold for the systems. This work is an extension of Sarkar (2007), where an improvement in order fill rate is shown by using ADI together with early fulfillment for pull-production systems controlled by two sets of Kanban cards. The first set of Kanban cards controls production, whereas the second set puts a limit on the frequency of early fulfillment. That work assumes lost sales instead of backordering, in case the demands cannot be fulfilled from the on-hand inventory.

4.1 Simulation Model

The algorithm for discrete-event simulation is developed in Visual C++ (Microsoft Corporation 2010). Figure 6 shows a basic flowchart of this algorithm, which employs three events: order arrival, order fulfillment (including early fulfillment), and production release. The model is validated by comparing the performance of existing control policies.

The specific logic for each event depends upon the policy being used (e.g., Policy 1 does not allow early fulfillment, Policy 2 cannot release production until order fulfillment, etc.).

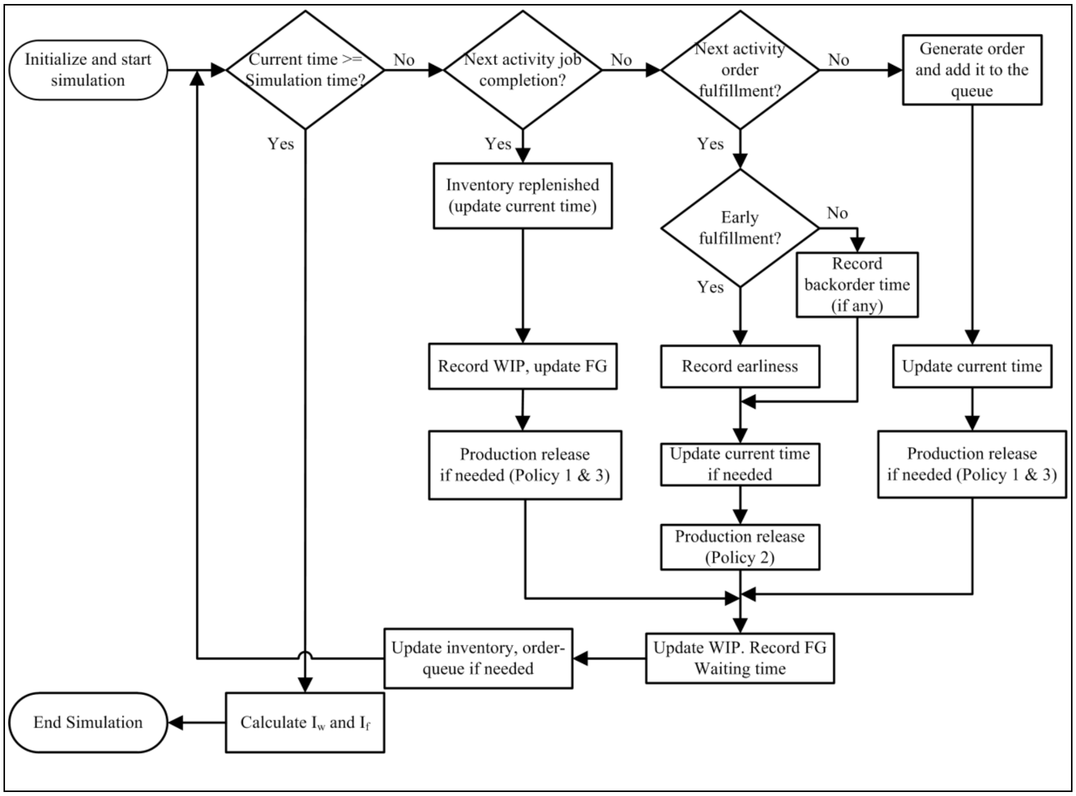


Figure 6. Flowchart of simulation algorithm for evaluating policy performance.

4.2 Experimental Design

Various settings (depending upon the external factors, such as, combined order arrival rate, DLT of both classes, cost structure, arrival ratio of order classes, SLT, etc.) are considered for performance evaluation. For each scenario, several experiment sets are formed. The objective for each experiment set is to identify the change in performance with a change in the settings for different policies. Each combination of the external factors constitutes a simulation experiment, and a group of experiments form a set. For each experiment, optimal values of the policy control parameters (K, Z and R, as applicable) are employed: these are established via simulation optimization. In other words, a range of design alternatives (determined by the control parameter values) are considered for each experiment. A fixed number of replications are performed for each design alternative in all experiment sets.

A total of nine experiment sets were used: values employed for establishing the different settings and cost structures are shown in Table 3[1]. As described earlier, LT is the average supply lead time for given λ and μ, when ADI is not employed: these values are found via simulation in the same manner as the main experiments described in this section. The service rate μ = 1 is assumed for each stage (except for Experiment Set 9, where expected supply lead time is changed). Some of the settings are parallel to those used in Claudio and Krishnamurthy (2009); but unlike the previous work, the optimal values of the parameters (K^*, Z^*, R^*, as applicable) are found in each setting instead of assuming arbitrary values for them. The relative values of the cost factors (w, h and b) are important, as the ranking of various policies can change as a result of large changes in cost for a given setting. Backorder cost, which is incurred in the form of reduced future orders because of damage to customer goodwill, is hard to estimate. A review of the literature was performed to determine suitable values of b/h (i.e., the relative importance of backorder cost compared to holding cost) to use. Choi et al. (2008) use five different values of b/h ratios: 1, 5, 15, 20 and 25, whereas Toktay and Wein (2001) use the values 10 and 50. Based upon these findings, three values are chosen, which are not too extreme: 10, 20 and 40. The w/h ratio (i.e., FG inventory cost to WIP holding cost ratio) is assumed to be ½, and the argument for validity of this value is as follows. Let w_0 be the cost of WIP at the beginning of a production process, whereas h is the cost when the production is completed. Using the arithmetic mean of these values as the average WIP cost seems reasonable. Assuming w_0 is zero, w/h = ½, which is the minimum possible value of this ratio. Thus, with three b/h ratios, and a single w/h ratio, two

[1] Experiment sets 1 and 2 are reported in Sarkar and Shewchuk (2012).

cost factor combinations are employed. Actual cost values are calculated, in all cases, assuming w = 1. In subsequent experiments, arrival ratios (r) and relative backorder costs (b_1, b_2) of the demand classes are changed. The ratio r is defined as the fraction of Class 1 orders in all arrived orders. Backorder cost b_1 corresponds to the Class 1 orders, whereas b_2 is that for Class 2.

Table 3. Experimental Design.

Scenario	Experiment Set	λ	τ_1	τ_2	μ	$(w:h:b)$ or $(w:h:b_1:b_2)$	r (%)
1	1	0.5, 0.6, 0.7, 0.8	0, 0.5LT, 1LT, 2LT	0	1	1:2:20	50
	2	0.5, 0.6, 0.7, 0.8	0, 0.5LT, 1LT, 2LT	0	1	1:2:40	50
	3	0.5, 0.6, 0.7, 0.8	0, 0.5LT, 1LT, 2LT	0	1	1:2:10	50
	4	0.5, 0.6, 0.7, 0.8	1LT	0	1	1:2:20	0, 25, 50, 75, 100
	5	0.5, 0.6, 0.7, 0.8	1LT	0	1	1:2:10:100, 1:2:10:10, 1:2:100:10	50
2	6	0.5, 0.6, 0.7, 0.8	1LT	1LT	1	1:2:10:100	25, 50, 75
	7	0.5, 0.6, 0.7, 0.8	1LT	1LT	1	1:2:10:100, 1:2:100:10	50
	8	0.5, 0.6, 0.7, 0.8	1LT, 1.15LT	1LT, 1.15LT	1	1:2:20:20	100
	9	0.5, 0.6, 0.7, 0.8	1LT	1LT	1.09,1.07, 1.06, 1.04	1:2:20:20	100

λ, order arrival rate; τ_1, demand lead time of Class 1 orders; τ_2, demand lead time of Class 2 orders; μ, service rate at each stage; w, cost of WIP; h, cost of FG inventory; b, backorder cost when order classes have equal priority; b_1 and b_2 are backorder costs for Class 1 and Class 2 orders respectively, when one class has higher priority than the other; all costs are measured per unit, per time; r, arrival ratio is the fraction of Class 1 orders among all orders; LT, average job flow-time. In Set 9, when λ = 0.5, μ = 1.09; λ = 0.6, μ = 1.07; λ = 0.7, μ = 1.06; λ = 0.8, μ = 1.04;

4.3 Implementation

For each simulation experiment, five replications are performed. For the most promising cases (as explained in the following section), each replication has a run length of 100,000 time units and a warm-up time of 100 units. Common random numbers are used as a variance reduction technique (e.g., Banks et al., 2010). For each experiment, 95% confidence intervals are established for optimal total costs of all policies, which are found to be within 2% of the mean estimates for respective optimal costs. The warm-up time is determined by using Welch's graphical procedure.

To perform simulation optimization, implicit enumeration is employed. The parameters are discrete, and the negative of the cost function is observed to be concave and unimodal. Thus, a relatively small set of possible combinations of parameter values needs to be tested. For example, Policy 1 has two parameters (K and Z). Let's assume K^* is optimal ($\forall\ Z$ in $K \geq Z \geq 0$) in a given setting and cost factor. The total cost is observed (Figure 7) to increase monotonically as K deviates further on either side of K^* ($\forall Z$ in $K \geq Z \geq 0$). Z is an integer with an upper limit K. So, the set $\boldsymbol{K}$ = $\{K_1,\ K_1 + 1,\ldots,\ K_2;\ K_1 < K^* < K_2\}$ is formed and the optimal parameter values are found by evaluating cost for each (K, Z) combination, where $K \in \boldsymbol{K}$ and $Z \in [0,K]$. The values of K_1, K_2 are determined by observing the trend of cost change with parameters for a few test runs. Similarly, optimal parameter values for Policy 2 and Policy 3 are determined. As the ranges for control parameter values widen, more efforts are needed for conducting these simulation experiments. For an example, if we consider the ranges of control parameter values to be $\boldsymbol{K}$ = [15,40], $\boldsymbol{Z}$ = [10,30], and $\boldsymbol{R}$ = [0,5], then we need to run 2556 number of cases. Thus, efficiency becomes a big concern particularly for the design alternatives with higher number of control parameters and larger ranges for these parameter values. While narrowing down the range by intuition or use of variance reduction methods can be helpful in reducing computational effort, it is still possible to further reduce it by implementation of an intelligent control of the simulation experiments. A more efficient procedure for discrete-event simulation-optimization is described in the following section, which is used in this research. This procedure, known as Optimal Computing Budget Allocation (OCBA) is outlined in Chen et al. (2000). Later, Lee et al. (2004) and Pujowidianto et al. (2009) extend it to multiple-objective and stochastic-constraint optimization problems respectively.

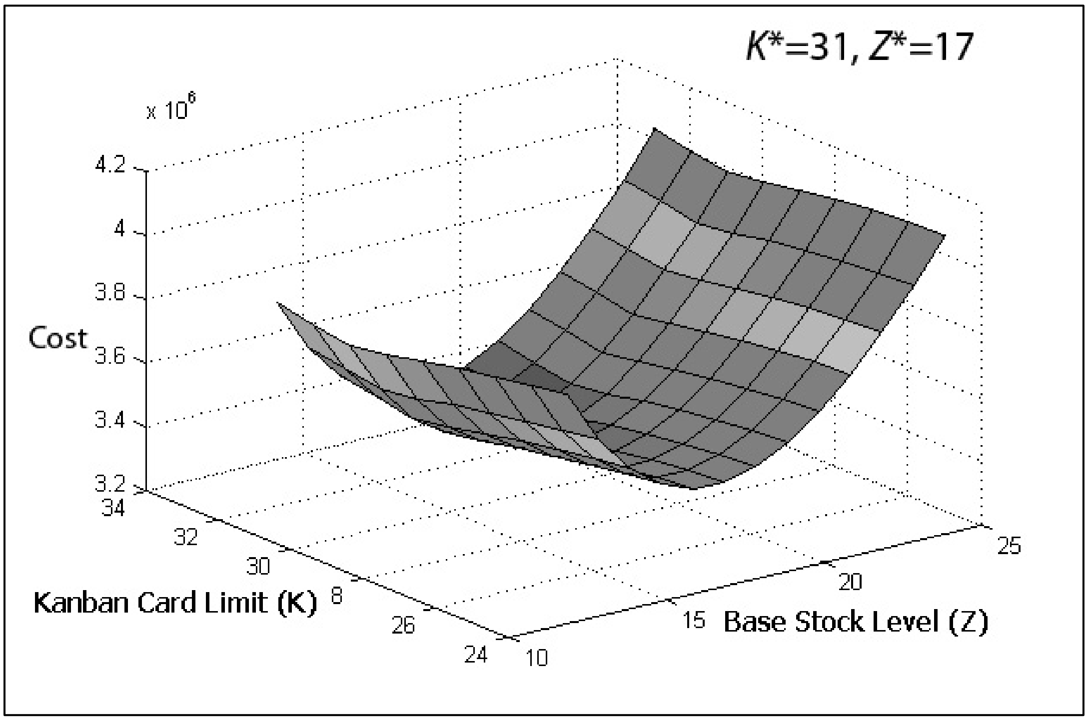


Figure 7. A typical response surface of total cost using various control parameter values.

4.3.1 Fundamentals of OCBA

This budget allocation procedure minimizes the computation time to select the best design from a discrete number of design alternatives. The underlying notion to achieve a high probability of correct selection is allocating a larger portion of the computing budget to more promising design alternatives. Using an iterative process based upon the information on both relative means and variances for the alternatives, it uses only a small fraction of the computation time needed for conventional equal allocation procedure.

A general problem statement for OCBA (as described in Chen et al., 2000) is the following:

$$\min_{\theta_i \in \Theta} J(\theta_i) \equiv E[L(\theta_i, \xi)] \tag{3}$$

Where,

Θ is the search space (finite set);

θ_i is design alternative i, i= 1, 2, …, k, where k number of alternatives are formed by various control parameter values.

J is the performance criterion (such as, total cost in our case);

L is the sample performance;

ξ is a random vector representating uncertainty in the systems.

It has been assumed that (i) the system constraints are implicitly involved in the simulation, and (ii) performance can be measured only through simulation.

$\bar{J}_i$, which is the estimated performance of design i, can be found by

$$\bar{J}_i \equiv \frac{1}{N_i} \sum_{j=1}^{N_i} L(\theta_i, \xi_{ij}) \tag{4}$$

The conventional simulation approach based upon equal allocation requires a large number of samples for all designs, resulting in more computational effort. However, OCBA determines the optimal number of simulation samples (or run length) for each design to maximize the probability of correct selection of the best alternative for a given total effort. The following theorem (Chen et al., 2000) governs the rule for OCBA:

Given a total number of simulation samples T to be allocated to k competing designs whose performance is depicted by random variables with means $J(\theta_1)$, $J(\theta_2)$, …, $J(\theta_k)$, and finite variances σ_1^2, σ_2^2, … , σ_k^2 respectively, as $T \rightarrow \infty$, the Approximate Probability of Correct Selection (APCS) can be asymptotically maximized when

(a) $$\frac{N_i}{N_j} = \left(\frac{\sigma_i / \delta_{b,i}}{\sigma_j / \delta_{b,j}}\right)^2, i,j \in \{1, 2, \dots, k\} and\ i \neq j \neq b \quad (5)$$

(b) $$N_b = \sigma_b \sqrt{\sum_{i=1, i \neq b}^{k} \frac{N_i^2}{\sigma_i^2}} \quad (6)$$

Where N_i is the number of samples allocated to design i, $\delta_{b,i}$=$\bar{J}_b - \bar{J}_i$, and $\bar{J}_b \leq \min_i \bar{J}_i$.

4.3.2 Implementation of the OCBA Algorithm

Based on the above theorem, an efficient iterative approach to select the best design among k alternatives can be formulated. In the first stage, less effort is spent on each of the k designs to get some estimation of the performance of each design. The sample mean and sample variance are computed from these initial sets of simulation runs. In the successive stages, incremental computing budget is allocated to each design based upon the above theorem. As mentioned earlier, five replications, each with 100,000 time units of run length is used in our experiments to get the necessary confidence interval. Because of the presence of smaller sample size with relatively longer run length, it is more efficient to use shorter run length (than using less number of replications) in the initial iterations for implementing the OCBA algorithm in the present case. The algorithm is summarized as follows:

Initialization:

$$(\alpha_1, \alpha_2, \dots, \alpha_k) = (\frac{l}{k}, \frac{l}{k}, \dots, \frac{l}{k}) \quad (7)$$

where l is the aggregate run length for one replication allocated to k designs, α_i is the run length of one replication for design i.

Step 1: run five replications (each with run length α_i) for each design i ($i \in k$).

Step 2: calculate $\bar{J}_i$, $\delta_{b,i}$ and σ_i.

Step 3: estimate the best system as $\hat{b} = argmin\{\bar{J}_i : i = 1,2, \dots, k\}$

Step 4: calculate $\alpha_1, \alpha_2, \dots, \alpha_k$ from the following equations ($\alpha_b$and σ_b^2 respectively denote the run length and the variance corresponding to the estimated best system):

$$\frac{\alpha_i}{\alpha_j} = \left(\frac{\sigma_i / \delta_{b,i}}{\sigma_j / \delta_{b,j}} \right)^2 \quad (8)$$

$$\alpha_b = \sigma_b^2 \sum_{i=1, i \neq b}^{k} \frac{\alpha}{\sigma_i^2} \quad (9)$$

Step 5: Increase the computing budget for the next stage by a factor c ($c > 1$), i.e., $l = c.l$. Stop when $l \geq C$, where C is the limit on the computing budget in the final stage.

5. Production-Inventory Systems: Results and Discussion

5.1 Experiment Sets 1, 2 and 3

Results of the Experiment Sets 1, 2, and 3 are shown in Table 4, Table 5, and Table 6. Each table shows, for a given cost factor combination, the optimal control parameter values (K^*, Z^*, R^*, as applicable) and resulting performance (WIP, FG inventory and backorder levels) for each setting (λ, τ_1) and policy (1, 2 or 3). Table 4 provides these results for the first cost factor combination (1:2:20), Table 5 for the second (1:2:40), and Table 6 for the third cost factor combination (1:2:10). For each cost factor combination, Figure 8, Figure 9 and Figure 10 show the total cost for each setting/policy. The lower bound on total cost for each setting/policy is also shown in each figure. Table 7 shows the order fill rates, average WIP, and FG in each policy when the control parameter values are optimal for the first cost factor combination.

Table 4. Results for comparison of three policies, first cost factor combination.

λ	τ_1	Policy 1 parameters and states		Policy 2 parameters and states		Policy 3 parameters and states	
		K^*, Z^*	$E(I_w)$, $E(I_f^+)$, $E(I_f^-)$	Z^*, R^*	$E(I_w)$, $E(I_f^+)$, $E(I_f^-)$	K^*, Z^*, R^*	$E(I_w)$, $E(I_f^+)$, $E(I_f^-)$
0.5	0	10, 7	3.0, 4.1, 0.13	7, 0	2.9, 4.1, 0.15	10, 7, 0	3.0, 4.1, 0.13
0.5	1LT	10, 5	3.0, 3.6, 0.14	6, 1	2.8, 3.2, 0.12	10, 5, 1	3.0, 2.4, 0.16
0.5	2LT	9, 4	2.9, 4.0, 0.09	5, 1	2.7, 2.3, 0.14	9, 4, 2	3.0, 2.0, 0.11
0.6	0	15, 9	4.5, 4.7, 0.27	10, 0	4.4, 5.6, 0.20	15, 9, 0	4.5, 4.7, 0.27
0.6	1LT	15, 7	4.5, 4.9, 0.22	8, 1	4.1, 3.8, 0.22	15, 7, 1	4.5, 3.1, 0.24
0.6	2LT	15, 4	4.5, 4.1, 0.26	8, 2	4.1, 3.9, 0.11	15, 6, 2	4.5, 2.7, 0.14
0.7	0	22, 14	7.0, 7.3, 0.36	14, 0	6.8, 7.2, 0.39	22, 14,0	7.0, 7.3, 0.36
0.7	1LT	22, 11	7.1, 7.3, 0.31	13, 2	6.6, 6.4, 0.24	22, 11, 2	7.1, 4.8, 0.32
0.7	2LT	22, 7	7.0, 6.5, 0.35	11, 2	6.3, 4.7, 0.25	21, 9, 2	7.0, 3.5, 0.24
0.8	0	30, 23	12, 11, 0.60	24, 0	12, 12, 0.53	30, 23,0	12, 11, 0.60
0.8	1LT	29, 19	12, 12, 0.52	22, 2	12, 10, 0.37	28, 20, 2	12, 8.8, 0.45
0.8	2LT	28, 13	12, 10, 0.65	19, 2	11, 7.9, 0.38	29, 16, 2	12, 5.8, 0.42

λ, order arrival rate; τ_1, demand lead time of Class 1 orders; Policy 1, make-to-order approach; Policy 2, early fulfillment approach; Policy 3, the proposed approach for using ADI; K^*, Z^* and R^* are optimal values of respective policy parameters to minimize total cost for cost factor combination w: h: b = 1:2:20. $E(I_w)$, expected WIP; $E(I_f^+)$, expected FG inventory; $E(I_f^-)$, expected backorder level for various settings; LT, average job flow-time. See Table 3 for experimental design.

Table 5. Results for comparison of three policies, second cost factor combination.

λ	τ_1	Policy 1 parameters and states		Policy 2 parameters and states		Policy 3 parameters and states	
		K^*, Z^*	E(I_w), E(I_f^+), E(I_f^-)	Z^*, R^*	E(I_w), E(I_f^+), E(I_f^-)	K^*, Z^*, R^*	E(I_w), E(I_f^+), E(I_f^-)
0.5	0	10, 8	3.0, 5.1, 0.08	8, 0	2.9, 5.0, 0.09	10 / 8 / 0	3.0, 5.1, 0.08
0.5	1LT	10, 6	3.0, 5.5, 0.05	7, 2	2.8, 4.2, 0.06	11 / 6 / 2	3.0, 3.4, 0.09
0.5	2LT	9, 4	2.9, 4.0, 0.10	6, 2	2.7, 3.3, 0.06	10 / 5 / 2	3.0, 2.7, 0.06
0.6	0	17, 11	4.5, 6.6, 0.13	11, 0	4.4, 6.6, 0.14	17 / 11 / 0	4.5, 6.6, 0.13
0.6	1LT	16, 9	4.5, 6.7, 0.10	10, 1	4.4, 5.6, 0.08	17 / 9 / 2	4.5, 4.8, 0.10
0.6	2LT	15, 6	4.5, 5.9, 0.12	9, 2	4.2, 4.8, 0.06	15 / 7 / 2	4.5, 3.3, 0.10
0.7	0	22, 17	7.0, 10, 0.16	17, 0	6.9, 10, 0.16	22 / 17 / 0	7.0, 10, 0.16
0.7	1LT	22, 13	7.0, 9.2, 0.18	14, 2	6.7, 7.3, 0.17	22 / 14 / 2	7.0, 7.4, 0.14
0.7	2LT	22, 9	7.0, 8.4, 0.20	13, 3	6.5, 6.5, 0.11	22 / 11 / 3	7.0, 5.1, 0.13
0.8	0	34, 27	12, 15, 0.30	27, 0	12, 15, 0.31	34 / 27 / 0	12, 15, 0.30
0.8	1LT	30, 23	12, 15, 0.27	24, 3	12, 12, 0.24	30 / 23 / 2	12, 11, 0.27
0.8	2LT	31, 18	12, 15, 0.28	22, 3	11, 11, 0.17	30 / 20 / 3	12, 9.0, 0.20

K^*, Z^* and R^* are optimal values of respective policy parameters for cost factor combination w: h: b = 1:2:40; Other notations are same as those in Table 4. See Table 3 for experimental design.

Table 6. Results for comparison of three policies, third cost factor combination.

λ	τ_1	Policy 1 parameters and states		Policy 2 parameters and states		Policy 3 parameters and states	
		K^*, Z^*	E(I_w), E(I_f^+), E(I_f^-)	Z^*, R^*	E(I_w), E(I_f^+), E(I_f^-)	K^*, Z^*, R^*	E(I_w), E(I_f^+), E(I_f^-)
0.5	0	10, 5	3.0, 2.4, 0.35	6, 0	2.8, 3.2, 0.24	10, 5, 0	2.9, 2.4, 0.35
0.5	1LT	11, 3	3.0, 1.9, 0.37	5, 1	2.6, 2.3, 0.21	10, 4, 1	2.9, 1.7, 0.24
0.5	2LT	7, 3	2.9, 2.9, 0.19	5, 1	2.6, 2.3, 0.12	8, 3, 1	2.9, 1.2, 0.22
0.6	0	12, 7	4.4, 3.1, 0.53	8, 0	4.1, 3.8, 0.42	12, 7, 0	4.4, 3.1, 0.53
0.6	1LT	12, 5	4.4, 3.2, 0.41	7, 1	4.0, 3.0, 0.30	13, 6, 1	4.4, 2.4, 0.30
0.6	2LT	12, 3	4.4, 3.3, 0.33	6, 1	3.8, 2.2, 0.31	13, 4, 1	4.4, 1.3, 0.32
0.7	0	20, 11	6.8, 4.7, 0.70	12, 0	6.5, 5.5, 0.61	20, 11, 0	6.9, 4.7, 0.69
0.7	1LT	17, 7	6.8, 4.2, 0.71	10, 1	6.1, 3.8, 0.52	20, 9, 1	6.9, 2.8, 0.45
0.7	2LT	18, 4	6.8, 4.6, 0.56	9, 1	5.9, 3.1, 0.42	20, 6, 1	6.9, 1.6, 0.45
0.8	0	25, 19	11, 8.1, 1.02	20, 0	11, 8.9, 0.95	25, 19,0	11.5, 8.1, 1.02
0.8	1LT	23,13	11, 8.0, 0.98	16, 1	10, 5.9, 0.87	25, 14, 1	11.5, 4.4, 0.87
0.8	2LT	25, 12	11, 7.2, 1.1	15, 2	10, 4.9, 0.56	27, 11, 2	11.6, 3.0, 0.51

K^*, Z^* and R^* are optimal values of respective policy parameters for cost factor combination w: h: b = 1:2:10; Other notations are same as those in Table 4. See Table 3 for experimental design.

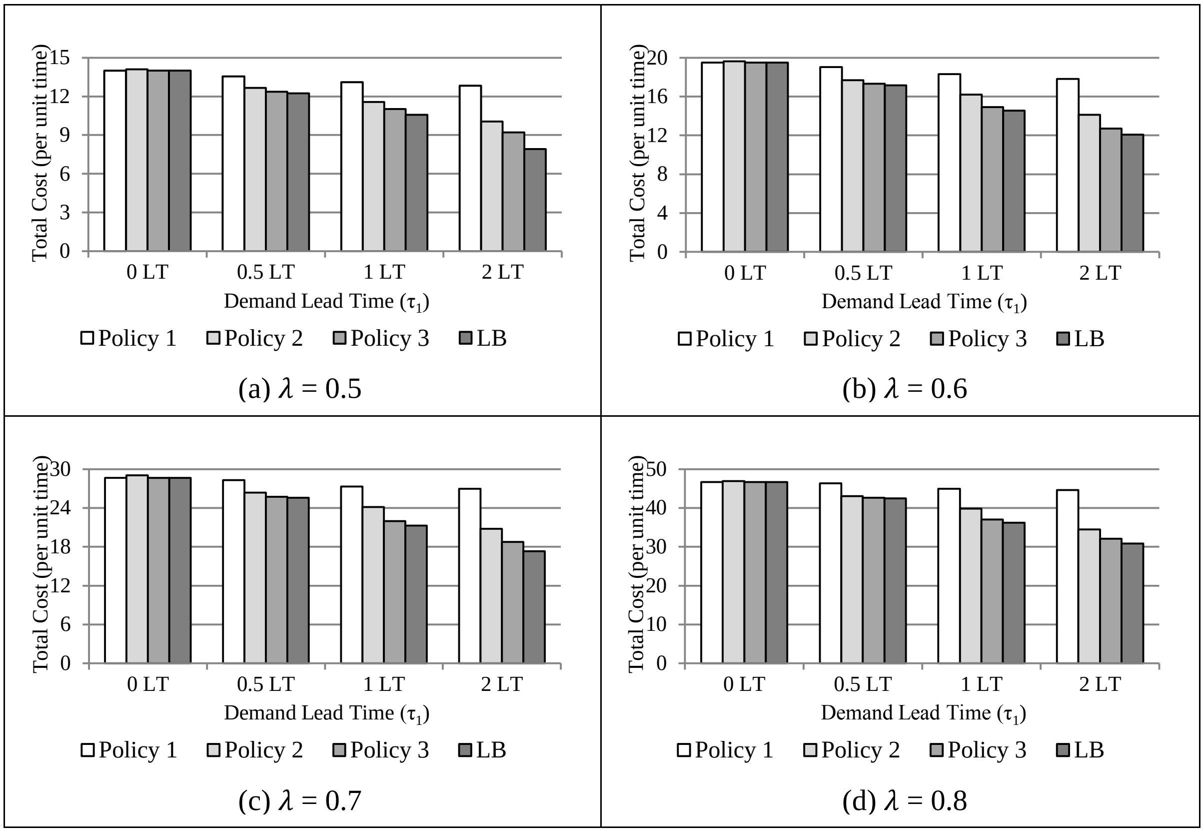


Figure 8. Optimal cost for various policies in different settings, first cost factor combination.

Table 7. Performance measures of three policies, first cost factor combination.

λ	τ_1	Policy 1 states Fill rate, avg. WIP, avg. FG	Policy 2 states Fill rate, avg. WIP, avg. FG	Policy 3 states Fill rate, avg. WIP, avg. FG
0.5	0	0.907, 3.00, 4.11	0.904, 2.91, 4.09	0.907, 3.00, 4.11
0.5	1LT	0.902, 3.00, 3.61	0.920, 2.80, 3.20	0.884, 3.00, 2.43
0.5	2LT	0.931, 2.94, 3.99	0.903, 2.66, 2.34	0.914, 2.97, 2.03
0.6	0	0.877, 4.53, 4.72	0.910, 4.39, 5.60	0.877, 4.53, 4.72
0.6	1LT	0.900, 4.53, 4.86	0.901, 4.18, 3.82	0.879, 4.53, 3.10
0.6	2LT	0.880, 4.53, 4.09	0.947, 4.07, 3.93	0.920, 4.53, 2.65
0.7	0	0.896, 7.05, 7.26	0.891, 6.80, 7.21	0.896, 7.05, 7.26
0.7	1LT	0.908, 7.05, 7.34	0.932, 6.62, 6.38	0.901, 7.05, 4.85
0.7	2LT	0.896, 7.05, 6.50	0.926, 6.31, 4.69	0.913, 7.04, 3.50
0.8	0	0.894, 12.05, 11.36	0.906, 11.79, 12.21	0.894, 12.05, 11.36
0.8	1LT	0.906, 12.01, 11.73	0.931, 11.53, 10.46	0.913, 11.98, 8.85
0.8	2LT	0.887, 11.97, 10.28	0.928, 11.11, 7.89	0.908, 12.02, 5.84

λ, order arrival rate; τ_1, demand lead time of Class 1 orders; values shown result in minimum total costs for respective policies, (cost factor combination w: h: b = 1:2:20); LT, average job flow-time. See Table 3 for experimental design.

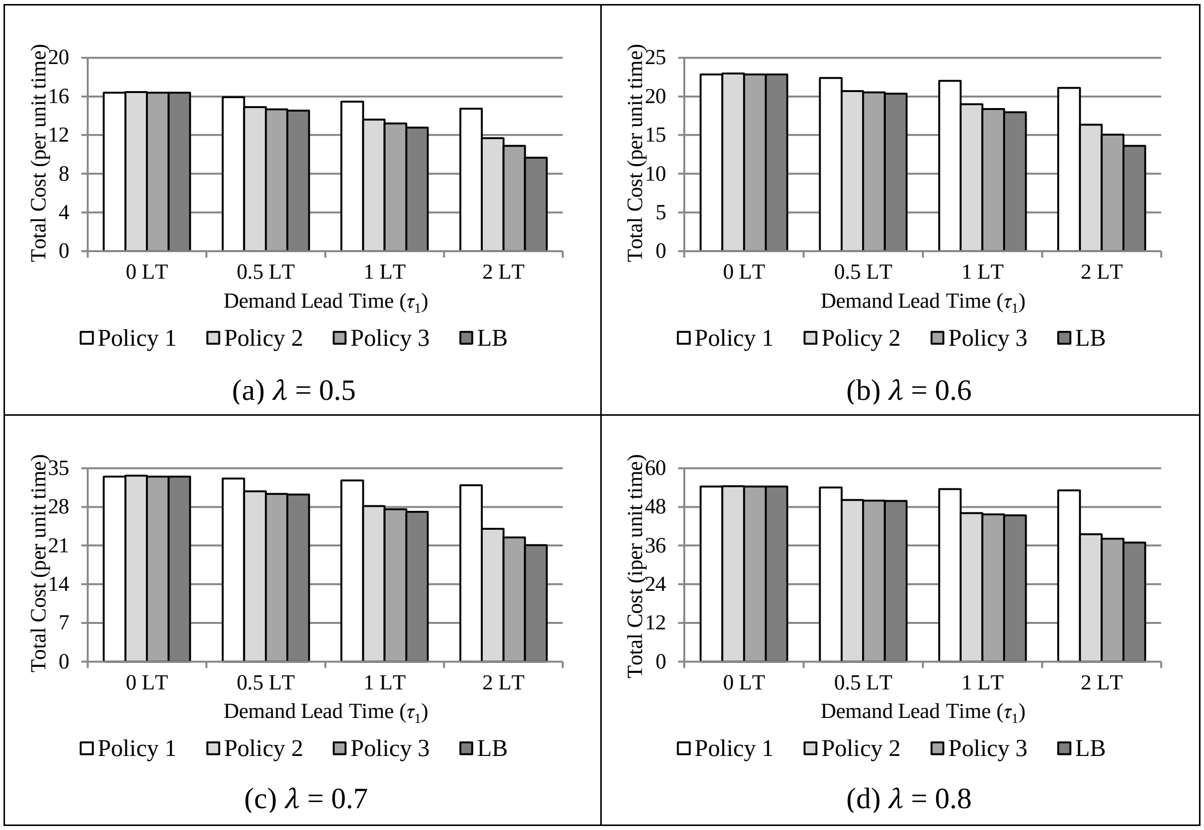


Figure 9. Optimal cost for various policies in different settings, second cost factor combination.

Table 4, Table 5 and Table 6 depict optimal configurations of the policies covering a broad range of order arrival rates, demand lead time and cost factors for three-stage facilities. It can be observed that the optimal number of Kanban cards (K^* in policies 1 and 3) increases with an increase in order arrival rate, but remains nearly unaffected as the amount of ADI changes. However, the base-stock level (Z^* in all policies) depends on both the arrival rate and ADI. This is because K^* is directly related to the average system utilization (the only way of keeping the system busier is more frequent production release), which cannot be changed by using ADI. However, ADI adds flexibility in the system that reduces safety stock requirement. The following is also observed:

$$K^* \geq \mathrm{E}(I_{\mathrm{w}}) + \mathrm{E}\left(I_{\mathrm{f}}^{+}\right) \geq Z^* \text{ for policies 1 and 3; } \mathrm{E}(I_{\mathrm{w}}) + \mathrm{E}\left(I_{\mathrm{f}}^{+}\right) = Z^* \text{ for Policy 2} \tag{10}$$

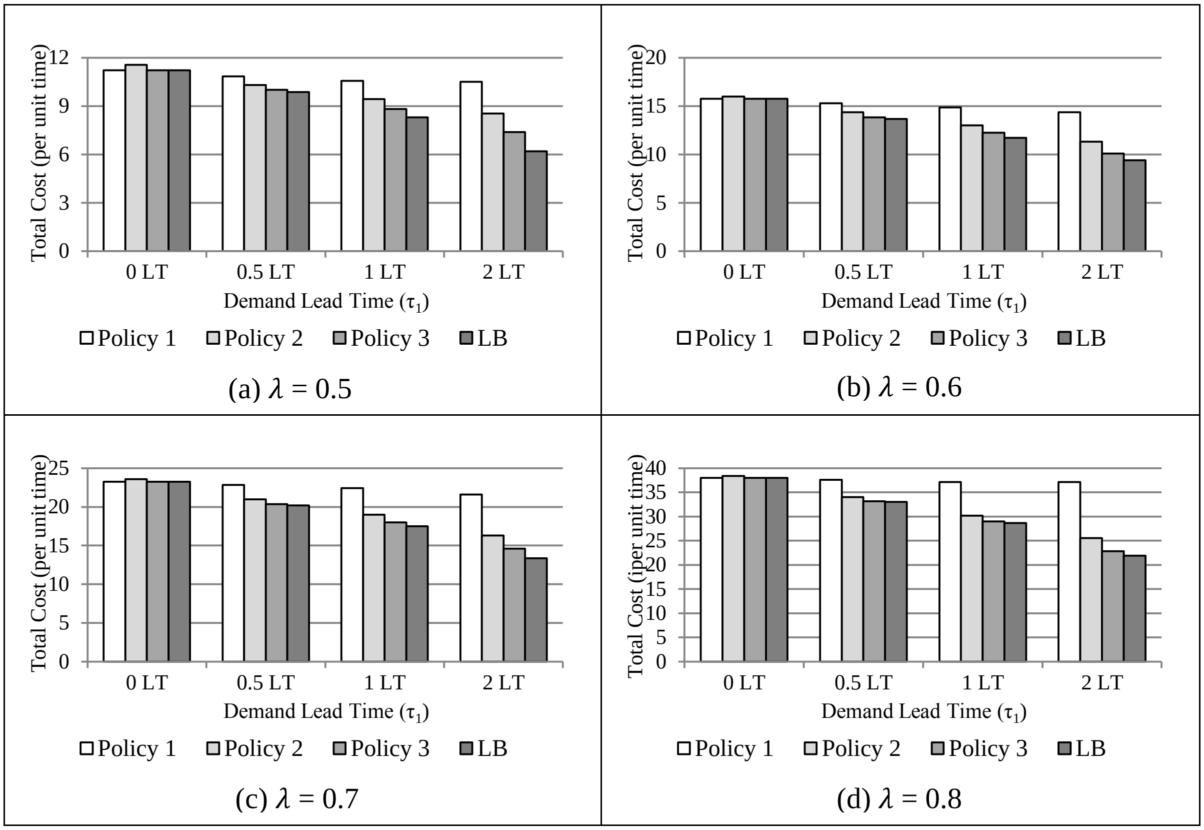


Figure 10. Optimal cost for various policies in different settings, third cost factor combination.

Policy 2 is a pure make-to-stock policy that performs early fulfillment under certain conditions. Thus, FG + WIP level always remains constant (equal to the base-stock level Z^*). Policies 1 and 3, however, are partial make-to-stock and partial make-to-order policies, respectively. Consequently, the average FG + WIP level varies between the base-stock and Kanban card limits. Additionally, the following inequality holds when $\tau_1 > 0$:

$$[\mathrm{E}(I_{\mathrm{w}})+\mathrm{E}(I_{\mathrm{f}}^{+})]_{\text{Policy 1}} > [\mathrm{E}(I_{\mathrm{w}})+\mathrm{E}(I_{\mathrm{f}}^{+})]_{\text{Policy 2}} > [\mathrm{E}(I_{\mathrm{w}})+\mathrm{E}(I_{\mathrm{f}}^{+})]_{\text{Policy 3}} \quad (11)$$

The benefit of pairing early fulfilment with make-to-order is evident in this inequality. The backorder levels resulting from the optimal configurations of the policies do not differ significantly. A substantial reduction in FG inventory volume is possible, however, using Policy 3. This finding is in agreement to that of Karaesman et al. (2004), who consider early fulfillment using a single demand class and conclude that the benefit of early fulfillment is reduced inventories that result from advance order deliveries. However, as the cost of WIP is considered here, an order-base-stock policy ($K = \infty$) is non-optimal. The results indicate that only a small amount of reserve stock ($R^* \in [1,3]$) is required for optimally fulfilling urgent orders: the remaining inventory is then made available for early fulfillment. When $\tau_1 = 0$, no ADI exists and hence there is no need for rationing. Table 7 shows the average order fill rate (fraction of demands fulfilled from the on-hand inventory), WIP, and FG inventory for the optimal configuration of each policy. It is observed that the key advantage of our proposed policy is a much lower average FG inventory compared with the other policies (when $\tau_1 > 0$), while the WIP and the fill rate are not greatly affected. As the overall cost is a function of the average WIP, FG, and backorders, the proposed policy emerges as superior to the competing policies, particularly at high system utilisation with longer demand lead time.

A comparison of total cost (Figure 8, Figure 9 and Figure 10) also depicts that early fulfillment is beneficial wherever ADI is available ($\tau_1 > 0$). Policy 1 does not use early fulfillment and performs the poorest in terms of total cost. In other words, it can be conjectured that early fulfillment with one-for-one replenishment is more beneficial than make-to-order for taking advantage of available ADI, and the relative benefit is increasing with τ_1. The benefits are even greater, however, when early fulfillment is combined with a suitable make-to-order strategy.

The plots show that the relative performance of the policies remains unchanged when the cost factor is changed (the b/h ratio is doubled in the second set of comparisons in Figure 9, and is halved in Figure 10) within the given range. In the absence of ADI ($\tau_1 = 0$), performance of all policies is the same (except for Policy 2, which has marginally higher cost due to its inability to make items to-order). As τ_1 increases, the benefit of the combined strategy (Policy 3) becomes more prevalent. This benefit is observed irrespective of the order arrival rates or cost factors used in the setting. Karaesmen et al. (2003) were not able to find the benefit of ADI using a make-to-order approach under high system utilizations. Claudio and Krishnamurthy (2009) also indicate *diminishing returns* of ADI as demand lead time increases. The experimental results for Policy 1 confirm both of these observations, as the total cost decreases slowly (especially under high system utilization) as τ_1 increases. In contrast, Policy 3 achieves a significant cost reduction in proportion to the amount of ADI available. At λ = 0.8, increasing τ_1 from 0 to 0.5LT, 0.5LT to 1LT and 1LT to 2LT reduces total cost by 8.7%, 9.2% and 17%, respectively (w: h: b = 1:2:20). At this cost factor combination and τ_1 = LT, an average cost reduction of 16.33% is observed for Policy 3 in comparison to Policy 1. The cost reduction becomes 29.21% when τ_1 = 2LT. Compared to Policy 2, Policy 3 achieves a cost reduction of 3.97% (at τ_1 = LT) and 8.76% (at τ_1 = 2LT). Results are similar using other cost factor combinations. No other policy in literature outperforms the proposed policy by using ADI in the context of production-inventory systems. The results presented here thus show a performance improvement over the existing policies. An average cost reduction of 3% is observed for Policy 3 in comparison to the better of the other two policies, considering all settings ($0 \leq \tau_1 \leq$ 2LT). For any firm, even a small percentage saving in cost may translate into a large percentage rise in profit. The optimality gap (defined as relative increase in total cost for Policy 3 in comparison to the lower bound) is found to be 2.89% on average. Although this gap is higher than that obtained by Wang and Toktay (2008) in the inventory system with constant replenishment time (which is 2% on average), the performance of Policy 3 seems close enough to the lower bound policy, which is difficult anyway to implement in reality. The outcome of these experiments indicates the superiority of combining make-to-order and rationing.

From the above sets of experiments, the performance of Policy 2 is observed to fall in between that for other two policies. Additionally, Policy 3 is an improvement over both Policy 1 and Policy 2 as it combines the benefits of both these approaches together. However, Policy 1 is suitable when early fulfilment is not allowed. In other words, depending on whether early fulfilment is allowed, one of Policy 1 and 3 is desirable. Therefore, from now onwards, performance measure of Policy 2 is discontinued. The comparison thus obtained, reflects the incremental benefits of allowing early fulfillment in given conditions (Experiment Set 4–5).

5.2 Experiment Set 4

Table 8 Shows the control parameter values, and Figure 11 shows the effect of changing the arrival ratio on the total cost for a range of system utilizations. Here arrival ratio denotes the fraction of orders providing ADI among all orders.

Table 8. Results for Experiment Set 4.

λ	r (%)	Policy 1 parameters and states		Policy 3 parameters and states	
		K^*, Z^*	E(I_w), E(I_f^+), E(I_f^-)	K^*, Z^*, R^*	E(I_w), E(I_f^+), E(I_f^-)
0.5	0	10, 6	3.0, 3.2, 0.21	10, 6, 0	3.0, 3.2, 0.21
0.5	25	10, 6	3.0, 3.9, 0.13	10, 6, 0	2.8, 3.7, 0.12
0.5	50	10, 5	3.0, 3.6, 0.14	10, 5, 1	3.0, 2.4, 0.16
0.5	75	10, 4	3.0, 3.4, 0.13	11, 4, 1	3.0, 1.8, 0.14
0.5	100	10, 3	2.9, 3.1, 0.13	10, 3, 1	3.0, 1.3, 0.12
0.6	0	15, 9	4.4, 5.0, 0.24	15, 9, 0	4.4, 4.8, 0.24
0.6	25	15, 8	4.4, 4.9, 0.23	15, 8, 2	4.4, 4.0, 0.22
0.6	50	15, 7	4.5, 4.9, 0.22	15, 7, 1	4.5, 3.1, 0.24
0.6	75	12, 5	4.4, 4.1, 0.24	12, 5, 1	4.4, 1.8, 0.25
0.6	100	12, 4	4.3, 4.2, 0.21	12, 4, 1	4.4, 1.5, 0.21
0.7	0	23, 14	6.9, 7.4, 0.32	23, 14, 0	6.9, 7.4, 0.30
0.7	25	24, 12	6.9, 7.3, 0.31	24, 12, 2	6.9, 5.7, 0.32
0.7	50	22, 11	7.1, 7.3, 0.31	22, 11, 2	7.1, 4.8, 0.32
0.7	75	20, 8	6.9, 6.6, 0.31	20, 8, 1	6.9, 2.7, 0.33
0.7	100	21, 6	6.9, 6.3, 0.30	20, 6, 1	6.9, 1.8, 0.31
0.8	0	32, 23	11.7, 11.7, 0.55	32, 23, 0	11.7, 11.7, 0.50
0.8	25	30, 20	11.7, 11.7, 0.53	30, 20, 2	11.7, 9.2, 0.48
0.8	50	29, 19	11.7, 10.7, 0.52	28, 20, 2	11.7, 8.8, 0.45
0.8	75	30, 13	11.7, 10.6, 0.52	30, 13, 1	11.7, 5.2, 0.44
0.8	100	28, 10	11.6, 10.5, 0.52	28, 10, 1	11.7, 3.4, 0.42

r, arrival ratio; Other notations are same as those in Table 4. See Table 3 for experimental design.

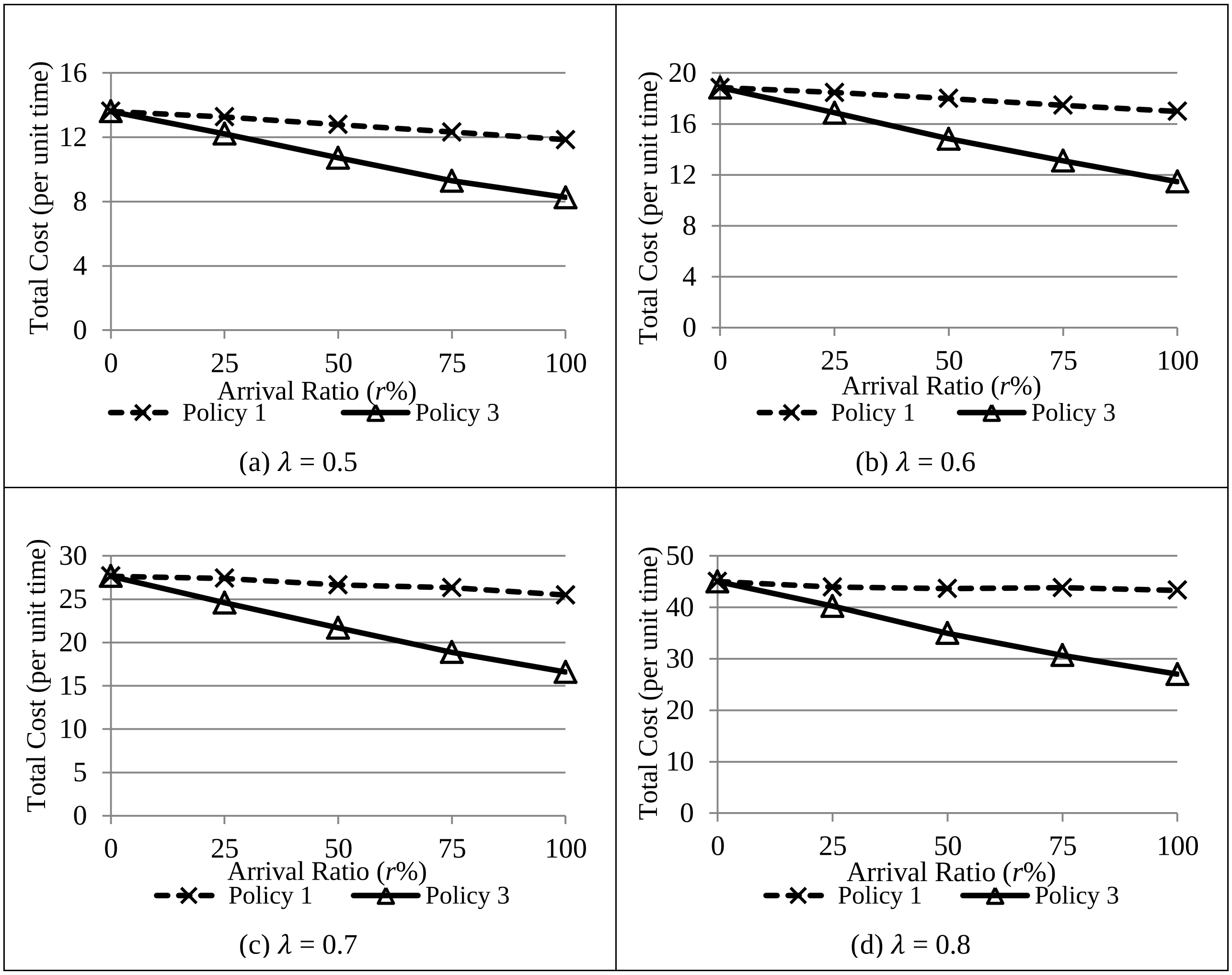


Figure 11. Variation of total cost with a change in arrival ratio (Experiment Set 4).

It can be observed for all system utilization values that the difference between Policy 1 and 3 widens as the arrival ratio increases. When all orders provide ADI (100% arrival ratio), the incremental benefits from early fulfillment varies from 30% cost reduction (for $\lambda = 0.5$) to 37.6% (for $\lambda = 0.8$).

Another interesting observation from this set of experiments is that the total cost of managing inventory reduces significantly when only a fraction of the orders provide ADI and early fulfillment is allowed for them. It turns out for each of the above plots that the cost is lower even when a quarter of the orders provide ADI with early fulfillment than all orders providing ADI with no early fulfillment. A managerial insight is that convincing a fraction of the customers to accept early fulfillment may worth more than convincing all customers to provide ADI.

5.3 Experiment Set 5

Results of changing the relative backorder costs are shown in Figure 12. Table 9 shows optimal control parameter values for three different backorder cost combinations at various system utilizations. Since the backorder costs for the demand classes are different, the class with higher backorder cost is the priority class. A static rationing level (R_1) is maintained, such that whenever the inventory level comes down to R_1, only the demands from the priority class is fulfilled. Another static rationing level R_2 determines early fulfilment of ADI orders. R_2 is applicable for Policy 3 only.

Table 9. Optimal control parameter values for Experiment Set 5.

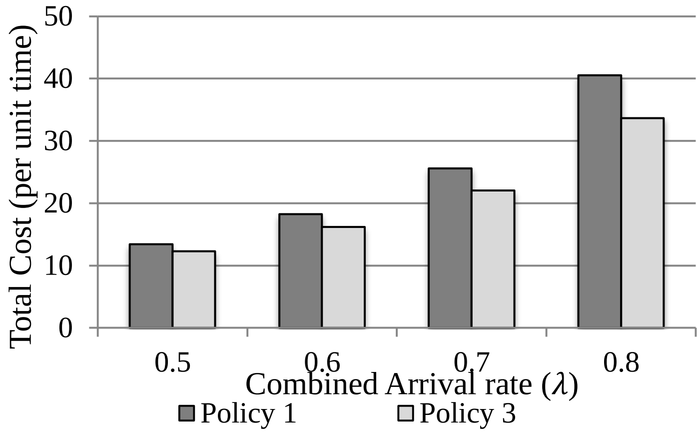
λ	b_1= 10, b_2= 100		b_1=10, b_2=10		b_1= 100, b_2= 10	
	Policy 1	Policy 3	Policy 1	Policy 3	Policy 1	Policy 3
	K^*, Z^*, R_1^*	K^*, Z^*, R_1^*, R_2^*	K^*, Z^*, R_1^*	K^*, Z^*, R_1^*, R_2*	K^*, Z^*, R_1^*	K^*, Z^*, R_1^*, R_2^*
0.5	9, 5, 1	8, 6, 1, 4	11, 3, 0	10, 4, 0 1	9, 5, 1	10, 5, 0, 0
0.6	14, 7, 1	12, 8, 1, 4	12, 5, 0	13, 6, 0, 1	13, 7, 1	13, 7, 1, 2
0.7	18, 10, 2	19, 11, 2, 4	17, 7, 0	20, 9, 0, 1	19, 9, 2	18, 10, 1, 2
0.8	29, 15, 2	29, 17, 2, 5	23, 13, 0	25, 14, 0, 1	29, 15, 2	29, 17, 2, 3

λ, order arrival rate; b_1 and b_2 are backorder costs for Class 1 and Class 2 orders respectively; K^*, Z^*, R_1^* and R_2^* are optimal values of respective policy parameters to minimize total cost. See Table 3 for experimental design.

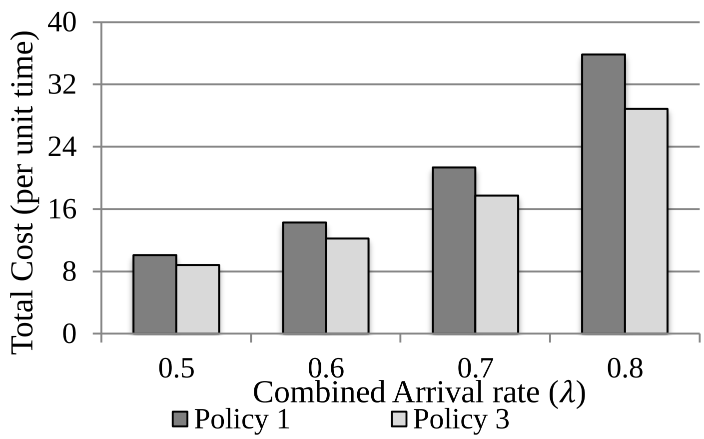
One important finding of these experiments is that the early fulfillment is beneficial even when no-ADI class is the priority class. For Policy 3, early fulfillment happens whenever R_2 is less than Z. It is observed from the optimal parameter values that R_2^* is positive when either of the demand classes is a priority class. However, actual level of early fulfilment depends on the type of the priority class (i.e., whether the class is ADI providing). When a backordering of no-

ADI demand is ten times more expensive that a backordering of ADI demand, still moderate amount of early fulfillment takes place for the optimal design alternatives (evidenced by the gap between R_2^* and Z^*). This gap widens when the demand classes have equal priority, or the ADI-class is at higher priority. When both demand classes have equal priority (b_1= 10, b_2= 10), R_1^* is obviously zero. R_1 takes positive values when either demand class becomes the priority class.

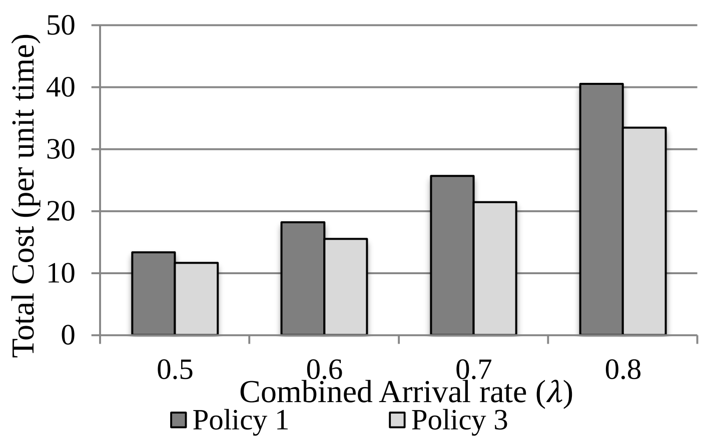
In Figure 12, total cost of managing inventory is appeared more for (a) or (c) than in (b), as the backordering cost for one class is higher than the other. Costs of using Policy 3 in (c) are marginally smaller than corresponding costs in (a) because of the advantage of obtaining ADI from higher priority class. This difference is more prominent for the costs associated with Policy 1. This observation is parallel to the outcome of Koçağa and Şen (2007), where incorporation of demand lead time is shown to be more valuable if the ADI providing class is the higher priority class. Significant difference between the costs of Policy 1 and Policy 3 for all conditions in (a), (b) and (c) indicates the robustness of early fulfillment irrespective of the relative priority of the demand classes. At high system utilization, a rapid increase in total cost is observed for all policies, indicating less benefit of ADI at high utilization.



(a) b_1=10, b_2=100



(b) b_1=10, b_2=10



(c) b_1=100, b_2=10

Figure 12. Optimal cost comparison of two policies (Experiment Set 5).

The remaining experiments (Sets 6–9) focus exclusively on early fulfillment. Policy 1, which does not use early fulfillment, is not further considered. These experiments show the performance-sensitivity of Policy 3 with a change in the backorder cost, an increase in the demand lead time, and an equal amount reduction in the supply lead time.

5.4 Experiment Set 6

Table 10 shows the effect of change in arrival rate when all orders provide ADI, but only a fraction of the orders accept early fulfillment (Scenario 2). Here arrival ratio indicates the fraction of orders that accept early fulfillment among all orders. Figure 13 shows the corresponding cost variations.

Table 10. Results for Experiment Set 6.

λ	r (%)	b_1= 10, b_2= 100	
		K^*, Z^*, R_1^*, R_2*	E(I_w), E(I_f^+), E(I_{f1}^-), E(I_{f2}^-)
0.5	25	9, 5, 1, 4	2.8, 4.4, 0.03, 0.022
0.5	50	9, 4, 1, 3	2.9, 3.2, 0.09, 0.014
0.5	75	10, 3, 1, 2	2.9, 2.0, 0.19, 0.006
0.6	25	11, 6, 1, 4	4.2, 5.1, 0.07, 0.044
0.6	50	11, 5, 1, 3	4.3, 3.6, 0.16, 0.023
0.6	75	11, 5, 1, 2	4.3, 2.8, 0.19, 0.006
0.7	25	17, 9, 1, 4	6.7, 7.6, 0.09, 0.049
0.7	50	15, 9, 1, 3	6.6, 6.2, 0.16, 0.018
0.7	75	17, 6, 1, 2	6.8, 2.9, 0.38, 0.010
0.8	25	24, 14, 1, 4	11.1, 11.2, 0.23, 0.090
0.8	50	23, 11, 2, 4	11.2, 7.0, 0.56, 0.020
0.8	75	27, 11, 1, 2	11.5, 4.8, 0.48, 0.010

r, arrival ratio; E(I_{f1}^-) and E(I_{f2}^-) are expected backorder levels for Class 1 and Class 2 orders. Other notations are same as those in Table 4 and Table 9. See Table 3 for experimental design.

It is observed from the plot that the cost reduction is almost linear with the increase in arrival ratio, particularly for low to moderate system utilization. At high system utilization, the incremental benefit of advance demand information reduces with an increase in arrival ratio. This is expected, as an acceptable service level of Class 2 demands makes early fulfillment of Class 1 demands less frequent at high system utilization.

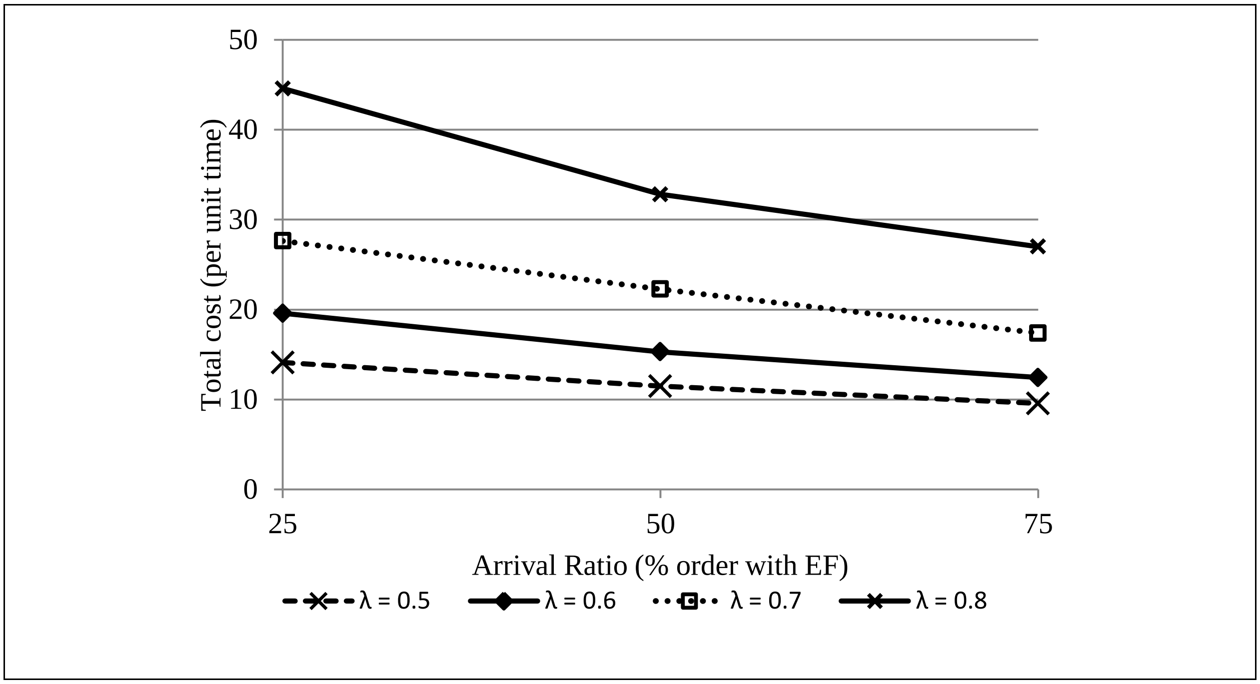


Figure 13. Variation of total cost with a change in arrival ratio (Experiment Set 6).

5.5 Experiment Set 7

Table 11 and Figure 14 show the results of experiments conducted for two different backorder cost combinations: (i) b_1= 10, b_2 = 100; (ii) b_1= 100, b_2 = 10. In (i), early fulfillment is not allowed for the higher priority class, whereas in (ii) that class accepts early fulfillment.

Table 11. Results for comparison of two policies (Experiment Set 7).

λ	No EF for Priority Class (b_1= 10, b_2 = 100)		EF allowed for Priority Class (b_1= 100, b_2 = 10)	
	K^*, Z^*, R_1^*, R_2*	E(I_w), E(I_f^+), E(I_{f1}^-), E(I_{f2}^-)	K^*, Z^*, R_1^*, R_2*	E(I_w), E(I_f^+), E(I_{f1}^-), E(I_{f2}^-)
0.5	9, 4, 1, 3	2.9, 3.2, 0.09, 0.014	11, 3, 0, 0	3.0, 1.9, 0.003, 0.26
0.6	11, 5, 1, 3	4.3, 4.0, 0.13, 0.016	13, 4, 0, 0	4.4, 2.4, 0.004, 0.42
0.7	15, 9, 1, 3	6.6, 6.2, 0.15, 0.018	18, 7, 0, 0	6.8, 4.3, 0.007, 0.48
0.8	23, 11, 2, 4	11.2, 7.0, 0.55, 0.020	23, 11, 1, 2	11.3, 6.6, 0.031, 0.51

E(I_{f1}^-) and E(I_{f2}^-) are expected backorder levels for Class 1 and Class 2 orders. Other notations are same as those in Table 4 and Table 9. See Table 3 for experimental design.

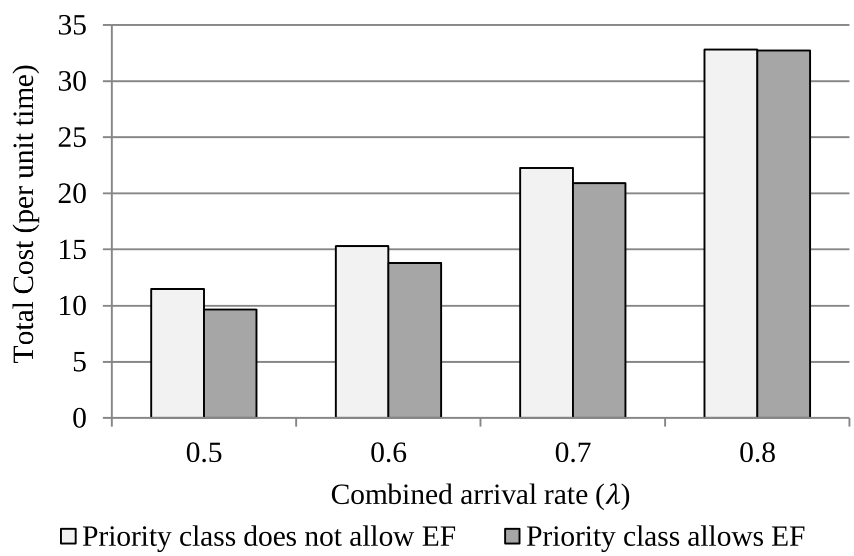


Figure 14. Change in total cost with the arrival ratio (Experiment Set 7).

Two important observations can be made from this set of experiments: (1) early fulfillment is most beneficial, when it is associated with the higher priority class (which is evident from the cost difference in the above plot), and (2) this cost difference reduces with an increase in system utilization. These are expected, as similar observations were made in Experiment Set 5.

5.6 Experiment Set 8 and 9

In Experiment Set 8, the effect of change in demand lead time is observed for different system utilizations. In Experiment Set 9, μ is adjusted such that the average supply lead times are reduced by the same amounts, as the demand lead time increments in the preceding set of experiments. Results are shown in Table 12 and Figure 15.

Table 12. Control parameter values and performance measures in Experiment Sets 8 and 9.

λ	E(SLT)= DLT (both equal to LT)		DLT increased by 15%		E(SLT) reduced by 15%	
	K^*, Z^*, R_2^*	$E(I_w), E(I_f^+), E(I_f^-)$	K^*, Z^*, R_2^*	$E(I_w), E(I_f^+), E(I_f^-)$	K^*, Z^*, R_2^*	$E(I_w), E(I_f^+), E(I_f^-)$
0.5	14, 3, 1	2.9, 2.0, 0.08	14, 3, 1	3.0, 2.2, 0.06	15, 2, 1	2.5, 1.5, 0.07
0.6	17, 5, 1	4.4, 3.3, 0.09	17, 4, 1	4.4, 2.7, 0.09	16, 3, 1	3.8, 2.1, 0.09
0.7	17, 7, 2	6.8, 4.4, 0.18	18, 6, 2	6.8, 4.1, 0.17	18, 5, 2	5.8, 3.6, 0.12
0.8	22, 12, 2	11.2, 7.3, 0.33	24, 10, 2	11.4, 6.4, 0.33	25, 8, 2	9.7, 5.4, 0.23

Notations are same as those in Table 4 and Table 9. See Table 3 for experimental design.

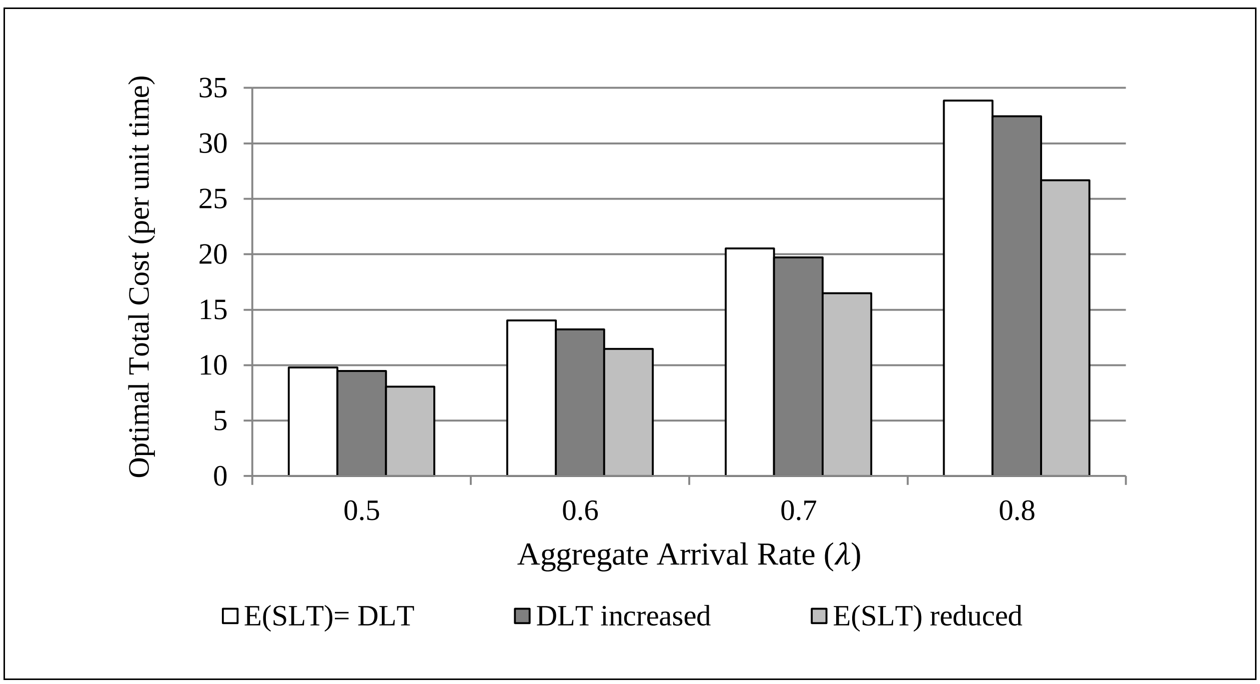


Figure 15. Effect of increasing the demand lead time and reducing the supply lead time.

For pure inventory systems, supply lead times are load-independent. For such systems with no early fulfillment, Hariharan and Zipkin (1995) shows that increasing the demand lead time has the exact same effect as reducing the supply lead time by an equal amount. Later, Wang and Toktay (2008) shows that allowing early fulfillment breaks this equality: increasing the demand lead time brings greater benefit than reducing the supply lead time by the same amount, as fulfillment flexibility increases in the first case. However, results in the current research are in contrast to the pure inventory systems, as increasing the demand lead time is observed to be worse than reducing the supply lead times. Outcome of these sets of experiments proves that the claim by Wang and Toktay in the context of pure inventory systems does not necessarily hold for production-inventory systems. This is so because of the queuing effect in the production-inventory systems (supply lead times are load-dependent).

6. Pure Inventory Systems: Analytical Investigation

6.1 Classical System

As described in Section 3.2, an analytical approach is used for characterizing the optimal policy in Scenario 3. The model detailed in this section is an extension of the classical periodic review pure inventory model to accommodate multiple demand class and advance demand information. The classical model (e.g., Porteus, 2002) is first outlined, which will be helpful later to characterize the current model.

This is a univariate system, where the state of the system can be completely defined by the inventory position x (thus, the state space is a real line). The decision (action) is conveniently represented by the level of inventory after ordering (y). The following notations are used:

C_{H} = unit holding cost per period (C_{H}>0)

c = unit procurement cost

C_{P} = unit shortage cost per period ($C_{\mathrm{P}} > C_{\mathrm{H}}$)

The expected loss function (i.e., one-period holding and shortage cost function) can be expressed in terms of y as follows:

$$\mathrm{L}(y) \coloneqq \int_{\xi=0}^{y} C_{\mathrm{H}}(\mathrm{y}-\xi)\phi(\xi)\mathrm{d}\xi + \int_{\xi=\mathrm{y}}^{\infty} C_{\mathrm{P}}(\xi - y)\phi(\xi)\mathrm{d}\xi \tag{12}$$

Then the optimality equation for this model becomes:

$$\mathrm{f_t}(x) = \min_{y \geq x}\left\{\mathrm{c}(y-x) + \mathrm{L}(y) + \alpha\int_{0}^{\infty} \mathrm{f_{t+1}}(y-\xi)\phi(\xi)\mathrm{d}\xi\right\} \tag{13}$$

The last period cost is assumed convex. It can be reasonably assumed that the backlogged demands at the end of period N are met immediately (at usual cost), but the excess inventory has no value. Therefore, the cost function is:

$$v_{\mathrm{T}}(x) = \begin{cases} -cx & \text{if } x \leq 0 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

which is a convex function.

Equation (13) can be expressed in terms of the decision variable y:

$$\mathrm{f_t}(x) = \min_{y \geq x} \{\mathrm{G_t}(y) - cx\} \tag{15}$$

Where,

$$\mathrm{G_t}(y) = \left\{ cy + \mathrm{L}(y) + \alpha \int_0^{\infty} \mathrm{f_{t+1}}(\mathrm{y} - \xi)\phi(\xi)\mathrm{d}\xi \right\} \tag{16}$$

As the last period cost function is convex and linearity preserves convexity, G_{t} is shown to be convex and coercive using backward recursion.

The above formulation indicates that the optimal inventory replenishment policy is a base-stock policy. For example, if G_{t} is shown in Figure 16 and S_{t} is a minimizer of G_{t}, then the optimal decision is to move up (to the right) of S_{t}, if the inventory position x is below the base-stock level S_{t}.

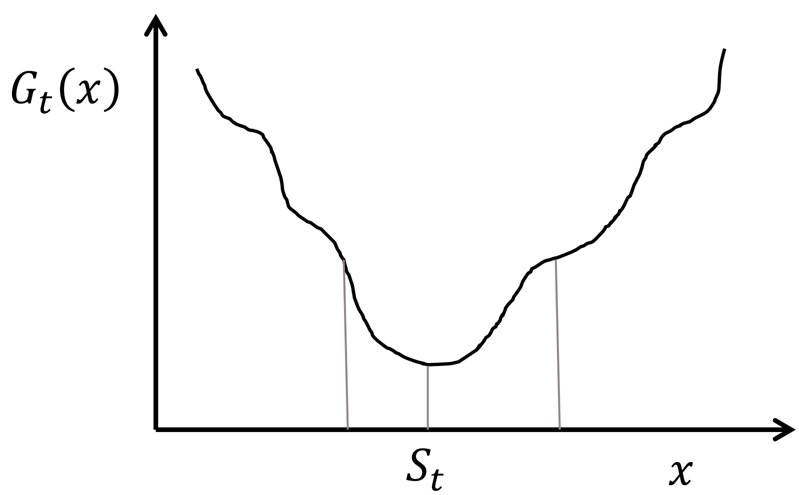


Figure 16. A case where a base-stock policy is the optimal.

6.2 Extension of the Classical Model to Accommodate ADI and Multiple Classes

When ADI and two demand classes are considered, inventory position alone is not sufficient to define the system state. At the beginning of any period t, the inventory manager takes replenishment order decision by considering all information available to her. These are: (a) the inventory level x_{t}, (b) replenishment orders placed since $(t-L)$ period, which are yet to be delivered, and (c) demands that need to be fulfilled in the current and future periods. The system state can be defined by these three variables together: $(x_{\mathrm{t}}, W_{\mathrm{t}}, V_{\mathrm{t}})$, where,

x_{t} = inventory level

W_{t} = supply pipeline

V_{t} = advance demand profile

W_{t} and V_{t} are vectors. $W_{\mathrm{t}} = (w_{\mathrm{t}},\ w_{\mathrm{t+1}},\ \ldots,\ w_{\mathrm{t+L-1}})$, where w_{t} is the replenishment order received in period t.

$$V_{\mathrm{t}} = \{V_{\mathrm{t,m}}, V_{\mathrm{t,n}}\} \tag{17}$$

$V_{\mathrm{t,m}}$ and $V_{\mathrm{t,n}}$ represent Class 1and Class 2 demand profiles respectively (m and n are suffixes to segregate demands from different classes).

$$V_{\mathrm{t,m}} = (v_{\mathrm{t,m}}^{\mathrm{t}}, v_{\mathrm{t,m}}^{\mathrm{t+1}}, \ldots, v_{\mathrm{t,m}}^{\mathrm{t+T_1-1}}) \tag{18}$$

$$V_{\mathrm{t,n}} = (v_{\mathrm{n}}^{\mathrm{t}}, v_{\mathrm{n}}^{\mathrm{t+1}}, \ldots, v_{\mathrm{n}}^{\mathrm{t+T_2-1}}) \tag{19}$$

Where, $v_{\mathrm{t,m}}^{\mathrm{k}}$ = Class 1 demands, which remain unfulfilled at the beginning of period t, and are due by period k. $v_{\mathrm{t,n}}^{\mathrm{k}}$ is similarly defined for Class 2 demand. Also, D_t denotes all orders arrived in period t. The term $d_{t,m}$ corresponds to Class 1 orders, whereas $d_{t,n}$ corresponds to Class 2 orders arrived in that period. Therefore, the following equations hold.

$$D_t = d_{t,m} + d_{t,n} \tag{20}$$

$$d_{t,m} = v_{t,m}^{t+T_1};\ d_{t,m} = v_{t,m}^{t+T_1} \tag{21}$$

The optimality equation (13) in classical case now becomes

$$f_{\mathrm{t}}(x_{\mathrm{t}}, W_{\mathrm{t}}, V_{\mathrm{t}}) = \min_{Q \geq z_{\mathrm{t}} \geq x} \{c(z_{\mathrm{t}}) + \mathrm{E}[\mathrm{L}(x_{\mathrm{t}}) + \alpha f_{\mathrm{t+1}}(x_{\mathrm{t+1}}, W_{\mathrm{t+1}}, V_{\mathrm{t+1}})]\} \tag{22}$$

Similar to the classical case, a global minimum exists if G_{t} can be shown as convex and coercive. However, convexity can be proved, when the evolution equations of state variables are predicted to be linear. Though W_t and $V_{\mathrm{t,n}}$ evolve linearly, the possibility of early fulfillment of class1 demands voids the guarantee of linear evolutions of x_{t} and $V_{\mathrm{t,m}}$.

In the following paragraphs, three different possibilities depending upon the values of T_1, T_2, and L are discussed. These three possibilities result in different outcomes for optimal policy characterization.

Case (*i*)*:* $T_1 \leq T_2 \leq L + 1$

Any type of demand arriving in period t is due before receiving the delivery of orders placed in that period. A schematic for the above condition is shown in Figure 17.

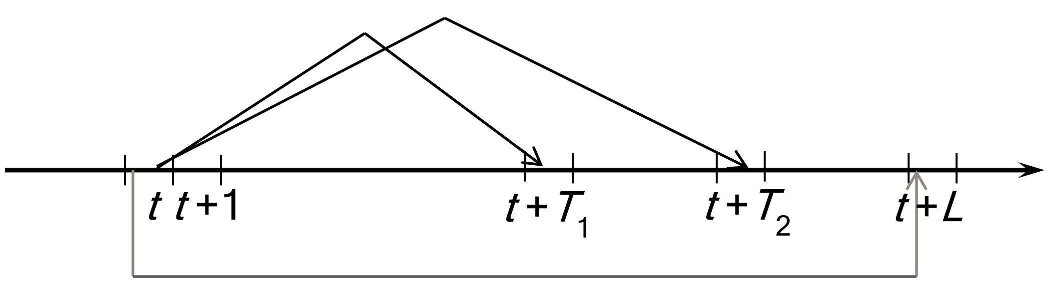


Figure 17. A time diagram for *Case* (*i*).

Case (*ii*)*:* $T_1 \leq L + 1 < T_2$

Supply lead time is more than the demand lead time for Class 1 orders, but less than that for Class 2 orders (Figure 18).

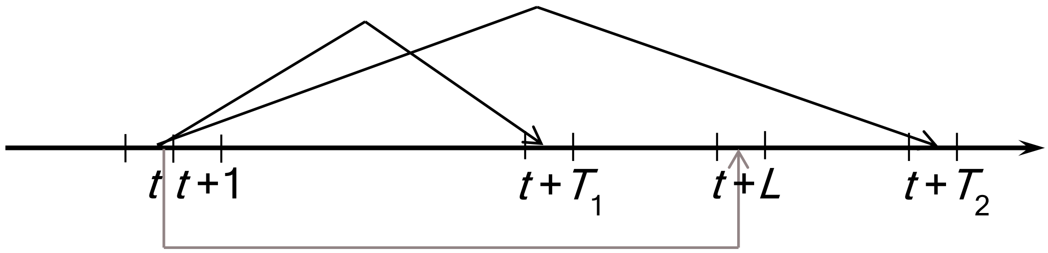


Figure 18. A time diagram for *Case* (*ii*).

Case (*iii*): $L + 1 < T_1 \leq T_2$

Supply lead time value is lower than the demand lead time for both order classes (Figure 19).

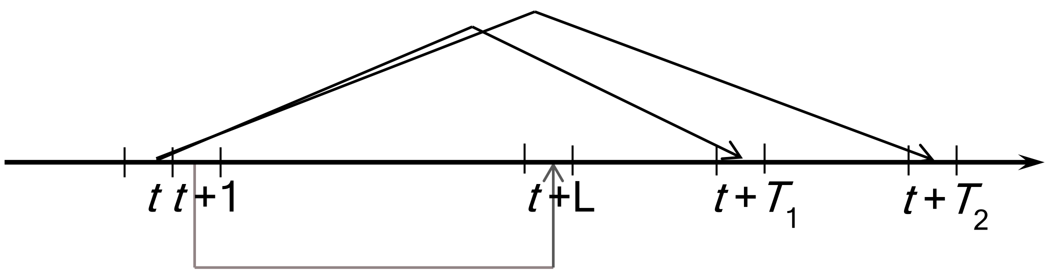


Figure 19. A time diagram for *Case* (*iii*).

Our characterization of the optimal policy is based upon two observations as described below:

Observation 1:

In the optimal policy, Class 1 demands are never early fulfilled at the expense of backordering of Class 2 demands.

This is intuitive as we have assumed ($C_P > C_H$), which typically holds in reality. Since the discount factor α = 1, preserving inventory to prevent future backordering always results in lower inventory cost than utilizing the same units in early fulfilment of current period demands.

Observation 2:

In the optimal policy, Class 1 demands are fulfilled in First-Come-First-Served basis.

This is also a straight forward observation, as the demands due sooner should be fulfilled (by either early or exact fulfillment) before fulfilling demands due later — which may reduce future backordering. There is no incentive to do otherwise.

Based upon the above observations, a notation $\hat{t}$ is introduced. At period t, the Class 1 demand profile till period (t + T_1) is known. The last $\hat{t}$ periods of these demands remain entirely unfulfilled (in period t). Therefore, v_m terms can be positive only from period (t +T_1-1- $\hat{t}$) to ($t + T_1$), which is a key observation for expressing the evolution of x.

An example of evaluating $\hat{t}$ is provided in Figure 20. Assume that T_1 =3. At period t, after fulfilling all demands in that period, it is better to early fulfill Class 1 demands due in period (t + 1) and period (t + 2). However, Class 1 demands due in period (t + 3) should not be fulfilled early in period t to avoid backordering of Class 2 demands in period (t + 1) and (t + 2). Therefore, last one period of Class 1 demand remains unfulfilled. Thus, $\hat{t}$ = 1 here.

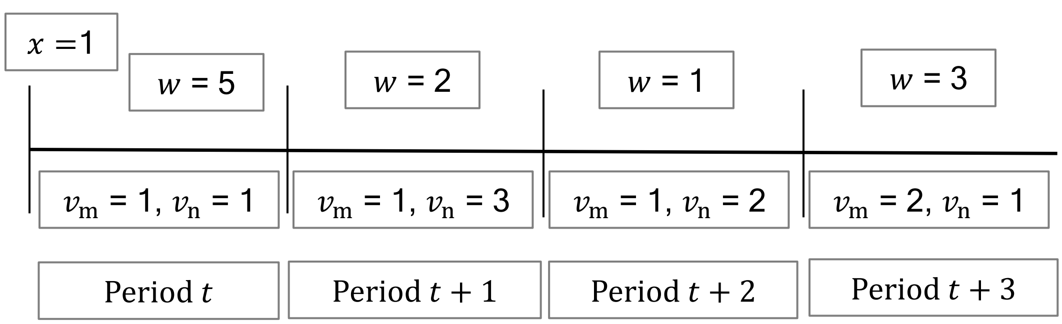


Figure 20. An example of finding $\hat{\boldsymbol{t}}$.

Then evolution of inventory level can be expressed as follows:

$$x_{t+1} = \left(x_t + w_t - \sum_{i=t}^{t+T_1-\hat{t}+1} v_{t,m}^{i} - v_n^t \right)^{+} - \left(x_t + w_t - v_{t,m}^{t} - v_n^t \right)^{-} \tag{23}$$

Where, $a^{+} = \max[a, 0]\,, a^{-} = \max[-a, 0]$

In the above formulation, $\hat{t}$ can be evaluated only when $T_1 \leq L + 1$. This indicates that characterization is possible for *Cases* (*i*) and (*ii*).

A new term — modified inventory position (as defined in Wang and Toktay, 2008) is used here:

$$u_t = \left(x_t + \sum_{i=t}^{t+L-1} w_i - \sum_{i=t}^{t+T_1-1} v_{t,m}^i - \sum_{i=t}^{t+T_2-1} v_n^i \right) \tag{24}$$

and

$$y_t = u_t + z_t \tag{25}$$

where u_t is the modified inventory position before ordering at period t, and y_t is that after ordering. The term z_t denotes the amount of replenishment order placed in period t. This modified inventory position incorporates ADI, so it is different from the standard definition of inventory position (thus the term *modified* is used). With the help of the modified inventory position, evolution of x_t can be expressed as the following.

Case (i):

$$x_{t+L+1} = \left(y_t - \sum_{i=t}^{t+L} d_{i,m} - \sum_{i=t}^{t+L-T_2} d_{i,n} + \sum_{i=t+L}^{t+L+T_1-\hat{t}} v_{t,m}^i \right)^+ - \left(y_t - \sum_{i=t}^{t+L-T_1} d_{i,m} - \sum_{i=t}^{t+L-T_2} d_{i,n} \right)^- \tag{26}$$

Case (ii):

$$x_{t+L+1} = \left(y_t - \sum_{i=t}^{t+L} d_{i,m} + \sum_{i=t+L}^{t+L+T_1-\hat{t}} v_{t,m}^i + \sum_{i=t}^{t+T_2-L-1} v_n^i \right)^+ - \left(y_t - \sum_{i=t}^{t+L-T_1} d_{i,m} + \sum_{i=t}^{t+T_2-L-1} d_{i,n} \right)^- \tag{27}$$

Case (iii):

As $\hat{t}$ is unknown, x_{t+L+1} is also unknown at period t.

Using the above expressions, the following proposition can be made:

Case (*i*)*:* $T_1 \leq T_2 \leq L + 1$

1. x_{t+L+1} can be expressed by MIP and upcoming demand.

2. $x_t - \sum v_m - \sum v_n$ evolves linearly, i.e., $u_{t+1} = u_t + w_t - D_t$.

3. A base-stock policy is optimal for this case.

Case (*ii*)*:* $T_1 \leq L + 1 < T_2$

1. x_{t+L+1} can be expressed by MIP, upcoming demand and $\sum_{i=t}^{t+T2-L-1} v_{i,n}$.

2. $u_{t+1} = u_t + w_t - D_t$ holds, but we need to remember $\sum_{i=t}^{t+T2-L-1} v_{i,n}$ terms.

3. A state-dependent base-stock policy is thus optimal.

Case (*iii*)*:* $L + 1 < T_1 \leq T_2$

1. As x_{t+L+1} is unknown by the above analysis, the optimal policy is difficult to find.

The above analysis suggests the optimality of a base-stock or state-dependent base-stock policy in some specific cases depending upon the demand lead time and supply lead time values.

7. Conclusions

In this research, an inventory management approach for satisfying demands with ADI has been developed. The approach is based upon early order fulfillment in the context of both production-inventory and pure inventory systems. Early fulfillment is shown to produce better inventory management policies in presence of multiple demand classes.

For production-inventory systems, two different scenarios were considered. In the first scenario, only a fraction of the demands provide ADI (and early fulfillment is acceptable for them). Three different policies were compared: (i) a free Kanban-based make-to-order policy (existing policy, termed as Policy 1), (ii) a policy using early fulfillment (Policy 2), and (iii) a new policy that combines the benefit of previous two policies (Policy 3). Unrestricted early fulfillment for one demand class is detrimental to the service level of the other demand class (which does not accept early fulfillment). Therefore, a balance in the form of rationing is required. A free Kanban-based policy starts production upon receiving ADI, thus helps to reduce average FG inventory. On the other hand, early fulfillment satisfied the order in the first opportunity, thus releasing the FG inventory. Therefore, these two approaches can work together from different angles to reduce inventory management cost. Thus, Policy 3 was found to consistently outperform the competing policies in a variety of settings, with a range of values for demand lead time, system utilization, cost-factor combination, arrival ratio and relative backorder costs. A simulation-based procedure for finding the lower bound on cost was also described. In the second scenario, the benefit of early fulfillment was investigated when all demands provide ADI, but only a fraction of the demands accepts early fulfillment. The outcome of the simulation experiments identified the benefit of using ADI along with an early fulfillment policy: this benefit can be significantly more valuable in comparison to both a conventional system with no ADI, and a system with ADI, but no early fulfillment. Utilizing early fulfillment is shown to improve performance even in high system utilizations, and the benefit is shown to be proportional to the demand lead time. Improvement from early fulfillment is so significant that receiving ADI (with early fulfillment allowed) from even a small fraction of customers seems to be better than all customers providing ADI, but none accepting early fulfillment. Also, in presence of two different priority classes with equal demand lead times and only one class accepting early fulfillment, this research found that inventory cost gets lower when the higher priority class accepts early fulfillment than the other class does it.

In pure inventory systems with two demand classes, the optimal replenishment policy is characterized for some specific cases. When the demand lead times of both classes are less than the supply lead time, the optimal replenishment policy is shown to be a base-stock policy. When the demand lead time of the class not accepting early fulfillment is more than the supply lead time, the optimal replenishment policy is shown to be a state-dependent base-stock policy.

This research on inventory management can be extended in several possible directions. One natural extension is consideration of more than two demand classes. Investigation of the combined benefit of early fulfilment with suitable postponement strategies is another possibility. Postponement strategies are effective for multiple product systems with some commonalities. Also, postponement helps the inventory manager to receive more information, which can be utilized to make better decisions on inventory allocation and shipping methods (see Xu et al., 2009). Another extension is to integrate ADI with pricing and revenue management. In reality, firms need to entice the customers in providing ADI (by some sort of price discount). Therefore, the discount level, order fulfillment and replenishment decisions should be considered together in the model to evaluate the benefit for the firm.

8. References

Banks, J., Carson II, J. S., Nelson, B. L. and Nicol, D. M., 2010. *Discrete-event system simulation*. Fifth Ed. Englewood Cliffs, NJ: Prentice Hall.

Bonvik, A. M., Couch, C. E. and Gershwin, S. B., 1997. A comparison of production-line mechanisms. *International Journal of Production Research*, 35(3), 789-804.

Cattani, K. D. and Souza G. C., 2002. Inventory rationing and shipment flexibility alternatives for direct market firm. *Production and Operations Management*, 11(4), 441-457.

Chen, C.-H., Lin, J., Yücesan, E., and Chick, S. E., 2000. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3): 251-270.

Chen, F., 2001. Market segmentation, advanced demand information, and supply chain performance. *Manufacturing & Service Operations Management*, 3(1), 53-67.

Choi, H.P., Blocher, J. D. and Gavirneni, S., 2008. Value of sharing production yield information in a serial supply chain. *Production and Operations Management*, 17(6), 614-625.

Claudio, D. and Krishnamurthy, A., 2009. Kanban-based pull systems with advance demand information. *International Journal of Production Research*, 47(12), 3139-3160.

Duran, S., Liu, T., Simchi-Levi, D. and Swann, J. L., 2007. Optimal production and inventory policies of priority and price-differentiated customers. *IIE Transactions*. 39(9), 845-861.

Duran, S., Liu, T., Simchi-Levi, D. and Swann, J. L., 2008. Policies utilizing tactical inventory for service-differentiated customers. *Operations Research Letters*. 36(2), 259-264.

Gallego, G. and Özer, Ö., 2001. Integrating replenishment decisions with advance demand information. *Management Science*, 47(10), 1344-1360.

Gallego, G. and Özer, Ö., 2003. Optimal replenishment policies for multiechelon inventory problems under advance demand information. *Manufacturing & Service Operations Management*, 5(2), 157-175.

Gayon, J., Benjaafar, S. and de Véricourt, F., 2009. Using imperfect advance demand information in production-inventory systems with multiple customer classes. *Manufacturing and Service Operations Management*, 11(1), 128-143.

Ha, A. H., 1997a. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science*, 43(8), 1093-1103.

Ha, A. H., 1997b. Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Research Logistics*, 44, 457-472.

Hadley, G. and Whitin, T.M., 1963. *Analysis of inventory systems*. Englewood Cliffs, NJ: Prentice-Hall.

Hariharan, R. and Zipkin, P., 1995. Customer-order Information, Leadtimes, and Inventories. *Management Science*, 41(10), 1599-1607.

Iravani, S. M. R., Liu, T., Luangkesorn, K. L. and Simchi-Levi, D., 2007. A produce-to-stock system with advance demand information and secondary customers. *Naval research Logistics*. 54, 331-345.

Kaplan, A., 1969. Stock rationing. *Management Science*, 15(5), 260-267.

Karaesmen, F., Buzacott, J. A. and Dallery, Y., 2002. Integrating advance order information in make-to-stock production systems. *IIE Transactions*, 34, 649-662.

Karaesmen, F., Liberopoulos, G. and Dallery, Y., 2003. Production/inventory control with advance demand information. In: J. G. Shanthikumar, D. D. Yao and W. H. M. Zijm, eds. *Stochastic modeling and optimization of manufacturing systems and supply chains*. The Netherlands: Kluwer Academic Publishers, 243-270.

Karaesmen, F., Liberopoulos, G. and Dallery, Y., 2004. The value of advance demand information in production/inventory systems. *Annals of Operations Research*, 126, 135-157.

Kim, J. H., Ahn, H. S. and Righter, R., 2009. Optimal production policies with multistage stochastic demand lead times. *Probability in the Engineering and Informational Sciences*. 23, 515-543.

Koçağa, Y. L. and Şen, A., 2007. Spare parts inventory management with demand lead times and rationing. *IIE Transactions*, 39, 879- 898.

Krishnamurthy, A. and Ge, D., 2009. Integrating kanban control with advance demand information: insights from an analytical model. In: Gerald Reiner, ed. *Rapid modeling for increasing competitiveness*. London: Springer, 131-141.

Lee, L. H., Chew, E. P., Teng, S., and Goldsman, D., 2004. Optimal computing budget allocation for multi-objective simulation models. In: R. G. Ingalls, M.D. Rosetti, J. S. Smith, and B. A. Peters, eds. *Proceedings of the 2004 Winter Simulation Conference*. Piscataway, New Jersey, 586-594.

Liberopoulos, G. and Koukoumialos, S., 2005. Tradeoffs between base stock levels, numbers of kanbans, and planned supply lead times in production/inventory systems with advance demand information. *International Journal of Production Economics*, 96, 213-232.

Liberopoulos, G., 2008. On the tradeoff between optimal order base-stock levels and demand lead-times. *European Journal of Operational Research*. 190, 136-155.

Lu, Y., Song, J. S. and Yao, D., 2003. Order fill rate, leadtime variability, and advance demand information in an assemble-to-order system. *Operations Research*. 51(2), 292-308.

Özer, Ö., 2003. Replenishment strategies for distribution systems under advance demand information. *Management Science*. 49(3), 255-272.

Özer, Ö. and Wei, W., 2004. Inventory control with limited capacity and advance demand information. *Operations Research*, 52(6), 988-1000.

Porteus, E. L., 2002. *Foundations of Stochastic Inventory Theory*. First Ed. Stanford, CA: Stanford Business Books.

Pujowidianto, N. A., Lee, L. H., Chen, C.-H., Yap, C. M., 2009. Optimal computing budget allocation for constrained optimization. In: M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds. *Proceedings of the 2009 Winter Simulation Conference*. Piscataway, New Jersey, 584-589.

Scarf, H., 1960. The optimality of (*s*, *S*) policies in dynamic inventory problems. In: K. Arrow, S. Karlin and P. Suppes, eds. *Mathematical Methods in the Social Sciences*. Standford University Press, Stanford, California.

Sarkar, S., 2007. *The effect of advance demand information on a pull production system with two customer classes.* Master's Thesis. Virginia Polytechnic Institute and State University.

Sarkar, S. and Shewchuk, J. P., 2012. Use of advance demand information in multi-stage production-inventory systems with multiple demand classes. *International Journal of Production Research*. In press.

Simchi-Levi,D., 2010 *Operations Rules: Delivering Customer Value Through Flexible Operations*. The MIT Press, Massachusetts.

Tan, T., Güllü, R. and Erkip, N., 2009. Using imperfect advance demand information in ordering and rationing decisions. *International Journal of Production Economics*. 121, 665-677.

Teunter, R. H. and Haneveld, W. K. K., 2008. Dynamic inventory rationing strategies for inventory systems with two demand classes, Poisson demand and backordering. *European Journal of Operational Research*, 190 (1), 156-178.

Toktay, L. B. and Wein, L. M., 2001. Analysis of a forecasting-production-inventory system with stationary demand. *Management Science*, 47 (9), 1268-1281.

Topkis, D. M., 1968. Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Management Science*, 15 (3), 160-176.

Veinott, A. F. Jr., 1965. Optimal policy in a dynamic, single product, nonstationary inventory model with several demand classes. *Operations Research*, 13(5), 761-778.

Wang, T. and Toktay, L. B., 2008. Inventory management with advance demand information and flexible delivery. *Management Science*, 54(4), 716-732.

Wang, Y., Cohen, M. A. and Zheng, Y., 2002. Differentiating customer service on the basis of delivery lead-times. *IIE Transactions*, 34, 979-989.

Wang, H. and Yan, H., 2009. Inventory management for customers with alternative lead times. *Production and Operations Management*, 18(6), 705-720.

Wijngaard, J. and Karaesmen, F., 2007. Advance demand information and a restricted production capacity: on the optimality of order base-stock policies. *OR Spectrum*. 29, 643-660.

Woodward, R., 2000. *Maestro: Greenspan's Fed and the American Boom*. Simon and Schuster, New York.

Xu, P. J., Allgor, R. and Graves, S. C., 2009. Benefits of reevaluating real-time order fulfillment decisions. *Manufacturing & Services Operations Management*. 11, 340-355.